

Application of Indigenous Knowledge, Machine Learning and
Satellite Imagery to Optimize Cropping Decisions by Small-Scale
Farmers: Case Study of uMgungundlovu District Municipality, South
Africa

JOHN MAKHETHA NYETANYANE

Thesis submitted in fulfilment of the requirements for the Degree

**DOCTOR OF PHILOSOPHY
IN INFORMATION TECHNOLOGY**

in the

Department of Information Technology
Faculty of Engineering, Built Environment and Information Technology

at the

Central University of Technology, Free State

Promoter: Prof. Muthoni Masinde

Co-Promoter: Prof Mabhaudhi Tafadzwanashe

BLOEMFONTEIN

26/09/2024

Copyright Notice

This doctoral thesis is to be used only for academic or non-commercial research purposes. The information contained in this thesis is to be published with acknowledgement of the source.

This doctoral thesis is published by the Central University of Technology, Free State, in terms of a non-exclusive licence granted to CUT by the author.

Disclaimer

The presentation of this thesis contains colour images meant to present the results visually in a more understandable form. While printing or viewing the thesis in greyscale (black and white) is possible, it is recommended that for clarity, the thesis is viewed (or printed) in full-colour format.

Dedication

This thesis is dedicated to my mother, brother and my late father for their unconditional support.

Declaration

This research as presented in this thesis is my original work and has not been presented for any other university award. Knowledge derived from other sources has been indicated, with acknowledgement and reference to the literature.

This study was conducted and completed under the guidance of Professor Muthoni Masinde, Department of Information Technology at Central University of Technology, Free State, South Africa, and co-supervised by Professor Mabhaudhi Tafadzwanashe, from the University of Kwazulu-Natal(UKZN) Department of Agricultural Science.

John Makhetha Nyetanyane

Signature:

Date: **26/09/2024**

In our capacity as the supervisors of this thesis, we certify that the above statements are true to the best of our knowledge.

Professor Muthoni Masinde

Signature: _____

Date: _____

Professor Mabhaudhi Tafadzwanashe

Signature: _____

Date: _____

Acknowledgements

This thesis stems from a PhD research study conducted between January 2019 and September 2024 at the Department of Information Technology, within the Faculty of Engineering, Built Environment, and Information Technology, at the Central University of Technology, Free State, South Africa. Its completion would not have been possible without the support of several individuals, to whom I extend my deepest gratitude.

First and foremost, I would like to thank Prof Masinde for the support and guidance through this long journey. This achievement would not have been possible without the role she played.

I would also like to extend my heartfelt thanks to my co-supervisor, Prof. M. Tafadzwa, for being a pillar of support and providing me with an opportunity to meet farmers in Swayimane region in uMgungundlovu. I would also like to pass my thanks to Khethiwe Naledi Mthethwa and her team for the support they offered me during my visits to KZN. I would also like to thank the farmers in Swayimane region for their warm welcome and the knowledge that they have shared with me.

I am also deeply grateful to my friends and colleagues who made the university feel like a home away from home. In particular, I would like to thank Mr. Moabi Kompi and Mr. Molefi Motsoenyane for their words of encouragement and wisdom. My heartfelt thanks also go to Ms Mpho Mbele, who has been a constant source of support throughout this research journey from start to finish.

I am deeply thankful to my brother, Nkopane Isaac Nyetanyane, my mother, Mpotseng Augustina Matsaisa, and my late father, Abraham Tsietsi Nyetanyane, for their unwavering support and assistance during my research period.

Finally, I express my deepest and wholehearted gratitude to the Almighty God and my ancestors for their loving kindness throughout my life.

Abstract

This thesis presents the methodology to optimise cropping decisions for small-scale farmers who depend primarily on rainfed agriculture for subsistence and market production. The research was motivated by the significant role small-scale farmers, particularly in sub-Saharan Africa (SSA), play in ensuring food security, despite often being overlooked by many technological developments and advancements. Although the use of chemicals to grow crops seems to expand in Africa, many smallholder farmers continue to use the Indigenous Knowledge (IK) system that by its originality is natural, organic and environmentally friendly, and most importantly food produced by using only IK methods is of high quality. Although IK is a crucial resource for small-scale farmers in food production and managing climate-related challenges, its value, accuracy, and presence have diminished over time due to climate change, market competition, deforestation, pollution, modernisation, and many more.

This thesis aims to restore and enhance the dignity and effectiveness of indigenous knowledge by integrating it with scientific data and methods to optimise cropping decisions for small-scale farmers. This integration seeks to improve farmers' traditional methods to predict rainfall patterns and temperature fluctuations throughout the agricultural season. This integration is further expanded to strengthen farmers' conventional methods of monitoring the health and growth of the crops. While indigenous and scientific knowledge systems have their strengths and limitations, a significant body of literature highlights their collaboration's benefits in enhancing the resilience of local communities to climate change. However, not much has been done to incorporate these knowledge domains together to optimise food production given that one is qualitative and locally based while the other is quantitative and generic.

In this research project, efforts are made to quantify the IK system farmers use to predict rainy season behaviour and monitor the health and growth of crops, then integrate it with scientific data and approaches (satellite imagery, climate data, and machine learning models) to come up with more robust and local-based information that farmers can use to improve their cropping decisions. These knowledge systems are integrated into a mobile-based technology designed for farmers' use. This hybrid technology is trained to recommend crops a farmer can grow based on anticipated seasonal behaviour. It is then evaluated using a dataset with unknown inputs, achieving an accuracy level of 83.3%.

List of Acronyms

ACF- Auto Correlation Function

ADF- Augmented Dickey-Fuller

ADTS- Additive Decomposition Time Series

AIC- Akaike Information Criterion

ANN- Artificial Neural Networks

AR- Auto Regressive

ARIMA- Auto Regressive Integrated Moving Average model

CF- Certainty Factor

CL- certainty level

CNN- Convolutional Neural Network

DES- Double Exponential Smoothing

DEWS- Drought Early Warning System

DTSF- Decomposition Time-Series Forecasting

EM- electromagnetic

ERD- entity relational diagram

EVI- Enhanced Vegetation Index

EWS- Early Warning System

FIRS- Fuzzy Inference Rules System

GHG- greenhouse gases

GIS- Geographic Information Systems

HCMGIS - Center for Applied GIS of Ho Chi Minh City

IADS- Intelligent Agro-climate Decision Support

IK- Indigenous Knowledge

IoT- Internet of Things

KNN- K-Nearest Neighbour

KZN- KwaZulu-Natal

LES- Linear Exponential Smoothing

LR- Linear Regression

LSTM- Long Short Term Memory

MA- Moving Average

MDTS- Multiplicative Decomposition Time Series

ML- Machine Learning

NaE- NDVI and EVI

NBD- Newton Backward Difference

NDDP- Newton Divided Difference Polynomial

NDMI- Normalised Difference Moisture Index

NDVI- Normalised Difference Vegetation Index

NDWI- Normalised Difference Vegetation Index

NFD- Newton Forward Difference

NIR- Near infrared

OBIA- object-based image analysis

PACF- Partial Auto Correlation Function

QGIS- Quantum Geographic Information System

RB-DEWES- Rule Based Drought Early Warning Expert System

RF- Random Forest

RGB- Red Green Blue

RMSE- Root Mean Squared Error

RNN- Recurrent Neural Network

RTF- Random Tree Forest

SARIMA- Seasonal Auto Regressive Integrated Moving Average model

SCFs- Seasonal Climate Forecasts

SCP- Semi-Automatic Classification Plugin

SES- Single Exponential Smoothing

SMA- Simple Moving Average

SSA- sub-Saharan Africa

SVM- Support Vector Machines

SWIR- Short Wave Infrared

TES- Triple Exponential Smoothing

UAV- Unmanned Aerial Vehicle

USGS-United States Geographical System

VAR- Vector Auto Regressive

VDVI- Visible-band Difference Vegetation Index

WDA- Weed Detection Accuracy

WLS- weighted list squares

WMA- Weighted Moving Average

Table of Contents

CHAPTER 1:	1
INTRODUCTION AND BACKGROUND INFORMATION	1
1.1. Background Information and Motivation	1
1.2. Problem Statement	2
1.3. Research Hypothesis.....	3
1.4. Research Questions	3
1.5. Research Aim and Objectives.....	4
1.6. Research Limitations and Scope	4
1.7. Structure of the Thesis	5
CHAPTER TWO: LITERATURE REVIEW	7
2.1. Introduction.....	7
2.2. Climate Change.....	7
2.3. Farmers' Cropping Decisions in the Face of Climate Change	7
2.4. The use of SCFs to support farmers' cropping decisions	10
2.5. Indigenous Knowledge.....	10
CHAPTER 3	37
RESEARCH METHODOLOGY	37
3.1 Introduction	37
3.2 Research Design	37
3.3 Research paradigm	38
3.4 Data Collection Methods	39
3.4.1 Focus group interviews	39
3.4.2 Survey.....	39
3.5 Data Types	40
3.6 Target Population	40
3.7 Research sampling.....	40
3.8 Reliability and Validity.....	40
3.9 Study Area.....	41
3.9.1 uMgungundlovu District Municipality.....	41
3.10 The proposed research framework	42
CHAPTER 4:	45
CROPPING DECISIONS' OPTIMISATION.....	45
	11

4.1	Introduction	45
4.2	Procedures followed to optimise farmers' cropping decisions.....	45
4.2.1	Primary data collection and processing.....	46
4.2.1.1.	Quantitative data collection	46
4.2.1.2.	Quantitative data processing and analysis.....	46
	Section A: Demographic distribution of the respondents.....	47
	Section B1: Indigenous Knowledge in Rainy Season Behaviour Predictions	47
	Section B2: IK in Crop Health and Development Assessment	49
	Conclusions from the quantitative data analysis	52
4.2.1.3.	Qualitative data collection methods and processing.....	53
	Focus group discussion	53
	Data collection using the survey	56
	Demographic distribution of the respondents.....	57
	Some qualitative data collected using the survey.....	57
4.2.2.	Systematic processing of the IK indicators and their observations	59
4.2.2.1.	Certainty Level Computation.....	64
4.2.3.	Secondary data collection, processing and analysis	67
4.2.3.1.	Satellite Imagery Data Collection	68
4.2.3.2.	Satellite Imagery Data Processing	68
	Pre-processing steps of the satellite images.....	68
	Atmospheric correction	69
	Cloud masking and mosaic.....	70
4.2.3.3.	Weather data collection and processing.....	71
	Weather Data Pre-Processing	71
4.2.3.4.	Time Series Analysis	73
4.2.3.5.	Time Series Forecasting	76
	Background information on the Holt-Winter Model	77
	Implementation of Holt-Winter's model in Python	78
	Background knowledge of SARIMA	80
	AR model	81
	MA Model.....	81
	ACF and PACF models.....	82
	ARIMA.....	82
	SARIMA.....	83
	Implementation of SARIMA in Python	83
		12

LSTM Background information	87
Implementation of LSTM	88
Models' Evaluations and future forecasts	91
4.2.4. Quantifying the season behaviour predictions made via IK system	92
4.2.4.1. Preparation of historical data to complement the IK system	92
Rainy season predictions using IK system	94
4.2.5. Integration of forecasts done via IK and science	99
4.2.6. Development of an optimisation model	100
4.2.6.1. Constructing a list of feature values in a presentable format to the farmer	104
4.2.6.2. Background knowledge of the proposed optimisation model	106
4.2.6.3. A deeper look into how the optimisation model was developed	107
Training the Optimisation Model	108
A deeper look into the learning process by an algorithm	116
Recommendations' prediction by the model	119
4.2.7. System output	120
4.2.8. Optimisation Model Evaluation	122
CHAPTER 5:DISCUSSIONS	128
CHAPTER 6: CONCLUSION	138
BIBLIOGRAPHY	140

List of Figures and Tables

Figure 2- 1: Electromagnetic Spectrum with different wavelengths and frequencies. (Source: Zwinkel, 2015)	16
Figure 2- 2: Example of SVM in a 2-dimensional space. (Source: Fletcher,2009)	22
Figure 2- 3: Decision tree classifier example. (Source: Charbuty and Abdulazeez, 2021)	23
Figure 2- 4: Example of Random Forest Classifier. (Source: Nakahara et al., 2017).....	24
Figure 3- 1: Kwazulu-Natal Province and the study site (Source: Author).....	42
Figure 3- 2:The framework followed to optimise farmers' cropping decisions.	43
Figure 4- 1:Presentation of respondents' results regarding foreseeing of the with season with fine weather (Source Author)	47
Figure 4- 2: Presentation of respondents' results regarding foreseeing the season with unfavourable weather (Source Author).....	48
Figure 4- 3: Presentation of respondents' results regarding farmers' satisfaction with IK (Source Author)	49
Figure 4- 4: Presentation of respondents' results regarding monitoring of health and growth of crops (Source Author)	50
Figure 4- 5 : Presentation of respondents' results regarding diagnosing and treating of crop problems (Source Author)	51
Figure 4- 6: Presentation of respondents' results regarding their satisfaction with IK in monitoring the health of crops (Source Author)	52
Figure 4- 7: First image of the participants in Swayimane High School (Source: Author)	53
Figure 4- 8: Second image of the participants in Swayimane High School (Source: Author).....	54
Figure 4- 9: ER diagram showing the relationship between indicators and observers (Source: Author)	61
Figure 4- 10: An algorithm that captures the registration of the IK indicators' observation (Source: Author)	62
Figure 4- 11: An algorithm that extracts distinct observed indicators (Source: Author)	64
Figure 4- 12: An algorithm that computes frequency of distinct observed indicators (Source: Author)	65
Figure 4- 13: An algorithm that computes number of observers for every observed indicator (Source: Author).....	65
Figure 4- 14: An algorithm that validates the qualifying indicators based on the given thresholds (Source: Author).....	66
Figure 4- 15: An algorithm that sums the weights based on season behaviour type (Source: Author)	66
Figure 4- 16: An algorithm that identifies the rainy season behaviour type with max value(Source: Author)	67
Figure 4- 17: Extracted pixels that represent the region of interest. (Source: Author)	68
Figure 4- 18: A surface reflection of Swayimane region after reducing the atmospheric effects. (Source: Author).....	69

Figure 4- 19: Depiction of healthy vegetation presented in green colour and non-vegetation presented in white colour. (Source: Author)	70
Figure 4- 20: Depiction of rainfall time series data. (Source: Author).....	74
Figure 4- 21: Depiction of temperature time series data. (Source: Author)	74
Figure 4- 22: Depiction of waterbodies time series data. (Source: Author)	75
Figure 4- 23: Depiction of soil moisture time series data. (Source: Author)	75
Figure 4- 24: Depiction of vegetation cover time series data. (Source: Author)	76
Figure 4- 25: Depiction of rainfall time series data with decoupled time series components. (Source: Author).....	76
Figure 4- 26: Temperature forecasts using Holt Winter's model (Source: Author)	79
Figure 4- 27: Rainfall forecasts using Holt Winter's model (Source: Author)	79
Figure 4- 28: Vegetation data forecasts using Holt Winter's model (Source: Author)	79
Figure 4- 29: Soil moisture data forecasts using Holt Winter's model (Source: Author)	80
Figure 4- 30: Water surfaces time series data forecasts using Holt Winter's model (Source: Author)	80
Figure 4- 31: A function that evaluates the stationarity of the series (Source: Author)	84
Figure 4- 32: A console screen demonstrating the stationarity of the vegetation time series data (Source: Author).....	84
Figure 4- 33: Rainfall forecast using SARIMA model (Source: Author).....	85
Figure 4- 34: Temperature forecast using SARIMA model (Source: Author).....	85
Figure 4- 35: Waterbodies forecast using SARIMA model (Source: Author).....	86
Figure 4- 36: Vegetation forecast using SARIMA model (Source: Author).....	86
Figure 4- 37: Soil moisture forecast using SARIMA model (Source: Author)	86
Figure 4- 38: Water surfaces forecast using LSTM model (Source: Author).....	89
Figure 4- 39: Vegetation health forecast using LSTM model (Source: Author)	90
Figure 4- 40: Rainfall forecast using LSTM model (Source: Author)	90
Figure 4- 41: Soil moisture forecast using LSTM model (Source: Author)	91
Figure 4- 42: Rainfall forecast using LSTM model (Source: Author)	91
Figure 4- 43: Future forecast using LSTM (Source: Author)	
Figure 4- 44: Illustration of how new data points will be normalised into the range of 1-6	93
Figure 4- 45: An algorithm that extracts valid historic season patterns based on an evaluated trend (Source: Author)	
Figure 4- 46: An algorithm that evaluates different learning parameters of SES model. (Source: Author)	97
Figure 4- 47: The SES model's fit() function that performs learning based on the given alpha. (Source: Author).....	98
Figure 4- 48: An algorithm that returns optimal model's learning rate parameter (Source: Author)..	98
Figure 4- 49: An algorithm that performs season behaviour forecast using historic records (Source: Author)	
Figure 4- 50: The SES model's prediction method (Source: Author).....	
Figure 4- 51: An ERD showing the relationship between features and recommendations for a given problem domain (Source: Author)	103
Figure 4- 52: An algorithm that normalizes the raw sensor data. (Source: Author)	105
Figure 4- 53: A fit() function of an optimisation model. (Source: Author).....	108
Figure 4- 54: A function that performs clustering of the records. (Source: Author)	110
Figure 4- 55: Illustration of centroids' starting positions based on K-parameter. (Source: Author)	111

Figure 4- 56: An algorithm that will return the optimal starting position of the centroids. (Source: Author)	112
Figure 4- 57: An algorithm that updates centroids' position. (Source: Author)	113
Figure 4- 58: Demonstration of features making up different questions. (Source: Author).....	114
Figure 4- 59: An algorithm that updates feature values of centroids . (Source: Author)	114
Figure 4- 60: An algorithm that returns an optimal value for a given feature. (Source: Author)	115
Figure 4- 61: An algorithm that computes distance between a data object and a centroid. (Source: Author)	116
Figure 4- 62: An algorithm that computes the distance for categorical features. (Source: Author)	117
Figure 4- 63: An algorithm that computes the distance for numerical features. (Source: Author)	117
Figure 4- 64: An algorithm that computes the clusters' variance. (Source: Author)	118
Figure 4- 65: An algorithm that retrieves the recommendations based on the given input. (Source: Author)	119
Figure 4- 66: An illustration of how the optimal cluster will be identified. (Source: Author)	119
Figure 4- 67: Menu screen for the user. (Source: Author)	120
Figure 4- 68: Screenshot of how the observation of the indicator will be registered. (Source: Author)	121
Figure 4- 69: Screenshot of photo capturing for environmental indicators. (Source: Author) ..	121
Figure 4- 70: Screenshot of the system's generated forecast for rainfall (Source: User)	122
Figure 4- 71: Screenshot of how features were created. (Source: Author)	124
Figure 4- 72: Screenshot of all features created. (Source: Author)	124
Figure 4- 73: Screenshot of all recommendations created. (Source: Author)	125
Figure 4- 74: Screenshot of how the features and recommendations are mapped. (Source: Author)	125
Figure 4- 75: Screenshot of the model's prediction of crops to grow based on features provided. (Source: Author)	126
Figure 4- 76: An algorithm that computes the model's accuracy for the input provide. (Source: Author)	127

List of Tables

Table 4- 1:Illustration of how time series data were split into training and test sets. (Source: Author).....	77
Table 4- 2:Presentation of models' performance (Source : Author).....	92
Table 4- 3: Example of historic, current and future time series observations. (Source: Author)	96
Table 4- 4:An illustration of forecasts presented in a tabular format (Source: Author)	100
Table 4- 5: An illustration of forecasts presented in a tabular format (Source: Author)	102
Table 4- 6: Illustration of how feature values that will be made from human observation will be modelled. (Source: Author).....	104
Table 4- 7: Illustration of how feature values from sensor reading will be modelled. (Source: Author).....	105
Table 4- 8: Illustration of how records will be stored inside the knowledge database. (Source: Author).....	107
Table 4- 9: Set of features and recommendations used to train the model. (Source: Author)	123

List of Publications

As an outcome of this research, the author of this thesis is the first author of the currently published papers listed below.

1. Nyetanyane, J., 2023, December. Integration of IK, Satellite Imagery Data, Weather Data and Time Series Models in Season Behaviour Predictions. Case of Swayimane, KZN, South Africa. In *International Conference on Emerging Technologies for Developing Countries* (pp. 165-184). Cham: Springer Nature Switzerland.
2. Nyetanyane, J. and Masinde, M., 2020. Integration of indigenous knowledge, climate data, satellite imagery and machine learning to optimise cropping decisions by small-scale farmers. A case study of uMgungundlovu district municipality, South Africa. In *Innovations and Interdisciplinary Solutions for Underserved Areas: 4th EAI International Conference, InterSol 2020, Nairobi, Kenya, March 8-9, 2020, Proceedings 4* (pp. 3-19). Springer International Publishing.
3. Nyetanyane, J., 2022, December. Indigenous Knowledge Mobile Based Application that Quantifies Farmers' Season Predictions with the Help of Scientific Knowledge. In *International Conference on Emerging Technologies for Developing Countries* (pp. 192-205). Cham: Springer Nature Switzerland.
4. Nyetanyane, J. and Masinde, M., 2022, May. Development of English-isiZulu Language Translator to Bridge Language Barrier Between EWS and Bantu Language Speaking Farmers. Case of Swayimane, Kwazulu-Natal, South Africa. In *2022 IST-Africa Conference (IST-Africa)* (pp. 1-11). IEEE.
5. Nyetanyane, J. and Masinde, M., 2020, November. Foresee Transition to Agricultural Season by Integrating Indigenous Knowledge, Satellite Imagery, Weather Data and ARIMA Family Models to Enable Good Crop Establishment by Small-Scale Farmers in Swayamani Region, KwaZulu-Natal, South Africa. In *2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)* (pp. 1-7). IEEE.
6. Nyetanyane, J., 2023, May. Indigenous Knowledge Crop Health Recommendation Expert System for the Upcoming Generations. In *2023 IST-Africa Conference (IST-Africa)* (pp. 1-8). IEEE.

CHAPTER 1:

INTRODUCTION AND BACKGROUND INFORMATION

1.1. Background Information and Motivation

Crop production optimisation is critical for ensuring food security and improving the livelihoods of small-scale farmers, particularly in regions like SSA, where rain-fed agricultural practices dominate. Effective cropping decisions are essential for maximising yield, maintaining soil health, and managing resources efficiently. However, farmers in SSA face significant challenges in optimising their crop production due to the variability and unpredictability of rainfall and the increasing frequency of extreme weather events (Hadlos et al., 2022).

Various efforts have been made across SSA to support farmers in making informed cropping decisions. However, these efforts often face challenges related to accessibility, credibility, interpretation, and scalability. For instance, SCFs provided by governments and meteorological organisations are usually regional and do not account for local variations in climatic factors, limiting their effectiveness for small-scale farmers (Atiah et al., 2020). The limitation and uneven distribution of weather stations and rain gauges exacerbate the inaccuracy of these forecasts (Lawens et al., 2019).

In South Africa, the government, through the South African Weather Service (SAWS), continues to provide SCFs to help farmers anticipate seasonal patterns and adjust their cropping strategies. Despite these efforts, numerous studies have highlighted farmers' dissatisfaction with these forecasts owing to similar issues of accessibility and scalability (Masinde et al., 2012; Dinku, 2017; Ncoyini et al., 2022). The regional nature of these forecasts means they often fail to account for the significant local variations in rainfall and other climatic factors, further limiting their usefulness for small-scale farmers.

Focusing on the uMgungundlovu district in KZN, South Africa, small-scale farmers face similar challenges. Their reliance on rain-fed agriculture makes them particularly vulnerable to the unpredictable and insufficient rainfall patterns exacerbated by climate change. This leads to crop loss and poor yields, significantly affecting their livelihoods (Nyetanyane and Masinde, 2022). To address these challenges, innovative approaches that integrate advanced

technologies with local expertise are needed to provide more accurate and localised predictions and management strategies.

Advancements in technology, particularly the use of satellite imagery and machine learning (ML) models, offer promising solutions for improving farmers' cropping decisions and crop production. ML models have been successfully applied to predict weather patterns and analyse remote sensing data, providing valuable insights into vegetation health, land use, and water resources (You et al., 2017). These technologies enable more precise monitoring and management of agricultural systems, allowing farmers to make data-driven decisions that enhance productivity and sustainability.

This thesis integrates satellite imagery, ML, climate data, and traditional knowledge to optimise cropping decisions for small-scale farmers in the uMgungundlovu district, KwaZulu-Natal, South Africa. By combining scientific methods with local expertise, this research aimed to develop a comprehensive decision-support tool that improves farmers' ability to predict rainfall and temperature patterns, monitor crop health, and manage their resources more effectively. The integration of these knowledge systems is implemented in a mobile application, providing farmers with an accessible and practical tool for optimising their cropping decisions.

1.2. Problem Statement

Small-scale farmers in the uMgungundlovu district, KwaZulu-Natal Province, South Africa, rely heavily on rain-fed agriculture for their crop production. Their agricultural systems face significant challenges owing to unpredictable and insufficient rainfall, leading to crop loss and poor yields (Nyetanyane and Masinde, 2022). To support farmers in making better cropping decisions, the South African government, in collaboration with meteorologists, provides SCFs. However, many local farmers find it difficult to incorporate these forecasts into their cropping strategies owing to issues of accessibility, credibility, interpretation, and scalability (Masinde et al., 2012; Dinku, 2017; Ncoyini et al., 2022). These forecasts are not localised to specific areas and do not incorporate farmers' practical knowledge, resulting in their limited adoption by smallholder farmers. Many local farmers continue to rely heavily on their IK system as a shield to tackle climate calamities and improve their cropping decisions, however, the value and strength of the IK continue to perish due to deforestation, climate stress, pollution, and many other human activities that negatively impact the environment and its habitats that make up the IK system.

To optimise crop production and address these challenges, there is a need for innovative approaches that integrate advanced technologies with local expertise. This research thus focuses on developing a comprehensive decision-support system that leverages satellite imagery, machine learning models, IK, and climate data to provide farmers with improved and localised predictions of rainfall and temperature patterns. Additionally, this system will assist farmers in monitoring crop health and managing their resources more efficiently, ultimately improving crop yields and ensuring sustainable agricultural practices.

By addressing the limitations of current forecasting methods and incorporating local knowledge, this research sought to provide small-scale farmers with a robust tool for optimising their cropping decisions, enhancing their resilience to climatic variability, and improving their overall productivity.

1.3. Research Hypothesis

Current IK systems face limitations in enabling farmers to accurately predict the onset and behaviour of rainfall and temperature throughout the rainy season, essential for optimising cropping decisions and maximising yield.

Hypothesis 1: The integration of IK, satellite imagery data, weather data, and ML time series models will provide farmers with improved and more localised predictions of rainy seasons, thus enabling more informed cropping decisions.

Farmers currently struggle to monitor the health and development of crops through traditional IK observations owing to the increasing complexity of crop health challenges intensified by climate change.

Hypothesis 2: Integrating IK with scientific data and methods will facilitate the development of a machine learning optimisation model that continually learns from IK experts and scientists, guiding farmers in effectively monitoring the health and growth of their crops.

1.4. Research Questions

From the research problem, the following research questions were formulated:

1. How do smallholder farmers who rely heavily on rain-fed agriculture utilise the indigenous knowledge system to support their cropping decision-making process?
2. To what extent is the IK system effective in the uMgungundlovu district of KwaZulu-Natal?

3. How can the weaknesses of the IK system be complemented with scientific knowledge to optimise farmers' cropping decisions?

1.5. Research Aim and Objectives

This research sought to integrate IK with ML, satellite imagery, and climate data to optimise cropping decisions for small-scale farmers in the uMgungundlovu district, South Africa. This aim was achieved through the following objectives:

Objective 1: Systematically collect and process IK indicators that farmers use to predict rainy season behaviour.

Objective 2: Collect and process historical weather and satellite imagery data using Geographic Information System (GIS) and ML imputation models.

Objective 3: Train and evaluate time series models on historical weather and satellite imagery data to generate rainy season forecasts representing the scientific perspective.

Objective 4: Quantify farmers' rainy season predictions using historical weather and satellite data, and a SES time series model.

Objective 5: Integrate farmers' predictions with scientific forecasts to develop more robust and localised rainy season predictions.

Objective 6: Develop and evaluate an optimisation model that continually learns from IK and scientific experts, guiding farmers in monitoring the health and development of their crops.

1.6. Research Limitations and Scope

Cropping decisions are extensive, encompassing activities at organisational, administrative, and marketing management levels. At the organisational level, decisions can be broadly classified into strategic and operational categories. This research is restricted to certain aspects of operational decision-making by small-scale farmers at the organisational level. This includes the prediction of rainy seasons and knowledge concerning the health and growth of crops. Given the broad scope of IK, this study was limited to the IK indicators that farmers use to predict the behaviour of rainfall and shifts in agricultural seasons, as well as their knowledge in monitoring crop health and growth. The IK indicators for predicting rainy seasons are

narrowed down to meteorological, astronomical, and environmental aspects, excluding indicators of a spiritual nature.

For satellite imagery data, only the land cover features of vegetation, waterbodies, and soil are extracted and manipulated from free weekly Sentinel-2 satellite images. The research employs univariate time series models exclusively to generate season behaviour forecasts from historical satellite imagery and weather data. The univariate time-series model assumes that the next forecast can be explained by the variable in question without considering other variables that may have either direct or indirect influence on the variable. The behaviour of the agricultural season or the rainy season discussed in this research is explained based only on rainfall and temperature changes. Furthermore, the optimisation model used in this research to assist farmers in monitoring crop health and growth is a modified version of the K-Prototype clustering algorithm, capable of working with varying sizes of input features to generate predictions. Moreover, this research is geographically confined to the uMgungundlovu district in KZN Province, South Africa, limiting the generalisability of the findings to other regions or broader agricultural contexts.

1.7. Structure of the Thesis

This thesis is organised into six chapters, including the introductory chapter.

Chapter One: Introduction

This chapter provides background information on the impact of climate change on small-scale farmers in the uMgungundlovu district, KZN Province, South Africa. It introduces the problem statement, research hypothesis, research questions, aim and objectives, and the scope of the study.

Chapter Two: Literature Review

Chapter two reviews existing literature on small-scale farmers' cropping decisions and the challenges they face due to climate change. It discusses adaptation and mitigation strategies employed by smallholder farmers, and the role of IK, ML, and satellite imagery technologies in optimising cropping decisions. This chapter provides a theoretical foundation and identifies gaps in the current research that this study sought to address.

Chapter Three: Research Methodology

This chapter details the methodological framework of the thesis. It explains the research design, data collection methods, and analytical techniques used. It also outlines the procedure for integrating IK with scientific data to develop predictive models.

Chapter Four: Methodology for Optimising Cropping Decisions

This chapter focuses on the specific methodologies used to optimise cropping decisions for smallholder farmers. It includes the processes for collecting and processing IK indicators, the application of time series models for forecasting rainy season behaviour, and the development and evaluation of the modified K-Prototype clustering algorithm that will aid farmers in monitoring the health and growth of crops.

Chapter Five: Discussion

This chapter outlines procedures followed to address the research aim and objectives. It also presents the results obtained, evaluates the strengths and limitations of the proposed tool, and offers recommendations for future research and improvements.

Chapter Six: Conclusion

Chapter six summarises the key findings of the research, and it highlights the study's contributions to the agricultural decision-making field. This chapter concludes the thesis by emphasising the potential impact of integrating IK with scientific methods to enhance the resilience of small-scale farmers in the face of climate change.

CHAPTER TWO: LITERATURE REVIEW

2.1. Introduction

This chapter highlights the challenges that small-scale farmers face in their cropping decision-making processes, and it also evaluates the technologies available to support these decisions, including their strengths and weaknesses. The focus is on IK, ML, and satellite imagery technologies, examining their potential to aid farmers' cropping decisions.

2.2. Climate Change

Climate change refers to significant and lasting changes in weather patterns, including prolonged heatwaves leading to global warming, and erratic rainfall causing droughts, floods, and cyclones, which adversely affect ecosystems and their inhabitants (Rani et al., 2020). Human activities contribute significantly to climate change, mainly through the emission of greenhouse gases (GHGs) associated with industrialisation, economic growth, and the burning of fossil fuels (Patz et al., 2007; Oreggioni, 2021). This phenomenon exacerbates food insecurity, particularly in vulnerable regions, leading to hunger and undernutrition (Braun, 2020).

The agricultural sector is severely impacted by climate change, which threatens food production and livelihoods globally, despite technological advancements such as irrigation systems (Wood et al., 2014). The threat is more pronounced in many African countries due to their reliance on the vulnerable agricultural sector and their limited adaptive capacity (Twomlow, 2008; Abebe, 2014).

In SSA, climate change has altered precipitation patterns and reduced water availability, causing significant challenges for rain-fed agriculture, which accounts for over 90% of farming in the region (Nuamah and Botchway, 2024; Lamptey, 2022). Droughts, one of the most severe impacts of climate change, have led to declines in water and food production, leaving millions without adequate water supply and vulnerable to starvation (Jiri et al., 2016).

2.3. Farmers' Cropping Decisions in the Face of Climate Change

Judicial implementation and optimisation of cropping decisions can improve food production and maintain livelihoods despite climate uncertainties and increased food and water demand. Cropping decisions at their basic level involve selecting crops, their acreage, and their location

within a given land (Dury et al., 2022). These decisions are central to farm management, and poor implementation can lead to poor yields and revenue loss. Cropping decisions occur at different management levels, namely: organisational, administrative, and marketing.

Organisational-level decisions include strategic and operational management. Strategic decisions involve significant investments with long-lasting effects, such as land purchases, machinery, and irrigation programs. Operational decisions are more frequent and involve smaller investments, addressing questions like what, how much, and when to produce. Administrative decisions focus on logistics and finance, while marketing decisions relate to buying and selling markets.

Despite their importance, cropping decisions are complex, especially in rainfed agriculture managed by small-scale farmers lacking resources to sustain crop development amid climate change. Strategies adopted by local farmers to increase yield and sustain biodiversity include cultivating drought or heat-tolerant crops, retaining ridges, mulching, and increasing soil organic matter to enhance water retention (Lawens et al., 2019).

Other techniques include crop rotation, use of crop residues, animal manure, legumes, green manure, off- and on-farm organic wastes, mechanical cultivation, mineral-bearing rocks, and biological pest control to maintain productivity and tilth, supply nutrients, and control weeds and pests (Sign et al., 2021). Crop row orientation helps suppress weeds by limiting sunlight, and increasing crop yield (Bayer et al., 2022). Crop rotation, the practice of growing different crops in succession on the same land, retains soil nutrients, breaks weed and disease cycles, and reduces dependence on external inputs (Dung et al., 2011; Chandrasekhar et al., 2022).

The slash-and-burn technique, widely used by small-scale farmers, clears land by cutting and burning old crops, improving soil fertility with nutrient-rich ash (John et al., 2021). However, this environmentally unfriendly approach can damage soil structure and organisms, leading to physical and chemical changes, increased greenhouse gas emissions, soil erosion, and other adverse effects (Tang and Yap, 2020; Fachin and Thomaz, 2023; Primo et al., 2023).

Staggered planting extends the harvesting period, minimises loss and labour by sowing seeds on small portions of land, and evaluating the rainy season behaviour before expanding cultivation (Nuamah et al., 2022). Intercropping, or mixed cropping, involves cultivating various crops together, with cover crops supporting cash crops. For example, combining maize and cowpea, where cowpea fixes nitrogen for maize, reduces the need for external inputs

(Mndzebele et al., 2020; Masete et al., 2022). Leguminous crops like *Canavalia* rejuvenate soil with natural fertilisers and improve soil aeration, aiding nutrient absorption (Mubiru and Coyene, 2009).

Fallowing, leaving land uncultivated during the sowing season, breaks pest and disease cycles and replenishes soil nutrients. Traditional tillage, aerating the soil, permits moisture and air penetration, encourages root growth, controls weeds, and integrates fertilisers (Nweke, 2018). Despite its benefits, traditional tillage conflicts with conservation agriculture, releasing conserved soil carbon into the atmosphere and potentially leading to soil compaction and nutrient leaching during heavy rains.

Rains are crucial for cropping plan decisions. Farmers determine planting times, locations, and crop types based on estimated rainfall behaviour. This behaviour is primarily anticipated by observing environmental, meteorological, astronomical, and spiritual IK indicators.

Some farmers plant immediately after the first rains, while others practice dry soil planting, involving sowing before rains to ensure seeds are in the soil when rains begin. This labour-intensive approach (dry soil planting) poses a risk if the expected rains are insufficient, which can result in premature germination and crop failure (Basdew et al., 2017). Drought-tolerant or short-cycle crops are planted when less rain is expected while water-demanding or long-cycle crops are planted if abundant rainfall is anticipated. Accurate rain predictions are thus critical for cropping decisions, but poor predictions can worsen food insecurity. In addition to rainfall estimates, planting periods should be timed to avoid periods of high insect populations that could harm crops during their early growth stages.

Market uncertainty also affects cropping decisions. A crop may be suitable for the climate but not the market, leading to losses for smallholder farmers. Crop staggering can help minimise such losses.

Despite organic methods to sustain biodiversity, growing populations, land degradation, food and water stress, and increased pests and diseases have forced farmers to use chemicals to meet food demands. Crop pests significantly constrain productivity and profitability, causing substantial pre- and post-harvest losses globally (Olatinwo and Hoegenboom, 2014). Chemical control remains essential for agricultural success, but its use poses environmental and health risks, especially for smallholder farmers in SSA who lack knowledge and resources including protective equipment (Demi and Sicchia, 2021).

2.4. The use of SCFs to support farmers' cropping decisions

SCFs provide long-term average seasonal conditions and agronomical guidelines to aid small-scale farmers, especially in rainfed agriculture, in making informed decisions to maximise yield. SCFs have proven useful in warning of above-normal rainfall, helping emergency aid providers, and improving crop management decisions, including crop choice, cultivar selection, fertiliser use, and tillage type (Lawens et al., 2019; Chagonda et al., 2015).

Training in SCFs interpretation and use has shown yield improvements among farmers compared to those who did not receive training (Chagonda et al., 2015). However, SCFs are perceived as less useful, credible, and legitimate in SSA compared to other regions, primarily due to scalability issues, and limited accessibility and interpretation (Joshua et al., 2011; Masinde et al., 2012; Wood et al., 2014; Dinku, 2017). According to Dinku (2017) and Ncoyini (2022), the accuracy of climate forecasts is hindered by the limited number of weather stations that are unevenly distributed with most being concentrated in cities and towns, while rural areas (where the data are needed most) remain underrepresented. The format and language used in the dissemination of SCFs further hinder understanding and acceptance (Asave-Nuamah and Botchway, 2019). Public agricultural extension services help convey and interpret SCFs information, but limited interaction with farmers remains a critical constraint (Ncoyini et al., 2022).

2.5 Indigenous Knowledge

Despite the climatic vagaries and poor technological adaptations in SSA, many small-scale farmers have utilised IK for centuries to adapt to and mitigate the impacts of climate change (Jiri et al., 2016). This knowledge has enabled African farmers to understand weather and climate patterns and make informed decisions about crop and farming practices. IK assists in pest management, soil fertility assessment, and various other agricultural aspects, thus improving yields without relying on chemical inputs (Mihale et al., 2009; Tongkoom et al., 2021). Enock (2013) accentuates that communities, particularly those in drought and flood-prone areas, have generated a vast body of knowledge on disaster prevention and mitigation through early warning and preparedness.

Indigenous knowledge, also known as traditional or local knowledge, is a body of knowledge existing within or acquired by small-scale farmers over time through the accumulation of experiences, society-nature relationships, community practices, and institutions, handed down through generations (Mercer et al., 2009; Turner and Clifton, 2009; Masinde et al., 2012). This place-based knowledge is rooted in local cultures and is typically associated with long-settled communities with strong ties to their natural environments (Orlone et al., 2009). IK results from cumulative experiences and observations tested in everyday life, developed through oral communication and repetitive engagements rather than formal instructions (Orlone et al., 2009; Pareek and Trivedi, 2011). It is an unstructured type of knowledge mainly stored in the minds of elderly farmers and residents, often not documented, and disseminated through various channels (Basdew et al., 2017).

While scientific knowledge is global, IK is local and dynamic, continuously influenced by internal creativity, experimentation, and contact with external systems (Mercer et al., 2010; Chepchirchir et al., 2019). IK is qualitative and geographically specific, unlike scientific knowledge, which is typically based on synchronic observations tending towards the quantitative and general (Mercer et al., 2010). IK can add value to developing sustainable climate change mitigation and adaptation strategies that are rich in local content and planned with small-scale farmers, making them cost-effective, participatory, and sustainable (Nyong et al., 2006; Mercer et al., 2009; Orlone et al., 2009; Masinde et al., 2012; Mistry and Berardi, 2016; Jiri et al., 2016; Akanbi, 2019).

Sustainable smallholder agricultural production cannot be achieved without local coping to current variability and change, and adaptive capacities for future climate change (Jiri et al., 2016). Studies have documented the effectiveness of IK in land management in various settings and shown that IK persists even in commercial land use and government land management contexts (Orlone et al., 2009). Small-scale farmers have developed many distinctive and successful lifeways through local resources and adaptation to their landscapes and environments (Turner and Clifton, 2009). Moreover, turning to indigenous people for help and insight makes great sense because they are long-term residents of a place, who have learned through systems of knowledge, practice, and belief to conserve, maintain, and promote their resources in situ, and have developed a capacity for resilience (Turner and Clifton, 2009).

The concept of indigenous knowledge has gained significant attention from academics and development practitioners for its crucial role in tackling various livelihood challenges in rural

communities and for supporting locally driven adaptation strategies (Jiri et al., 2016). Although IK has prompted calls for its use in disaster reduction, it has not yet been fully accepted as a knowledge system that is capable of bringing change in adaptation and mitigation strategies nationally and internationally (Mercer et al., 2009; Pareek and Trivedi, 2011; Owiny et al., 2014). IK is oppressed in numerous ways because of marginalisation, exploitation, powerlessness, cultural imperialism, violence, and denial of existing knowledge placed upon its bearers (Mercer et al., 2009; Ross et al., 2016; Magni, 2017). Academic debates, whether in science or history, too often evaluate local expertise as data or evidence rather than as knowledge or theory that might contribute different perspectives to academic questions (Cruikshank, 2001; Ansa and Mji, 2013; Ross et al., 2016).

Little has been done to incorporate IK into formal climate change mitigation and adaptation strategies (Nyong et al., 2006; Ansa and Mji, 2013; Ross et al., 2016). Factors that limit the integration of IK with science include the dynamic nature of IK, especially as cultures or communities develop and grow, and is subjected to changing environmental, cultural, physical, and economic stressors (Tharakan, 2017). Makwara (2013) and Basdew et al. (2017) indicate that this knowledge is thinly transmitted to upcoming generations, who are more attached to modern life than to cultural life. Many younger generations move from rural to urban areas to seek employment, and when elders, the main custodians of this knowledge, pass away, the accumulated knowledge is often lost (Makwara, 2013, Nyetanyane, 2023). This loss occurs because the knowledge has not been systematically documented and technologically exploited to add more meaning and value for upcoming generations. Climate change impacts, human actions such as pollution, industrialisation, and deforestation also sabotage the value of IK by destroying environmental habitats such as birds, insects, and trees that farmers use as evidence in their IK systems. Another concern is that IK is intertwined with communities' cultural and religious beliefs, making it difficult to integrate with modern scientific reasoning (Masinde et al., 2012).

Small-scale farmers have reported that IK is becoming less common as they lose confidence in the reliability of rains and their ability to predict them (Orlone et al., 2009). This decline is a result of climate change and other various human activities that threaten the survival of the fauna and flora that farmers rely on to develop their IK systems.

