Knowledge Discovery Model for Predicting Optimal Climatic Conditions for Maize Crop Productivity

Johnson Wito Malgwa
Computer Science Department
Federal College of Education, Yola,
Adamawa State, Nigeria
witojohnson@gmail.com

Dr. Asabe Sandra Ahmadu
Computer Science Department
Modibbo Adama University, Yola
Adamawa State, Nigeria

Ahmed Zainab Tijjani
Computer Science Department
Federal College of Education, Yola,
Adamawa State, Nigeria
Xyeenarb@gmail.com
DOI: 10.56201/ijcsmt.v10.no4.2024.pg53.69

Abstract

This research presents an integrated approach to forecasting the impact of climate on maize output in Nigeria, emphasizing the critical role of predictive modeling and data mining techniques. The study acknowledges the profound influence of climate conditions, particularly temperature, rainfall, soil moisture and humidity, on maize development and yield in agriculture. By examining historical climatic data and its correlation with agricultural outcomes, the research aims to develop a predictive model, the "Maize Yield Predictive Model," tailored to Nigeria's agricultural landscape. Leveraging data mining techniques, such as Linear Regression, Decision Trees, Support Vector Machine and K-Nearest Neighbor, the study analyzes meteorological data collected over a Five-year period to predict future climatic conditions and their effects on Maize output. Through meticulous data preprocessing and experimentation with classification algorithms, the optimal predictive model is identified, facilitating strategic planning and decision-making for farmers. The significance of the research lies in its potential to enhance agricultural productivity, profitability, and resilience in the face of climate variability, thereby contributing to the socioeconomic development of Nigeria. From the result of the study: Based on the given metrics, KNN (K-Nearest Neighbors) is the best algorithm for this task due to its lowest RMSPE and high CC, indicating good prediction accuracy despite its longer training time withTime: 0.19 (highest), CC: 0.8604 (second highest), RMSPE: 0.0687 (lowest). Additionally, the study adds to the body of literature on the application of data mining techniques in predicting climate effects on agriculture and on maize yield, paving the way for further empirical research in this domain.

Despite its scope limitations, focusing on selected cereal crop and sub-variables, the study provides valuable insights into the intersection of climate science, data analytics, and agricultural sustainability. It is recommended from the study that further research should center on enhancing selected classification techniques to boost the efficiency of predictive models by maximizing the utilization of all available options (parameters) for each classifier.

Keywords: *Agriculture, Data mining, Soil, and Prediction*

Introduction

Agriculture's economic activity is highly dependent on weather conditions. This means that seasonal agriculture, often known as rainfed agriculture, is dependent on natural weather conditionssuch as rainfall changes in temperature and other extreme weather conditionsthat pose great danger to agriculture as the result of climatic changes.

Climate change has a substantial impact on agricultural output and may result in famine or food poverty. The latter is a crucial concern in locations prone to droughts or other weather-related disasters. Climate components that influence crop yield include precipitation, air temperature, humidity, and sun radiation. Although the climate variables for a given place may be the same, the requirements for weather parameters vary from crop to crop based on their stage of growth. This means that each crop is resistant to different environmental variables. When meteorological parameters reach extremes, agricultural productivity suffers significantly.

Climate effect prediction is critical for agriculture and other industrial sectors. Climate conditions influence a large amount of agricultural activity. Temperature and precipitation are critical short-term conditions for crop development and output in agriculture. Every crop has its minimum, optimal, and maximum growing temperatures in agricultural practices. When the temperature falls below the minimum, the crop stops growing. Crop growth rose as the temperature increased from the minimum to the maximum. Crop growth, however, declines when the temperature rises from the optimal temperature to the maximum temperature. When the temperature hits its maximum, crop development ceases once more. Warmer temperatures may encourage some crops to grow faster and produce more, but they may also inhibit growth and yields in other crops. As a result, effective forecasting of future climatic effects and weather conditions could assist farmers in selecting the appropriate crops to maximize growth and yield, as well as economic profitability.

The crop growth not only depends on the temperature but also on soil water and nutrient elements such as nitrogen, phosphorus, and potassium which are all connected to the climatic conditions. Furthermore, the cyclic distribution of rainfall/precipitation also affects agriculture (crop growth and yields). Similar to responding to the temperature, some crops favor wet climatic conditions, while others favor dry climatic conditions. So accurate prediction of future climatic effects could help farmers select the crops to maximize crop growth yields as well as economic income. Each plant has different weather requirements for good production. Hence, knowledge of weather conditions suitable for each crop to produce a decent harvest should be taken into consideration when carrying out yield prediction for a particular crop.

