

Prifysgol Cymru

Y Drindod Dewi Sant

University of Wales

Trinity Saint David

Ragi Crop Yield Prediction Using Data Mining Techniques

NAME: SURAJ RAMESH BABU

ID NO: 2513205

LECTURER: Prof. Nik Whitehead

MODULE CODE: ACCA7015

MODULE NAME: Data Mining & Visualisation

DATE: 15/05/2026

Table of Contents

1) Introduction	4
1.1 Background	4
1.2 Problem Statement	5
1.3 Aim	5
1.4 Objectives	5
1.5 Research Questions	5
2. Literature Review	5
2.1 Data Mining in Agriculture	5
2.2 Machine Learning in Crop Prediction	6
2.3 Clustering in Agriculture	6
2.4 Association Rule Mining	6
2.5 Research Gap	7
3. Dataset Description	7
3.1 Dataset Overview	7
3.2 Attribute Description Table	8
3.3 Dataset Characteristics	8
4. Data Understanding and Visualization	8
4.1 Attribute Distributions	9
4.2 Correlation Observations	9
4.3 Data Quality Issues	9
4.4 Dataset Source	9
5. Data Preprocessing	10
5.1 Handling Missing Values	10
5.2 Attribute Selection	11
5.3 Normalization	12

5.4 Discretization.....	13
5.5 Data Formatting.....	14
6. Methodology	15
6.1 CRISP-DM Framework.....	16
6.2 Algorithms Used.....	17
7. Results and Discussion	17
7.1 Random Forest Results.....	17
7.2 Linear Regression Results	18
7.3 Random Tree Results	19
7.4 Model Comparison	20
7.5 Clustering Results.....	21
7.6 Apriori Association Rule Results	23
8. Discussion.....	23
8.1 Interpretation of Findings	24
8.2 Strengths of the Project	24
8.3 Limitations of the Project	24
9. Conclusion.....	25
10. Recommendations and Future Work	25
11. References	26

Abstract

This work used an agricultural dataset with 18,000 records and 14 characteristics to forecast Ragi crop productivity using data mining and machine learning approaches. Preprocessing, including handling missing values, normalization, discretization, and attribute selection, was done using WEKA software. 10-fold cross-validation was used to implement and assess algorithms including Random Forest, Linear Regression, Random Tree, K-Means Clustering, and Apriori. With a correlation coefficient of 0.8883, Random Forest performed the best. While Apriori found connections between rainfall, temperature, season, and productivity, K-Means clustering found similar farming conditions. The results demonstrate how data mining can enhance agricultural forecasting and decision-making.

1.1 Background

The production of food and the economies of many developing nations depend heavily on agriculture. Increasing agricultural production is essential to addressing environmental issues and the rising need for food. However, crop yield is impacted by variables like climate change, erratic rainfall, temperature fluctuations, problems with the soil, and inadequate resource management. Conventional farming techniques frequently rely on manual estimation, which can result in erroneous forecasts and monetary losses. In order to anticipate crop yield, this project applies data mining techniques using WEKA software. Agricultural data was analyzed using machine learning, clustering, and association rule mining to find patterns that could help farmers make better decisions and increase productivity.

1.2 Problem Statement

Due to their inability to effectively assess extensive agricultural and environmental data, traditional agricultural forecast systems are frequently wrong. Farmers typically rely on manual estimation and historical data, which might not yield accurate forecasts in situations where the climate is changing. To improve agricultural decision-making, sophisticated data mining and machine learning methods that can handle massive datasets, spot hidden patterns, and produce precise crop output forecasts are required.

1.3 Aim

The aim of this project is to apply data mining and machine learning techniques to predict crop yield and identify hidden agricultural patterns using a large agricultural dataset.

1.4 Objectives

This project's primary goals are:

- To use WEKA technologies for preprocessing and cleaning the agriculture dataset
- Using machine learning algorithms to forecast crop yields
- To evaluate how well various forecasting models perform
- To find comparable agricultural circumstances using clustering analysis
- To create principles of correlation between agricultural characteristics
- To assess the correctness and effectiveness of the models that have been implemented

1.5 Research Questions

The following research questions are the focus of this project:

1. What is the best accurate machine learning algorithm for predicting agricultural yield?
2. What effects do temperature and rainfall have on agricultural productivity?
3. Can commonalities between agricultural circumstances be found using clustering techniques?
4. What connections exist between crop yield and farming attributes?
5. To what extent may data mining approaches enhance agricultural decision-making?

2. Literature Review

2.1 Data Mining in Agriculture

Food production and economic expansion depend on agriculture. Artificial intelligence, machine learning, and data mining are examples of technology that have been widely applied in recent years to increase farming production and efficiency. In order to facilitate improved decision-making, data mining assists in the analysis of agricultural data, such as rainfall, temperature, soil quality, and crop productivity [1]. Modern technologies are used in precision agriculture to minimize waste and maximize the usage of herbicides, fertilizers, and water [2]. IoT devices, sensors, and meteorological data are being used in smart farming to increase productivity [3]. Predictive analytics and machine learning enhance efficient agricultural planning and resource management by forecasting crop output and environmental changes [4][5].

2.2 Machine Learning in Crop Prediction

Crop forecast and agricultural analysis heavily rely on machine learning techniques. These algorithms can forecast future crop production, disease prevalence, and soil conditions by learning from past data. For agricultural prediction problems, numerous researchers have used supervised learning algorithms such Random Forest, Decision Trees, Support Vector Machines, and Neural Networks [6].

2.3 Clustering in Agriculture

An unsupervised learning method called clustering is used to organize comparable data instances according to shared attributes. Clustering does not require labeled data, in contrast to classification techniques. Clustering is used in agriculture to find areas or crops that have similar farming practices, production patterns, or environmental circumstances [7].

K-Means clustering is one of the most popular clustering methods. K-Means uses similarity metrics to separate the dataset into several groups. Data points are assigned to the closest cluster centroid by the algorithm, which then iteratively updates the centroid values until convergence is reached.

K-Means clustering has been applied in agriculture for crop grouping, soil categorization, rainfall pattern analysis, and identifying high-yield farming areas [8].

One data mining method for finding correlations between variables in big datasets is association rule mining. It uses metrics like support, confidence, and lift to find trends and relationships between various characteristics. The Apriori algorithm is one of the most widely used algorithms for association rule mining.

2.4 Research Gap

Numerous research have used data mining and machine learning methods in agriculture to estimate crop yields and analyze the environment. Prediction models like Random Forest, Decision Trees, and Neural Networks are the main focus of most research. Additional research examines association rule mining or clustering algorithms independently.

3. Dataset Description

3.1 Dataset Overview

This research uses an agricultural dataset with over 18,000 instances and 14 attributes pertaining to crop production and farming circumstances. Numerical and categorical variables representing various agricultural, seasonal, and environmental aspects influencing crop yield are included in the dataset. The dataset's primary goal is to facilitate data mining techniques for agricultural pattern analysis and crop yield prediction.

3.2 Attribute Description Table

Attribute	Description
Year	Year of cultivation
District	Location of farming
Season	Crop season (Kharif, Rabi, Summer)
Rainfall_mm	Annual rainfall (mm)
Temperature_C	Average temperature (°C)
Soil_Type	Type of soil
Area_hectares	Cultivated area
Fertilizer_kg_per_ha	Fertilizer usage
Irrigation_Type	Irrigation method
Seed_Variety	Type of seed
Pesticide_kg_per_ha	Pesticide usage
Production_tons	Total production
Yield_ton_per_hectare	Crop yield

3.3 Dataset Characteristics

The dataset is appropriate for data mining and machine learning analysis due to a number of significant features. First, the size of the dataset roughly 18,000 agricultural records improves the predictive models' accuracy and dependability. Big datasets make it easier for machine learning algorithms to find trends and lessen the chance of biased outcomes.

Second, both numerical and categorical attributes are present in the dataset. While categorical variables define farming parameters including soil type, season, irrigation technique, and seed variety, numerical attributes offer quantifiable environmental and productivity information. Multi-dimensional agricultural analysis is supported by the combination of various characteristics

4. Data Understanding and Visualization

Before using machine learning models, data comprehension is a crucial step in the CRISP-DM process since it aids in determining the dataset's structure, quality, and attributes. The WEKA Explorer visualization and preprocessing tools were used in this study to investigate the dataset. To comprehend the distribution of variables and the connections between agricultural factors influencing crop production, various visualizations and attribute summaries were examined.

4.1 Attribute Distributions

The distribution of the dataset's numerical and categorical variables was examined using WEKA's "All Attributes" visualization. The majority of records had uniformly distributed yield values appropriate for machine learning analysis, balanced temperature values around 27°C, and moderate rainfall.