2.6 Integration of Indigenous and Scientific knowledge to enhance climate forecast

Despite the shortcomings of both the scientific and IK systems, previous research has emphasised that the two can be integrated to complement each other, leading to robust predictions that enhance the resilience of small-scale farmers against climate change. Integrating the two can help farmers make informed farming decisions and improve the adoption and acceptance of weather forecasts by small-scale farmers (Masinde et al., 2012; Changonda et al., 2015). Although the use of IK alongside scientific knowledge is increasingly advocated, the framework on how the two can be merged is still a challenge given their conflicting characteristics.

Many EWSs have been developed to aid farmers, but they often overlook communities' knowledge, possibly because many scientists have not yet explored deeply how farmers' knowledge can be collected, processed, and integrated with scientific knowledge. What limits the integration of the two is their fundamentally different natures. One is local and qualitative, while the other is global and quantitative. According to Agrawal (2014), integrating the two is possible if IK is collected systematically. Although further efforts are required to merge these knowledge systems, their integration has already proven feasible across various fields, including medicine, engineering, artificial intelligence, machine learning, and the social and economic sciences. (Kala et al., 2004; McBride et al., 2016; Gerke and Evans, 2018).

Several studies have demonstrated the successful integration of IK and scientific knowledge to strengthen farmers' cropping decisions. For instance, JongKoom et al. (2021) proposed a way to combine farmers' knowledge and scientific knowledge on tree species and soil fertility assessment to improve cropping decisions. Pareek and Triredi (2011) integrated science and IK to diagnose hazards in diverse environmental and cultural settings. Masinde et al. (2012) developed the Integration of IK, Artificial Intelligence, and Information Technology (ITIKI) to forecast the early onset of droughts in Western Africa and enhance communities' adaptive capacities. Masinde and Thothela (2019) extended ITIKI to ITIKI Plus, a mobile-based technology that integrates IK, scientific data, and information to provide micro-level drought forecasts and cropping decision information to improve crop production in extreme weather events. Masinde and Akanbi (2016) proposed a middleware that integrates heterogeneous data sources with IK based on a unified ontology for an accurate Internet of Things (IoT) Based Drought Early Warning System (DEWS).

McBride et al. (2016) proposed participatory geographic information systems as an organisational platform for integrating traditional and scientific knowledge in contemporary fire and fuel management. Jones et al. (2015) emphasised the integration of local knowledge and science to address the economic consequences of driftwood harvest in a changing climate. Thothela et al. (2021) proposed the Intelligent Agro-climate Decision Support(IADS) tool for small-scale farmers, investigating the effectiveness of integrating IK with fuzzy inference systems, mobile phone technology, and smart sensors to enhance decision support systems. Thothela et al. (2023) further developed a framework for an Intelligent Agro-climate Decision Support tool incorporating machine learning multi-class classification algorithms and IK systems to predict suitable crops for cultivation given climate constraints.

Wenisch et al. (2010) built a Fuzzy Inference Rules System (FIRS), integrating the farmers' knowledge of soil and fertility, motivated by the idea that fuzzy technologies can quantify farmers' IK. Archieng (2022) explores the potential for digitalising IK to minimise its erosion and disappearance in agricultural productivity, food preservation, and other resources, further highlighting barriers to IK conservation, such as various communication methods and languages impeding documentation processes. Limpo et al. (2022) explored the integration of indigenous and scientific knowledge systems to support rice farming decision-making, hoping to improve crop farming and preserve local wisdom. Akanbi and Masinde (2018) developed a Rule-Based Drought Early Warning Expert System (RB-DEWSES) to forecast drought using local knowledge from small-scale farmers in Swayimane, KwaZulu-Natal Province.

Mwagha and Masinde (2015) proposed artificial intelligence techniques (fuzzy cognitive maps) to validate farmers' IK on weather lore scientifically. This integration stimulates opportunities to incorporate consistent weather lore with modern weather prediction systems, enhancing decision support applications reliant on weather effects.

Despite growing evidence of the potential for integrating indigenous and scientific knowledge, incorporating IK with satellite imagery, climate data, and machine learning technologies to optimise small-scale farmers' cropping decisions has not yet been addressed.

2.7 Remote Sensing

Remote sensing is a technology that can significantly aid small-scale farmers in their cropping decisions. It involves acquiring information about the Earth's surface without direct contact

with it (Lie et al., 2017). This technology has proven effective in tracking and monitoring crop vigour, density, health, development, and growth (Al-Kindi et al., 2017). Remote sensing can be applied in various areas of computational sustainability, such as species distribution modelling, poverty mapping, climate modelling, and natural disaster prevention (You et al., 2017). One prominent example of remote sensing technology is satellite imagery.

2.7.1 Satellite Imagery

Satellite imagery is a powerful tool equipped with advanced sensors and cameras that collect images of the Earth's surface (Jean et al., 2016). These sensors record the electromagnetic radiation emitted from the sun or any other source, reflected from the ground or objects on the ground.

The main advantages of satellite imagery include its synoptic view (global coverage), which allows large areas to be covered at once; repeatability (temporal resolution), which means the same area can be revisited at different times; multispectral capabilities, where sensors measure energy at wavelengths beyond human vision; near real-time data collection; systematic data collection; accessibility, and multipurpose applications where one image can serve various purposes.

Unlike traditional digital image sensors or cameras that capture light energy only in the visible spectrum (RGB: Red, Green, and Blue), satellite imagery sensors capture light energy at multiple wavelength bands, both visible and non-visible to the human eye. Hence, these images are known as multispectral satellite images which range from high to low wavelengths and frequencies. This is illustrated in Figure 2-1 below.

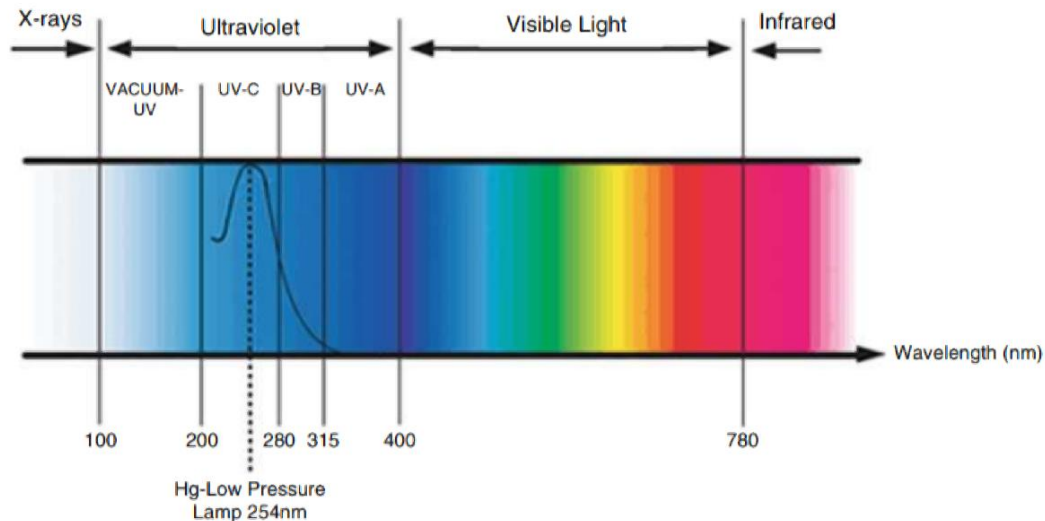


Figure 2- 1: Electromagnetic Spectrum with different wavelengths and frequencies (Source: Zwinkel, 2015).

From Figure 2-1 above, the bands range from the infrared wave, which has a high wavelength and low frequency to X-rays which has a lower wavelength and high frequency.

The satellite imagery technology is equipped with different resolutions. These are temporal, spatial, radiometric, and spectral resolutions.

2.7.1.1 Spatial resolution

The spatial resolution of a satellite determines the spatial coverage of each pixel and is the ability to distinguish closely spaced objects in an image (Somya et al., 2018). It is often measured in meters or kilometres, indicating the ground area represented by one pixel at a given wavelength band. For instance, a spatial resolution of 10m means a 10m by 10m ground area is represented by one pixel value. Higher spatial resolution provides a clearer view of ground objects but covers a smaller portion of the Earth's surface, whereas lower spatial resolution covers larger areas but with less detail.

2.7.1.2 Temporal resolution

Temporal resolution is the frequency with which a sensor revisits a given location or the time between successive images of the same area (Kammili et al., 2014; Inglada et al., 2016). There is an inverse relationship between spatial and temporal resolution; higher spatial resolution generally means lower temporal resolution owing to the time it takes for the sensor to revisit the same area.

2.7.1.3 Radiometric resolution

Radiometric resolution refers to the number of bits used to store information about an image, representing the light reflectance of objects on the Earth's surface (Vibute et al., 2016). Higher radiometric resolution means more detailed information about an object's light reflectance can be recorded. For instance, an 8-bit image can detect light variations from 0 to 255, whereas an 11-bit image ranges from 0 to 2048, providing more detailed information.

2.7.1.4 Spectral resolution

Spectral resolution indicates the number and width of spectral bands that a satellite sensor can measure. Multispectral images are composed of layers, with each layer representing a two-dimensional image captured at a specific wavelength band. A higher spectral resolution enables more accurate characterization and identification of land cover types (Benedetti et al., 2023). For example, hyperspectral sensors, which can record over 100 contiguous spectral bands, have significant applications in fields such as precision agriculture for monitoring crop health, moisture status, and maturity (Shafi et al., 2019).

2.7.2 Image pre-processing

Although satellite imagery is a well-known and widely used technology, the images obtained from satellites are not perfect and require correction. According to Somya et al. (2018), an image must undergo four main processes before it can be utilised efficiently. These processes are image pre-processing, enhancement, transformation and classification.

Image pre-processing deals with correcting radiometric, atmospheric, and geometric distortions present in raw image data. According to Stow (2017), radiometric correction involves reducing the influence of errors or inconsistencies in image brightness values that may limit one's ability to interpret or quantitatively process and analyse digital remotely sensed images. These errors are generally corrected by the operator before an image is supplied for commercial or public use. Atmospheric correction minimises distortion generated when solar energy penetrates through the atmosphere to the Earth's surface and reflects back to the satellite. Some light energy is lost due to the atmosphere's thickness. Geometric distortion refers to errors in the brightness values of the pixels caused by satellite motion, its scanners, and the Earth's movement. These distortions can be systematic or non-systematic. Systematic errors created by sensor scanners include cross-track scanner distortion, mirror velocity variation, and scan skew. Non-systematic errors may be due to Earth's rotation, altitude variation, pitch variation,

spacecraft velocity variation, roll variation, and yaw variation. All these errors can be corrected using the geo-referencing technique, also known as image-to-world transformation, which converts image coordinates to real-world coordinates using a base map or any image with geographic coordinates. Enhancement techniques are applied to pre-processed images to prepare them for visual presentation. This is done through contrast manipulation, spatial feature manipulation, and multi-image manipulation. Image classification, a part of computer vision, involves grouping pixels into finite classes based on their data values (Thakur and Maheshwari, 2017).

2.7.2.1 Band Ratioing

Band ratioing is among the techniques performed under multi-image manipulation enhancement. This technique involves the mathematical manipulation and computation of different spectral bands of a satellite image to generate new information not carried in any single band. Ratio application is also used to remove shadow or topographic effects due to sensor position and solar illumination, improving image interpretation (Shahtahmassebi et al., 2013). Various land cover indices are generated through ratioing, including those for vegetation, soil moisture, water surface, and others.

2.7.2.2 Vegetation indices

Spectral vegetation indices are mathematical combinations of different spectral bands, mostly in the visible and near-infrared regions of the electromagnetic (EM) spectrum (Wu, 2014). These combinations improve the assessment of vegetation properties and enable reliable spatial and temporal comparisons of terrestrial photosynthetic activity and variations in canopy structure (Huete et al., 2002). They are also designed to find functional relationships between crop characteristics and remote sensing data (Chlingaryan, 2018). Using vegetation indices, it is now possible to discriminate healthy vegetation from stressed vegetation based on their spectral signatures. Healthy vegetation reflects more in the near-infrared (NIR) part of the spectrum, while stressed vegetation, lacking chlorophyll, reflects more in the red part of the EM spectrum. Vegetation indices measure vegetation greenness, a composite property of leaf chlorophyll, leaf area, canopy cover, and canopy architecture. Among the existing indices, the Normalised Difference Vegetation Index (NDVI) is a well-known index used to measure and assess vegetation greenness, growth, and health. NDVI measures the optical reflectance of sunlight by an object in the red and NIR wavelengths, expressed mathematically as:

$$ndvi = \frac{NIR-red}{NIR+red} \quad (\text{Equation 2-1})$$

where Red and NIR stand for the spectral reflectance measurements acquired in the Red (visible) and NIR regions, respectively. This technique can also be extended to extract different land cover features other than vegetation greenness. For instance, NDVI values range between -1 and +1, where values <0 represent waterbodies; values ≥ 0 and ≤ 0.2 correspond to barren areas of rock, sand, snow, or soil; values > 0.2 to < 0.5 represent shrubs and grasslands, and high NDVI values above 0.4 correspond to dense vegetation such as forests or crops at their peak growth stage with high chlorophyll content (Ahmad, 2013; Zhang et al., 2013; Ghandi et al., 2015).

2.7.2.3 Role of NDVI in agriculture

The NDVI technique is a powerful tool used to understand and analyse changes in vegetation cover, addressing climate change impacts, water budget, ecological balance, and necessary conservation measures (Reddy and Prasad, 2018). NDVI helps predict crop yield, assess natural disaster damage, and develop new protection strategies (Ghandi et al., 2015; Davey, 2018). The NDVI index is used for various vegetation properties, including calculating the surface covered in leaves index, chlorophyll concentration in leaves, plant productivity, land coverage fraction with vegetation, and the quantity of water accumulated (Minai, 2015).

Despite its usefulness, NDVI has limitations. It is sensitive to atmospheric conditions such as sunshine, shadow, temperature, wind, aerosol, clouds, precipitation, soil variations, and backgrounds (Matsushita et al., 2017; Huang et al., 2017). Research has also indicated NDVI's limited capacity to predict senescent (stressed) vegetation and in areas with low vegetation cover where exposed soil surfaces may affect radiometric measurements (Tsalyuk et al., 2017). NDVI accuracy decreases with dense vegetation owing to saturation, making canopy variation assessments challenging.

2.7.2.4 Enhanced Vegetation Index

EVI is an optimised vegetation index designed to address NDVI's limitations. It enhances vegetation signals with improved sensitivity in high biomass regions, reduces atmospheric influences, and is more responsive to canopy structure variations, canopy type, plant physiognomy, and canopy architecture (Benedetti et al., 2023). EVI is calculated using

atmospherically corrected bands, a canopy background adjustment factor (L), aerosol correction factors (C1 and C2), and a gain factor (G) with a value of 2.5.

$$EVI = G \times \frac{(NIR - Red)}{NIR + C1 \times Red - C2 \times Blue + L} \quad (\text{Equation 2-2})$$

2.7.2.5 Water Indices

Water is a crucial resource for sustaining life, ecosystems, and human civilisation (Ozelkan, 2020). Effective water resource management is critical and achievable through continuous monitoring using remote sensing technology. This helps decision-makers plan and mitigate droughts, reducing economic losses. Among existing water indices, NDVI, NDWI, and NDMI play vital roles in detecting moisture levels in vegetation, waterbodies, and soil. NDVI is included, as its normalised reflectance values below zero indicate waterbodies. These indices are effective in areas without weather stations and rain gauges, downscaling precipitation data to increase spatial resolution (Ezzine et al., 2017). Vegetation, waterbodies, and soil moisture are sensitive to rainfall and temperature, showing a strong correlation with these weather parameters (Ezzine et al., 2017; Muradyan et al., 2019).

2.7.2.6 McFeeter's Normalised Difference Water Index

The NDWI introduced by McFeeters is commonly used to detect open surface waterbodies, initially created using green and NIR spectral bands of Landsat TM (Ozelkan, 2020). It maximises water reflectance using the green wavelength and minimises low NIR reflectance by water features, enhancing water surfaces while suppressing land and vegetation features (Xu, 2006). Reflectance values range between -1 and 1, with positive values indicating waterbodies and zero or negative values indicating non-waterbodies such as soil and vegetation. It is mathematically explained as:

$$ndwi = \frac{green - NIR}{green + NIR} \quad (\text{Equation 2-3})$$

2.7.2.7 Normalised Difference Moisture Index

The NDMI, also known as Gao's NDWI, assesses water content in vegetation and soil using NIR and SWIR spectral bands. SWIR is sensitive to leaf water content due to electromagnetic

(EM) energy absorption in this wavelength, with reflectance decreasing as water content increases (Ji et al., 2011). Wang et al. (2023) noted that while SWIR is used for plant moisture assessment, it also indicates soil moisture content owing to high reflectivity in this spectral range. High NDMI values may indicate saturated soil moisture, while low values denote low soil moisture. Similarly to other water indices, its reflectance values range from -1 to +1, where values close to negative indicate low-moist areas and values close to one indicate highly moist areas. NDMI applications include vegetation drought monitoring, waterlogged area assessment, crop growth assessment, and more. It is mathematically explained as:

$$ndmi = \frac{NIR-SWIR}{NIR+SWIR} \quad (\text{Equation 2-4})$$

2.8 Machine Learning

The integration of ML and remote sensing technologies has been extensively applied in the agricultural sector to improve crop yield. ML is the ability of machines to learn from the data by extracting and analysing features without strictly being programmed. ML techniques have been extensively applied in the agricultural fields, including bioinformatics, biochemistry, medicine, meteorology, economic sciences, robotics, and many more with commendable results.

The ML process can be categorised into supervised and unsupervised. In supervised learning, the model is presented with training data (input and corresponding output). The model will learn from the data by creating a mathematical rule that will map an input to the output. The model will be evaluated with the test data (training data examples with hidden output) and the trained model will be required to use the knowledge acquired to analyse the input examples from the test data and predict the corresponding output. The error rate will be calculated to determine the model's accuracy. It is computed as a difference between the model's predictions and what is required.

In unsupervised learning, the model is not provided with any data to learn from. It is only provided with data that is unlabelled and it will learn by discovering the hidden patterns from the data.

2.8.1 Supervised Learning

Supervised learning is mainly categorised into classification and regression. In classification, the model is provided with input features and is expected to predict discrete values or classes. This can be binary type of questions where the expected outcome is yes or no, true or false, 0 or 1, and so on. In regression, the model is provided with input features and is expected to predict the continuous values. Unlike classification where the outcome is set to known classes, the regression outcome is not known. Below are the classification and regression models that are widely used.

2.8.1.1 Support Vector Machines

Support Vector Machines (SVM) is a classification technique that tries to separate data in the best possible way. Data can be features extracted from the image or any form of source of information. It uses support vectors as constraints when drawing the separation line and it tries to maximise the separation line in such a way that margins between the support vectors (points close to the boundary surface) are equal. The greater the spacing between categories, the better the generalisation, and the largest classification surface is the optimal classification surface (Wang et al., 2019). Figure 2-2 below is the SVM classification of 2 classes in a two-dimensional (2D) space with a maximised boundary line between vector points of different classes.

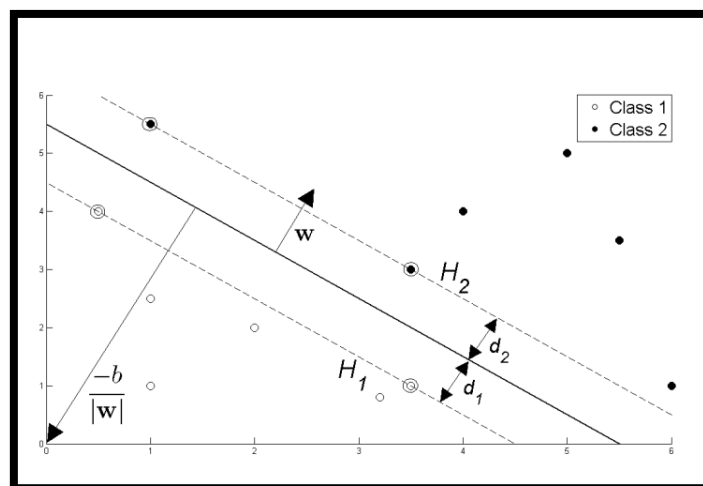


Figure 2- 2: Example of SVM in a 2-dimensional space (Source: Fletcher, 2009).

2.8.1.2 Decision Tree

Decision trees are one of the efficient techniques that are widely used in various areas, like machine learning, image processing, and pattern recognition (Li et al., 2019). The model builds a tree from features, variables or attributes provided to a model during the training phase. The decision-making of the tree is binary (0 or 1) based and it is mainly modelled from top to bottom where the top node is the root, successive nodes are branches, and the terminal nodes are leaves as illustrated in Figure 2-3 below.

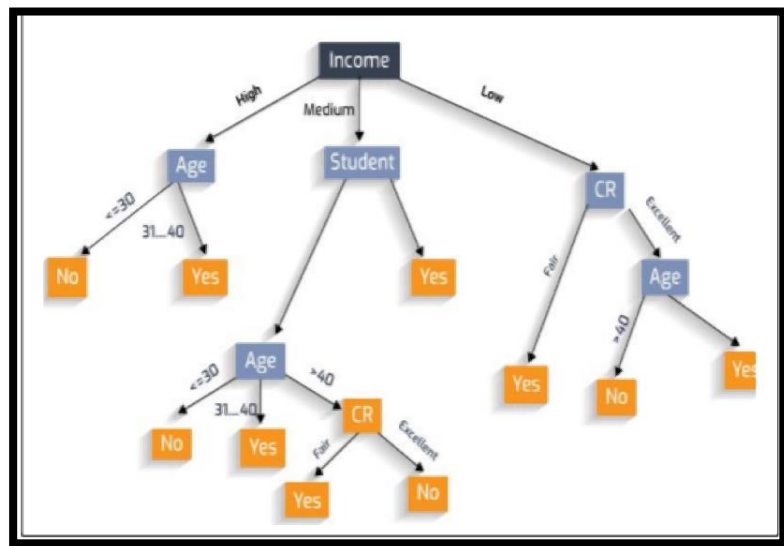


Figure 2- 3: Decision tree classifier example (Source: Charbuty and Abdulazeez, 2021).

The model is capable of drawing decisions using either categorical or numerical data. Given the attributes or features, the model will do some arithmetic operations to figure out the best node that will be used as a root node. The best node is the node that can well split the data towards the desired classification outcome also known as a node with a minimum level of impurity. A root node or competent node has a high information gain and lower entropy level (Lata and Khan, 2020). The information gain and entropy level are computed to determine the level of impurity of nodes as indicated below.

$$\text{Entropy} = H(C) = -p(+)\log_2 p(+) - p(-)\log_2 p(-) \quad (\text{Equation 2-5})$$

where $p(+)$ is the probability of the positive class and $p(-)$ is the probability of the negative class.

$$\text{Gain} = H(S, A) = H(C) - \sum_{V \in VAL} \frac{|S_v|}{|S|} H(S_v) \quad (\text{Equation 2-6})$$

where S represents the subset of a given node (A); S_v is the subset after splitting; $H(C)$ is the entropy level of the subset of the given node, and $H(S_v)$ is the entropy of the subset of the node after splitting.

The entropy level ranges between 0 and 1, where 0 represents complete purity and 1 represents complete impurity. The information gain of a node is the average of the entropy on the right- and left-hand sides of the splits.

2.8.1.3 Random Forest

This algorithm uses multiple trees to improve the accuracy level of individual trees. This algorithm combines the simplicity of decision trees with flexibility resulting in a vast improvement in accuracy. Among various ML algorithms, this has been regarded as one of the most precise prediction methods for classification and regression, as it can model complex interactions among input variables and is relatively robust to outliers (Zhou et al., 2016). Figure 2-4 below is an example of a Random Forest classifier that combines various decision trees to make a judgement.

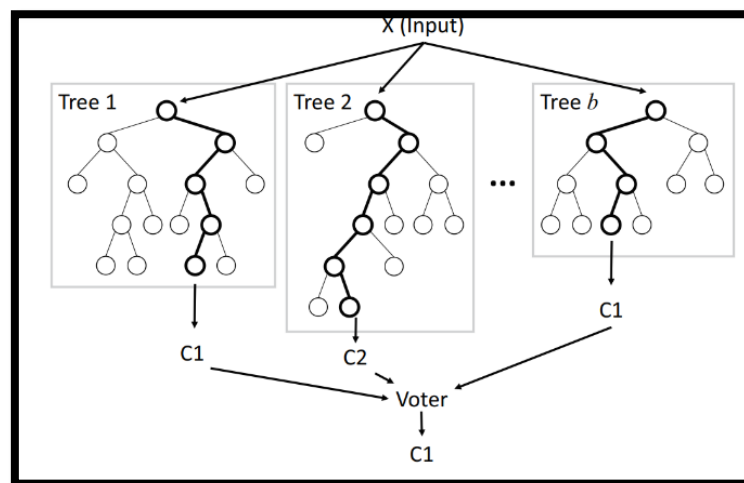


Figure 2- 4: Example of Random Forest Classifier (Source: Nakahara et al.,2017)

The model creates the bootstrap data sets from the original data set provided to a model. It randomly selects samples from the original table to fill the bootstrap table. Some of the samples can be selected more than once while others may not be included in the bootstrap table. The model builds trees from the bootstrap sample by randomly selecting the subset of variables or columns, computing the level of impurity, and selecting the best variable as a root node. The same process will be done for the branches and leaves of the tree. The algorithm will build multiple trees through the process of bagging (bootstrapping aggregating).

The data that were not bootstrapped are known as out-of-bag data sets. The model uses the out-of-bag sample to improve its accuracy. The sample is classified by trees and the final decision will depend on the majority rule (Altman and Krzywinski, 2017). The error rate will be calculated by comparing the model's classification with the known classification. The model will improve its accuracy by repeating the process of a random selection of the subset of variables but with either an increased or decreased number of variables in a subset, building other trees, using out-of-bag samples to validate the accuracy, and selecting the best forest to classify the data.

Using more samples reduces the variance of the fit, but because many bootstrap samples are similar, at some point owing to random selection, more bootstraps will merely increase computation time without improving the estimates (Altman and Krzywinski, 2017).

2.8.1.4 Time Series Models

Forecasting is the process of feeding the machine with historical data and allowing the machine to learn the pattern or the trend from the data to predict new or future values. Most of the time series prediction approaches are based on the analysis of historical data and assume that past and current data patterns can be used to forecast future data points. Forecasting can also be explained as a planning tool to help people in different domains such as politics, economics, econometrics, marketing, agriculture, and so forth in their attempts to cope with the uncertainty of what might happen by manipulating past and current data (Hyndman and Anthanasopoulos, 2018).

Before performing forecasting, time series analysis must progress. Time series analysis is a technique that is used to determine the change of a variable (series) over time. It can also be explained as the analysis of the temporal ordering of observations (Chatfield, 2016). Prema and

Rao (2015) explain the time series as a set of statistical data collected at regular intervals and or sequence of observations on a variable measured over successive points in time. The analysis of the time series data is considered important as it can unleash the hidden characteristics inherent in the time series that will help in identifying a suitable forecasting model.

When analysing the time series data, the statistical characteristics (mean and the variance - cyclical component) of the variable must be evaluated. Both mean and variance can be constant, otherwise mean can change over time but the oscillation of the series around the mean might be constant (constant variance). In other instances, the volatility process on time series can be observed. This is a process where the mean is constant but has non-constant variance. The series is declared complicated when both mean, and variance are not constant.

The time series can be univariate or multivariate. In univariate time series analysis, the focus is on observing the behaviour of a single variable over time. The term "univariate time series" refers to a time series that consists of single (scalar) observations recorded sequentially over equal time increments while a multivariate time series focuses on the observation of multiple variables over time (Aboagye-Sarfo et al., 2018).

This research is limited to only univariate time series. Under univariate time series, the researcher will focus on several models used to learn the pattern in the data and perform forecasting. These models are DTSF, Linear Regression (LR), SMA, WMA, SES, DES, and ARIMA family models.

Decomposition time series (DTS) technology

Decomposition forecasting is a well-known and more understandable model than complicated technologies such as ARIMA family models. The use of this model is motivated by the data that have seasonality also known as fluctuations in the movement of the observations of the series data. This model can account for different types of variabilities such as seasonal change, gradual change, and abrupt change (Wang et al., 2015). It aims to understand the pattern in the time series data by decomposing the series into seasonality and trend components. It makes some mathematical computations using those components to extrapolate future values. For instance, it will multiply or add the components together. It is used in conjunction with other smoothing forecasting models such as linear regression and moving average models that will help to estimate the trend and simultaneously reduce the noise in the data. Once the trend is determined, it will be extracted(removed) from the time series data, and what will be left is

known as seasonal irregularities (cyclical component and the error term). The cyclical component will then be extracted from the seasonal irregularities, leaving the error terms, also known as residuals, which are purely random with no pattern inherent in them. The forecast will be made by either multiplying or adding the trend and the cyclical component together as demonstrated below with Multiplicative Decomposition Time Series (MDTS) and Additive Decomposition Time Series (ADTS) forecasting models.

$$MDTS = seasonal * trend * error \quad (\text{Equation 2-7})$$

$$ADTS = seasona + trend + error \quad (\text{Equation 2-8})$$

MDTS model is used when the data is non-stationary (the series has an increasing or decreasing trend) while ADTS is used when the trend in the time series is constant over time (Prema and Rao, 2015).

Linear Regression (LR) model

Most of the time series models assume that there is a linear correlation of observations of the tested variable. The observations that are stochastically (random) assigned are difficult to analyse and predict. Lopez-Paz (2013) explains that measuring statistical dependence between random variables is a fundamental problem in statistics because of the radically larger amount of possible association patterns.

The linear model attempts to model the relationship between two variables by fitting the linear equation to the observed data. One variable is declared independent (predictor) and another dependent (response). In the case of time series analysis, only one variable (dependent variable) is modelled against its regressed observations (independent variable). The series is linearly correlated when each observation is closed to its regressed observations and there is a persistent increase or decrease direction in the pattern of the data. Forecasting will be done when the best-fit line has been determined. The LR model in time series data is applied when there is a homoscedasticity relationship between two variables. This relationship is observed when there is a finite or constant seasonal variation, and after fitting the model, it results in the constant mean of the error terms (Schmidt and Finan, 2018). Most simply, the homoscedastic of

observations can be observed when there is a minimum deviation of points from the best-fit line or the estimated line, and the error terms are summed to 0.

The model does worse when there is a violation of homoscedasticity, resulting in what is known as heteroscedasticity (non-constant seasonal variations) in the time series, resulting in an inconsistent mean of the error terms (Schmidt and Finan, 2018). Data transformation models such as weighted list squares (WLS) regression are normally used to alter and transform data to homoscedasticity. WLS assigns similar weight to observations with constant seasonal variations and less weight to observations with non-constant seasonal variations to pull or adjust them to become constant (Pregiborne and Hastie, 2017). The relationship between homoscedasticity and heteroscedasticity is centred around the movement of the variance or the fluctuations of the series.

The LR model performs prediction by fitting a straight line across the time series. The model uses the formula ($y=mx+c$) to fit a straight line. The gradient (m) and the intercept (c) of the line are two important coefficients that must be determined on the data before the prediction can be made (Srinivasa and Mupalla, 2015).

The gradient also known as the slope is calculated as

$$m = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} \quad (\text{Equation 2-9})$$

where $\bar{x}$ represents the mean of the x variable and $\bar{y}$ represents the mean of the y variable. The intercept (the starting position of the best-fit line) is calculated as

$$c = \bar{y} - m(\bar{x}) \quad (\text{Equation 2-10})$$

The linear regression model is highly significant and widely used in many forecasting models; however, it tends to underperform with seasonal data due to its lack of flexibility in capturing seasonal variations.

Simple Moving Average (SMA)

This is the simplest method that performs forecasting by calculating an average from the past data towards the recent data. Like any other forecasting model, this model aims to track the trend in the data and simultaneously smooth the data by reducing the seasonal effects for accuracy and prediction purposes. From the declaration of Azami et al. (2012) and Prema and

Rao (2015), SMA is the easiest digital filter to understand and use, and it is the most common filter to use in the analysis of a signal.

The model has to be parameterised with the number of steps it should take towards forecasting future values. The algorithm will start to calculate from the past data towards the recent data. It will compute the moving average by dropping the oldest observations and including the next observation (Prema and Rao, 2015). For example, if the four-step moving average is utilised, the model will start to forecast at the 5th period. This forecast will be calculated as the average of the four regressed periods (t_4, t_3, t_2, t_1). The forecast for the 6th period will be the average of the subset of the four regressed observations (t_5, t_4, t_3, t_2) where there is an offset of one period relative to each other.

The limitation of this model lies in its forecasting approach, which focuses only on recent data while disregarding historical data. In addition, the model applies equal weights to the regressed data when predicting the next observation. The formula is :

$$f_t = \frac{1}{k} \sum_{j=1}^k Y_{t-j} \quad (\text{Equation 2-11})$$

where k represents the order or the step of the moving average, f_t is the current forecast and Y_t is the current observation.

Weighted Moving Average

This model like SMA, makes use of the order movement. The predicted value is highly related to the subset of the previous values. However, each regressed data point in each subset is manually assigned a higher weight that is within the range of 0 to 1, and it will decay towards the oldest observation in the subset. This approach is based on the philosophy that in linearly correlated data, recent observations are more closely related to future forecasts, and should therefore be prioritised over older observations.

Conventionally, these weights are summed to 1. The weakness of this model, just like that of SMA, is that it focuses only on the recent subset of data, ignoring the past data. The weights are determined among the users. The formula is :

$$f_t = \frac{1}{k} \sum_{j=1, r=1}^k Y_{t-j} W_r \quad (\text{Equation 2-12})$$

where W is the array of user-defined weights based on the order process of the WMA.

Single Exponential Smoothing

Unlike Simple Moving Average and Weighted Moving Average models that focus only on the given window of observations, the SES considers all the observations of the time series to generate the next forecast. Ravinder (2013) explains the SES as one of the simplest and most popular forecasting techniques that are included in popular forecasting software products. Like any other forecasting model, it prioritises the recent data by utilising a weighting scheme that assigns high weight to recent data and decays as observations get older. It automates the weight assignment process, and it does not utilise the order movement and its forecasting mechanism does not ignore the past data. To automate the weight assignment, it utilises the summation of exponents and a constant known as alpha that is assigned a value greater than 0 and less than 1.

The formula for simple exponential smoothing is presented below, where f_{t+1} is the next forecasted value; Y_t is the constant term, and F_t is the current forecast:

$$f_{t+1} = \alpha Y_t + \alpha(1 - \alpha)F_t \quad (\text{Equation 2-13})$$

The formula can be expanded up until the nth period as illustrated below.

$$f_{t+1} = \alpha F_t + \alpha(1 - \alpha)F_{t-1} + \alpha(1 - \alpha)^2 F_{t-2} + \dots + \alpha(1 - \alpha)^{n-1} F_1 + (1 - \alpha)^n f_t \quad (\text{Equation 2-14})$$

The formula can be generalised as:

$$f_{t+1} = \alpha F_t + \sum_{i=1}^n \alpha(1 - \alpha)^i F_{t-i} \quad (\text{Equation 2-15})$$

The major disadvantage of the SES is a failure to account for the time series with the trend and seasonality effects.

2.8.1.4.1 Holt's Linear Method

This model is also known as the Double Exponential Smoothing (DES) model because it takes care of the level and trend components of the time series. This method, introduced by Charles Holt in 1957, is an extension of Single Exponential Smoothing (SES) to Linear Exponential Smoothing (LES) to allow forecasting of data with trends. The forecast for this method is performed by first estimating the two smoothing constants (α and β) that range between 0 and 1. The formula is:

$$f_{t+1} = L_t + T_t = [\alpha f_t + (1 - \alpha)(L_{t-1} + T_{t-1})] + [\beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1}] \quad (\text{Equation 2-16})$$

where L_t is the level ; T_t is the trend, and α and β are the smoothing constants for the level and trend respectively.

The more powerful time series algorithms such as Triple Exponential Smoothing, LSTM, and SARIMA models are discussed later in this thesis.

2.8.2 Unsupervised Learning

The unsupervised learning technique is a process in which the machine comes into contact with data objects and classifies them based on some hidden features they possess. Clustering is defined as an unsupervised learning where the objects are grouped based on some similarity inherent in them. This approach is similar to the classification approach. However, the difference is that in classification, the class (label) of an object (pattern) is known and it is given in advance to a model whereas in clustering problem, the class label is not tagged to an object and the machine must figure out the class to which an object belongs by manipulating its properties or characteristics. Clustering is a very essential component of various data analysis or machine learning-based applications like regression, prediction, data mining, and so forth (Saxena et al., 2017). It is considered to be more difficult than supervised classification as there is no prior knowledge given to the classifier about the data objects. The main aim of data clustering is to search the real grouping(s) of a set of instances, points, or objects. The clustering can also be explained as a statistical classification method for finding whether each of the patterns comes into various groups by making quantitative comparisons of different features. In most cases, the number of clusters or groupings that must be performed by the machine is user-defined. Furthermore, in most scenarios, the features of the objects are numeric. The most common approach to define the similarity is measuring the distance among the objects. The lower the distance between the two objects, the higher the similarity, and vice versa. The K-Means, K-Mode, and K-prototype clustering models are discussed later in this report.

2.9 Optimisation

Optimisation is also known as mathematical programming. It is a collection of mathematical principles and methods used for solving quantitative problems in various disciplines such as physics, mathematics, computer science, economics, and many more. It involves the selection of the best solution out of the set of possible solutions based on the global optimal value in the

entire region that is restricted by a set of constraints. Optimisation techniques are highly applied where there is a need to maximise profit and minimise costs by efficiently and effectively utilising the resources available. The optimisation techniques are used in solving combinatorial problems such as traveling salesman problems, job-shop scheduling problems, vehicle routing problems, graph colouring problems, network routing problems, timetabling, finding an optimal location of an industry, production planning, controlling and scheduling, and many more (Deng et al., 2019). Furthermore, the optimisation models are used in digital image processing and compression, fault detection and feature selection, trial neural network, classification and clustering, parameter optimisation of SVM, scheduling, data analysis and mining, finance and economics, cost reduction, digital signal processing, constraint mechanical design optimisation problems, thermal, electrical, chemistry and chemical, mechanical, civil engineering, and many more disciplines (Khan and Sighh, 2017). Before optimisation takes place, there must be a clear formulation of the optimisation problem associated with specific decision variables, objectives, constraints, and variable bounds (Branke et al., 2014). Mathematically, optimisation involves maximising or minimising the objective function ($f(x)$) using decision variable(s) that can either be continuous or discrete, that is bounded or unbounded, and further constraint to other limitations. The optimal value is determined within the feasible region known as a region where constraints of all decision variables are satisfied. A formula below is an example of the optimisation problem that needs to be maximised given decision variables (x_1 and x_2) that are bounded between 0 and 30 and further subjected to the following constraints when plotted into these formulas: $\frac{\pi x_1^2 x_2}{4} \geq 300$ and $x_1 - 3x_2 \geq 400$.

$$\begin{array}{ll}
 \text{Max } f(x) = c \left[\frac{\pi x_1^2}{\sin 2} * \sqrt{\pi x_1 x_2} \right] & \\
 \text{s.t.} & \frac{\pi x_1^2 x_2}{4} \geq 300 \\
 & x_1 - 3x_2 \leq 400 \\
 & 0 \leq x_1 \leq 30 \\
 & 0 \leq x_2 \leq 30
 \end{array} \quad \left. \vphantom{\begin{array}{l} \text{Max } f(x) = c \left[\frac{\pi x_1^2}{\sin 2} * \sqrt{\pi x_1 x_2} \right] \\ \text{s.t.} \end{array}} \right\} \text{(Equation 2-17)}$$

Even if the optimal value is obtained, when there is a violation of at least one constraint, the solution is declared infeasible.

The decision variable can either be continuous or discrete. The continuous variable can take any value within the finite or infinite interval, while the discrete variable is selective (can take distinct, separate, unconnected or fixed values). Below is the generic presentation of the optimisation problem that can take either continuous variable (x) or discrete variable (y) subjected to either inequality constraints ($g(x, y) \leq 0$) or equality constraints $h(x, y) = 0$ and either bounded ($x^m \leq x \leq x^n$) or unbounded ($x \in R$).

$$\begin{aligned}
 & \text{Min } Z = F(x, y) \\
 & \text{S.t. } \quad g(x, y) \leq 0 \\
 & \quad \quad h(x, y) = 0 \\
 & \quad \quad x \in X, y \in Y \\
 & \quad \quad X = \{x | x \in R^n, x^l \leq x \leq x^n\} \\
 & \quad \quad Y = \{y | y \in R^n, x^l \leq y \leq x^n\}
 \end{aligned}
 \quad \left. \vphantom{\begin{aligned} & \text{Min } Z = F(x, y) \\ & \text{S.t. } \quad g(x, y) \leq 0 \\ & \quad \quad h(x, y) = 0 \\ & \quad \quad x \in X, y \in Y \\ & \quad \quad X = \{x | x \in R^n, x^l \leq x \leq x^n\} \\ & \quad \quad Y = \{y | y \in R^n, x^l \leq y \leq x^n\} \end{aligned}} \right\} \text{(Equation 2-18)}$$

The objective function responds differently depending on whether the decision variable increases or decreases. An objective function is considered monotonic if it consistently increases or decreases in response to changes in the decision variable. Furthermore, the objective function can be declared mainly as unimodal, and multimodal. Unimodal function is observed when there is only one optimal peak on an entire feasible region, while multimodal involves multiple peaks on a feasible region. The multimodal function is observed when there are fluctuations in the movement of the optimisation function based on the increase or decrease of the decision variable. The multimodal function has local optima (optimal values in a neighbourhood region) and global optimum (an optimal value in an entire region). The global optimum is a desired optimum and many algorithms are failing to determine it (Mott et al., 2017 and Deng et al., 2019). Real-world optimisation problems are usually multimodal and require optimisation algorithms to keep a balance between exploration and exploitation, and it is one of the main opportunities as well as one of the main challenges for algorithms (Li and Tan, 2017).

When there are conflicts of constraints, the feasible region cannot be determined and, as a result, the optimal value can also not be determined (Flippaki et al., 2011). The multi-objective optimisation problem is also a challenge especially where satisfaction of one objective is a

deterioration of another. In this case, there is more than one solution that cannot eliminate others, known as Pareto optimal solutions or non-dominating solutions (Branke et al., 2014). However, only one solution needs to be selected.

2.10 Cropping Decision Optimisation using Remote Sensing and ML

Cropping decision optimisation is primarily conducted using mathematical models; however, this thesis focuses on the optimisation techniques employed under the umbrella of machine learning and remote sensing techniques. The driving factors behind the optimisation of cropping decisions and crop production include the continuous decline in food production caused by climate change and the rapid rise in the human population and land degradation. Some of the technologies developed are highlighted below.

Adebiyi et al. (2020) developed a machine learning-aided mobile system for farmland optimisation using various inputs such as location, crop type, soil pH, and farm size. The system employs a Random Tree Forest (RTF) classifier and BigML to evaluate inputs and produce outputs such as crop class, soil type suitability, crop recommendations, cultivation directions, and crop yield estimations. This system showed high accuracy owing to the utilisation of multiple decision trees, enabling farmers to make informed decisions based on numerous parameters.

Lata and Khan (2020) used seasonal environmental data (rainfall, temperature, and humidity) to develop and train a decision tree algorithm for estimating different crop yields. The model estimates low or high yields based on input environmental data. During training, information gain and entropy levels were computed to determine the best feature for the root node. Rainfall was used as the root node, effectively classifying crops that rely heavily on rain.

Prasad et al. (2016) developed a crop recommendation system using CNN and SVM that takes soil images as input and produces soil class type as output. The soil class type, combined with geographic parameters like latitude and longitude, is fed into SVM to recommend suitable crops a farmer can grow. However, the system's accuracy was limited because of insufficient training data and neural networks.

De Castro et al. (2017) developed a robust and innovative automatic object-based image analysis (OBIA) algorithm utilising Unmanned Aerial Vehicle (UAV) images to detect weeds between and within crop rows during the early growing stages. Features such as NIR/G, mean

Red, mean Green, mean NIR, brightness, and NIR standard deviation were extracted for every weed and crop object in the training set. The RTF model achieved high Weed Detection Accuracy (WDA) in most classified maps, indicating accurate weed identification and spatial location.