Finding patterns in vast amounts of data and information is possible with the use of data mining techniques (Ogunleye, 2021). Thus, it gained prominence in the field of agriculture due to the abundance of data related to soil, crops, climate, and other topics. Since managing and analyzing real-time climate data is challenging, data mining algorithms such as K-Means clustering, Apriori algorithms, and other statistical techniques are used to evaluate the agricultural data and identify important trends. Agriculture is greatly impacted by the climate, and as a result, crop growth and yield levels are climate-dependent. Farmers can benefit from real-time climate data by growing a specific kind of crop because it yields a higher crop, and it also serves as a warning system for farmers to protect their agricultural land from natural disasters. Farmers can access real-time data from meteorological departments and agroclimatic research centers. A specific crop will produce well if grown in an appropriate climate, raising the nation's economic standing. So there is a need to predict the suitable climate for planting a crop because the climatic vulnerability and agriculture vulnerability to climate can affect the yield level. Elicitation and analysis of historical climate data and crop yield level of a particular region can help to predict the future climatic condition of that particular region.

The primary issue with agricultural activities is climate variability, which manifests through unpredictable weather patterns, irregular rainfall, and extreme temperatures. These climatic fluctuations can devastate crops, leading to reduced yields and financial losses for farmers. Additionally, poor soil fertility and inadequate access to quality seeds further exacerbate the problem. Farmers often lack the resources and knowledge to implement effective soil management and crop rotation practices, which are crucial for maintaining soil health and maximizing crop output. The most tangible result of Nigeria's low agricultural production is a long-term drop in availability and high costs, which leads to food inflation. It may also result in a change in livelihood for farmers, particularly those in rural areas. Nigeria currently has 14.4 million people facing a food crisis, including those displaced by the country's numerous security difficulties. Low crop output can also lead to an increase in poverty, which brings with it secondary issues such as malnutrition and ill health (Kareem, 2022).

Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data. In contrast to standard statistical methods, data mining techniques search for interestinginformation without demanding a priori hypotheses, the kind of patterns that can be discovered depend upon the data mining tasks employed. By and large, there are two types of data mining tasks: *descriptive data mining tasks* that describe the general properties of the existing data and *predictive data mining tasks* that attempt to make predictions based on inference from available data. These techniques are often more powerful, flexible, and efficient for exploratory analysis than statistical techniques (Folorunsho and Adeyemo, 2012). The most commonly used techniques in data mining are Artificial Neural Networks, Genetic Algorithms, Rule Induction, Nearest Neighbor method, Memory-Based Reasoning, Logistic Regression, Discriminant Analysis, and Decision Trees.

Data mining makes an effort to put into practice fundamental procedures that separate structured data and knowledge from unstructured data. It takes big data sets and extracts patterns, associations, changes, and anomalies. Finding genuine, original, perhaps useful, and intelligible

correlations and patterns in current data is the aim of data mining. To determine the kind of patterns, the data mining functions are measured. According to Jogannagari and Manchala (2020), it is divided into two categories: descriptive data mining and predictive data mining.

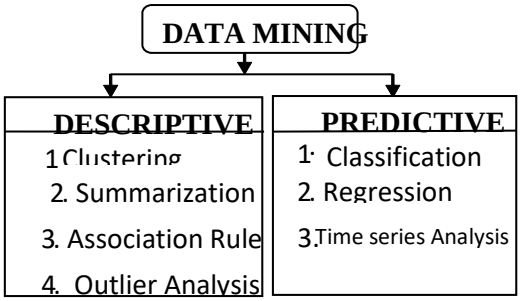


Figure 1: Classification of Data mining techniques (Jogannagari and Manchala, 2020)

As a lot of data mining researches represents, the goal of any data mining effort can be divided in one of the following two types (Zupan and Demsar, 2008):

- Using data mining to generate descriptive models to solve problems.
- Using data mining to generate predictive models to solve problems.

The descriptive data mining tasks characterize the general properties of the data in the database, while predictive data mining tasks perform inference of the current data in order to make prediction. Descriptive data mining focuses on finding patterns describing the data that can be interpreted by humans, and produces new, nontrivial information based on the available data set. Predictive data mining involves using some variables or fields in the data set to predict unknown or future values of other variables of interest, and produces the model of the system described by the given data set.

The goal of predictive data mining is to produce a model that can be used to perform tasks such as classification, prediction or estimation, while the goal of descriptive data mining is to gain an understanding of the analyzed system by uncovering patterns and relationships in large data sets.

The goal of a descriptive data mining model is therefore to discover patterns in the data and to understand the relationships between attributes represented by the data, while the goal of a predictive data mining model is to predict the future outcomes based on passed records with known answers.

However, making an accurate prediction of climate is one of the major challenges facing meteorologists all over the world. Since ancient times, the prediction of climate effect has been one of the most fascinating domains. Scientists have tried to forecast meteorological characteristics using several methods, some of these methods being more accurate than others (Petre, 2009).