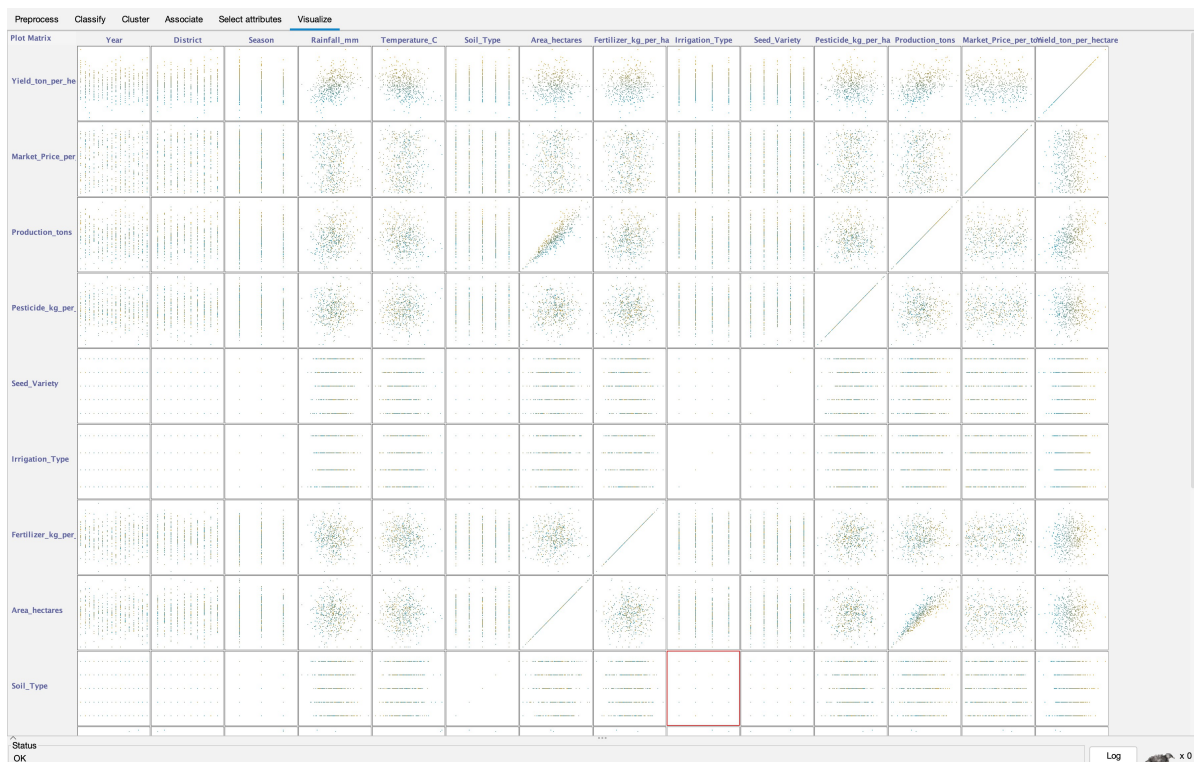


Figure 4.1: Distribution of agricultural attributes in WEKA Explorer.

4.2 Correlation Observations

During data visualization and analysis, a number of correlations between attributes were found. The strong correlation between crop yield and rainfall was one obvious tendency. Crop yield values were generally greater in agricultural records with moderate to high rainfall values. This implies that crop productivity benefits from enough rainfall. The connection between production and farmed area was another significant finding. Higher output levels were typically seen in records with bigger farmed fields. This suggests that the total crop output is significantly influenced by land area.

4.3 Data Quality Issues

During dataset analysis, a number of difficulties with data quality were found, such as missing values, inconsistent attribute scales, and errors with categorical formatting. The Replace Missing data filter in WEKA was used to manage missing data, and normalization was used to normalize numerical variables like production and rainfall into a uniform range. Discretization was employed for Apriori analysis, and categorical attributes were appropriately structured. These preprocessing procedures increased model dependability, decreased training mistakes, and improved dataset consistency.

4.4 Dataset Source

The dataset was collected from Kaggle and contains agricultural production records from multiple districts and cultivation seasons.

<https://www.kaggle.com>

5. Data Preprocessing

Because the accuracy and dependability of machine learning models are directly impacted by the quality of the dataset, data preparation is a crucial step in the data mining process. Missing values, uneven formatting, and attributes with varying value ranges are common in real-world agricultural datasets. In order to clean and prepare the dataset for classification, clustering, and association rule mining, preprocessing techniques were used in WEKA.

At first, the dataset was kept in Microsoft Excel (.xlsx) format. The Excel dataset was first converted to CSV format before being loaded into WEKA Explorer because WEKA primarily supports CSV and ARFF formats. The dataset was saved in ARFF (Attribute Relation File Format), the common format used by WEKA for machine learning analysis, after it had been loaded. Both data and attribute information, including nominal and numerical data types, are stored in the ARFF format.

5.1 Handling Missing Values

A number of missing values were found in the dataset during the preliminary analysis. Because many algorithms are unable to properly interpret incomplete records, missing data can have a detrimental effect on machine learning performance. The Replace Missing Values filter in WEKA was used to address this issue.

The filter used was:

Filter → Unsupervised → Attribute → Replace Missing Values

The mean value of the corresponding variable was used to substitute missing data for numerical attributes like rainfall, temperature, fertilizer usage, and productivity. The mode, which stands for the most common category value, was used to replace missing values for categorical parameters including soil type, irrigation type, and seed variety. This preprocessing stage decreased mistakes during model training and preserved dataset consistency. In addition to preventing needless data loss from deleting incomplete records, replacing missing values increased forecast reliability.

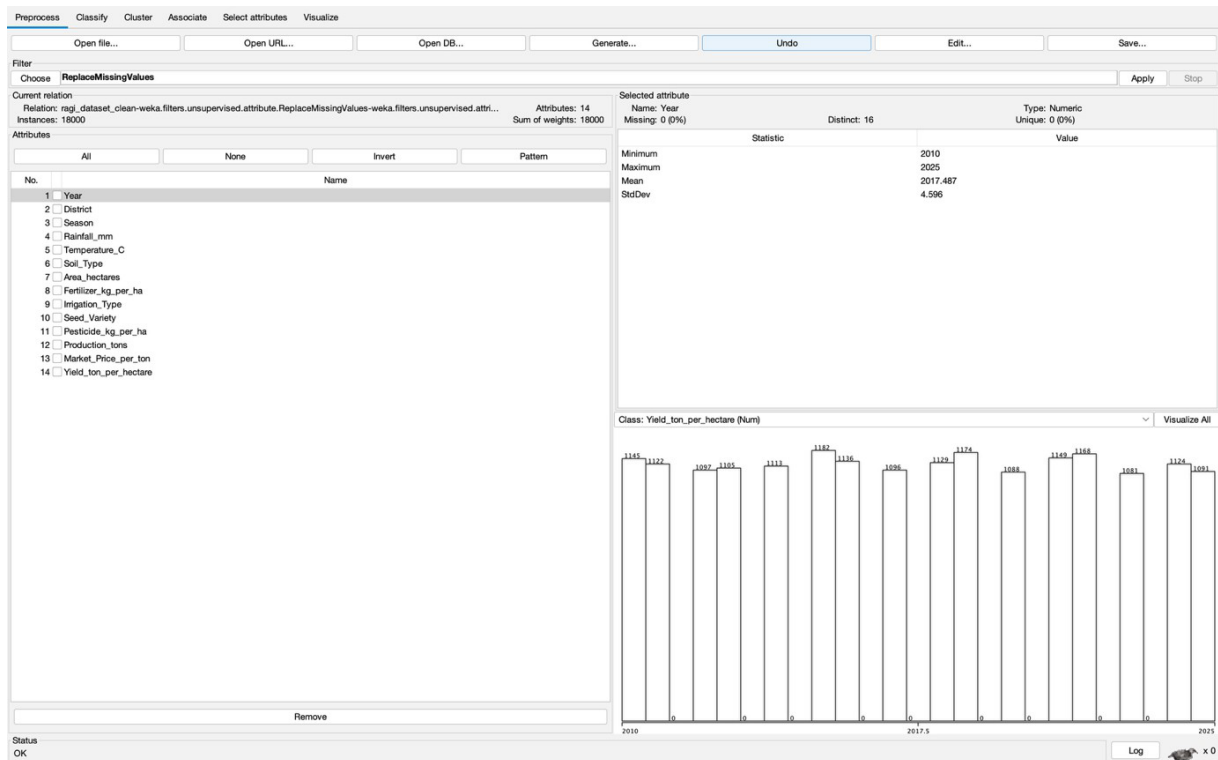


Figure 5.1: Handling missing values using the ReplaceMissingValues filter in WEKA Explorer.