Akbani et al. (2020) used RTF and Sentinel-2 satellite images to map and monitor different crop types using various feature sets, including spectral bands, spectral indices, variations of spectral indices, texture, and feature combinations. The system achieved over 90% overall accuracy with a Kappa coefficient of over 0.8.

Zhou et al. (2016) utilised the Random Forest Regression (RFR) algorithm and remote sensing data to estimate wheat biomass using spectral vegetation indices as features for the RFR model. The RFR model produced more accurate estimates than SVM and Artificial Neural Networks (ANN) for every test performed.

Xu et al. (2021) developed a cotton yield estimation model based on time series Unmanned Aerial Vehicle (UAV) remote sensing data. The model developed was a Bayesian regularization BP (backpropagation) neural network that receives various spectral indices including NDVI, EVI, Visible-band Difference Vegetation Index (VDVI), and others as input to the network. The network makes some computations and predict the cotton yield. The correlation coefficient (R^2) of 0.854 was obtained when comparing the actual and predicted yields.

Meraj et al. (2022) utilised Random Forest (RF) and SVM classifiers to estimate wheat acreage and yield in one of the largest wheat-producing states in India (Uttar Pradesh) using Sentinel 2 satellite data, temperature, rainfall, and solar energy data as input features to the classifiers. The results revealed a magnificent performance of RF with a kappa hat coefficient of 0.84 in contrast with SVM which produced the kappa hat coefficient of 0.68.

The study of Wang et al. (2022) examined the integration of ML, deep learning, and remotely sensed time series data for identifying multiple crop type classification over a large-scale region. The NDVI, EVI, and NaE (combined NDVI and EVI) time series data were adopted as input features. Further, the ML models (RF, SVM, KNN, and their integration (Stacking)) and deep learning models (One-dimensional CNN and LSTM) were compared to examine the performance of the classification of multiple crop types in the North China Plain. The NaE was identified as an optimal feature to use towards the desired classification results. Furthermore,

the Stacking ML model (that integrates RF, SVM, and KNN) outperformed other algorithms with the highest accuracy of 77.12%.

Crop diseases account for a significant percentage of crop loss globally, and early detection can help minimise losses. Rangaragan (2018) developed a tomato disease classification algorithm using deep learning architectures, namely AlexNet and VGG16Net. The models were trained using images of tomato leaves from the Pant Village dataset, achieving classification accuracies of 97.49% and 97.23% respectively for 13,626 images.

Picon et al. (2014) developed a mobile application to detect and classify different crop diseases (Septoria, Tan spots, and Rust) at early stages using Deep CNN. The model was trained and tested using 8,178 images of crop diseases. It achieved an overall accuracy of 87% for early and late disease stages.

2.11 Knowledge gap

Currently, there are no crop production optimisation models that integrate the use of indigenous knowledge with ML, climate data, and satellite imagery technologies to cater for small-scale farmers in their local settings. This gap highlights the need for a comprehensive approach that combines traditional knowledge with advanced technological tools to enhance cropping decision-making processes for small-scale farmers.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter is the summary of the research plan followed to optimise farmers' cropping decisions. In a nutshell, the chapter discusses the research design followed to collect the data, the type of data collected, the population and study area where the data were collected, the instruments used to collect the data, and approaches used to analyse the data. This chapter further introduces the framework followed to integrate the primary data with the scientific data and models to optimise farmers' cropping decisions.

3.2 Research Design

Research is an effort to discover a logical solution to a problem identified by a researcher systematically and scientifically. It can also be elucidated as an art of scientific investigation towards answering a research question, solving a problem, or generating new knowledge through the systematic and orderly collection, organisation, and analysis of information with an ultimate goal of making the research useful in decision making (Ho et al., 2006 and Boon and Ballen, 2018)). The methodology is the rationale behind the research and the lens through which the results will be analysed. The research methodology is a set of specific procedures or methods used to identify, select, process, and analyse information about the topic or the problem identified (Wilkinson, 2002). Several research methodology designs are available such as basic research design, correlational research design, explanatory research design, and applied research design. The basic research attempts to generate and expand the fundamental knowledge about the social world and it has no practical value or little direct impact on action, performance, or policy decision. Applied research seeks to solve practical problems by systematically looking at a problem and trying various solutions to see how effective they are. Correlational research refers to the systematic investigation or statistical study of relationships among two or more variables without necessarily determining cause or effect. Explanatory research is similar to correlational research but in this case, the researcher tries to find the cause-and-effect relationship between the tested variables through conducting experiments.

Given the nature of this research, the action-based (applied) research design was undertaken with the rationale of measuring the effectiveness of the IK in supporting farmers' cropping

decisions, identifying the existing gaps, and then trying to bridge those gaps by integrating IK with the scientific knowledge to optimise farmers' cropping decisions.

3.3 Research paradigm

Research paradigm or philosophy refers to a way of understanding the reality of the world and studying it (Arman and Alharthi, 2017). There are two major research paradigms, namely: quantitative or positivism and qualitative or constructivism. The quantitative research paradigm or positivism research philosophy is described as a good explanation of the phenomena by collecting numerical data that are analysed using a mathematical-based approach. This paradigm is attributed to the scientific truth, experimentation, facts, objectivity, and quantitative methodology. It supports the philosophy of strict empiricism by exploring real situations for the prediction and formulation of general laws. The ontological belief of this paradigm is that reality is always tangible, and the epistemological belief of this paradigm is that reality can be gained through a scientific approach. In summary, the belief behind positivism or quantitative research paradigm is that only one truth exists (realism ontology), and that it can be measured (positivism epistemology).

The constructivism, qualitative or interpretivism research paradigm is described as research that aims at gathering insight from the phenomena that can be explained in words rather than in numbers. This is a philosophical perspective interested in finding ways in which human beings individually or collectively interpret or construct the social and philosophical world in specific social, linguistic, and or historical contexts (Sha and Bargi, 2013)). It is based on critical reality, and it combines empirical evidence with logical reasoning. The ontological belief of this paradigm is that more than one truth exists and it can be interpreted rather than measured.

In this research, the mixed-method research paradigm, also known as pragmatism research philosophy, was adopted where both positivism and constructivism research paradigms are compounded together to measure (quantitative) the effectiveness of the IK in support of farmers' cropping decisions, identify the existing gaps, then attempt to bridge the existing gaps by exploring (qualitative or explorative) the use of IK to support farmers' cropping decisions and try to find possible ways to integrate IK (which is qualitative) with scientific knowledge (which is quantitative) to optimise farmers' cropping decisions.

3.4 Data Collection Methods

Data collection involves systematically gathering information on relevant variables from appropriate sources to address research questions, test hypotheses, and evaluate results (Zageye et al., 2009). Data collection methods are devices for obtaining information relevant to the given research project and can be categorised into primary and secondary data collection methods. The primary data collection methods are divided into two groups, namely: quantitative and qualitative. In this research, the focus group data collection method was used to gather qualitative data while the survey data collection method was used to gather both qualitative and quantitative data. The secondary data are the data that are readily available from various sources such as websites, books, and newspapers among other sources. In this research, the historical weather and satellite imagery data are the secondary data collected, processed, and integrated with the primary data to generate new knowledge. The weather data were collected from the district weather station close to the selected region of interest while the satellite imagery data were collected within the region of interest.

3.4.1 Focus group interviews

Focus group discussion is a qualitative technique involving the use of in-depth group interviews in which participants are selected on the criteria that they would have something to say on the topic, are within the age range, have similar socio-characteristics and would be comfortable talking to the interviewer and each other (Rabiee, 2004). The selection of this method was based on exploring the use of the IK in predicting the rainy season and monitoring the health and growth of crops.

3.4.2 Survey

Surveys are commonly classified as descriptive research methods in which respondents are asked a series of questions in a standard manner so that responses can be easily quantified and analysed statistically. This technique was selected to quantitatively assess the effectiveness of the IK in predicting rainy season behaviour and monitoring crop health and growth. The survey was also used to collect qualitative data from the respondents who were not part of the focus group discussion.

3.5 Data Types

Two different types of data were collected from this research. These are primary and secondary data. The procedures followed to collect, process and analyse the data are discussed in the next chapter.

3.6 Target Population

The target population are total group of individuals from which the sample is drawn. In this research, the target population is small-scale farmers in uMgungundlovu district KZN South Africa, who primarily do crop farming for either subsistence or market purposes, and rely heavily on rainfed agriculture and the IK system to support their cropping decisions.

3.7 Research sampling

The sample is a group of individuals who participate in the research. Research sampling is a statistical procedure to identify the group of participants from the target population. It can be categorised into probability and non-probability. Probability sampling means the sample is selected using random methods. Non-probability sampling means the sample is selected in a non-random way.

The non-probability (purposive) data sampling method was used to select the small-scale farmers in Swayimane region in uMgungundlovu district that primarily do crop farming and rely heavily on rain-fed agriculture and IK to support their cropping decisions.

3.8 Reliability and Validity

Reliability is concerned with the consistency of the measuring tool while validity is about whether a tool measures what is supposed to measure. These two components work together to ensure that knowledge or conclusions that will be made and generalised to the population are objective and free from bias. If the measuring tool is reliable, it will continue to produce the same results repeatedly in different spatial and temporal intervals. However, it will be pointless if it measures the data that are not valid. Similarly, data measured with the non-reliable tool, do not help much. In this research, these two concepts are addressed in methods used to gather and process both the primary and secondary data. The pilot study approach was used to minimise ambiguity between the respondents and what was required by the researcher during the primary data collection. Both qualitative and quantitative data collected from the respondents were carefully documented and analysed.

As for the secondary data collection, the historic weather data, collected from the district weather station, were prepared by filling in the empty cells. This was achieved by employing ML imputation methods and validating their accuracy to come up with the ones that can well impute the data. The collected satellite images were processed by reducing atmospheric effects as well as systematic and non-systematic distortions. The cloudy images were also corrected with the help of other images. The algorithms deployed to perform season behaviour predictions were also evaluated on the weather and satellite data, and the ones with high accuracy were used to make future predictions.

3.9 Study Area

3.9.1 uMgungundlovu District Municipality

Pietermaritzburg, also known as uMgungundlovu District, is situated in the KZN Province in South Africa's eastern seaboard. It is one of the 10 municipal districts that make up the province and it shares borders with Mozambique, Eswatini, and Lesotho. The uMgungundlovu District Municipality is classified as a Category C municipality and is in the KwaZulu-Natal Midlands. The name uMgungundlovu originates from the Zulu phrase *umgungu we ndlovu*, meaning "the secret conclave of the elephant". The district includes seven local municipalities, namely: Msunduzi, Impendle, Umshwathi, Mkhambathini, Mpofana, Umngeni, and Richmond local municipality. The province receives an average annual rainfall ranging from 650 to 1,400 mm and has highly cultivated lands covering 6.5 million hectares suitable for farming. Sugarcane is the province's primary crop, contributing over 80% of South Africa's total sugar production (Buthelezi et al., 2024). The province is endowed with ample water resources from six major rivers and five significant dams. This abundance of water is one of the reasons uMgungundlovu ranks among the top regions in the country for agricultural yield potential. The Swayimane region, where the study was conducted, is situated approximately 65 km east of Pietermaritzburg in the uMshwathi local municipality. Covering an area of 32 km², this rural community relies heavily on farming as a primary economic activity, supporting numerous livelihoods both within the community and in neighbouring areas such as Mambedwini, Engoleleni, KwaHlwemini, and others. The region is primarily used for agricultural production due to its favourable rainfall (500 to 800 mm annually) and fertile soils. Figure 3-1 below shows the Swayimane region in KZN, which is highlighted with light grey.

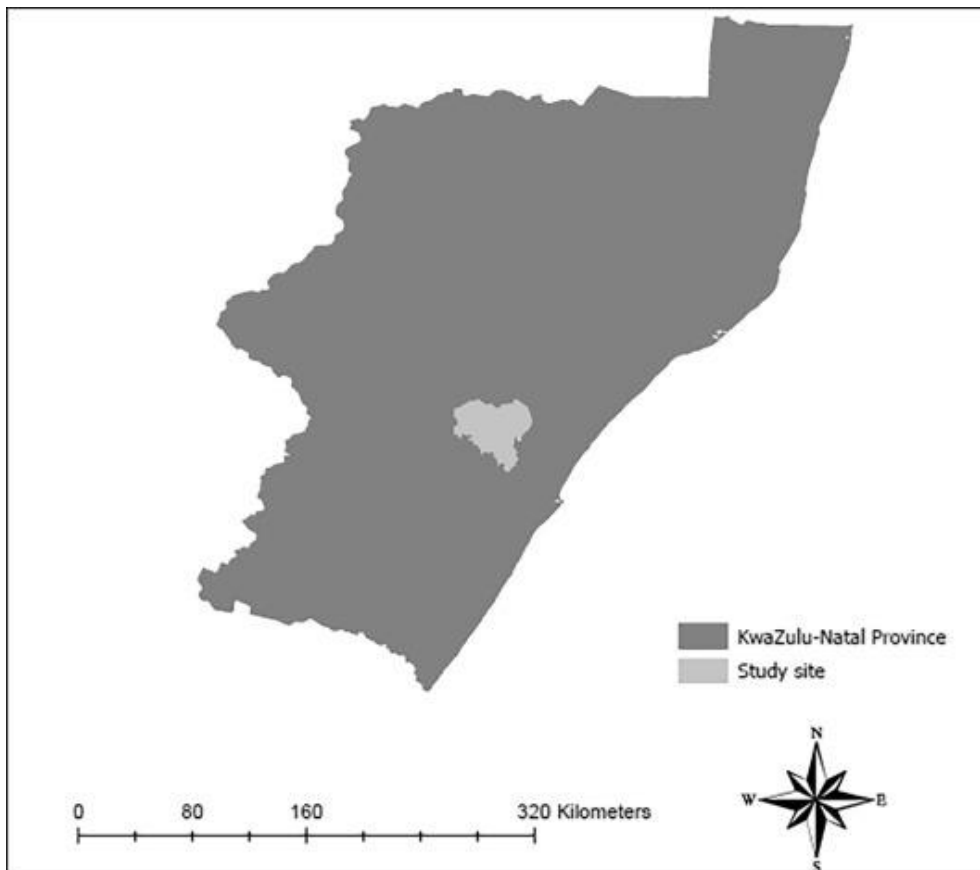


Figure 3- 1: Kwazulu-Natal province and the study site (Source: Author).

3.10 The proposed research framework

The proposed framework is presented in Figure 3-2 below.

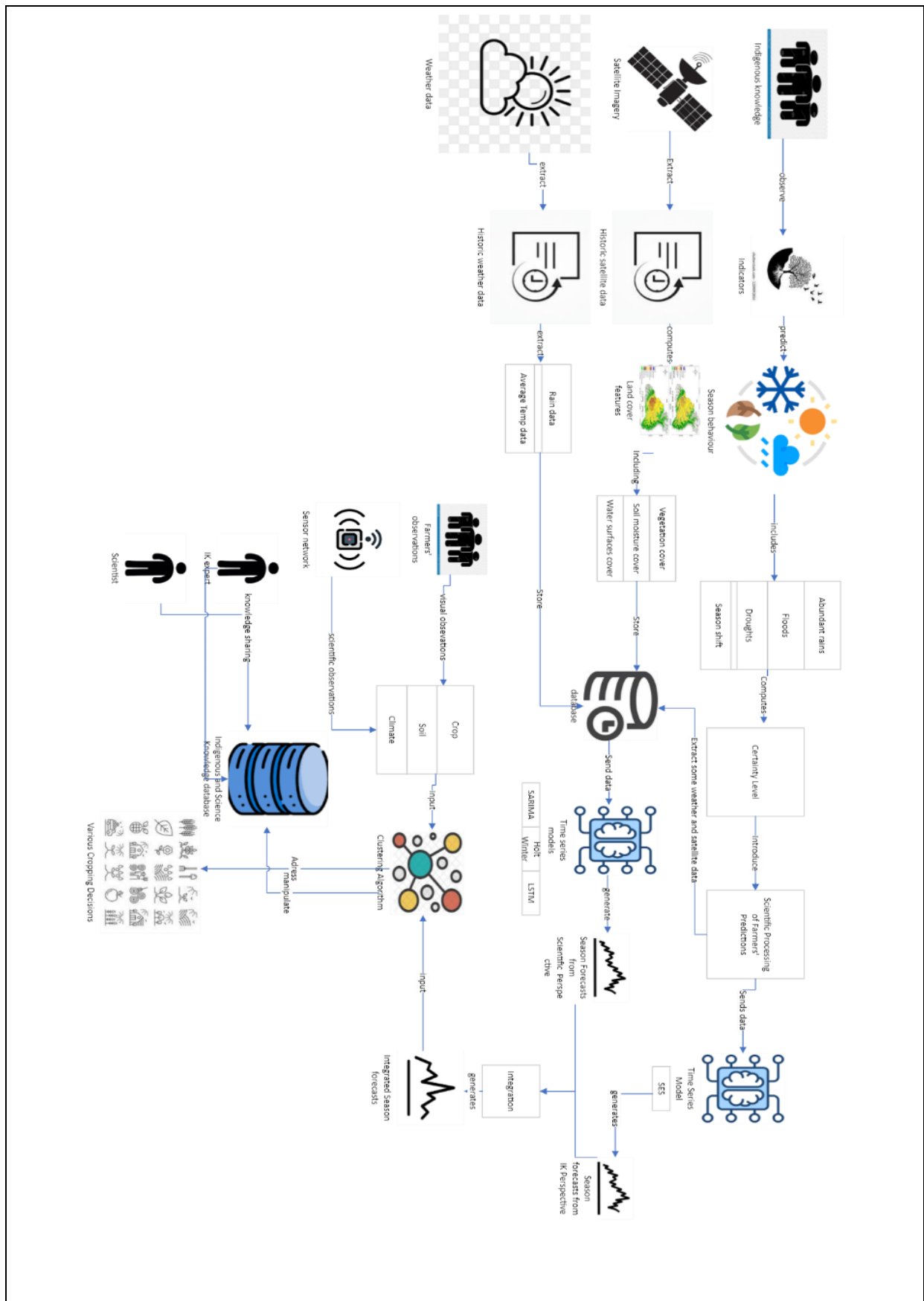


Figure 3- 2: The framework followed to optimise farmers' cropping decisions (Source: Author).

The framework above summarises the methodology undertaken to optimise the farmers' cropping decisions. This was achieved by first collecting historic weather and satellite imagery data from various sources. The weather data are initially prepared by imputing the missing data using imputation models. The satellite imagery data are prepared by extracting the pixel intensity values that represent vegetation greenness cover, waterbodies cover, and soil moisture cover for every image. Both the processed weather and satellite data are passed to the univariate time-series models that will be trained and validated on the data. The best-performing models are used to make the next rainy season behaviour predictions. These predictions will then represent the scientific perspective. As for the IK, the mobile-based application was developed to enable a farmer to register the IK indicators that they use to foresee the behaviour of rains (which can either be excessive, scarce, or abundant (normal or enough rains to enable good establishment of crops)) and temperature changes throughout the rainy season. The indicators observed are systematically processed to calculate the farmers' certainty level (CL) of the anticipated rainy season behaviour. For instance, if farmers are certain of abundant rains to come, the historic weather (rainfall and temperature) and satellite imagery data (water surface cover, vegetation greenness cover, and soil moisture cover) of periods when the abundant rains were experienced will be extracted and passed to the SES time series model to perform the next season behaviour predictions. These predictions will represent the IK's perspective. The season behaviour predictions made via the IK and the ones made via the scientific perspective will be integrated to come up with more robust and localised season behaviour predictions or forecasts. These forecasts are further passed to the algorithm that will describe every data point in simple words that farmers can understand. The generated forecasts will assist farmers in understanding the possible onset, distribution, cessation, intensity, duration, and amount of rainfall, as well as temperature patterns. This knowledge is deemed to improve farmers' cropping decisions. However, given the crop complications intensified by climate change, the ML optimisation model was developed to guide farmers in monitoring the health and growth of their crops. It receives as an input, farmers' observations (this includes season, crop, soil, and climate information) and data from the soil and climate sensors. This model will use the knowledge it has learned from the IK and science experts, and share recommendations on various topics such as crops a farmer can grow, fertiliser selection, treatment for various crop health challenges, and many more.

CHAPTER 4: CROPPING DECISIONS' OPTIMISATION

4.1 Introduction

This chapter dwells on processes followed to optimise cropping decisions for small-scale farmers. This is achieved by integrating the IK with science (weather data, satellite imagery, and ML time series models) to predict rainy season behaviour. These predictions aimed to provide a farmer with knowledge on the possible onset, distribution, quantity, magnitude, duration, and cessation of rains and temperature behaviour throughout the season. This knowledge will help farmers to employ all necessary strategies or methods to obtain better crop yield. However, given the wide range of crop health challenges arising because of climate change, the researcher further develops the ML optimisation model that will continually learn from indigenous and science experts with the support of wireless sensors. This model will share recommendations to help farmers with diverse knowledge to tackle various cropping challenges that they may encounter such as crop selection, fertiliser selection, pests and diseases, and other crop and soil health-related challenges. The model receives as an input, farmers' observations, as well as soil and climate readings from wireless sensors. It then uses the knowledge acquired to generate a set of recommendations to answer the question (input) provided.

4.2 Procedures followed to optimise farmers' cropping decisions

The action-based (applied) research methodology was conducted to evaluate the effectiveness of the IK in supporting farmers' cropping decisions, identify the existing gaps, and try to patch or bridge those gaps with the help of scientific data and approaches. The IK system is primarily used by farmers to predict seasonal changes, monitor rainfall patterns during the agricultural season, and track the health and growth of crops, including early diagnosis and treatment of various crop problems.

To evaluate and strengthen the effectiveness of IK, both the primary and secondary data were collected, processed and integrated to optimise farmers' cropping decisions.

4.2.1 Primary data collection and processing

The primary data were collected using both qualitative and quantitative data collection methods. The quantitative method was initially applied to evaluate the effectiveness of the IK in rainy season behaviour predictions and monitoring the health and growth of crops. The knowledge gained helped the researcher to understand the weaknesses and strengths of the IK. The qualitative data collection methods were used as a second tool to get in-depth knowledge on how the IK is used by farmers in predicting agricultural season behaviour and monitoring the health and growth of crops. The knowledge gained from the qualitative methods helped in developing a prototype that will integrate the IK with science to optimise the farmers' cropping decisions.

4.2.1.1. Quantitative data collection

The survey data collection method was employed to assess the effectiveness of indigenous knowledge in predicting rainy season patterns and monitoring crop health and growth. The pilot study was initially conducted on a small number of farmers before the actual dissemination of the questionnaire to a larger population. The motive behind the pilot study was to identify and clarify questions that were ambiguous to the respondents. Afterwards, 100 questionnaires were disseminated to farmers in Swayimane region in the uMgungundlovu district. The region was chosen because it is a well-known area in KZN with many smallholder farmers who depend heavily on rainfall to grow their crops and who continue to use the IK system to guide their cropping decisions. The questionnaire was primarily comprised of close-ended questions to numerically assess the effectiveness of the IK in predicting rainy season behaviour and monitoring the health and growth of the crops. Out of 100 questionnaires that were disseminated, 12 questionnaires were disregarded owing to their incompleteness and inability to contribute towards the objective of the research.

4.2.1.2. Quantitative data processing and analysis

The Microsoft Office Excel 2016 application was used to process the data collected from the questionnaires. The data that were processed included demographic data and close-ended questions. Section A of the questionnaire focused on demographic information such as farmers' gender, age group, educational level, and years they have been in the study region among other elements. Section B aimed to evaluate the effectiveness of the IK in rainy season behaviour predictions and monitoring the health and growth of crops.

Section A: Demographic distribution of the respondents

The gender of the respondents was as follows: 42 respondents accounting for 47.73% of the sample were females, 35 (39.77%) respondents were males, 6 (6.81%) respondents described their gender as *other* while 5 (5.68%) respondents ignored the question. Only 10 (11.36%) respondents fell in the age bracket of 18 to 35 years; 25 (28.41%) were in the age bracket of 36 to 45 years; 35 (39.77%) were in the age bracket of 46 to 55 years; 15 (17.05%) were in the age bracket of 56 to 65, and 3 (03.41%) respondents were in the age bracket of 66 and above.

Additionally, 11.36% of the respondents never attended school; 17.05% ended in primary education; 54.55% ended in secondary school, and 17.05% reached a post-secondary level of education. Respondents were asked to report the number of years they had been in the study area. Of these, 6.82% indicated less than 5 years; 11.36% reported between 5 and 10 years; 34.09% had spent between 11 and 20 years in the area, while the remaining 47.73% stated they had been in the study area for over 20 years. Lastly, 59.09% of the respondents claimed to have smartphones; 29.55% of the respondents claimed to have non-smartphones, and 11.36% of the respondents claimed not to have phones at all.

Section B1: Indigenous Knowledge in Rainy Season Behaviour Predictions

The first question evaluated the effectiveness of the IK in predicting the season with good rains and temperature before a farmer can grow their crops. The results of the first question are presented in Figure 4-1 below.

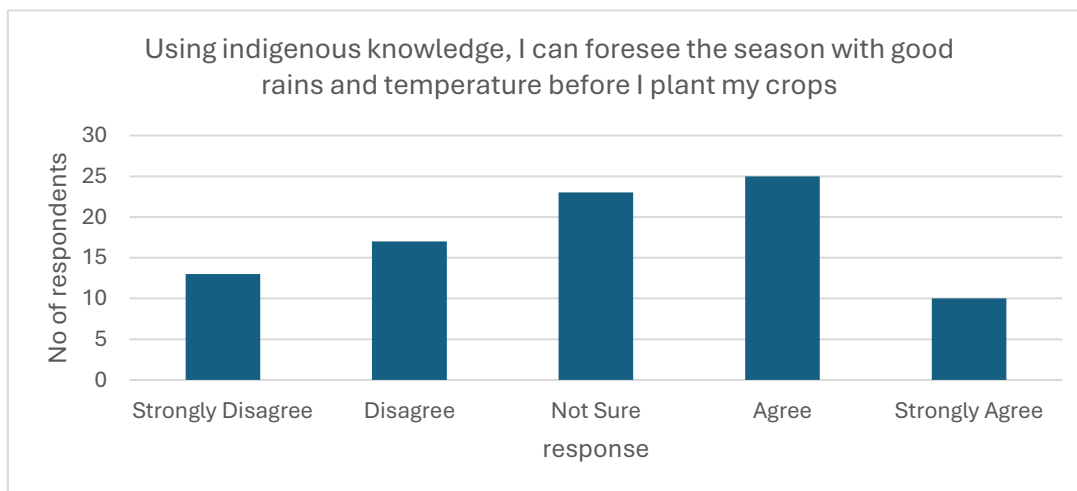


Figure 4- 1: Presentation of respondents' results regarding foreseeing the season with favourable weather (Source Author).

From Figure 4-1 above, the prompt was: Using indigenous knowledge, I can foresee the season with good rains and temperature before I plant my crops. The results indicate that 13 (14.77%) respondents strongly disagreed; 17 (19.32%) respondents disagreed; 23 (26.14%) respondents were not sure; 25 (28.41%) respondents agreed, and 10 (11.36%) respondents strongly agreed.

The second question assessed the effectiveness of the IK in predicting an undesirable season (for instance, one with scarce rains and undesired temperatures) before a farmer can cultivate. The results of the second question are presented in Figure 4-2 below.

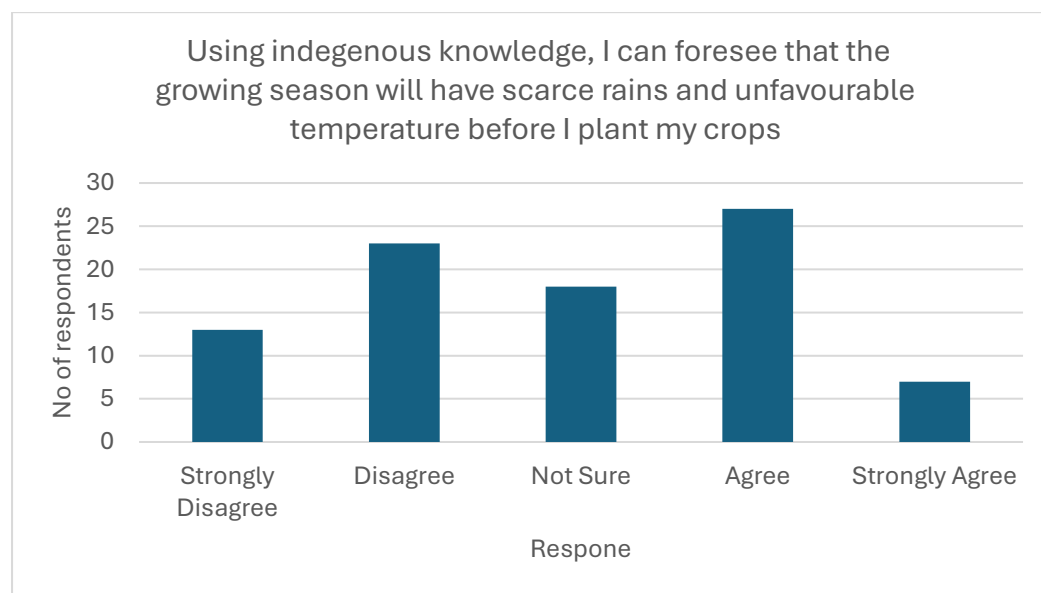


Figure 4- 2: Presentation of respondents' results regarding foreseeing the season with unfavourable weather (Source Author).

From Figure 4-2 above, the prompt to the respondents was: Using indigenous knowledge, I can foresee that the growing season will have scarce rains and unfavourable temperature before I plant my crops. Results indicate that 14.77% of the respondents strongly disagreed; 26.14% disagreed; 20.45% were not sure; 30.68% agreed, and 7.98% of the respondents strongly agreed.

The third question assessed farmers' satisfaction with the use of the IK system alone without any support from the scientific approach in predicting the agricultural season. The results of the third question are presented in Figure 4-3 below.

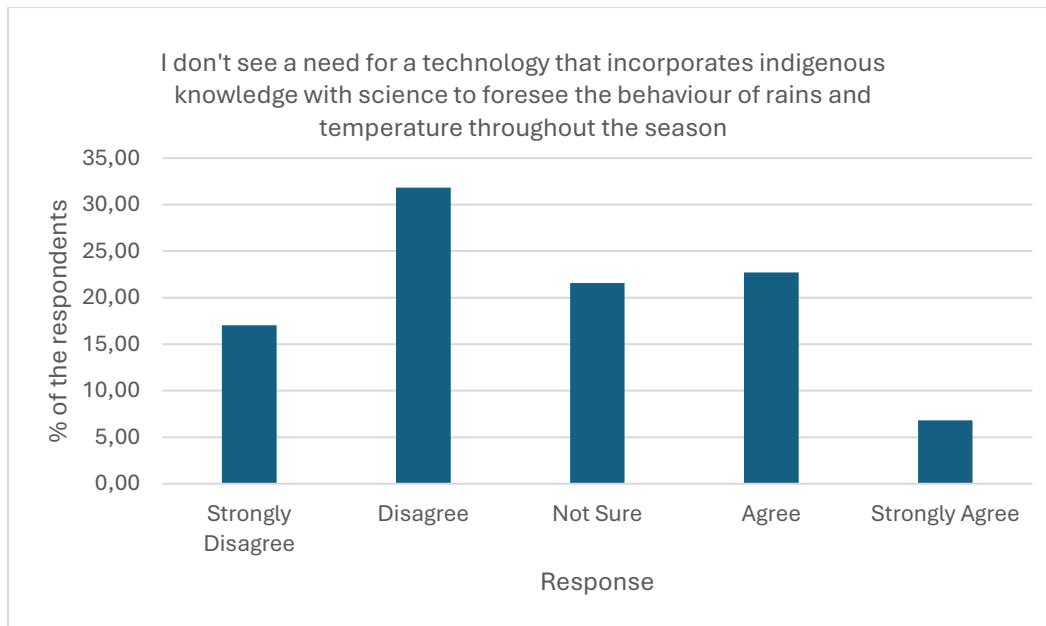


Figure 4- 3: Presentation of respondents' results regarding farmers' satisfaction with IK (Source Author).

From Figure 4-3 above, the prompt to the respondents was: I don't see a need for a technology that incorporates indigenous knowledge with science to foresee the behaviour of rains and temperature throughout the season. On this, 17.05% of the respondents strongly disagreed; 31.82% disagreed; 21.59% were not sure; 22.73% agreed, and 6.82% of the respondents strongly agreed.

Section B2: IK in Crop Health and Development Assessment

In this phase, the effectiveness of the IK in monitoring the health and growth of crops was evaluated. The first question was evaluating the effectiveness of the IK in assessing the health and development of crops. Figure 4-4 below presents the results from the first question.

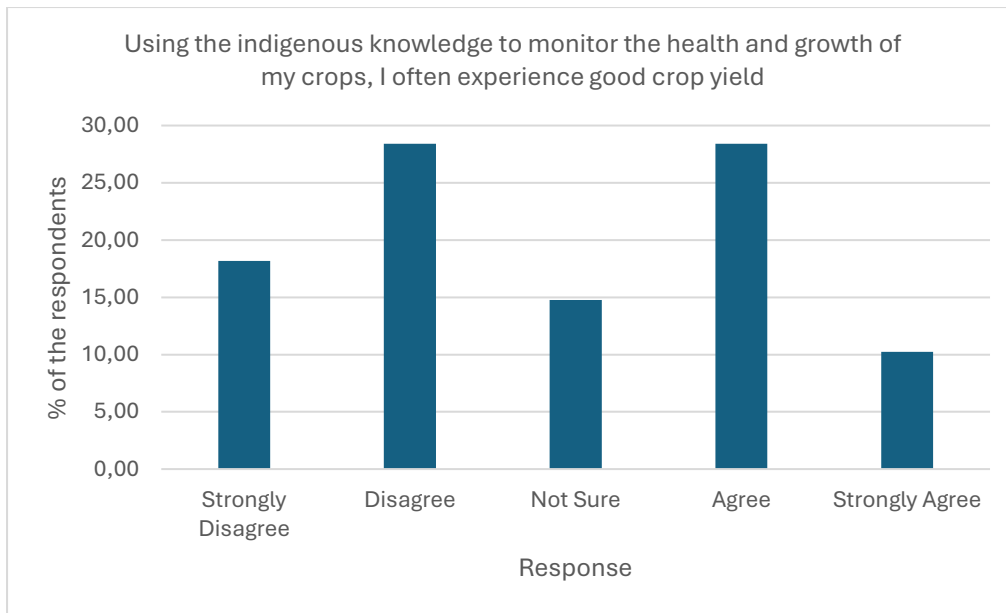


Figure 4- 4: Presentation of respondents' results regarding monitoring of health and growth of crops (Source Author).

From Figure 4-4 above, the prompt to the respondents was: Using the indigenous knowledge to monitor the health and growth of my crops, I often experience good crop yield. In responding, 18.18% of the respondents strongly disagreed; 28.14% disagreed; 14.77% were not sure; 28.41% agreed, and 10.23% of the respondents strongly agreed.

The second question was a follow-up to the first question. It evaluated the effectiveness of the IK in early diagnosing and treating different crop health problems. Figure 4-5 below shows the results from the second question.

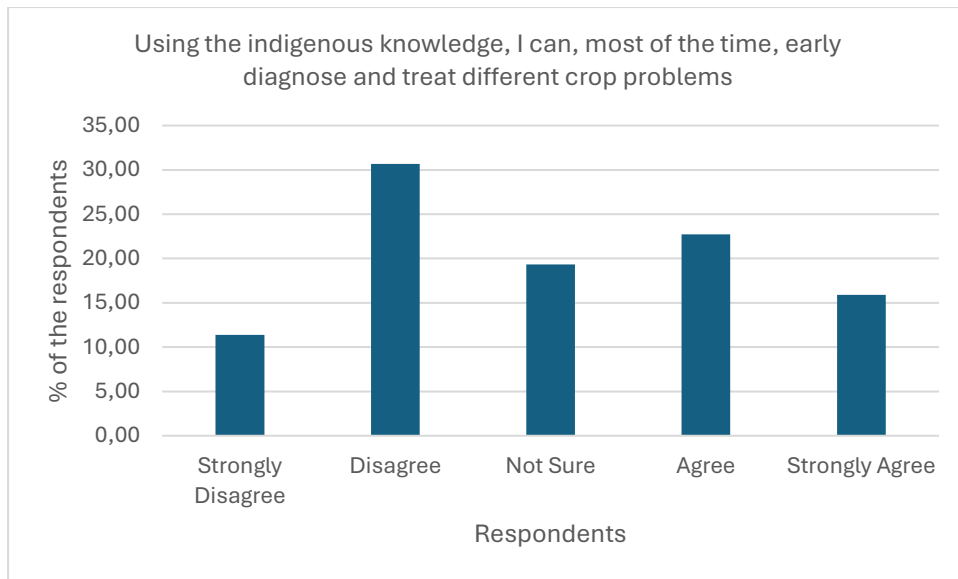


Figure 4- 5 : Presentation of respondents' results regarding diagnosing and treating of crop problems (Source Author).

From Figure 4-5 above, the prompt to the respondents was: Using the indigenous knowledge, I can, most of the time, early diagnose and treat different crop problems. To this prompt, 11.36% of the respondents strongly disagreed; 30.68% disagreed; 19.32% were not sure; 22.73% agreed, and 15.91% of the respondents strongly agreed.

The third question was evaluating farmers' satisfaction with using only the IK to monitor the health and development of their crops. Figure 4-6 below depicts the results obtained from the third question.

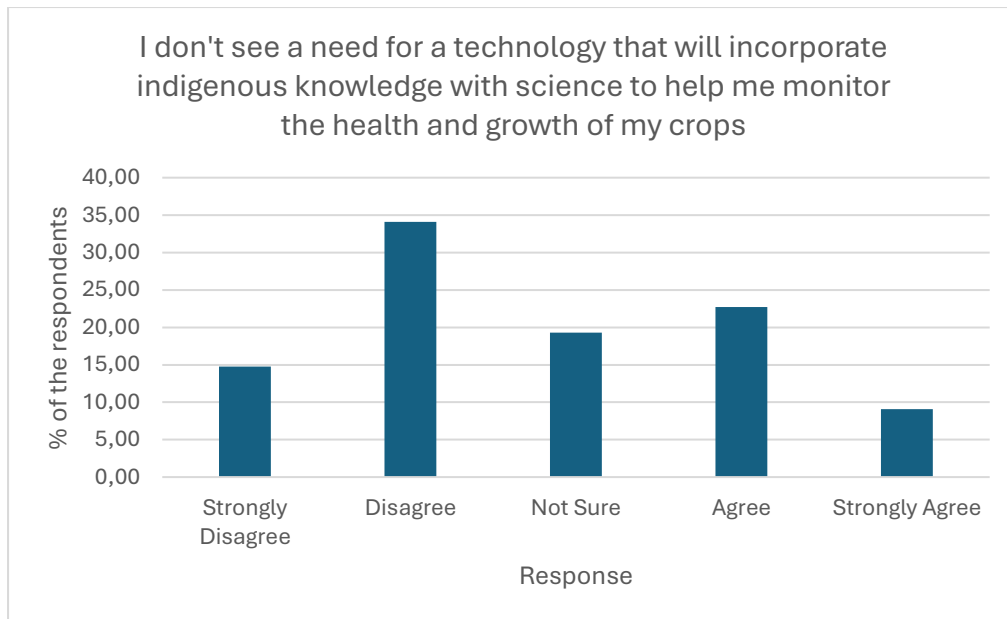


Figure 4- 6: Presentation of respondents' results regarding their satisfaction with IK in monitoring the health of crops (Source Author).

From Figure 4-6 above, the prompt to the respondents was: I don't see a need for a technology that will incorporate indigenous knowledge with science to help me monitor the health and growth of my crops. As the results show, 14.77% of the respondents strongly disagreed; 34.09% disagreed; 19.32% were not sure; 22.73% agreed, and 9.09% of the respondents strongly agreed.

Conclusions from the quantitative data analysis

Despite the value of the IK in predicting season behaviour and monitoring the health and growth of crops, it became evident to the researcher after the analysis of the data that the strength of the IK system continues to deplete. Although many respondents agreed on the effectiveness of the IK in season behaviour prediction and monitoring of crop health, a significant portion of the respondents disagreed while the remaining portion was not certain about the effectiveness of the IK. The depreciation of the IK's effectiveness in support of farmers' cropping decisions became crystal clear when a large number of respondents seemed to support the idea of a technology that will incorporate their local expertise with science towards the agricultural season behaviour prediction and monitoring of health and growth of the crops. This information enabled a researcher to dig deep into understanding how the IK is

used by small-scale farmers in their cropping decisions, and to come up with an approach to complement the knowledge of farmers with science and technology.

4.2.1.3. Qualitative data collection methods and processing

The qualitative data were collected primarily through the focus group discussions followed by the survey to reach out to other farmers who were not part of the focus group discussion.

Focus group discussion

The focus group interview was conducted to get in-depth knowledge on the IK indicators and their observations towards season behaviour predictions, planting periods' estimations, crop selection, the indigenous methods employed by farmers to early diagnose and treat various crop problems and many more. The purposive sampling technique was adopted to select 20 local farmers aged 30 and above, who have been in the study region for over 20 years, primarily do crop farming for subsistence or market purposes, and rely heavily on rain-fed agriculture and IK to support their cropping decisions. The focus group discussion took place outside the Swayimane High School on the 29th of August in 2023. The images presented in Figure 4-7 and Figure 4-8 were captured during the interview.



Figure 4- 7: First image of the participants at Swayimane High School (Source: Author).



Figure 4- 8: Second image of the participants at Swayimane High School (Source: Author).

The images were captured with the consent of the respondents, including the voice recordings that were made that day.

Before the researcher could explore the knowledge of IK indicators in-depth, respondents were asked to provide information on the start of their agricultural seasons, as well as the usual timing of their land preparations and seed sowing. The respondents expounded that owing to season patterns that have been altered because of climate change, they do not have the exact periods to start with land preparations and sowing. One of the respondents said:

Planting is now a risky process because of extreme and unpredictable weather events that can lead to poor germination of the seeds.

However, they indicated that they often start in August with land ploughing and cultivation of warm-season crops conditioned to the availability of rains and fine temperatures. The respondents further expressed that when temperatures start to go down around March, they start considering the cultivation of crops that can withstand cold temperatures. This also depends on the availability of rain.

Respondents further stressed the importance of IK indicators as a pillar that can tell them when to start sowing, what to cultivate, and the behaviour of the agricultural season (iQola). One of the respondents said,

When we start observing strong winds associated with extreme lightning in August, we know that iQola is about to start.

Another one said:

I used to evaluate the behaviour of the moon around August, when it is half, I know that I need to grow crops like green beans and when it is full, I will grow carrots.

Another one said,

There is a big bird that used to visit us, if it sings in a certain manner, we know that heavy rains are coming and if it puts its wings under the water, we know that these rains will last longer.

Another respondent said,

Given that the season changes can be deceiving, I used to evaluate the behaviour of locusts and other insects, in the winter season, they hide themselves deep into the soil, and when the fine weather begins, you will see them coming out.

The IK indicators that were shared during the focus group discussion were documented by the researcher. The researcher further, requested the respondents to share knowledge on how the observation of these indicators impacts their confidence level towards the season behaviour they anticipate. The respondents stressed that even though some indicators are more trusted than others, they are no longer as reliable as they were before due to climate change impacts. However, what brings a glimmer of hope, is when these indicators are observed repeatedly in different locations or timestamps depending on their nature.

