Developing an Early Warning System for Banana Xanthomonas Wilt (BXW) in Rwanda

Caroline Akinyi Owuor

Adm. No. 149457

Supervisor

Dr. Kennedy Senagi

Submitted in Partial Fulfilment of the Requirements of the Master of Science in Data Science and Analytics at Strathmore University

Institute of Mathematical Sciences and iLab Africa

Strathmore University

Nairobi, Kenya

April 2024

Declaration and Approval

I confirm that this dissertation report has not been previously submitted for a degree, either at this institution or any other academic institution. To the best of my knowledge and belief, the research report does not contain any material that has been previously published or authored by another individual, except where proper acknowledgment is provided within the report.

Student: Caroline Akinyi Owuor

Student Number: 149457

Student Signature:  Date: April 05, 2024

This dissertation report has been reviewed and approved by Dr. Kennedy Senagi.

Supervisor Signature:  Date: 5th April 2024

Acknowledgment

I would like to express my sincere appreciation to my dissertation supervisors, Dr. Kennedy Senagi and Dr. Louise Leroux, for their invaluable guidance, support, and encouragement throughout this research. Their expertise and insightful feedback have been instrumental in shaping the direction and quality of this dissertation.

I am also grateful to Strathmore University and the Consortium of International Agricultural Research Centers (CGIAR) Excellence in Agronomy Initiative for providing the resources and conducive environment necessary for carrying out this study, and to ICT4BXW that provided the data for use in this research. Your contribution has greatly enriched the findings and analysis presented in this dissertation.

Lastly, I extend my gratitude to everyone who has played a role, big or small, in completing this dissertation. Your assistance and encouragement, whether from family, friends, or others, have been invaluable to me.

Abstract

Bananas are crucial for the agricultural economy of the African Great Lakes region, including countries like Kenya, Uganda, Tanzania, Burundi, Rwanda, and parts of the Democratic Republic of the Congo, with an annual production exceeding 22 million tonnes. However, banana productivity faces significant threats from pests and diseases such as the Banana Xanthomonas Wilt (BXW), caused by the *bacterium Xanthomonas campestris pv. Musacearum*. In this study, machine learning techniques were employed to develop an early warning system for BXW. Various classification models, including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), and Gradient Boosting Machine (GBM), were trained and evaluated for predicting BXW occurrence. RF outperformed the other models with an accuracy of 94%, followed by GBM (89%), KNN (87%), and SVM (83%). In terms of the area under the curve (AUC), RF outperformed the other models with a score of 96%, followed by GBM (95%), KNN (94%), and SVM (90%). This highlights RF's effectiveness in creating habitat suitability maps and establishing an early warning system for BXW. The RF model was used to develop a BXW habitat suitability map for Rwanda, aiding agricultural stakeholders in identifying high-risk areas. Furthermore, a Short Message Service (SMS)-based early warning system was implemented to provide timely alerts to farmers, thereby, enhancing BXW mitigation efforts. Additionally, a web portal for real-time BXW risk prediction and analysis was developed, providing accessible information to stakeholders for proactive management strategies.

Keywords: BXW, Early Warning System, Rwanda, Remote Sensing, Machine Learning.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

API Application Programming Interface

AUC Area Under Curve

BBTV Banana Bunchy Top Virus

BFW Banana Fusarium Wilt

BPNN Back Propagation Neural Networks

BXW Banana Xanthomonas Wilt

CGIAR Consortium of International Agricultural Research Centers

CRISP-DM Cross-Industry Standard Process for Data Mining

CRS Coordinate Reference System

CSV Comma Separated Values

CSS Cascading Style Sheets

DOM Document Object Model

DRC Democratic Republic of the Congo

EDA Exploratory Data Analysis

ERD Entity Relationship Diagram

EWS Early Warning System

GIZ German Corporation for International Cooperation GmbH

GBM Gradient Boosting Machine

HTML Hypertext Markup Language

ICT Information and Communications Technology

IITA International Institute of Tropical Agriculture

JS JavaScript

KNN K-Nearest Neighbors

ML Machine Learning

Maxent Maximum Entropy

NASA National Aeronautics and Space Administration

NDVI Normalized Difference Vegetation Index

NN Neural Network

PCA Principal Component Analysis

RAB Rwanda Agriculture and Animal Resources Board

RF Random Forest

SAR Synthetic Aperture Radar

SDG Sustainable Development Goals

SDM Species Distribution Models

SMS Short Message Service

SRTM Shuttle Radar Topography Mission

SVM Support Vector Machine

TIFF Tag Image File Format

UAV Unmanned Aerial Vehicles

UML Unified Modeling Language

URL Uniform Resource Locator

Chapter 1: Introduction

1.1 Background

In line with the 2022 Sustainable Development Goals (SDG) report, over 30% of the global population lacked consistent access to an adequate food supply in 2020. During this period, between 720 million and 811 million individuals worldwide experienced hunger, marking an increase of approximately 161 million people compared to the previous year (UN, 2022). With the global population on the rise, there is a corresponding increase in the demand for food (Kilwenge et al., 2023).

Small-scale farmers constitute a dominant force, comprising 84% of the global farming community and contributing 32% to the world's food supply, thus playing a crucial role in ensuring food security for human populations (Ritchie, 2021). Despite utilizing only about 24% of the world's agricultural land, these smaller farms, known for their efficiency in achieving higher yields, manage to produce a slightly larger quantity of crops compared to the land they occupy (Ricciardi et al., 2021). Hence, there is an urgent need to enhance the agricultural productivity of small-scale farmers to meet growing global food demands (Kilwenge et al., 2023).

The ongoing threats posed by pests and diseases persistently hinder agricultural productivity, with infectious crop diseases capable of causing substantial declines in yields, thereby negatively affecting the socioeconomic status of developing nations (Vurro et al., 2010). In certain instances, pests and diseases can result in total yield losses of up to 100%. Thus, it is essential to maintain vigilant monitoring of pests and diseases. Additionally, the implementation of effective control measures becomes imperative to curb their spread and mitigate their adverse impacts (Vurro et al., 2010).

Bananas play a vital role, serving as both a primary source of income and sustenance for many farmers in the African Great Lakes Region, while also holding significant cultural importance within the local population (Damme et al., 2014). Rwanda distinguishes itself as a prominent banana producer in the Eastern and Central African region, with smallholder farmers contributing a significant 90% to its total output. Notably, it holds the distinction of being the world's second-largest consumer of bananas, with individuals consuming an average of approximately 144kg annually. A considerable percentage, roughly 32%, of Rwandan households heavily rely on bananas as a staple crop, constituting nearly half of their dietary intake (Jackson et al., 2015).

Beyond their nutritional significance, bananas hold profound cultural importance in Rwandan society. They are exchanged as gifts, enjoyed in the form of banana wine during cultural celebrations, used for ornamental purposes at weddings and public events, and serve as a source of financial stability during community crises (Rietveld et al., 2020).

A significant contributor to the decline in banana productivity is Banana Xanthomonas Wilt (BXW), a bacterial disease caused by the *bacterium Xanthomonas campestris pv. Musacearum* (Jackson et al., 2015). The emergence of BXW was initially reported in 1968 in Ethiopia and subsequently spread to other countries within the African Great Lakes Region (McCampbell et al., 2018b). This disease inflicts significant damage on banana-centric farming systems due to its rapid and easily transmissible nature within plants, the absence of cultivars resistant to the disease, and ultimately, the death of affected plants (Ainembabazi et al., 2015).

Despite coordinated endeavors to address it, BXW has remained an ongoing challenge in Rwanda since its initial identification in 2002. The Rwanda Ministry of Agriculture is particularly interested in pioneering approaches that can provide a enduring resolution to BXW (McCampbell et al., 2018a). Information gleaned from (McCampbell et al., 2018a) reveals that 67% of farmers have faced BXW on their farms, with 59% of these instances occurring within the past year (McCampbell et al., 2018a).

Several technologies have been introduced in the African Great Lakes Region to mitigate the spread of BXW. These strategies include either the complete removal of infected plants or a more moderate approach involving the removal of only the affected stems (Blomme et al., 2017). While these technologies have made significant contributions to reducing the impact of BXW, there remains a need for further innovation in preventive measures (Rietveld et al., 2020).

1.2 Problem Statement

Promptly identifying and responding to outbreaks of BXW in Rwanda is a challenge, especially for smallholder banana farmers. This calls for a need to develop an early warning system to predict the risk of occurrence of BXW for purposes of preparedness, control, and mitigating outbreaks. Such measures can reduce crop loss and improve food security.

1.3 Research Aim

The primary focus of this study was to synthesize existing knowledge on Machine Learning (ML)-based applications for BXW risk surveillance and control, with a particular emphasis on its relevance to smallholder farmers in Rwanda.

1.4 Research Objectives

This research aimed to address the following objectives:

(a) To conduct a review of literature to identify existing ML-based approaches for BXW risk surveillance and control, highlighting any gaps in current research.

(b) To develop and test the effectiveness of ML-based models tailored for early warning on the risk of BXW among smallholder farmers in Rwanda.

(c) To evaluate the performance and reliability of the developed ML models.

(d) To deploy the models and visualize BXW risk maps on a web browser.

1.5 Research Questions

The research questions addressed in this study were as follows:

(a) What are the existing ML-based approaches for BXW risk surveillance and control, as reported in literature, and what gaps are there in current research?

(b) How can ML-based models be effectively developed and tested to provide early warning on the risk of BXW among smallholder farmers in Rwanda?

(c) What methods are employed to test and evaluate the performance and reliability of developed ML models for BXW risk surveillance and control under real-world conditions?

(d) How can the deployment of ML models and visualization of BXW risk maps on a web browser contribute to enhanced BXW surveillance and control efforts?

1.6 Scope and Limitations of the Study

1.6.1 Scope

The study included the following key components:

(a) Identification and selection of remote sensing data: An evaluation of spatial and temporal data was conducted to select the most suitable remote sensing information for early detection of BXW.

(b) Construction of a probability of occurrence map: Species Distribution Models (SDM) alongside ML models were employed to construct a map indicating the likelihood of BXW occurrence based on prevailing environmental conditions.

(c) Model assessment: The developed models were applied to a validation dataset, and accuracy metrics were extracted to assess their performance.

1.6.2 Limitations

Here are some of the limitations of this research work:

(a) Operationalizing research findings: Incorporating the findings of the study into BXW management in Rwanda posed a challenge, as it required engagement with the Rwanda Ministry of Agriculture and other stakeholders. This limited the effective translation of research findings into practical applications for managing BXW.

(b) Reliance on self-reported data: Another limitation of the study stemmed from its reliance on self-reported data collected through the ICT4BXW app, as detailed in Chapter 3, Section 3.2.1. This dependence introduced the possibility of response bias, which could have compromised the accuracy of the study's findings.

1.7 Research Justification

This research was crucial for agricultural stakeholders, including smallholder farmers, policymakers, and agricultural extension services, as it addressed a pressing issue impacting banana production and food security in Rwanda. By developing an effective Early Warning System (EWS) for BXW, this work provided practical solutions to mitigate the economic and social consequences of the disease, thereby enhancing the resilience of agricultural systems and livelihoods in the region.

Chapter 2: Literature Review

The literature review was structured around four main themes: the impact of BXW in East and Central Africa, environmental factors influencing plant diseases and pests, applications of remote sensing and ML in banana cropland mapping, and applications of remote sensing and ML in plant diseases and pests. These themes provided a comprehensive overview of past research efforts in understanding and addressing agricultural challenges, particularly focusing on disease management and technological innovations in crop monitoring and surveillance. Understanding these themes was essential for developing an effective ML-driven early warning system, as it required insights into disease dynamics, environmental influences, and advanced technologies for accurate prediction and timely intervention in agricultural settings.

2.1 Effects of BXW in Eastern and Central Africa

Findings from a household survey carried out in the Kagera region of Tanzania, coupled with investigations across 16 provinces of Burundi and 12 districts in Rwanda from 2009 to 2011, highlight a considerable decrease in banana yields attributed to the prevalence of BXW (Jackson et al., 2015). This decline resulted in a notable 35% reduction in banana production following the emergence of BXW. As a result, the affected countries encountered a significant economic setback, with losses amounting to $14.05 million in 2012 (Jackson et al., 2015).

The potential of BXW to result in complete yield losses, reaching up to 100%, further underscores its substantial threat to both food security and the livelihoods of smallholder farmers (Blomme et al., 2017). From Bloomme's work, the presence of BXW emerged as a significant challenge to Rwanda's food security, particularly impacting households where 32% relied on bananas for more than half of their daily dietary needs. The adverse effects of the disease were evident, with 38% of survey participants reporting disruptions in their household diets due to BXW (Jackson et al., 2015).

BXW has effects that transcend banana production alone; it also significantly impacts the entire banana value chain. An example of this is evident in a case study conducted in Central Uganda, which examined how BXW influences the banana-beer value chain. The study revealed that a substantial number of individuals derive their livelihoods not only from cultivating bananas but also from processing and selling banana-beer products (Rietveld et al., 2020).

2.2 Environmental Factors Influencing Plant Diseases and Pests

Research has indicated that the inclusion of diverse environmental variables, sourced from various outlets, enhances the performance of SDM (Araújo et al., 2019). Combining both climate-related factors and non-climate variables such as land cover and topography has been demonstrated to improve the accuracy and effectiveness of SDM. By integrating multiple types of variables, SDM can capture a broader range of ecological influences, leading to more robust predictions of species distributions (Burns et al., 2020).

Various methods have been utilized for selecting relevant environmental variables to use in SDM (Lin and Chiu, 2020). Minimizing collinearity between variables is essential in this process. Taking into account the correlation between variables helps ensure the robustness and accuracy of SDM (Pradervand et al., 2014).

In the realm of predictive modeling, understanding the influential variables is paramount for accurate forecasts. (Domingues et al., 2022) investigated the weather variables that impact the forecast performance, shedding light on their significance. Conversely (Lasso et al., 2020) identified the optimal time period window for each weather variable and crop-related feature crucial for predicting the occurrence of coffee leaf rust disease in coffee crops.

(Small et al., 2015) developed a web-based tool using weather data, crop resistance information from literature and field trials, and management tactics to predict disease outbreaks, particularly late blight in tomato and potato crops.

Deutsch et al. (2018) underscores temperature as a key driver of insect development, significantly impacting both their metabolic rate and population growth. Through their findings, it becomes evident that temperature variations play a pivotal role in shaping insect biology and population dynamics. Understanding this relationship is essential for comprehending insect ecology and devising effective pest management strategies (Deutsch et al., 2018).

The interplay of elements such as soil composition, humidity levels, rainfall patterns, and moisture content has been identified as significant factors influencing crop diseases and yields. These environmental factors play crucial roles in shaping the overall health and productivity of crops, affecting various stages of growth and development (Patil et al., 2019).

Several studies employing regression models and weather data have elucidated the significant impact of humidity on disease and pest development. Through comprehensive analyses, these

studies have revealed the intricate relationship between humidity levels and the proliferation of diseases and pests in agricultural ecosystems (Xiao et al., 2019).

Below is a summary table outlining the main environmental factors identified to contribute to SDM in the research reviewed:

Table 1: Main environmental factors contributing to SDM

Environmental Factor	**Description**
Temperature	Influences physiological processes and species distributions
Precipitation	Affects water availability and plant growth
Soil Type	Determines nutrient availability and root growth
Land Cover	Influences habitat suitability and species interactions
Topography	Influences microclimate and habitat structure

2.3 Applications of Remote Sensing and ML in Banana Cropland Mapping

A study by (Selvaraj et al., 2020) conducted in the Democratic Republic of the Congo (DRC) and the Republic of Benin showcased the effectiveness of integrating high-resolution satellite imagery (including Sentinel 2, PlanetScope, and WorldView-2) with Unmanned Aerial Vehicles (UAV) equipped with Sense RedEdge sensors and ML-based mobile applications for detecting and establishing a surveillance system for mapping banana plants and associated diseases. The study highlighted the superiority of employing a Random Forest (RF) model for pixel-based banana classification, especially when incorporating vegetation indicators and Principal Component Analysis (PCA). This approach yielded improved outcomes in mapping bananas across diverse African environments. Furthermore, the study emphasized the superior performance of high-resolution sensors compared to medium-resolution satellites, particularly in scenarios involving mixed cultivation (Selvaraj et al., 2020).

In a study conducted in Uganda's East African highlands, the goal was to map the distribution of banana crops and assess changes in production zones spanning the period from 1958 to 2016. To collect data on the geographical locations of banana plantations, an online survey based on high-resolution satellite imagery was utilized. These satellite images were combined with independent covariates and subsequently input into ensemble ML models, which were employed to predict the current distribution of banana crops. The ensemble model was constructed using training results from RF, Gradient Boosting Machine (GBM), and Neural Network (NN). The study revealed that the number of covariates incorporated into the ensemble model had

a discernible impact on its performance. The ensemble models consistently outperformed the individual methods when trained with both 12 and 17 variables, achieving higher Area Under Curve (AUC) values (AUC = 0.895 and 0.907, respectively), in contrast to RF (AUC = 0.883 and 0.901), GBM, and NN (AUC = 0.870 and 0.890) (Ochola et al., 2022).

In a study conducted by (Alabi et al., 2022) in Ogun State, Nigeria, the application of remote sensing and ML models for identifying banana farms and conducting targeted surveillance of Banana Bunchy Top Virus (BBTV) occurrence was investigated. The study established a comprehensive framework for detecting bananas in smallholder agricultural systems, utilizing UAV, Sentinel 2A optical, and Synthetic Aperture Radar (SAR) data. UAV images were used to create spectral ortho mosaics and models for digital surface, digital terrain, and canopy height. Vegetation indices were derived from the UAV's spectral features, and RF and Support Vector Machine (SVM) models were developed, both with and without canopy height information, to differentiate bananas from other land cover types. The ML models achieved an average accuracy score of 93% when incorporating vegetation height features, while models without canopy height attained a lower accuracy score of 78%. Consequently, the study concluded that structural height attributes are essential for crop delineation when utilizing UAV-based predictors (Alabi et al., 2022).

Analysis through a logistic regression model revealed several significant factors influencing banana distribution. Notably, yearly precipitation, bulk density, soil organic carbon, soil pH, and slope gradient were positively correlated with banana distribution, while mean annual temperature and precipitation seasonality exhibited a negative association (Ochola et al., 2022). Furthermore, the most crucial variables connected with geographical variations were biophysical attributes tied to soil water availability. This underscores the importance of irrigation and soil water conservation as essential measures to alleviate the consequences of climate change-induced temperature increases and dry spells (Ochola et al., 2022).

In a study focusing on Costa Rican banana crops, researchers employed a fixed-wing UAV equipped with sensors to analyze photosynthetic activity patterns. These patterns, identified through the Normalized Difference Vegetation Index (NDVI), were then compared with both soil quality and banana fruit production patterns. The study unveiled notable positive correlations between NDVI patterns and several fruit yield and quality parameters, such as bunch weight, number of hands per bunch, length of the largest finger, and overall yield. Surprisingly,

NDVI also demonstrated a significant negative correlation with the proportion of rejected bananas due to poor quality. However, no statistically significant relationship was detected between NDVI patterns and physical soil quality patterns. These findings underscore the potential of employing UAV systems within banana plantations to map fruit quality and yield trends, providing valuable insights into the spatial dynamics of production factors and ultimately improving agricultural efficiency (Machovina et al., 2016).

In summary, the literature predominantly featured the utilization of RF, SVM, GBM and Logistic Regression as the main ML models for predicting species distributions. The primary evaluation metric across these studies was the AUC.

2.4 Application of Remote Sensing and ML for Plant Pests and Diseases Surveillance

Globally, agriculture and forestry are grappling with formidable challenges stemming from plant diseases and pests. There is a pressing need for the widespread adoption of economical, effective, and remote techniques to detect and monitor these issues over extensive geographical regions. Such methods have the capacity to greatly enhance efforts in safeguarding plant health and productivity (Zhang et al., 2019).

A study was conducted with the aim of improving the detection and monitoring of Banana Fusarium Wilt (BFW). The research aimed to develop optimal classification models for different infection stages by analyzing the spectral characteristics of banana canopies affected by BFW. Multispectral imagery was obtained from a banana plantation with BFW infections using a RedEdge-MX camera mounted on an UAV. Three visible-band images, five multispectral-band images, and various vegetation indices were utilized as distinct spectral features. To identify BFW-infected canopies, several supervised techniques, including SVM, RF, Back Propagation Neural Networks (BPNN), and logistic regression, were employed, along with unsupervised techniques such as hotspot analysis and the Iterative Self-Organizing Data Analysis Technique Algorithm (ISODATA). Classification results demonstrated outstanding accuracy for most techniques. The best-performing model among supervised techniques achieved an overall accuracy of 97.28% within a shorter timeframe, using RF with five multispectral bands of data. Hotspot analysis, utilizing specialized indices derived from the red and Near-Infrared (NIR) bands, achieved excellently balanced accuracies exceeding 95% with unsupervised approaches. For slightly enhanced accuracy, the study recommended utilizing hotspot analysis

for identifying BFW, particularly in late-stage infections, and RF for early-stage infections. This research offers valuable methodologies for effective plant disease detection and carries significant implications for banana plantation management (Zhang et al., 2022).

In a study conducted by (Ye et al., 2020), researchers utilized UAV to conduct surveillance on banana crops for BFW. They employed multispectral imagery captured by UAVs to differentiate between banana patches afflicted by the disease and those that remained disease-free. The investigation highlighted the effectiveness of employing UAV-based remote sensing, particularly with a focus on the red-edge band, which significantly improved the identification of Fusarium wilt. Additionally, the study delved into the impact of different spatial resolutions by conducting simulations with satellite imagery. Various spatial resolutions from different satellite systems were examined, including 0.5m for the WorldView series, 1m for GF-2, 5m for RapidEye, and 10m for Sentinel-2. Notably, the study found that higher spatial resolutions, specifically those of 2 meters and above, exhibited greater accuracy in disease recognition, owing to the spacing of the banana plants (Ye et al., 2020).

In their study, (Kilwenge et al., 2023) conducted an evaluation of BXW) risk in Rwanda under both current and projected climatic conditions. The researchers utilized the Maximum Entropy (Maxent) model, analyzing 1,022 georeferenced sites and 20 environmental variables to predict the occurrence of BXW. The study achieved a mean validation AUC ranging from 0.79 to 0.85 in predicting BXW occurrence. Notably, elevation, the average maximum monthly temperature, and precipitation during the coldest quarter emerged as the most reliable predictors. The research identified Rwanda's western, northern, and southern regions, characterized by elevations between 1350-2000 meters, annual precipitation of 900-1700 mm, and average temperatures ranging from 14-20°C, as the most vulnerable areas to BXW risk (Kilwenge et al., 2023).

Below is a summary table showcasing the SDM and ML models utilized in the reviewed studies, along with the corresponding evaluation metrics.

Table 2: Summary of ML models and evaluation metrics used in reviewed studies

Study	ML Models Used	Evaluation Metrics
Selvaraj et al. (2020)	RF	Accuracy
Alabi et al. (2022)	RF and SVM	Accuracy
Ochola et al. (2022)	Ensemble, RF, GBM and NN	AUC
Zhang et al. (2022)	SVM, RF, BPNN and Logistic Regression	Accuracy
Kilwenge et al. (2023)	Maxent	AUC

2.5 Summary

The literature discussed environmental factors that significantly influence the occurrence of pests and diseases in agricultural ecosystems. Variables such as temperature, humidity, rainfall, and soil moisture were identified as important indicators for modeling and predicting outbreaks. Various ML models commonly employed for this purpose included SVM, RF, and Logistic Regression, each offering unique advantages in terms of predictive accuracy, computational efficiency, and interpretability.