5.2 Attribute Selection

Attribute selection was performed to identify the most relevant features affecting crop yield prediction. Feature selection is important because irrelevant or redundant attributes may reduce model performance and increase computational complexity.

The main attributes selected for analysis included:

- Rainfall_mm
- Temperature_C
- Soil_Type
- Fertilizer_kg_per_ha
- Seed_Variety
- Pesticide_kg_per_ha
- Production_tons

Select Attributes Tab-Evaluator → CfsSubsetEval
Search Method → BestFirst

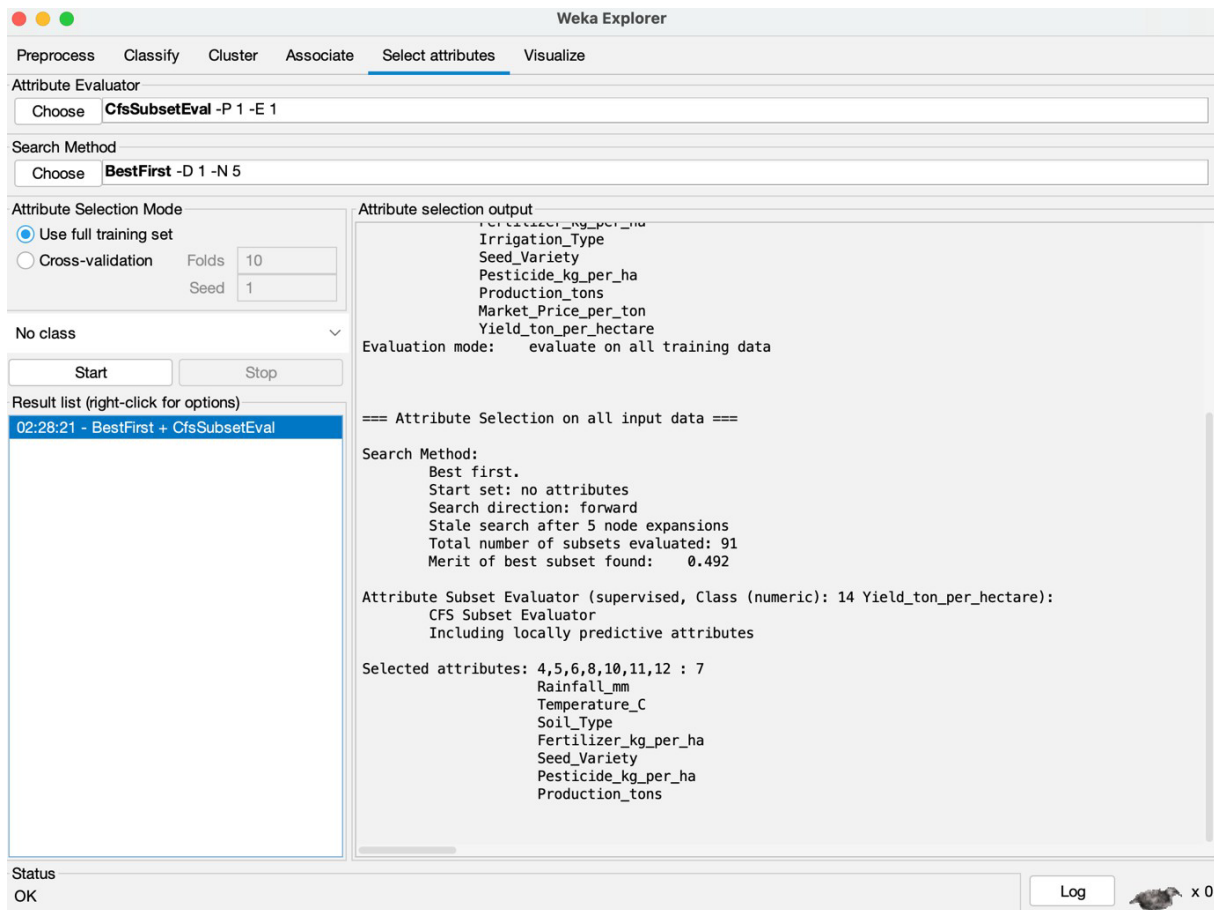


Figure 5.2 here – Select Attributes Screenshot

Certain characteristics were eliminated because they added unnecessary complexity or made a smaller contribution to prediction accuracy. By lowering dataset noise, choosing just significant features enhances machine learning performance.

Feature selection provides several advantages:

- Improves prediction accuracy
- Reduces training time
- Simplifies model interpretation
- Decreases computational cost
- Helps reduce overfitting

Overfitting occurs when a model memorizes training data instead of learning general patterns. By removing unnecessary attributes, the model becomes more efficient and performs better on unseen data.

5.3 Normalization

Numerical attributes were scaled into a common range using normalization. Rainfall, farmed area, fertilizer use, and productivity all exhibited substantially distinct value ranges in the dataset. In distance-based analysis, large-scale characteristics may predominate over smaller-scale characteristics.

Because K-Means clustering computes the distances between data points and cluster centroids, normalization is very crucial. Larger numerical values may have a greater impact on the clustering process than smaller ones in the absence of normalization.

To guarantee fair scaling across numerical characteristics, WEKA's Normalize filter was used. This enhanced clustering analysis's stability and fairness and made it easier for the algorithm to find significant agricultural patterns. Normalization

increases model performance, but because the data is converted into scaled ranges, it may make the original numerical values less interpretable. Filter → Unsupervised → Attribute → Normalize

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter: Choose **Normalize -S 1.0 -T 0.0** [Apply] [Stop]

Current relation: Relation: ragi_dataset_clean-weka.filters.unsupervis...
Instances: 18000 | Attributes: 14 | Sum of weights: 18000

Attributes: All | None | Invert | Pattern

No.	Name
1	☒ Year
2	☐ District
3	☐ Season
4	☐ Rainfall_mm
5	☐ Temperature_C
6	☐ Soil_Type
7	☐ Area_hectares
8	☐ Fertilizer_kg_per_ha
9	☐ Irrigation_Type
10	☐ Seed_Variety
11	☐ Pesticide_kg_per_ha
12	☐ Production_tons
13	☐ Market_Price_per_ton
14	☐ Yield_ton_per_hectare

[Remove]

Selected attribute: Name: Year
Missing: 0 (0%) | Distinct: 16 | Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.499
StdDev	0.306

Class: Yield_ton_per_hectare (Num) [Visualize All]

Status: OK | Log | x 0

5.3 Normalization of numerical attributes in WEKA.

5.4 Discretization

The process of transforming numerical quantities into nominal or categorical values is known as discretization. This preprocessing stage was primarily necessary for the Apriori algorithm's association rule mining.

Because it creates association rules between nominal values and finds frequent itemsets, the Apriori method performs better with categorical data. As a result, WEKA preprocessing filters were used to group continuous numerical attributes including temperature, production, and rainfall.