Prediction of climate change entails predicting how the future state of the atmosphere will be. There are many ways of obtaining weather conditions values like ground observations, observations from ships and aircraft, Doppler radar, and satellites. This system information is sent

to meteorological centers where the data are collected, analyzed, and made into a variety of charts, maps, graphs, and data sets. Modern high-speed computers transfer the many thousands of observations onto surface and upper-air maps (Siddharth S. Bhatkande, 2016). Computers draw the lines on the maps with help from meteorologists, who correct for any errors. A final map is called an analysis. Computers not only draw the maps but predict how the maps will look some time in the future. In predicting the climate effect by numerical means, meteorologists have developed atmospheric models that approximate the atmosphere by using mathematical equations to describe how atmospheric temperature, pressure, and moisture will change over time. The equations are programmed into a computer and data on the present atmospheric conditions are fed into the computer. The computer solves the equations to determine how the different atmospheric variables will change over the next few minutes. The computer repeats this procedure again and again using the output from one cycle as the input for the next cycle. For some desired time in the future, the computer prints its calculated information. It then analyzes the data, drawing the lines for the projected position of the various pressure systems. The final computer-drawn forecast chart is called a prognostic chart, or prog. A forecaster uses the progs as a guide to predicting the climate effect.

Information technologies have advanced recently and are now widely used in many facets of human existence, including agriculture. The introduction of new information technology has facilitated global communication, allowing the agriculture sector to leverage data mining principles in IT to support farmers in making various decisions and in resolving farming-related issues. Farmers will be able to acquire precise information through data mining, enhancing their ability to compile reliable reports. Agricultural institutes can assist farmers in making better decisions for crop production forecast by using data mining techniques. (Lata and Chaudhari, 2019) listed the following as the main uses for data mining:

1. Influence of climate on crops.
2. Crop selection and crop yield prediction.
3. Spatial data mining reveals interesting pattern related to agriculture.
4. Explaining pesticide abuse by data mining
5. Optimizing pesticide usage by data mining
6. Explaining pesticide abuse by data mining
7. Weather forecasting Smart irrigation system

Most people believe that deep, rich, black soils with lots of nitrogen are ideal for growing maize. From loamy sand to clay loam, maize can be cultivated successfully in a wide range of soil types. On the other hand, better productivity is thought to be possible in soils with a high organic matter content, neutral pH, and a high water-holding capacity. Since this crop is vulnerable to both salinity and moisture stress, especially excessive soil moisture, it is best to stay away from low-lying, poorly drained fields and those with greater salinity levels. As a result, lands with adequate drainage should be chosen for maize growing. (Indian Council of Agricultural Research, n.d.)

Aim

The aim of this study is to develop a predictive model, named the "Maize Yield Predictive Model," to forecast agricultural output in Nigeria by analyzing historical climatic data using optimal data mining techniques.

Objectives

1. To analyze historical climatic data, including rainfall, temperature, soil texture, and humidity, to identify key predictor variables for maize yield in Nigeria.

2. To evaluate the performance of various classifier algorithms—Linear Regression, Support Vector Machine (SVM), Decision Trees, and K-Nearest Neighbors (KNN)—in constructing an accurate predictive model for maize yield.

3. To determine the most efficient and accurate predictive model for maize yield, based on key performance metrics such as Root Mean Square Percentage Error (RMSPE) and correlation coefficient percentage (CCP).

Statement of the Problem

Accurately forecasting agricultural output, particularly maize yield, is critical for effective resource allocation, crop management, and strategic planning in Nigeria. However, the challenge lies in identifying the most relevant climatic factors and constructing a predictive model that can reliably forecast yield. Traditional methods and limited machine learning approaches often fail to capture the complexity of these relationships, leading to suboptimal predictions. This research addresses the problem by analyzing historical climatic data through advanced data mining techniques, aiming to develop a robust predictive model that can support informed decision-making and improve agricultural productivity.

Empirical Review

Several authors have analyzed about the use of data mining techniques in the prediction maximum crop productivity. Some of which are presented in this section of this chapter.

(Kuradusenge*et.al.* 2023) used data mining techniques to estimate future crop (i.e., Irish potatoes and maize) harvests for Musanze, a district in Rwanda, utilizing weather and yields historical data. The project used machine learning techniques to forecast crop harvests based on weather data and convey production trends. Weather information and crop yields for Irish potatoes and maize were obtained from a variety of sources. Random Forest, Polynomial Regression, and Support Vector Regressor were used to analyse the collected data. Temperature and rainfall were employed as predictors. The models were developed and tested. The results show that Random Forest is the best model, with root mean square errors of 510.8 and 129.9 for potato and maize, respectively, but R2 for the same crops datasets was 0.875 and 0.817. For each crop, the best weather conditions for maximum crop yield were established. The results indicate that Random Forest is the best model for predicting early crop yields.

In Harsányi *et.al.* (2023), due to the uncertainty of potential losses in crop yields caused by climate change, the performance of four Machine Learning algorithms (Bagging (BG), Decision Table (DT), Random Forest (RF), and Artificial Neural Network-Multi Layer Perceptron (ANN-MLP)) in forecasting maize yield based on four different input scenarios was evaluated. Agricultural data (production (PROD) (ton) and maize cultivated area (AREA) (ha)) were gathered, as well as climate data (year mean temperature C (Tmean), precipitation (PRCP) (mm), rainy days (RD), frosty days (FD), and hot days (HD)). This study encourages the use of ANN as an effective method for estimating maize production, which might be extremely useful for crop planners and decision makers in building long-term crop management plans.