Respondents were then requested to share knowledge on how they monitor the health and development of crops including the strategies they use to early diagnose and treat various crop health problems such as crop pests and diseases, and malnutrition among others. Respondents shared various symptoms of crop inflicted by both biotic (pests and diseases) and abiotic factors (heat stress and wind, chemicals and so forth) such as leaf tips wilting, leaf curling, leaf yellowing, leaf margins browning, leaf spots, lower leaves falling, stunted growth, browning of roots, and many more. Some of the farmers have shown a significant amount of knowledge that they gained from the visits made by the scientific personnel. Among others, is when farmers stated that the purpling of crop leaves is because of phosphorus deficiency. They further shared their knowledge of how they map these symptoms to various crop problems and

the organic methods they use to treat these crop problems. Among the methods mentioned are the growing of garlic or mixing it with water to repel insects; using soapy water to fight pests on the soil, crop rotation, companion planting, staggered cropping, and many more. Farmers were further requested to share knowledge on how they map different crops to different soils, properties that they use to evaluate the soil health, how they estimate crops to grow, planting periods and many more based on season behaviour, climate and soil characteristics.

According to farmers, the impacts of climate change continue to demoralise their cropping decisions and the yield. Furthermore, the use of the IK system alone is no longer effective in addressing various crop health challenges that they face. Those who are financially well-heeled, have now introduced the use of chemicals in their farms, and they also take their soils to laboratories for soil tests before they can start planting. Others are continuing to use only the IK methods due to financial constraints while others are totally against the use of chemicals in their farm. One of the respondents accentuated that:

Sometimes when I don't have money to buy chemicals, I grow legume crops such as cowpea to feed my soil with nitrogen.

Some of the farmers who are using chemicals have indicated that they are aware of the danger associated with the use of chemicals to their crops, their soil, their health, and also to the environment. However, there isn't much they can do since it has become a matter of survival. Furthermore, the market is interested in products that look big and attractive. Hence, the use of chemicals can give farmers the produce that is ready for the market. The farmers further claimed to minimise the use of chemicals by mixing them with organic manure such as cow dung, chicken waste and plant residue.

Data collection using the survey

The knowledge gathered was further extended by issuing 50 questionnaires to reach out to farmers who were not part of the focus group discussion. The questionnaire was constructed using more open-ended questions to enable the respondents to share knowledge on how they use IK to support their cropping decisions. The survey was built of questions from the interview guide used during the focus group discussion. However, it was fine-tuned using the knowledge gathered during the focus group discussion. The questionnaire was organised into three parts. The first part was about the demographic information of the respondents, the second part focused on the IK in season behaviour predictions, and the third part focused on the IK in

monitoring the health and growth of the crops. The purposive sampling technique was adopted to select 50 farmers aged 30 and above who primarily do crop farming and rely heavily on rainfed agriculture and the IK system to support their cropping decisions.

The researcher with the help of the youths from Swayimane worked together in disseminating the questionnaires to the participants and also ensuring that they were properly filled. These youths were invited to the focus group discussion to acclimatize themselves to what was discussed. They were appointed on the basis that they have tertiary qualifications and they have been in the study area for over 10 years. These youths were also provided with training to enable them to address any concerns that may occur during the data collection.

Demographic distribution of the respondents

The gender of the respondents was as follows: 30 (60%) respondents were females; 17 (34%) respondents were males, and 3 (6%) declared their gender as other. In terms of age, 14% of the sample fell in the age bracket of 30 to 35 years; 22% were in the age bracket of 36 to 45 years; 52% of the sample were in the age bracket of 46 to 55 years; 12% were in the age bracket of 56 to 65, and none of the respondents aged 66 and above participated in this survey. Further, 10 (20%) of the respondents never went to school; 15 (30%) respondents ended up at the primary school level; 42% ended up at the secondary school level, and 4 (8%) respondents reached tertiary education. As high as 92% of the respondents indicated to have been in the study area for over 10 years, while the rest had been in the study area for 10 years or less. Lastly, 28 respondents claimed to have smartphones, 16 to have non-smartphones and 6 not to have phones of any kind.

Some qualitative data collected using the survey

Among the IK indicators respondents stressed in the questionnaire are:

- The flowering of peach trees, pear trees, Hantela trees and blooming roses is a good indication of the end of the cold temperatures and the signal to start with land ploughing.
- The calling of birds named Phezukomhkono and Insingiza indicate abundant rains to come.
- The first rains in August signal the onset of iQola.
- The increase in insect and birds species signal the beginning of the warm season.
- The sighting of worms in the soil indicates the soil fertility.

- The croaking of frogs signals abundant rains to come.
- Sighting of the half-moon facing east and the visible phases of the moon signal scarcity of rain.
- The formation of dark clouds signals heavy rains coming.
- The movement of stars from west to east at night signals that the rains coming soon.
- The appearance of red and too many black ants is a sign of good rains coming.
- Long days and short nights indicate the beginning of the warm season.

The respondents further stressed that crops such as corn, amadumbe, butternut, pumpkin, potatoes, tomato, kidney beans, black beans, pinto beans, peppers and others are cultivated during the warm season and vegetables such as spinach, cabbages, peas, radish, kale and others are cultivated during the cold season. Other crops such as lettuce, radish, carrots, cabbage and beetroot can be grown in all seasons. Some of these including cabbage are preferred to be cultivated during the cold season as they are highly susceptible to various crop insects including *iBhungezi* (beetle) when cultivated during the warm season.

The *isiKhuthaza* is the type of homemade fertiliser mentioned in various settings that farmers use to treat various unhealthy crops including tomatoes that have dark spots on the fruits and the leaves. The respondents mentioned that the health of vegetable crops is improved by mixing the soil with ashes and *Shiboshi* (a type of medication used for health and cleansing purposes). The combination of *Unguba* (cow dung) and lime, not only neutralises the acidic level in the soil, but also improves the soil moisture, water retention capacity of the soil, and organic fertilisers. The respondents further emphasised the mix of various household items (such as lemon juice, tobacco dust, vinegar, and dish soap) with various vegetables such as chills and garlic, and animal and chicken manure, to improve the soil fertility, minimise weeds in the soil and, more importantly, to combat various sorts of crop insects.

The knowledge gathered from the focus group discussion and questionnaires helped the researcher in thinking of possible ways to systematically integrate the knowledge of farmers and science to improve their cropping decisions. The next sections dwell on how the knowledge gathered was processed (quantified) and integrated with the scientific data and scientific models to help farmers predict season behaviour and monitor the health and growth of the crops.

4.2.2. Systematic processing of the IK indicators and their observations

The content analysis approach was used to analyse the information acquired through the questionnaire, the researcher's notes and audio recordings that were made during the focus group discussion. Content analysis is a well-known technique used to organise, interpret and make valid inferences from research data. According to Krippendorff (2018), it involves the systematic coding and assessment of qualitative data, and determining the existence and frequency of certain words, and phrases, among others in the data, which are then categorised and organised into themes. This approach enabled the researcher to better understand the IK indicators and their observations by reviewing and analysing the researcher notes, the data collected from the questionnaire, and replaying the audio recordings. The logic was to find a possible solution to systematically register the IK indicators and their observations. The registration of indicators' observation became very complex as the researcher had to think of ways to systematically distinguish between single observation and multiple observations of the IK indicators considering their nature which can either be meteorological, astronomical or environmental. This was motivated by what farmers accentuated during the focus group discussion that the IK indicators are no longer reliable; hence, multiple observations of them bring a glimmer of hope towards their confidence level of an anticipated season behaviour.

An indicator can be systematically described as an object or entity made up of various sets of attributes to explain it. These attributes are:

- **ID:** To uniquely identify it from others.
- **Description:** To explain the indicator.
- **Season type:** Some indicators can be observed either during the warm season or cold season while others can be observed in any season.
- **Prediction Type:** Indicators can be used to forecast normal rainfall, floods, and droughts, as well as to predict seasonal shifts between warm and cold periods.
- **Nature type:** Indicators can be categorised into either astronomical, meteorological or environmental.
- **Weight:** This field measures the amount of impact the indicator has towards the season behaviour it predicts. It reflects what was discussed during the focus group meeting, that some indicators are considered more reliable than others.

The IK indicator object is further annotated with other properties to distinguish between single and multiple observations of the indicator, preventing duplicate entries during their registration. To systematically register the IK indicator's observation, the location and timestamp of the indicator's observation need to be considered. This may not be a problem for astronomical and meteorological indicators given that most of the time, their occurrence can be witnessed by everyone in the study area such as *very hot or cold temperatures during the dry season or visible or dark phases of the moon at the beginning of the rainy season*. In the case of astronomical and meteorological indicators, the farmer's certainty level will increase as they start to observe them on different days or timestamps.

Regarding the environmental indicators, more knowledge needs to be gained because it is less likely that they can be witnessed by everyone in the region. Furthermore, the actors making up these indicators are living environmental organisms such as birds, insects, animals and plants with different natures and characteristics. An environmental indicator can be an organism itself or an action performed by an organism. If the indicator is an organism, then it becomes important to understand the nature of an organism, whether it is fixed or it can move from one place to the other. If it is fixed such as plants, then it will be pointless to evaluate the timestamp distance between the two observations of an organism given that it is likely that where it is today it will still be in the next day or timestamp. Therefore, farmers' certainty level will only increase if the same type of organism is observed at different locations. If they are non-fixed organisms (that can move from one place to another) such as birds and insects, then it means farmers' certainty level (CL) can increase if the next observation of the same type of organism is made in the same location at different timestamps (days), or if it is made at different locations within a given timestamp (day). If an indicator is an action performed by an organism, such as *a black-brown bird named umGugwane starts singing during the dawn or cattle jumping up and down with their tails held high at the beginning of the rainy season*, it is not guaranteed that in the next timestamp, the occurrence of the indicator will take place. Hence, farmers' CL will increase if the same observation repeats itself at the next timestamp either at the same or different locations.

The two additional fields were created to try to capture the role of environmental indicators as shown below:

- **organismOractionByOrganism:** This attribute tries to understand if the environmental indicator is an organism or an action performed by an organism.

- **fixedOrNotfixed:** If the indicator is an organism, this attribute will capture the nature of the indicator whether it can move from one place to the other or is stationary.

The relationship between the indicators and the observers is presented below in Figure 4-9 using an entity relational diagram (ERD).

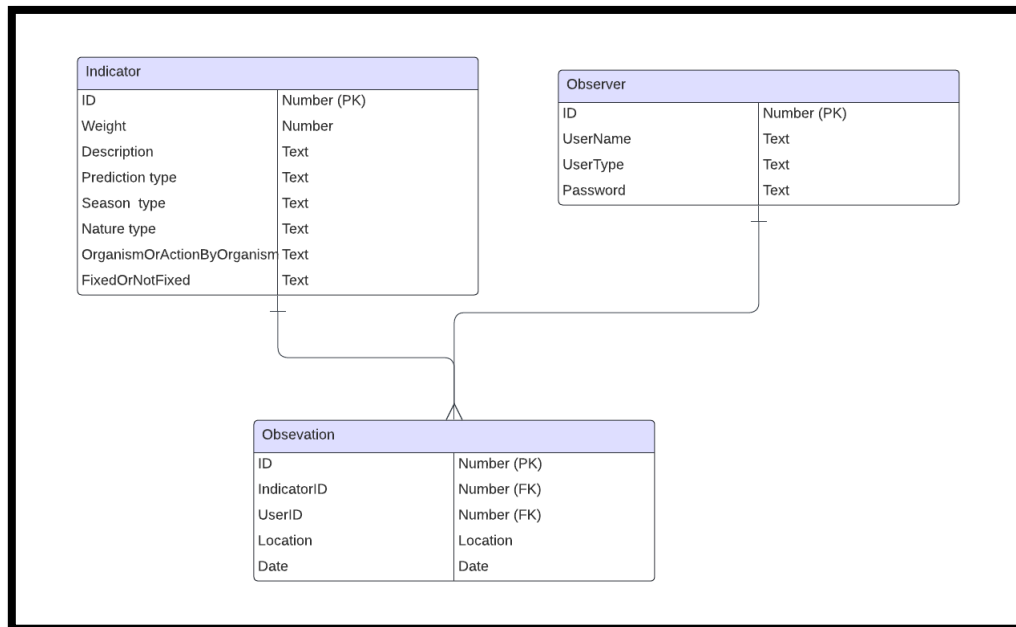


Figure 4- 9: ERD showing the relationship between indicators and observers (Source: Author).

From the diagram in Figure 4-9 above, a farmer (observer) can observe one or many indicators, and one indicator can be observed once or multiple times by one or many farmers. The observation entity was created as a bridge to handle the ambiguity of the many-to-many relationship between the observer and the indicator. This entity is comprised of the observer, the indicator observed, the location where the observation was made, and the date when the observation was made.

An algorithm presented in Figure 4-10 below was developed to evaluate every indicator's observation before it can be registered into the database.

```

boolean ValidateAndInsertObservation(Observation observation)
Declarations
num timeStampDistance
Boolean firstTime = false
num locationDistance
conn = new DbsConnection()
List<Observation> obsList
if observation.getLocation().toLowerCase().contains(regionOfInterest) then
    flag = true
    obsList = conn.readObservationsForIndicator(observation.getIndicator.getIndicatorId())
    if obsList.size() < 1 then
        conn.createObservation(observation)
        firstTime = true
    else
        if NOT(observation.indicator.getNature().equalsIgnoreCase("environmental")) then
            for num x = 0, x < obsList.size(), x++
                timeStampDistance = getTimeStampDistance(obsList.get(x).getDateCreated(), observation.getDateCreated())
                if timeStampDistance = 0 then
                    flag = false
                endif
            endfor
        else
            for num x = 0, x < obsList.size(), x++
                if observation.indicator.getActionOrOrganism() = "action" then
                    timeStampDistance = getTimeStampDistance(obsList.get(x).getDateCreated(), observation.getDateCreated())
                    locationDistance = getLocDistance(obsList.get(x).getLatitude(),
                                                    obsList.get(x).getLongitude(), observation.getLatitude(), observation.getLongitude())
                    if timeStampDistance = timeThreshold && locationDistance < locationThreshold then
                        flag = false
                    endif
                else if observation.indicator.getNatureByOrganism() = "non-fixed" then
                    timeStampDistance = getTimeStampDistance(obsList.get(x).getDateCreated(), observation.getDateCreated())
                    locationDistance = getLocDistance(obsList.get(x).getLatitude(), obsList.get(x).getLongitude(),
                                                    observation.getLatitude(), observation.getLongitude())
                    if timeStampDistance = timeThreshold && locationDistance < locationThreshold then
                        flag = false
                    endif
                else
                    locationDistance = getLocDistance(obsList.get(x).getLatitude(), obsList.get(x).getLongitude(),
                                                    observation.getLatitude(), observation.getLongitude())
                    if locationDistance < locationThreshold then
                        flag = false
                    endif
                endif
            endfor
        endif
    endif
    if flag = true && firstTime = false then
        conn.createObservation(observation)
    endif
return flag

```

Figure 4- 10: An algorithm that captures the registration of the IK indicators' observation (Source: Author).

The algorithm in Figure 4-10 above, receives the proposed observation of an indicator made by the farmer (observer) as a parameter. The location of the observation is evaluated to ensure that it is registered within the region of interest (in this context will be Swayimane). An indicator's observation will be disqualified if it is not made within the specified area of interest, otherwise (if the observation is made within the required area) all the observations made for

that indicator will be extracted from the cloud database. If the indicator is observed for the first time, then it will be accepted and registered into the database. Otherwise, if the indicator was observed before (meaning there are one or more observations of the indicator that exist in the database), the nature of the indicator will be evaluated. If the indicator observed is an environmental indicator, the algorithm will first check if the indicator represents an organism or an action performed by an organism. If the indicator is an organism, then based on the nature of the organism (whether it is fixed or non-fixed), the algorithm will loop through the observations and compute either the location distance or both the location and timestep distances between the proposed observation and the observations already made of that particular indicator. If the indicator is an action performed by an organism, then, the algorithm will calculate both the location and timestamp distances between the proposed observation and others already made. For astronomical and meteorological indicator observations, the algorithm will only consider the timestamp distance as they can be observed by anyone in the region.

An observation will be accepted and recorded in the database if it is not a duplicate, following the assessment of the temporal and spatial distance parameters discussed.

The Haversine equation presented below is utilised by the algorithm to compute the shortest location distance between two points on an Earth's surface using their latitudes and longitudes:

$$d = 2r \sin^{-1} \left(\sqrt{\sin^2 \left(\frac{x_2 - x_1}{2} \right) + \cos(x_1) \cos(x_2) \sin^2 \left(\frac{y_2 - y_1}{2} \right)} \right) \quad (\text{Equation 4-1})$$

where

r represents the radius

x_1 represents the latitude of point A

x_2 represents the latitude of point B

y_1 represents the longitude of point A

y_2 represents the longitude of point B

The timestamp distance is computed as a difference of days between two dates. The location and timestamp distance thresholds will be adjusted by the administrator of the proposed tool based on the farmers' requirements. These thresholds will be used to determine the qualifying and non-qualifying indicators' observations.

The algorithm stressed above in Figure 4-10, focused on minimising the ambiguity in the registration of the indicators' observations. The next concern was how the farmers' certainty level would be processed from the observations registered given that indicators have different weights and can predict different rainy season behaviours. The certainty level also has to be evaluated against the threshold that will be adjusted based on farmers' requirements. If the certainty level is high (meaning above the threshold), the IK will be considered successful and will be integrated with the scientific knowledge. Otherwise, it will be disregarded.

4.2.2.1. Certainty Level Computation

Several algorithms were developed to manipulate the IK indicators' observations in the database to compute the farmers' certainty level.

The first algorithm presented below in Figure 4-11 receives a list of the observed IK indicators from the database. It will run through the list and return distinct observed indicators.

```

Collection getDistinctIndicators(Collection observations)
  Declarations
    Collection distinctIndicators = new Collection()
    Collection indicatorIds = new Collection()
    Text id
  for num x = 0, x < observations.size(), x++
    id = observations.get(x).getIndicator().getObjectId()
    if (NOT)indicatorIds.contains(id)
      indicatorIds.add(id)
      distinctIndicators.add(observations.get(x).getIndicator());
    endif
  endfor
  return distinctIndicators

```

Figure 4- 11: An algorithm that extracts distinct observed indicators (Source: Author).

The next algorithms answer the question of how many times each indicator was observed and by how many observers.

The first one presented below in Figure 4-12 receives as input a list of the observed IK indicators from the database and a list of distinct indicators extracted from the same list of the observed IK indicators. The function will iterate through the lists, perform arithmetic operations, and return a frequency list showing the number of times each indicator was observed.


```

Collection getFrequencyofDistinctIndicators(Collection distinctIndicators, Collection
observations)
Declarations
Text obsIndicatorId
Collection frequencyList = new Collection()
num count = 0
for num x = 0, x < distinctIndicators.size(), x++
Indicator indicator = distinctIndicators.get(x)
for num y = 0, y < observations.size(), y++
obsIndicatorId = observations.get(y).getIndicator().getObjectId()
if indicator.getObjectId() = obsIndicatorId then
count += 1
endif
endfor
frequencyList.add(count)
count = 0
endfor
return frequencyList

```

Figure 4- 12: An algorithm that computes the frequency of distinct observed indicators (Source: Author).

The second algorithm in Figure 4-13 below receives a list of IK indicators' observations and a distinct indicators' list. The algorithm will loop through the lists, analyse the data, and return the frequency list showing the number of observers for each indicator.

```

Collection getFrequencyOfObservers(Collection distinctIndicators, Collection observations)
Declarations
Text observationIndicatorId, observationUserId
Collection userListDistinct = new Collection(), userCountList = new Collection()
num countUser = 0
foreach Indicator indicator in distinctIndicators
foreach Observation observation in observations
observationIndicatorId = observation.getIndicator().getObjectId()
if indicator.getObjectId() = observationIndicatorId then
observationUserId = observation.getUser().getObjectId()
if NOT userListDistinct.contains(observationUserId) then
userListDistinct.add(observationUserId)
endif
endif
endfor
countUser = userCountList.size()
userListDistinct = new Collection()
userCountList.add(countUser)
endfor
return userCountList

```

Figure 4- 13: An algorithm that computes the number of observers for every observed indicator (Source: Author).

The next algorithm in Figure 4-14 determines the qualifying indicators by validating the frequency of their observations and observers using the thresholds that will be adjusted by the

administrator of the system based on farmers' requirements. This algorithm accepts a list of distinct indicators and a frequency list of observations and observers for each indicator. The algorithm will make validations and return the observed indicators that qualify. An indicator qualifies if its number of observations and observers is above a given threshold.

```

Collection getQualifyingIndicators(Collection distinctIndicators, Collection frequencyOfIndicators,
                                   Collection frequencyOfUsers)

Declarations
    num countIndicator, countObservers
    Collection qualifyingIndicators = new Collection()
    num count = 0
while count < distinctIndicators.size()
    countIndicator = frequencyOfIndicators.get(count)
    countObservers = frequencyOfUsers.get(count)
    if countIndicator >= thresholdForObservations AND countObservers >= thresholdForObservers then
        qualifyingIndicators.add(distinctIndicators.get(count))
    endif
    count += 1
endwhile
return qualifyingIndicators

```

Figure 4- 14: An algorithm that validates the qualifying indicators based on the given thresholds (Source: Author).

Considering that farmers can observe different indicators that predict different season behaviours (rainfall behaviour (floods, drought, abundant rains) and season change from either warm to cold or cold to warm), the next step will be to determine the rainy season behaviour that is highly voted by the qualifying observed indicators. This is achieved by evaluating the weights of the qualifying indicators based on the season behaviour type they predict. The predicted season behaviour type with high weight will be considered. The algorithm in Figure 4-15 below sums the weights of qualifying indicators based on the season behaviour type they predict.

```

num getWeight(Collection qualifyingIndicators, Text seasonBehaviourType)
    num weightSum = 0
    foreach Indicator indicator in qualifyingIndicators
        if indicator.getType = seasonBehaviourType then
            weightSum += indicator.getWeight()
        endif
    endfor
    return weightSum

```

Figure 4- 15: An algorithm that sums the weights based on season behaviour type (Source: Author).

To find the season behaviour type with the maximum sum of weights, the algorithm presented below in Figure 4-16 receives an array of the summed weights. It will make validations and return the maximum weight.

```

num findMax(List weights)
  Declarations
    num max = weights.get(0)
    for num count = 0, count < weights.size(), count++
      if (max < weights.get(count)) then
        max = weights.get(count)
      end
    endfor
  return max

```

Figure 4- 16: An algorithm that identifies the rainy season behaviour type with max value (Source: Author).

The season behaviour type with the maximum weight will be identified as a winner. So, the farmers' CL of an oncoming rainy season behaviour will be computed based on the selected season behaviour type. The CL is calculated as a percentage obtained after dividing the sum of the weights of observed indicators by the sum of weights of all indicators that predict the same season behaviour as shown below:

$$CL = \left(\frac{1}{\sum_{i=1}^n F_i} * \sum_{b=1}^k L_b \right) * 100 \quad (\text{Equation 4-2})$$

where L_b represents the sum of observed indicators' weights and F_i represents the sum of all indicators' weights that predict the same season behaviour.

4.2.3. Secondary data collection, processing and analysis

This section entails the processes followed to collect, process, and analyse the historic satellite imagery and weather data. The weekly Sentinel 2 satellite images of the area of interest (Swayimane region) ranging from 2013 to 2023 were collected primarily from the following web sources: Earth Explorer - United States Geographic System (USGS) and Scihub Copernicus. The daily rain and maximum temperature data ranging from 1993 to 2023 were extracted from the district weather station that is close to the region of interest.

To process satellite imagery data, pixel intensity values of healthy vegetation, waterbodies, and soil moisture covers were extracted from each satellite image and were documented for further processing. The time series analysis and forecasts were made on weekly historic satellite imagery and daily weather data. The daily weather data were aggregated into weeks to enable

weekly forecasts. The SARIMA, Holt's Winter, and LSTM time series models were evaluated on the data and best performing models were used to forecast the next 48 weeks' time series data.

4.2.3.1. Satellite Imagery Data Collection

Over 450 weekly Sentinel 2 satellite images ranging from 2013 to 2023 were collected primarily from Earth Explorer-USGS and SciHub Copernicus websites. The multispectral bands of satellite images ranged from band-1 (Coastal aerosol) to band-12 (SWIR). The temporal resolution of the satellite images was roughly 7 days with a spatial resolution of 10 meters in the visible and NIR multi-spectral bands.

4.2.3.2. Satellite Imagery Data Processing

Quantum Geographic Information System (QGIS) was used to process the satellite images. The Semi-Automatic Classification Plugin (SCP) and Basemap plugin were installed into QGIS to aid in the pre-processing of the satellite images.

Pre-processing steps of the satellite images

For every image, only a set of required multispectral bands were loaded into the QGIS. Due to the high synoptic view and coarse spatial resolution of the satellite images, the Basemap (HCMGIS) plugin and SCP tools were installed. The Basemap plugin was used to address the geometric deficiency by mapping the thematic (satellite) image's coordinates with the Earth's coordinates presented on the map. The thematic image was overlaid on the Earth's map to locate the region of interest. The raster clipping was performed to extract the subset of pixels that represent Swayimane region as shown below in Figure 4-17.



Figure 4- 17: Extracted pixels that represent the region of interest (Source: Author).

Atmospheric correction

This pre-processing step is about calibrating bands from digital numbers to surface reflectance values where atmospheric distortions are removed or minimised for quality interpretation and quantity analysis of the data.

The satellite images' clipped bands were loaded from their directory into QGIS. The metadata file that comes with each satellite image containing all the necessary information regarding the computation of atmospheric parameters was also loaded for each corresponding set of clipped bands. The atmospheric correction was then applied to the clipped bands.

Figure 4-18 below is the satellite image representing the surface reflectance of the Swayimane region after applying the atmospheric correction.

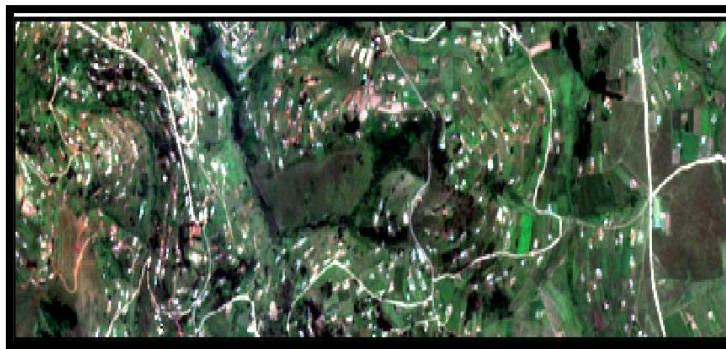


Figure 4- 18: A surface reflection of Swayimane region after reducing the atmospheric effects (Source: Author).

The raster calculator was used to compute the healthy vegetation cover represented by NDVI values of 0.3 and above using Red and NIR bands.

From the image presented below, the NDVI reflectance values that are greater or equal to 0.3 are assigned to a green class and those that are below the threshold are assigned to a white class. This is presented in the image depicted in Figure 4-19 below.

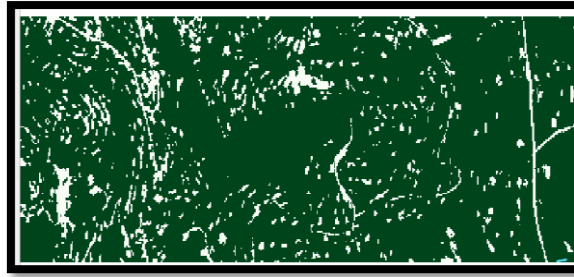


Figure 4-19: Depiction of healthy vegetation presented in green colour and non-vegetation presented in white colour (Source: Author).

The green class's pixel intensity cover (representing vegetation cover) was computed and recorded together with the acquisition date of each satellite image. The soil moisture and waterbody pixel covers were also computed following the same approach, and their pixel intensity values were recorded for every image. The McFeeters NDWI was used to estimate the water surface cover for every image using green and NIR bands. The reflectance values greater than 0, represent water surfaces while those that are zero or below are considered as non-waterbodies. NDMI was used to calculate the soil moisture cover using NIR and SWIR spectral bands. Similar to McFeeters' NDWI, values above 0 are interpreted as indicating the presence of soil moisture, while values of 0 or below suggest its absence. It is also worth noting that NDWI values greater than 0, not only reflect moisture in the soil but also in the vegetation.

Because land cover areas can be presented by a large number (thousands) of pixel intensity values, the normalisation or scaling down of them was performed to simplify the analysis and further processing of them. The satellite imagery data were stored in the computer's directory for further processing.

Cloud masking and mosaic

Some satellite images collected during the rainy season were distorted by high cloudy pixels, making it difficult for interpretation and analysis of land cover features. The cloud masking and mosaic were performed on those images. These processes helped the researcher in removing cloudy pixels in one image and replacing them with non-cloudy pixels from a neighbouring image. This was achieved using the SCP tools and cloud classification vector file that comes with every image.

4.2.3.3. Weather data collection and processing

The daily rain and maximum temperature data ranging from 1993 to 2023 were extracted from the district weather station in Pietermaritzburg which is close to the region of interest. The rain data were in millimeters (mm) and the temperature data were in degrees Celsius. The data had a very low rate of missing entries which can become detrimental to most of the time series models if they are ignored. To fill in those missing entries, the imputation algorithms were deployed.

Weather Data Pre-Processing

Before generating forecasts, the data imputation techniques were rolled out to address data columns with missing data. This was important because most of the time series forecasting models, need to work with the data that is free from the null entries. The time series data were initially deseasonalised because imputation methods used in this thesis do not account for the seasonality variations in the data. The data were deseasonalised by first estimating the average seasonal indices in the data, and removing them from every observation.

Where there are missing entries, the subset of preceding and succeeding data points was extracted and used to train various imputation models. These models learn by generating the rules that can aid in predicting unknown data entries. The models that were trained are Newton Forward Difference (NFD), Newton Backward Difference (NBD), Newton Divided Difference Polynomial (NDDP) model and Polynomial regression models.

The NFD is mathematically explained as:

$$y = y_0 + \frac{u}{1!} \Delta y_0 + \frac{u(u-1)}{2!} \Delta^2 y_0 + \frac{u(u-1)(u-2)}{3!} \Delta^3 y_0 + \dots + \frac{u(u-1)(u-2)\dots(u-n+1)}{n!} \Delta^n y_0 \quad (\text{Equation 4-3})$$

where u is expressed as : $u = \frac{x - x_0}{h = x_1 - x_0 = x_2 - x_1 = x_3 - x_2 \dots x_n - x_{n-1}}$ (Equation 4-4)

y_0 is the first value from the dependent variable; x_0 is the first value from the independent variable (time), and h is the time interval between observations.

When using the NFD model, the coefficients : $\Delta y_0, \Delta^1 y_0, \Delta^2 y_0 \dots \Delta^n y_0$ have to be estimated using the difference table.

The NBD model is computed as follows:

$$y = y_n + \frac{u}{1!} \Delta y_{n-1} + \frac{u(u-1)}{2!} \Delta^2 y_{n-2} + \frac{u(u-1)(u-2)}{3!} \Delta^3 y_{n-3} + \frac{u(u-1)(u-2)(u-3)}{4!} \Delta^4 y_{n-4} + \dots + \frac{u(u-1)(u-2)\dots(u-n+1)}{n!} \Delta^i y_{n-i} \quad (\text{Equation 4-5})$$

where u is calculated as : $u = \frac{x-x_n}{h=x_1-x_0=x_2-x_1=x_3-x_2\dots\dots\dots x_n-x_{n-1}}$ (Equation 4-6)

i is the number of differences that will be made.

When using the NBD model, the coefficients $\Delta y_{n-1}, \Delta^2 y_{n-2}, \Delta^3 y_{n-3} \dots \Delta^i y_{n-i}$ also have to be estimated using the difference table. The NBD is the opposite of NFD where the interpolation process takes place through the backward instead of forward differencing.

The Newton Divided Difference Polynomial model is mathematically explained as:

$$f_{n-1}(x) = b_1 + b_2(x - x_1) + \dots + b_n(x - x_1)(x - x_2)(x - x_{n-1}) \quad (\text{Equation 4-7})$$

where $b_1, b_2, \dots, b_n$ are the coefficients estimated through both the division and forward differencing approaches made on both dependent and independent variables.

Unlike NFD and NBD models that are restricted to only observations captured on an equal time interval, the NDDP model can work on both equidistance and non-equidistance observations.

The polynomial regression is mathematically explained as:

$$y = b_1 + b_2x + b_3x^2 + \dots + b_mx^m \quad (\text{Equation 4-8})$$

where $b_1, b_2, b_3 \dots b_n$ are the coefficients that are commonly estimated using polynomial matrix sequences.

The number of polynomial coefficients reflects the type of model employed, depending on the pattern observed in the data. For instance, the first-order polynomial degree model (Linear Regression) uses only two coefficients (that represent the gradient and the slope) to fit a straight line to the data. Hence, it is suitable for the data that signify a linear dependence pattern. The second-order, third-order, and so forth polynomial degree models (presented by more than two coefficients) can work well on the data points that signify non-linear dependence. This research was restricted to only first-, second-, and third-order polynomial regression models.

The models presented above were trained using different subsets of the data that house missing data entries. They were further evaluated using Root Mean Squared Error (RMSE) presented below, to identify ones that can well understand the pattern in the data.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - p_i)^2}{n}} \quad (\text{Equation 4-9})$$

From the equation above, n is the number of observations, y_i is the actual value and p_i is the predicted value.

The best-performing model for the given subset of the data was used to estimate the missing entries. After filling the gaps in the data, the data were transformed back to its original form by adding the seasonality component that was earlier removed.

4.2.3.4. Time Series Analysis

The time series data (rain, maximum temperature, waterbodies, vegetation greenness, and soil moisture cover values) were loaded from the computer's directory using Anaconda Integrated Development Environment (IDE). Data were plotted to observe the following time series components: trend, seasonal, and irregular components. The trend is a long-term movement of the time series. The trend can either increase or decrease over time unless the series is stationary. The seasonal component is known as the regular periodic rhythmic movement of the data over time. It can also be explained as intra-year stable fluctuations repeatable over the entire length of the time series (Atzberger and Eilers, 2011). Irregular component is a white noise or residuals left after the trend and seasonal components are removed from the series. The irregular component is presented by fluctuations that are purely random, erratic, unforeseen, and unpredictable.

Below is the plotted time series data of rainfall, temperature, waterbody, soil moisture and vegetation cover.

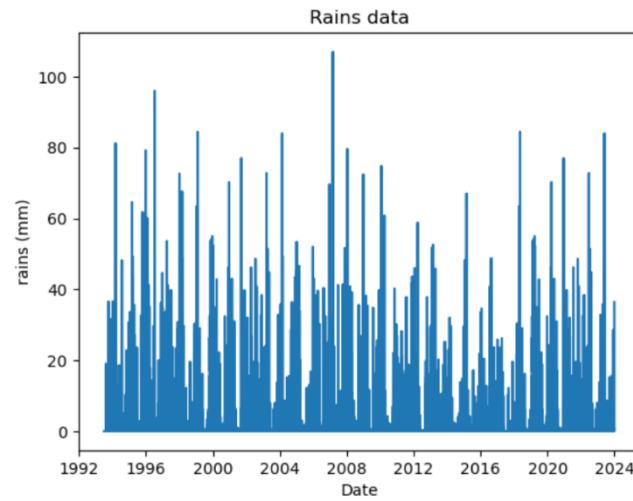


Figure 4- 20: Depiction of rainfall time series data (Source: Author).

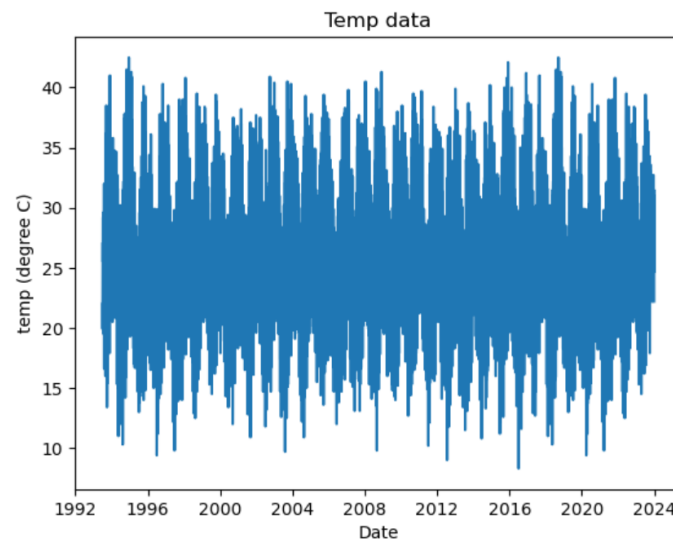


Figure 4- 21: Depiction of temperature time series data (Source: Author).

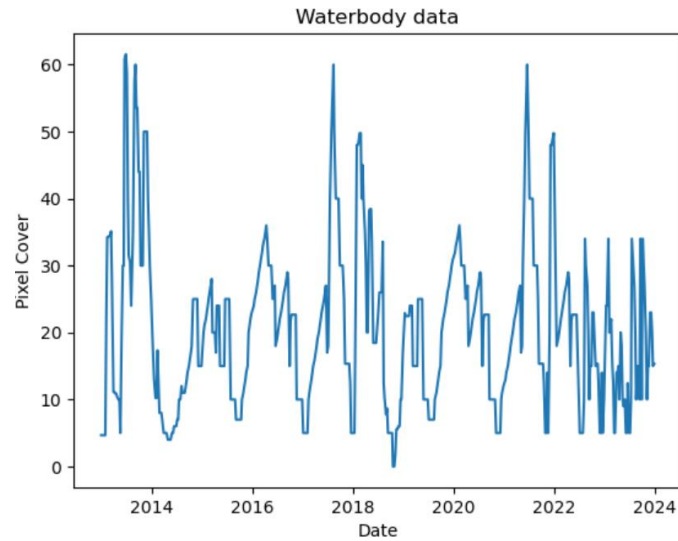


Figure 4- 22: Depiction of waterbodies time series data (Source: Author).

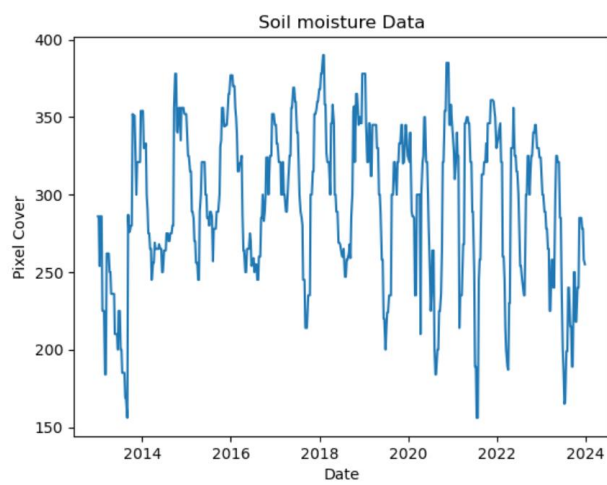


Figure 4- 23: Depiction of soil moisture time series data (Source: Author).

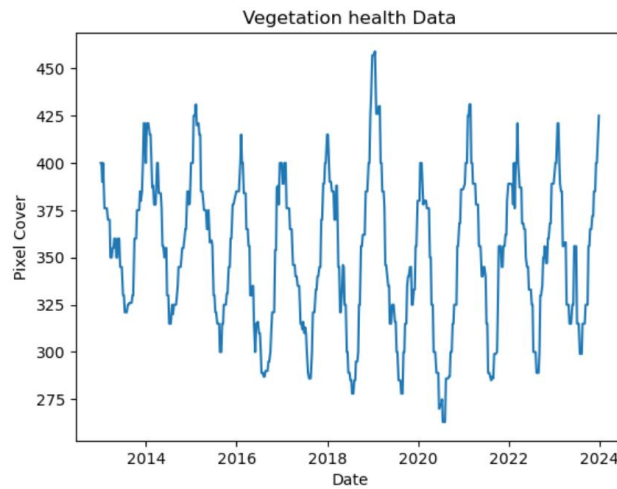


Figure 4- 24: Depiction of vegetation cover time series data (Source: Author).

Figure 4-25 presented below is an example of the rainfall time series data that have been decoupled to observe the trend, seasonality and white noise components.

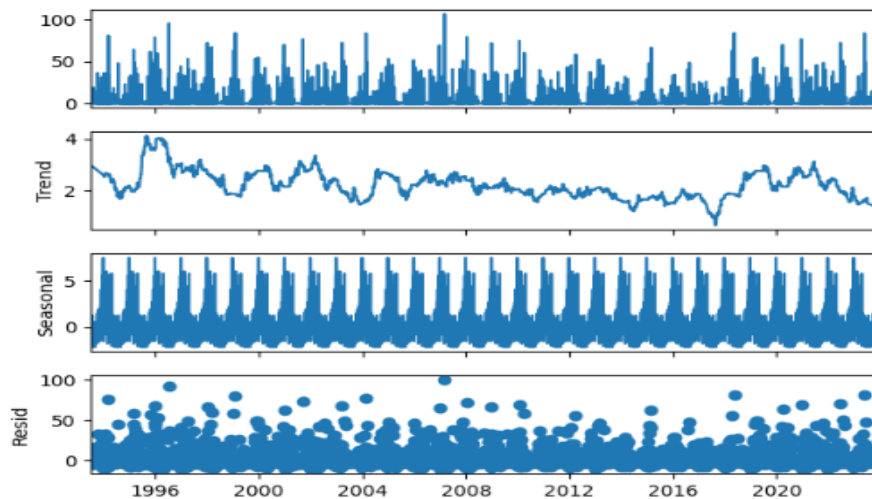


Figure 4-25: Depiction of rainfall time series data with decoupled time series components (Source: Author).

4.2.3.5. Time Series Forecasting

Three well-known time series models, namely Holt-Winter's model, SARIMA, and LSTM were evaluated. The selection of these models was motivated by their capacity to handle the time series with trend and seasonality effects. Before the models were used, the time series data were split into two parts (training and test sets) as indicated in Table 4-1 below.

Table 4- 1: Illustration of how time series data were split into training and test set (Source: Author).

Data	Training	Testing
Weather data	1993-2017	2018-2023
Satellite imagery data	2013-2017	2018-2023

The researcher first provides some background knowledge on every time series model followed by the approach undertaken to deploy each model to generate forecasts.

Background information on the Holt-Winter Model

This model, also known as the TES model, is an extension of DES and SES explained in Chapter 2. It was given the name TES due to its capacity to handle the series with level, trend, and seasonality components. Before performing the forecast using this model, the behaviour of the trend and seasonality must first be identified. Both trend and seasonality can either be additive or multiplicative. An additive trend is observed when the rate of change in data is constant over time and can be represented by a linear graph. The multiplicative trend is observed when the rate of change in data is either increasing or decreasing exponentially over time and it can be represented by fitting a parabola (non-linear graph). The formula below denotes how the current additive trend (T_t) is calculated by adding the difference obtained from the current (L_t) and previous (L_{t-1}) levels, with the previous trend (T_{t-1}) controlled by the trend's learning rate (β). The levels (L_t and L_{t-1}) are the smoothed values (forecasts) obtained using the SES model:

$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1} \quad (\text{Equation 4-10})$$

For a multiplicative trend, the level difference is multiplied by the previous trend as illustrated below:

$$T_t = \beta(L_t - L_{t-1}) * (1 - \beta)T_{t-1} \quad (\text{Equation 4-11})$$

An additive seasonality is observed when the magnitude of the seasonal pattern is constant over time and multiplicative seasonality is observed when the magnitude of the seasonal pattern increases or decreases over time. Additive seasonality and multiplicative seasonality are presented below where γ is the learning rate coefficient for the seasonality element of the series:

$$\text{Additive seasonality: } S_t = \gamma(y_t + L_t) \quad (\text{Equation 4-12})$$

$$\text{Multiplicative Seasonality: } S_t = \gamma(y_t * L_t) \quad (\text{Equation 4-13})$$

The forecast can also be additive or multiplicative. The additive forecast is observed when all components (level, trend, and seasonality) are added together and the multiplicative forecast is observed when all components are multiplied together.