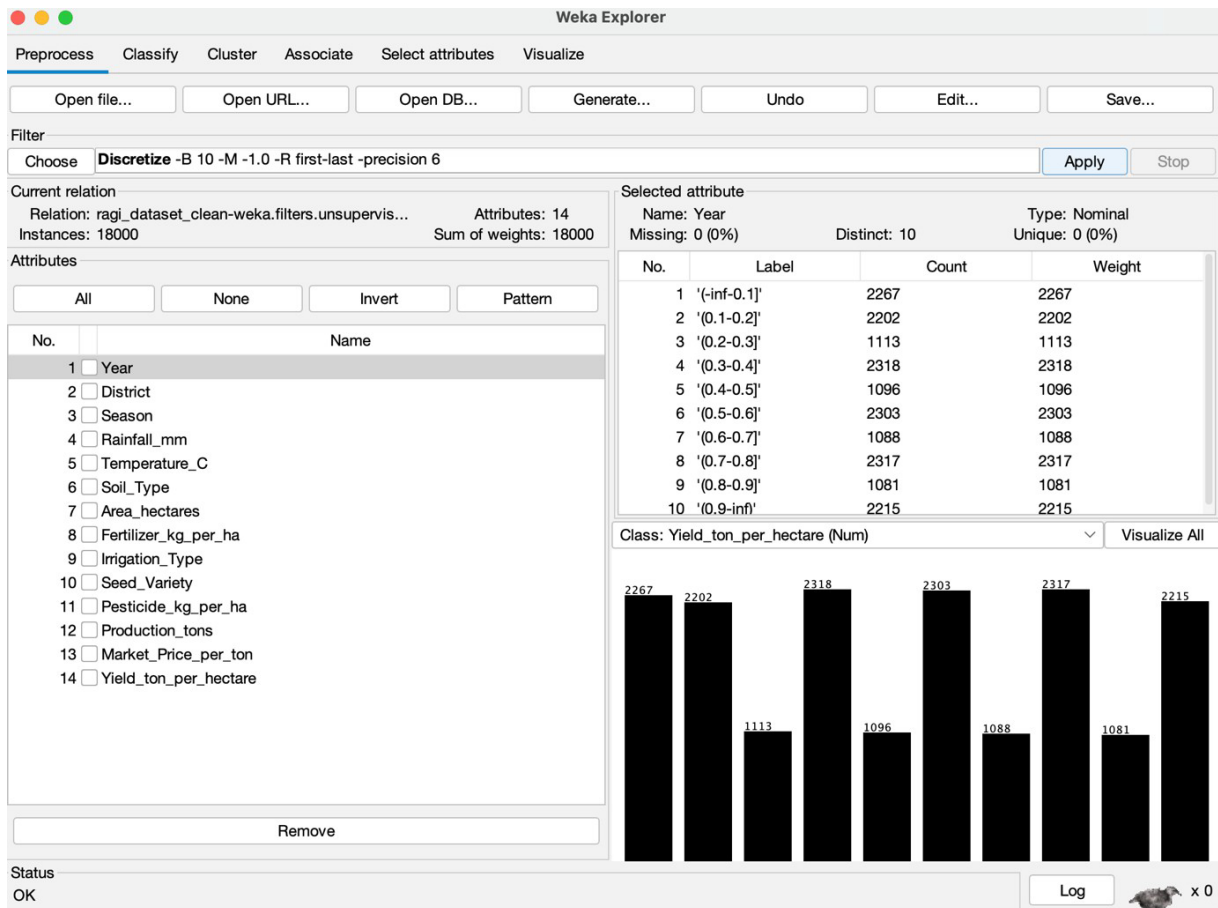


Figure 5.4 Discretization of numerical attributes for Apriori association rule mining.

5.5 Data Formatting

The dataset contained both numerical and categorical attributes. Proper formatting was necessary to ensure compatibility with WEKA algorithms and preprocessing tools.

Categorical attributes such as:

- Soil_Type
- Irrigation_Type
- Seed_Variety
- Season

Were kept in nominal format since category representation is necessary for many association rule and machine learning techniques. The effectiveness of classification, clustering, and association rule creation was enhanced by proper formatting.

6. Methodology

6.1 CRISP-DM Framework

1. Business Understanding

Understanding the agricultural issue and establishing the project's goals were the main goals of the first phase. Due to shifting environmental factors including temperature, rainfall, and soil quality, farmers frequently struggle to forecast crop yields. This project's primary goal was to use data mining techniques to forecast crop productivity and find hidden agricultural patterns that can help farmers make better decisions.

2. Data Understanding

During this stage, WEKA Explorer tools were used to examine the dataset. To comprehend the dataset's structure and quality, attribute summaries, histograms, and visualizations were examined. Important agricultural parameters were identified by observing relationships between variables like crop productivity, fertilizer use, and rainfall.

3. Data Preparation

A number of preprocessing methods were used to enhance the performance of the model and the quality of the data. To make the dataset compatible with WEKA, it was first saved in Excel format and then converted to ARFF format. The Replace Missing Values filter was used to deal with missing values. To find significant characteristics influencing crop output, feature selection was carried out. While discretization was used to transform numerical attributes into categorical ranges for association rule mining, normalization was used to scale numerical values between 0 and 1.

4. Modelling

Different machine learning and data mining algorithms were applied to the prepared dataset. Classification algorithms such as Random Forest, Random Tree, and Linear Regression were used for crop prediction. K-Means clustering was applied to group similar agricultural records, and the Apriori algorithm was used to discover association rules among farming attributes.

5. Evaluation

Metrics like accuracy, precision, recall, F-measure, and Root Mean Square Error (RMSE) were used to assess the models' performance. To guarantee accurate results, cross-validation methods were also applied. To determine which prediction model was the most accurate, the performance of each algorithm was compared.

6. Deployment

The final stage focused on interpreting the results and identifying how the developed models could support agricultural decision-making. The findings of the project can help farmers and agricultural organizations improve crop management and production planning.

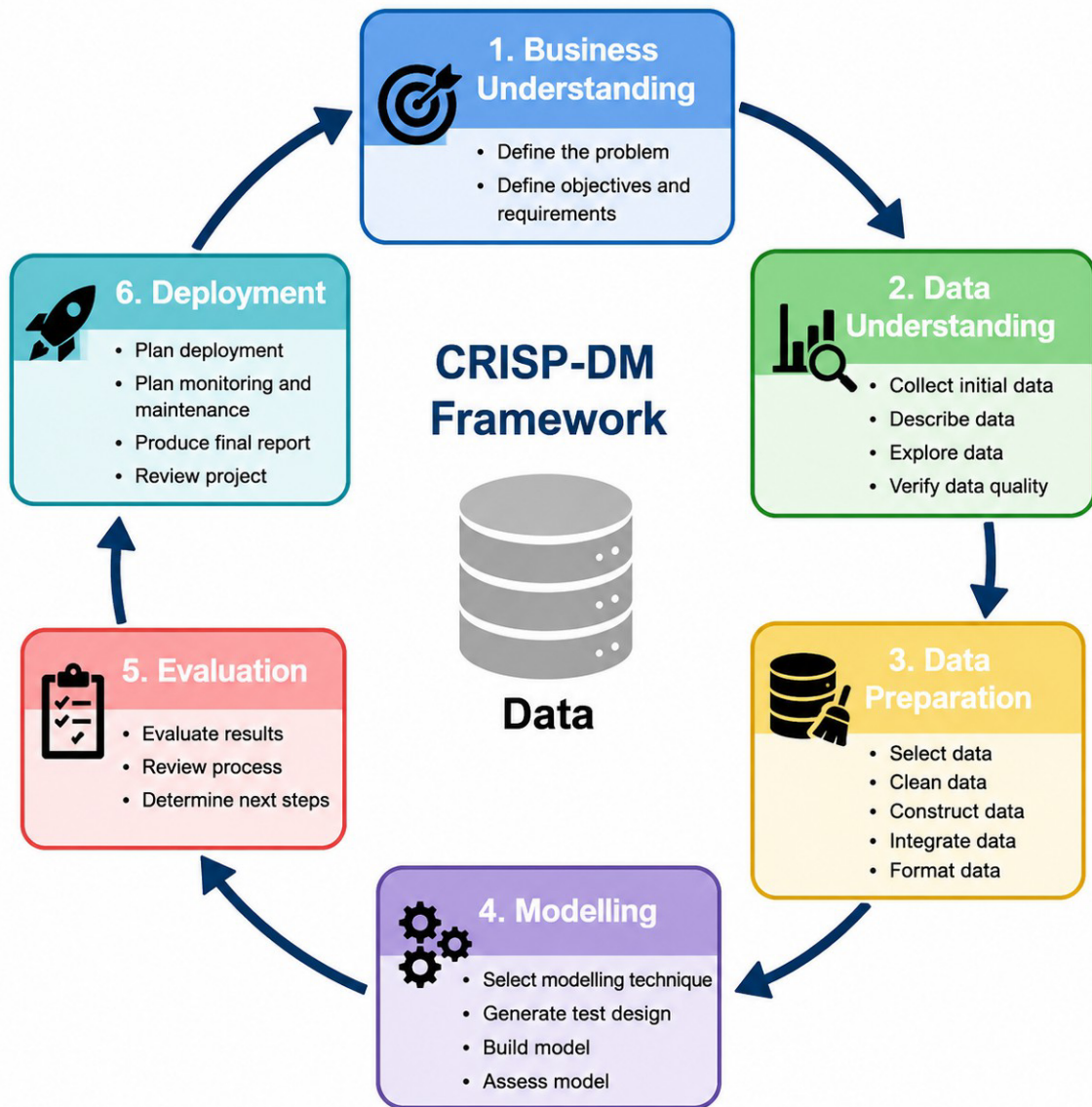


Figure 6.1: CRISP-DM lifecycle followed in the data mining project.

6.2 Algorithms Used

1) Random Forest

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy. Each tree is trained using different subsets of the dataset, and the final prediction is generated through voting or averaging. Random Forest is highly effective for large datasets because it reduces overfitting and improves generalization performance. In this project, Random Forest achieved the highest prediction accuracy among all classification algorithms.