While the literature provided valuable insights into the choice of ML models and environmental variables it was better to explore an array of ML algorithms due to the broader range of techniques and approaches available within this realm. In this study, we leveraged the ML classification models identified in the literature, specifically RF, SVM, and GBM. Additionally, the literature guided us in selecting relevant environmental variables, highlighted in Table 4, to incorporate into our predictive models.

Through our research, we endeavored to address existing gaps in pest and disease prediction methodologies, with a specific focus on improving the accuracy and applicability of predictive models in agricultural settings. The aim was to develop a human-centered early warning system driven by ML. By leveraging insights from previous studies and integrating novel approaches, our research aimed to enhance our understanding of pest and disease dynamics, ultimately contributing to more effective and sustainable agricultural practices.

Chapter 3: Methodology

The study followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, a widely adopted framework for ML and data mining projects. CRISP-DM comprises several essential phases organized in an iterative process.

1. **Business Understanding:** During this stage, objectives and business goals are defined to ensure alignment with desired outcomes.
2. **Data Understanding:** This phase involved collecting and examining the data utilized in the project, with the aim of obtaining a thorough understanding of the dataset.
3. **Data Preparation:** Activities like feature engineering and selection were undertaken to prepare the data for modeling.
4. **Modeling:** Diverse models are employed on the prepared data to construct predictive or descriptive models.
5. **Evaluation:** The fitted models undergo evaluation to gauge their accuracy and performance in alignment with the project's objectives.
6. **Deployment:** After obtaining a satisfactory model, it is deployed and integrated into the appropriate business processes to extract practical insights and facilitate decision-making.

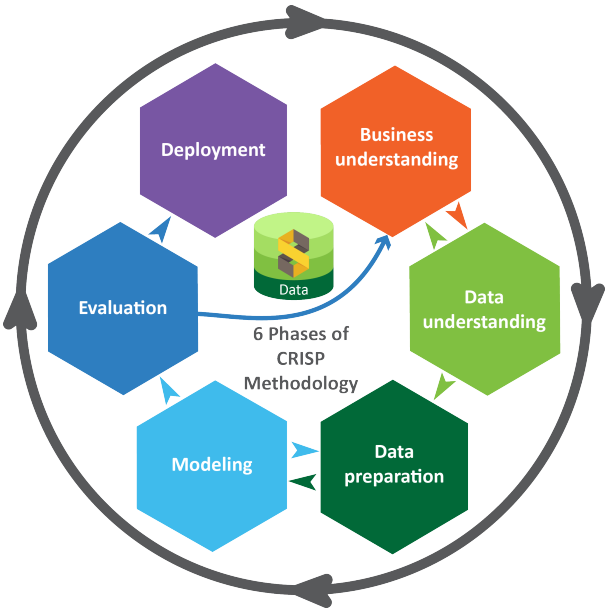


Figure 3.1: CRISP-DM framework (RuchaReads, 2021)

3.1 Business Understanding

The business understanding stage of this dissertation centered on addressing the need for an effective EWS to mitigate the impact of BXW in Rwanda's agricultural sector. Various aspects of the problem required understanding to achieve this goal. The key components of the business understanding included:

1. **Identifying Stakeholders:** It was crucial to understand who stands to benefit from this work. Reviewing the banana production ecosystem and understanding the players was important for developing a human-centered tool. Stakeholders such as smallholder farmers, agricultural extension officers, and the Rwanda Ministry of Agriculture tend to benefit from the early warning system, as highlighted in the Unified Modeling Language (UML) diagram represented in Figure 4.1.

2. **Defining Objectives:** Clear research objectives were defined to guide the research. These objectives have been outlined in Chapter 1, Section 1.3.

3. **Determining Requirements:** Informed by the research objectives, the specific requirements of the EWS were outlined. System designs and architecture were clearly established to provide an overview of the system even before implementation. Low-fidelity wireframes and an Entity Relationship Diagram (ERD) were developed to visualize the web portal and the database. These visualizations can be found in Chapter 4, Section 4.2.2 and Figure 4.2 respectively.

Overall, the business understanding stage provided a comprehensive understanding of the problem domain, the needs of stakeholders, and the objectives and requirements for developing a human-centered EWS for BXW in Rwanda.

3.2 Data Understanding

In the data understanding stage of this dissertation, we delved into the exploration and comprehension of two critical datasets: BXW occurrence data and environmental conditions. The BXW occurrence data provided insights into the historical occurrences of BXW within Rwanda, offering essential information for understanding the spatial and temporal distribution of the disease. Similarly, the environmental conditions dataset encompassed various environmental factors such as climate and topographic data, which were known to influence the spread and sever-

ity of BXW outbreaks, as highlighted in Chapter 2, Section 2.4. By exploring these datasets, we aimed to gain valuable insights into the factors contributing to BXW occurrences and enhance our ability to develop an effective EWS for BXW mitigation in Rwanda's agricultural sector.

3.2.1 BXW Occurrence

The data used in this study was collected using the ICT4BXW android application. The app was developed by the ICT4BXW [`https://ict4bxw.com/`] project, a collaborative effort supported by German Corporation for International Cooperation GmbH (GIZ) funding, involving partnerships between the Rwanda Agriculture and Animal Resources Board (RAB), International Institute of Tropical Agriculture (IITA), and various organizations such as IAMO, Viamo, ARIFU, and Linking Pin Africa. The project aimed to create and implement Information and Communications Technology (ICT)-based tools to enhance sustainable banana cultivation in Rwanda (Kilwenge et al., 2021).

A citizen science approach was used to collect the BXW occurrences. This approach involved engaging local community members, including smallholder farmers and agricultural extension officers, in the process of collecting data on BXW occurrences. Through training and equipping community members with mobile applications and necessary tools, they were empowered to identify and report instances of BXW-infected banana plants in their respective areas. By leveraging the knowledge and observations of individuals directly involved in banana cultivation, this approach not only increased the quantity of available data but also promoted community engagement and ownership of the research process.

The variables in the BXW occurrence data are highlighted in Table 3.

Table 3: BXW diagnosis data variables

Variable	**Data Type**	**Description**
Date of creation	DateTime	The date when the data entry was made
Gender	String	The farmer's gender, categorized as either male or female
District	String	A geographical administrative division within Rwanda
Sector	String	A smaller administrative unit within a district
Village	String	A small settlement or community within a sector
Cell	String	A smaller administrative subdivision within a village
Latitude	Float	Latitude coordinates of the banana farm
Longitude	Float	Longitude coordinates of the banana farm
HAS.BXW	Char	BXW diagnosis reported as either YES or NO

The data used in the study was collected between December 2019 and February 2024. The data contained 14, 118 records of BXW diagnosis collected across different regions in Rwanda. Exploratory Data Analysis (EDA) was performed to understand the spatial and temporal distribution patterns of the historical BXW records. The findings of this analysis are discussed in detail in Chapter 6, Section 6.1.1.

3.2.2 Environmental Factors

The comprehensive WorldClim dataset, comprising temperature and precipitation records across diverse geographical scales and temporal resolutions, served as a valuable resource for this study. Specifically tailored to Rwanda, this dataset was instrumental in generating bioclimatic data crucial for ecological modeling (WorldClim, 2022). These bioclimatic variables, derived from monthly temperature and rainfall data, provided physiologically relevant parameters commonly employed as predictive factors in SDM (Gardner et al., 2019). In this research, a total of 19 bioclimatic variables were utilized, accessed through the WorldClim's Application Programming Interface (API) on R. These raster files were acquired at a resolution of 30 seconds, corresponding to approximately 1km. Rwanda's climatic conditions are elaborated upon in Chapter 6, Section 6.1.4

In addition to bioclimatic data, topographic data obtained through remote sensing technology played a vital role in enhancing the accuracy of the SDM. Recent advancements in remote sensing have facilitated the acquisition of high-resolution environmental data, enabling the depiction of detailed species micro-habitats. By integrating topographic data with SDM, researchers can achieve improved prediction accuracy, particularly for fine-scale biodiversity management initiatives (Pradervand et al., 2014). Rwanda's topography has been discussed in Chapter 6, 6.1.3

The topographic data employed in the study included key variables sourced from WorldClim. Elevation, derived from Shuttle Radar Topography Mission (SRTM) elevation data, was obtained and served as the foundation for computing other topographic parameters.

Slope, a measure of the steepness of the terrain, was derived from the elevation rasters using Equation 1.

$$Slope = \arctan\left(\sqrt{\left(\frac{\partial Z}{\partial x}\right)^2 + \left(\frac{\partial Z}{\partial y}\right)^2}\right) \tag{1}$$

where $\frac{\partial Z}{\partial x}$ and $\frac{\partial Z}{\partial y}$ are the partial derivatives of elevation with respect to the x and y coordinates, respectively.

Aspect, indicating the direction in which the terrain faces, was calculated from the elevation data using Equation 2

$$Aspect = \arctan\left(\frac{\partial Z}{\partial y}, \frac{\partial Z}{\partial x}\right) \tag{2}$$

Where, $\frac{\partial Z}{\partial x}$ and $\frac{\partial Z}{\partial y}$ are the partial derivatives of elevation with respect to the x and y coordinates.

Hillshade, representing the shading of the terrain, was generated using both slope and aspect data according to the below equation:

$$Hillshade = \arctan\left(\frac{cos(Z_{zenith}) + cos(Slope) + sin(Slope) \times cos(Azimuth - Aspect)}{1 + \left(\frac{\partial Z}{\partial x}\right)^2 + \left(\frac{\partial Z}{\partial y}\right)^2}\right) \tag{3}$$

where, Z_{zenith} denotes the zenith angle, which indicates the angle between the sun's rays and the vertical direction. An azimuth, representing the direction measured in degrees clockwise from north on an azimuth circle, is denoted by $Azimuth$. Additionally, $Slope$ and $Aspect$ represent the computed slope and aspect values, respectively.

These topographic variables, combined with the bioclimatic data, played a crucial role in enhancing the accuracy and precision of the SDM developed in the study.

Table 4 provides a summary of the environmental data utilized in the research.

Table 4: Environmental predictors

Category	Code	Variable	Unit
Climatic	Bio1	Annual Mean Temperature	°C
	Bio2	Mean of monthly (max temp - min temp)	°C
	Bio3	Isothermality (BIO2/BIO7) (×100)	°C
	Bio4	Temperature Seasonality (standard deviation ×100)	°C
	Bio5	Max Temperature of Warmest Month	°C
	Bio6	Min Temperature of Coldest Month	°C
	Bio7	Temperature Annual Range (BIO5-BIO6)	°C
	Bio8	Mean Temperature of Wettest Quarter	°C
	Bio9	Mean Temperature of Driest Quarter	°C
	Bio10	Mean Temperature of Warmest Quarter	°C
	Bio11	Mean Temperature of Coldest Quarter	°C
	Bio12	Annual Precipitation	mm
	Bio13	Precipitation of Wettest Month	mm
	Bio14	Precipitation of Driest Month	mm
	Bio15	Precipitation Seasonality (Coefficient of Variation)	mm
	Bio16	Precipitation of Wettest Quarter	mm
	Bio17	Precipitation of Driest Quarter	mm
	Bio18	Precipitation of Warmest Quarter	mm
	Bio19	Precipitation of Coldest Quarter	mm
Topographic	Elev	Elevation	m
	SLP	Slope	°
	ASP	Aspect	°
	HS	Hillshade	°

3.3 Data Preparation

Data preparation played a crucial role in this dissertation, as it directly influenced the performance of the models by ensuring data quality and suitability. This phase encompassed various tasks aimed at transforming raw data into a format suitable for training ML algorithms. For the BXW occurrence data, tasks such as handling missing values, outliers, class imbalances, and feature scaling were conducted to ensure data cleanliness and balance, as discussed in Section 3.3.1. Similarly, to prepare the environmental raster files for species distribution modeling using ML, cleaning procedures were implemented to ensure consistency in extent and resolution across the raster files, as elaborated in Section 3.3.2.

3.3.1 BXW Occurrence Data

(a) **Feature Selection:**

A subset of the data was extracted, focusing solely on the latitude, longitude, and BXW diagnosis variables.

(b) **Handling Missing Values:**

Upon inspection, no missing values were detected in the dataset, ensuring data completeness and reliability for subsequent analysis.

(c) **Duplicate Values:**

Records containing duplicate coordinates were identified and removed, while retaining the first occurrence of the duplicate records to maintain data integrity and avoid redundancy.

(d) **Binary Encoding:**

Binary encoding was applied to the response variable (BXW diagnosis), wherein instances labeled as YES were encoded as 1, while those labeled as NO were encoded as 0, facilitating computational analysis.

(e) **Extent Adjustment:**

The dataset was refined by cropping and masking the data points to the geographical extent of Rwanda, excluding any occurrences outside the country's boundaries. Additionally, to enhance precision, the data points were restricted to Rwanda's cropland, ensuring relevance to the study's focus on agricultural contexts within the country.

(f) **Class Imbalance:**

The dataset underwent an examination to assess class imbalance, achieved through count plot visualization of YES (1) and NO (0) diagnoses. These findings are detailed in Chapter 6, Section 6.2.1. Subsequently, three techniques were explored to address this imbalance: Euclidean distance spatial undersampling, stratified spatial undersampling, and utilization of background points for the absence of data. The outcomes of each sampling method are elaborated upon in Chapter 6, Section 6.2.1.

In the Euclidean distance spatial undersampling technique, negative instances (NOS) were selected based on their distance from positive instances (YESES) using Euclidean distance metrics. This approach aimed to ensure a spatially balanced distribution of NOS

across the study area, thereby enhancing the representativeness of the dataset for training ML models. The results for the euclidean distance undersampling are discussed in Chapter 6, Section 6.2.1, Section (a).

For the stratified spatial undersampling method, predefined strata were established based on geographic units, particularly districts in the study area. Random NO points were then sampled from each district to maintain the distribution of different groups within the dataset and ensure its representativeness. The results for this approach are discussed in Chapter 6, Section 6.2.1, Item (b)

Lastly, the background points technique entailed discarding all NO records from the dataset and selecting random background points to represent the absence data. These background points were chosen randomly from various locations across the study area, ensuring a diverse representation of absence. This technique aimed to create a balanced dataset by providing sufficient negative instances for model training while mitigating class imbalance effects.

Given that the background absence technique yielded a more balanced and well-represented dataset as shown in Chapter 6, Section 6.2.1, Item (c), it was selected for implementation in the model development phase.

3.3.2 Environmental Rasters

To prepare the environmental data for analysis and modeling, several steps were undertaken as outlined below:

(a) **Acquisition of Rwanda Shapefile and Cropland Raster**

The Rwanda shapefile was obtained from the HumData portal [`https://data.humdata.org/dataset/cod-ab-rwa?`] to facilitate the cropping of various raster datasets, ensuring that the data aligns with the geographic extent of the study area.

Additionally, to confine the analysis and modeling to agricultural land, a cropland raster, as depicted in Figure 6.13 and defined by National Aeronautics and Space Administration (NASA) (Xiong et al., 2017), was employed to mask the occurrence predictions.

(b) **Retrieval and Cropping of Climate Data**

Climate data was retrieved directly from WorldClim using an R API. Subsequently, it was

cropped using the Rwanda cropland raster file to confine the data to the study area. The resolution was adjusted to 0.0133 by 0.0133, and the Coordinate Reference System (CRS) used was "$+proj = longlat + datum = WGS84 + no\_defs$". Individual bioclimatic variables were then extracted from the raster stack and saved for modeling purposes as shown in Figure 6.8.

(c) **Preprocessing of Elevation Data**

The elevation raster was sourced from the WorldClim website and imported into an R script for preprocessing. It underwent cropping and masking to match the Rwanda shapefile, and its resolution and CRS were updated to align with the climate rasters.

Bilinear resampling was applied to ensure consistency in extent between the climate and elevation data.

(d) **Extraction of Slope and Aspect Information**

The R terrain function was employed to derive slope information from the elevation data, utilizing 8 neighbors and degrees, as per the slope equation highlighted in Equation 1. Moreover, the terrain function was utilized to extract aspect information of the study region, employing Equation 2.

(e) **Generation of Hillshade Data**

Hillshade data was generated using the hillshade function, leveraging aspect and slope data as described in Equation 3. Additionally, specific parameters such as a 40-degree angle and 270 for direction were specified.

(f) **Standardization of Environmental Data**

Z scaling, also known as standardization, was employed to normalize the environmental data in preparation for ML algorithms. This process involves transforming the data so that it has a mean of 0 and a standard deviation of 1. The equation for Z scaling is given by Equation 4.

$$Z = \frac{x - \mu}{\sigma} \tag{4}$$

where Z is the standardized value, x is the original value, μ is the mean of the dataset and σ is the standard deviation of the dataset.

(g) **Extraction and Storage of Predictors**

All generated rasters were exported as Tag Image File Format (TIFF) files and stored locally for analysis and modeling purposes.

The environmental conditions for each diagnosis location were extracted from the previously obtained rasters, and the data was stored in a Comma Separated Values (CSV) format for further analysis and modeling.

3.4 Machine Learning Modeling

Before fitting the ML models, the preprocessed data was divided into training and testing sets, with 80% allocated for training and 20% for testing. The "diagnosis" column was transformed into a factor and designated as the target variable, while the environmental conditions were used as the predictors. A 10-fold cross-validation was applied across all models.

Four classification ML models were employed for modeling. The selection of these models was based on their significance in species distribution modeling, as discussed in Chapter 2, Section 2.4. The models included SVM, K-Nearest Neighbors (KNN), RF and GBM.

3.4.1 Support Vector Machine

SVM is a supervised learning algorithm used for classification and regression tasks. It works by finding the hyperplane that best separates classes in feature space. SVM aims to maximize the margin between the hyperplane and the nearest data points from each class. This is achieved by solving a constrained optimization problem, where the margin is maximized subject to the constraint that all data points are correctly classified (Tan, 2021).

The optimization problem for SVM can be formulated as shown in Equation 5.

$$\textit{minimize} : Q(w) = \frac{1}{2}||w||^2 \tag{5}$$

$$\text{subject to } y_i(wx_i - b) \geq 1,\ \forall (x_i, y_i) \in D$$

where w is the vector of coefficients of the hyperplane, $||w||^2$ is the length of the weight vector w

In essence, the goal of SVM is to maximize the margin, by minimizing $||w||^2$.

In this study, SVM radial was employed, and a validation plot was generated to evaluate its performance. The outcomes of this are elaborated upon in Chapter 6, 6.3.1. SVM radial is a popular algorithm used for classification tasks. It is particularly effective when dealing with non-linearly separable data by transforming the input space into a higher-dimensional space, where a linear decision boundary can be found (Ding et al., 2021). The radial basis function measures the similarity between two data points based on their distance from each other. This kernel function is defined as:

$$K(x_i, x_j) = \exp(-\lambda||x_i - x_j||^2) \tag{6}$$

where x_1 and x_j are data points, $||x_i - x_j||^2$ represents the Euclidean distance between them and λ is a parameter that determines the influence of each training example on the decision boundary

3.4.2 K-Nearest Neighbors

KNN is a simple and intuitive classification algorithm that works based on the principle of similarity. It classifies data points by identifying the majority class among their k nearest neighbors in feature space. The choice of k determines the level of smoothness in the decision boundary (Zhang et al., 2022).

Various distance metrics can be employed to compute distances in KNN, including Euclidean distance (Equation 7) and Manhattan distance (Equation 8), among others. Euclidean distance is a measure of the straight-line distance between two points in an Euclidean space, which is the shortest distance between them, as shown in Equation 7. Manhattan distance, also known as taxicab or city block distance, is a metric that calculates the distance between two points in a grid-based system. It is computed as the sum of the absolute differences between the coordinates of the points along each dimension (Suwanda et al., 2020), as shown in Equation 8.

$$d(p, q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{7}$$

where p, q = two points in Euclidean n-space

$$\text{Manhattan Distance} = \sum_{i=1}^{n} |x_i - y_i| \tag{8}$$

3.4.3 Random Forest

RF is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Each decision tree is built on a random subset of the training data and a random subset of the features. RF aims to reduce overfitting and improve generalization performance by averaging the predictions of multiple trees (Tan, 2021).

The RF classifier adapts the Gini Index, a metric used to assess the impurity of an attribute based on the classes present. Given a training set T, where a case is randomly selected and assigned to a class , the Gini index can be expressed using Equation 9 (Tan, 2021):

$$\sum\sum_{j \neq i} \left(\frac{fC_i, T}{|T|}\right) \left(\frac{f(C_i, T}{|T|}\right) \tag{9}$$

where $\frac{fC_i,T}{|T|}$ represents the probability that the selected case applies to the class C_i

3.4.4 Gradient Boosting Machine

GBM is another ensemble learning method that builds a sequence of weak learners (typically decision trees) in a forward stage-wise manner. Each new learner focuses on minimizing the residual errors of the previous learners. GBM combines the predictions of all learners to make the final prediction. Unlike RF, GBM builds trees sequentially and relies on the gradient descent optimization algorithm to minimize the loss function.

3.5 Machine Learning Model Evaluation and Optimization

In assessing the effectiveness of the models in classifying the occurrence of BXW, a variety of metrics were utilized, encompassing AUC, recall, precision, F1-score, and accuracy. These metrics, chosen based on their relevance and widespread use in plant disease classification studies as outlined in Chapter 2, Section 2.4, offered distinct insights into the models' performance. By utilizing a combination of these metrics, an evaluation of the models' effectiveness was attained, enabling an examination of their classification capabilities.

3.5.1 Accuracy

Accuracy measures the overall correctness of a model's predictions, representing the proportion of correctly classified instances (both true positives and true negatives) among all instances. While accuracy is an intuitive metric, it may not be suitable for imbalanced datasets, where the majority class dominates. However, in balanced datasets, accuracy provides a straightforward evaluation of the model's performance (Luque et al., 2019). As class imbalance was addressed during data preparation, in Chapter 2, Section 3.3.1, accuracy was included as one of the evaluation metrics. The confusion matrix was extracted for each model, and accuracy was calculated from it using the Equation 10.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

where TP represents True Positives, TN represents True Negatives, FP represents False Positives, and FN represents False Negatives.

The accuracy results of the models are elaborated upon in Chapter 6, Section 6.4.1

3.5.2 Area Under the Curve

AUC is a widely used metric for evaluating the performance of binary classification models. It represents the probability that the model ranks a randomly chosen positive instance higher than a randomly chosen negative instance. AUC is advantageous as it provides a single scalar value that summarizes the model's ability to discriminate between the positive and negative classes across all possible decision thresholds (Obi, 2023). A higher AUC value indicates better model performance. The performance of the models used in this study in relation to the AUC is discussed in Chapter 6, Section 6.4.2.

3.5.3 Recall

Recall, also known as sensitivity or true positive rate, measures the proportion of actual positive instances that are correctly identified by the model. In the context of our classification problem, recall indicates the model's ability to correctly identify instances of the target class (BXW diagnosis). High recall is desirable when the cost of missing positive instances (false negatives) is high (Obi, 2023).