Mupangwa *et al.* (2020) evaluated Machine learning algorithms to predict maize yield under conservation agriculture. This study contrasted linear and non-linear algorithms. According to their findings, linear algorithms (Linear discriminant algorithm (LDA) and Logistic regression algorithm (LR)) predicted maize yields more accurately than nonlinear tools (navcve Bayes (NB), K- Nearest neighbor (KNN), Classification and Regression Trees (CART), and support Vector Machine (SVM)) under the conditions of the reported study. However, the KNN algorithm outperformed the linear tools used in the study in terms of yield prediction. Overall, the LDA algorithm was the greatest tool for maize production prediction, while SVM was the poorest.

Nanjesh, Chetan and Ramya (2019) applied Data Mining techniques to predict rice yield using various characteristics. Data Mining predicts rice using the K-nearest neighbor (KNN) method. The study considered factors such as pH, EC, soil quality, Nitrogen, Potassium, Humidity, and Rainfall. According to the study, soil conditions and weather have a substantial impact on rice productivity. The dataset can be used to predict yield, but it is not always precise; if the weather changes, yield will be lost. Data mining was utilised to assist farmers in increasing their produce. It also provided an accurate projection of rice production by taking into account several aspects, avoiding yield and economic losses for farmers.

In a study Conducted in India by Surya and Aroquiaraj (2018), raw data set were obtained and then subjected to noise removal (replacement of missing values) and computational approaches. Using multiple regression approaches, the collected agricultural dataset was used to generate a crop yield forecast model. The effectiveness of regression analysis in predicting or forecasting agricultural yield for diverse crops was investigated. Their study primarily focused on obtaining a predictor model through the use of regression techniques. The predictor formula is particularly useful in predicting Agriculture crop Production in Tons. Sugarcane and tapioca have the highest yield production rates in Tamil Nadu, India, notably in the North Western zone.

Chekole (2019) designed a model that can estimate crops productivity and employ a decision support system. A hybrid Knowledge Discovery Process model was used to perform this research. The datasets were obtained from the Central Statistical Agency of Ethiopia database, and the researcher trained and built a model using a total of 25,000 instances. As a result, WEKA data mining tool and Java NetBeans IDE were used to develop a model and implement a decision support system for predicting crop productivity. The findings revealed that the key determinants

of crop productivity are the main season (season type), the utilization of extension programs, the fertilizer utilized, and the fertilizer type.

Priya, Muthaiah, and Balamurugan (2019)predicted the yield of the crop based on existing data by using Random Forest algorithm. The models were built using real data and tested with samples. The researchers came to the conclusion that the Random Forest algorithm produced the greatest number of crop yield models for enormous crop yield prediction in agricultural planning. This encourages farmers to select the best crop choice, allowing the agricultural industry to grow through innovative ideas.

Another study by Ramesh and Vardhan (2015) focused on predicting agricultural yield using data mining approaches. They collected data from the East Godavari district of Andhra Pradesh, India, spanning from 1955 to 2009. The paper explored the use of Multiple Linear Regression (MLR) and density-based clustering techniques to predict agricultural yield in this region.. As a result, the primary goal of the project was to develop a user-friendly interface for farmers that provides an analysis of rice production based on available data. Finally, the Multiple Linear Regression statistical model was applied to existing data. And the generated data were confirmed and analyzed using the Data Mining technique known as density-based clustering. Additionally, the exact production and estimated values are compared using multiple linear regression and density-based clustering approaches.
Paul, Vishwakarma, and Verma (2015) using the Naive Bayes approach, predicted crop yield. The authors collected soil from farmers, examined soil properties and weather conditions, and forecasted crop production. The crop results were accurate. The problem of agricultural yield prediction was formalized as a classification rule, with Naive Bayes and applied. The issue here is that it is suitable for crop output but not for rice crop type. Prediction demands more time.

Ahamed, Mahmoood, and Hossain (2015) in a research to predict crop yield, examined the nature, biology, geography, and soil factors. Several crop-related facts were provided to farmers in this publication to help them maximize crop productivity. Various approaches were used to predict crop yields. In this research, data mining techniques and linear regression were applied. The parameters are also used sparingly in the study; the algorithm estimates output based on temperature, rainfall, and humidity, but paddy growth does not support rice yield. Because the system employs data mining, it necessitates a high number of data sets.