Below is Holt-Winter's additive forecast composed of an additive trend and multiplicative seasonality:

$$F_t = L_t + T_t + S_t = \left[\alpha \frac{y_t}{s_{t-m}} + (1 - \alpha)L_{t-1} + T_{t-1} \right] + [\beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1}] + \left[\alpha \frac{y_t}{L_t} + (1 - \gamma)S_{t-m} \right] \quad (\text{Equation 4-14})$$

Implementation of Holt-Winter's model in Python

The model was imported from the Statistical Models library. To fit the model, the training set, the trend which can either be additive or multiplicative, and the seasonal type which can either be additive or multiplicative were inputted to the model. The trend and seasonality types were applied interchangeably for each time series data and the Akaike Information Criterion (AIC) metric was computed.

For each training set, the best model was selected based on the minimum AIC value. The best models were used to forecast the number of units or rows equivalent to the ones of the testing sets. Both the models' forecast and test sets are plotted for each time series data as illustrated in the figures below.

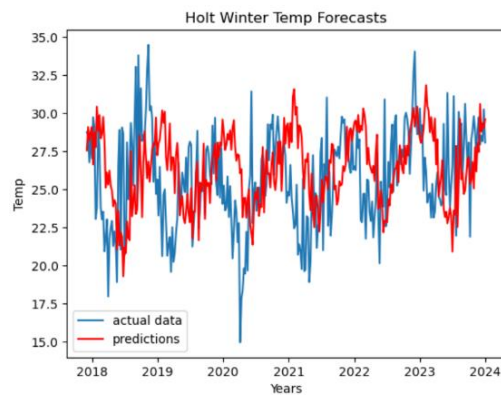


Figure 4- 26: Temperature forecasts using Holt Winter's model (Source: Author).

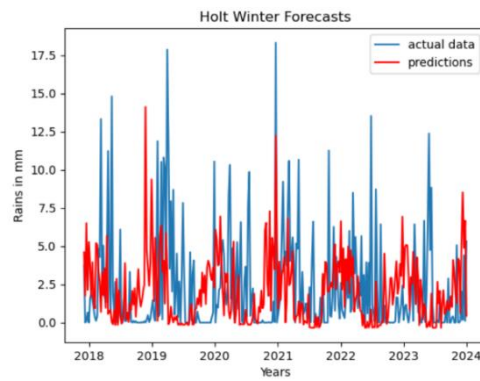


Figure 4- 27: Rainfall forecasts using Holt Winter's model (Source: Author).

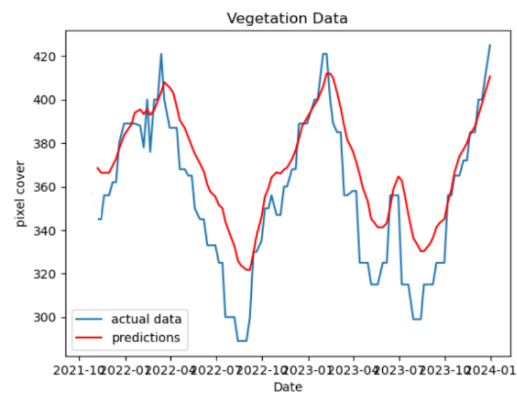


Figure 4- 28: Vegetation data forecasts using Holt Winter's model (Source: Author)

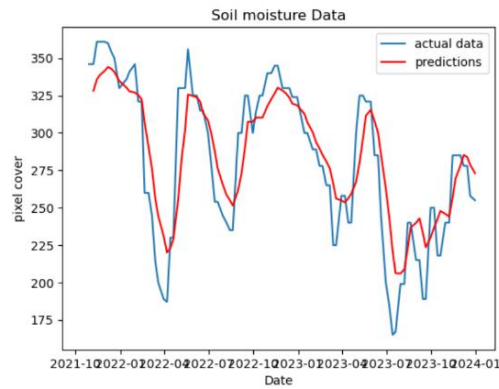


Figure 4- 29: Soil moisture data forecasts using Holt Winter's model (Source: Author).

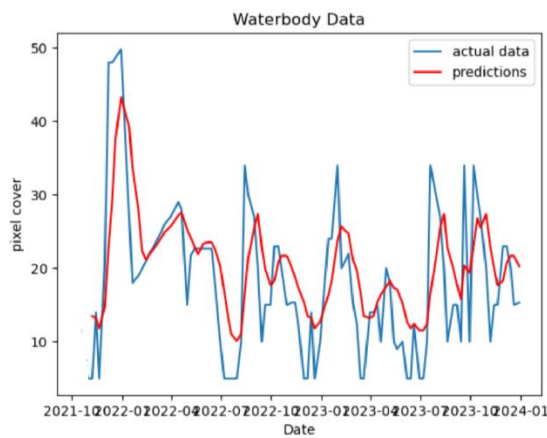


Figure 4- 30: Water surfaces time series data forecasts using Holt Winter's model (Source: Author).

The RMSE was computed using the forecast and the actual data. The results are discussed later in this chapter.

Background knowledge of SARIMA

SARIMA is an ARIMA model that takes care of the seasonality effects of a time series. An ARIMA model is composed of two distinct models which explain the behaviour of the series from two different angles. These are the Autoregressive (AR) and Moving Average (MA) models. The integration part denoted as (I) reveals how many times the differencing was performed to stationarise the series. A stationary series is a series that has no increasing or decreasing trend and has constant mean and variance over time. According to Perez and Ceballos (2019), most of the time series data are not stationary by nature; hence, they have to go through the transformation process called Box-Cox transformation or differencing also

known as differencing to stationarise them. Differencing is a gap between the two data points computed by subtracting the current value (y_t) from the previous value (y_{t-1}) (Verma and Kaur, 2017). The researcher explains the models (AR and MA) separately and their integration afterwards.

AR model

AR model performs forecasting by first establishing the level of linear dependence between the lags of the series. It is given the name “auto” because it assumes that the current value of a time series is a function of its past values. The series or variable has a higher linear dependence on its recent regressed values and a lower dependence on its past regressed values for as long as data points are not randomly distributed. The model has a parameter that holds the number of lags that the series is highly correlated with. For instance, if the AR model has a two-lag parameter, it implies that today’s forecast f_t can be explained by yesterday’s forecast (f_{t-1}) and the day before yesterday (f_{t-2}) plus an error term as shown below:

$$AR(2): f_t = \mu + \theta_1 f_{t-1} + \theta_2 f_{t-2} + \epsilon_t \quad (\text{Equation 4-15})$$

where μ is a constant term; θ_1 and θ_2 are correlation coefficients computed by the model for each regressed lag, and ϵ_t is the error term at time (t). The $AR(1)$, $AR(2)$ and $AR(p)$ models are presented below.

$$AR(1): f_t = \mu + \theta f_{t-1} + \epsilon_t \quad (\text{Equation 4-16})$$

$$AR(2): f_t = \mu + \theta_1 f_{t-1} + \theta_2 f_{t-2} + \epsilon_t \quad (\text{Equation 4-17})$$

$$AR(p): f_t = \mu + \sum_{i=1}^p \theta_i f_{t-i} + \epsilon_t \quad (\text{Equation 4-18})$$

The number of lags for the AR model is precisely identified by the PACF model.

MA Model

The MA model is similar to the AR model; however, the forecasted value of today (f_t) is determined by the previous or regressed error terms or past forecasted mistakes. The ACF model is used to precisely determine the correlation coefficients and the order movement of this model. The $MA(1)$, $MA(2)$ and $MA(p)$ are presented below.

$$MA(1): f_t = \epsilon_t + \theta \epsilon_{t-1} \quad (\text{Equation 4-19})$$

$$MA(2): f_t = \epsilon_t + \theta \epsilon_{t-1} + \theta \epsilon_{t-2} \quad (\text{Equation 4-20})$$

$$MA(p): f_t = \epsilon_t + \sum_{i=1}^p \theta_i \epsilon_{t-i} \quad (\text{Equation 4-21})$$

ACF and PACF models

The ACF and PACF models are estimation approaches that are used to estimate the parameters of the ARIMA models and the correlation coefficients of the models (Perez and Ceballos, 2019). The ACF also known as serial correlation states that if the data is not randomly distributed, then each value is correlated with its previous lag. This means that the results of each lag are influenced by the results of the preceding lags. The correlation of the two tested lags, for instance, lag Y_i and lag $Y_{i+(n-l)}$, is influenced by the intermediate lags. Using the ACF model, it is difficult to determine the correlation between two non-adjacent lags because each lag is influenced by neighbouring lags. The PACF model comes in handy. This model is also known as the conditional correlation model since it performs a correlation based on a condition. Given lag Y_i and Y_{i-n} , it will take out all the intermediate effects in between the tested lags.

To determine the parameters of the AR model using the correlogram tool, the ACF model is known to be significant to the n th lag while the PACF is geometric (lags slowly decay or slowly converge towards 0). However, for the MA model, the PACF is significant to the n th lag while the ACF is geometric. When both ACF and PACF are geometric, the series is declared non-stationary and it must be transformed to stationarity through differencing or any other transformation technique.

ARIMA

The ARIMA model is rendered with the notation (p, d, q) where p denotes the order process of the AR model; d denotes the number of times differencing was applied to stationarise the series, and q denotes the order movement of the MA model (Verma and Kaur, 2017). These ARIMA family models assume that the series in question is stationary. If the series is not stationary, it has to be stationarised before using the models. The $ARIMA(1,0,1)$ or $ARMA(1,1)$ model presented below shows today's forecast (f_t) to be explained by yesterday's forecast (F_{t-1}) and yesterday's forecast mistakes (ϵ_{t-1}):

$$ARMA(1,1): F_t = \mu + y_1 F_{t-1} + \epsilon_t + d_1 \epsilon_{t-1} \quad (\text{Equation 4-22})$$

Below is the general equation of the ARMA model:

$$ARMA(p, d, q): F_t = \mu + \sum_{i=1}^p y_i F_{t-i} + \epsilon_t + \sum_{j=1}^q d_j \epsilon_{t-j} \quad (\text{Equation 4-23})$$

SARIMA

The current problem with the ARIMA model is that it does not account for the series that has the seasonality effects; hence, the SARIMA model comes in handy. The SARIMA is an ARIMA model that takes care of the seasonality effects of a time series. SARIMA is presented with 7 parameters as shown below:

$$ARIMA(p, d, q)(P, D, Q)_m \quad (\text{Equation 4-24})$$

where (m) is the seasonal factor (number of periods it takes for seasonality to repeat).

P -number of seasonal AR terms;

D -number of seasonal differences;

Q -number of seasonal MA terms;

For instance, $SARIMA(2,0,1)(2,1,1)_{12}$ can be explained as

$$ARIMA(2,0,1)(2,1,1)_{12}: F_t = \mu + y_1 F_{t-1} + y_2 F_{t-2} + \epsilon_t + y_3 F_{t-12} + y_3 F_{t-24} + d_1 \epsilon_{t-1} + d_2 \epsilon_{t-12} + \epsilon_t \quad (\text{Equation 4-25})$$

The seasonality component (m) controls the shift of P, D, and Q notations.

The general equation of SARIMA is presented below:

$$ARIMA(p, d, q)(P, D, Q)_m = \mu + \sum_{i=1}^p y_i f_{t-i} + \sum_{i=1}^q \theta_i \epsilon_{t-i} + \sum_{i=1}^P \phi_j f_{t-si} + \sum_{i=1}^Q \mu_j \epsilon_{t-sj} + \epsilon_t \quad (\text{Equation 4-26})$$

Implementation of SARIMA in Python

The SARIMA model, identified as SARIMAX was imported together with ACF and PACF libraries from the Statistical Models library. The daily rain and maximum temperature were aggregated into weeks using the *MEAN()* function. This was performed to minimise the model fitting process which is considered a resource-heavy computation.

The first step was to evaluate the stationarity of the time series data to determine if transformation has to be done or not. The Augmented Dickey-Fuller (ADF) test was performed to test the stationarity of each time series data. To conclude that the series is stationary, the null hypothesis must be rejected, and the alternative hypothesis must be accepted. The null hypothesis states that the series in question is not stationary by nature and the alternative hypothesis states that the series is stationary by nature. The P-statistic value is the one that is

primarily used to evaluate the stationarity of the series. If it is less than 5%, the series is considered stationary, otherwise, it is considered not stationary and, as a result, the transformation techniques will be implemented to stationarise the series. Figure 4-31 below is the function that receives the time series data and tests its stationarity.

```
def adfuller_test(timeseries_data)
    result=adfuller(timeseries_data)
    label=['ADF test statistics','p-value','#lags used','Number of observations used']
    foreach value,label in zip(result,label)
        print(label+' : '+str(value))
        if result[1] <=0.05 then
            print("Reject null hypothesis, series stationary")
        else
            print("Accept null hypothesis, series not stationary")
        end
    endfor
    return
```

Figure 4- 31: A function that evaluates the stationarity of the series (Source: Author)

Each time series data set was passed to the function to evaluate its stationarity as demonstrated in Figure 4-32 below with vegetation data.

```
Vegetation Data
-----
ADF test statistics : -6.314577030449437
Reject null hypothesis, series stationery
p-value : 3.1735769964374085e-08
Reject null hypothesis, series stationery
#lags used : 10
Reject null hypothesis, series stationery
```

Figure 4-32: A console screen demonstrating the stationarity of the vegetation time series data (Source: Author)

The researcher did not apply any transformation techniques, given that all the time series data were stationary. The next step was to determine the order movement of the SARIMA model for each time series. The autocorrelation plots (ACF and PACF) were created for all the time series data. However, the accurate estimation of the lags for SARIMA through visual observation was complicated due to a massive number of parameters that need to be estimated. To get around the selection of optimal model parameters, the researcher wrote a function that uses nested loops to test a wide range of candidate values representing model parameters. For each iteration, model fitting is performed, and the AIC matrix is computed.

The model for each time series that showed the minimum error rate was selected. Each model was used to forecast the number of units or weeks equivalent to test sets.

The figures presented below are the plotted graphs of the time series data predictions where blue represents the actual data and red represents the predictions.

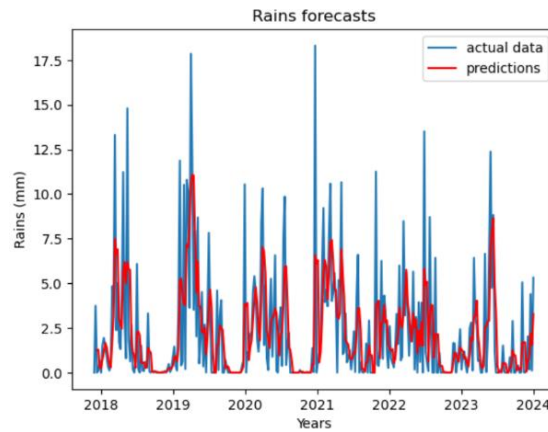


Figure 4- 33: Rainfall forecast using SARIMA model (Source: Author)

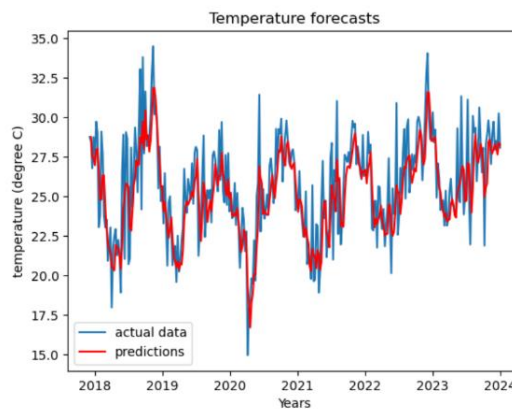


Figure 4- 34: Temperature forecast using SARIMA model (Source: Author)

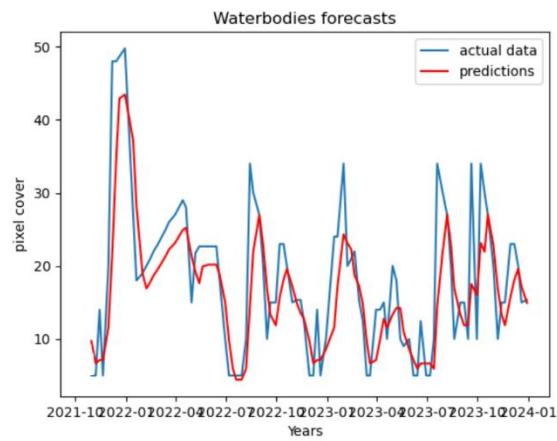


Figure 4- 35: Waterbodies forecast using SARIMA model (Source: Author)

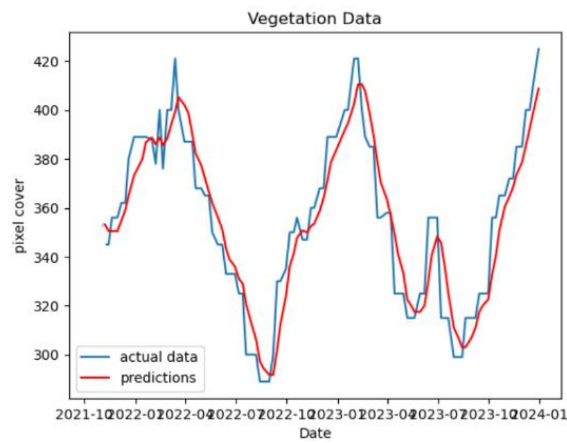


Figure 4- 36: Vegetation forecast using SARIMA model (Source: Author)

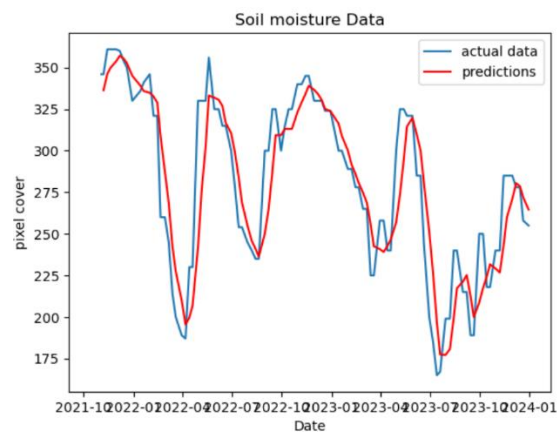


Figure 4- 37: Soil moisture forecast using SARIMA model (Source: Author)

The RMSE was also computed for each time series data. The results will be discussed later in this chapter.

LSTM Background information

The LSTM model is an extension of a Recurrent Neural Network (RNN) model that suffers from the exploding and vanishing gradient problems during the backpropagation process used by the network to train itself. The RNN suffers from these problems due to several multiplications that occur resulting in the model's inability to learn long-term dependencies (Kong et al., 2017). This is where LSTM comes in place by eliminating the traditional RNN problems. The LSTM network consists of the internal memory cell state (that represents the long-term memory), the hidden state (that represents the short-term memory), and the gates that solve the vanishing and exploding gradient problems. The first gate of the LSTM is known as the forget gate (f_t), is the one that determines how much of the previous long-term memory to remember going forward. This gate uses the sigmoid (σ) activation function that squashes any value into the range of 0 to 1. If it is 0, it simply means all the knowledge kept in the previous cell state will be thrown or forgotten moving forward (given that every value multiplied by 0 will be 0). If it is 1, it means all the knowledge kept in the previous memory cell of the network will be considered or remembered moving forward. Hence, values close to 0 will be thrown while those close to 1 will be kept. The mathematical presentation of the forget gate is illustrated below:

$$f_t = \sigma(W_{fx}X_t + W_{fh}h_{t-1} + b_f) \quad (\text{Equation 4-27})$$

where the input at the given timestep (X_t) and the previous short-term memory (hidden state h_{t-1}) adjusted by the weights (W_{fx} , W_{fh}) and bias b_f are fed to the sigmoid activation function.

The second gate of the network is known as the input gate (i_t) that also uses the sigmoid activation function to determine how much of the potential long-term memory (that is newly created using the candidate cell state) to remember or forget going forward. This gate is computed using the current input and the previous hidden state adjusted by weights and bias as illustrated below:

$$i_t = \sigma(W_{ix}X_t + W_{ih}h_{t-1} + b_i) \quad (\text{Equation 4-28})$$

The potential long-term memory of the LSTM also known as the candidate cell state (g_t) is calculated using the $\tanh$ activation function that squashes any input value into the range of -1 to +1. The candidate cell state is calculated using the current input and previous hidden state adjusted by the weights and the bias that will be fed to the $\tanh$ activation function as illustrated below:

$$g_t = \tanh(W_{gx}X_t + W_{gh}h_{t-1} + b_g) \quad (\text{Equation 4-29})$$

The current cell state (c_t) that represents the new long-term memory of the network is calculated by adding the previous long-term memory c_{t-1} (that has been adjusted by the forget gate) with the potential long-term memory g_t (that has been adjusted by the input gate) as illustrated below:

$$c_t = g_t \cdot i_t + c_{t-1} \cdot f_t \quad (\text{Equation 4-30})$$

The output gate (o_t) is the last gate of the network that also uses the sigmoid activation function. It determines how much of the current long-term memory (c_t) to remember or consider to create the network's current short-term memory. Similarly to other gates, it uses the current input and previous short-term memory adjusted by weights and bias as displayed below:

$$o_t = \sigma(W_{ox}X_t + W_{oh}h_{t-1} + b_o) \quad (\text{Equation 4-31})$$

The current short-term memory (h_t) is created using the current long-term memory (c_t) that has to first pass through the $\tanh$ activation function before it is adjusted by the output gate as illustrated below.

$$h_t = \tanh(c_t) \cdot o_t \quad (\text{Equation 4-32})$$

The current short-term memory or hidden state (h_t) and the current long-term memory or cell state (c_t) are then carried out to the next timestep ($t+1$).

The key reason for using LSTM is that the cell state sum activities over time which can overcome vanishing gradient and better capture long-term dependencies of time series (Qin et al., 2017).

Implementation of LSTM

The LSTM model was also loaded from the Statistical Models library. Given that the LSTM is sensitive to the scale of the data, the *Minmaxscaler()* function was applied to transform the time series data into a range of 0 to 1. The smallest values were close to 0, while the largest

values were close to 1. To prepare data for model fitting, training, and test sets were further processed by splitting each sample into dependent and independent features based on a given timestep that will be provided as an argument to the function. The timestep describes the number of previous steps that have to be considered to make the next prediction. For instance, to predict y_t considered as a dependent variable, the model will first learn from the following regressed independent features $y_{t-1}, y_{t-2}, y_{t-3}, \dots, y_{t-n}$. where n is the number of timesteps provided to the model. The LSTM will be expected to learn the independent features and predict the dependent feature for both training and testing samples of every time series. The function that splits the given sample into two subsamples based on the timestep provided was developed and it was invoked by passing both training and testing samples for a given time series. Before model fitting, the training and testing samples were reshaped into three-dimensional form (sample, timestamp, input features). The input parameter was assigned the value of 1 because the model was expected to predict one value for every iteration. The stacked LSTM model with three hidden layers and an output layer was created. The model was compiled with the RMSE loss function and ‘Adam’ optimiser. The model training was performed for every time series’ generated samples. The model was used to predict the dependent features in both training and test sets of every time series. To evaluate the model’s performance, the data were rescaled back to the original form.

The figures presented below are the plots of the time series forecasts made for all variables.

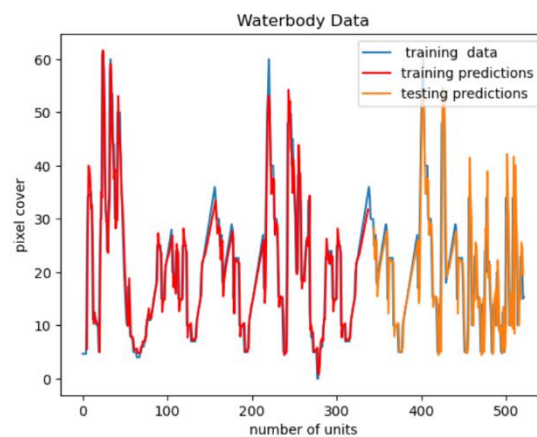


Figure 4- 38: Water surfaces forecast using LSTM model (Source: Author).

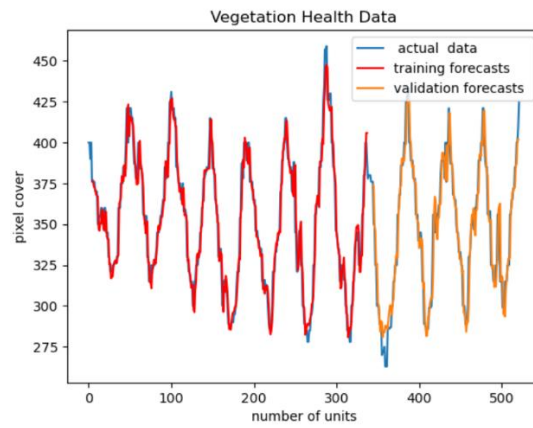


Figure 4- 39: Vegetation health forecast using LSTM model (Source: Author).

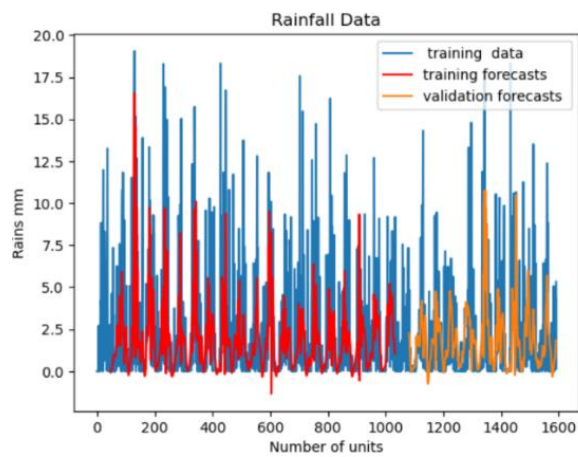


Figure 4- 40: Rainfall forecast using LSTM model (Source: Author)

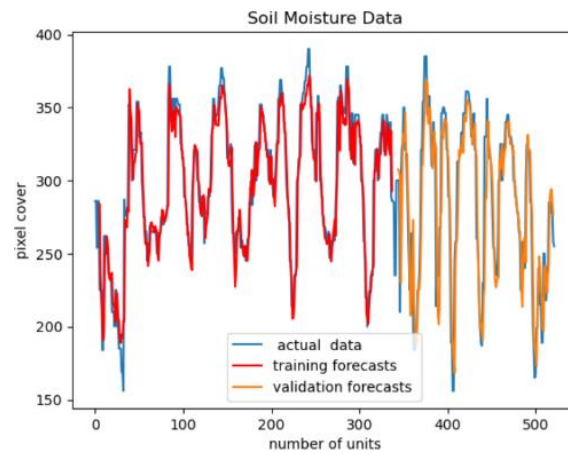


Figure 4- 41: Soil moisture forecast using LSTM model (Source: Author).

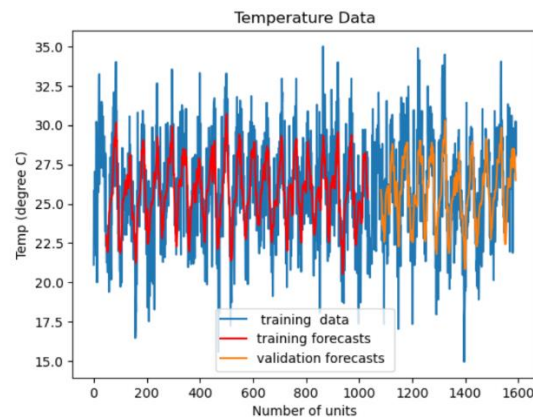


Figure 4- 42: Temperature forecast using LSTM model (Source: Author).

The RMSE was computed for time series data. It will be discussed in the next section

Models' Evaluations and future forecasts

Table 4-2 below, is the summary of the models' performances in various knowledge domains.

Table 4- 2: Presentation of models' performance (Source: Author).

Models	RMSE				
	Weather Data		Satellite Data		
	Rain	Max Temp	Vegetation Cover	Waterbodies Cover	Soil Moisture Cover
Holt Winter's model	9.34	10.21	5.89	6.44	4.71
SARIMA(1,0,1)(1,0,2) ₄₈	2.66	3.34	2.99	3.37	4.24
LSTM	5.45	4.56	2.43	3.26	4.11

The SARIMA(1,0,1)(1,0,2)₄₈ model was the most effective for forecasting weather data, achieving an RMSE of 2.66 for rainfall and 3.34 for temperature. In contrast, the LSTM excelled in processing satellite imagery data, with RMSE values of 2.43 for vegetation cover, 3.26 for waterbody coverage, and 4.11 for soil moisture coverage.

These models were used to forecast the next 48 weeks' time series data. Before forecasting, the training and testing samples were merged back together and the *forecast()* function was used for the SARIMA model. Given that the LSTM performs predictions based on the timesteps, the algorithm that uses the LSTM model to make one-step predictions using previous timesteps was developed. So, as this algorithm shifts to predict the next values, it appends itself using the previously made forecast. The drawback of this prediction approach is that as the model continues to make further predictions, the reliability of future forecasts decreases. This is because the predictions increasingly rely on previous forecasts rather than actual data.

The models' forecasts were stored in the cloud database for later manipulation.

4.2.4. Quantifying the season behaviour predictions made via IK system

4.2.4.1. Preparation of historical data to complement the IK system

The historic weather and processed satellite data (time series data of land cover features) of the warm rainy season (August - December) and the cold rainy season (March - May) were extracted from their original data sample. The idea of extracting these data was to quantify the predictions that will be made from the IK system and be able to integrate them with the scientific forecasts. For instance, if farmers are certain of abundant rains coming based on the

observation of the IK indicators, then the historic weather and satellite data of seasons when abundant rains were experienced will be extracted and manipulated to forecast the possible behaviour of rains. This idea came about as a result of the limitation the IK system has in revealing insights regarding the possible onset, distribution, cessation and magnitude of rainfalls throughout the season.

So, for this idea to be possible, the data for every agricultural season had to be extracted, visualised, and categorised into either abundant, drought or floods based on how data behave throughout the season. At this stage, the extracted historic data of seasons were still in raw format (in mm for rain data, in degrees Celsius for temperature data, and in pixel intensity values for satellite data). As such, it became difficult for the researcher to conclude on the overall behaviour of every seasonal data. To get away with this problem, the algorithm responsible for normalising (scaling down) the data was developed. The algorithm receives the time series data for a given knowledge domain (vegetation, soil moisture, waterbodies, rainfall and temperature), it loops through the data and searches for the minimum and maximum values from the time series data. The medium is then calculated as the mean of the minimum and maximum. These values are further partitioned to come up with other centre values that will assist in classifying every extracted season behaviour data point into the range of 1 to 6 where 1 represents very low and 6 represents very high as illustrated in Figure 4-43 below.

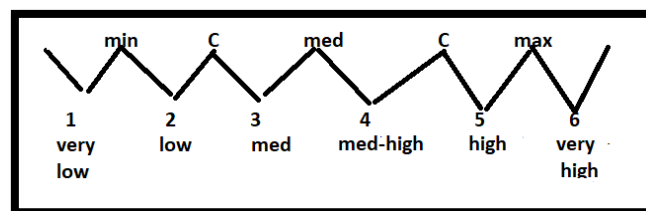


Figure 4- 43: Illustration of how new data points will be normalised into the range of 1-6 (Source: Author).

The algorithm was used to loop through every extracted season behaviour data, to analyse each data point and classify it accordingly.

This approach also helped the researcher to merge the normalised rain data, soil moisture data and waterbody data to become a single time series data that represents water. This was done to simplify their visualisation and classification process. Every normalised season behaviour data

set was further classified into either abundant, floods or drought seasons by focusing more on the changes in water and vegetation time series movement. The season behaviour patterns that became difficult to manually classify based on the observation of the time series data were ignored.

The classified or categorised data were stored in the cloud database for further manipulation.

Rainy season predictions using IK system

When a farmer uses a mobile application to predict the season behaviour, the farmers' CL of an oncoming rainy season behaviour will be calculated using the observed IK indicators and will also be evaluated against a given threshold. So, if farmers are certain of for instance, abundant rains coming soon, the categorised historic weather and satellite data of periods when abundant rains were experienced will be extracted from the cloud database. The extraction of these data also depends on the timestamp the predictions are made by the farmer. For example, if a farmer uses a mobile application to predict the abundant rains in the middle of the rainy season, then, the extracted historical data that signify abundant rains in the middle of the season will be considered while ignoring those that may signify abundant rains either early or late in the season.

The algorithm that evaluates the behaviour of the next weekly data points based on the current week the predictions are made was developed. These weekly data points are evaluated based on their knowledge domains (water, green vegetation cover and temperature) and the predicted season behaviour (scarce, excessive or normal rains or temperature changes). This algorithm helps in filtering only the historical season patterns that are relevant to the predicted season behaviour. The number of the next data points that will be evaluated is determined by the given knowledge domain. For instance, the next two data points can be considered for water and temperature data, and three for vegetation data. This is because the vegetation cover is slow to respond to the effects of both rain and temperature unlike water surfaces and soil moisture. However, the number of values representing the number of weeks to be considered will be adjusted based on farmers' experience with their environment and the season. The increasing or decreasing trend in the selected set of data will be evaluated by the algorithm considering the anticipated rainy season behaviour and the knowledge domain being evaluated. For example, If good or normal or abundant rains are expected to come soon (in the next coming weeks), so, based on the current week the forecast is made, the next two weeks' data points of water and the next three weeks' data points of the vegetation cover will be evaluated to identify

increase in them. Conversely, when the drought season is expected, the decline in the selected subset of the next weeks' data for water and vegetation time series, will be evaluated. When heavy rains (floods) are expected, the increase in water and the decrease in vegetation will be evaluated from the given subset of data points. This is because heavy rains are expected to flush off the crops and other vegetation. Furthermore, when farmers are certain of the seasonal changes from cold to warm season, the data points that indicate the increase in temperature and green vegetation cover will be considered and visa-versa to when the season shifts from warm to cold.

The increasing or decreasing pattern in the given subset of the data is computed by summing the differences of the observations starting from the future predictions towards the recent predictions as illustrated below:

$$rate = \sum_{i=n}^1 E_i - E_{i-1} \quad (\text{Equation 4-33})$$

where n is the number of observations in the selected subset.

If the rate is greater than 0, it indicates an increase in the observations. If it is less than 0, it indicates a decrease. Otherwise, the rate is considered constant.

The algorithm presented below in Figure 4-44 receives the historical data collected based on the current timestamp the forecast is made using the application, the season behaviour type being predicted, the knowledge domain being tested, and the count of the next data points (weeks) that need to be considered. The algorithm will loop through every record, extract the next data points and make a validation. If the record is valid, it will be added to the list of valid records. The successful records will be submitted to the SES time series model to perform the next season's forecast to quantify the seasonal predictions made via the IK system.

```

Collection getDataChanges(Collection data, Text SeasonType, num countOfDataPoints, Text knowledgeDomain)
Declarations
  Collection validData = new Collection
  boolean flag
  num[] values
  num value
for Object d : data
  values = new num[countOfDataPoints]
  for num c = 0, c < countOfDataPoints, c++
    value = ((FarmersCategorisedData) ((ArrayList) d).get(c+1)).getValue()
    values[c] = value
  endfor
  flag = isValid(values, seasonType, knowledgeDomain)
  if flag then
    validData.add(d)
  endif
endfor
return validData

```

Figure 4-44: An algorithm that extracts valid historic season patterns based on an evaluated trend (Source: Author).

The SES model will predict the next value for every record of the historic data as shown in Table 4-3 below where $t+1$ is the next season forecast that will be generated from the historic season data (past observations indicated as $t, t-1, t-2, t-3 \dots t-n$).

Table 4- 3: Example of historic, current, and future time series observations (Source: Author).

	t-5	t-4	t-3	t-2	t-1	t	t+1
Week 1	0	3	2	4	4	0	?
Week 2	1	0	0	3	4	0	?
.	.	.	.	.	.	.	?
.	.	.	.	.	.	.	?
Week 20	2	4	0	3	5	1	?

The SES model was developed and tailored to fit the requirements of this thesis. Its selection was driven by the fact that it is both simple to implement and a robust time series model, capable of working with small datasets. Unlike SMA and WMA models discussed in Chapter 2, it considers all past observations, and it has a weighting scheme that prioritises the most recent data. In addition, its learning rate can be optimised to produce more accurate predictions. Although the model does not perform well with data that exhibits trend or seasonality effects, the researcher considers it the best option compared to more complex models like DES and Holt-Winters. These advanced models require the estimation of multiple parameters, which increases computational time and slows down the system's response.

The algorithm presented in Figure 4-45 below runs through the SES learning rate's candidate values that range from 0.1 to 0.9, and the historical data that is organised into rows and columns. The algorithm calls the SES model's $fit()$ function to perform model fitting passing every candidate learning rate (known as α) and every record extracted from the

historical data. The model's *fit()* function returns the average error rate for every evaluated record. The error rate is calculated as a squared difference between the model's prediction and the actual data. The errors will be accumulated for every model's learning rate candidate value and will be returned. The extreme values of the learning rate (0 and 1) are ignored to avoid under-smoothing and over-smoothing where the model does not learn. For instance, if α is 1, past values do not influence forecasts (under-smoothing) and if α is 0, past values have equal influence on forecasts (over-smoothing).

```
Collection fitting(num[][] weeklyRecords)
  Declarations
    num error = 0, alpha, col, row
    Collection errors=new Collection(), colElements = new Collection(),
      errors =new Collection()
    ExponentialSmoothing model = new ExponentialSmoothing()
  for alpha = 0.1, alpha < 1, alpha += 0.1
    for col = 0, col < weeklyRecords[0].length, col++
      for row = 0, row < weeklyRecords.length, row++
        colElements.add(weeklyRecords[row][col])
      endfor
      error += model.fit(colElements, alpha)
    endfor
    errors.add(error)
  endfor
  return errors
```

Figure 4- 45: An algorithm that evaluates different learning parameters of SES model (Source: Author).

The SES model's *fit()* function is presented in Figure 4-46 below. This function receives a record from the historic data and the candidate learning rate as parameters. In the first iteration, predictions are performed starting from the old observations towards the recent observations. The model predicts the next forecast as today's forecast plus an error term of today adjusted by the learning rate. The second iteration compares the model's predictions with the expected values to get the squared error terms. The squared errors are averaged and returned.

```

num fit(Collection dataPoints, num alpha){
  Declarations
  num error, actual, count, prediction, sumError=0
  Collection predictions=new Collection()
  for num count = 1, count < dataPoints.size(),count ++
    if count = 1 then
      prediction = dataPoints.get(0)
    else
      prediction = prediction + alpha * (dataPoints.get (count - 1) – prediction)
    endif
    predictions.add(prediction)
  endfor
  for num count = 0, count < predictions.size(), count++
    actual = dataPoints.get(count + 1)
    prediction = predictions.get(count)
    error = Math.pow(actual – prediction,2)
    sumError +=error
  endfor
  return sumError / predictions.size()

```

Figure 4- 46: The SES model's *fit()* function that performs learning based on the given alpha (Source: Author).

The algorithm in Figure 4-47 below receives the list of the squared error terms for every learning rate. It will find the optimal learning rate (one with minimum squared errors) that will soon be used to make the next forecasts.

```

num getBestAlpha (Collection errorRates)
  num min = errorRates.get(0)
  num bestAlpha = 0.1
  for num alpha = 0.1, count = 0; alpha < 1 && count < errorRates.size(); alpha += 0.1, count++
    if min > errorRates.get(count) then
      min = errorRates.get(count)
      bestAlpha = alpha
    endif
  endfor
  return bestAlpha

```

Figure 4- 47: An algorithm that returns optimal model's learning rate parameter (Source: Author)

To prepare data for the next season's predictions using the optimum learning rate, the following algorithm in Figure 4-48 receives the historical data and the optimum learning rate as parameters. The function will loop through the historical data, take every record and send it to the model's *predict()* function together with the best learning rate. The model's *predict()* function will return the next forecast for a given record. The returned forecast will be added to the prediction list that the function will later return.

```

Collection getIKPredictions(num bestAlpha, num[][] weeklyRecords) {
  Declarations
  Collection predictList = new Collection(), colElements = new
Collection()
  num prediction
  ExponentialSmoothing model = new ExponentialSmoothing()
  for num col = 0, col < weeklyRecords[0].length, col++
    for num row = 0, row < weeklyRecords.length, row++
      colElements.add(weeklyRecords[row][col])
    endfor
    prediction = model.predict(colElements, bestAlpha)
    predictList.add(prediction)
  endfor
  return predictList
}

```

Figure 4- 48: An algorithm that performs season behaviour forecast using historic records (Source: Author).

The model's *predict()* function is presented below in Figure 4-49. The function receives a historical data record and the optimum learning rate as parameters. The model will learn from the given record and predict the next value. The predicted value will be returned.

```

num predict(Collection dataPoints, num alpha)
  Declarations
  num count, prediction
  for count = 1, count <= dataPoints.size(), count ++
    if count = 1 then
      prediction = dataPoints.get(0)
    else
      prediction = prediction + alpha * (dataPoints.get(count - 1) - prediction)
    endif
  endfor
  return prediction
}

```

Figure 4- 49: The SES model's prediction method (Source: Author).

4.2.5. Integration of forecasts done via IK and science

Given that the scientific forecasts are still in raw format (in degrees Celsius for temperature, in millimetres for rainfall data and pixel intensity values for satellite data), the algorithm explained in the previous section will be used to normalise (scale down) the data into the range of 1 to 6. The scaled-down scientific rainfall, water surface cover and soil moisture cover forecasts will be merged to represent the water time series. The integration of both the normalised IK and science forecasts will be made, and the integrated time series forecasts will be passed to the data interpretation algorithm. The integration is made by finding the mean of the two observations at a given timestep. The data interpretation algorithm aims to explain every forecasted data point based on its knowledge domain in a simple way a farmer can understand. It will receive the time series data, time series type (water or temp or vegetation

cover), starting date and ending date of every predicted week. The algorithm will describe forecasts in the form of words. For instance, *from the 29th of September to the 05th of October, no rains will be expected; from the 06th of October to the 12th of October, below-normal rains will be expected*, and so on. The forecasts will also be presented in a tabular format as shown in Table 4-4 below with water time series.

Table 4- 4: An illustration of forecasts presented in a tabular format (Source: Author).

Week number	Week Date	Forecasted values	Description
1	29 Sept to 5 Oct	2	Below normal
2	06 Oct to 12 Oct	1	No rains
3	13 Oct to 19 Oct	1	No rains
4	20 Oct to 26 Oct	3	Below normal to Normal
5	27 Oct to 02 Nov	4	Normal rains
6	03 Nov to 09 Nov	5	Above normal rains
7	10 Nov to 16 Nov	6	Extreme rains

Although the farmer will receive forecasts in the form of a table and a text description, the advanced display option of the season behaviour will be available, to enable farmers to visualise the predicted season behaviour by observing the line graph representing the time series movement of water, healthy vegetation cover and temperature.

4.2.6. Development of an optimisation model

Although season behaviour predictions are important to give farmers insight into ways cropping decisions can be improved, farmers are still faced with numerous biotic and abiotic factors that they have to compete with throughout the lifecycle of their crops. Although farmers have stressed how they map various crop symptoms to various crop problems, they have also mentioned that most of the time it becomes difficult through just an ordinary observation to make precise decisions given that a crop problem(s) can be attributed to a wide range of biotic and abiotic factors driven by climate change. Further, a wrong diagnosis may result in the wrong treatment that can exacerbate an already inflicted stress on a crop such as overfertilization or overwatering. Farmers have also declared that lack of guidance and support from the scientists or extension officers is also a concern, particularly to those farmers who have now introduced synthetic inputs in their farms. The knowledge gathered made the researcher fully understand that the value and strength of the IK continue to deplete as farmers' trust shifts to the use of chemicals. Furthermore, what is currently standing between farmers and better crop yield is a lack of information and guidance that is primarily contained by the

IK experts, agricultural or crop scientists and the use of wireless sensors. So, to bridge this knowledge barrier, the researcher developed the ML optimisation model that continually learns from the IK and crop science experts with the support of the soil and climate sensors. The model uses the knowledge it has gained to guide farmers in addressing various cropping challenges such as crop selection, fertiliser selection, diagnosis and treatment of various crop problems and many more. The proposed optimisation model works together with various algorithms that scale down raw sensor data into a format that farmers can understand and for quick processing by the ML optimisation model.