Recall was calculated using Equation 11:

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

where TP represents True Positives and FN represents False Negatives

The recall results of the models are discussed in Chapter 6, 6.4.3

3.5.4 Precision

Precision, depicted in Equation 12, quantifies the proportion of positive predictions made by the model that are actually correct (Obi, 2023). In our classification problem, precision, discussed in Chapter 6, Section 6.4.4, reflects the accuracy of the model in identifying positive instances (presence of BXW) among all instances predicted as positive. Precision is particularly important in scenarios where the cost of false positives is high, such as in disease diagnosis, fraud detection or spam filtering.

$$Precision = \frac{TP}{TP + FN} \tag{12}$$

where TP represents True Positives and FN represents False Negatives.

3.5.5 F1-Score

The F1-score, representing the harmonic mean of precision and recall, offers a balanced evaluation of the model's performance by considering both false positives and false negatives. It serves as a comprehensive metric, especially valuable in datasets where class imbalances exist, as it accounts for both Type I and Type II errors (Obi, 2023). This single score strikes a balance between precision and recall, making it suitable for comparing models with different trade-offs between false positives and false negatives. The F1-score results are discussed in Chapter 6, Section 6.4.5, where we delve into the implications of this metric in our classification problem. The F1-score is calculated using the following equation:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{13}$$

In summary, the chosen evaluation metrics - AUC, recall, precision, F1-score, and accuracy

were well-suited for assessing the performance of the classification models. AUC provided an overall measure of discrimination capability, while recall, precision, and F1-score offered insights into the models' ability to correctly classify positive instances and avoid false positives. Accuracy complemented these metrics by providing a holistic view of overall prediction correctness.

3.5.6 Model Optimization

In Chapter 6, Section 6.4.5, it was observed that the RF model outperformed the other models in classifying BXW occurrences. To enhance its performance, we optimized the model further through hyperparameter tuning.

The RF model was customized by specifying parameters such as **mtry**, the number of variables randomly sampled as candidates at each split and **ntree**, the number of trees in the forest. To ensure comprehensive evaluation, train control settings were configured for cross-validation using the "**repeatedcv**" method with 10 folds and 3 repeats. Furthermore, parallel processing was enabled to streamline the tuning process.

When generating a tuning grid, various combinations of hyperparameters for the RF model were explored. The range for **mtry** was set from 1 to 15, and specific values for **ntree** (500, 1000, 1500, 2000) were considered. The evaluation of this tuning process is discussed in detail in Chapter 6, Section 6.4.6

During the tuning process, the optimal parameters were identified as setting **mtry** to 1 and using 1500 trees, as illustrated in Figure 6.23. Subsequently, an optimized RF model was constructed using these parameters, and its performance on the test data was assessed. The performance was then compared with that of the RF model using default parameters and results discussed in Chapter 6, Section 6.4.6.

3.6 Deployment

After model evaluation, RF emerged as the top-performing model, as detailed in Chapter 6, Section 6.4.5. Consequently, it was employed in subsequent deployment stages for various purposes.

This section delineates the utilization of the RF model to discern the environmental factors influencing the occurrence of BXW and their respective impacts. Furthermore, it explores the

development of a habitat suitability map derived from the model's predictions to assess the risk of BXW in Rwanda. Lastly, the section outlines the deployment of the model to the web interface, facilitating user interaction for real-time BXW prediction and early warning alerts through Short Message Service (SMS).

3.6.1 BXW Environmental Drivers

To determine the environmental factors contributing to the occurrence of BXW, we utilized the RF model to assess variable importance. Variable importance measures the contribution of each predictor variable in the model to predicting the outcome. By analyzing variable importance, we identified the significant environmental factors influencing the presence or absence of BXW. Chapter 6 Section 6.5.1 provides detailed insights into the results obtained from this analysis.

3.6.2 Impact of Environmental Factors on BXW

Partial plots were instrumental in exploring the relationship between BXW occurrence and various environmental factors identified as significant contributors. These plots provide a nuanced understanding of how BXW responds to key environmental drivers, shedding light on the complex interplay between these factors and the incidence of the disease. The findings from these analyses are discussed in more detail in Chapter 6, Section 6.5.2.

3.6.3 Mapping BXW Habitat Suitability

The RF model was utilized to forecast the likelihood of BXW occurrence across the entirety of Rwanda's cropland. Leveraging the insights gained from the model, a comprehensive habitat suitability map was crafted, as shown in Chapter 6, Section 6.5.3. This risk map serves as a valuable resource, enabling researchers and stakeholders to delve into BXW prevalence patterns across Rwanda's diverse landscapes.

By discerning regions with heightened susceptibility to BXW, this map empowers decision-makers with the knowledge needed to implement targeted interventions and preventive measures effectively. Through its detailed analysis, the habitat suitability map facilitates a deeper understanding of BXW distribution dynamics, guiding strategic efforts toward the mitigation and management of this devastating plant disease.

3.6.4 Developing an Early Warning System

To establish an ML-driven EWS, the RF model was stored, and a Plumber API was developed to enable seamless interaction with a web portal developed using Django, a Python web framework. This infrastructure allowed for real-time access to the RF model's predictive capabilities, facilitating timely interventions in response to potential BXW outbreaks. Furthermore, the early warning alerts were disseminated through the utilization of SMS, ensuring swift communication of critical information to relevant stakeholders. For a comprehensive understanding of the implementation process and the seamless integration of the various components comprising the platform, further details are discussed in Chapter 5.

Chapter 4: System Design and Architecture

4.1 System Modeling

The system modeling process in this study employed the UML, a standardized methodology renowned in software engineering for its effectiveness in visualizing and documenting system design. UML serves as a universal language, enabling system architects, developers, and stakeholders to communicate and comprehend complex systems efficiently (Koc et al., 2021).

Among the array of UML diagrams available, the ***use case diagram*** was utilized in this study. This diagram, a type of behavioral diagram, illustrates the interactions between users (actors) and the system, offering a high-level overview of the system's functionalities and the roles of the actors involved (El Miloudi and Ettouhami, 2018). Through this depiction, stakeholders gained valuable insights into the system's operational dynamics.

Figure 4.1 illustrates the UML diagram, showcasing the various components, relationships, and behaviors within the system.

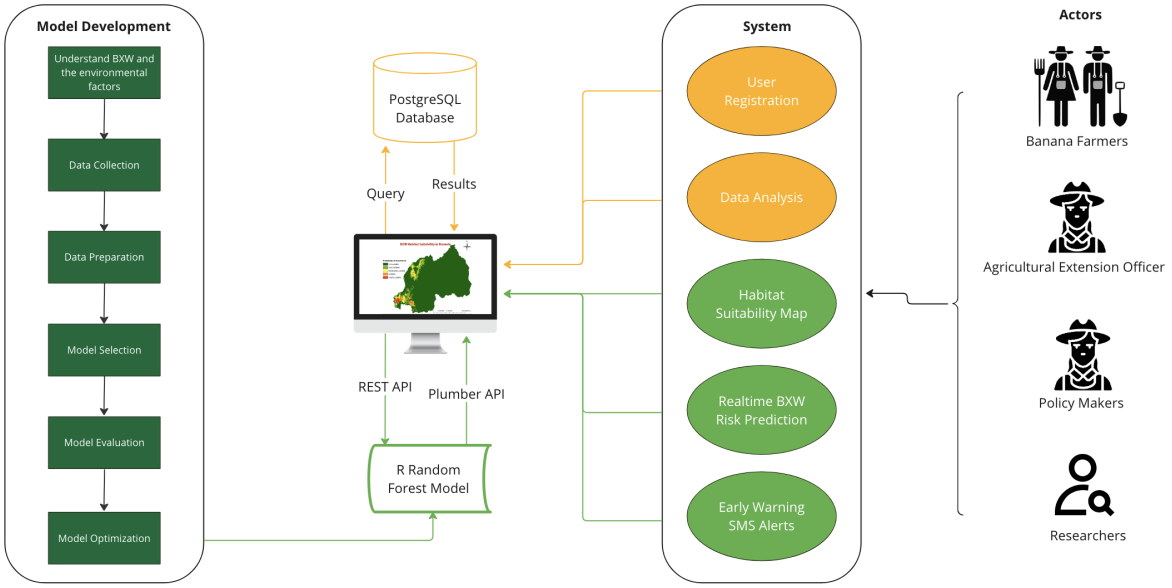


Figure 4.1: UML diagram

4.2 System Components

During the study, the system developed comprised three main components: a database, a web portal, and an SMS notification system.

4.2.1 Database

In this study, the database architecture incorporated the use of a Snowflake schema to optimize data organization and query performance. The Snowflake schema, characterized by its centralized fact table surrounded by multiple dimension tables (Karmani et al., 2020), facilitated enhanced data integrity and streamlined querying processes. Dimension tables, including Province, District, Sector, User, BXW Diagnosis, and SMS, were structured in a normalized manner, ensuring data integrity and facilitating efficient querying.

An ERD, shown in Figure 4.2 visually depicted the relationships between these entities, outlining their attributes and connections within the database system. This ERD served as a blueprint for understanding the database structure and ensuring the integrity of data relationships.

Through the combined utilization of the Snowflake schema and ERD, the database architecture was structured to handle diverse data needs. This approach ensured that the database system could efficiently manage BXW-related data and support mitigation efforts effectively.

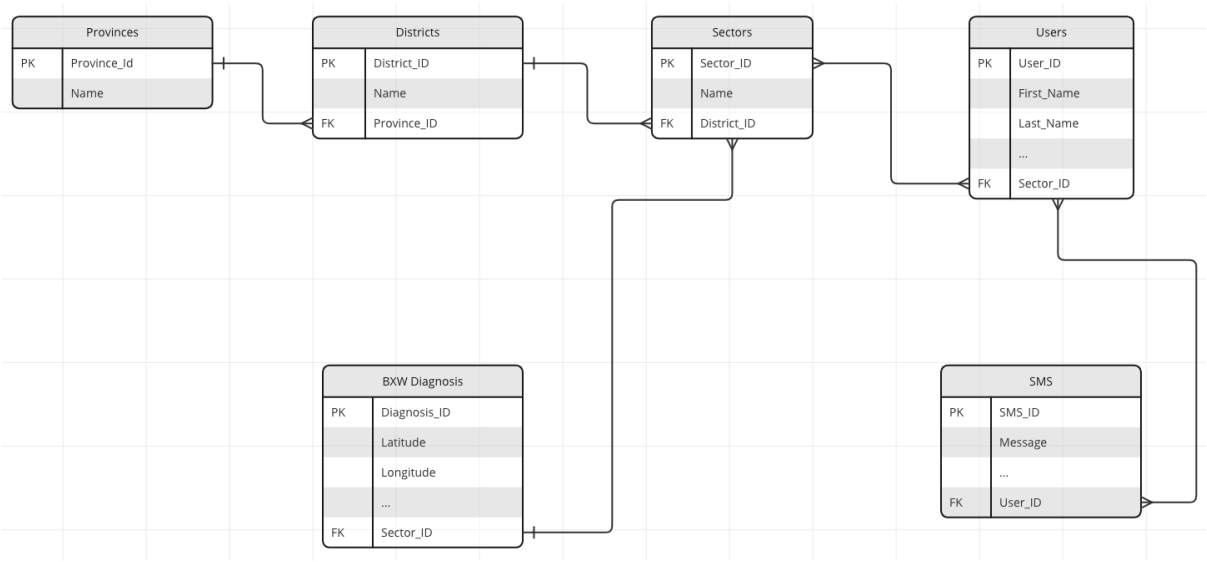


Figure 4.2: Entity relationship diagram

Table 5 offers a comprehensive breakdown of each database table that has been created.

Table 5: Database tab

Table	Field Name	Data Type	Description
Province	Province_ID	int	Unique identifier for province
	Name	char (100)	Name of the province
District	District_ID	int	Unique identifier for district
	Name	char (100)	Name of the district
	Province_ID	int	Foreign key referencing the Province table
Sector	Sector_ID	int	Unique identifier for sector
	Name	char (100)	Name of the sector
	District_ID	int	Foreign key referencing the District table
BXW Diagnosis	Diagnosis_ID	int	Unique identifier for occurrence
	Latitude	float	Latitude coordinates of occurrence
	Longitude	float	Longitude coordinates of occurrence
	Diagnosis	int	BXW diagnosis encoded as binary
	Occurrence_Date	DateTime	Date of the diagnosis
	Sector_ID	int	Foreign key referencing the Sector table
User	User_ID	int	Unique identifier for user
	First_Name	char (50)	User's first name
	Last_Name	char (50)	User's last name
	Phone_Number	char (15)	User's phone number, including country code
	Latitude	float	Latitude coordinates of the user's farm
	Longitude	float	Longitude coordinates of the user's farm
	Registered_Date	DateTime	Date of user enrollment in the alerts system
	Sector_ID	int	Foreign key referencing the Sector table
SMS	SMS_ID	int	Unique identifier for SMS notification
	Message	text	The message content, including the prediction
	Response	text	The API's response to query
	SMS_Date	DateTime	The date and time the SMS was sent
	User_ID	int	Foreign key referencing the User table

4.2.2 Web Portal

The web portal was organized into four primary sections, each addressing different aspects of BXW concerns. These sections included a *Home Page* offering a summary of the problem and proposed solutions, a *Data Analysis* section for reviewing historical occurrences of BXW, a *BXW Sentinel* Page for real-time susceptibility prediction, and a *User Registration* Page for early warning sign-ups. To visualize the user experience within the portal, a sitemap and wireframes were created as per below.

Sitemap

A sitemap is a visual representation of the structure and hierarchy of a website, outlining its various pages and the relationships between them. It provides users with a clear overview of the website's layout and organization, helping them navigate through different sections and find the information they need more efficiently. Sitemaps are essential for improving website usability and enhancing the user experience by ensuring that all pages are easily accessible and logically organized.

Figure 4.3: Sitemap of the web portal

Wireframes

Wireframes are simple, schematic representations of a website or application's layout, illustrating the placement of elements such as navigation menus, content sections, and interactive features without focusing on design details. They serve as blueprints for the user interface design, helping to visualize the structure and functionality of the final product before investing time and resources in full-fledged development. Miro, a collaborative online platform [`https://miro.com/`], was used to create the wireframes for the website design.

(a) **Home Page Wireframe**

The home page offers a concise overview of the BXW problem, showcases key features of the portal, and provides easy navigation to other sections.

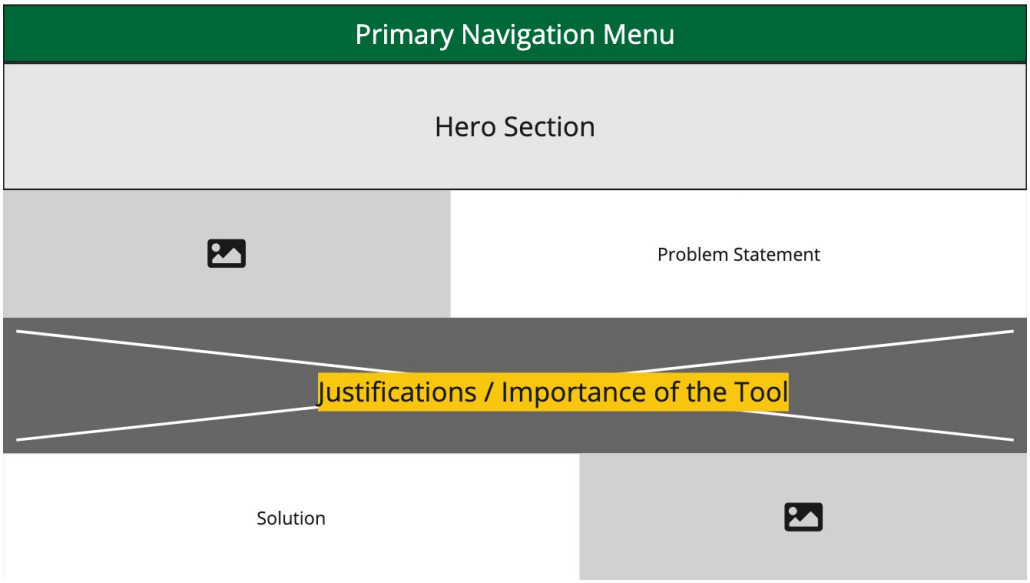


Figure 4.4: The home page wireframe

(b) **Data Analysis Wireframe**

The Data Analysis page provides insights into BXW occurrences, encompassing diagnostic counts, a geographical distribution map, and correlations with diverse factors including location, time, and environmental conditions.

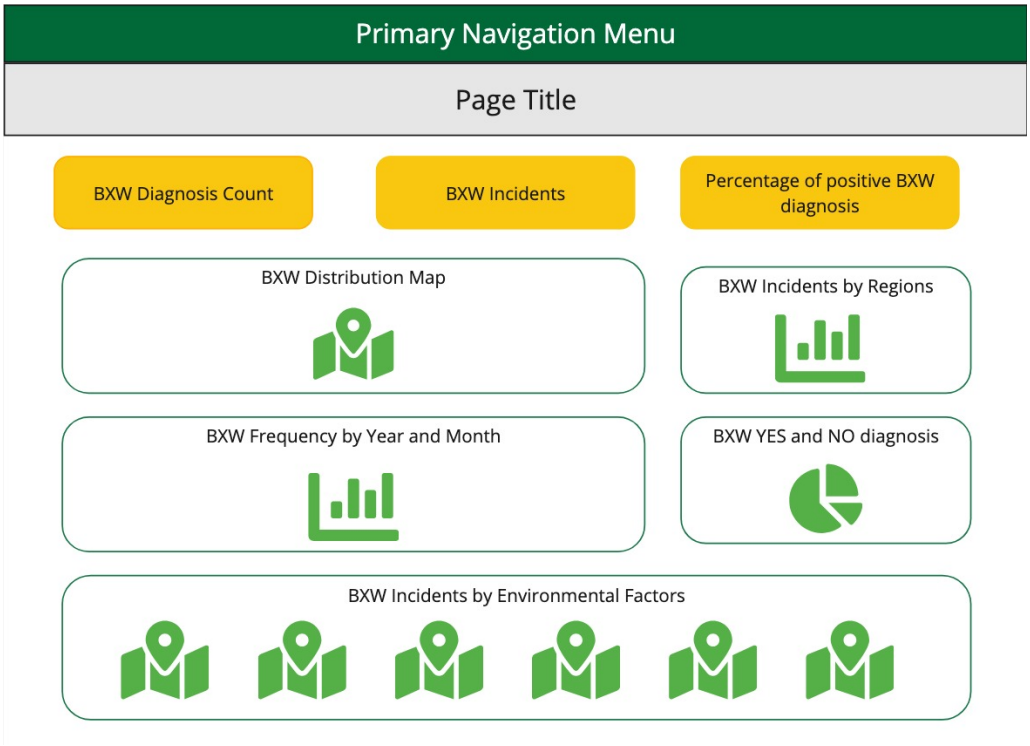


Figure 4.5: Data analysis wireframe

(c) **BXW Sentinel Wireframe**

The BXW Sentinel page provides real-time monitoring of BXW susceptibility, featuring a habitat suitability map, a predictive form for obtaining BXW susceptibility forecasts, and graphical representations of BXW responses to environmental conditions.

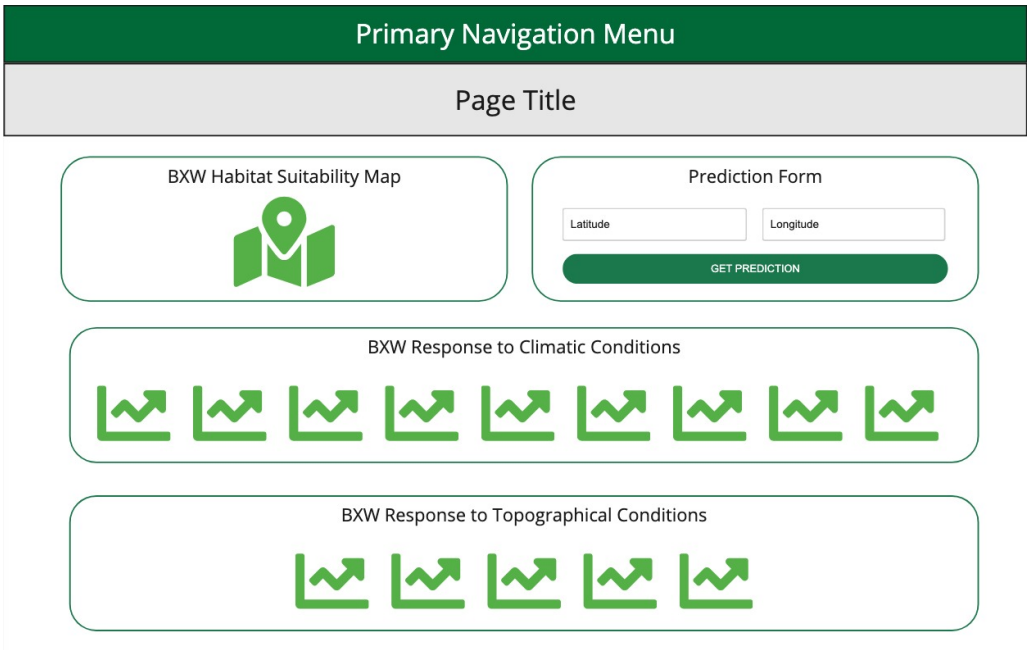


Figure 4.6: BXW sentinel wireframe

(d) **User Registration Wireframe**

Users have the option to sign up for alerts regarding BXW susceptibility via the user registration page. They are prompted to input their names, contact information, and location to facilitate accurate predictions.

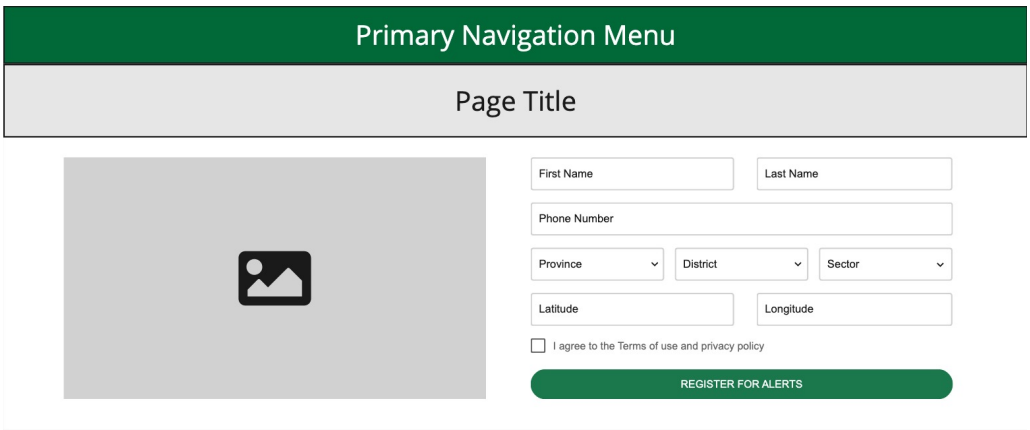


Figure 4.7: User registration wireframe

4.2.3 SMS Notifications

The early warning alert SMS system employed the architecture depicted in Figure 4.8 to ensure seamless communication with users.