2) Linear Regression

Linear Regression is a statistical algorithm used to identify the relationship between independent variables and the target variable. It predicts output values using a linear equation based on the input features. Linear Regression was used as a baseline predictive model to compare the performance of advanced machine learning algorithms.

3) Random Tree

A decision tree-based approach called Random Tree uses randomly chosen features to create categorization trees. Data is divided into branches according to attribute conditions in a tree-structured decision-making process. Although the algorithm's prediction accuracy is typically lower than that of ensemble techniques like Random Forest, it is computationally quick and simple to understand.

4) K-Means Clustering

An unsupervised machine learning approach called K-Means is used to cluster comparable records. The approach locates cluster centroids and uses distance computations to allocate data points to the closest centroid. To find various agricultural patterns and commonalities among farming records, the value of K was set to 4 in this study.

5) Apriori Algorithm

An association rule mining method called the Apriori algorithm is used to find recurring itemsets and connections between characteristics in a dataset. The algorithm uses metrics like lift, confidence, and support to find patterns. Apriori was used in this effort to find connections between agricultural production trends, fertilizer use, and environmental factors.

7. Results and Discussion

7.1 Random Forest Results

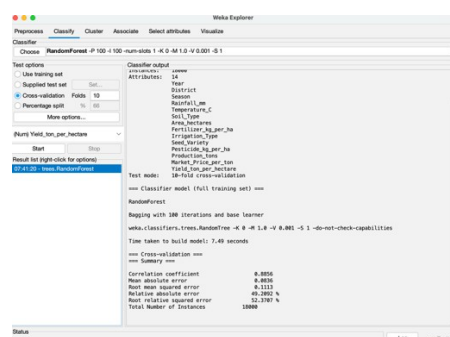
Out of all the predictive models utilized in this research, the Random Forest algorithm yielded the best results. Several decision trees are combined in the Random Forest ensemble learning technique to increase prediction accuracy and decrease overfitting. The WEKA agriculture dataset was used to train and evaluate the model.

The performance results obtained are shown below:

- Correlation Coefficient = 0.8883
- Mean Absolute Error (MAE) = 0.0834
- Root Mean Squared Error (RMSE) = 0.111

These findings show that the Random Forest model produced predictions with extremely high accuracy and minimal error. The anticipated crop yield and the actual crop yield values are strongly correlated, as seen by the high correlation coefficient. In a similar vein, minimum prediction errors are shown by low MAE and RMSE values.

Random Forest's ability to manage big datasets, nonlinear correlations, and several agricultural attributes at once is mostly responsible for its excellent performance. Because agricultural productivity is dependent on numerous interrelated elements, including temperature, soil type, fertilizer use, and rainfall, Random Forest successfully captured these intricate relationships.



In order to assess the impact of ensemble size on prediction performance, the Random Forest model was also trained with varying numbers of decision trees, such as 100, 200, and 300 trees. By lowering variance and enhancing generalization, it was found that adding more trees increased the model's stability and accuracy. Higher tree counts among the studied configurations resulted in more accurate predictions with reduced error rates, albeit the gain was slow after a given amount

of time. This indicates that the size of the ensemble has a significant impact on how well Random Forest models forecast agricultural productivity.

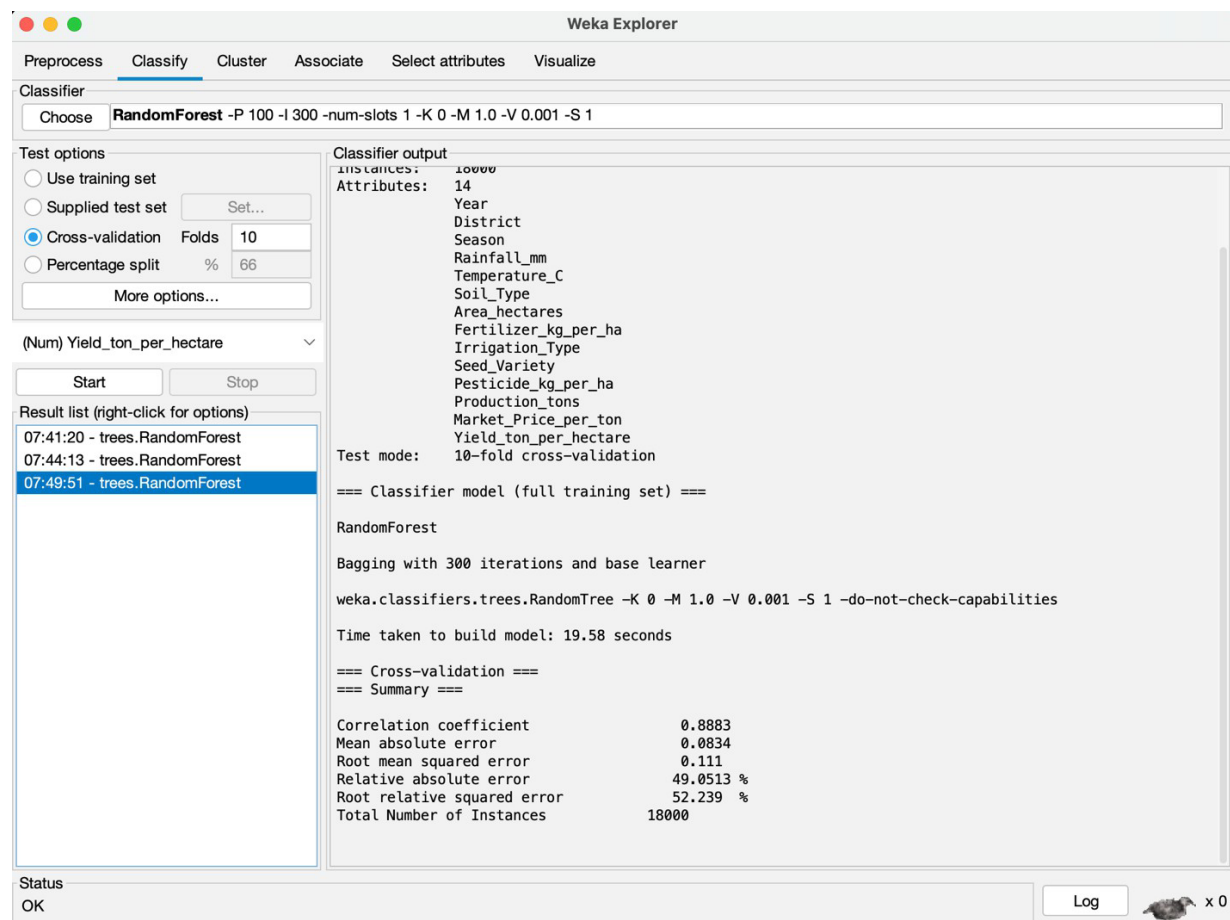


Figure 7.1 here – Random Forest Result Screenshot

7.2 Linear Regression Results

Linear Regression was used as a baseline prediction model in this project. The algorithm attempts to establish a linear relationship between the independent variables and the crop production output.

The obtained results are:

- Correlation Coefficient = 0.5021
- RMSE = 0.1837

The accuracy of the Linear Regression model was much lower than that of Random Forest. Only a moderate link between expected and actual values is shown by the correlation coefficient. Larger prediction mistakes are also indicated by a greater RMSE score.

The assumption of a linear relationship between variables is the primary drawback of linear regression. However, because a variety of farming and environmental factors interact to affect crop output, agricultural records typically contain nonlinear connections. Consequently, the model failed to adequately capture intricate patterns.