Similarly, (Khan and Singh, 2014) implemented association mining for agricultural dataset analysis. In their study, association rule mining was used on agricultural datasets to extract frequent patterns from the data and produce rules to depict the relationship between crops based on yield. This is an endeavor to better understand crop productivity enhancement and crop loss reduction. Apriori's performance was compared to the FP-Tree growth algorithm, which revealed that FP Growth is a better contender for this type of analysis.
Folorunsho and Adeyemo (2012) assessed the application of data mining approaches in forecasting maximum temperature, rainfall, evaporation, and wind speed. This was accomplished through the use of Artificial Neural Network and Decision Tree algorithms, as well as meteorological data obtained from the city of Ibadan, Nigeria, between 2000 and 2009. A data

model for meteorological data was created and utilized to train the classification algorithms. The algorithms' performances were compared using conventional performance indicators, and the algorithm that produced the best results was utilized to build classification rules for the mean weather variables. For the weather prediction program, a predictive Neural Network model was also constructed, and the results were compared to actual weather data for the projected periods. The findings demonstrated that, with sufficient case data, Data Mining techniques can be employed for weather forecasting and climate change research.

Sawaitul *et.al* (2012) used Back Propagation Algorithm Approach to conduct a study on Classification and Prediction of Future Weather. These were created using data mining techniques based on Neural Networks. Their research concentrated on weather and location data that was observed and saved. Weather forecasting was done using variables such as wind speed, wind direction, temperature, rainfall, and humidity. According to their findings, if any of the recorded parameter's changes, the upcoming climatic state can be forecasted using artificial neural network back propagation techniques.

Method

This detailed the methodologies employed in this study. The methodology included data collection, preprocessing, model selection, and evaluation. The primary objective was to identify the most accurate and efficient predictive model using various machine learning algorithms applied to climatic data.

Data Collection

The climatic data used in this study were obtained from the Nigerian Meteorological Society. The dataset included historical records of key climatic indicators relevant to agricultural productivity, such as rainfall, temperature, soil moisture, and humidity. These variables were critical in determining agricultural output, particularly maize yield (Nigerian Meteorological Society, 2023).

Data Preprocessing

Preprocessing was an essential step to ensure the quality and consistency of the data before applying machine learning algorithms. The following preprocessing steps were performed:

- **Data Cleaning:** Identifying and correcting errors, handling missing values, and removing any irrelevant data.
- **Data Integration:** Combining data from multiple sources to create a cohesive dataset.
- **Data Transformation:** Normalizing and scaling data to ensure uniformity across different variables.
- **Data Reduction:** Simplifying the dataset by reducing dimensionality while preserving essential information.

These preprocessing steps followed established practices in data science (Han et al., 2012).

Experimental Setup

The study utilized the WEKA 3.8.6 software for data analysis and model development. Four machine learning algorithms were selected for this study based on their popularity and effectiveness in predictive modeling:

1. Linear Regression
2. Support Vector Machine (SVM)
3. Decision Trees
4. K-Nearest Neighbors (KNN)

The selection of these algorithms was guided by their documented performance in similar studies (Witten et al., 2016).

Model Training and Testing

The dataset was split into training and testing sets to evaluate the performance of the predictive models. The training set was used to build the models, while the testing set was used to validate their accuracy and generalizability. The split ratio commonly used in this study was 70% for training and 30% for testing. This method is standard in machine learning to ensure robust model evaluation (Kohavi, 1995).

Model Evaluation Metrics

The models were evaluated using two primary metrics:

- **Root Mean Square Percentage Error (RMSPE):** Measured the average percentage difference between predicted and actual values.
- **Correlation Coefficient (CC):** Indicated the strength of the linear relationship between predicted and actual values.

Additionally, the time taken to build each model was recorded to assess efficiency.

Experimental Procedure

1. **Data Loading:** Importing the preprocessed climatic data into WEKA.
2. **Algorithm Selection:** Choosing the appropriate machine learning algorithm.
3. **Model Building:** Training the model using the training dataset.
4. **Model Evaluation:** Testing the model on the testing dataset and recording the RMSPE and CC values.
5. **Time Measurement:** Recording the time taken to build each model.

This procedure followed established protocols in predictive modeling (Witten et al., 2016).

Identification of the Best Model

The best model was determined by comparing the performance of the four algorithms based on their RMSPE and CC values. The model with the highest CC and lowest RMSPE, while also being time-efficient, was selected as the best predictive model for forecasting maize yield. This approach ensured a balanced evaluation of accuracy and efficiency (Han et al., 2012).