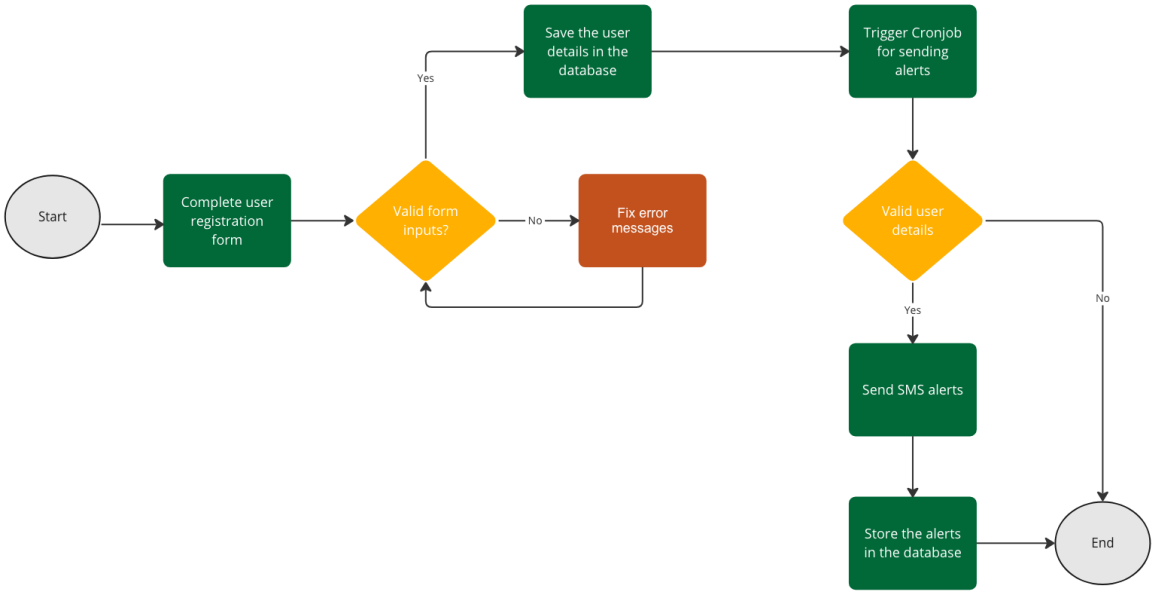


Figure 4.8: SMS notification flowchart

Chapter 5: System Implementation and Testing

This chapter provides an in-depth exploration of the development process and technical aspects of the BXW EWS. We delve into the architecture, design decisions, and implementation details of the database, web portal and SMS notification system. By examining the methodologies, tools, and technologies employed, this chapter offers insights into the systematic approach taken to address the challenges posed by BXW in Rwanda. Through an analysis of the system's components and implementation strategies, a deeper understanding is gained of how the system was conceptualized, designed, and implemented to effectively mitigate the impact of BXW on agricultural practices.

Additionally, this chapter will discuss testing methods aimed at evaluating the usability of the system and determining whether the problem statement has been adequately addressed.

5.1 System Implementation

5.1.1 Database

In this study, the database implementation relied on PostgreSQL, an open-source relational database management system. PostgreSQL was chosen for its robust features, reliability, and scalability, making it well-suited for handling the complexities of the data involved in addressing BXW concerns.

To structure the database, Django models were utilized, providing a high-level abstraction for defining the data models and their relationships within the application. The tables developed included Province, Sector, District, BXW Diagnosis, User, and SMS, as highlighted in the ERD shown in Figure 4.2.

The database design underwent a normalization process to ensure optimal data organization, eliminate redundancy, and maintain data integrity. This involved breaking down data into smaller, more manageable entities and establishing primary and foreign key constraints to define relationships between tables. By adhering to normalization principles, the database schema was streamlined and optimized for efficient data retrieval and manipulation.

To populate the database with Rwanda's administrative data and historical BXW occurrences, the data was prepared in CSV files, and primary-foreign key constraints were defined. Django scripts were then developed to facilitate the loading of the existing data into the database.

5.1.2 Web Portal

The implementation of the web portal leveraged several key technologies, each serving a distinct role in the development process. Django, a high-level Python web framework, was employed for the backend infrastructure. Django provides a robust framework for building web applications, offering features such as Uniform Resource Locator (URL) routing, template rendering, and database interaction. Its object-relational mapping capabilities streamline database access and manipulation, facilitating efficient data handling and storage.

On the frontend side, Hypertext Markup Language (HTML), Cascading Style Sheets (CSS), and jQuery played pivotal roles in developing the user interface and enhancing interactivity. HTML served as the foundation of web pages, defining the structure and content of the user interface elements. It provided a standardized markup language for organizing text, images, links, and other multimedia content.

CSS complemented HTML by enabling the styling and presentation of web pages. With CSS, we were able to control the layout, colors, fonts, and visual aspects of the user interface, ensuring a cohesive and visually appealing design across different devices and screen sizes.

jQuery, a fast and concise JavaScript (JS) library, was utilized to enhance the frontend functionality and user experience. jQuery simplified common JS tasks, such as event handling, Document Object Model (DOM) manipulation, and AJAX requests, which made it easier to implement dynamic and interactive features on web pages. Its extensive library of plugins and utilities further streamlined development tasks, allowing for rapid prototyping and iteration.

(a) Homepage

The homepage, illustrated in Figure 5.1 was carefully crafted to provide a brief summary of the problem and proposed solution, as well as reasons why using the tool is important. Additionally, it featured navigational links to facilitate seamless transitions between various sections of the website. Furthermore, the page included acknowledgments to the projects that supported the research, namely ICT4BXW [https://ict4bxw.com/], an initiative focused on leveraging ICT solutions for managing banana diseases, particularly BXW, and the CGIAR Excellence in Agronomy Initiative [`https://eia.cgiar.org/`], which aims to enhance agricultural practices and productivity through research and innovation.

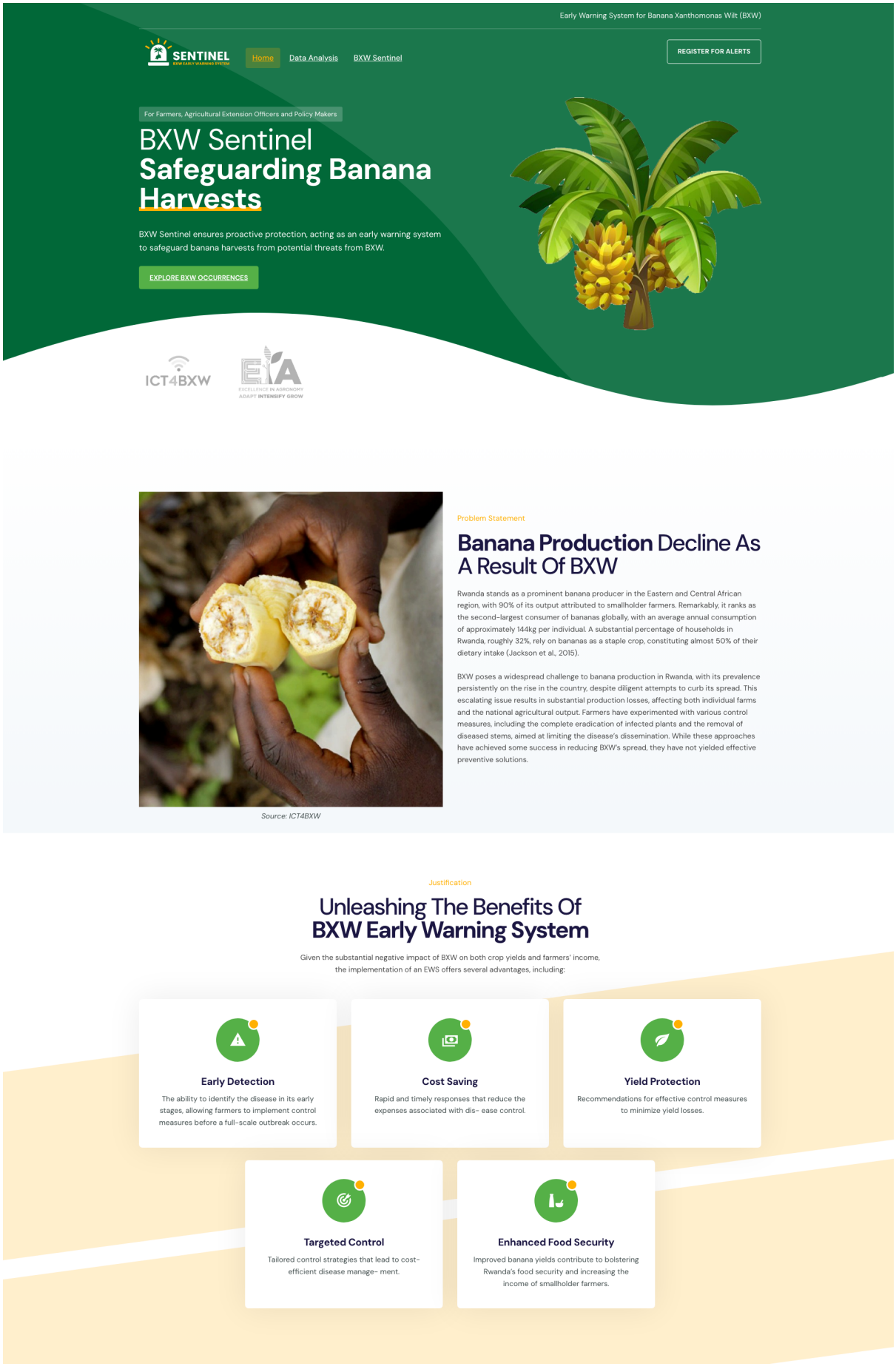


Figure 5.1: Homepage

(b) Data Analysis Page

The data analysis page, highlighted in Figure 5.2 provided insights into the historical occurrences of BXW. Django views were used to interact with the database to retrieve stored data, which was then passed to frontend templates for visualization. Caching was employed to optimize data retrieval.

Highcharts, a JS library for creating interactive charts and graphs, was utilized for plotting various charts, including the total counts of diagnosis, the number of positive BXW diagnoses, and their percentage. Additionally, the page featured a spatial distribution map of occurrences, bar charts illustrating occurrences per province, and class distribution counts.

A drilldown chart was also developed to enable users to explore occurrences per year and month. Furthermore, the page included plots of BXW occurrences under different climatic and topographical conditions, offering valuable insights into BXW's response to varying environmental factors.

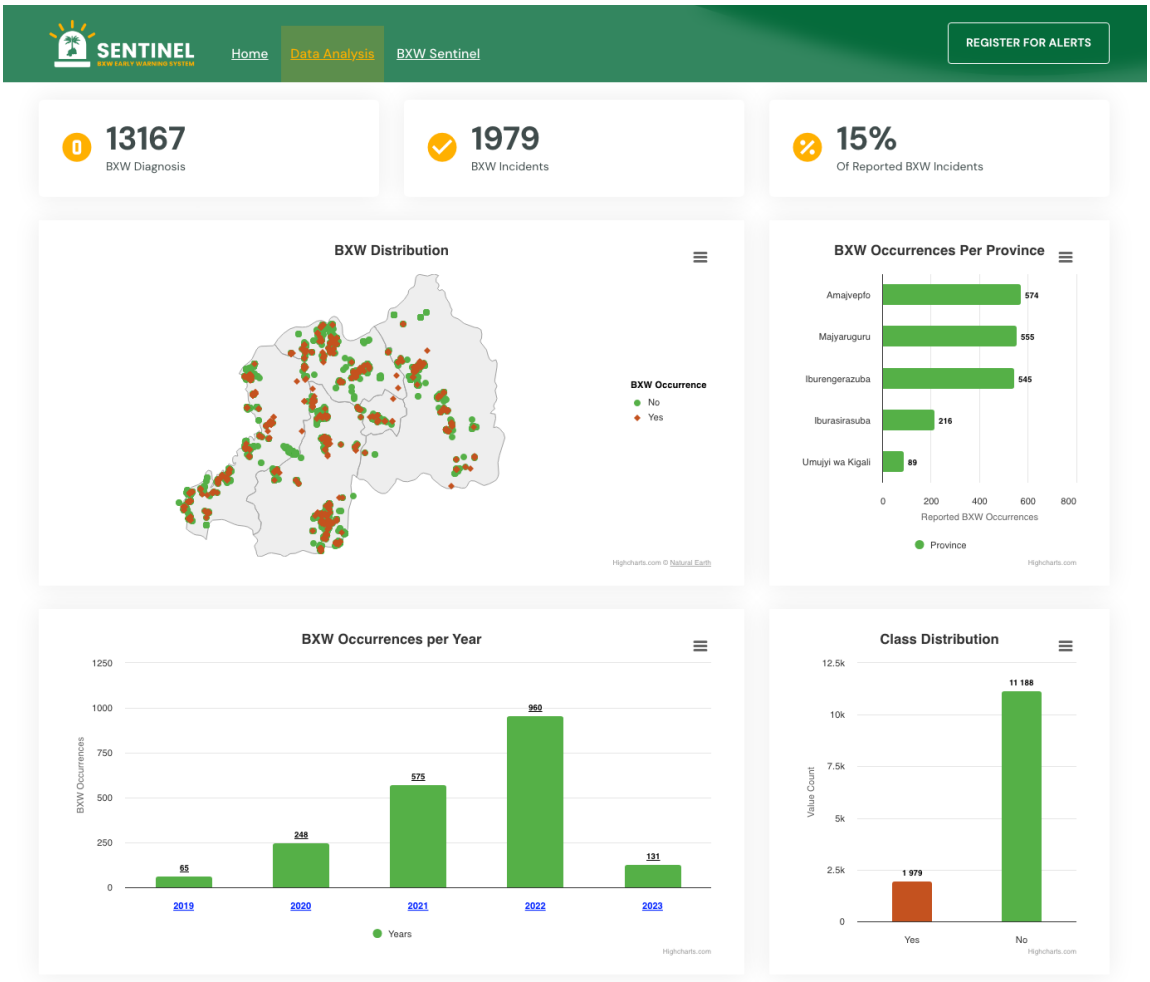


Figure 5.2: Data analysis page snippet

(c) BXW Sentinel

The BXW sentinel, showcased in Figure 5.3 was developed to interact with the RF model for predicting BXW occurrences. The R model was stored and loaded into an R microservice to enable faster predictions. Containerizing the R microservice using Docker enhanced its portability and reuse. An API endpoint was developed using the R Plumber package to facilitate interaction with the model from the web portal.

The habitat suitability map for BXW, depicted in Figure 6.26, was generated from the modeling process and integrated into the platform. It provides an overview of BXW susceptibility in different regions of Rwanda.

For real-time predictions from the model, users are prompted to input coordinates through an HTML form. Upon clicking the "Get Prediction" button, the user inputs are passed to a Django view. Django then sends a GET API request to the R microservice and provides the coordinates. The microservice checks if the coordinates belong in Rwanda, extracts the environmental conditions, and passes them to the RF model for prediction. The microservice returns the probability of BXW occurrence to Django through the Plumber API, which is then displayed to the user. If the location is outside Rwanda, a message requests valid coordinates from the user.

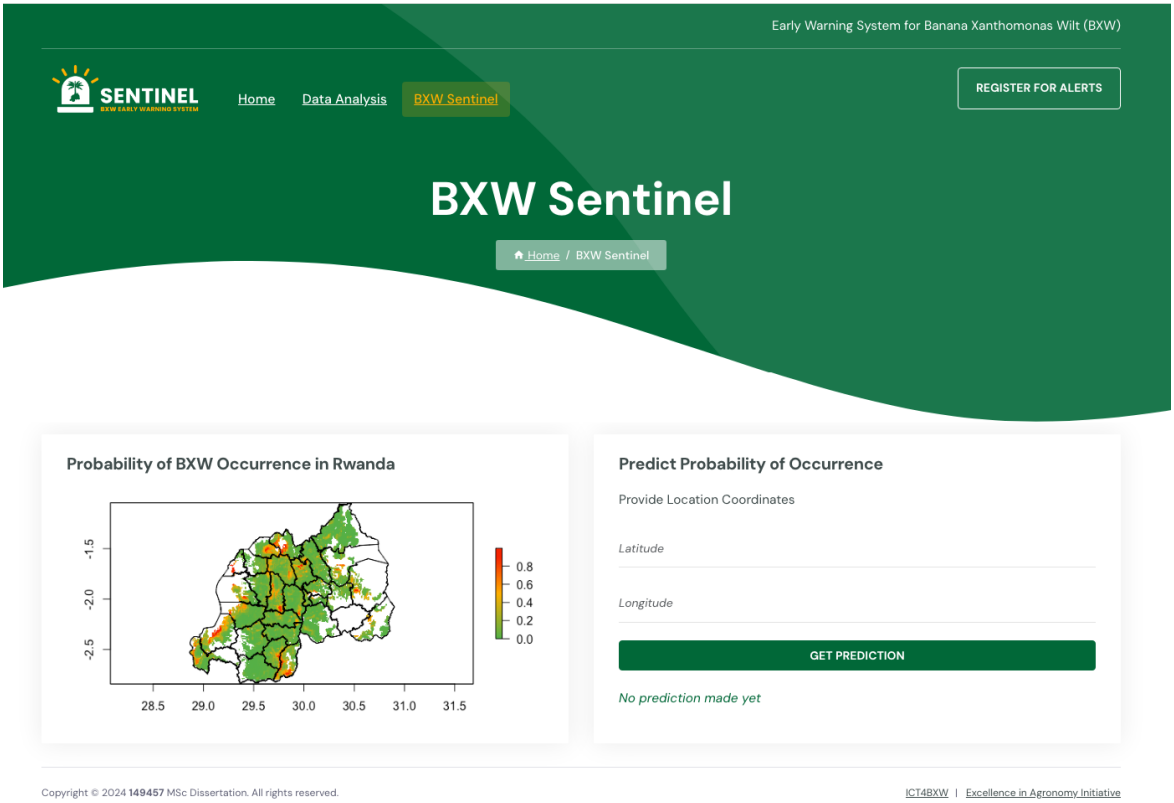


Figure 5.3: BXW sentinel page

(d) User Registration

A user registration module was developed to streamline the process of enrolling for BXW early warnings. This module enables users to input their first and last names, phone numbers, and select their respective provinces, districts, and sectors from interdependent dropdown menus. The dropdown menus dynamically adjust based on the user's selection, ensuring accuracy and efficiency in data entry. Additionally, users provide the coordinates of their banana farms to enhance the precision of the EWS. Upon submission of the registration form by clicking the "Register for Alerts" button, the system validates the form fields, verifying that the coordinates fall within Rwanda's bounding box and that the phone number includes the required country code.

In cases where the form inputs are found to be invalid, error messages are promptly displayed to guide users in rectifying the errors before resubmitting their details. Once the user registration process is successfully completed and all necessary criteria are met, the user details are securely stored in the database. These details serve as crucial data points utilized by the system for sending timely and accurate early warning messages, thereby enhancing the effectiveness of the BXW mitigation efforts.

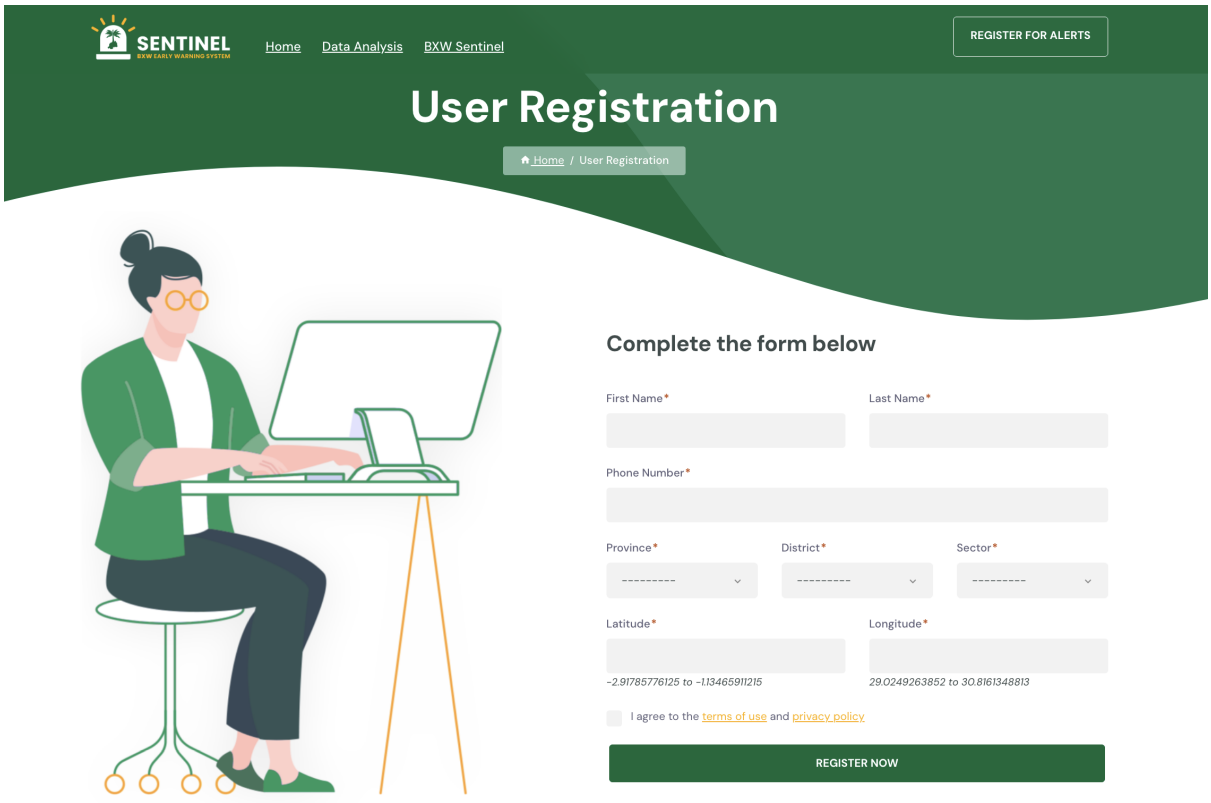


Figure 5.4: User registration page

5.1.3 SMS Notifications

The early warning component of the study was implemented using a bulk SMS system, leveraging the Africa's Talking [`https://africastalking.com/`] API to facilitate seamless communication. The user management system discussed in the preceding section played a crucial role in overseeing the complete process of user registration and enrollment in the SMS alert system.

To automate the process of sending early warning alerts, a CronJob was developed to handle message scheduling. These alerts were scheduled to be sent out biweekly, ensuring timely dissemination of critical information. During the scheduled intervals, the CronJob seamlessly executed its task, pulling data of enrolled users from the system's database. Subsequently, an API call was triggered to the R model microservice, which provided predictions of BXW occurrence risk for each user based on their specified locations. The resulting predictions were then incorporated into personalized messages, as shown in Figure 5.5, effectively conveying the probability of BXW occurrence to the respective users.