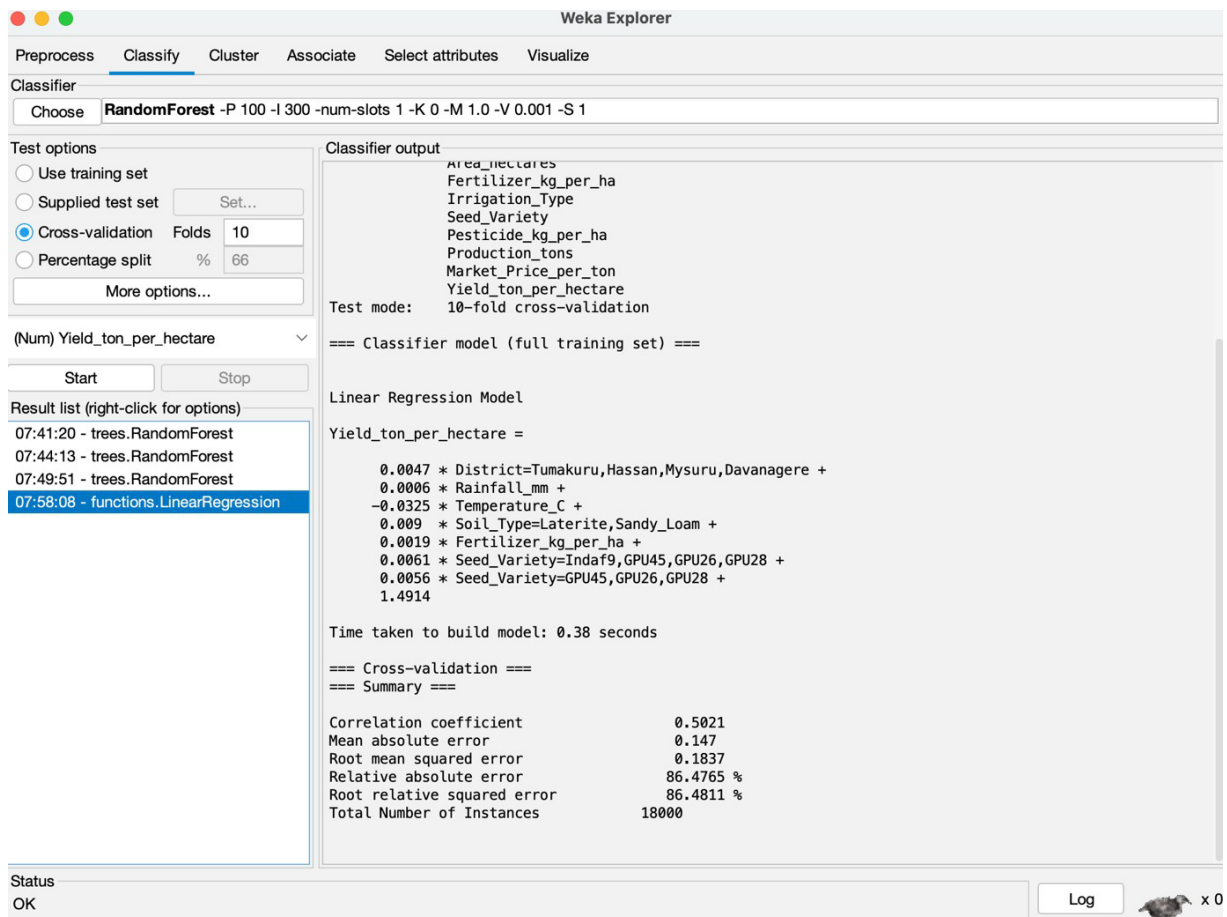


Figure 7.2 here – Linear Regression Screenshot

Although Linear Regression is simple and computationally efficient, it is less suitable for complex agricultural prediction problems.

7.3 Random Tree Results

The Random Tree algorithm was also implemented for crop yield prediction. Random Tree is a decision tree-based algorithm that creates branches based on attribute selection and decision rules.

The obtained performance results are:

- Correlation Coefficient = 0.5841
- MAE = 0.1419
- RMSE = 0.1921

Random Forest outperformed Random Tree, but Random Tree outperformed Linear Regression. When compared to regression, the model's ability to identify certain nonlinear correlations in the dataset increased prediction accuracy. However, because one decision tree can readily overfit the training data, the performance was still inferior than Random Forest. When a model learns the training data too precisely and is unable to effectively generalize to new data, this is known as overfitting. By merging several trees rather than depending just on one, Random Forest gets around this restriction.

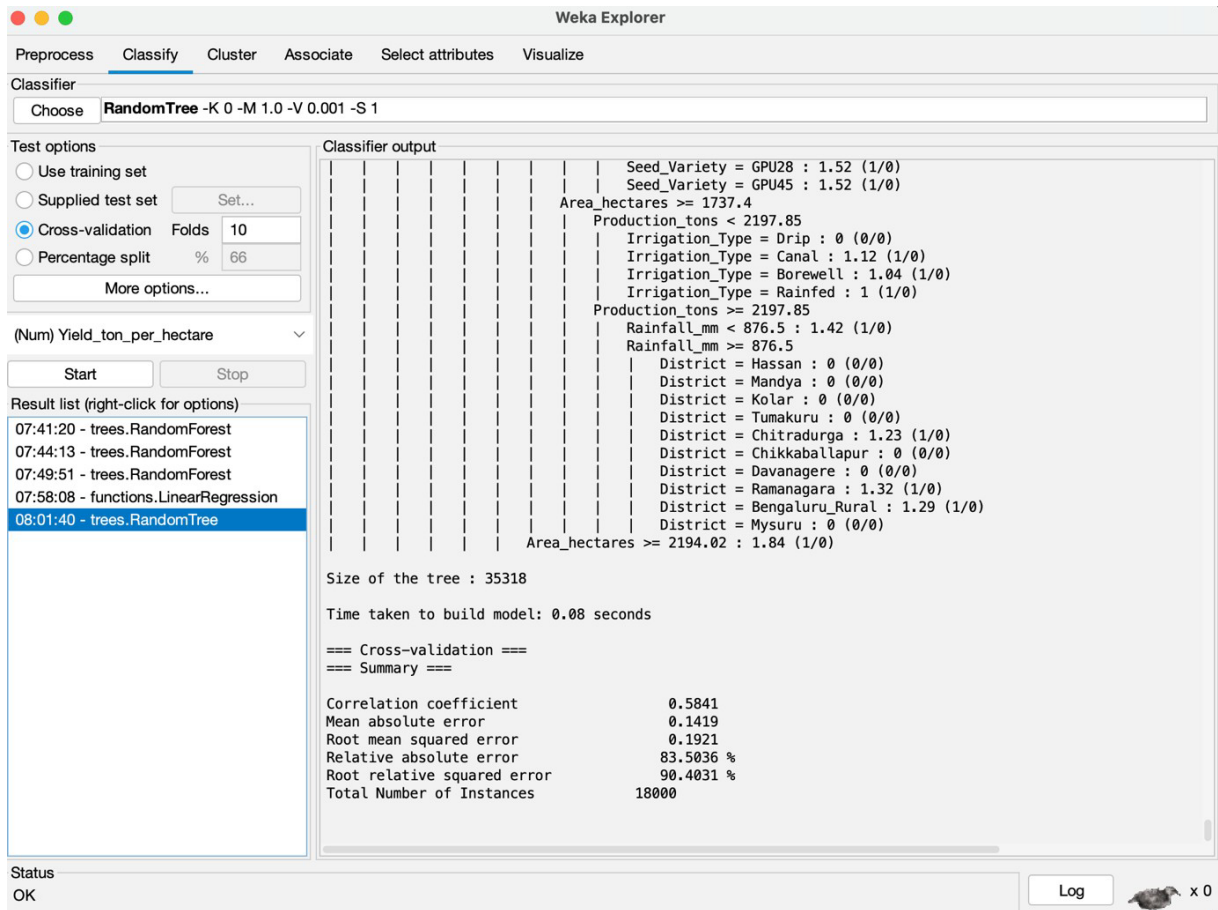


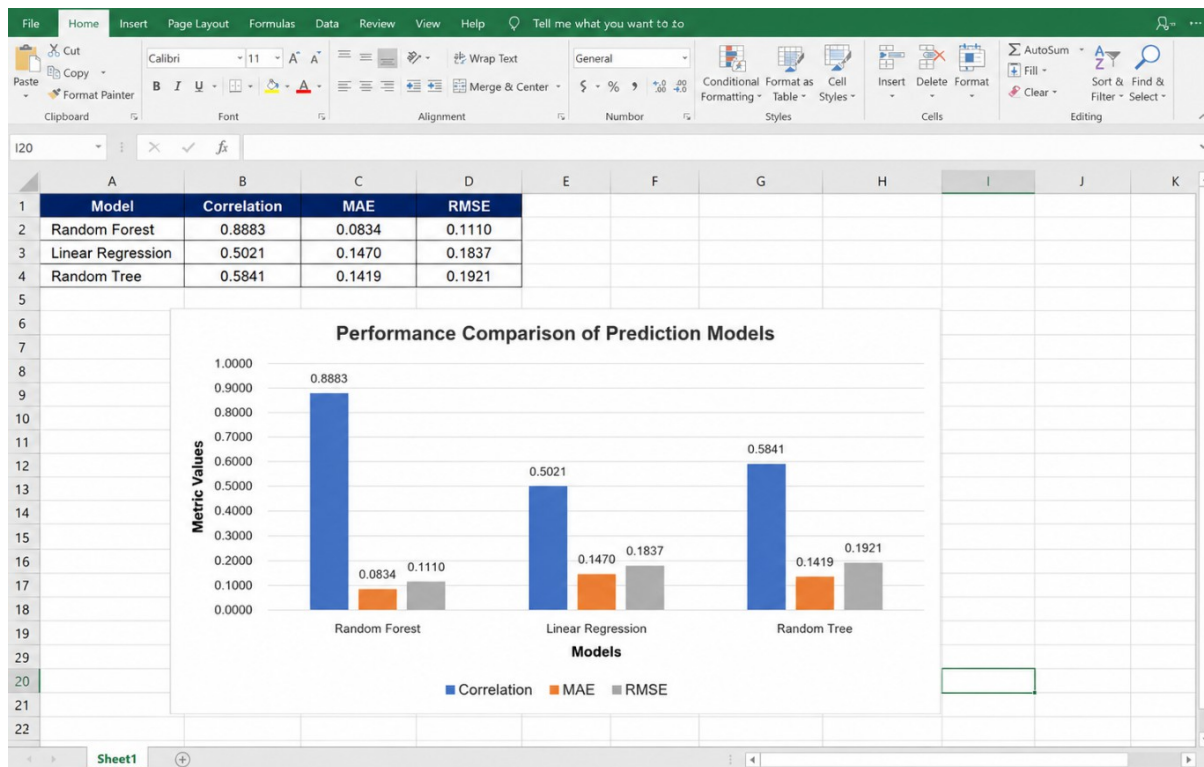
Figure 7.3 here – Random Tree Screenshot