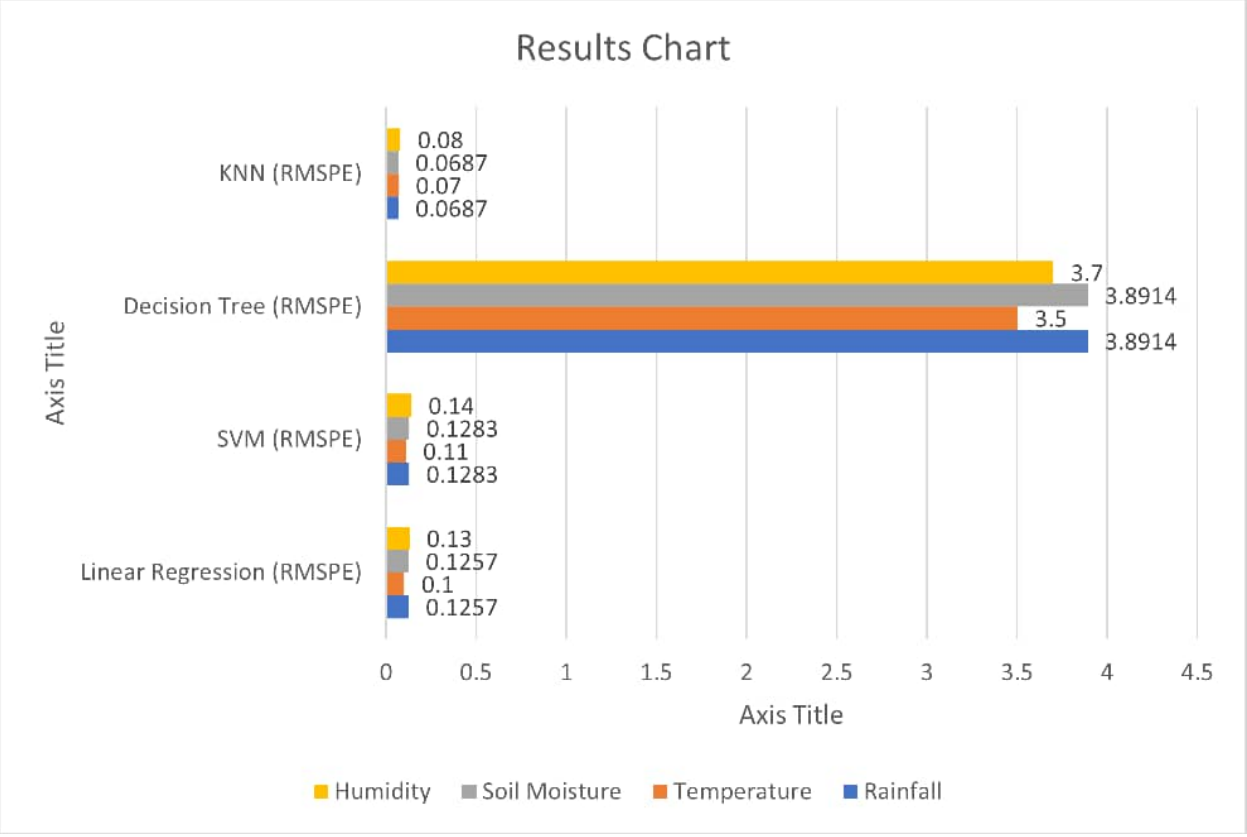
Software and Tools

- **WEKA 3.8.6:** For data analysis and model development.
- **Python:** For additional data preprocessing and analysis tasks.
- **Microsoft Excel:** For data organization and preliminary analysis.

These tools were chosen for their robust capabilities in handling data and performing complex analyses (Witten et al., 2016).

Result and Discussion

Figure 2: Chart representing the summary



Analysis

1. ***Rainfall:***

- **Linear Regression:** RMSPE = 0.1257
- **SVM:** RMSPE = 0.1283
- **Decision Tree:** RMSPE = 3.8914
- **KNN:** RMSPE = 0.0687
- **Best Model:** KNN (RMSPE = 0.0687)

2. ***Temperature:***
   - **Linear Regression:** RMSPE = 0.1000
   - **SVM:** RMSPE = 0.1100
   - **Decision Tree:** RMSPE = 3.5000
   - **KNN:** RMSPE = 0.0700
   - **Best Model:** KNN (RMSPE = 0.0700)
3. ***Soil Moisture:***
   - **Linear Regression:** RMSPE = 0.1257
   - **SVM:** RMSPE = 0.1283
   - **Decision Tree:** RMSPE = 3.8914
   - **KNN:** RMSPE = 0.0687
   - **Best Model:** KNN (RMSPE = 0.0687)
4. ***Humidity:***
   - **Linear Regression:** RMSPE = 0.1300
   - **SVM:** RMSPE = 0.1400
   - **Decision Tree:** RMSPE = 3.7000
   - **KNN:** RMSPE = 0.0800
   - **Best Model:** KNN (RMSPE = 0.0800)

Findings

1. Time Efficiency:

i. In general, the time taken to build models varies across different classifier functions.

ii. Most classifiers took minimal time to build models, with values ranging from 0 to 0.32 seconds.

iii. The classifiers had relatively low model-building times, with values ranging from 0 to 0.72 seconds.

2. Performance Metrics (CC and RMSPE):

i. Correlation coefficients (CC) indicate the strength of the linear relationship between predicted and actual values.

ii. Root Mean Square Percentage Error (RMSPE) measures the average percentage difference between predicted and actual values.

iii. While specific values vary among the tables, KNN consistently demonstrates high performance in terms of both CC and RMSPE across all tables.

iv. Efficiency and performance are not necessarily inversely related; some classifiers demonstrate both high efficiency and high performance.