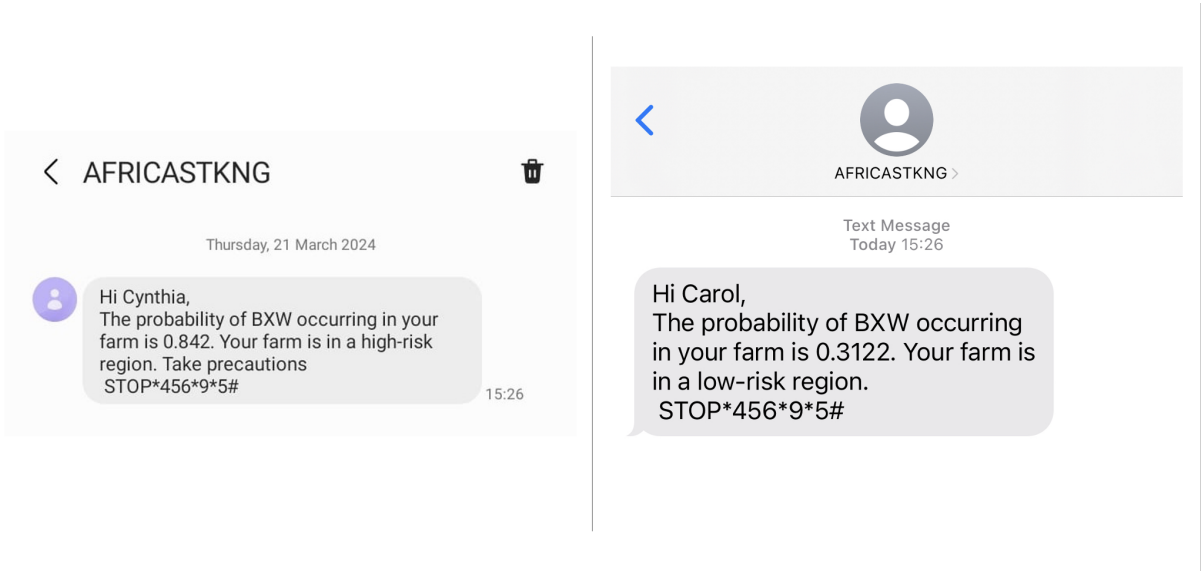


Figure 5.5: Early warning SMS

5.2 Testing

The testing phase of the implemented BXW early warning system aimed to validate its functionality, usability, compatibility, security, and effectiveness in addressing the identified problem statement. Through a series of testing procedures, including functionality, usability, compatibility, security, and validation testing, the system's performance and reliability were evaluated

to ensure its readiness for deployment in real-world scenarios.

5.2.1 Functionality Testing

Functionality testing was a crucial phase in evaluating the BXW EWS's performance. The accuracy of BXW diagnosis was assessed to guarantee precise identification of BXW occurrences by comparing system-generated diagnoses with known historical occurrences. Furthermore, the real-time BXW susceptibility prediction feature underwent thorough validation against current environmental conditions and historical data.

In addition to BXW diagnosis accuracy and real-time prediction, functionality testing also encompassed the evaluation of user registration and SMS delivery processes. The user registration process was scrutinized to ensure that users could seamlessly enroll for early warning alerts by providing their information. Similarly, SMS delivery functionality was rigorously tested to confirm that users received SMS notifications promptly at scheduled times.

Overall, functionality testing played a pivotal role in assessing the BXW EWS's operational capabilities. By rigorously evaluating BXW diagnosis accuracy, real-time prediction functionality, user registration process, and SMS delivery functionality, the system's readiness for deployment in real-world scenarios was confirmed.

5.2.2 Usability Testing

The usability testing was conducted in-house to assess various aspects of the web portal's functionality and usability. This included evaluating the ease of navigating through different sections of the portal's user interface and observing users as they performed tasks. The goal was to ensure that the portal's navigation was intuitive and user-friendly, allowing users to access information and features effortlessly.

Additionally, the usability of registration forms was tested by having users complete them and providing feedback on any difficulties encountered. This process aimed to identify and address any potential issues that could hinder users from successfully registering for the BXW EWS. By gathering feedback directly from users, the testing helped ensure that the registration process was streamlined and accessible to all users.

Furthermore, testing was conducted to assess the real-time prediction of BXW on the sentinel page. Users were tasked with obtaining BXW susceptibility forecasts using the predictive form,

and their interactions were closely monitored. This testing aimed to verify that users could easily access and utilize the real-time prediction feature, enabling them to make informed decisions based on the latest BXW susceptibility information.

In general, conducting testing in-house allowed for a thorough evaluation of the web portal's functionality and usability. By identifying and addressing any issues early in the development process, the portal could be optimized to provide users with an intuitive and seamless experience.

5.2.3 Compatibility Testing

Compatibility testing was conducted to ensure the web portal's functionality across various platforms. To verify cross-browser compatibility, the web portal was accessed using different web browsers, including Chrome, Firefox, Safari, and Edge. Testers navigated through different sections of the portal and performed tasks to assess its performance and functionality. Any discrepancies or issues encountered during this process were documented for further analysis and resolution.

Similarly, mobile responsiveness testing was carried out to evaluate the web portal's display and functionality on different mobile devices. Testers accessed the portal using smartphones and tablets with varying screen sizes and resolutions. They interacted with the portal to determine if all features were accessible and if the layout adapted effectively to different screen sizes. Any instances of misalignment, distortion, or functionality issues were noted and addressed to ensure a seamless user experience across mobile devices.

5.2.4 Security Testing

In the study, security testing was conducted to ensure the robustness of the system's data protection measures and API security protocols. To assess user data protection, testers thoroughly examined the database architecture and access controls to verify the secure storage of user information. This involved scrutinizing encryption practices, access permissions, and data storage protocols to prevent unauthorized access or data breaches. Additionally, rigorous testing was carried out to evaluate the resilience of API endpoints against potential security threats.

5.2.5 Validation Testing

Validation testing was conducted to assess the system's impact on smallholder banana farmers in Rwanda, in line with the problem statement highlighted in chapter 1, subsection 1.2. This involved expert validation by domain experts in BXW detection and mitigation. These experts tested the system to ensure that it accurately reflected actual BXW occurrences on the ground and provided feedback on its performance and effectiveness in identifying and predicting BXW outbreaks. Their feedback was crucial in assessing the system's usability, accuracy, and overall effectiveness in addressing BXW concerns among banana farmers in Rwanda.

Chapter 6: Discussion of Results

This chapter offers a review of the outcomes achieved across the different stages of the CRISP-DM framework, from data understanding to model deployment. This segment presents the insights gathered from exploring and analyzing the dataset, highlighting the patterns and trends identified during data preprocessing and modeling. Moreover, it showcases the performance metrics and effectiveness of the developed models in achieving the research objectives highlighted in Chapter 1, Section 1.3.

6.1 Data Understanding

6.1.1 Spatial Distribution of BXW

The map shown in Figure 6.1 illustrates the spatial distribution of BXW occurrences across Rwanda. Each green point indicates a location where BXW diagnosis was negative, while red points denote positive diagnoses of BXW on banana plants. This visualization offers an overview of BXW outbreaks' geographic spread throughout the country, highlighting disease hotspots. When analyzed alongside the bar plot in Figure 6.2, the map provides insights into the provinces in Rwanda more susceptible to BXW occurrences.

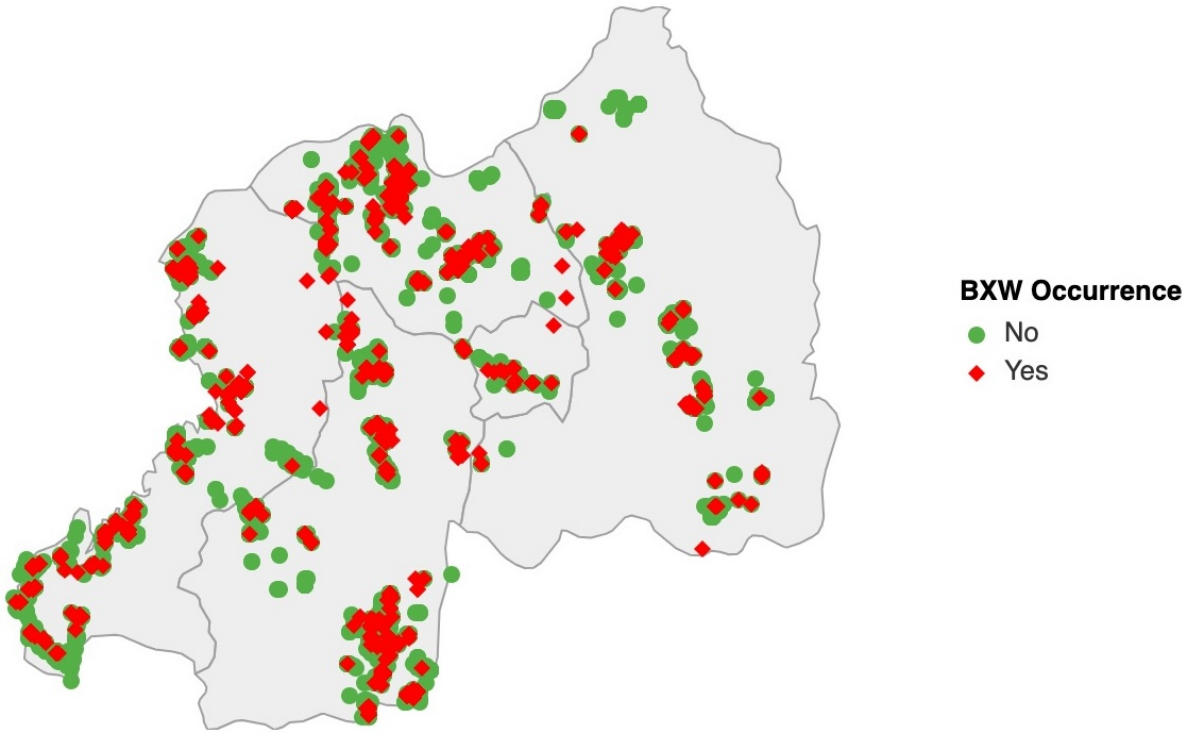


Figure 6.1: BXW spatial distribution

The results in Figure 6.2 reveal notable disparities in the occurrence of BXW across different provinces. Umujyi wa Kigali (City of Kigali) stands out with the highest percentage of BXW occurrence at 42.14%, indicating a significant prevalence of the disease within the province. Following behind, Majyaruguru (Northern Province) exhibits a substantial percentage of 16.89%, highlighting considerable BXW incidence in the Northern Province. Similarly, Amajvepfo (Southern Province) demonstrates a moderate occurrence of the disease at 15.98%.

In contrast, Iburengerazuba (Western Province) and Iburasirasuba (Eastern Province) display relatively lower percentages of BXW occurrence at 13.51% and 13.14% respectively. These findings underscore the varying levels of BXW prevalence across different provinces, with the City of Kigali notably experiencing a significantly higher incidence compared to the other provinces.

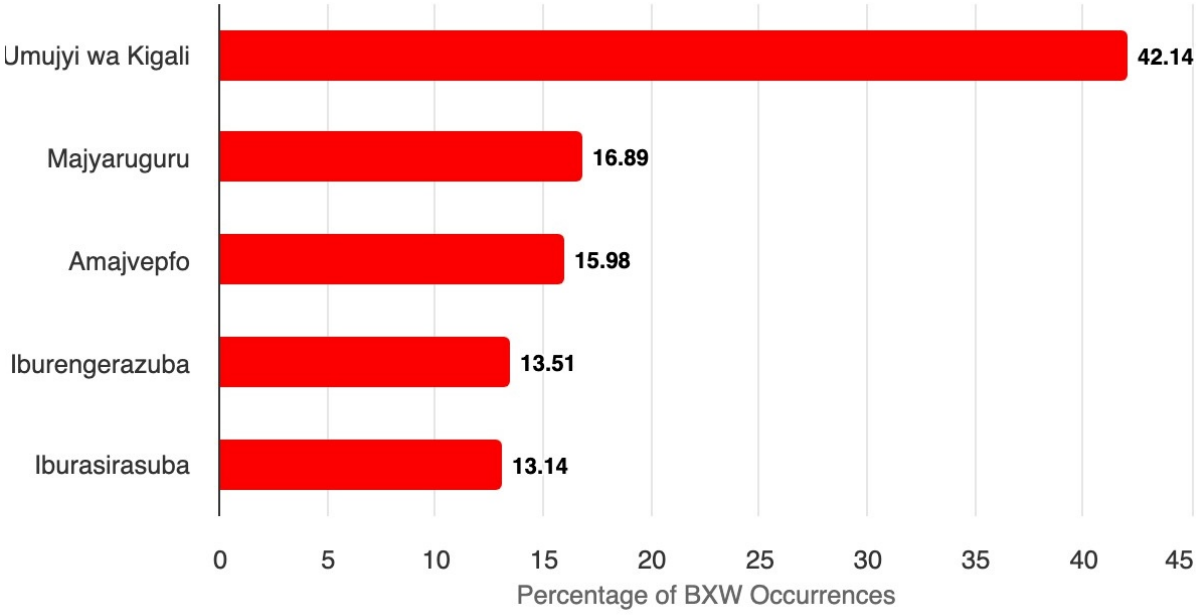


Figure 6.2: BXW occurrences per province

6.1.2 Temporal Distribution of BXW

The temporal distribution of BXW, highlighted in Figure 6.3 and Figure 6.4 provide valuable insights into the dynamic nature of its occurrences over time. Analysis of the data highlights fluctuations in BXW outbreaks across various months and years, revealing distinct seasonal patterns and long-term trends. Observations indicate a notable increase in BXW occurrences from 2020 to 2022, followed by a decrease in 2023.

While it's evident that the occurrences in 2019 and 2024 are relatively low, it's crucial to acknowledge that the dataset for these years was incomplete, covering only specific months (December 2019 and January to February 2024). This limitation emphasizes the need for careful interpretation.

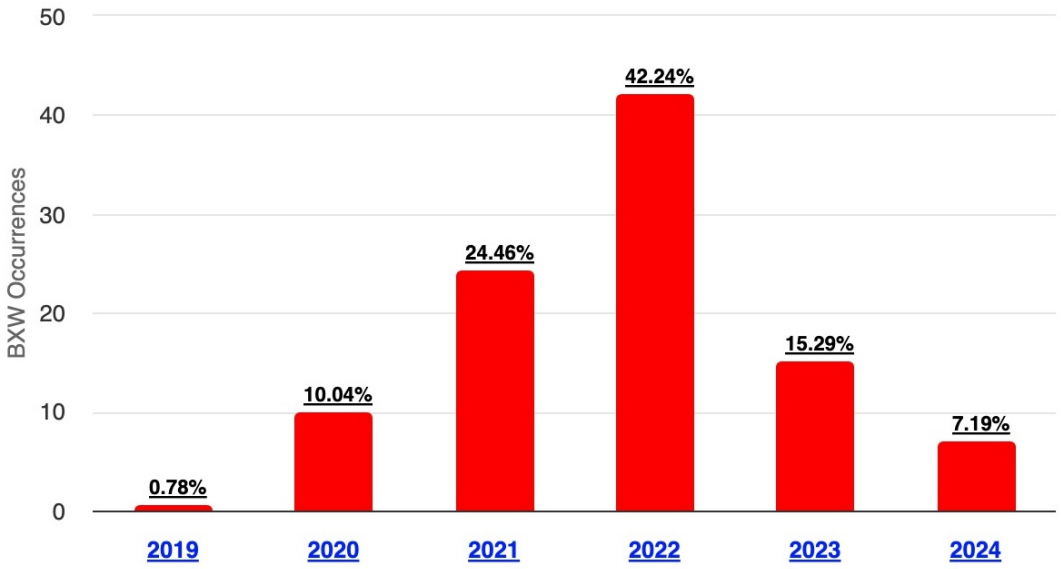


Figure 6.3: BXW occurrences per year

The analysis shown in Figure 6.4 illustrates the percentage of BXW occurrence for each month across the years 2020 to 2023. In 2020, peaks in BXW occurrence were observed in February (23.39%), March (15.60%), and November (12.05%), while lower occurrences were recorded in April (1.38%) and July (2.29%). In 2021, notable spikes were seen in November (27.87%), December (15.82%), and October (13.37%), with lower occurrences noted in January (3.01%) and February (2.45%). In 2022, a more varied distribution was observed, with May (12.76%) exhibiting the highest percentage of occurrence. Lastly, 2023 witnessed a significant spike in BXW occurrence in October (28.92%) alongside notable percentages in June (12.95%) and November (12.05%), while January (0.6%) and May (0.6%) showed the lowest occurrences.

The fluctuations in BXW occurrences observed across different months and years highlight the dynamic nature of the disease and its response to various environmental factors. Factors such as climate conditions, agricultural practices, and disease management strategies may contribute to the observed temporal patterns. Additionally, the variations in BXW occurrences underscore the need for continuous monitoring and adaptive management strategies to address emerging challenges and ensure the resilience of banana cultivation systems.

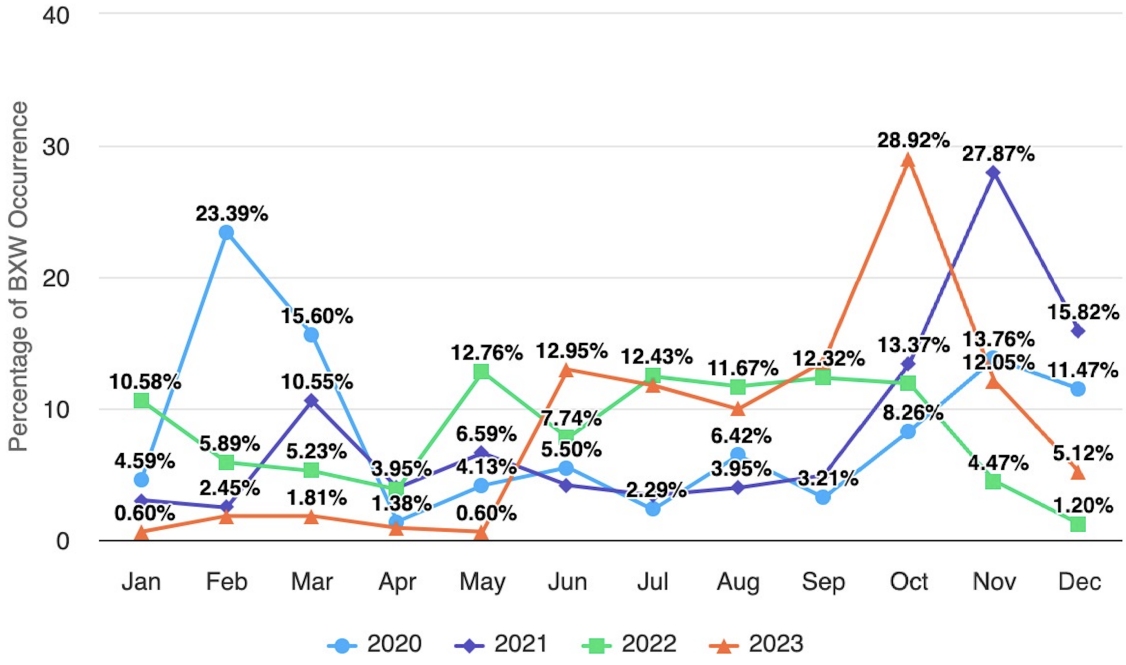


Figure 6.4: BXW occurrences per month

6.1.3 Rwanda's Topography

The analysis of Rwanda's topographic features, including elevation, slope, aspect, and hill-shade, offers valuable insights into the country's diverse landscape and its implications for ecological processes.

Rwanda's elevation exhibits notable variations across different regions. The yellow and green areas in Figure 6.5, showing higher elevations, mainly represent mountainous regions, particularly in the Northern, Southern, and Western Provinces. Conversely, lower elevations, mainly found in the Eastern Province and the City of Kigali, are depicted by lowland regions. This diversity in elevation influences local climate dynamics, the distribution of vegetation, and agricultural activities.

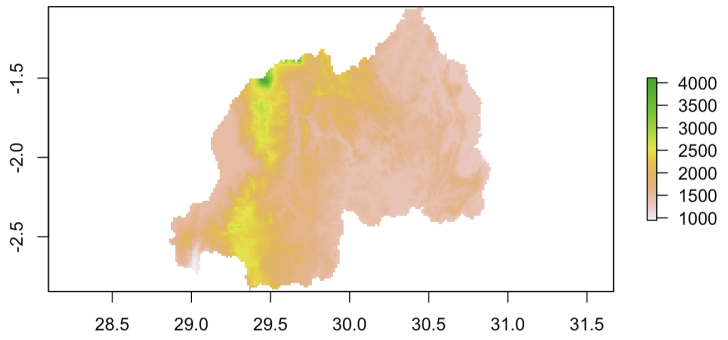


Figure 6.5: Rwanda's elevation (in m)

The slope analysis reveals the steepness of Rwanda's terrain, with steeper slopes often associated with mountainous regions and gentler slopes prevalent in lowland areas. Aspect, indicating the direction that the terrain faces, plays a crucial role in determining exposure to sunlight and prevailing winds, thus influencing microclimates and habitat suitability for various species.

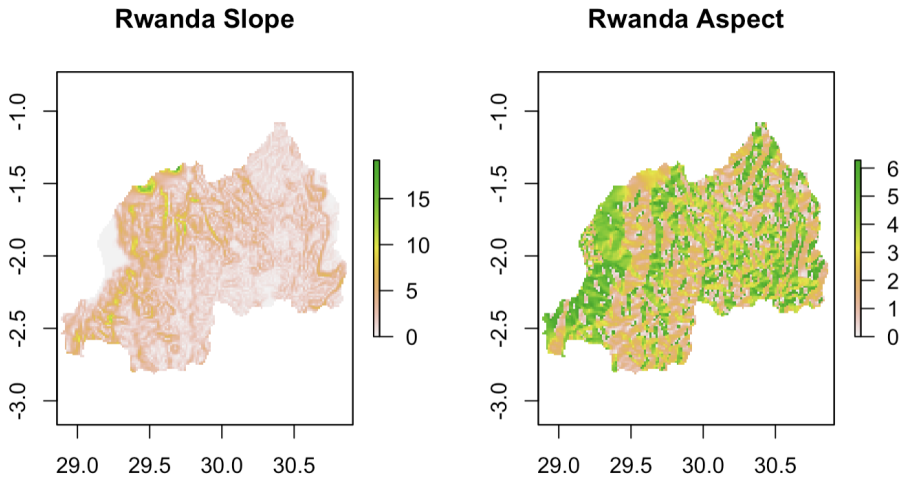


Figure 6.6: Rwanda's slope (in °) and aspect (in °)

The color scale in Figure 6.7 illustrates shading intensity across Rwanda's landscape. Negative values (shades of red) indicate areas with less illumination or shadowed regions, while positive values (shades of green) signify brighter spots. This shading gradient offers insights into terrain relief and slope characteristics, impacting factors like solar radiation exposure and temperature distribution. Understanding these patterns helps assess Rwanda's ecological diversity and the influence of topographic features on its ecosystems and biodiversity.

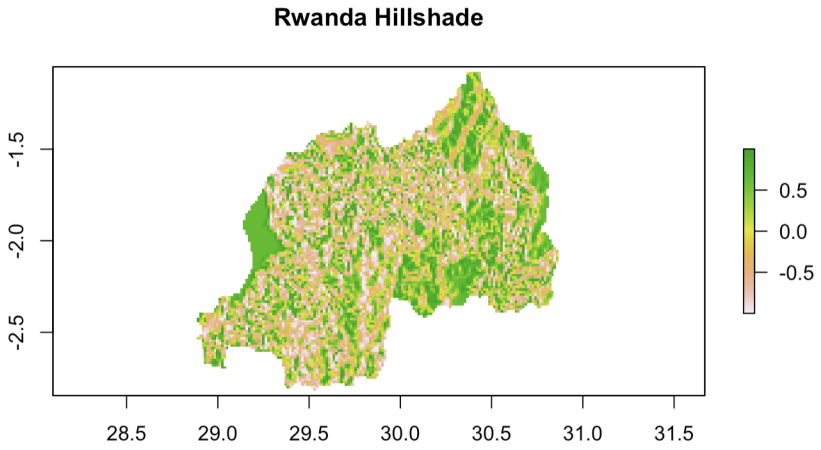


Figure 6.7: Rwanda's hillshade

6.1.4 Rwanda's Climate

Rwanda's climate, as inferred from the analysis of bioclimatic variables, exhibits diverse characteristics across different regions of the country. Here's a summary of Rwanda's climate based on the values of the bioclimatic variables obtained during the study and shown in Figure 6.8.