Overall, Random Tree showed moderate predictive capability but lacked the stability and accuracy of the Random Forest model.

7.4 Model Comparison

Model	Correlation	MAE	RMSE
Random Forest	0.8883	0.0834	0.1110
Linear Regression	0.5021	0.1470	0.1837
Random Tree	0.5841	0.1419	0.1921

Table 6.1: Performance Comparison of Prediction Models



With the best correlation coefficient and the lowest error values, Random Forest outperformed all other models, according to the comparative results. Due to its inability to capture the intricate nonlinear interactions between agricultural inputs and crop output, linear regression provided the worst results. Overfitting and instability plagued Random Tree, despite its considerable accuracy. With a high correlation coefficient of 0.8883 (88.83%), Random Forest was shown to be the best model for predicting Ragi crop yield.

7.5 Clustering Results

Similar agricultural records were grouped according to farming and environmental factors using K-Means clustering. The dataset was split into four clusters in this research since the value of K was set to 4.

Agricultural conditions are diverse and multifaceted, as evidenced by the clustering visualization's moderate overlap between certain groupings. The cluster development was controlled by a variety of conditions, including temperature, fertilizer use, and rainfall. Additionally, the clustering study revealed that areas with better fertilizer application and more rainfall typically belonged to clusters linked to higher output levels. This implies that, under appropriate farming conditions, rainfall has Beneficaial impact on crop productivity.

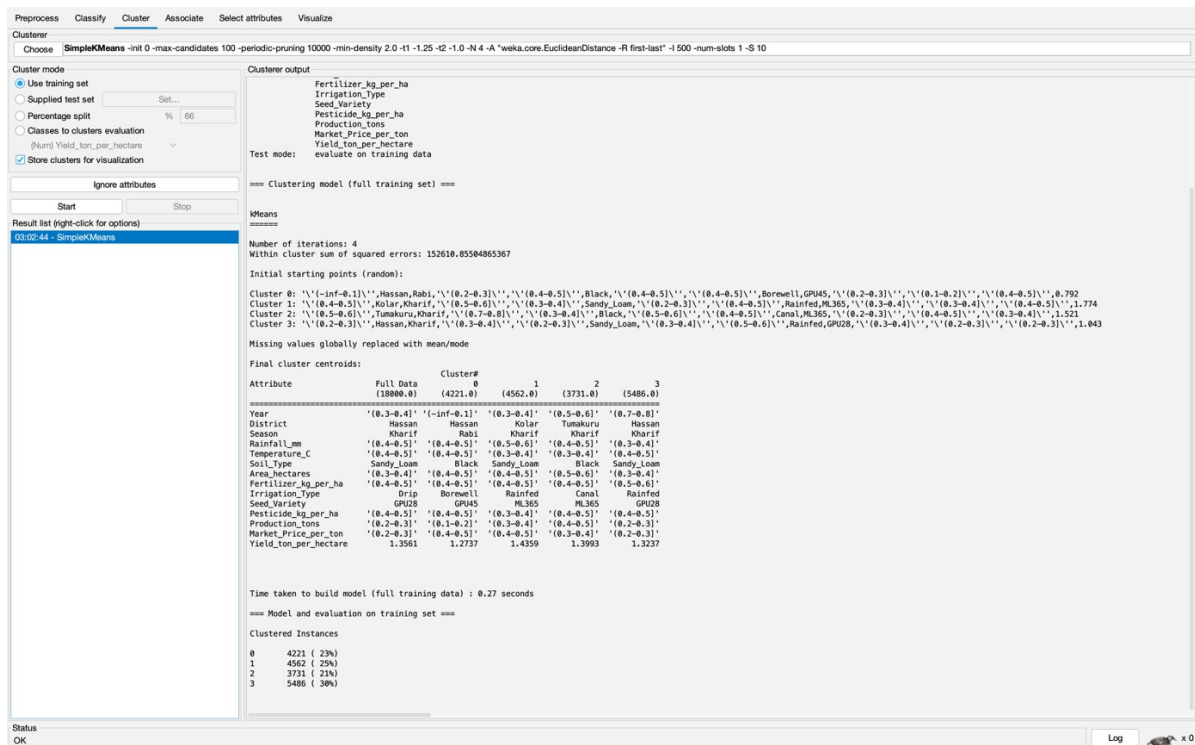


Figure 7.5 – K-Means Clustering Results



Figure 7.5 – K-Means Clustering Visualization

The clustering process helped identify hidden patterns and similarities within the agricultural dataset, which can support better farming decision-making and regional agricultural planning.

1. The Axis Variables (X and Y)
 - **X-Axis (Rainfall_mm):** This represents the annual rainfall ranging from **400 mm** to approximately **1326.9 mm**.
 - **Y-Axis (Yield_ton_per_hectare):** This represents the crop yield output, ranging from **0.69** to **2.19** tons per hectare.
 - **Correlation:** We can observe a general upward "slope" in the data points, which visually confirms the finding that higher rainfall typically leads to higher Ragi yields.

2. The Color-Coded Clusters

The algorithm has assigned each of the 18,000 records to one of four clusters, represented by different colours at the bottom of the screen:

- **Cluster 0 (Blue)**
- **Cluster 1 (Red)**
- **Cluster 2 (Green)**
- **Cluster 3 (Cyan)**

3. Key Interpretations of the Visual

- **Cluster Overlap:** The colors in the plot's center significantly overlap. This suggests that agricultural conditions are complex; rainfall and yield by themselves do not constitute a cluster; temperature, soil type, fertilizer use, and other hidden elements also play a role in the formation of these clusters.
- **Density:** The majority of the dataset reflects moderate agricultural circumstances, as seen by the highest concentration of dots in the middle of the graph (around 863 mm of rainfall and 1.44 tons of produce).
- **Outliers:** The most productive farming situations are represented by the dispersed points at the very top (high yield) and far right (high rainfall), which your Apriori rules probably associate with the Kharif season or ideal fertilizer use.

7.6 Apriori Association Rule Results

The agricultural dataset's hidden links and recurring patterns were found using the Apriori method. The dataset was discretized prior to using Apriori since association rule mining necessitates nominal or categorical data.

A number of significant agricultural links were identified using the created association rules. Production levels and farmed area were found to be strongly correlated. Furthermore, a number of criteria indicated a positive correlation between increased agricultural production and the Kharif season.

The quality of association rules was evaluated using support, confidence, and lift values:

- **Support** measures how frequently a rule appears in the dataset.
- **Confidence** measures how often the rule is correct.
- **Lift** measures the strength of the relationship between attributes.

Several rules produced a lift value greater than 1, indicating positive associations between farming conditions and crop yield outcomes.