This model was developed in a way that it can continually learn from a set of questions and answers in different problem domains similar to humans. When a new question is provided, the model will use its knowledge to provide answer(s) to the question. Questions are structured in a way the machine can understand without employing sophisticated natural language processing tools. A question comprises one or more features generated through human observation or wireless sensor technology. An answer is one or more recommendations provided for a question.

The expert system that houses the optimization model provides a platform where questions and answers can be categorised into different problem domains to simplify the learning process by the model. This expert system further provides a platform where information can be created by different users with different roles in the system. For instance, the system experts are the ones responsible for creating features and recommendations and being able to map them to create new knowledge with the guidance of IK experts and scientists.

Below is a set of questions and guidelines on how every question will be structured for a given problem domain.

Q1: What crops can I plant when abundant rains are expected to start in the middle of the season given that my soil texture is loam to sandy with low nitrogen, low potassium, and high phosphorus?

Q2: How can I help my six-weeks-old tomato crop that starts to show signs of leaf tips wilting, leaf yellowing and stunted growth planted in an environment that is highly humid with high temperatures and low wind movement?

Q3: What fertiliser can I apply to improve the health of my four-weeks-old maize crop given that the nitrogen in the soil is very low, and potassium and phosphorus are moderate?

Q3.1. If now I know the fertiliser to use, how much of this fertiliser should I apply on my 1-hectare land size?

Table 4-5 below illustrates how the different questions above, rooted in various problem domains, will be modelled using the features. It is also worth considering that some of the features will be generated from the sensor data and will also be scaled down for easier interpretation by the farmer.

Table 4- 5: An illustration of forecasts presented in a tabular format (Source: Author)

Q1	Q2	Q3	Q3.1.
Problem Domain: Crop Selection	Problem Domain: Crop health problem	Problem Domain: Fertiliser selection	Problem Domain: Quantity of fertiliser
Features	Features	Features	Features
Rains type(magnitude): normal Rains onset: middle of the season Soil texture: loam to sandy Soil nitrogen: low Soil potassium: low Soil phosphorus: high	Crop name: tomato Crop age: six weeks Leaf tips: wilting Leaf colour: yellowing Crop growth: poor Humidity: high Temperature: high Wind speed: low	Crop name: maize Crop age: four weeks Soil nitrogen: very low Soil potassium: moderate Soil phosphorus: moderate	Fertiliser name: ? Crop name: maize Crop age: young Land size: 1 h

The knowledge that will be fed (inputted) to the machine is segmented into three parts: problem domain, question, and recommendations. A question is made up of one or many features and can be answered by one or many recommendations. A recommendation or answer can be provided to one or many questions. Both questions and recommendations strictly belong to one problem domain. The relationship between various entities making up knowledge is presented in Figure 4-50 below.

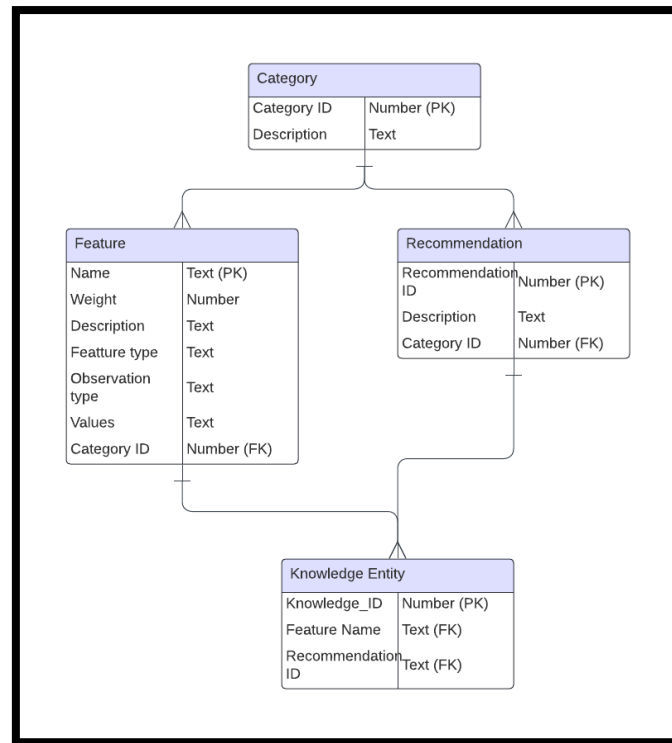


Figure 4- 50: An ERD showing the relationship between features and recommendations for a given problem domain (Source: Author)

From the diagram in Figure 4-50 above, the knowledge entity is a bridge entity that handles many-to-many relationship that exists between features and recommendations. Both features and recommendations are categorised into a certain problem domain.

A feature entity or object is a fundamental bit of information that makes up a question. It is made up of the following properties:

- **Feature name:** A unique name provided for a feature.
- **Weight:** This property will help in distinguishing and prioritising features that are very important over the ones that are less important for the given problem domain.
- **Description:** This field explains the feature and can also be used as a guideline to explain to farmers how the feature is evaluated such as soil texture, soil compaction and so forth.
- **Feature type:** Feature can either be numerical or categorical, this will help the model in deploying the right arithmetic operations for the feature.

- **Observation type:** This field tries to explain if the feature is made from human observation or the sensor.
- **Values:** This field presents possible values that make up the feature. This field is thoroughly discussed later in the chapter.
- **Problem Domain (Category):** This field will cluster the feature into a specific problem domain.

The recommendation entity is composed of an **identifier** field that uniquely identifies the recommendation from others, the **description** field that describes the recommendation, and the **problem domain** to which the recommendation falls. The problem domain entity has the **identifier** field and the **description** field.

4.2.6.1. Constructing a list of feature values in a presentable format to the farmer

The method of constructing a list of feature values varies depending on the type of observation for each feature. If the feature will be made from human observation, then its values can be separated by a comma delimiter as illustrated in Table 4-6 below for both numerical and categorical features.

Table 4- 6: Illustration of how feature values that will be made from human observation will be modelled (Source: Author).

Numerical Feature	Categorical Feature
Very low, low, moderate, high, very high	Yes, no

A list of sensor readings is constructed differently to enable the algorithm to structure the raw sensor data into the format that the farmer can be able to understand such as: very low, low, moderate, high and very high. This will be guided by the template the expert user will write for the given feature.

Table 4-7 below illustrates how the temperature and humidity readings from the sensor can be modelled by the expert user of the system.

Table 4- 7: Illustration of how feature values from sensor reading will be modelled (Source: Author).

Temperature	Humidity(%)
∞ - 10: very cold, 11 - 20: cold, 21 - 28: warm, 29 - 32: hot, 32 - ∞ : very hot	0 - 30: very low, 31 - 49: low, 50 - 65: medium, 66 - 80: high, 81 - 100: very high

The information provided in Table 4-7 above is an example of a template that will be generated by the expert user to guide the algorithms to calibrate (normalise) the raw sensor data into a format a farmer can relate to. For instance, if the temperature is -5 degrees Celsius, the model will categorise the temperature as very cold, and if the humidity level is 70%, the model will categorise the humidity as high.

The function presented below in Figure 4-51 is used to calibrate or normalise the sensor reading to become well presentable to the farmer.

```

Collection getCalibratedReading(Text fName, num rawValue)
  Declarations
    num min, max;
    ArrayList calibratedValue = new ArrayList()
    Text featureTemplate = getValues(fName)
    StringTokenizer token = new StringTokenizer(featureValues, ",")
    Collection data = new Collection(), valueRanges = new Collection(),
      valueCategories = new Collection(), mins = new Collection(), maxs = new Collection()
    while token.hasMoreTokens()
      data.add(token.nextToken())
    endwhile
    for Object d : data
      token = new StringTokenizer("'" + d, ":")
      valueRanges.add(token.nextToken().trim())
      valueCategories.add(token.nextToken().trim())
    endfor
    for Object d : valueRanges
      token = new StringTokenizer(d.toString(), "-")
      mins.add(token.nextToken().trim())
      maxs.add(token.nextToken().trim())
    endfor
    if mins.get(0) = "" then
      mins.set(0, "0");
    endif
    for int x = 0, x < mins.size(), x++
      try
        min = mins.get(x)
        max = maxs.get(x)
        if rawValue >= min AND rawValue <= max then
          calibratedValue.add(0, valueCategories.get(x))
        endif
      catch Exception ex
        calibratedValue.add(0, valueCategories.get(valueCategories.size() - 1))
      endtry
    endfor
    return calibratedValue
  
```

Figure 4-51: An algorithm that normalizes the raw sensor data. (Source: Author)

The algorithm presented above in Figure 4-51 receives the feature name, and the raw sensor value extracted from the real-time cloud database. The function aims to identify a suitable category that will explain the raw sensor data using the guidelines (template) set by the expert user of the system. The function will first extract the template for the given feature. This template is comprised of records separated by a comma delimiter where each record represents the sensor reading range and the category for that range. The first task of the algorithm will be to separate these records for easier manipulation using a comma (,) delimiter. For each record, the sensor reading range is decoupled from its category using the colon (:) delimiter. The sensor reading range is further partitioned into minimum and maximum values using a hyphen (-) delimiter. These minimum and maximum values will be evaluated to determine the suitable range for the given raw sensor data passed to the function. When the suitable range is found, the category for that range will be returned by the algorithm.

4.2.6.2. Background knowledge of the proposed optimisation model

The proposed optimisation model is a customised K-prototype algorithm. This clustering algorithm is among very few that are capable of handling both numerical and categorical data features. Similarly, to any clustering approach, it aims to find the optimal grouping of closely related data patterns based on some similarities or characteristics inherent in them. It uses the K-parameter (that represents the number of splits(clusters) an algorithm needs to perform) as a decision variable that will minimise the objective function which in this context is the mean of clusters' variance. The cluster's variance is determined by finding the squared distance between the members of the cluster and the centroid (centre point) of the cluster. Therefore, this model aims to evaluate different K-parameters and identify the one that can optimally group closely related data objects.

This model is a hybrid approach that integrates both the K-means clustering algorithm for numerical data and the K-mode clustering algorithm for categorical data. It is mathematically explained as:

$$d(X_i, C_j) = d_r(X_i, C_j) + d_c(X_i, C_j) \quad (\text{Equation 4-34})$$

where X_i represents the input feature at a given index, C_j represents a cluster's centroid at a given index, d_r represents numerical features and d_c represents categorical features.

The calculation of numerical features is shown below using K-means clustering where the distance matrix between two objects (data object (X_{ij}) and cluster's centroid object (C_{ij})) is

computed as the squared difference between numerical features of the data object and the centroid.

$$d_r(X_i, C_j) = \sum_{l=1}^p |X_{il} - C_{jl}|^2 \quad (\text{Equation 4-35})$$

The formula of categorical features is shown below using the K-Mode algorithm where the distance matrix between two objects (data object (X_{il}) and cluster's centroid object (C_{jl})) is calculated by comparing the categorical features of the data object and the centroid. The penalty is assigned when there is dissimilarity:

$$d_r(X_i, C_j) = \sum_{l=p+1}^m \delta(X_{il}, C_{jl}) \quad (\text{Equation 4-36})$$

$$\delta(X_{il}, C_{jl}) = \begin{cases} 0, & \text{when } X_{il} = C_{jl} \\ 1, & \text{when } X_{il} \neq C_{jl} \end{cases} \quad (\text{Equation 4-37})$$

The K-Prototype clustering algorithm initializes the centroids to any random data sample. It then loops through all the data records and classifies each by comparing centroids' features with the ones of the tested data record. A data record is classified to a cluster with the shortest distance from the data record to the cluster's centroid. After all the data records have been clustered, the centroids will be updated by computing the mode of values in every categorical feature and the mean of values in every numerical feature. The centroids will be initialised to the categorical values with high mode and the mean of the numerical values. The classification process will be made again using the updated centroids' position. The process will be repeated until the centroids' position is fixed. Hence, optimum clustering will be reached.

4.2.6.3.A deeper look into how the optimisation model was developed

The proposed optimisation model is a modified K-prototype algorithm that is flexible enough to work with varying lengths of features making up different questions. The model learns by building clusters of the homogenous set of data objects. This model was built using two main functions *fit()* and *predict()*. The *fit()* function is used to train the model by evaluating different K-parameters and identifying the one that can optimally group the data objects based on their characteristics. A *predict()* function is used to extract relevant recommendations based on the input provided. A data object or a record is a piece of information composed of a question (a set of features) and an answer (one or more recommendations) from a given problem domain.

Training the Optimisation Model

The *fit()* function of the optimisation model is presented below in Figure 4-52.

```
Collection fit(Collection records) {
    Declarations
    conn=new DbsConnection();
    Collection optimalCentroids=new Collection(),errors = new Collection(),
        k_states = new Collection(),difs=new Collection(),
        identities=conn.getFeatureRecommendationIds(records)
    num error,k = 2,kmax, iteration = 1,changelnCentroidsMovement=0
    num[] clusterNumbers
    compute=new Computation()
    boolean change = true
    kmax = identities.size() / threshold
    while k < kmax
        do
            recs = compute.KProtoLearning(k, records, iteration);
            clusterNumbers=conn.getClusterNumbers(recs,records.size());
            k_states.add(clusterNumbers);
            if iteration>1 then
                changelnCentroidsMovement= compute.comparison(k_states.get(iteration - 2), k_states.get(iteration-1))
                difs.add(changelnCentroidsMovement)
            endif
            iteration++
        while changelnCentroidsMovement!=100.00 AND iteration<learningThreshold
        endwhile
        optimumCentroids.add(compute.getClusterentroids())
        error = compute.getClusterErrors(k)
        errors.add(error)
        iteration = 1
        k_states=new ArrayList()
        difs=new ArrayList()
        changelnCentroidsMovement=0
        k+=1
    endwhile
    bestK = compute.getMinimumPosition(errors) + 1
    optimalCentro= optimumCentroids.get(bestK-1)
    return optimalCentro
}
```

Figure 4-52: A *fit()* function of an optimisation model. (Source: Author)

The function presented above in Figure 4-52, receives the records (questions and answers) already filtered based on the selected problem domain from the knowledge database. The K-value (parameter) is initially set to 2, to start with two clusters and it will increment towards the maximum number of clusters (K-values) determined by the algorithm based on the sample size. The higher the sample size the higher the number of K-values to be evaluated. However, the selection of the number of K-values is also made considering the avoidance of overfitting and too much computation time. Every clustering performed by the algorithm is being optimised. This is done by adjusting the movement of the centroids to a level where the

centroids are fully representing the members of the clusters. This is a climax (convergence) stage where records are no longer moving from one cluster to another as the learning process continues. For every learning process, the records' movement from cluster to cluster is evaluated, and for as long as there is a change in the movement, the learning process will continue. However, the learning process is also set to a certain limit to avoid infinite learning. The frequency of records' movement between clusters decreases as the learning process progresses (as the centroids adapt to the sample). The movement of records from cluster to cluster is due to the movement of the centroids as they continue searching for an optimal spot. The learning process will proceed to a level where centroids are no longer moving. At that stage, the learning process will stop, and the clusters' variance will be computed. It is computed as the average distance between the centroids and clusters' members. There is also a penalty associated with the number of clusters the model needs to generate. The higher the number of clusters the higher the penalty to minimise overfitting. The learning process will take place for every K-parameter, and clusters' variance will also be computed for every K-parameter. Finally, the K-parameter that has resulted in an optimal split based on minimal clusters' variance, will be used to predict the recommendations for the input or question provided.

The function that performs the actual learning of the data is presented below in Figure 4-53. This function is invoked by the *fit()* function presented above in Figure 4-52 passing the K-parameter, records, and the number that tracks its iteration which is initially set to 1. This number aims to monitor the learning process by the model, and it is constrained to a certain limit to avoid infinite learning.

```

Collection KProtoLearning(num k, Collection historic_records, num iteration)
Declarations
Collection cluster = new Collection(), randomClusters = new Collection(),
cluster_errors = new Collection(), records = new Collection(), record = new Collection();
records = getOrganisedRecords(historic_records)
boolean exceptionOccurred = false
num counter = 1, distance = 0, minPosition, randomIteration, pos = -1, penalty = 0
Random random = new Random()
if iteration == 1 then
randomIteration = records.size() / iterationThreshold
while randomIteration >= 0
pos += 1
positions.add(pos)
exceptionOccurred = false
for num c = pos, c <= records.size() AND exceptionOccurred = false, c++
try
record = records.get(c)
if counter <= k then
for Object r : record
((Knowledge) r).setClusterNumber(counter)
Cluster cls = new Cluster(((Knowledge) r).getFeatureName(), ((Knowledge) r).getFeatureValues(),
((Knowledge) r).getRecommendations(), ((Knowledge) r).getCategoryName(),
((Knowledge) r).getFeatureRecId(), ((Knowledge) r).getClusterNumber())
cluster.add(cls)
endfor
centroids.add(cluster)
cluster = new ArrayList()
else
for Object cr : centroids
penalty += Math.pow(Math.abs(((Collection)cr).size() - record.size()), sizeThreshold)
for Object cc : (Collection)cr
penalty += ((Cluster) cc).computePenalty2(record)
endfor
cluster_errors.add(penalty)
penalty = 0
endfor
minPosition = getMinimumPosition(cluster_errors)
cluster_errors = new Collection()
for Object r : record
((Knowledge) r).setClusterNumber(minPosition + 1)
endfor
endif
counter++
catch Exception ex
exceptionOccurred = true
for c = 0, c < pos, c++
record = records.get(c)
for Object cr : centroids
penalty += Math.pow(Math.abs(((Collection)cr).size() - record.size()), sizeThreshold)
for (Object cc : (Collection) cr)
penalty += ((Cluster) cc).computePenalty2(record)
endfor
cluster_errors.add(penalty)
penalty = 0
endfor
minPosition = getMinimumPosition(cluster_errors)
cluster_errors = new ArrayList()
for Object r : record
((Knowledge) r).setClusterNumber(minPosition + 1)
endfor
endfor
endtry
endfor
ArrayList hold = findDistributionOrRecords(records, k)
randomClusters.add(hold)
hold = new ArrayList()
centroids = new ArrayList()
counter = 1;
randomIteration--
endwhile
bestSplit = getOptimumSplit(randomClusters, records)
records = fitWithBestSplit(records, bestSplit, k)
centroids = new ArrayList()
optimumCentroids = updateCentroids(records, k)
positions = new ArrayList()
else
Iterator iterator = records.iterator()
while iterator.hasNext()
record = (Collection) iterator.next()
for Object cr : optimumCentroids
penalty += Math.pow(Math.abs(((Collection)cr).size() - record.size()), sizeThreshold)
for (Object c : (ArrayList) cr)
penalty += ((Cluster) c).computePenalty2(record)
endfor
cluster_errors.add(penalty);
penalty = 0;
endfor
minPosition = getMinimumPosition(cluster_errors)
cluster_errors = new ArrayList()
for Object r : record
((Knowledge) r).setClusterNumber(minPosition + 1)
endfor
endwhile
optimumCentroids = updateCentroids(records, k)
endif
return records

```

Figure 4- 53: A function that performs clustering of the records. (Source: Author)

The *learning()* function in Figure 4-53 above, initially groups every set of features by their question numbers to simplify their manipulation. In the first iteration of its learning process, the algorithm evaluates different data records to determine the best ones that can represent the optimal starting position of the centroids. Although this approach may become time-consuming for the algorithm, it minimises the local optima problem that many clustering and other optimisation models suffer. In clustering, this is a type of problem where different starting positions of the centroids result in different formations of clusters despite the adjustments performed by the centroids. Consequently, this results in a failure of the clustering algorithm to determine the optimal grouping (Vora et al.,2013).

To evaluate different centroids' starting positions, the algorithm selects data objects that will be assigned to the centroids at their initial stage of learning. The number of data objects selected by the model is based on the data sample's size. The higher the sample size the more the data objects that will be selected. However, the selection is also made with the mind of finding a balance between the accuracy and the machine's response time. This is because the higher the number of the selected data objects the higher the chances of improved accuracy, but the lower the response time due to massive operations that the machine will undergo. Figure 4-54 below demonstrates how the algorithm will go about selecting different starting data objects for a given K-value.

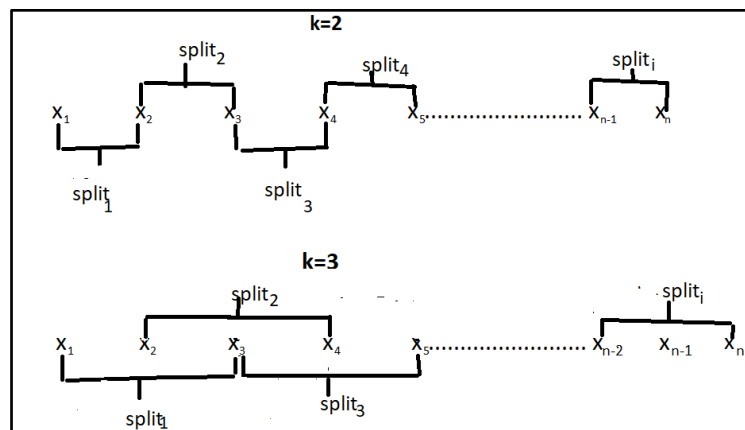


Figure 4-54: Illustration of centroids' starting positions based on K-parameter (Source: Author).

From Figure 4-54 presented above, in the first iteration of the learning process, the centroids will be attached to the starting data objects in the sample, and they will continue to shift with an offset of 1 position as the learning process continues. After the learning has been performed using different centroids' starting positions, the generated groupings (splits) of data objects will

be evaluated to determine the one that has resulted in the normal distribution of centroids across different clusters. This is a type of distribution where the difference in the number of members in one cluster is minimal when compared to other cluster(s). Given that the distribution of the data objects in the database is not known, the algorithm assumes that the initial split that can well separate the data objects is the one that can help in getting closer to the global optimal split.

The function presented in Figure 4-55 below is invoked by the *learning()* function presented above in Figure 4-53. It receives as a parameter, a collection of partitions (splits) made for different centroids' starting positions together with the data sample size.

```

num getOptimumSplitPosition(Collection partitions, Collection records)
  Declarations
    num min,max, difference = 0,total=0,minDifference,bestSplitPosition
    ArrayList differences=new ArrayList()
  for Object partition : partitions
    min = getMinimum((Collection) partition)
    max = getMax((Collection) partition)
    for int x = 0, x < ((Collection) partition).size(), x++
      total += ((Collection) obj).partition(x) / min * max
    endfor
    differences.add(total / records.size())
    total = 0
  endfor
  minDifference = getMinimumPosition(differences)
  bestSplitPosition = positions.get(minDifference)
  return bestSplitPosition

```

Figure 4- 55: An algorithm that will return the optimal starting position of the centroids.
(Source: Author)

The algorithm presented above in Figure 4-55 will make some mathematical operations to find the optimal partition of the data objects as illustrated with the formula below.

$$total = \frac{\sum_{i=1}^n \frac{x_i}{min_i} max_i}{j} \quad (\text{Equation 4-38})$$

where n represents the number of partitions

min_i represents a cluster for a given partition that has a minimum number of data objects.

max_i represents a cluster for a given partition that has a maximum number of data objects.

x_i represents the number of data objects in the cluster for a given partition.

j represents the data sample size (number of data records in all clusters)

The algorithm above is designed in such a way that a partition that has an equal number of members in every cluster, will result in 1. An increase from 1, will become an increase in variability.

The learning process discussed earlier was performed by the model to identify the best starting position of the centroids. The actual learning process will now take place using the optimal centroids' starting position (that has signified optimal separation of the data). The data records are grouped based on the minimal distance between them and the centroids. After the grouping of data objects to their respective clusters, the centroids' position is updated from the generated clusters. The method that updates the centroids' position is presented in Figure 4-56 below.

```

Collection updateCentroids(Collection records, int k)
    Declarations
    num count=0,size,highest,pos
    Text featureName
    Collection recordsPerCluster = new Collection(),averageOfFeaturesPerCluster = new Collection(),
    distinctFeatureNames=new Collection(),frequencyOfFeatureNames=new Collection(),recs=new Collection,
    centroids = new Collection(), optimumFeatureNamesPerCluster = new Collection(), optimumFeatureNames = new
    Collection(),list = new Collection(), listForK = new Collection()

    for num c = 1, c <= k, c++
        recs = getRecordsInCluster(c, records)
        recordsPerCluster.add(recs)
        averageOfFeaturesPerCluster.add(getCountOfObjectsPerCluster(recs))
    endfor
    distinctFeatureNames = getDistinctFeatureNames(recordsPerCluster)
    frequencyOfFeatureNames = getFrequencyOfFeatureNames(distinctFeatureNames, recordsPerCluster)
    for num x = 0; x < k, x++
        for Object obj : frequencyOfFeatureNames
            list.add(((Collection) obj).get(x));
        endfor
        listForK.add(list)
        list = new ArrayList()
    endfor
    for Object obj : listForK
        size = averageOfFeaturesPerCluster.get(count)
        if size>distinctFeatureNames.size() then
            size=distinctFeatureNames.size()
        endif
        for num x = 0; x < size, x++
            highest = getMax(((Collection) obj)
            pos = getMaxPosition(highest, (Collection) obj)
            optimumFeatureName = ""+distinctFeatureNames.get(pos)
            optimumFeatureNamesPerCluster.add(optimumFeatureName)
            ((Collection) obj).set(pos, -1)
        endfor
        count += 1
        optimumFeatureNames.add(optimumFeatureNamesPerCluster)
        optimumFeatureNamesPerCluster = new Collection()
    endfor
    centroids = UpdateValuesOfOptimalFeatures(recordsPerCluster, optimumFeatureNames);
    return centroids

```

Figure 4- 56 An algorithm that updates centroids' position. (Source: Author)

The algorithm presented in Figure 4-56 above receives the clustered records and the K-parameter. The algorithm will initially group the data records according to the clusters they belong for easier manipulation. Given that each data record is made up of varying sizes of features as illustrated in Figure 4-57 below:

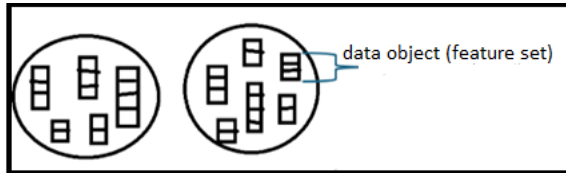


Figure 4- 57: Demonstration of features making up different questions (Source: Author).

The average length of features in every cluster will be computed to represent the length (number of features) each centroid will have. The frequency of distinct features in every cluster will be calculated and the ones that have high frequency will be identified and will be slotted into the body of the given centroid. The algorithm in Figure 4-58 below is invoked to provide the optimal values for the optimal features that will be used to build the centroids.

```
Collection UpdateValuesOfOptimalFeatures (Collection recordsPerCluster,Collection optimumFeatureNames)
Declarations
num count = 0
Text fName,bestValue
Cluster cluster
Collection centroid = new Collection()
for Object obj : optimumFeatureNames
for Object obj2 : (Collection) obj)
fName = ""+obj2
bestValue = getBestValue(fName, (ArrayList) recordsPerCluster.get(count))
cluster= new Cluster()
cluster.setFeatureName(fName)
cluster.setFeatureValue(bestValue)
cluster.setClusterN(count + 1)
centroid.add(cluster)
endfor
centroids.add(centroid)
centroid = new Collection()
count++
endfor
return centroids
```

Figure 4- 58: An algorithm that updates feature values of centroids . (Source: Author)

The algorithm in Figure 4-58 above, receives as a parameter the grouped data records per cluster from the database and the selected features per cluster. The algorithm will use this

information provided to it to build the complete centroids made up of optimal feature names and feature values. This algorithm invokes another function presented below in Figure 4-59 to come up with an optimal value for a given feature.

```
Text getBestValue(Text fName, Collection recordsPerCluster)
Declarations
Text bestValue="", type = conn.getFeatureType(fName)
Collection values, distinctValues, frequencyList
num position, mean
if type="categorical" then
values = conn.getCatValues(fName, recordsPerCluster)
if values.size()>0 then
distinctValues = conn.getDistinctValues(values)
frequency = getFrequencyOfValues(distinctValues, values)
position = findMaxPos(frequency)
bestValue = "" + distinctValues.get(position)
endif
else
values = conn.getNumericalValues(fName, recordsPerCluster)
mean = findMean(values)
bestValue = "" + mean
endif
return bestValue
```

Figure 4-59: An algorithm that returns an optimal value for a given feature. (Source: Author)

The function in Figure 4-59 above receives the feature name and the data records in each cluster. It will loop through the data records to get the values of a given feature. If the feature type is numerical, the mean of the values will be calculated. The mean will be the optimal value that represents the numerical feature. If the values are categorical, the frequency of distinct values will be calculated and the value with high frequency will be used as an optimal value for the given categorical feature.

In the next iteration of the learning process, the new position of the centroids will be evaluated to determine if they truly reflect only members in the clusters that they represent. If there is a disparity, the centroids' position will be updated using newly formed clusters. This process will continue to a level where centroids' position is optimally adjusted such that the members from a given cluster are homogenous to the cluster they belong and heterogenous to the members in other clusters.

A deeper look into the learning process by an algorithm

The learning process is initiated by running through the data objects and computing the distance between each data object and the centroids. The data object will be classified to the centroid with minimal distance. In this case, given that data objects (questions) are made up of one or many features, the initial process of computing the distance between the data object and the centroid is to first measure the feature set's length in the data object and the one in the centroid. The penalty will be assigned based on the magnitude of the difference. The distance between the centroid and the data object is further computed using the centroid's function presented below in Figure 4-60.

```
num computePenalty(Collection<Knowledge> record)
  Declarations
    boolean flag=false
    Text type
    num penalty=0
  for Knowledge knowledge:record
    if knowledge.getFeatureName().equalsIgnoreCase(featureName) then
      type=conn.getFeatureType(knowledge.getFeatureName())
      flag=true
      if type ="numerical" then
        penalty+=computeNumericalDistance(featureValue,knowledge.getFeatureValue(),knowledge.getFeatureName())
      else
        penalty+=computeCategoricalDistance(featureValue,knowledge.getFeatureValue(),knowledge.getFeatureName())
      endif
    endif
  endfor
  if NOT flag then
    penalty+=conn.getFeatureWeight(featureName)*threshold
  endif
  return penalty
```

Figure 4- 60: An algorithm that computes the distance between a data object and a centroid (Source: Author).

The function in Figure 4-60 above receives the data object to evaluate as a parameter. This function will loop through the features that make up the data object. It will compute the distance between the centroid's features and the ones held by the data object. Before computing the distance, it will first check if the data object's feature at a given index exists inside the centroid. If it does not exist, the magnitude of a penalty assigned to the centroid will be high. If the feature exists, the type of the feature will be evaluated, if it is numerical, the function responsible for computing the numerical distance between the feature values will be invoked, otherwise, the function responsible for computing the categorical distance will be invoked.

The functions that compute the numerical and categorical distances are presented below in Figures 4-61 and 4-62 respectively.

```
num computeCategoricalDistance(Text fValueData, Text fValueCluster, Text  
featureName)  
  Declarations  
    num distance;  
  if fValueData = fValueCluster then  
    distance=conn.getFeatureWeight(featureName)*-1;  
  else  
    distance=conn.getFeatureWeight(featureName);  
  endif  
  return distance
```

Figure 4- 61: An algorithm that computes the distance for categorical features. (Source: Author)

```
num computeNumericalDistance(num fValueData, num fValueCluster, Text featureName)  
  Declarations  
    num distance=fValueData-fValueCluster  
  if distance = 0 then  
    distance+=conn.getFeatureWeight(featureName)*-1  
  else  
    distance+=distance*conn.getFeatureWeight(featureName)  
  endif  
  return distance
```

Figure 4- 62: An algorithm that computes the distance for numerical features. (Source: Author)

The function presented in Figure 4-61 above, responsible for calculating the categorical distance, receives the feature name, the centroid's feature value, and the data object's feature value. It will compare the values, if they are similar, a point (+1) will be assigned to the centroid which will reduce the penalty a centroid has. Otherwise, the penalty (-1) will be assigned. The function presented above in Figure 4-62, responsible for computing the numerical distance between the two feature values, receives the same information as the function that computes the categorical distance. However, instead of comparing the values, it will compute the squared distance between two values. If the distance is 0, then a point will be assigned to a centroid. Otherwise, the penalty will be assigned based on the magnitude of the distance between the two values.

The assignment of penalties and points is largely influenced by the feature weights. The high the feature weight the high the penalty or point assigned to a centroid and visa-versa.

When the learning process for the given K-value is complete, the clusters' variance will be computed using the following algorithm in Figure 4-63.

```

num getClusterErrors(int k, Collection centroids, Collection recordsPerCluster)
Declarations
    num average = 0, count = 0, penalty = 0
    Collection centroid
for Object obj : recordsPerCluster
    for Object record : (ArrayList) obj)
        centroid = (Collection) centroids.get(count)
        penalty += Math.pow(Math.abs(((Collection) centroid).size() - record.size()), threshold)
        for Object c : centroid
            penalty += ((Cluster) c).computePenalty((ArrayList<Knowledge>)record)
        endfor
    endfor
    count++
    average += penalty / ((ArrayList) obj).size()
    penalty = 0
endfor
return average + 2*k

```

Figure 4- 63: An algorithm that computes the clusters' variance. (Source: Author)

The function in Figure 4-63 above receives the K-value, grouped records per cluster, and the optimum centroids. The function will compute the average absolute distance between the members and the centroid of a given cluster. Similarly, to the AIC model, the penalty will be assigned based on the number of clusters the model needs to generate (K-value parameter) to minimise overfitting. The higher the number of clusters formed by the model, the higher the penalty and visa-versa.

All the processes discussed above concerning the fitting or training of the K-Prototype algorithm will be repeated for every evaluated K-parameter. Finally, the clusters' variance for every K-parameter will be assessed to identify the parameter that has optimally clustered the data records. The centroids generated using the optimal K-parameter will be saved in the cloud database and used to deliver relevant recommendations for the given question within a given problem domain.

Recommendations' prediction by the model

The function that performs prediction is presented below in Figure 4-64:

```
Collection predict(Collection inputRecord, Collection optimumCentroids, Collection<Knowledge> historic_records)
Declarations
    num counter = 2, iteration = 1
    Collection optimumRecords, recs = compute.getOptimumClusters(optimumCentroids, historic_records),
    bestCluster = compute.getBestCluster(inputRecord, optimumCentroids)
    optimumRecords = compute.findRecords(bestCluster, recs),
    recommendations = compute.getAllRecommendations(optimumRecords),
    distinctRecommendations = compute.getDistinctRecommendations(recommendations)
    frequencyOfRecommendations = compute.getFrequency(distinctRecommendations, recommendations)
    descriptionList = compute.getDescriptions(distinctRecommendations)
return descriptionList
```

*Figure 4- 64: An algorithm that retrieves the recommendations based on the given input.
(Source: Author)*

The function presented above in Figure 4-64 receives the question (input) from the user, the optimal centroids, and the data records; all from the problem domain selected by the user. The function will initially partition the records using the optimal centroids. The algorithm will further get the optimal cluster by computing the distance between the input (question) and the centroids as illustrated below in Figure 4-65:

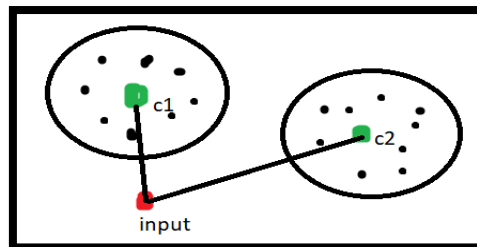


Figure 4- 65: An illustration of how the optimal cluster will be identified (Source: Author).

where $c1$ is the centre point for the first cluster, and $c2$ is the centre point for the second cluster. The input will be classified to the cluster with the minimal distance from its centre point to the input. The algorithm will loop through the members in the selected (optimal) cluster and extract the recommendation identities. The recommendation identities extracted from the optimal cluster are further manipulated to identify the ones that are highly supported by the members in a cluster. This is achieved by identifying distinct recommendation identities in a cluster and computing the frequency of each. The recommendation descriptions will be extracted and will

be displayed in descending order based on their frequency percentage. Meaning the recommendations that are highly supported will be on top of the list. In this way, it will become easier for farmers to draw attention to recommendations that have more weight than those with less weight.

4.2.7. System output

The algorithms explained in this chapter were fused into the mobile-based application that farmers will use to optimise their cropping decisions. A few mobile application screenshots are displayed. The first menu screen presented below in Figure 4-66 gives a user an option to register the IK indicator that they have observed, make season behaviour predictions or monitor the health and growth of their crops.

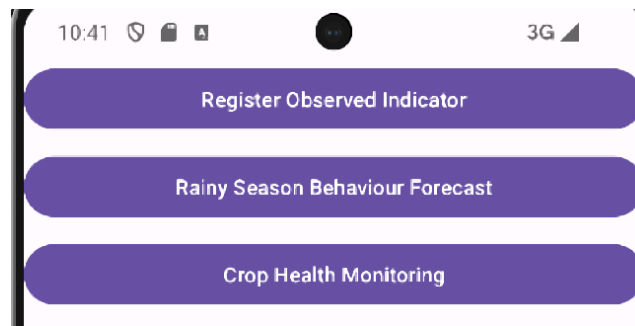


Figure 4- 66: Menu screen for the user (Source: Author).

The menu screen for an administrator will display additional options such as management of IK indicators and their observations, management of thresholds and so forth, where the user can create, read, update and delete the information. The menu screen of an expert user (considered as the one with more rights than the administrator) will include tasks that can be done by an ordinary user and administrator plus additional roles such as the management of features and recommendations, and their mapping to create new knowledge.

The screenshot presented below in Figure 4-67 is displayed when the observation of the IK indicator is registered into the database.

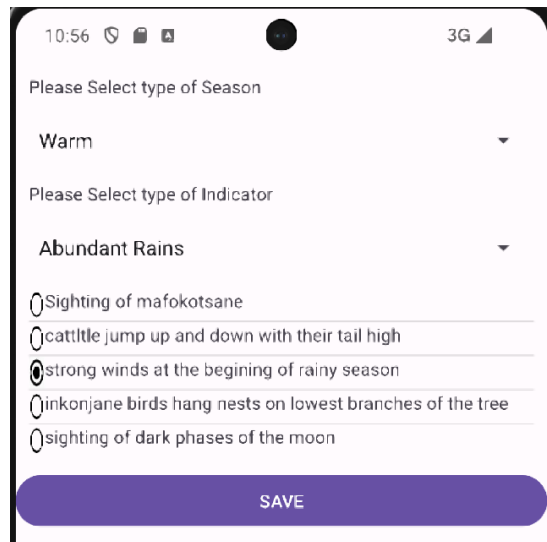


Figure 4- 67: Screenshot of how the observation of the indicator will be registered (Source: Author).

From Figure 4-67 above, the user is provided with an interface to filter the IK indicators based on the season type of their occurrence (warm or cold or any season indicators) and the rainy season behaviour type that they predict such as season changes, drought, floods or abundant rains. The list of filtered IK indicators will be displayed to the user. The user will further select the indicator they have observed, and save the observation to the cloud database. The observation of the IK indicator will be evaluated before it is registered. When registering an environmental indicator, as demonstrated in Figure 4-68 below, a camera option will be available to enable a user to capture the indicator as proof of their observation.



Figure 4- 68: Screenshot of photo capturing for environmental indicators (Source: Author).

The screenshot presented in Figure 4-69 below is an illustration of season behaviour forecasts that will be displayed to the user.

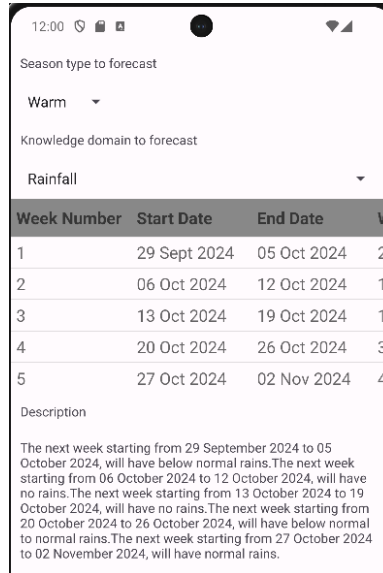


Figure 4- 69: Screenshot of the system’s generated forecast for rainfall (Source: Author)

From the screenshot presented in Figure 4-69 above, the user is provided with an option to select the type of the season (warm or cold rainy season) they would like to predict, and the knowledge domain (rainfall (water), temperature, vegetation health, and all domains) that they would like to observe their behaviour throughout the season. The forecasts are presented in a tabular format and text description explaining each observation. The line graph showing the time series movement of all the knowledge domains will be displayed when the all domains option is selected.

4.2.8. Optimisation Model Evaluation

The proposed optimisation model was trained and evaluated to come up with crops a farmer should grow based on the possible behaviour of rainfall and soil characteristics inputted into the system. This evaluation was made possible through the integration of crop science knowledge and insights obtained from IK experts during primary data collection. The model was trained and evaluated with the concern that farmers cultivate immediately after the first rains.

Several features and recommendations were selected as demonstrated in Table 4-8 below. The recommendations, in this case, are the crops a farmer will cultivate based on the question (set of features) provided to the system.

Table 4- 8: Set of features and recommendations used to train the model (Source: Author).

features for Rainy Season behaviour	Soil features	List of crops
Rains onset > early Aug, middle Aug, late Aug, early Sept, middle Sept, late Sept, early Oct, middle Oct, late Oct, early Nov, middle Nov, late Nov, early Dec, middle Dec, late Dec. Rains cessation > early Aug, middle Aug, late Aug, early Sept, middle Sept, late Sept, early Oct, middle Oct, late Oct, early Nov, middle Nov, late Nov, early Dec, middle Dec, late Dec. Rains distribution > one week apart, two weeks apart, three weeks apart, four weeks apart, five weeks apart Rainfall type(magnitude) >very scarce, scarce, normal, normal to excessive, excessive	Soil texture > sandy, loam to sandy, loam, loam to clay, clay Soil colour > red, brown, black, dark grey, light grey Soil salinity > very low, low, moderate, high, very high Soil compactness > very low, low, moderate, high, very high Ph-scale > very acidic, acidic to neutral, neutral, alkaline to neutral, very alkaline Nitrogen level > very low, low, moderate, high, very high Potassium level > very low, low, moderate, high, very high Magnesium level > very low, low, moderate, high, very high	Sugar beans, amadumbe ,butternut, watermelon, cowpea, carrots, sugarcane, maize, radishes, kale, spinach, potatoes, soybeans, peanuts

Before the model could be trained, features and recommendations were created under the crop selection category. The screenshot presented in Figure 4-70 below illustrates how features were created.

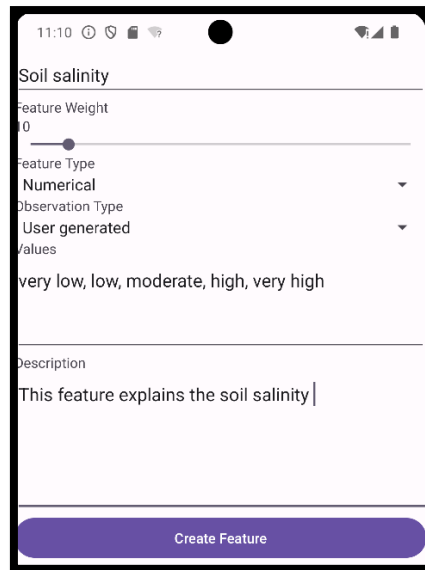


Figure 4- 70: Screenshot of how features were created (Source: Author).

The screenshots presented in Figures 4-71 and 4-72 below depict a list of features and crop names that were created for model evaluation.