(a) **Temperature:** Rwanda experiences a relatively mild climate with moderate temperatures throughout the year. However, there are notable variations in temperature across different elevations and geographic regions. Higher elevations, such as mountainous areas, tend to have cooler temperatures, especially during the coldest quarters (bio11), while lowland areas generally experience warmer conditions, particularly in the warmest quarters (bio10). The temperature annual range (bio7) indicates the extent of temperature fluctuations between seasons, with higher values suggesting greater variability.

(b) **Precipitation Patterns:** Precipitation distribution in Rwanda varies seasonally, with distinct wet and dry periods. Regions with higher precipitation of the wettest month (bio13) and lower precipitation seasonality (bio15) are likely to experience more consistent rainfall throughout the year, supporting lush vegetation and agricultural productivity. Conversely, areas with lower precipitation of the wettest month and higher precipitation seasonality may face challenges related to water availability and drought risk.

(c) **Temperature and Precipitation Extremes:** Rwanda's climate is characterized by occasional temperature and precipitation extremes, as indicated by variables such as temperature annual range (bio7) and precipitation seasonality (bio15).

(d) **Spatial Variability:** Rwanda's diverse topography, including highlands, plateaus, and lowland areas, results in spatial variations in climate. Elevation and slope influence temperature gradients and precipitation patterns, leading to the formation of distinct climatic zones. For example, mountainous regions may experience cooler temperatures and higher precipitation, supporting montane forests and unique biodiversity.

In summary, Rwanda's climate is characterized by moderate temperatures, seasonal rainfall patterns, and spatial variability influenced by topography as depicted in Figure 6.8.

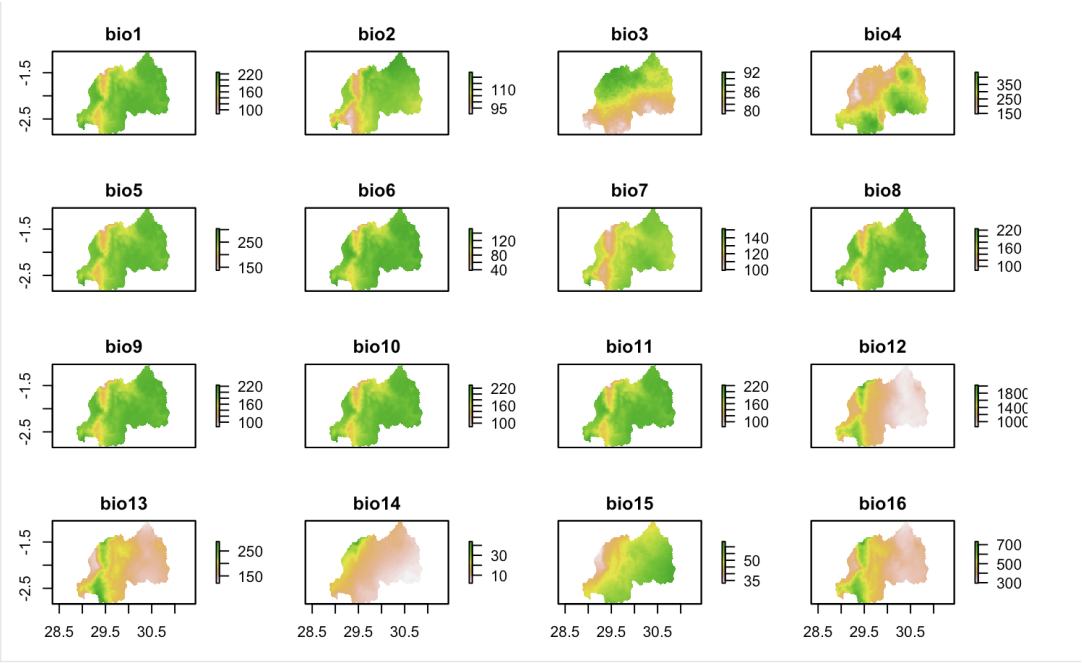


Figure 6.8: Bio1 - Bio16 Rwanda climate rasters

6.2 Data Preparation

6.2.1 Class Imbalance

The examination of class imbalance in the dataset revealed a significant disproportion between positive instances (YES) and negative instances (NO), with 2171 YES counts representing approximately 15.4% of the dataset, and 11947 NO counts accounting for approximately 84.6% of the dataset. This substantial class imbalance, illustrated in Figure 6.9 posed a challenge for ML model development, as it could lead to biased predictions and reduced model performance, particularly in accurately identifying positive instances.

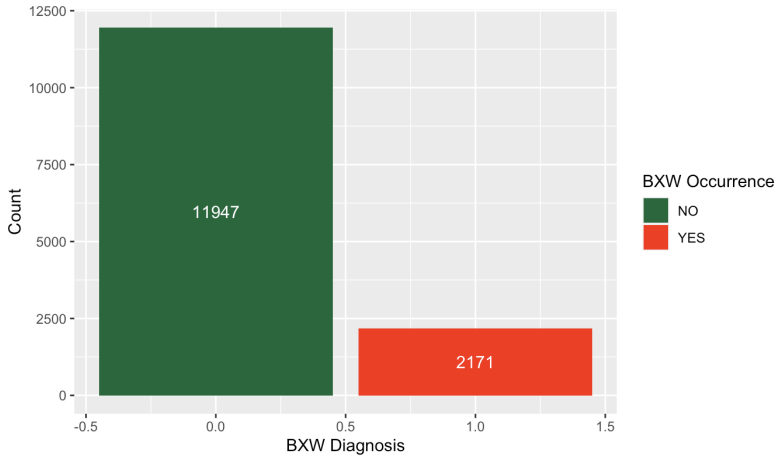


Figure 6.9: BXW diagnosis distribution

(a) **Euclidean Spatial Undersampling**

Equal numbers of NO points (2171) were selected to match the count of YES points. Upon plotting only the NO points in Figure 6.10, as the YES points were retained, it became evident that the chosen points exhibited a bias towards the north and eastern provinces. While the original map displays numerous NO points in the southwest region, the undersampled plot lacks any points in that area. This discrepancy suggests that the undersampling method failed to achieve geographical balance in the distribution of points.

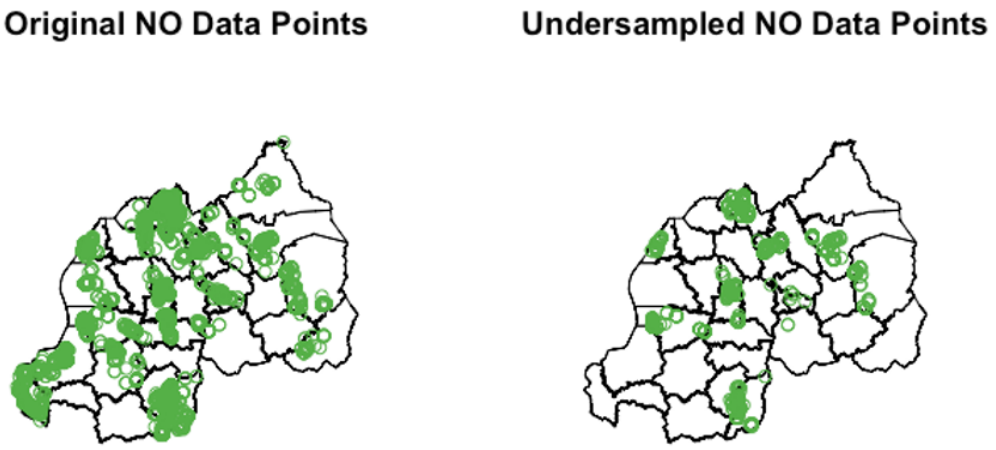


Figure 6.10: Euclidean spatial undersampling

(b) **Stratified Spatial Undersampling**

In the stratified sampling approach, an equal number of NO points were selected from different strata to match the count of YES points, while regions with more points were picked more to retain the original distribution. Upon examining the maps before and after sampling, illustrated in Figure 6.11, it is apparent that representation from each stratum was achieved. Before sampling, the NO points were distributed unevenly across the map, with some regions having more points than others. However, after stratified sampling, the distribution of NO points became more balanced. This suggests that the stratified sampling method effectively addressed the class imbalance issue while ensuring representation from each geographic region.

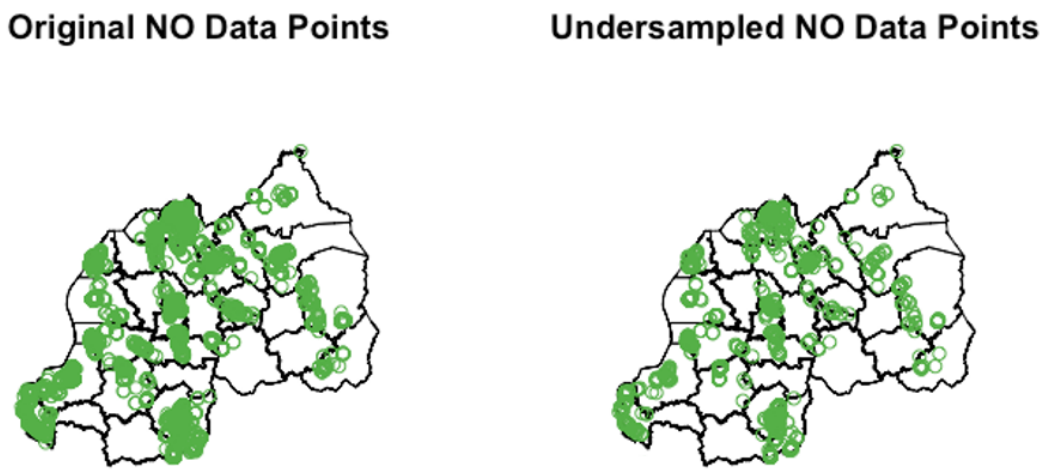


Figure 6.11: Stratified spatial undersampling

(c) **Background Points**

In the background points technique, an equal number of NO points were randomly selected from the agriculture cropland depicted in Figure 6.13, matching the count of YES points. The selected points were evenly distributed across all regions, resulting in a balanced representation as seen in Figure 6.12. This approach effectively addressed the class imbalance issue while ensuring that the dataset encompassed a comprehensive geographical coverage. By randomly selecting points from the agriculture cropland, the background points technique preserved the spatial diversity of the dataset, contributing to a more robust and representative sample for model training.

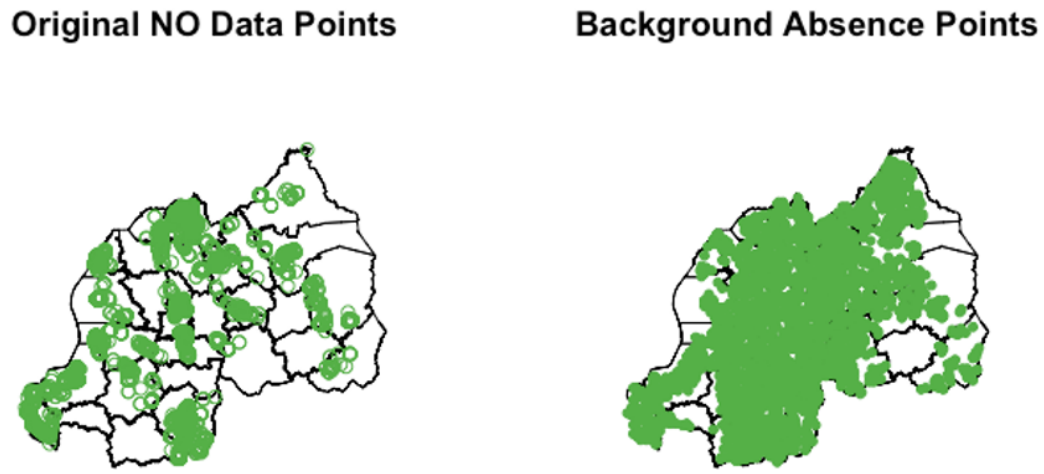


Figure 6.12: Background absence points

6.2.2 Rwanda's Cropland

Figure 6.13 below cropland raster for Rwanda provides a visual representation of agricultural activity across the country. In this dataset, white areas denote regions with no agricultural activity, indicating non-agricultural land cover or areas where cropland is not present. Conversely, areas with non-white colors represent regions where agricultural activities occur, highlighting the spatial extent of cropland areas.

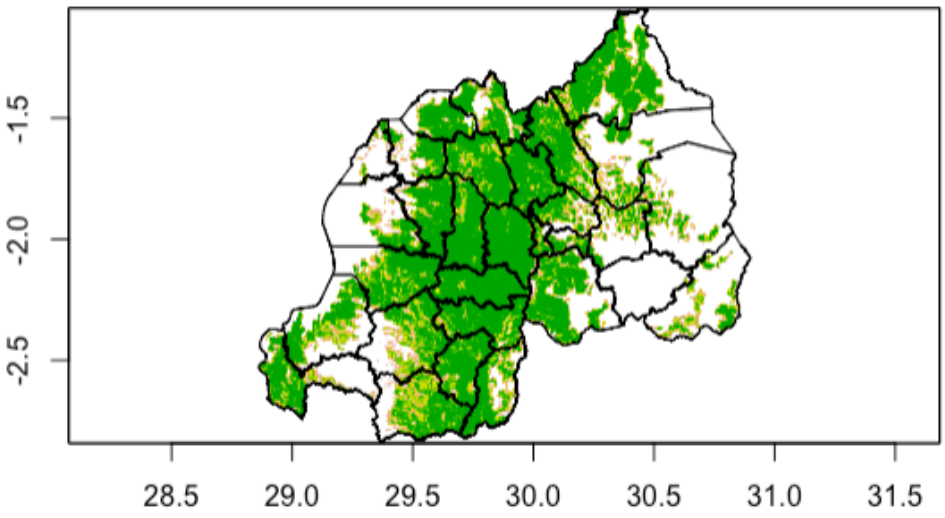


Figure 6.13: Rwanda's cropland (Source: `https://croplands.org`)

6.3 Machine Learning Modeling

6.3.1 Support Vector Machine

The results from the SVM validation plot in Figure 6.14 reveal a clear trend between the cost parameter and the model's accuracy. Initially, there is a noticeable increase in accuracy as the cost parameter rises, indicating improved classification performance. However, beyond a certain point (around 0.5), the rate of accuracy improvement slows down, suggesting diminishing returns with higher cost values. Despite this, the accuracy continues to show a slight but consistent increase as the cost parameter further rises. These findings emphasize the importance of selecting an appropriate cost parameter to strike a balance between model complexity and predictive accuracy in SVM classification.

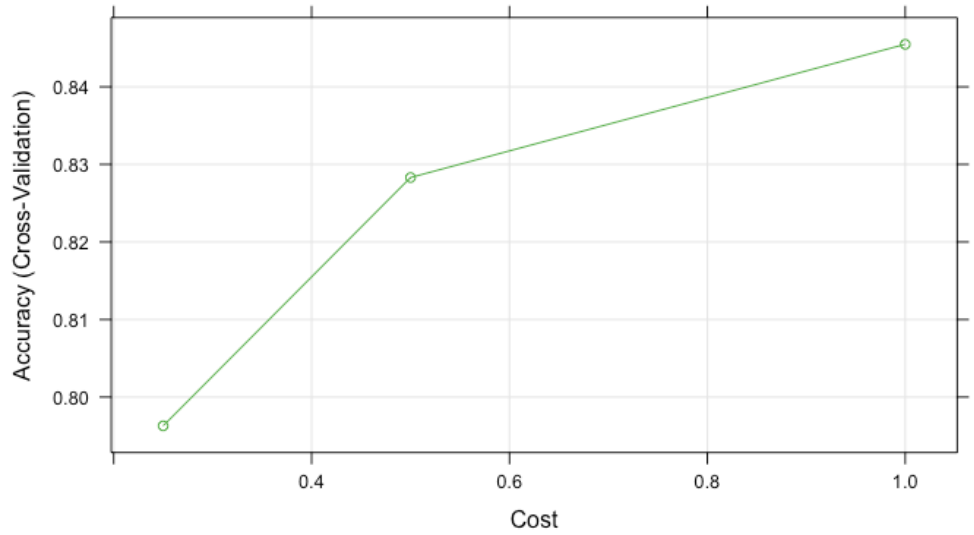


Figure 6.14: SVM validation plot

6.3.2 K-Nearest Neighbors

The KNN validation plot, in Figure 6.15, displays a clear relationship between the number of neighbors and the model's accuracy. Initially, with a lower number of neighbors, the accuracy tends to be higher, indicating that the model is capturing the local patterns more effectively. However, as the number of neighbors increases, the accuracy starts to decline steadily. This decline suggests that including more neighbors in the classification process introduces more noise or irrelevant information, leading to a decrease in predictive performance. Therefore, it is crucial to carefully select the number of neighbors to achieve the optimal balance between bias and variance in the KNN model.

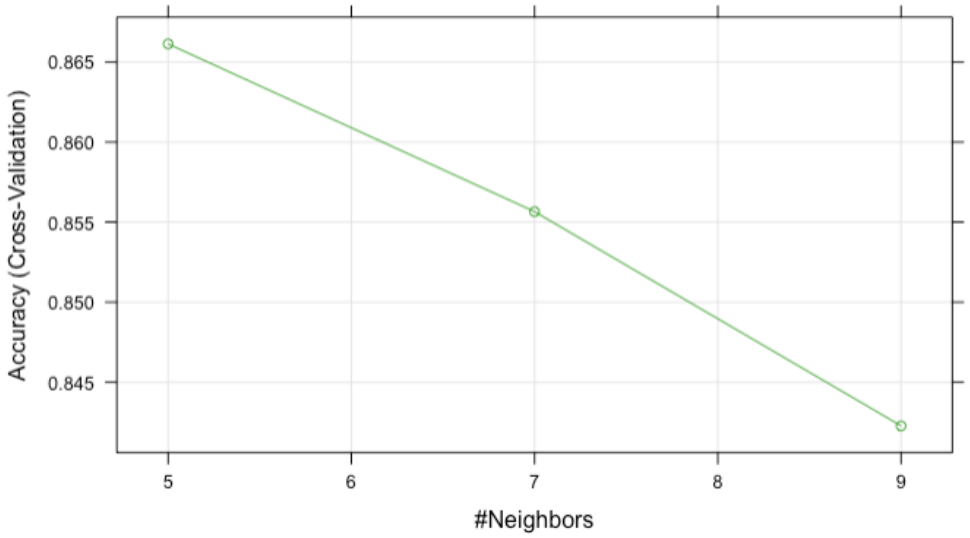


Figure 6.15: KNN validation plot

6.3.3 Random Forest

In the RF validation plot, shown in Figure 6.16, an apparent trend is observed wherein the accuracy decreases as the number of randomly selected predictors increases. Initially, with a smaller number of predictors, the model tends to have higher accuracy, indicating that a more focused selection of predictors contributes to better predictive performance. However, as the number of predictors increases, the accuracy gradually declines. This decline suggests that including more predictors introduces more noise or irrelevant features into the model, leading to a reduction in accuracy. Consequently, it is essential to strike a balance between the number of predictors and the model's predictive performance to achieve optimal results with RF.

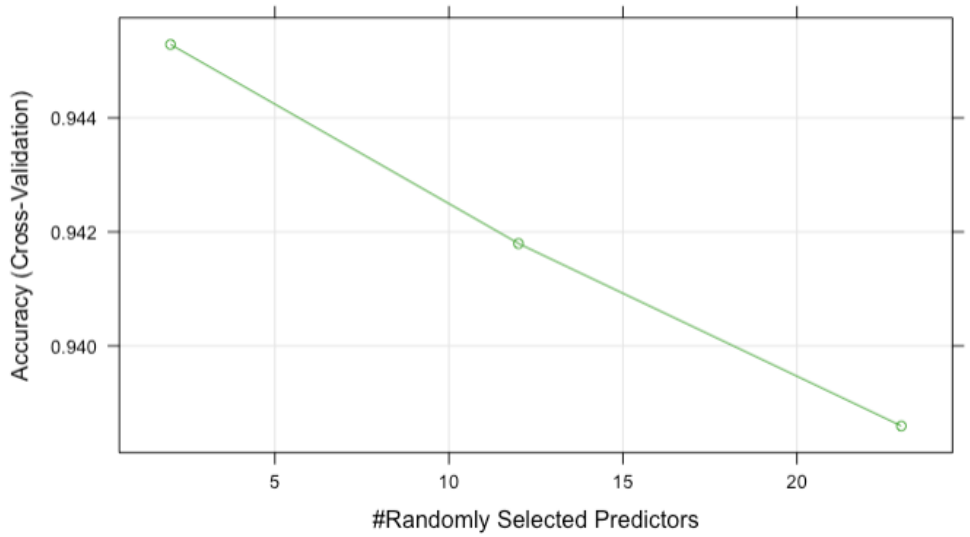


Figure 6.16: RF validation plot

6.3.4 Gradient Boosting Machine

A notable trend emerges in the GBM model, where accuracy shows consistent improvement with each iteration and as the maximum tree depth increases. This dynamic is distinctly observed across the plotted iterations in Figure 6.17, revealing a discernible pattern of enhanced accuracy over the course of model iterations. However, regardless of the maximum tree depth setting, the plotted results consistently depict an increasing trend in accuracy as the iterations progress. This observation underscores the beneficial impact of deeper trees and continued iterations on the model's predictive performance, reinforcing the effectiveness of the GBM algorithm in capturing complex relationships within the data.

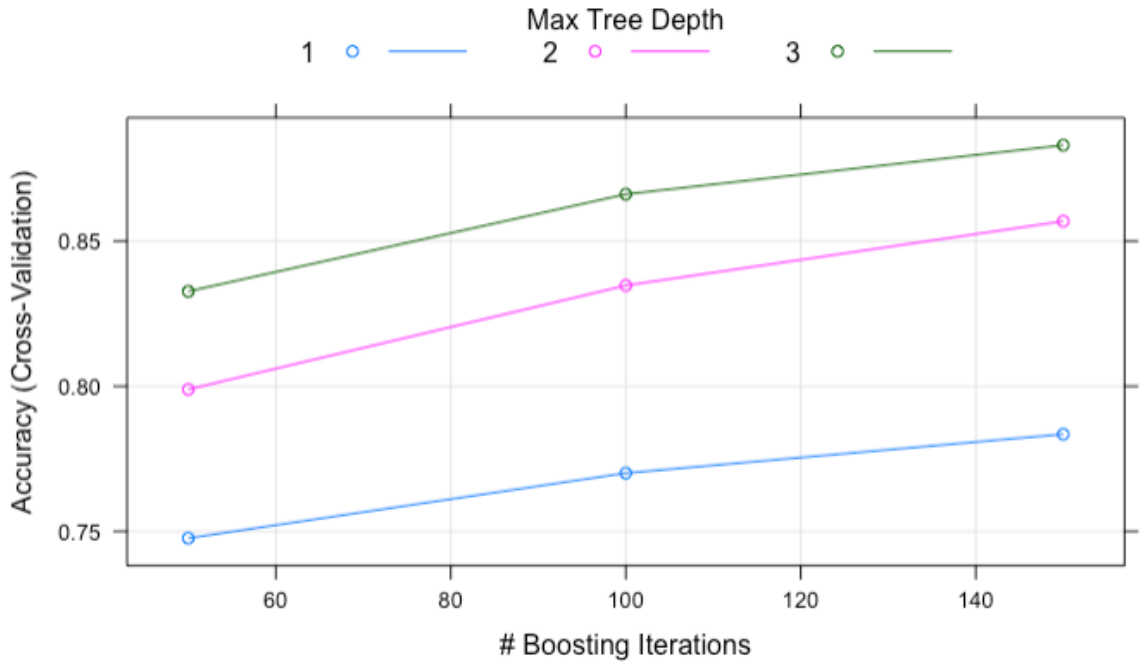


Figure 6.17: GBM validation plot

6.4 Model Evaluation and Optimization

6.4.1 Accuracy

The accuracy scores, depicted in Figure 6.18 shed light on the models' performance in classifying instances of BXW presence or absence. RF leads with an accuracy score of 0.94 (94%), showcasing its ability to make precise predictions by capturing intricate data patterns effectively. Following closely, GBM achieves an accuracy score of 0.89 (89%), demonstrating robust predictive performance and proficiency in the BXW classification task. KNN, with an accuracy score of 0.87 (%), exhibits moderate effectiveness in BXW classification, though slightly trailing RF and GBM.