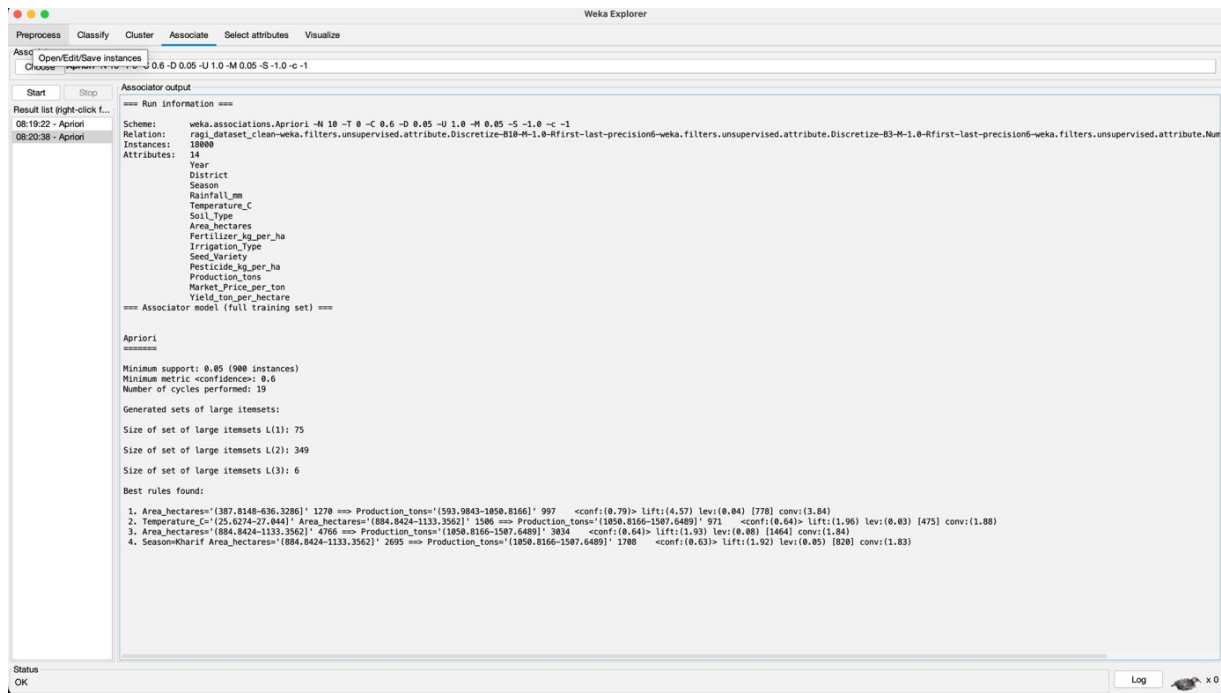


Figure 7.6 – Apriori association rules generated in WEKA showing relationships between cultivated area, season, temperature, and production levels.

The Apriori analysis successfully identified meaningful agricultural patterns that can help farmers and policymakers understand the factors contributing to improved crop production.

Best Rules Found:

- **Rule 1 (Production vs. Area):** This rule shows that when the farmed area is between **387.8 and 636.3 hectares**, there is a strong correlation (confidence) with production levels between **593.9 and 1050.8 tons**. The **Lift of 4.57** is very high, indicating a very strong positive relationship.
- **Rule 2 (Climate & Area):** When the temperature is between **25.6°C and 27.0°C** and the farmed area is moderate to large, production typically falls into the higher range.

8. Discussion

8.1 Interpretation of Findings

The project's findings demonstrate that environmental and farming-related factors have a significant impact on agricultural yield. Rainfall has a significant favorable impact on agricultural yield, according to the data. Higher output levels were typically found in areas or records with greater rainfall conditions. In a similar vein, increased agricultural productivity was greatly influenced by both farmed area and fertilizer use.

8.2 Strengths of the Project

This project contains several important strengths:

- Multiple data mining techniques were applied, including prediction, clustering, and association rule mining.
- The Random Forest model achieved strong predictive performance with high correlation and low error values.
- Data preprocessing techniques such as normalization, missing value handling, and discretization improved data quality and model performance.

8.3 Limitations of the Project

Lastly, the WEKA software platform was used exclusively for all tests. Programming platforms like Python or R offer more flexibility, scalability, and sophisticated visualization capabilities, even though WEKA is useful for academic study. All things considered, the research effectively illustrated the value of data mining methods for agricultural analysis and prediction while also pointing up opportunities for future development.

9. Conclusion

Using an agricultural dataset of 18,000 records, this study effectively used machine learning and data mining approaches for Ragi crop yield prediction. WEKA was used to implement a number of algorithms, such as Random Forest, Linear Regression, Random Tree, K-Means clustering, and Apriori. Random Forest outperformed all other models in terms of prediction accuracy and error levels. Important agricultural trends and linkages were also found by association rule mining and clustering. Overall, the results demonstrate how data mining may help with agricultural decision-making, boost production, and promote smart farming.

10. Recommendations and Future Work

In order to increase prediction accuracy, future research can enhance the system by utilizing sophisticated deep learning models like ANN, CNN, and LSTM. Real-time agricultural support can be achieved by combining mobile applications, weather predictions, and IoT sensor data. These enhancements have the potential to increase resource management, productivity, smart farming, and decision-making.

11. References:

- [1] References[1] M. Javaid, A. Haleem, I. H. Khan and R. Suman, "Understanding the potential applications of Artificial Intelligence in Agriculture Sector," vol. 2, no. 1, pp.15–30, 2023, doi: 10.1016/j.aac.2022.10.001. Available at: <https://www.sciencedirect.com/science/article/pii/S277323712200020X>.
- [2] M. Javaid, A. Haleem, I. H. Khan and R. Suman, "Understanding the potential applications of Artificial Intelligence in Agriculture Sector," vol. 2, no. 1, pp.15–30, 2023, doi: 10.1016/j.aac.2022.10.001. Available at: <https://www.sciencedirect.com/science/article/pii/S277323712200020X>.
- [3] V. Choudhary, P. Guha, G. Pau and S. Mishra, "An overview of smart agriculture using internet of things (IoT) and web services," vol. 26, pp.100607, 2025, doi: 10.1016/j.indic.2025.100607. Available at: <https://www.sciencedirect.com/science/article/pii/S2665972725000285>.
- [4] S. Castillo-Girones, S. Munera, M. Martínez-Sober, J. Blasco, S. Cubero and J. Gómez-Sanchis, "Artificial Neural Networks in Agriculture, the core of artificial intelligence: What, When, and Why," *Comput.Electron.Agric.* vol. 230, pp.109938, 2025, doi: 10.1016/j.compag.2025.109938. Available at: <https://www.sciencedirect.com/science/article/pii/S0168169925000444>.
- [5] I. K. Nti, A. Zaman, O. Nyarko-Boateng, A. F. Adekoya and F. Keyeremeh, "A predictive analytics model for crop suitability and productivity with tree-based ensemble learning," vol. 8, pp.100311, 2023, doi: 10.1016/j.dajour.2023.100311. Available at: <https://www.sciencedirect.com/science/article/pii/S2772662223001510>.
- [6] I. K. Pious, et al., "Enhancing Prediction Accuracy Through Random Forest in Classification and Regression", 2024 *International Conference on Smart Technologies for Sustainable Development Goals (ICSTSDG)*, pp. 1–6. [Online].doi: 10.1109/ICSTSDG61998.2024.11026358.
- [7] M. A. Javed and M. A. Azmi Murad, "Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability," vol. 10, no. 24, pp.e40836, 2024, doi: 10.1016/j.heliyon.2024.e40836. Available at: <https://www.sciencedirect.com/science/article/pii/S2405844024168673>.
- [8] D. K. Pandey and R. Mishra, "Towards sustainable agriculture: Harnessing AI for global food security," vol. 12, pp.72–84, 2024, doi: 10.1016/j.aiia.2024.04.003. Available at: <https://www.sciencedirect.com/science/article/pii/S2589721724000151>.
- [9] N. Aijaz, H. Lan, T. Raza, M. Yaqub, R. Iqbal and M. S. Pathan, "Artificial intelligence in agriculture: Advancing crop productivity and sustainability," vol. 20, pp.101762, 2025, doi: 10.1016/j.jafr.2025.101762. Available at: <https://www.sciencedirect.com/science/article/pii/S2666154325001334>.

- [10] V. Choudhary, P. Guha, G. Pau and S. Mishra, "An overview of smart agriculture using internet of things (IoT) and web services," vol. 26, pp.100607, 2025, doi: 10.1016/j.indic.2025.100607. Available at: <https://www.sciencedirect.com/science/article/pii/S2665972725000285>.
- [11] P. Kotak and H. Modi, "Enhancing the Data Mining Tool WEKA," 2020 5th International Conference on Computing, Communication and Security (ICCCS), Patna, India, 2020, pp. 1-6, doi: 10.1109/ICCCS49678.2020.9276870.
- [12] A. Suresh, P. Ganesh Kumar and M. Ramalatha, "Prediction of major crop yields of Tamilnadu using K-means and Modified KNN," 2018 3rd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2018, pp. 88-93, doi: 10.1109/CESYS.2018.8723956.