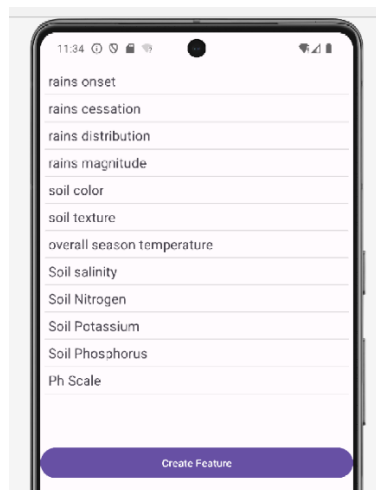


Figure 4- 71: Screenshot of all features created (Source: Author).

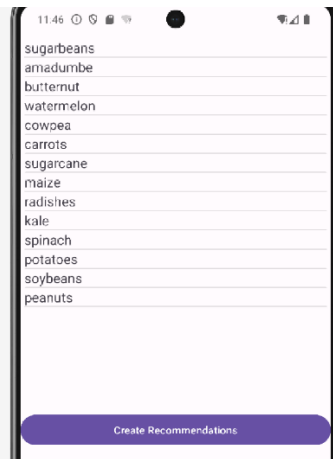


Figure 4- 72: Screenshot of all recommendations created (Source: Author).

The weight of the features was the same for all, to avoid bias in predictions.

From the features above in Figure 4-71, many questions were created. The questions were structured from more general to more narrowed questions. The mobile's screenshot presented below in Figure 4-73 shows how the questions were created and mapped to their recommendations:

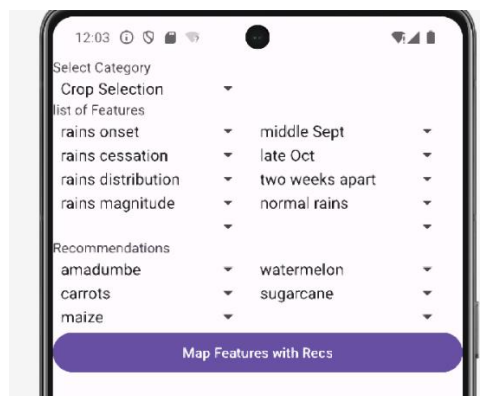
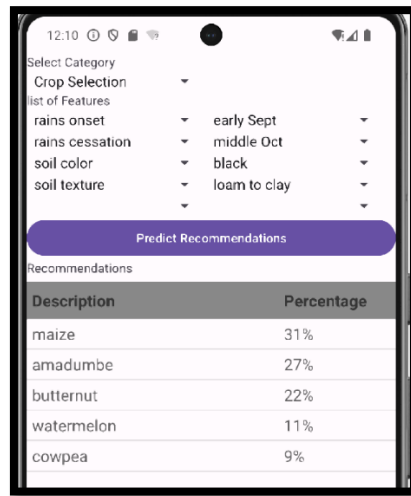


Figure 4- 73: Screenshot of how the features and recommendations are mapped (Source: Author).

The screenshot in Figure 4-73 above tries to address the following question:

- What crops can I grow when abundant (normal) rains are expected to start in the middle of September and end in late October, given that their distributions are roughly two weeks apart?

Over 200 sets of questions and answers were created to train the optimisation model. Only 20 new records were used to validate the accuracy of the model. For every question the model answers, the approach presented in section 4.2.8.1 below was used to evaluate the accuracy of the predictions against what was required by the researcher. The model's predictions or recommendations (crops to grow) are associated with percentages (each explains the frequency of each recommendation inside the optimal cluster) that were treated as a weight towards the certainty of the model (how sure the model is for the given recommendation). This is illustrated in Figure 4-74 below.



The screenshot shows a mobile application interface. At the top, there is a status bar with the time 12:10 and various icons. Below the status bar, there is a 'Select Category' dropdown menu. Underneath, there is a 'Crop Selection' dropdown menu. Below that, there is a 'list of Features' section with five rows, each containing a feature name and a dropdown menu: 'rains onset' (early Sept), 'rains cessation' (middle Oct), 'soil color' (black), and 'soil texture' (loam to clay). Below the features section, there is a purple button labeled 'Predict Recommendations'. Below the button, there is a section titled 'Recommendations' which contains a table with two columns: 'Description' and 'Percentage'.

Description	Percentage
maize	31%
amadumbe	27%
butternut	22%
watermelon	11%
cowpea	9%

Figure 4- 74: Screenshot of the model's prediction of crops to grow based on features provided (Source: Author).

Figure 4-74 above illustrates the recommendations provided by the model for the following question:

- What crops can I grow when rains are expected to start in early September and end in the middle of October, provided that my soil colour is black with a loam-to-clay texture?

4.2.8.1. An approach followed to compute the accuracy of the recommendation(s) for a given question

Initially, the length of the recommendations generated by the model is compared to the length of the recommendations generated by the researcher. The penalty (-1) is assigned to the model based on the magnitude of the difference. Further, the model's predictions are compared to the expected ones to find the match. If the expected recommendation (crop) is listed by the model,

a point (+1) will be assigned to the model, however, the weight of the point will be determined by the certainty of the model. The higher the model's certainty the higher the weight of the point. If the expected recommendation is not listed by the model, the penalty (-1) is assigned.

The approach is structured such that, if all the recommendations generated by the model are the same as the expected recommendations, the score of 1 (100%) will be the final accuracy measure, given that the weights of the machine's recommendations are summed to 100%. The dissimilarity will cause deviation from +1 towards negative values.

The algorithm presented below in Figure 4-75 was used to automate the evaluation process. It receives the list of the expected recommendations, the list of the model's recommendations, and the list of the weights for the model's recommendations. It will loop through the lists and perform all the necessary computations explained earlier. It will return the model's accuracy for a given question.

```

num accuracyMeasure(Text[] answers, Text[] predictions, num[] weights)
  Declarations
    num accuracy, weight=0
  accuracy=Math.abs(answers.length - predictions.length)*-1
  for num i =0, i<answers.length, i++
    for num d=0, d<predictions.length, d++
      if answers[i]=predictions[d] then
        weight=weights[d]
        break
      endif
    endfor
    if weight=0 then
      accuracy+=-1
    else
      accuracy+=(1*weight/100)
    endif
  endfor
  return accuracy

```

Figure 4- 75: An algorithm that computes the model's accuracy for the question provided.
(Source: Author)

The overall accuracy was computed as the sum of accuracy measures divided by 20 which is the number of questions used to evaluate the model. The 83.3% average accuracy was obtained. The accuracy was deemed to improve with an increased number of examples provided to the model.

CHAPTER 5:DISCUSSIONS

5.1. Introduction

This chapter explores the main reasons that inspired the development of this research project. It also summarises the methodologies employed to tackle the research problem, discusses the opportunities and limitations of the proposed developments, and outlines potential directions for future research.

5.1.1. Motivation behind the developed tool

Climate change impacts continue to cripple the rain-fed agricultural systems that make up 80% of the world's agricultural land area, and that also generate 58% of the world's staple food (Reddy, 2015). SSA also known as the Horn of Africa is highly vulnerable due to overdependence on rainfed agriculture (given that over 90% of its farmed land is rainfed (Nuamah and Botchway, 2024; Lamptey, 2022)) and low adaptive capacities by small-holder farmers who constitute the majority of the labour force in the agriculture sector (Peter and Botchway, 2019). Limited access to reliable and accurate climate information is another limiting factor that emanates from the poor technological and financial background experienced by many undernourished countries in SSA including South Africa.

Many small-scale farmers continue to use the IK system as a shield to tackle climate change impacts and support their cropping decisions. This knowledge has been acquired through the strong ties with the dwellers and their environment and passed on through generations. Even though the IK system is a pillar that supports many local farmers' cropping decisions, the value and strength of this knowledge system are being washed out by climate stress, pollution, modernisation including industrialisation and many more. The policies and technological developments in South Africa to support the small-holder farmers' adaptation capacity, seem to have minimal impact due to the financial unfitness of the country and lack of participation by the local communities during their development life cycle. Additionally, data analysis revealed that agriculture is primarily practised by elderly farmers, with fewer young people involved. Many young individuals, especially men from rural areas, continue to migrate to urban areas in search of jobs. Consequently, much of the agricultural knowledge held by their elders is not optimally passed down or utilised by the younger generation. When these elderly farmers pass away, the knowledge they've accumulated over centuries is often lost, without being systematically preserved or utilised.

All these problems have sparked an interest in the researcher to come up with a technology that will integrate the farmers' knowledge with scientific data and models to optimise farmers' cropping decisions. These cropping decisions are considered critical because of their ability to modify farm productivity and profitability (Dury et al., 2012). Optimisation on the other side, is concerned with full exploitation of the limited resources by the farmer to maximise yield.

The farmers' cropping decisions that were addressed in this thesis include the agricultural season behaviour predictions and monitoring of the health and growth of the crops. The rationale behind the selection of these cropping decisions was due to their utmost importance in rain-fed agricultures led by under-resourced farmers. In addition, knowledge from both indigenous and scientific approaches has already been acquired on both season behaviour predictions and crop health monitoring. So, this research project focused on building a bridge that will enable the interaction and integration of these knowledge domains to optimise the farmers' cropping decisions, and thus, increases their resilience to climate stress.

5.1.2. Research methodology followed to address the research aim

The action-based (applied) research methodology was deployed to identify the problem with the current IK system towards farmers' cropping decisions and come up with a solution to mitigate or minimise the effects of the problem. The targeted population was small-scale farmers in the uMgungundlovu district who primarily do crop farming for either subsistence or market purposes, and rely heavily on rainfed agricultures and IK to leverage their cropping decisions. The population was further refined to focus exclusively on small-scale farmers in the Swayimane region of the uMgungundlovu district. The mixed data collection methods were utilised to extract sufficient knowledge about the IK in support of farmers' cropping decisions. Initially, the quantitative data collection method was used to measure the effectiveness of the IK in season behaviour predictions and monitoring the health and growth of the crops. The knowledge gathered provided the researcher with insights into the strengths and weaknesses of the IK system. Subsequently, the qualitative data collection methods were employed to gain a deeper understanding of how the IK is used by farmers to make seasonal behaviour predictions and monitor the health and growth of their crops, to complement the IK system with scientific knowledge.

The knowledge acquired from farmers and the secondary data acquired from scientific sources helped the researcher in addressing the main aim of this research project which was to integrate

the IK with satellite imagery, climate data, and machine learning models to optimise farmers' cropping decisions. The research aim was guided by a set of objectives.

The first objective (**objective 1**) was to collect and systematically process the IK indicators farmers use to predict season behaviour. To address this objective, farmers during the qualitative data gathering (focus group discussion followed by the survey) were requested to share their knowledge on the IK indicators that fall in the category of meteorological, astronomical and environmental, that they use to predict the season behaviour. They were also requested to provide insights on how the observation of these indicators impacts their certainty level of an anticipated rainy season behaviour. Among the critical issues, respondents raised is a lack of trustworthiness of the IK indicators due to their poor rainfall and season change predictions. However, multiple observations of these indicators either in different locations or timestamps depending on their nature, bring a glimmer of trust towards the predicted season behaviour. The knowledge gained and the analysis conducted helped the researcher understand the dynamics surrounding the observation of IK indicators concerning their nature, which can be environmental, astronomical, or meteorological. The algorithm was developed that uses the location and timestamp distance parameters to systematically register the observations of the IK indicators while minimising erroneous data (duplicate observations). The algorithm aimed to systematically discriminate between single and multiple observations of the IK indicators considering their nature.

The second objective (**objective 2**) was to collect and process the historical satellite imagery and weather data using GIS and machine learning imputation models. The collection and processing of these secondary data were performed to supplement farmers' indigenous knowledge in season behaviour predictions. The historic weather data (daily rainfall and temperature data) were collected from the district weather station in Pietermaritzburg close to the Swayimane region. These data were pre-processed by filling up the missing entries with the help of various ML imputation methods. The free historic Sentinel 2 satellite images were extracted from various satellite sources including Earth Explorer-USGS and SciHub Copernicus websites. The QGIS was used to process the satellite data. For every image, the region of interest was located with the help of Basemap and Semi Classification Plugin tools, the area of pixel intensity values that represent healthy vegetation, soil moisture, and waterbodies covers were computed using vegetation and water indices, and recorded into a

cloud database for further analysis. The selection of these land cover features was because of their sensitivity to climate conditions. Therefore, they had the potential to complement the climate knowledge, particularly in the areas that are remote from the weather stations. In addition, these land cover features form part of the locally based environmental indicators that farmers continue to assess towards their cropping decisions. For example, the change in vegetation greenness can reflect on the quantity of rain and the season transition from either warm to cold or cold to warm. Furthermore, when water surfaces remain flooded after the first rains, it is anticipated by farmers that more rains are coming.

The third objective (**objective 3**) of the study was to train and evaluate the time series models on historical weather and satellite data to generate the next season's behaviour forecasts representing the scientific perspective. The historic weather and satellite data (vegetation, soil moisture and waterbodies land area covers) were analysed and were used to train the well-known time series models (SARIMA, Holt Winter and LSTM). These models were selected based on their capacity to handle the time series data that have trend and seasonality effects. These supervised models were trained and evaluated to identify the ones with optimal performance. The best-performing models were used to generate the next season's behaviour predictions. This includes the weekly time series forecasts of rainfall, temperature, change in vegetation, water surface and soil moisture covers.

The fourth objective (**objective 4**) of the study was to quantify the season behaviour prediction made via the IK knowledge system with the help of scientific data and time series models. This was done to prepare the IK's predictions to be integrated with the scientific forecasts. The key limitation with the predictions made via the IK system is the failure to provide insight knowledge on how rains and temperature will behave throughout the season such as rain's onset, distribution, magnitude, duration and so forth.

The algorithms were developed to aid in quantifying farmers' season behaviour predictions. These algorithms are responsible for extracting the observed IK indicators from the cloud storage, and use them to evaluate the farmers' certainty level of the rainy season behaviour being anticipated. The certainty level is calculated using the weights of the observed indicators. If farmers are certain of, for instance, abundant rains coming, the already prepared historical weather and satellite data from the cloud storage that signify the abundant rains season will be extracted. The algorithms will further refine the extracted data based on the timestamp the forecasts are made by the farmer. For instance, if the forecasts are made in the middle of the

rainy season, only historical data that signify the abundant rains in the middle of the season will be considered. This is done by evaluating the change in the given subset of the time series values based on the season behaviour type predicted. For instance, if the drought season is predicted, the decrease in healthy vegetation cover and water will be evaluated from the given subset of the historical time series data.

The qualifying historical data is passed to the SES time series model which will make the next season's behaviour predictions. This time series model is not effective on the data that signify trend and seasonality effects. However, it was selected based on its simplicity to develop and customise to meet the needs of this research. This model is utilised during the system's runtime, so, given that only one learning parameter has to be optimised unlike with other sophisticated exponential smoothing models that have more than one learning parameter, the system's response time can be minimised especially when working with large data sets.

The fifth objective (**objective 5**) of this research, was to integrate the forecasts that will be generated via the IK's system with the ones made via the scientific perspective. The forecasts are integrated by calculating the mean of the time series forecasts for every time step. The integrated forecasts are submitted to the algorithm that explains every forecasted value in words to make it easy for farmers to comprehend. The forecasts are presented in a table form and also in text description where each forecasted observation is explained in words. The advanced display of the forecasts is available to enable farmers to visualize the movement of time series observations on a line graph.

The sixth objective (**objective 6**) of the study was to develop a machine learning clustering (optimisation) model that will continually learn from the IK and science experts with the support of wireless sensors to guide farmers in monitoring the health and development of their crops. The development of this optimisation model was motivated by an increasing number of crop complications due to climate stress, and a lack of guidance and support from the extension officers or agricultural scientists in farmers' struggle to sustain better crop yield. During the qualitative data collection, the respondents were requested to share their knowledge on how they monitor the health and growth of their crops. This included the crop and soil properties or characteristics that they use to early diagnose crop problems, methods they use to treat various crop problems and improve soil nutrition, knowledge on how they map various crop symptoms to various crop stresses, knowledge on how they select the crops to grow based on the anticipated season behaviour, soil status, and many more. From the knowledge gathered it

became evident to the researcher that the use of the IK system in monitoring the health and growth of the crops is no longer as effective as it was before, as farmers start to introduce external inputs into their farms.

The knowledge gathered helped in the development of the proposed optimisation model. This model is a customised K-prototype clustering algorithm capable of handling any question or input made up of one or many features provided in any sequence. A feature represents a bit of information that can be made via a human observation or a sensor. The model is continually trained by the IK and science experts to improve its learning capacity. It learns from various sets of questions and recommendations and can optimally group information (questions and recommendations) into clusters based on some similarities possessed by the information. When a new question is provided as input to the model, the question will be mapped to the cluster that contains a set of questions that relate to the input. The recommendations within the identified cluster will be extracted and will be provided to answer the question or input provided. The answer is structured such that the recommendations in the cluster that have high frequency are being prioritised to make it easier for farmers to allocate more of their focus and resources to where they are needed most. The knowledge (recommendations) that will be presented to the farmer or the user of the application is deemed to expound the indigenous and scientific approaches to tackle a given problem. The IK by its nature, is simple to employ, it is inexpensive and it is environmentally friendly. Hence, there is a high possibility that a farmer will first try the indigenous recommendations before shifting to the scientific ones.

This proposed optimisation tool can be considered as an artificial IK and science expert that can stand in the absence of human experts to address the needs of the farmers in monitoring the health and development of their crops.

This model was evaluated under the crop selection domain and has received over 80% accuracy. This accuracy was also deemed to increase for as long as more knowledge was generated.

5.1.3. Opportunities and challenges of the proposed research, along with directions for future studies

All the objectives in this research were successfully met and were fused into an Android mobile technology that farmers will use to make season behaviour predictions and monitor the health and growth of their crops. The selection of Android mobile-based technology was because of its alarming usage increase due to affordability, usability, flexibility and so forth. Smartphone

usage has expanded to local areas, with over 57% of farmers who participated in this research project reporting to have smartphones.

The mobile application was designed in such a way that it is decoupled from the data storage to enhance its ability to accommodate different farmers from different geographical regions in the uMgungundlovu district who will also inject their indigenous knowledge system.

The development of the learning expert tool (addressed by **objective 6**) was also driven by the vast amount of information or knowledge embedded by local communities, particularly the elderly ones, in fields like medicine, politics, agriculture, and education, much of which remains unexplored and undocumented. The proposed tool aims not only to preserve this knowledge but also to utilise its expertise for effective sharing. Although this learning expert tool was used to monitor the health and growth of crops, its use is not confined to only the agricultural domain. It can be applied in different disciplines including nutrition and medicine to address various human and animal health challenges. Additionally, the tool is not limited to English; it can be trained in the native languages of communities to enhance its effectiveness. With the knowledge stored in the cloud database, the model can be continuously trained and tested by farmers anytime and anywhere, providing an opportunity to enhance its accuracy.

Despite the achievement attained in this research project, there are several limitations that the researcher wants to outline to spark interest towards further research and developments to improve the current project. The first limitation relates to the use of remote sensing technology to extract valuable knowledge about the Earth's surface. This research project was conducted using free Sentinel 2 satellite images with a coarse spatial resolution that was hindering the observation and monitoring of small objects on the ground surface that could have added value to this research project. This includes the likes of small water surfaces that are formed during the wet season. Although changes in vegetation density serve as a useful indicator of rainfall and temperature variations, it can be challenging for farmers to correlate vegetation behaviour with rainfall during the cold rainy season, as many plant species, like grass and shrubs, are sensitive to low temperatures.

Changes in the selected ground features (vegetation, soil moisture, and water surfaces) are vulnerable to human activities like deforestation, residential development, sewage issues, and others, which can be misinterpreted as seasonal or weather-related changes. Wrong

interpretation or capturing of these land cover features will lead to erroneous forecasts that will negatively impact farmers' cropping decisions.

The atmospheric distortions including clouds and mist also impact the quality of the satellite data. For instance, some of the images acquired during the rainy season were distorted with cloudy pixels. These cloudy pixels have fixed values that cannot be altered further to extract the hidden information. Although the pre-processing techniques are available to reduce the noise in the image filled with cloudy pixels, they are replacing the pixels, and not fixing them.

A farmer is presented with a weekly season behaviour forecast due to the temporal resolution of the satellite data. Therefore, these forecasts are failing to unpack the daily variations in the forecasts. The pre- and post-processing of satellite images is a man-machine process where human involvement is highly required, and, hence prone to mistakes that can yield wrong analysis of the data.

The quality of data interpretation of satellite data is also impacted by both systematic and non-systematic errors, including sensor characteristics, stray lights, illumination variation, light source sensor geometry, and other environmental factors such as atmospheric effects (Jafarbiglu and Pourreza, 2023).

This research project further focuses only on the univariate time series models which try to explain the time series concerning its historical information (or past observations) while ignoring other factors or variables that may influence the changes in the time series data. For instance, rainfall forecast has to incorporate other variables such as air temperature, air pressure, wind speed and direction, humidity, cloud coverage, and so forth that can highly contribute towards more accurate predictions. This paves the way for the use of multivariate time series models, including Facebook Prophet, LSTM, Vector Auto Regression (VAR), and others. The idea of multivariate forecasting is mathematically presented below using a system of regression models (one-order movement of VAR) where the forecast in one variable depends on the historical observation of that variable and the interdependent variable(s).

$$y_{1t} = \mu_1 + \theta_{11}y_{1,t-1} + \theta_{12}y_{2,t-1} + \epsilon_1 \quad (\text{Equation 4-44})$$

$$y_{2t} = \mu_2 + \theta_{21}y_{1,t-1} + \theta_{22}y_{2,t-1} + \epsilon_2 \quad (\text{Equation 4-45})$$

From the equation presented above, the forecast of a given variable on a given timestep (y_{1t}) depends on its previous timestep $y_{1,t-1}$ and the previous timestep of the other dependent variable $y_{2,t-1}$ and an error term.

The multivariate time series models are powerful and can provide noteworthy weather and satellite forecasting results that can increase the farmers' resilience towards climate change.

The optimisation model presented in this research is a customised K-prototype clustering algorithm that has some limitations. The first limitation is the misclassification of a new data point, which may be assigned to a cluster based solely on its short distance to the cluster's centre. In a situation of varying clusters' sizes and densities, the points may be misclassified just because they are far from their original centres. The hierarchical clustering methods can improve the partitioning of the data by creating the hierarchical decomposition of the given set of data objects. This approach is also less sensitive to noise, and it is capable of clustering the data objects without having to manually specify the number of clusters that should be created by the model (Kaushik and Mathar, 2014). The second limitation as discussed earlier in this research project is the failure to reach a global optimum given that every starting position of the centroids may result in different formations of the clusters. Although this research project demonstrated efforts to mitigate the local optima problem, it only manipulates and evaluates a limited set of data objects used as starting positions for the centroids. Because the distribution of the data in the database is not known, the algorithm assumes that the best starting position for centroids is the one that can result in fair or normal distribution of data records.

The optimization tool lacks natural language processing (NLP) capabilities, which could improve communication between farmers and the system. Currently, creating features and recommendations before generating knowledge the model can learn from is time-consuming and cumbersome. Additionally, integrating image capturing and processing could streamline communication and enhance knowledge sharing between farmers and the system.

Many algorithms discussed in this research project including the clustering algorithm, were developed from scratch by the researcher and were tailored accordingly to meet the needs of this research. However, they still need to be leveraged by IT and computer science professionals to enhance their effectiveness, particularly by optimising search algorithms to reduce system response time when handling large datasets.

Other factors that can hinder the use of the proposed technology by small-scale farmers include the following: first, the utilisation of smartphones by elderly farmers has not yet been fully optimised. Secondly, for the application to function, an internet connection is required to allow the app to access and retrieve data from cloud storage. Finally, for the tool to be effectively used in monitoring crop health and growth, soil and climate sensor data must be accessible to enable the farmer to make informed decisions. The Swayimane region among others in uMgungundlovu is equipped with a local weather station that can be integrated into the proposed mobile-based technology to enhance farmers' cropping decisions. This weather station provides various parameters, including solar radiation, air temperature, relative humidity, precipitation, wind speed and direction, air pressure, visibility, and cloud cover.

CHAPTER 6: CONCLUSION

Numerous studies conducted have dwelled on the importance of the integration of IK and scientific knowledge to tackle the climate-related issues and improve the resilience and livelihood of the small-scale farmers. However, little has been done to incorporate these two knowledge domains due to among others their nature. One is local and unstructured while the other is general and structured. So, in this project, efforts have been made to develop a mobile-based technology that integrates these two knowledge domains to optimise farmers' cropping decisions. The focus was on how the IK in rainy season behaviour predictions and monitoring the health and development of the crops can be systematically processed to become quantitative and, hence, integrated with the existing scientific data and models to benefit the farmer.

Despite the remarkable potential inherent in the proposed mobile-based technology that facilitates the integration of these knowledge systems, it has not yet been deployed on the ground level where it will be used by small-scale farmers to support their cropping decisions. As a way forward, the proposed mobile application will be installed on farmers' smartphones, and they will be requested to use the application for a given time frame. The survey will then be issued to farmers to share their experience with the use of the tool including its strengths and limitations. The knowledge gathered will be used to improve the proposed tool to leverage the farmers' cropping decisions.

In conclusion, local communities possess a wealth of knowledge across various fields, including medicine, agriculture, nutrition, politics, and education, much of which has not been fully preserved or utilised. Fully tapping into this knowledge could reduce the heavy reliance on chemicals for food production, and treatment of various health issues. Many health problems people face today stem from chemically laden food, and ironically, these health issues are often treated with the same chemicals that caused them (Baillie-Hamilton, 2005, Cohen and Vom Saal, 2020 and Reid and Price, 2023). This highlights the importance of IK in addressing these challenges.

The existence and importance of the IK are not only impacted by climate change and other human-related activities discussed in this thesis but also the mindset. We are living in a world driven by money, and market competition that continues to suppress and shadow the IK which is noncompetitive, natural and ecological and further all about quality of food production than quantity (Vogliano et al., 2020 and Gomez, 2022). Due to climate stress, an alarming increase in human population and a decrease in food production, the balance between quantity and

quality is being missed. Further, quantity continues to suppress the quality. This was noted during the focus group discussion that the use of chemicals can help farmers double their production and bring the produce ready for the market.

So, the integration of the IK and science is to try to optimise or find a balance between quantity (science) and quality (IK) while trying to increase food production and also conserving the IK that seems to perish in the coming years.

BIBLIOGRAPHY

1. Abebe, M.A., 2014. Climate change, gender inequality and migration in East Africa. *Wash. J. Env'tl. L. & Pol'y*, 4, p.104.
2. Aboagye-Sarfo, P., Mai, Q., Sanfilippo, F.M., Preen, D.B., Stewart, L.M. and Fatovich, D.M., 2015. A comparison of multivariate and univariate time series approaches to modelling and forecasting emergency department demand in Western Australia. *Journal of biomedical informatics*, 57, pp.62-73.
3. Achieng, M.S., 2022. Digitalisation of indigenous knowledge of agricultural practices: A strategy for food security in Africa. *Journal of Public Administration*, 57(4), pp.802-818.
4. Adebiyi, Marion Olubunmi, Roseline Oluwaseun Ogundokun, and Aneoghena Amarachi Abokhai. "Machine learning-based predictive farmland optimisation and crop monitoring system." *Scientifica* 2020 (2020).
5. Agrawal, A., 2014. Indigenous and scientific knowledge: some critical comments. *Antropologi Indonesia*.
6. Agricultural Survey 2017. Statistics South Africa
7. Ahmad, F., 2013. Landsat ETM+ and MODIS EVI/NDVI data products for climatic variation and agricultural measurements in Cholistan Desert. *Global Journal of Human-Social Science Research*, 12(13-B).
8. Aina, L.O., 2007, August. Globalisation and small-scale farming in Africa: what role for information centres. In *World Library and Information congress: 73rd IFLA general conference and council* (pp. 1-8).
9. Akanbi, A.K. and Masinde, M., 2015, December. Towards semantic integration of heterogeneous sensor data with indigenous knowledge for drought forecasting. In *Proceedings of the Doctoral Symposium of the 16th International Middleware Conference* (p. 2). ACM.
10. Akanbi, A.K. and Masinde, M., 2018, August. Towards the development of a rule-based drought early warning expert systems using indigenous knowledge. In *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)* (pp. 1-8). IEEE.

11. Akbari, E., Darvishi Boloorani, A., Neysani Samany, N., Hamzeh, S., Soufizadeh, S. and Pignatti, S., 2020. Crop mapping using random forest and particle swarm optimisation based on multi-temporal Sentinel-2. *Remote Sensing*, 12(9), p.1449.
12. Al-Bakri, J., Suleiman, A., Abdulla, F. and Ayad, J., 2011. Potential impact of climate change on rainfed agriculture of a semi-arid basin in Jordan. *Physics and Chemistry of the Earth, Parts A/B/C*, 36(5-6), pp.125-134.
13. Al-Kindi, K.M., Kwan, P., Andrew, N.R. and Welch, M., 2017. Remote sensing and spatial statistical techniques for modelling *Ommatissus lybicus* (Hemiptera: Tripiduchidae) habitat and population densities. *PeerJ*, 5, p.e3752.
14. Altman, N. and Krzywinski, M., 2017. Ensemble methods: bagging and random forests. *Nature Methods*, 14(10), pp.933-935.
15. Andersson, J.C.M., Zehnder, A.J., Jewitt, G.P. and Yang, H., 2009. Water availability, demand and reliability of in situ water harvesting in smallholder rain-fed agriculture in the Thukela River Basin, South Africa. *Hydrology and Earth System Sciences*, 13(12), pp.2329-2347.
16. Asare-Nuamah, P. and Botchway, E., 2019. Comparing smallholder farmers' climate change perception with climate data: the case of Adansi North District of Ghana. *Heliyon*, 5(12).
17. Asare-Nuamah, P., Mandaza, M.S. and Amungwa, A.F., 2022. Adaptation strategies and farmer-led agricultural innovations to climate change in Mbire District of Zimbabwe. *International Journal of Rural Management*, 18(2), pp.206-231.
18. Atiah, W.A., Tsidu, G.M. and Amekudzi, L.K., 2020. Investigating the merits of gauge and satellite rainfall data at local scales in Ghana, West Africa. *Weather and Climate Extremes*, 30, p.100292.
19. Atzberger, C. and Eilers, P.H., 2011. A time series for monitoring vegetation activity and phenology at 10-daily time steps covering large parts of South America. *International Journal of Digital Earth*, 4(5), pp.365-386.
20. Azami, H., Mohammadi, K. and Bozorgtabar, B., 2012. An improved signal segmentation using moving average and Savitzky-Golay filter. *Journal of Signal and Information Processing*, 3(01), p.39.
21. Baillie-Hamilton, P., 2005. Toxic overload: a doctor's plan for combating the illnesses caused by chemicals in our foods, our homes, and our medicine cabinets. Penguin.

22. Basdew, M., Jiri, O. and Mafongoya, P.L., 2017. Integration of indigenous and scientific knowledge in climate adaptation in KwaZulu-Natal, South Africa. *Change and Adaptation in Socio-Ecological Systems*, 3(1), pp.56-67.
23. Benedetti, Y., Callaghan, C.T., Ulbrichová, I., Galanaki, A., Kominos, T., Abou Zeid, F., Ibáñez-Álamo, J.D., Suhonen, J., Díaz, M., Markó, G. and Bussière, R., 2023. EVI and NDVI as proxies for multifaceted avian diversity in urban areas. *Ecological Applications*, 33(3), p.e2808.
24. Boon, Mieke, and Sophie Van Baalen. "Epistemology for interdisciplinary research—shifting philosophical paradigms of science." *European journal for philosophy of science* 9.1 (2019): 16.
25. Braun, J. V. (2020). Climate change risks for agriculture, health, and nutrition. In *Health of people, health of planet and our responsibility* (pp. 135-148). Springer, Cham.
26. Brede, B., Suomalainen, J., Bartholomeus, H. and Herold, M., 2015. Influence of solar zenith angle on the enhanced vegetation index of a Guyanese rainforest. *Remote sensing letters*, 6(12), pp.972-981.
27. Bringezu, S., 2017. Visions of a sustainable resource use. In *Sustainable resource management* (pp. 155-215). Routledge.
28. Buthelezi, S.P., Nxumalo, L. and Ngema, X.T., 2024. Challenges faced by small scale sugarcane growers: An exploration of the impact of social unrest on sugarcane farming in Kwazulu-Natal.
29. Carvalho, F.P., 2017. Pesticides, environment, and food safety. *Food and energy security*, 6(2), pp.48-60.
30. Carvallo, H.O., Holzapfel, E.A., Lopez, M.A. and Marino, M.A., 1998. Irrigated cropping optimisation. *Journal of irrigation and Drainage Engineering*, 124(2), pp.67-72.
31. Chagonda, I., Mugabe, F.T., Munodawafa, A., Mubaya, C.P., Masere, P. and Murewi, C., 2015. Engaging smallholder farmers with seasonal climate forecasts for sustainable crop production in semi-arid areas of Zimbabwe. *African Journal of Agricultural Research*, 10(7), pp.668-676.
32. Chandrasekhar, S., Pandey, V.L. and Sahoo, S., 2022. Land use and cropping decisions of agricultural households in India. *Food Security*, pp.1-15.

33. Charbuty, B. and Abdulazeez, A., 2021. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(01), pp.20-28.
34. Chatfield, C., 2018. *Statistics for technology: a course in applied statistics*. Routledge.
35. Chen, X., Xiang, S., Liu, C.L. and Pan, C.H., 2014. Vehicle detection in satellite images by hybrid deep convolutional neural networks. *IEEE Geoscience and remote sensing letters*, 11(10), pp.1797-1801.
36. Chepchirchir, Sally, and Tom Kwanya. "An Analysis of Indigenous Knowledge Legislation and Policies in Kenya." (2019).
37. Chlingaryan, A., Sukkariéh, S. and Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and electronics in agriculture*, 151, pp.61-69.
38. Cohen, A. and vom Saal, F., 2020. *Non-toxic: Guide to Living Healthy in a Chemical World*. Dr Weil's Healthy Living Guides.
39. Community Survey 2016 Agricultural households. Statistics South Africa
40. Davey, J., 2018. *Sugar cane monitoring and analysis using remote sensing*. Wooding Geospatial Solutions,
41. De Castro, A.I., Torres-Sánchez, J., Peña, J.M., Jiménez-Brenes, F.M., Csillik, O. and López-Granados, F., 2018. An automatic random forest-OBIA algorithm for early weed mapping between and within crop rows using UAV imagery. *Remote Sensing*, 10(2), p.285.
42. Demi, S.M. and Sicchia, S.R., 2021. Agrochemicals use practices and health challenges of smallholder farmers in Ghana. *Environmental Health Insights*, 15, p.11786302211043033.
43. Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D. and Raska, R., 2018, June. Deepglobe 2018: A challenge to parse the earth through satellite images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 172-17209). IEEE.
44. Dhanachandra, N., Manglem, K. and Chanu, Y.J., 2015. Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, 54, pp.764-771.
45. Dinku, T., 2019. Challenges with availability and quality of climate data in Africa. In *Extreme hydrology and climate variability* (pp. 71-80). Elsevier.

46. Dury, J., Schaller, N., Garcia, F., Reynaud, A. and Bergez, J.E., 2012. Models to support cropping plan and crop rotation decisions. A review. *Agronomy for sustainable development*, 32(2), pp.567-580.
47. Eastman, J.R., Sangermano, F., Machado, E.A., Rogan, J. and Anyamba, A., 2013. Global trends in seasonality of normalized difference vegetation index (NDVI), 1982–2011. *Remote Sensing*, 5(10), pp.4799-4818.
48. Enock, C.M., 2013. Indigenous knowledge systems and modern weather forecasting: exploring the linkages. *Journal of Agriculture and Sustainability*, 2(2).
49. Ezzine, H., Bouziane, A., Ouazar, D. and Hasnaoui, M.D., 2017. Downscaling of open coarse precipitation data through spatial and statistical analysis, integrating NDVI, NDWI, elevation, and distance from sea. *Advances in Meteorology*, 2017.
50. Fachin, P.A. and Thomaz, E.L., 2023. Fire severity in slash-and-burn agriculture in southern Brazil: an overview. *Scientia Agricola*, 80, p.e20220042.
51. Fernández-González, María, Francisco Javier Rodríguez-Rajo, Victoria Jato, María Jesús Aira, Helena Ribeiro, Manuela Oliveira, and Ilda Abreu. "Forecasting ARIMA models for atmospheric vineyard pathogens in Galicia and Northern Portugal: Botrytis cinerea spores." *Annals of Agricultural and Environmental Medicine* 19, no. 2 (2012).
52. Fletcher, T., 2009. Support vector machines explained. Tutorial paper, 1118, pp.1-19.
53. Gandhi, G.M., Parthiban, S., Thummalu, N. and Christy, A., 2015. NDVI: vegetation change detection using remote sensing and GIS—A case study of Vellore District. *Procedia Computer Science*, 57, pp.1199-1210.
54. Gandhi, G.M., Parthiban, S., Thummalu, N. and Christy, A., 2015. NDVI: vegetation change detection using remote sensing and GIS—a case study of Vellore District. *Procedia Computer Science*, 57, pp.1199-1210.
55. Gerke, S. and Evers, H.D., 2018. Globalizing local knowledge: Social science research on Southeast Asia, 1970–2000. *SOJOURN: Journal of social issues in Southeast Asia*, 33(S), pp.S242-S263.
56. Girard, N., 2015. Knowledge at the boundary between science and society: a review of the use of farmers' knowledge in agricultural development. *Journal of Knowledge Management*, 19(5), pp.949-967.
57. Gaudreault, C., Lama, I. and Sain, D., 2020. Is the beneficial use of wood ash environmentally beneficial? A screening-level life cycle assessment and uncertainty analysis. *Journal of Industrial Ecology*, 24(6), pp.1300-1309.

58. Goldblatt, A., 2010. Agriculture: Facts & Trends: South Africa. Ceo Wwf-Sa.
59. Gomez, C., White, J.C. and Wulder, M.A., 2016. Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 116, pp.55-72.
60. Gómez-Baggethun, E., 2022. Is there a future for indigenous and local knowledge?. *The Journal of Peasant Studies*, 49(6), pp.1139-1157.
61. Gupta, A.D., 2012. Way to study indigenous knowledge and indigenous knowledge system. *Cultural Anthropology*, 373, p.393.
62. Hadlos, A., Opdyke, A. and Hadigheh, S.A., 2022. Where does local and indigenous knowledge in disaster risk reduction go from here? A systematic literature review. *International journal of disaster risk reduction*, 79, p.103160.
63. Hastie, T.J. and Pregibon, D., 2017. Generalized linear models. In *Statistical models in S* (pp. 195-247). Routledge.
64. Ho, D.Y.F., Ho, R.T.H. and Ng, S.M., 2006. Investigative Research as a Knowledge-Generation Method: Discovering and Uncovering. *Journal for the Theory of Social Behaviour*, 36(1), pp.17-38.
65. Huang, S., Ming, B., Huang, Q., Leng, G. and Hou, B., 2017. A case study on a combination NDVI forecasting model based on the entropy weight method. *Water Resources Management*, 31(11), pp.3667-3681.
66. Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X. and Ferreira, L.G., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote sensing of environment*, 83(1-2), pp.195-213.
67. Hyndman, R.J. and Athanasopoulos, G., 2018. *Forecasting: principles and practice*. OTexts.
68. Inglada, J., Vincent, A., Arias, M. and Marais-Sicre, C., 2016. Improved early crop type identification by joint use of high temporal resolution SAR and optical image time series. *Remote Sensing*, 8(5), p.362.
69. Jafarbiglu, H. and Pourreza, A., 2023. Impact of sun-view geometry on canopy spectral reflectance variability. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196, pp.270-286.

70. Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B. and Ermon, S., 2016. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301), pp.790-794.
71. Ji, L., Zhang, L., Wylie, B.K. and Rover, J., 2011. On the terminology of the spectral vegetation index (NIR– SWIR)/(NIR+ SWIR). *International journal of remote sensing*, 32(21), pp.6901-6909.
72. Jiri, O. and Mafongoya, P.L., 2018. Managing vulnerability to drought and enhancing smallholder farmers resilience to climate change risks in Zimbabwe. *Handbook of climate change resilience*, pp.1-17.
73. Jiri, O., Mafongoya, P.L., Mubaya, C. and Mafongoya, O., 2016. Seasonal climate prediction and adaptation using indigenous knowledge systems in agriculture systems in Southern Africa: a review. *Journal of Agricultural Science*, 8(5), pp.156-172. Jiri, O., Mafongoya, P.L., Mubaya, C. and Mafongoya, O., 2016. Seasonal climate prediction and adaptation using indigenous knowledge systems in agriculture systems in Southern Africa: a review. *Journal of Agricultural Science*, 8(5), pp.156-172. Jiri, O., Mafongoya, P.L., Mubaya, C. and Mafongoya, O., 2016. Seasonal climate prediction and adaptation using indigenous knowledge systems in agriculture systems in Southern Africa: a review. *Journal of Agricultural Science*, 8(5), pp.156-172. Jiri, O., Mafongoya, P.L., Mubaya, C. and Mafongoya, O., 2016. Seasonal climate prediction and adaptation using indigenous knowledge systems in agriculture systems in Southern Africa: a review. *Journal of Agricultural Science*, 8(5), pp.156-172.
74. Jiri, O., Mafongoya, P.L., Mubaya, C. and Mafongoya, O., 2016. Seasonal climate prediction and adaptation using indigenous knowledge systems in agriculture systems in Southern Africa: a review. *Journal of Agricultural Science*, 8(5), pp.156-172.
75. Johan, P.D., Ahmed, O.H., Omar, L. and Hasbullah, N.A., 2021. Phosphorus transformation in soils following co-application of charcoal and wood ash. *Agronomy*, 11(10), p.2010.
76. Jones, C., Kielland, K., Hinzman, L. and Schneider, W., 2015. Integrating local knowledge and science: economic consequences of driftwood harvest in a changing climate. *Ecology and Society*, 20(1).
77. Jørgensen, J.B., Huusom, J.K. and Rawlings, J.B., 2011, December. Finite horizon MPC for systems in innovation form. In 2011 50th IEEE Conference on Decision and Control and European Control Conference (pp. 1896-1903). IEEE.