Conversely, SVM presents the lowest accuracy score of 0.83 (89%). This could be attributed to the data's non-linear nature and the choice of hyperparameters. Despite its lower accuracy, SVM still shows promise in predictive performance, indicating potential for refinement.

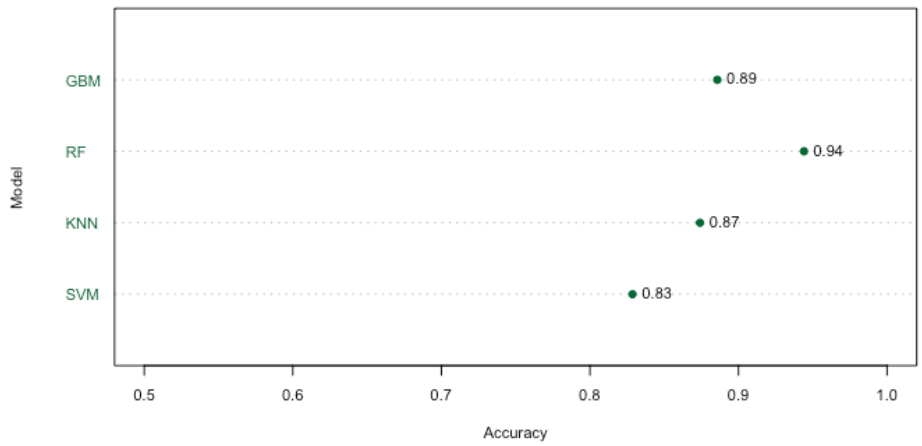


Figure 6.18: Accuracy comparison

6.4.2 Area Under the Curve

With an outstanding AUC score of 0.97 (97%), the RF model emerges as the top performer, showcasing its exceptional ability to correctly rank positive instances higher than negative ones across various thresholds. Following closely behind, the GBM model achieves an impressive AUC score of 0.96 (96%), indicating robust discriminative performance and proficiency in accurately distinguishing between positive and negative instances of BXW. KNN also demonstrates strong discriminative capabilities with an AUC score of 0.94 (94%), showcasing its effectiveness in accurately classifying instances of BXW presence or absence based on their predicted probabilities. SVM, with an AUC score of 0.91(91%), presents slightly lower discriminative performance compared to the other models. However, SVM still exhibits reasonable discriminative power, suggesting potential for further optimization and improvement.

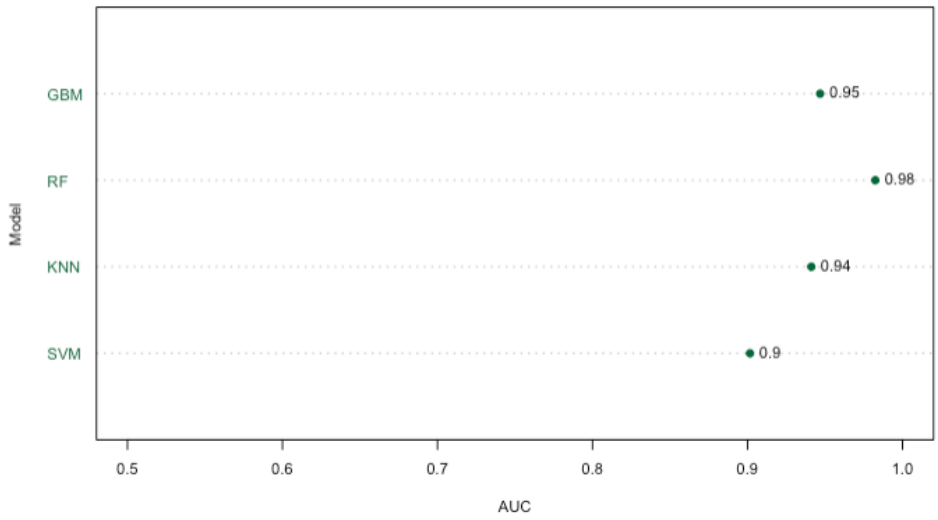


Figure 6.19: AUC comparison

6.4.3 Recall

The recall scores, shown in Figure 6.20 provide valuable insights into the models' abilities to correctly identify positive instances of BXW presence among all actual positive instances. RF and KNN stands out with a recall score of 0.94 (94%), indicating their high capability in capturing most positive instances of BXW presence, thereby minimizing false negatives. GBM follows closely with a recall score of 0.91 (91%), demonstrating its effectiveness in correctly identifying a substantial proportion of positive instances while maintaining a relatively low rate of false negatives. SVM, with a recall score of 0.84 (84%), demonstrates moderate performance in identifying positive instances of BXW presence, with some room for improvement compared to the other models.

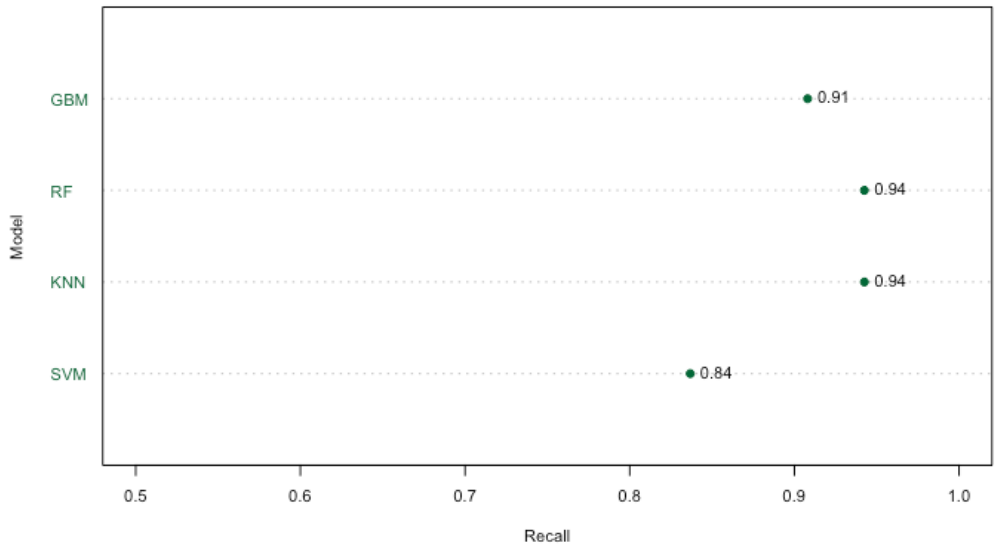


Figure 6.20: Recall comparison

6.4.4 Precision

RF achieves the highest precision score of 0.95 (95%), indicating its strong capability in making positive predictions that are actually correct. This suggests that RF effectively minimizes the occurrence of false positives while maximizing true positive predictions. Following closely behind is GBM with a precision score of 0.87 (87%), demonstrating its effectiveness in accurately classifying positive predictions, albeit with a slightly higher false positive rate compared to RF. KNN and SVM both exhibit precision scores of 0.83 (83%), indicating their moderate performance in avoiding false positives and accurately classifying positive predictions.

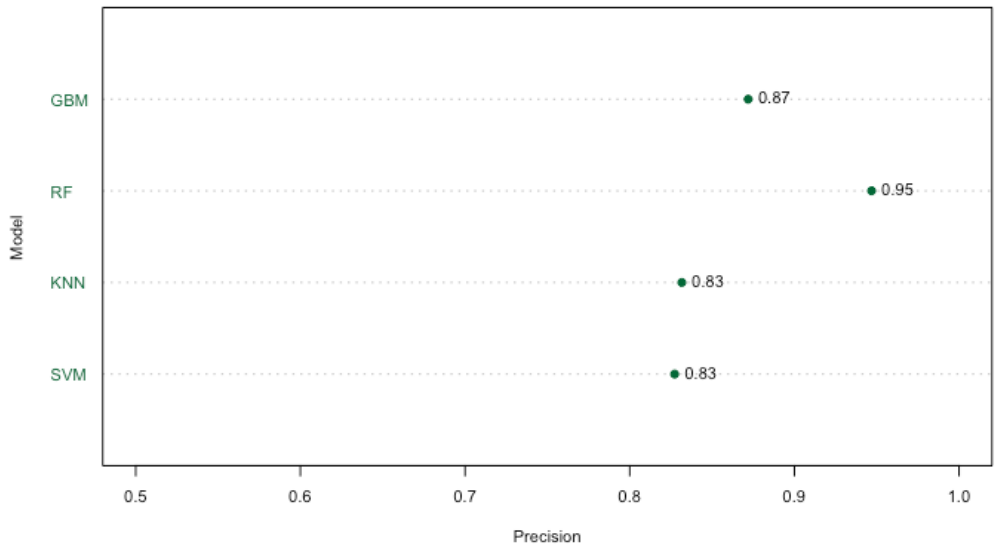


Figure 6.21: Precision comparison

6.4.5 F1-Score

RF achieves the highest F1-score of 0.94 (94%), indicating its ability to achieve a balance between precision and recall. This suggests that RF effectively minimizes both false positives and false negatives, leading to a robust overall performance. Following RF is GBM with an F1-score of 0.89 (89%), demonstrating its strong performance in balancing precision and recall. KNN achieves an F1-score of 0.88 (88%), indicating its moderate performance in balancing precision and recall, although slightly lower compared to RF and GBM. SVM exhibits the lowest F1-score of 0.83 (83%) among the evaluated models, suggesting its relatively lower effectiveness in achieving a balance between precision and recall. Despite its lower F1-score, SVM still demonstrates reasonable performance as shown in Figure 6.22.

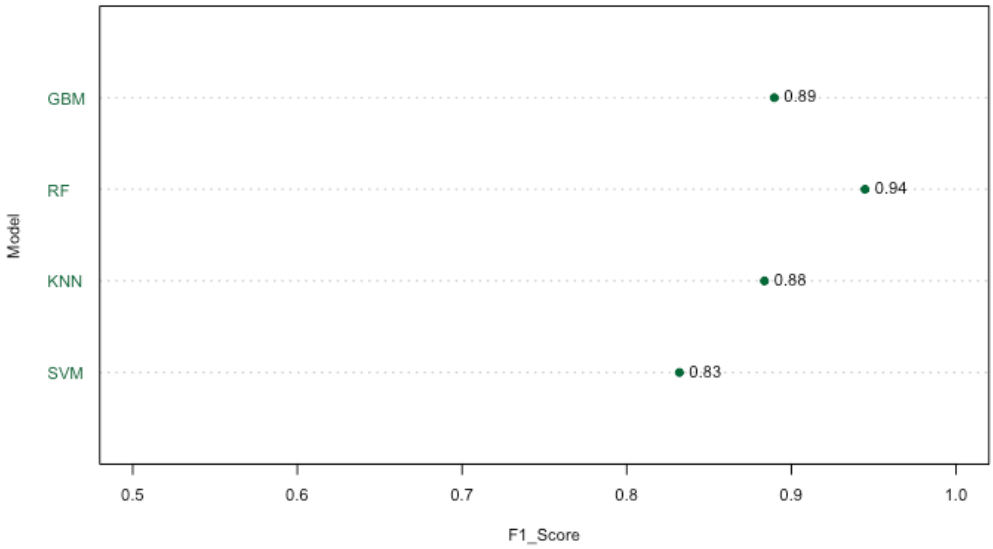


Figure 6.22: F1-Score comparison

RF emerged as the top-performing model across all the evaluation metrics. With an accuracy score of 96%, RF demonstrated exceptional capability in correctly classifying instances of BXW presence or absence. Additionally, RF achieved an AUC of 94%, indicating strong discriminatory power in distinguishing between positive and negative instances. Furthermore, RF exhibited impressive recall, precision, and F1-score values of 94%, 95%, and 94%, respectively. These high scores across multiple metrics underscore the robustness and effectiveness of the RF algorithm in capturing intricate patterns within the dataset and making accurate predictions.

6.4.6 Model Optimization

As RF emerged as the top-performing model across all metrics, further optimization was conducted to enhance its performance. The plot in Figure 6.23 shows how different values of **mtry** and **ntree** impact the model's accuracy. We can note that the accuracy peaks when the number of trees is set to 1500. Interestingly, increasing the **mtry** parameter consistently lowers accuracy regardless of the number of trees. This suggests that including more features in the random selection process during tree construction could lead to overfitting. Therefore, finding the right balance between **mtry** and **ntree** is essential for achieving optimal model performance.

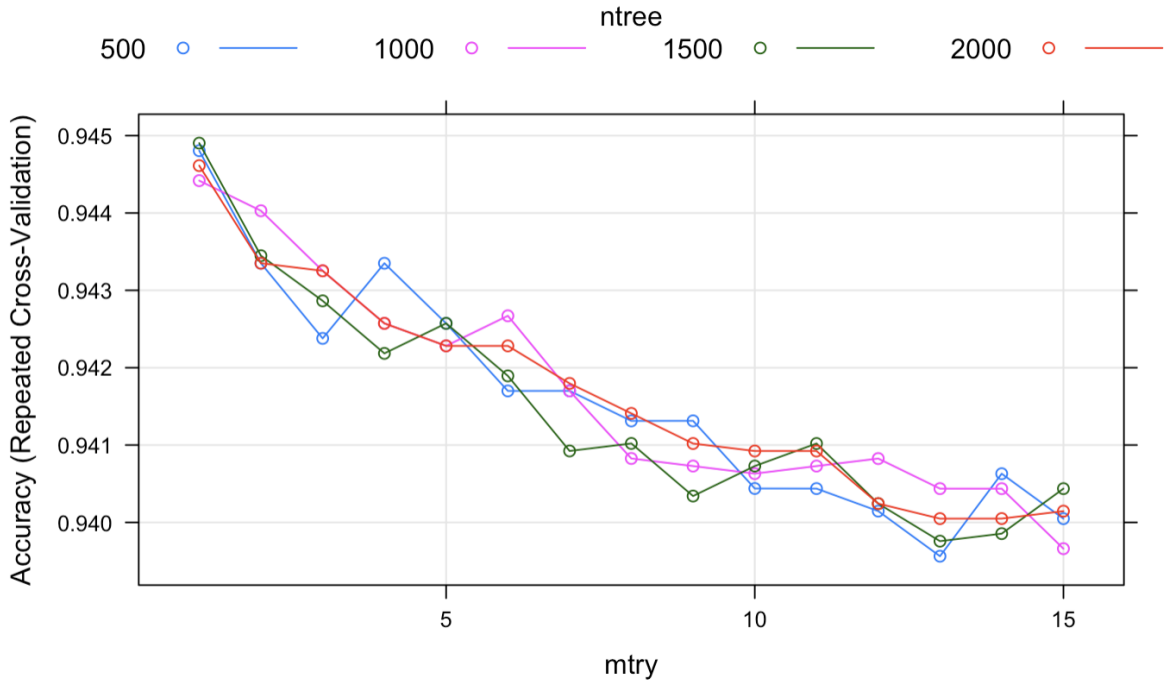


Figure 6.23: Effect of **mtry** and **ntree** on model accuracy

After optimization, the RF model showed slight improvements across various performance metrics. The AUC increased marginally from 98.19% to 98.23%, indicating improved overall model discrimination. The model's Recall remained consistent at 94.25%, while Precision slightly improved from 93.82% to 94.25% after optimization. This indicates that the model maintained its ability to correctly identify the majority of actual positive instances, with a Recall rate of 94.25%. The increase in Precision suggests that the model became more accurate in identifying true positive instances while reducing false positives. Consequently, the F1 Score, representing the harmonic mean of Recall and Precision, as shown in Equation 13, increased from 94.04% to 94.25%, indicating a more balanced performance in capturing both false positives and false negatives.

Moreover, the Accuracy of the RF model observed a notable increase from 93.94% to 94.29% after optimization. This signifies an enhanced ability of the model to correctly classify both positive and negative instances of BXW occurrences, leading to a higher overall prediction accuracy.

Overall, these results indicate that the optimization process refined the RF model, resulting in improved performance metrics across the board, thereby strengthening its efficacy in predicting and classifying BXW occurrences.

6.5 Deployment

6.5.1 BXW Environmental Drivers

The analysis of variable importance highlighted in Figure 6.24 reveals the significant environmental factors influencing the occurrence of BXW. Among these factors, precipitation of the wettest month emerges as the most influential predictor, with a variable importance score of 100. This suggests that the amount of rainfall during the wettest month plays a crucial role in determining the presence or absence of BXW.

Following closely behind is elevation, with a variable importance score of 97.03, indicating its strong influence on BXW occurrence. Elevation likely affects BXW distribution by influencing factors such as temperature and moisture levels, which are known to impact disease development.

Other notable environmental factors include the maximum temperature of the warmest month, mean diurnal range, and annual mean temperature, which all exhibit substantial variable importance scores above 70. These variables reflect the importance of temperature-related factors in BXW occurrence, highlighting the sensitivity of the disease to climatic conditions.

Furthermore, factors related to precipitation, such as precipitation of the warmest and wettest quarters, as well as annual precipitation, also demonstrate considerable importance, underscoring the role of rainfall patterns in BXW epidemiology.

The analysis also identifies topographic factors like slope and hillshade as influential predictors, suggesting that terrain characteristics may influence the spread and severity of BXW.

The analysis further reveals environmental factors with a variable importance score of 0, indicating negligible influence on BXW occurrence. These factors include temperature seasonality, isothermality (uniformity of temperature throughout the year), and aspect. While these variables may have some impact on local environmental conditions, their contribution to BXW distribution appears to be minimal compared to other factors identified in the analysis.

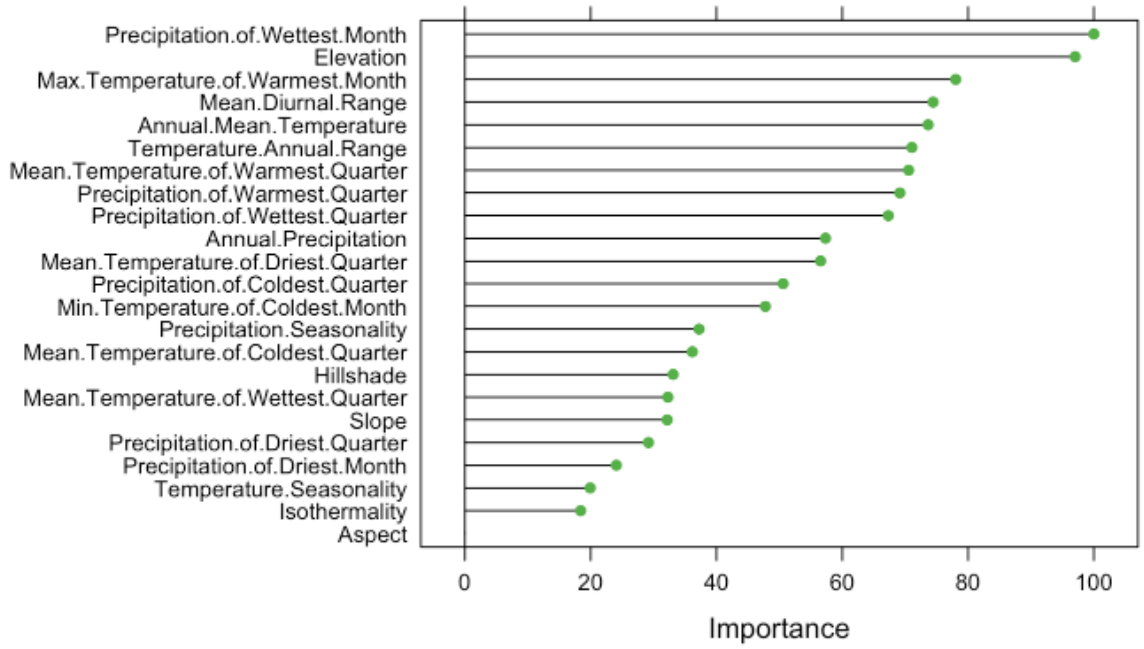


Figure 6.24: BXW environmental drivers

Overall, the variable importance analysis provides valuable insights into the environmental drivers of BXW occurrence, informing strategies for disease management and mitigation efforts.

6.5.2 Impact of Environmental Factors on BXW

The interaction between BXW and environmental variables unveils distinctive trends shedding light on its occurrence patterns as illustrated in Figure 6.25. Specifically selected for their significant variable importance scores exceeding 60, partial plots were generated to explore how BXW occurrence responds to these influential factors. Analyzing these partial plots uncovers valuable insights into the dynamics of BXW risk across a range of environmental conditions:

a) Precipitation of Wettest Month: The plot suggests that BXW risk remains relatively stable at lower precipitation levels but sharply decreases beyond a threshold. This observation indicates a potential mitigation of BXW risk with higher precipitation levels, as excessive moisture might be unfavorable for the pathogen's survival or spread.

b) Elevation: BXW risk exhibits an interesting pattern concerning elevation. It shows a gradual rise with increasing elevation until a certain threshold, after which there's a slight decrease and stabilization. This trend implies that BXW prevalence might be influenced by specific

altitudinal ranges, possibly due to variations in climate or ecological conditions.

c) Max Temperature of Warmest Month and Annual Mean Temperature: Both partial plots depict similar trends, with BXW risk increasing with temperature up to a certain threshold, decreasing afterward, and stabilizing. This observation underscores the critical role of temperature in influencing BXW prevalence, with optimal conditions existing within a lower temperature range.

d) Mean Diurnal Range: The partial plot reveals a pattern where BXW risk increases up to a certain range of mean diurnal temperature variation, decreases afterward, and stabilizes. This pattern suggests that BXW occurrence may be influenced by variations in temperature between day and night, with optimal conditions existing within the higher range.

e) Temperature Annual Range: The partial plot demonstrates fluctuations in BXW risk in response to temperature variability, indicating a nonlinear relationship. This suggests that temperature variability, in addition to mean temperatures, plays a role in influencing BXW occurrence, possibly interacting with other environmental factors.

f) Mean Temperature of Warmest Quarter and Mean Temperature of Driest Quarter: Both partial plots show BXW risk increasing until a certain temperature threshold, decreasing afterward, and stabilizing. This highlights the critical influence of temperature during specific periods in shaping BXW prevalence, with optimal conditions existing within specific temperature ranges.

g) Annual Precipitation and Precipitation of Wettest Quarter: These partial plots depict BXW risk stabilizing until a certain precipitation level, followed by fluctuations and stabilization. This suggests a nonlinear relationship between precipitation levels and BXW occurrence.

h) Precipitation of Coldest Quarter: The partial plot indicates fluctuations in BXW risk in response to precipitation during cold periods, with optimal conditions existing within specific ranges.