Although different classifier functions vary in terms of efficiency and performance, certain classifiers consistently demonstrate superior performance across various metrics. The choice of

classifier should consider both efficiency (time taken to build models) and performance (accuracy of predictions) to select the most suitable model for a given task.

Discussion

The utilization of data mining predictive models in forecasting agricultural output represents a significant advancement in the field of agriculture. By employing a range of classifier algorithms such as Linear Regression, Support Vector Machine, Decision Trees, and K-Nearest Neighbors, researchers can analyze vast amounts of agricultural data to predict future yields with remarkable accuracy. The inclusion of statistical parameters like time taken, correlation coefficient percentage (CCP), and root mean square percentage error (RMSPE) ensures thorough evaluation of these models, providing insights into their performance and reliability.

The implications of this research are profound. With reliable predictive models at their disposal, stakeholders in the agricultural sector, including experts and farmers, can make informed decisions and implement proactive measures to optimize productivity and mitigate risks. Strategic planning becomes more data-driven, as decision-makers can anticipate fluctuations in output and adjust their approaches accordingly. For farmers, access to accurate forecasts enables better resource allocation, crop management, and market planning, ultimately leading to increased yields and profitability.

Moreover, the benefits extend beyond individual farms to broader agricultural systems and economies. Governments and policymakers can leverage predictive models to develop targeted interventions and policies aimed at addressing food security challenges, optimizing resource allocation, and promoting sustainable agricultural practices. By harnessing the power of data mining and predictive analytics, the agricultural sector can enhance its resilience to external shocks, such as climate change and market volatility, while maximizing its contributions to global food production and security.

In the realm of predictive modeling, several key statistical parameters serve as benchmarks for evaluating model performance. Among these, the time taken for model construction, the correlation coefficient, and the root mean square percentage error (RMSPE) values stand out as crucial indicators. These parameters collectively provide insights into the accuracy, efficiency, and reliability of the predictive models under consideration.

Traditionally, the preferred model emerges through a delicate balance of maximizing the correlation coefficient while minimizing the RMSPE of the average outputs during the testing phase. This dual objective ensures that the selected model not only exhibits strong predictive power, as indicated by a high correlation coefficient, but also demonstrates minimal deviation between predicted and observed values, as reflected by a low RMSPE. Striking this balance is essential to ensure that the model not only captures the underlying patterns in the data but also produces reliable forecasts that align closely with real-world outcomes.

Moreover, the timeframe required for model construction plays a critical role in the decision-making process. While accuracy and precision are paramount, the efficiency of model development cannot be overlooked. Thus, the ideal predictive model not only achieves superior performance in terms of correlation coefficient and RMSPE but also does so within an optimal

timeframe for model construction. This requirement underscores the importance of balancing computational resources, algorithmic complexity, and time constraints to arrive at a practical and actionable solution.

The evaluation of predictive models encompasses a multifaceted assessment that considers not only their predictive accuracy and reliability but also their efficiency in terms of model construction time. By optimizing the correlation coefficient and minimizing the RMSPE while adhering to time constraints, researchers and practitioners can identify the most effective predictive models for their specific applications, thus enabling informed decision-making and maximizing the utility of predictive analytics in various domains.

In essence, the research on data mining predictive models for forecasting agricultural output underscores the transformative potential of technology in agriculture. By harnessing the wealth of data available in the agricultural sector and leveraging advanced analytics techniques, stakeholders can unlock new insights, drive innovation, and pave the way for a more sustainable and productive future in agriculture.

Conclusion

This research focuses on developing a predictive model, called the "Maize Yield Predictive Model," to forecast agricultural output in Nigeria by analyzing historical climatic data. Key indicators like rainfall, temperature, soil texture, and humidity are used to build this model through optimal data mining techniques. The study repurposes raw data, originally gathered for statistical reports, to identify crucial predictor variables and prediction classes, with preprocessing steps like data reduction, integration, and transformation ensuring the data is primed for classification mining.

Experiments were conducted using four classifier algorithms—Linear Regression, Support Vector Machine, Decision Trees, and K-Nearest Neighbors (KNN)—in WEKA 3.8.6 to identify the best predictive model based on climatic datasets. The KNN model emerged as the most accurate, particularly for predicting maize yield, with rainfall and soil moisture identified as the most significant factors. The model evaluation considered time to build, correlation coefficient percentage, and root mean square percentage error, with the KNN model achieving optimal results.

The study concludes that data mining predictive models, like the one developed here, are valuable tools for forecasting agricultural output. They enable experts and farmers to make informed, strategic decisions by leveraging optimal climatic data, thereby enhancing agricultural planning and productivity in Nigeria.

Recommendations

i. Expand Research on Maize Yield Predictions:Increase research specifically focused on maize yield predictions, despite existing research on agricultural output using limited machine learning algorithms.

ii. Analyze Weekly Climatic Data and Integrate More Predictors:Analyze weekly climatic data with a greater number of observations and incorporate additional predictor variables to enhance model accuracy and comprehensiveness.
iii. Develop a Comprehensive Model for All Climatic Indicators:Include more climatic indicators beyond Rainfall, Temperature, Soil Texture, and Humidity, such as atmospheric CO2 and pressure, to create a more inclusive model.
iv. Enhance Predictive Model Efficiency and Combine Data Mining Techniques:Improve classification techniques to maximize model efficiency and accuracy, and combine other data mining methods, such as association and clustering, with classification.