78. Kala, C.P., Farooquee, N.A. and Dhar, U., 2004. Prioritization of medicinal plants on the basis of available knowledge, existing practices and use value status in Uttaranchal, India. *Biodiversity & Conservation*, 13(2), pp.453-469.
79. Kammili, R., Chandu, V.R. and Valurouthu, K.P., 2018. A Novel LISS-III Swath Extension through Wider AWiFS Swath and SVM Training and Learning. *International Journal of Simulation--Systems, Science & Technology*, 19(6).
80. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.
81. Kiadtikornthaweeeyot, W & Tatnal,A (2016). Region of interest detection based on histogram segmentation for satellite image. *The international archives of the photogrammetry, remote sensing and spatial information science*.
82. Krippendorff, K., 2018. *Content analysis: An introduction to its methodology*. Sage publications.
83. Lamptey, S., 2022. Agronomic practices in soil water management for sustainable crop production under rain fed agriculture of Drylands in Sub-Sahara Africa. *African Journal of Agricultural Research*, 18(1), pp.18-26.
84. Lata, K. and Khan, S., 2020, June. Seasonal Environmental Data-Sets Simulations for Optimising the Crop Yield. In *2nd International Conference on Communication & Information Processing (ICCIP)*.
85. Lawens, M., Eckhardt, H. and Gramel, S., 2019. Core aspects of international water supply and wastewater disposal with a focus on developing countries in the Asia and Africa region. *Water Supply*, 19(6), pp.1809-1815.
86. Lei, L.I.U., ZHOU, L.L. and BAO, H.F., 2017. Remote sensing image segmentation algorithm based on multi-agent and fuzzy clustering. *DEStech Transactions on Environment, Energy and Earth Sciences*, (icesee).
87. Li, M., Xu, H. and Deng, Y., 2019. Evidential decision tree based on belief entropy. *Entropy*, 21(9), p.897.
88. Liao, Z., He, B. and Quan, X., 2015. Modified enhanced vegetation index for reducing topographic effects. *Journal of Applied Remote Sensing*, 9(1), p.096068.
89. Limpo, S.Y., Fahmid, I.M., Fattah, A., Rauf, A.W., Surmaini, E., Muslimin, Saptana, Syahbuddin, H. and Andri, K.B., 2022. Integrating indigenous and scientific knowledge

- for decision making of rice farming in South Sulawesi, Indonesia. *Sustainability*, 14(5), p.2952.
90. Lobell, D.B. and Asseng, S., 2017. Comparing estimates of climate change impacts from process-based and statistical crop models. *Environmental Research Letters*, 12(1), p.015001.
 91. Lopez-Paz, D., Hennig, P. and Schölkopf, B., 2013. The randomized dependence coefficient. In *Advances in neural information processing systems* (pp. 1-9).
 92. Lu, H.D., Xue, J.Q. and Guo, D.W., 2017. Efficacy of planting date adjustment as a cultivation strategy to cope with drought stress and increase rainfed maize yield and water-use efficiency. *Agricultural Water Management*, 179, pp.227-235.
 93. Magni, G., 2017. Indigenous knowledge and implications for the sustainable development agenda. *European Journal of Education*, 52(4), pp.437-447.
 94. Mairura, F.S., Mugendi, D.N., Mwanje, J.I., Ramisch, J.J., Mbugua, P.K. and Chianu, J.N., 2007. Integrating scientific and farmers' evaluation of soil quality indicators in Central Kenya. *Geoderma*, 139(1-2), pp.134-143.
 95. Masete, Florence M., Lawrence Munjonji, Kingsley K. Ayisi, and Moshibudi P. Mopape-Mabapa. "Cowpea growth and nitrogen fixation performance under different mulch treatments." *Agriculture* 12, no. 8 (2022): 1144.
 96. Masih, I., Maskey, S., Mussá, F.E.F. and Trambauer, P., 2014. A review of droughts on the African continent: a geospatial and long-term perspective. *Hydrology and Earth System Sciences*, 18(9), p.3635.
 97. Masinde, E.M., 2012. Bridge between African Indigenous knowledge and modern science on drought prediction (Doctoral dissertation, UNIVERSITY OF CAPE TOWN).
 98. Masinde, M. and Thothela, P.N., 2019, March. ITIKI Plus: A mobile based application for integrating indigenous knowledge and scientific agro-climate decision support for Africa's small-scale farmers. In *2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT)* (pp. 303-309). IEEE.
 99. Masinde, M., Bagula, A. and Muthama, N.J., 2012, March. The role of ICTs in downscaling and up-scaling integrated weather forecasts for farmers in sub-Saharan Africa. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development* (pp. 122-129).

100. Maskey, S., Mussá, F.E.F. and Trambauer, P., 2014. A review of droughts on the African continent: a geospatial and long-term perspective. *Hydrology and Earth System Sciences*, 18(9), p.3635.
101. Matsushita, B., Yang, W., Chen, J., Onda, Y. and Qiu, G., 2007. Sensitivity of the enhanced vegetation index (EVI) and normalized difference vegetation index (NDVI) to topographic effects: a case study in high-density cypress forest. *Sensors*, 7(11), pp.2636-2651.
102. McBride, B.B., Sanchez-Trigueros, F., Carver, S.J., Watson, A.E., Stumpff, L.M., Matt, R. and Borrie, W.T., 2016. Participatory geographic information systems as an organizational platform for the integration of traditional and scientific knowledge in contemporary fire and fuels management. *Journal of Forestry*, 115(1), pp.43-50.
103. Meraj, G., Kanga, S., Ambadkar, A., Kumar, P., Singh, S.K., Farooq, M., Johnson, B.A., Rai, A. and Sahu, N., 2022. Assessing the yield of wheat using satellite remote sensing-based machine learning algorithms and simulation modeling. *Remote Sensing*, 14(13), p.3005.
104. Mercer, J., Kelman, I., Taranis, L. and Suchet-Pearson, S., 2010. Framework for integrating indigenous and scientific knowledge for disaster risk reduction. *Disasters*, 34(1), pp.214-239.
105. Mercer, J., Kelman, I., Taranis, L. and Suchet-Pearson, S., 2010. Framework for integrating indigenous and scientific knowledge for disaster risk reduction. *Disasters*, 34(1), pp.214-239.
106. Mihale, M.J., Deng, A.L., Selemani, H.O., Kamatenesi, M.M., Kidukuli, A.W. and Ogendo, J.O., 2009. Use of indigenous knowledge in the management of field and storage pests around Lake Victoria basin in Tanzania. *African Journal of Environmental Science and Technology*, 3(9).
107. Mistry, J. and Berardi, A., 2016. Bridging indigenous and scientific knowledge. *Science*, 352(6291), pp.1274-1275.

108. Mndzebele, B., Ncube, B., Nyathi, M., Kanu, S. A., Fessehazion, M., Mabhaudhi, T., ... & Modi, A. T. (2020). Nitrogen fixation and nutritional yield of cowpea-amaranth intercrop. *Agronomy*, 10(4), 565.
109. Mubiru, D.N. and Coyne, M.S., 2009. Legume cover crops are more beneficial than natural fallows in minimally tilled Ugandan soils. *Agronomy journal*, 101(3), pp.644-652.

110. Muradyan, V., Tepanosyan, G., Asmaryan, S., Saghatelian, A. and Dell'Acqua, F., 2019. Relationships between NDVI and climatic factors in mountain ecosystems: A case study of Armenia. *Remote Sensing Applications: Society and Environment*, 14, pp.158-169.
111. Mwagha, S.M. and Masinde, M., 2015, May. Scientific Verification of Weather Lore for Drought Forecasting–The Role of Fuzzy Cognitive Mapping. In *Proceedings of the IST-Africa 2015 Conference*, Lilongwe, Malawi (pp. 6-8).
112. Nakahara, H., Jinguji, A., Sato, S. and Sasao, T., 2017, May. A random forest using a multi-valued decision diagram on an FPGA. In *2017 IEEE 47th international symposium on multiple-valued logic (ISMVL)* (pp. 266-271). IEEE.
113. Ncoyini, Z., Savage, M.J. and Strydom, S., 2022. Limited access and use of climate information by small-scale sugarcane farmers in South Africa: A case study. *Climate Services*, 26, p.100285.
114. Nweke, I.A., 2018. The good, the bad and the ugly of tillage in agricultural sustainability-A review. *Greener Journal of Agricultural Sciences*, 8(9), pp.217-250.
115. Nyetanyane, J., 2023, May. Indigenous Knowledge Crop Health Recommendation Expert System for the Upcoming Generations. In *2023 IST-Africa Conference (IST-Africa)* (pp. 1-8). IEEE.
116. Nyetanyane, J. and Masinde, M., 2022, May. Development of English-isiZulu Language Translator to Bridge Language Barrier Between EWS and Bantu Language Speaking Farmers. Case of Swayimane, Kwazulu-Natal, South Africa. In *2022 IST-Africa Conference (IST-Africa)* (pp. 1-11). IEEE.
117. Nyoni, T., 2018. Box-Jenkins ARIMA approach to predicting net FDI inflows in Zimbabwe.
118. Oreggioni, G.D., Ferraio, F.M., Crippa, M., Muntean, M., Schaaf, E., Guizzardi, D., Solazzo, E., Duerr, M., Perry, M. and Vignati, E., 2021. Climate change in a changing world: Socio-economic and technological transitions, regulatory frameworks and trends on global greenhouse gas emissions from EDGAR v. 5.0. *Global Environmental Change*, 70, p.102350.
119. Owiny, S.A., Mehta, K. and Maretzki, A.N., 2014. The use of social media technologies to create, preserve, and disseminate indigenous knowledge and skills to communities in East Africa. *International Journal of Communication*, 8, p.14.

120. Özalkan, E., 2020. Water body detection analysis using NDWI indices derived from landsat-8 OLI. *Polish Journal of Environmental Studies*, 29(2), pp.1759-1769.
121. Pareek, A. and Trivedi, P.C., 2011. Cultural values and indigenous knowledge of climate change and disaster prediction in Rajasthan, India.
122. Patz, J.A., Gibbs, H.K., Foley, J.A., Rogers, J.V. and Smith, K.R., 2007. Climate change and global health: quantifying a growing ethical crisis. *EcoHealth*, 4(4), pp.397-405.
123. Pavlovsky, G., 2017. *AGRICULTURAL RUSSIA: On the Eve of the Revolution*. Routledge.
124. Pelletier, C., Valero, S., Inglada, J., Champion, N. and Dedieu, G., 2016. Assessing the robustness of Random Forests to map land cover with high resolution satellite image time series over large areas. *Remote Sensing of Environment*, 187, pp.156-168.
125. Perez, E.G. and Ceballos, R.F., 2019. Malaria Incidence in the Philippines: Prediction using the Autoregressive Moving Average Models. *International Journal of Engineering and Future Technology* (2019), 16(4), pp.1-10.
126. Picchioni, F., 2017. Monitoring the effects of changing food prices on food and nutrition security: the minimum calorie expenditure share (MCES) (Doctoral dissertation, SOAS University of London).
127. Picon, A., Alvarez-Gila, A., Seitz, M., Ortiz-Barredo, A., Echazarra, J. and Johannes, A., 2019. Deep convolutional neural networks for mobile capture device-based crop disease classification in the wild. *Computers and Electronics in Agriculture*, 161, pp.280-290.
128. Pörtner, H.O. and Peck, M.A., 2010. Climate change effects on fishes and fisheries: towards a cause-and-effect understanding. *Journal of fish biology*, 77(8), pp.1745-1779.
129. Prema, V. and Rao, K.U., 2015. Development of statistical time series models for solar power prediction. *Renewable energy*, 83, pp.100-109.
130. Primo, A.A., de Araújo Neto, R.A., Zeferino, L.B., Fernandes, F.É.P., de Araújo Filho, J.A., Cerri, C.E.P. and de Oliveira, T.S., 2023. Slash and burn management and permanent or rotation agroforestry systems: A comparative study for C sequestration by century model simulation. *Journal of Environmental Management*, 336, p.117594.

131. Qiu, J., Wang, X., He, S., Liu, H., Lai, J. and Wang, L., 2017. The catastrophic landside in Maoxian county, Sichuan, SW China, on June 24, 2017. *Natural Hazards*, 89(3), pp.1485-1493.
132. Rahman, A. and Ahmar, A.S., 2017, September. Forecasting of primary energy consumption data in the United States: A comparison between ARIMA and Holter-Winters models. In *AIP Conference Proceedings* (Vol. 1885, No. 1, p. 020163). AIP Publishing.
133. Rajaram, T. and Das, A., 2010. Modeling of interactions among sustainability components of an agro-ecosystem using local knowledge through cognitive mapping and fuzzy inference system. *Expert Systems with Applications*, 37(2), pp.1734-1744.
134. Rangarajan, A.K., Purushothaman, R. and Ramesh, A., 2018. Tomato crop disease classification using pre-trained deep learning algorithm. *Procedia computer science*, 133, pp.1040-1047.
135. Rani, G., Kaur, J., Kumar, A. and Yogalakshmi, K.N., 2020. Ecosystem health and dynamics: An indicator of global climate change. *Contemporary environmental issues and challenges in era of climate change*, pp.1-32.
136. Ravinder, H.V., 2013. Determining the Optimal Values of Exponential Smoothing Constants--Does Solver Really Work?. *American Journal of Business Education*, 6(3), pp.347-360.
137. Reddy, D.A., Dadore, B. and Watekar, A., 2019. Crop recommendation system to maximise crop yield in ramtek region using machine learning. *International Journal of Scientific Research in Science and Technology*, 6(1), pp.485-489.
138. Reddy, D.S. and Prasad, P., 2018. Prediction of vegetation dynamics using NDVI time series data and LSTM. *Modeling Earth Systems and Environment*, 4(1), pp.409-419.
139. Reddy, P.P. and Reddy, P.P., 2015. Impacts of climate change on agriculture. *Climate resilient agriculture for ensuring food security*, pp.43-90.
140. Reid, K. and Price, B., 2023. *Fat, Stressed, and Sick: MSG, Processed Food, and America's Health Crisis*. Rowman & Littlefield.
141. Rohrbach, D.D. and Okwach, G.E., 1998. Modeling crop performance or cropping decisions. *Risk Management for Maize Farmers in Drought-Prone Areas of Southern Africa*. ICRISAT and CIMMYT, Mexico, DF, pp.47-58.

142. Roncoli, C., Ingram, K. and Kirshen, P., 2002. Reading the rains: local knowledge and rainfall forecasting in Burkina Faso. *Society & Natural Resources*, 15(5), pp.409-427.
143. Ross, A., Sherman, K.P., Snodgrass, J.G., Delcore, H.D. and Sherman, R., 2016. *Indigenous peoples and the collaborative stewardship of nature: knowledge binds and institutional conflicts*. Routledge.
144. Sabini, M., Rusak, G. and Ross, B., 2017. *Understanding Satellite-Imagery-Based Crop Yield Predictions*. Technical Report. Stanford University. <http://cs231n.stanford.edu/reports/2017/pdfs/555.pdf> [Accessed on 23th Oct 2017]. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. and Dieleman, S., 2016. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587), pp.484-489.
145. Saini, V., Arora, M.K. and Gupta, R.P., 2016, May. Relationship between surface temperature and SAVI using Landsat data in a coal mining area in India. In *Land Surface and Cryosphere Remote Sensing III* (Vol. 9877, p. 987711). International Society for Optics and Photonics.
146. Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O.P., Tiwari, A., Er, M.J., Ding, W. and Lin, C.T., 2017. A review of clustering techniques and developments. *Neurocomputing*, 267, pp.664-681.
147. Schmidt, A.F. and Finan, C., 2018. Linear regression and the normality assumption. *Journal of clinical epidemiology*, 98, pp.146-151.
148. Shafí, U., Mumtaz, R., García-Nieto, J., Hassan, S.A., Zaidi, S.A.R. and Iqbal, N., 2019. Precision agriculture techniques and practices: From considerations to applications. *Sensors*, 19(17), p.3796.
149. Shah, S.R. and Al-Bargi, A., 2013. Research Paradigms: Researchers' Worldviews, Theoretical Frameworks and Study Designs. *Arab World English Journal*, 4(4).
150. Shahtahmassebi, A., Yang, N., Wang, K., Moore, N. and Shen, Z., 2013. Review of shadow detection and de-shadowing methods in remote sensing. *Chinese geographical science*, 23, pp.403-420.

151. Sharma, N. and Singhvi, R., 2017. Effects of chemical fertilisers and pesticides on human health and environment: a review. *International journal of agriculture, environment and biotechnology*, 10(6), pp.675-680.
152. Shashikant, V., Mohamed Shariff, A.R., Wayayok, A., Kamal, M.R., Lee, Y.P. and Takeuchi, W., 2021. Utilising TVDI and NDWI to classify severity of agricultural drought in Chuping, Malaysia. *Agronomy*, 11(6), p.1243.
153. Shi, W., Tao, F. and Zhang, Z., 2013. A review on statistical models for identifying climate contributions to crop yields. *Journal of geographical sciences*, 23(3), pp.567-576.
154. Sowmya, D.R., Rao, S.A., Shenoy, P.D. and Venugopal, K.R., 2018, August. Generation of Digital Elevation Map for Steep Terrain Region Using Landsat-7 ETM+ Imagery. In *2018 International Conference on Data Science and Engineering (ICDSE)* (pp. 1-5). IEEE.
155. Srinivasa, K.G. and Muppalla, A.K., 2015. Case Study III: Regression Analysis using Scalding and Spark. In *Guide to High Performance Distributed Computing* (pp. 219-259). Springer, Cham.
156. Stow, D.A., 2017. Radiometric Correction of Remotely Sensed Data.
157. Swain, K.C., Zaman, Q.U., Schumann, A.W., Percival, D.C. and Bochtis, D.D., 2010. Computer vision system for wild blueberry fruit yield mapping. *Biosystems Engineering*, 106(4), pp.389-394.
158. Tang, K.H.D. and Yap, P.S., 2020, September. A systematic review of slash-and-burn agriculture as an obstacle to future-proofing climate change. In *The Proceedings of The International Conference on Climate Change* (Vol. 4, No. 1, pp. 1-19).
159. Tesfamariam, B.Y., Owusu-Sekyere, E., Emmanuel, D. and Elizabeth, T.B., 2018. The impact of the homestead food garden programme on food security in South Africa. *Food security*, 10(1), pp.95-110.
160. Thakur, N. and Maheshwari, D., 2017. A Review of Image Classification Techniques. *International Research Journal of Engineering and Technology (IRJET)*, 4(11), pp.1588-1591.
161. Tharakan, J., 2017. Indigenous Knowledge Systems for Appropriate Technology Development. *Indigenous People*, p.123.

162. Tharwat, A., Hassanien, A.E. and Elnaghi, B.E., 2017. A BA-based algorithm for parameter optimisation of support vector machine. *Pattern recognition letters*, 93, pp.13-22.
163. Tongkoom, K., Marohn, C., Piepho, H.P. and Cadisch, G., 2021. Combining farmers' and scientists' tree species and soil fertility assessment for improved cropping decisions in swidden systems of Northern Thailand. *Ecological Indicators*, 127, p.107719.
164. Twomlow, S., Mugabe, F.T., Mwale, M., Delve, R., Nanja, D., Carberry, P. and Howden, M., 2008. Building adaptive capacity to cope with increasing vulnerability due to climatic change in Africa—A new approach. *Physics and Chemistry of the Earth, Parts A/B/C*, 33(8-13), pp.780-787.
165. Ugbaje, S.U., Odeh, I.O. and Bishop, T.F., 2019. Fuzzy measure-based multicriteria land assessment for rainfed maize in West Africa for the current and a range of plausible future climates. *Computers and Electronics in Agriculture*, 158, pp.51-67.
166. Valeriano, M.D.M., Sanches, I.D.A. and Formaggio, A.R., 2016. Topographic effect on spectral vegetation indices from Landsat TM data: is topographic correction necessary?. *Boletim de Ciências Geodésicas*, 22(1), pp.95-107.
167. Verbesselt, J., Hyndman, R., Newnham, G. and Culvenor, D., 2010. Detecting trend and seasonal changes in satellite image time series. *Remote sensing of Environment*, 114(1), pp.106-115.
168. Verma, N. and Kaur, H., 2017. TRAFFIC ANALYSIS AND PREDICTION SYSTEM BY THE USE OF MODIFIED ARIMA MODEL. *International Journal of Advanced Research in Computer Science*, 8(8).
169. Verma, U., Koehler, W. and Goyal, M., 2012. A study on yield trends of different crops using ARIMA analysis. *Environment and Ecology*, 30(4A), pp.1459-1463.
170. Vibhute, A.D., Dhumal, R.K., Nagne, A.D., Rajendra, Y.D., Kale, K.V. and Mehrotra, S.C., 2016. Analysis, classification, and estimation of pattern for land of Aurangabad region using high-resolution satellite image. In *Proceedings of the Second International Conference on Computer and Communication Technologies* (pp. 413-427). Springer, New Delhi.

171. Vogliano, C., Raneri, J.E., Maelaua, J., Coad, J., Wham, C. and Burlingame, B., 2020. Assessing diet quality of indigenous food systems in three geographically distinct solomon islands sites (Melanesia, Pacific Islands). *Nutrients*, 13(1), p.30.
172. Vora, P. and Oza, B., 2013. A survey on k-mean clustering and particle swarm optimisation. *International Journal of Science and Modern Engineering*, 1(3), pp.24-26.
173. Wang, J., Lin, Y., McFarquhar, G.M., Park, E., Gu, Y., Su, Q., Fu, R., Lee, K.W. and Zhang, T., 2023. Soil moisture observations from shortwave infrared channels reveal tornado tracks: A case in 10–11 December 2021 tornado outbreak. *Geophysical Research Letters*, 50(6), p.e2023GL102984.
174. Wang, W.C., Chau, K.W., Xu, D.M. and Chen, X.Y., 2015. Improving forecasting accuracy of annual runoff time series using ARIMA based on EEMD decomposition. *Water Resources Management*, 29(8), pp.2655-2675.
175. Wang, X., Zhang, J., Xun, L., Wang, J., Wu, Z., Henchiri, M., Zhang, S., Zhang, S., Bai, Y., Yang, S. and Li, S., 2022. Evaluating the effectiveness of machine learning and deep learning models combined time-series satellite data for multiple crop types classification over a large-scale region. *Remote Sensing*, 14(10), p.2341.
176. Wang, Y., Shi, F., Cao, L., Dey, N., Wu, Q., Ashour, A.S., Sherratt, R.S., Rajinikanth, V. and Wu, L., 2019. Morphological segmentation analysis and texture-based support vector machines classification on mice liver fibrosis microscopic images. *Current Bioinformatics*, 14(4), pp.282-294.
177. Wang, Z., Xia, J., Wang, L., Mao, Z., Zeng, Q., Tian, L. and Shi, L., 2018. Atmospheric correction methods for GF-1 WFV1 data in hazy weather. *Journal of the Indian Society of Remote Sensing*, 46(3), pp.355-366.
178. Wheeler, T. and Von Braun, J., 2013. Climate change impacts on global food security. *Science*, 341(6145), pp.508-513.
179. Wood, S.A., Jina, A.S., Jain, M., Kristjanson, P. and DeFries, R.S., 2014. Smallholder farmer cropping decisions related to climate variability across multiple regions. *Global Environmental Change*, 25, pp.163-172.
180. Wu, W., 2014. The generalized difference vegetation index (GDVI) for dryland characterization. *Remote Sensing*, 6(2), pp.1211-1233.
181. Xie, S. and Tu, Z., 2015. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 1395-1403).

182. Xu, H., 2006. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International journal of remote sensing*, 27(14), pp.3025-3033.
183. Xu, W., Chen, P., Zhan, Y., Chen, S., Zhang, L. and Lan, Y., 2021. Cotton yield estimation model based on machine learning using time series UAV remote sensing data. *International Journal of Applied Earth Observation and Geoinformation*, 104, p.102511.
184. You, J., Li, X., Low, M., Lobell, D. and Ermon, S., 2017. Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data. In *AAAI* (pp. 4559-4566).
185. Yu, H., Xie, T., Paszczyński, S. and Wilamowski, B.M., 2011. Advantages of radial basis function networks for dynamic system design. *IEEE Transactions on Industrial Electronics*, 58(12), pp.5438-5450.
186. Zhang, Q., Wang, K., Guo, H., Yang, S. and Li, C., 2018, October. Base on ETS model for Forecasting Emergency Department Visits. In *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)* (pp. 2148-2151). IEEE.
187. Zhang, Y., Gao, J., Liu, L., Wang, Z., Ding, M. and Yang, X., 2013. NDVI-based vegetation changes and their responses to climate change from 1982 to 2011: A case study in the Koshi River Basin in the middle Himalayas. *Global and Planetary Change*, 108, pp.139-148.
188. Zhao, Y., Li, Z., Liu, G., Jiang, Y., Maurer, H.P., Würschum, T., Mock, H.P., Matros, A., Ebmeyer, E., Schachschneider, R. and Kazman, E., 2015. Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. *Proceedings of the National Academy of Sciences*, 112(51), pp.15624-15629.
189. Zhao, Y., Wang, X., Novillo, C.J., Arrogante-Funes, P., Vázquez-Jiménez, R. and Maestre, F.T., 2018. Albedo estimated from remote sensing correlates with ecosystem multifunctionality in global drylands. *Journal of arid environments*, 157, pp.116-123.
190. Zhou, X., Zhu, X., Dong, Z. and Guo, W., 2016. Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *The Crop Journal*, 4(3), pp.212-219.

191. Zwinkels, J., 2015. Light, electromagnetic spectrum. Encyclopedia of Color Science and Technology, 8071, pp.1-8.

Appendix A

```
private ArrayList<FarmersKnowlegde> getHistoricData(String indicatorType, String
seasonType) {
    ArrayList data = conn.readFarmersData();

    ArrayList<FarmersKnowlegde> newData = new ArrayList<>();
    for (Object d : data) {
        if (((FarmersKnowlegde) d).getSeasonType().equalsIgnoreCase(seasonType) &&
            ((FarmersKnowlegde) d).getType().equalsIgnoreCase(indicatorType)) {
            newData.add((FarmersKnowlegde) d);
        }
    }
    return newData;
}
```

```
public ArrayList findDistributionOrRecords(ArrayList records, int k) {
    int count = 0;
    int countK = 1;
    ArrayList clusters = new ArrayList();
    while (countK <= k) {
        for (Object obj : records) {
            for (Object obj2 : (ArrayList) obj) {
                if (((FeatureRecommendation) obj2).getClusterNumber() == countK) {
                    count++;
                }
                break;
            }
        }
        clusters.add(count);
        count = 0;
        countK += 1;
    }
}
```

```

    return clusters;
}

private int getOptimumSplit(ArrayList randomClusters, ArrayList records) {
    double min,max, difference = 0,total;
    ArrayList differences=new ArrayList();
    total = 0;
    for (Object obj : randomClusters) {
        min = getMinimum((ArrayList) obj);
        max = getMax((ArrayList) obj);
        for (int x = 0; x < ((ArrayList) obj).size(); x++) {
            total += Double.parseDouble("" + ((ArrayList) obj).get(x)) / min * max;
        }

        differences.add(total / records.size());
        total = 0;
    }
    int minDifference = getMinimumPosition(differences);
    int bestSplit = (int) positions.get(minDifference);
    return bestSplit;
}

```

```

private ArrayList updateCentroids(ArrayList records) {
    ArrayList<Cluster> cluster = new ArrayList<>();
    int count = 0;
    Cluster c = new Cluster();
    ArrayList clusters = new ArrayList<>();
    DbsConnection conn = new DbsConnection();
    ArrayList categoricalNames = new ArrayList(), numericalNames = new ArrayList(),
    valuesPerCat = new ArrayList(),
        averageOfNumericalValues = new ArrayList(), catValues, clusterNumbers,

```



```

bestValues = new ArrayList(), frequencyOfValues = new ArrayList();
Cluster fc = new Cluster();
clusterNumbers = conn.getClusterNumbers(records);
for (Object clusterN : clusterNumbers) {
    categoricalNames = conn.getCategoryNames(records, clusterN);
    if (categoricalNames.size() > 0) {
        for (Object catname : categoricalNames) {
            catValues = conn.getCatValues(records, (String) catname, (int) clusterN);
            valuesPerCat.add(catValues);
        }
        count = 0;
        for (Object vpc : valuesPerCat) {
            for (Object obj : (ArrayList) vpc) {
                frequencyOfValues.add(conn.getFrequency((String) obj, clusterN, records,
categoricalNames.get(count).toString()));
            }
            String bestValue = conn.getBestValue((ArrayList) vpc, frequencyOfValues);
            bestValues.add(bestValue);
            frequencyOfValues = new ArrayList();

            count++;
        }

        count = 0;
        for (Object name : categoricalNames) {
            c.setFeatureName((String) name);

            c.setFeatureValues("" + bestValues.get(count));
            c.setClusterN((int) clusterN);
            cluster.add(c);
            c = new Cluster();
            count++;
        }
    }
}

```

```

    }

    numericalNames = conn.getNumericalNames(records, clusterN);
    if (numericalNames.size() > 0) {
        for (Object name : numericalNames) {
            double average = conn.getAverageOfValues((String) name, (int) clusterN,
records);
            averageOfNumericalValues.add(average);
        }
        count = 0;
        c = new Cluster();
        for (Object name : numericalNames) {
            c.setFeatureName((String) name);
            c.setFeatureValues("" + averageOfNumericalValues.get(count));
            c.setClusterN((int) clusterN);
            cluster.add(c);
            c = new Cluster();
            count++;
        }
    }

    clusters.add(cluster);
    bestValues = new ArrayList();
    valuesPerCat = new ArrayList();
    averageOfNumericalValues = new ArrayList();
    cluster = new ArrayList<>();

}
return clusters;
}

private ArrayList updateCentroids2(ArrayList records, int k) {
    recordsPerCluster = new ArrayList();

```

```

ArrayList averageOfFeaturesPerCluster = new ArrayList();
ArrayList distinctFeatureNames = new ArrayList();
ArrayList frequencyOfFeatureNames = new ArrayList();
optimumCentroids = new ArrayList();
centroids = new ArrayList();

for (int c = 1; c <= k; c++) {
    ArrayList recs = getRecordsInCluster(c, records);
    recordsPerCluster.add(recs);
    averageOfFeaturesPerCluster.add(getCountOfObjectsPerCluster(recs));
}
distinctFeatureNames = getDistinctFeatureNames(recordsPerCluster);
frequencyOfFeatureNames = getFrequencyOfFeatureNames(distinctFeatureNames,
recordsPerCluster);
ArrayList optimumFeatureNamesPerCluster = new ArrayList(), optimumFeatureNames =
new ArrayList();

int count = 0;
ArrayList list = new ArrayList(), listForK = new ArrayList();

for (int x = 0; x < k; x++) {
    for (Object obj : frequencyOfFeatureNames) {
        list.add(((ArrayList) obj).get(x));
    }
    listForK.add(list);
    list = new ArrayList();
}
for (Object obj : listForK) {
    long size = (long) averageOfFeaturesPerCluster.get(count);
    if(size>distinctFeatureNames.size())
    {
        size=distinctFeatureNames.size();
    }
}

```

```

for (int x = 0; x < size; x++) {
    double highest = getMax((ArrayList) obj);
    int pos = getMaxPosition(highest, (ArrayList) obj);
    String optimumFeatureName = distinctFeatureNames.get(pos).toString();
    optimumFeatureNamesPerCluster.add(optimumFeatureName);
    ((ArrayList) obj).set(pos, -1);
}
count += 1;
optimumFeatureNames.add(optimumFeatureNamesPerCluster);
optimumFeatureNamesPerCluster = new ArrayList();
}
ArrayList centroids = updateValuesOfOpimumFeatureNames(recordsPerCluster,
optimumFeatureNames);
return centroids;
}

```

```

public ArrayList getOrganisedRecords(ArrayList<FeatureRecommendation>
historic_records) {
    ArrayList records = new ArrayList(), record = new ArrayList();
    ArrayList identities = conn.getFeatureRecommendationIds(historic_records);
    for (Object id : identities) {
        for (FeatureRecommendation fr : historic_records) {
            if (Integer.parseInt(id.toString()) == fr.getFeatureRecId()) {
                record.add(fr);
            }
        }
        records.add(record);
        record = new ArrayList();
    }
    return records;
}

```

```

public double comparison(int[] l1, int[] l2) {
    int count = 0;
    int countRec = 0;
    boolean flag = false;
    double sum = 0;
    for (int c : l1) {

        if (c == l2[count]) {
            sum += 1;
        }
        count++;
    }

    return sum / l1.length * 100;
}

```

```

public double getClusterErrors(int k) {
    double average = 0;
    int count = 0;
    double penalty = 0;
    for (Object obj : recordsPerCluster) {
        for (Object record : (ArrayList) obj) {
            ArrayList centroid = (ArrayList) centroids.get(count);
            penalty += Math.abs((centroid.size() - ((ArrayList) record).size())) * 2;
            for (Object c : centroid) {
                penalty += ((Cluster) c).computePenalty2((ArrayList<FeatureRecommendation>)
                    record);
            }
        }
        count++;
        average += penalty / ((ArrayList) obj).size();
    }
}

```

```

        penalty = 0;
    }
    centroids = new ArrayList();
    return average ;//+ 2*k;
}

public Object getClusterentroids() {
    return centroids;
}

public ArrayList getOptimumClusters(ArrayList optimumCentroids,
ArrayList<FeatureRecommendation> historic_records) {
    ArrayList cluster = new ArrayList<>(), randomClusters = new ArrayList(),
        cluster_errors = new ArrayList<>(), records = new ArrayList<>(), record = new
ArrayList<>();
    records = getOrganisedRecords(historic_records);
    boolean exceptionOccured = false;
    double penalty = 0;
    int counter = 1, distance = 0, minPosition;
    Random random = new Random();
    int randomIteration, pos;
    Iterator iterator = records.iterator();

    while (iterator.hasNext()) {
        record = (ArrayList) iterator.next();

        for (Object cr : optimumCentroids) {
            penalty += Math.pow(Math.abs(((ArrayList) cr).size() - record.size()), 3);
            for (Object c : (ArrayList) cr) {
                penalty += ((Cluster) c).computePenalty2(record);
            }
            cluster_errors.add(penalty);

```

```

        penalty = 0;
    }
    minPosition = getMinimumPosition(cluster_errors);
    cluster_errors = new ArrayList();
    for (Object r : record) {
        ((FeatureRecommendation) r).setClusterNumber(minPosition + 1);
    }
}

return records;
}

public ArrayList getBestCluster(ArrayList inputRecord, ArrayList optimumCentroids) {
    double penalty = 0;
    ArrayList cluster_errors = new ArrayList();
    for (Object cr : optimumCentroids) {
        penalty += Math.pow(Math.abs(((ArrayList) cr).size() - inputRecord.size()), 3);
        for (Object c : (ArrayList) cr) {
            penalty += ((Cluster) c).computePenalty2(inputRecord);
        }
        cluster_errors.add(penalty);
        penalty = 0;
    }
    return (ArrayList) optimumCentroids.get(Integer.parseInt("" +
getMinimumPosition(cluster_errors)));
}

public ArrayList findRecords(ArrayList bestCluster, ArrayList historicRecords) {
    ArrayList recordsInCluster = new ArrayList();
    for (Object obj : historicRecords) {
        for (Object obj2 : (ArrayList) obj) {
            if (((FeatureRecommendation) obj2).getClusterNumber() == ((Cluster)

```

```
bestCluster.get(0)).getClusterN()) {
    recordsInCluster.add(obj);
}
break;
}
}
return recordsInCluster;
}
```

```
public double getAverageDistance(ArrayList inputRecord, ArrayList
recordsInOptimuCluster) {
    double penalty = 0, distance = 0;

    for (Object cr : recordsInOptimuCluster) {
        penalty += Math.pow(Math.abs(((ArrayList) cr).size() - inputRecord.size()), 3);
        for (Object c : (ArrayList) cr) {
            penalty += ((FeatureRecommendation) c).computePenalty2(inputRecord);
        }
        distance += penalty;
        penalty = 0;
    }
    return Math.round(distance / recordsInOptimuCluster.size());
}
```

```
public ArrayList getCloseRecords(ArrayList inputRecord, ArrayList
recordsInOptimuCluster, double average) {
    double penalty = 0, distance = 0;
    ArrayList recordsBelowAverage = new ArrayList();
    for (Object cr : recordsInOptimuCluster) {
        penalty += Math.pow(Math.abs(((ArrayList) cr).size() - inputRecord.size()), 3);
        for (Object c : (ArrayList) cr) {
            penalty += ((FeatureRecommendation) c).computePenalty2(inputRecord);
        }
    }
}
```



```
    }  
    if (penalty <= average) {  
        recordsBelowAverage.add(cr);  
    }  
    penalty = 0;  
}  
return recordsBelowAverage;  
}
```

```
public ArrayList getAllRecommendations(ArrayList optimumRecords) {  
    String recommendations = "";  
    ArrayList recommendationList = new ArrayList();  
    String recommendation;  
    for (Object rec : optimumRecords) {  
        for (Object r : (ArrayList) rec) {  
            recommendations += ((FeatureRecommendation) r).getRecommendations();  
            break;  
        }  
    }  
    StringTokenizer tokenizer = new StringTokenizer(recommendations, ",");  
    while (tokenizer.hasMoreTokens()) {  
        recommendationList.add(tokenizer.nextToken());  
    }  
    return recommendationList;  
}
```

```
public ArrayList getDistinctRecommendations(ArrayList recommendations) {  
    String rec;  
    ArrayList newRecs = new ArrayList();  
    for (Object r : recommendations) {  
        rec = "" + r;
```

```
        if (!(newRecs.contains(rec))) {  
            newRecs.add(rec);  
        }  
    }  
    return newRecs;  
}
```

```
public ArrayList getFrequency(ArrayList distinctRecommendations, ArrayList  
recommendations) {  
    int count = 0;  
    ArrayList frequency = new ArrayList();  
    for (Object d : distinctRecommendations) {  
        for (Object r : recommendations) {  
            if (d.toString().equalsIgnoreCase(r.toString())) {  
                count++;  
            }  
        }  
        frequency.add(count);  
        count = 0;  
    }  
    return frequency;  
}
```

```
public ArrayList getDescriptions(ArrayList distinctRecommendationIds) {  
    int count = 0;  
    ArrayList desc = new ArrayList();  
    for (Object obj : distinctRecommendationIds) {  
        String desc = conn.getDescription((String) obj);  
        desc.add(desc);  
    }  
  
    return desc;  
}
```

Appendix B

Consent form

Date:

Title of the research project: **Integration of Indigenous Knowledge, Satellite Imagery, Climate Data and Machine Learning to optimise cropping decisions by small scale farmers. Case of uMgugundlovu district, Kwazulu-Natal South Africa.**

Student name and surname: **John Nyetanyane**

Student number:

Student contact number:

Supervisor(s): Prof M Masinde

Co-supervisor(s): Prof M Tafadzwa

Introduction to the study

The Department of Information Technology at the Central University of Technology (Free State, South Africa) is conducting research aimed at developing a cropping decisions' optimisation model to help local farmers in uMgugundlovu district to increase their crop yields. This is by way of integrating the indigenous knowledge with technology to foresee rains and onset of agricultural seasons to enable farmers to determine perfect times to perform crop cultivation. Further, the indigenous knowledge will be integrated with science and technology to assess the health and development of the crops.

Consent to participate in the research project

I, _____ (participant name), confirm to participation in this research project and that, the person asking my consent to take part in this research has introduced the study to me and highlighted its purpose.

I have read (or had explained to me) and understood the study as explained in the research study information and consent form. I have had an opportunity to ask questions and am prepared to participate in the study. I understand that my participation is voluntary and that I am free to withdraw at any time without penalty. I am aware that the findings of this study will be processed into a research report, journal publications, and/or conference proceedings. I am fluent in English and understand these documents or I have had the information explained to me in a language I understand well.

Full Name of Participant:

Signature of Participant: _____ Date:

Appendix 3

Collection of the Indigenous Knowledge Indicators in different domains

Indicators Nature	Warm Rainy Season Indicators (August to December)
Meteorological	Cold temperatures during warm dry season
	Hot temperatures during cold dry season
	Strong winds in August
	good rains at the beginning of the season
	Distributed clouds, clear sky, but heavy clouds appearing on the eastern side of the sky
	Low cloud after rains
	Strong haze(mist) in September
	Red sky in the west towards sunset
	Hot and damp conditions after rains
	Cold winds after rains
	bright sun but with no much heat output
	bright sun and very hot conditions in September
	Very strong North to west wind direction
	Presence of dew(water vapour) in the morning
	Delay of rains for a long time (about 2 weeks to a month)
	High temperatures before or after rains
	Heavy fall of rain before actual rains
Astronomical	Visible phases of the moon
	Dark phases of the moon
	Appearance of dark or small few clouds on the sky
	Pattern of stars forming hoe striking the ground
	Pattern of stars forming a roofing mat
	Dark clouds underneath and white clouds on top
	Disposition or Slanted position of the moon in September
	Presence of mist during the dry season
	When the sky is very clear
	When it is very sunny and hot
	Appearance of rain clouds at daytime and disappearance at night
	A very small circle around the sun and no circle around the moon
	A pattern of 7 stars appearing in the East
	Sky is full of stars at night
	Appearance of circle or circles around the moon and the sun
	Few stars in the sky
	Shading of the sun
Environmental(Birds)	The moon moves from East to West
	A pattern of stars remains at the central position
	Sighting of inkonjane birds at the beginning of rainy season
	Movement of bees and birds from east to west and north to south
	Appearance of many birds in groups
	Sighting of an owl at daytime at the beginning of the rainy season
	Birds named Speros singing
	When the bird named ijuba hang nests low the tree
	When the bird named ijuba hang nests very high the tree
Environmental(Insects)	Sighting of Large number of white and black birds named (Mashohori) at the beginning of the season
	A black-brown bird named umgugwane start singing during the dawn
	When the red chested bird start singing in the afternoon at the beginning of the rainy season
	Appearance of dragon flies, white butterflies and other insects at the beginning of rainy season
	Hasty spiders struggling to get indoors or into hiding places
	Invading of locusts
	Ants moving in a straight line

	Sighting of black and brown ants collecting food in large numbers in the house
	Ant named inkuba singing in large numbers
	Appearance of red ants named umkothane
	Appearance of termites (flying insects)during the rainy season
	Appearance of army worms on trees towards the beginning of the rainy season
	Occurrence of large number of grass-green grasshoppers during the rainy season
	Presence of too much honey in the hives indicates
	Appearance of insects and birds in less number
	Disappearance of certain birds/insect species (e.g. weaver birds)
	Millipede, ilulwane(bat) and crickets making noise at night
	Appearance of certain ants such as louse(umkhuphe), aphids(i-afidi)
	Invasion of locusts
Environmental(Animals)	Frogs making lots of sound
	Frogs fail to make noise
	Mating of local animals
	Appearance of white frogs in the grass
Environmental(Trees)	When the umhlaba tree starts to flower and generate new leaves at the beginning of warm rainy season
	Germinating of plants
	When the peach tree gives a lot of sap during the dry season
	The shooting of sausage tree
	Significant flowering and good fruit bearing of umvumvu tree
	Unusual shedding of the flowers of apricot tree
	Flowering of the plant named umphafa before onset of rains
	When the leaves of umnqumo plant turns yellow at the beginning of the rainy season
	Good harvest of certain trees

Appendix D

Pool of questions prepared for focus group discussion

1. How does the observation of IK indicators impact your certainty of the anticipated season behaviour?
2. How do you know that it is time to plant?
3. How do you estimate crops to cultivate for the predicted season behaviour?
4. How do you estimate the arrival of a high crop insect population concerning the anticipated season behaviour?
5. What crops do you cultivate during warm rainy and cold rainy seasons?
6. What crop, weather and other external properties are you observing to determine the health of a crop
7. How do you know that it is time to apply fertilisers?
8. How will the knowledge of anticipated rains and temperature assist in evaluating the health of a crop?
9. How will the knowledge of rain predictions assist in applying pesticides and fertilisers to your crops? For instance, if I know that rain is not coming soon, can I continue to apply fertilisers and pesticides to my crops?
10. What organic fertilisers do you apply to your crops and soil?
11. What symptoms observable on a tomato crop or any other crop that needs fertilisers?
12. What are the symptoms of crops suffering from over-fertilisation, heat stress or water imbalance (overwatering or underwatering)?
13. How can one help the small sorghum crop or any other crop that is suffering from over-fertilisation or heat stress?
14. How do you know it is time to apply pesticides?
15. What properties do you use to assess the health of the soil?
16. What organic methods do you use to lessen the use of chemicals in treating crop intruders (weeds, insects and diseases) and improving soil fertility?
17. What kind of crops can be cultivated on clay/sandy/loam soil when there is an expectation of below/normal/above normal rains throughout the season?
18. How will you tell if the crop in question is suffering from bacterial or fungi or virus disease?
 - a. How will you help that particular crop?

19. What kind of crops that can be cultivated if normal rains are expected to start late during the rainy season (for instance late December)?
20. How can I help the development of my small crops given that there is an expectation of above or below-normal rains coming soon
21. How can I help the development of my small crops that start to show signs of leaf wilting given that there is an expectation of no rains and higher temperatures in the coming weeks?
22. How can I help the development of my small crops that are cultivated on sandy soil given that there is an expectation of no rains and higher temperatures in the coming weeks?
23. What can be done to help the small maize crop that is showing signs of leaf wilting given that the crop is exposed to hot temperatures, low wind speed and high humidity?
24. What can be done to help the 4-weeks-old pumpkin crop that is showing signs of leaf wilting, leaf swelling and leaf discolouration?
25. What can be done to improve the growth of potato crops and other crops that are suffering from stunted growth (slow growth)?
26. What can be done to improve the growth of soybeans, or any other crop given that they are exposed to hot temperatures with low humidity and low wind speed? Further, the crop is cultivated on loam soil that is red and compacted.
27. How can you tell if the given crop problems are attributed to disease attacks rather than external factors such as heat?