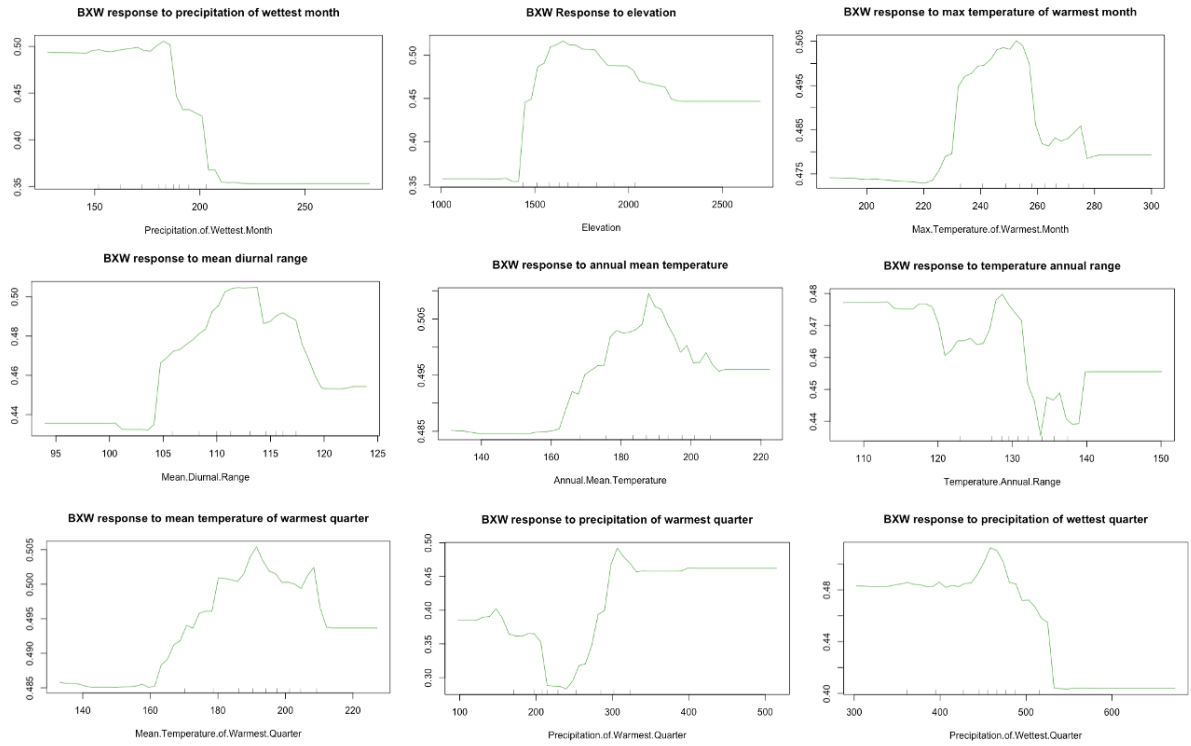


Figure 6.25: BXW response to environmental factors

6.5.3 Mapping BXW Habitat Suitability

The habitat suitability map for BXW in Rwanda provides valuable insights into the distribution and risk levels associated with BXW occurrence across the country. The color scheme employed in the map effectively categorizes regions based on their risk levels, with green indicating low risk, yellow signifying medium risk, and red representing high risk. Additionally, the presence of white areas on the map indicates regions where agricultural activities are absent.

Upon analysis of the habitat suitability map, it is evident that the majority of Rwanda exhibits a low risk of BXW occurrence, as indicated by the prevalence of green areas. However, there are notable exceptions, particularly in specific regions characterized by red shading, which denotes areas of heightened risk. Specifically, districts such as Burera, Musanze, Gisagare, Nyamasheke, and Rusizi emerge as hotspots with elevated BXW risk levels. These regions warrant closer attention and targeted intervention strategies to mitigate the potential impact of BXW outbreaks on agricultural productivity and livelihoods.

Furthermore, the eastern part of Rwanda shows lower risk levels for BXW, suggesting less favorable conditions for the disease. This understanding of BXW risk distribution is vital for effective resource allocation, disease management, and policy decisions to protect agricultural

productivity and ensure food security.

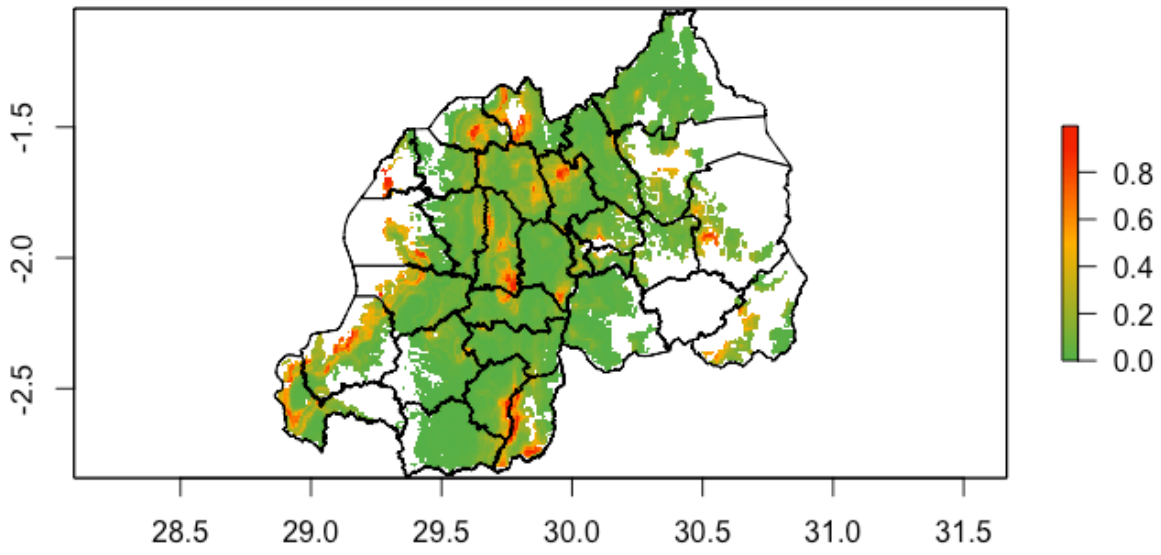


Figure 6.26: BXW habitat suitability map

6.6 Summary

The results indicate that RF outperformed all other models across various evaluation metrics, demonstrating its suitability for constructing a ML-driven EWS for BXW. Further tuning of the model with parameters such as **mtry** and **ntree** led to slight performance improvements.

The analysis of variable importance in Section 6.5.1 unveiled several environmental factors with significant influence on BXW occurrence, particularly those with importance scores above 55. These factors include Precipitation of Wettest Month, Elevation, Max Temperature of Warmest Month, Mean Diurnal Range, Annual Mean Temperature, Temperature Annual Range, Mean Temperature of Warmest Quarter, Precipitation of Warmest Quarter, and Annual Precipitation. These findings closely correspond with previous research by (Kilwenge et al., 2023), who likewise identified elevation, annual precipitation, and quarterly temperature and precipitation metrics as pivotal drivers of BXW occurrence.

The habitat suitability map presented in Figure 6.26 highlights regions in Rwanda that are particularly susceptible to BXW, notably the northern, western, and southern areas. These findings corroborate those of (Kilwenge et al., 2023), further emphasizing the importance of these regions in BXW management efforts.

Chapter 7: Conclusions, Recommendations and Future Work

7.1 Conclusion

Our study highlights the significant role of ML in addressing the specific agricultural challenge of combating BXW in Rwanda. Through the utilization of RF models, we successfully crafted a habitat suitability map, delineating regions with varying degrees of susceptibility to BXW outbreaks. This precise tool equips stakeholders with actionable insights, enabling targeted resource allocation, the implementation of tailored disease management protocols, and the formulation of strategic policies aimed at strengthening agricultural productivity and fortifying food security. Furthermore, our implementation of a ML-driven early warning system, seamlessly integrated with APIs and SMS alerts, exemplifies the transformative potential of technology in fostering proactive disease surveillance and management strategies. By harnessing these tools, we can empower farmers to anticipate and mitigate the spread of BXW, thereby safeguarding crop yields and livelihoods. The integration of ML methodologies into agricultural frameworks holds immense promise in fortifying resilience against emerging threats. By leveraging advanced technologies, we can forge a path toward sustainable food production practices, ensuring the long-term viability of agricultural systems and the well-being of communities reliant upon them. This study serves as a testament to the pivotal role that ML can play in addressing complex agricultural issues, ushering in a future characterized by enhanced efficiency, adaptability, and resilience in the face of evolving challenges.

7.2 Recommendations

Recommendations for the implementation of the study's findings in real-life scenarios are as follows: Firstly, fostering stakeholder engagement and collaboration, including government agencies, agricultural organizations, and local communities, is essential for effective dissemination and adoption of the developed tools and strategies for combating BXW. Secondly, providing training programs and capacity-building initiatives tailored to empower farmers, extension workers, and policymakers with the knowledge and skills to utilize the habitat suitability map and early warning system effectively is crucial. Thirdly, integrating the developed tools into existing agricultural systems and decision-making processes, such as land-use planning and disease surveillance networks, will enhance their impact. Customization of the tools to suit local contexts, considering factors like agroecological zones and socio-economic conditions,

is also necessary. Lastly, advocating for supportive policies and institutional frameworks that recognize the importance of ML-driven solutions in agriculture can facilitate their adoption.

7.3 Future Works

In considering avenues for future research, several potential areas for expansion and refinement emerge from this study. These include geographical expansion to regions facing similar agricultural challenges, necessitating comprehensive data collection and adaptation of ML models to account for variations in environmental factors, agricultural practices, and disease dynamics. Furthermore, the exploration of more sophisticated algorithms or ensemble methods presents an opportunity to enhance the accuracy and robustness of the habitat suitability mapping and early warning system. Techniques such as deep learning could refine predictive models, while ensemble methods could mitigate algorithm limitations and yield more reliable predictions. Additionally, the integration of real-time data sources such as satellite imagery and weather forecasts offers promising prospects for enabling proactive disease management strategies. Satellite imagery can provide insights into vegetation health, while weather forecasts can predict disease outbreaks by identifying favorable conditions for pathogen spread. Leveraging real-time data streams enables timely interventions to mitigate the impact of BXW and other agricultural diseases. These future research directions aim to expand the applicability and effectiveness of our ML-driven approaches, contributing to the development of more resilient and sustainable agricultural systems globally.

Bibliography

Ainembabazi, J. H., Tripathi, L., Rusike, J., Abdoulaye, T., and Manyong, V. (2015). Ex-ante economic impact assessment of genetically modified banana resistant to xanthomonas wilt in the great lakes region of africa. *PLOS ONE*, 10(9):1–21.

Alabi, T. R., Adewopo, J., Duke, O. P., and Kumar, P. L. (2022). Banana mapping in heterogenous smallholder farming systems using high-resolution remote sensing imagery and machine learning models with implications for banana bunchy top disease surveillance. *Remote Sensing*, 14.

Araújo, M. B., Anderson, R. P., Barbosa, A. M., Beale, C. M., Dormann, C. F., Early, R., Garcia, R. A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R. B., Zimmermann, N. E., and Rahbek, C. (2019). Standards for distribution models in biodiversity assessments. *Science Advances*, 5(1):eaat4858.

Blomme, G., Ocimati, W., Sivirihauma, C., Vutseme, L., Mariamu, B., Kamira, M., van Schagen, B., Ekboir, J., and Ntamwira, J. (2017). A control package revolving around the removal of single diseased banana stems is effective for the restoration of xanthomonas wilt infected fields. *European Journal of Plant Pathology*, 149:385–400.

Burns, P., Clark, M., Salas, L., Hancock, S., Leland, D., Jantz, P., Dubayah, R., and Goetz, S. J. (2020). Incorporating canopy structure from simulated gedi lidar into bird species distribution models. *Environmental Research Letters*, 15(9):095002.

Damme, J. V., Ansoms, A., and Baret, P. V. (2014). Agricultural innovation from above and from below: Confrontation and integration on rwanda's hills. *African Affairs*, 113:108–127.

Deutsch, C. A., Tewksbury, J. J., Tigchelaar, M., Battisti, D. S., Merrill, S. C., Huey, R. B., and Naylor, R. L. (2018). Increase in crop losses to insect pests in a warming climate. *Science*, 361(6405):916–919.

Ding, X., Liu, J., Yang, F., and Cao, J. (2021). Random radial basis function kernel-based support vector machine. *Journal of the Franklin Institute*, 358(18):10121–10140.

Domingues, T., Brandão, T., and Ferreira, J. C. (2022). Machine learning for detection and prediction of crop diseases and pests: A comprehensive survey. *Agriculture*, 12(9).

El Miloudi, K. and Ettouhami, A. (2018). A multiview formal model of use case diagrams using z notation: Towards improving functional requirements quality. *Journal of Engineering*, 2018:6854920.

Gardner, A. S., Maclean, I. M., and Gaston, K. J. (2019). Climatic predictors of species distributions neglect biophysiologically meaningful variables. *Diversity and Distributions*, 25(8):1318–1333.

Jackson, N., William, T., Gertrude, N., Nicholas, N., Wellington, J., Innocent, N., Leornard, M., Privat, N., Celestin, N., Svetlana, G., Ivan, R., Fina, O., and Eldad, K. (2015). Adverse impact of banana xanthomonas wilt on farmers livelihoods in eastern and central africa. *African Journal of Plant Science*, 9:279–286.

Karmani, M. Z., Mustafa, G., Sarwar, N., Wajid, S., Qureshi, J., and Siddque, S. (2020). *A Review of Star Schema and Snowflakes Schema*, pages 129–140.

Kilwenge, R., Adewopo, J., Manners, R., Mwizerwa, C., Kabirigi, M., Gaidashova, S., and Schut, M. (2021). Ict4bxw — ict4bxw.com. `https://www.ict4bxw.com/`. [Accessed 26-10-2023].

Kilwenge, R., Adewopo, J., Manners, R., Mwizerwa, C., Kabirigi, M., Gaidashova, S., and Schut, M. (2023). Climate-related risk modeling of banana xanthomonas wilt disease incidence in the cropland area of rwanda. *Plant Disease*.

Koc, H., Erdoğan, A., Barjakly, Y., and Peker, S. (2021). Uml diagrams in software engineering research: A systematic literature review. *Proceedings*, 74:13.

Lasso, E., Corrales, D. C., Avelino, J., de Melo Virginio Filho, E., and Corrales, J. C. (2020). Discovering weather periods and crop properties favorable for coffee rust incidence from feature selection approaches. *Computers and Electronics in Agriculture*, 176:105640.

Lin, C.-T. and Chiu, C.-A. (2020). Comparison of predictor selection procedures in species distribution modeling: A case study of fagus hayatae. *CERNE*.

Luque, A., Carrasco, A., Martín, A., and de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91:216–231.

Machovina, B., Feeley, K., and Machovina, B. (2016). Uav remote sensing of spatial variation in banana production. *Crop and Pasture Science*, 67.

McCampbell, Mariette, and Adewopo, J. (2018a). Ict4bxw report on baseline survey of banana farmers in rwanda 2 ict4bxw baseline report.

McCampbell, M., Schut, M., den Bergh, I. V., van Schagen, B., Vanlauwe, B., Blomme, G., Gaidashova, S., Njukwe, E., and Leeuwis, C. (2018b). Xanthomonas wilt of banana (bxw) in central africa: Opportunities, challenges, and pathways for citizen science and ict-based control and prevention strategies. *NJAS - Wageningen Journal of Life Sciences*, 86-87:89–100.

Obi, J. C. (2023). A comparative study of several classification metrics and their performances on data. *World Journal of Advanced Engineering Technology and Sciences*, 8(1):308–314.

Ochola, D., Boekelo, B., van de Ven, G. W., Taulya, G., Kubiriba, J., van Asten, P. J., and Giller, K. E. (2022). Mapping spatial distribution and geographic shifts of east african highland banana (musa spp.) in uganda. *PLoS ONE*, 17.

Patil, D. N. N., Department of Computer Engineering, R. C. Patel Institute of Technology, Shirpur, India, Saiyyad, M. M. A. M., and Department of Computer Engineering, R. C. Patel Institute of Technology, Shirpur, India (2019). Machine learning technique for crop recommendation in agriculture sector. *Int. J. Eng. Adv. Technol.*, 9(1):1359–1363.

Pradervand, J.-N., Anne, D., Pellissier, L., Guisan, A., and Randin, C. (2014). Very high resolution environmental predictors in species distribution models: Moving beyond topography? *Progress in Physical Geography*, 38:79–96.

Ricciardi, V., Mehrabi, Z., Wittman, H., James, D., and Ramankutty, N. (2021). Higher yields and more biodiversity on smaller farms. *Nature Sustainability*, 4(7):651–657.

Rietveld, A. M. R., Blomme, P., Gaidashova, G., Ocimati, S., and Ntamwira, J. W. (2020). A superior technology to control banana xanthomonas wilt (bxw) in rwanda.

Ritchie, H. (2021). Smallholders produce one-third of the world's food, less than half of what many headlines claim. *Our World in Data*. https://ourworldindata.org/smallholder-food-production.

RuchaReads (2021). Crisp dm framework. https://ruchareads.wordpress.com/2021/03/29/1-crisp-dm-framework/. [Accessed 05-04-2024].

Selvaraj, M. G., Vergara, A., Montenegro, F., Ruiz, H. A., Safari, N., Raymaekers, D., Ocimati, W., Ntamwira, J., Tits, L., Omondi, A. B., and Blomme, G. (2020). Detection of banana plants and their major diseases through aerial images and machine learning methods: A case study in dr congo and republic of benin. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:110–124.

Small, I. M., Joseph, L., and Fry, W. E. (2015). Development and implementation of the blight-pro decision support system for potato and tomato late blight management. *Computers and Electronics in Agriculture*, 115:57–65.

Suwanda, R., Syahputra, Z., and Zamzami, E. M. (2020). Analysis of euclidean distance and manhattan distance in the k-means algorithm for variations number of centroid k. *Journal of Physics: Conference Series*, 1566(1):012058.

Tan, H. (2021). Machine learning algorithm for classification. *Journal of Physics: Conference Series*, 1994(1):012016.

UN (2022). The sustainable development goals report.

Vurro, M., Bonciani, B., and Vannacci, G. (2010). Emerging infectious diseases of crop plants in developing countries: impact on agriculture and socio-economic consequences. *Food Security*, 2(2):113–132.

WorldClim (2020-2022). WorldClim — worldclim.org. https://worldclim.org/. [Accessed 23-03-2024].

Xiao, Q., Li, W., Kai, Y., Chen, P., Zhang, J., and Wang, B. (2019). Occurrence prediction of pests and diseases in cotton on the basis of weather factors by long short term memory network. *BMC Bioinformatics*, 20(25):688.

Xiong, J., Thenkabail, P. S., Gumma, M. K., Teluguntla, P., Poehnelt, J., Congalton, R. G., Yadav, K., and Thau, D. (2017). Automated cropland mapping of continental africa using google earth engine cloud computing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 126:225–244.

Ye, H., Huang, W., Huang, S., Cui, B., Dong, Y., Guo, A., Ren, Y., and Jin, Y. (2020). Recognition of banana fusarium wilt based on uav remote sensing. *Remote Sensing*, 12(6).

Zhang, J., Huang, Y., Pu, R., Gonzalez-Moreno, P., Yuan, L., Wu, K., and Huang, W. (2019). Monitoring plant diseases and pests through remote sensing technology: A review. *Computers and Electronics in Agriculture*, 165:104943.

Zhang, S., Li, X., Ba, Y., Lyu, X., Zhang, M., and Li, M. (2022). Banana fusarium wilt disease detection by supervised and unsupervised methods from uav-based multispectral imagery. *Remote Sensing*, 14(5).

Appendices

Appendix A: Ethical Review Approval

Strathmore UNIVERSITY

1st March 2024

Ms Owuor Caroline,
caroline.owuor@strathmore.edu

Dear Ms Owuor,

RE: Developing an Early Warning System for Banana Xanthomonas Wilt (BXW) in Rwanda

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** research proposal. Your application reference number is **SU-ISERC2020/24.** The approval period is from **1st March 2024 to 28th February 2025.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

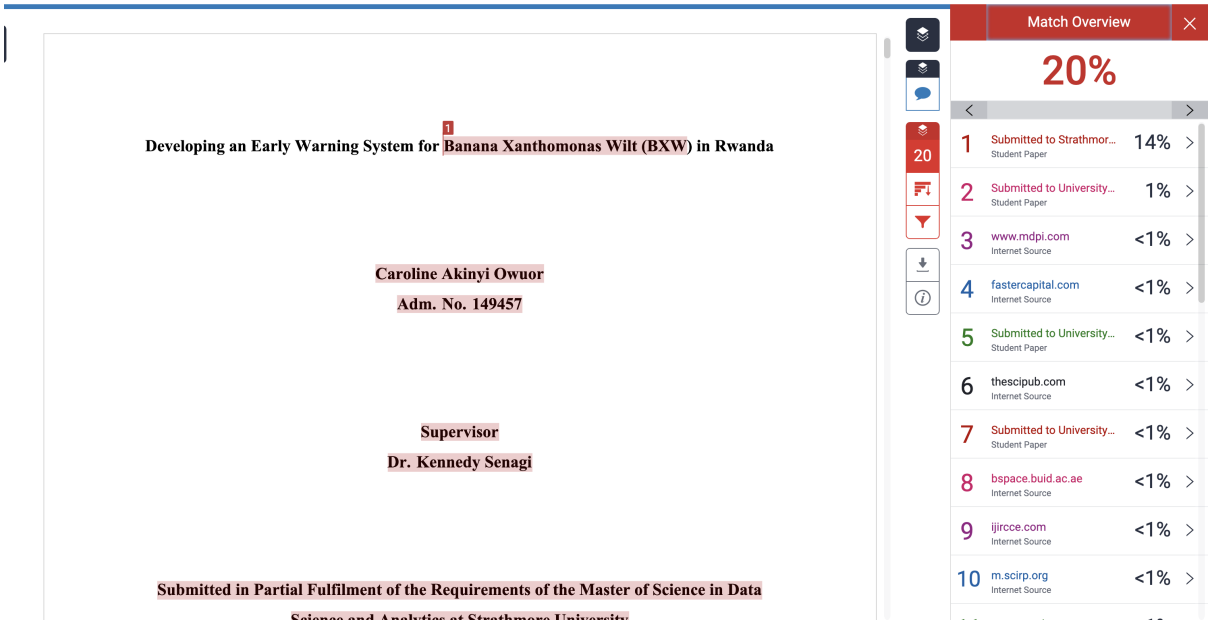
Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

STRATHMORE UNIVERSITY INSTITUTIONAL SCIENTIFIC AND ETHICAL REVIEW COMMITTEE
(SU-ISERC)
01-Mar-2024
Emailiethicsreview@strathmore.edu
P.O BOX 59857-00200
NAIROBI-KENYA

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

Appendix B: Plagiarism Report



Developing an Early Warning System for Banana Xanthomonas Wilt (BXW) in Rwanda

Caroline Akinyi Owuor

Adm. No. 149457

Supervisor

Dr. Kennedy Senagi

Submitted in Partial Fulfilment of the Requirements of the Master of Science in Data Science and Analytics at Strathmore University

Match Overview

20%

1	Submitted to Strathmor... Student Paper	14%
2	Submitted to University... Student Paper	1%
3	www.mdpi.com Internet Source	<1%
4	fastercapital.com Internet Source	<1%
5	Submitted to University... Student Paper	<1%
6	thescipub.com Internet Source	<1%
7	Submitted to University... Student Paper	<1%
8	bspace.buid.ac.ae Internet Source	<1%
9	ijircce.com Internet Source	<1%
10	m.scirp.org Internet Source	<1%