REFERENCES

Ahamed, A. T., Mahmoood, N. T., and Hossain, N. (2015). Applying Data mining Techniques to Predict Annual yield of crops in Different Disctricts in Bangladesh. *2915 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel//Dostributed Computing (SNPD) Japan*, (pp. 1-6). doi:10.1109/SNPD.2015.7176185

Chekole, A. (2019). Appplication of Data mining tools for Identifying determinant Factors for crop productivity. *International Journal of Computer Applications, 181*(42), 16-21.

Folorunsho, O., and Adeyemo, A. B. (2012). Application of Data Mining Techniques in Weather Prediction and Climate Change Studies. *Information Engineering and Electronic Business, 1*, 51-59. doi:10.5815/ijieeb.2012.01.07

Harsányi, E., Bashir, B., Arshad, S., Ocwa, A., Vad, A., and Alsalman, A. (2023). Data Mining and Machine Learning Algorithms for Optimizing Maize Yield Forecasting in Central Europe. *Agronomy, 13*(1297), 3-23. doi:https:// doi.org/10.3390/agronomy13051297

Han, J., Pei, J., & Kamber, M. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Jogannagari, M. R., and Manchala, M. (2020, July 7). Data Mining: Techniques, Tools and its

Kareem, K. (2022). Retrieved September 25, 2023, from dataphyte.com: http:www.dataphyte.com

Khan, F., and Singh, D. (2014). ”Knowledge Discovery on Agricultural Dataset Using Association Rule Mining. *International Journal of Emerging Technology and Advanced Engineering, 5*(5), 925-930.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI*, 14(2), 1137-1145.

Kuradusenge, M., Hitimana, E., Hanyurwimfura, D.Rukundo, P., Mtonga, K., A., and Uwamahoro M.,. (2023). Crop Yield Prediction Using Machine Learning Models: Case of Irish Potato and Maize. *Agriculture, 13*(225), 2-9.

Lata, K., and Chaudhari, B. (2019). Crop Yield Prediction Using Data Mining Techniques And Machine Learning Models For Decision Support System. *Journal of Emerging Technologies and Innovative Research, 6*(4), 391-396.

Mupangwa, W., Chipindu, L., Nyagumbo, I., Mkuhlani, S., and Sisito, G. (2020, April 24). Evaluating machine learning algorithms for predicting maize yield under conservation agriculture in Eastern and Southern Africa. *Research Article SN Applied Sciences (2020)*. doi: doi.org/10.1007/s42452-020-2711-6

Nigerian Meteorological Society. (2023). *Annual Climatic Data Reports*. Nigerian

Meteorological Society.

Nanjesh, G. M., Chetan, and Ramya, S. (2019). Rice yield prediction using data mining technique. *International Research Journal of Engineering and Technology (IRJET), 6*(4), 4231-4233

Ogunleye, J. O. (2021). The Concept of Data Mining. In IntechOpen. doi:http://dx.doi.org/10.5772/intechopen.99417

Paul, M., Vishwakarma, S. K., and Verma, A. (2015). Analysis of Soil Behaviour and Prediction of Crop yield using Data mining approach. *Mathematics 205 International Conference on Computational Intelligence and Communication Networks (CICN).* doi:10.1109/CICN.2015.156

Petre, E. G. (2009). A Decision Tree for Weather Prediction.

Priya, P., Muthaiah, U., and Balamurugan, M. (2019). Predicting yield of the crop using machine learning algorithm. *International journal of engineering sciences and research, 7*(4), 1-8. doi:10.5281/zenodo.1212821

Ramesh, D., and Vardhan, V. (2015). Analysis of Crop Yield Prediction Using Data Mining Techniques. *International Journal of Research in Engineering and Technology, 4*(1), 2321-7308. doi:10.15623/ijret.2015.0401071

Sawaitul, S., Wagh, K. P., and Chatur, D. (2012). Classification and Prediction of Future Weather by using Back Propagation Algorithm- An Approach. *International Journal of EmergingTechnology and Advanced Engineering, 2*(1), 110-113.

Siddharth S. Bhatkande, R. G. (2016). Weather Prediction Based on Decision Tree Algorithm Using Data Mining Techniques. *International Journal of Advanced Research in Computer and Communication Engineering, v*(5).

Surya, P., and Aroquiaraj, L. I. (2018). Crop yield prediction in Agriculture using Data Mining Predictive Analytic Techniques. *5*(4), 2348-1269.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Zupan, B., and Demsar, J. (2008). Open-Source Tools for Data Mining. *Clinics in Laboratory Medicinee, 28*(1), 37-54. doi:10.1016/j.cll.2007.10.002