Smart Crop Prediction for Agricultural Transformation

Prepared for

Reza Ebrahimi

University of South Florida - Muma College of Business

by

Safiya Joseph

Chi Phuong Diep

Carolina Aldana Yabur

Bhargav Rishi Medisetti

Venkata Sai Niharika Allu

Divya Sharmila Penumaka

November 20, 2024

Executive Summary

Agriculture remains a cornerstone of human survival and economic stability, yet it faces mounting challenges from global population growth, climate change, and resource constraints. In response, precision agriculture and advanced technologies are emerging as critical solutions to improve crop productivity and sustainability. This report focuses on developing a **Smart Crop Prediction System** using machine learning to identify the "most suitable crop" or the crop that would give the largest yield with the given conditions.

Utilizing a dataset with features such as soil pH, electrical conductivity, nutrient levels, moisture, and temperature, the study applied various machine learning models, including Naive Bayes, Support Vector machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, Logistic Regression, Neural Networks, and Random Forest. Each model was evaluated using robust metrics, including accuracy, precision, recall, and F1-score, supported by cross-validation to ensure reliability.

Key findings reveal that the **Decision Tree** model emerged as the most effective, achieving an accuracy of **91.07%**, with superior precision, recall, and F1-score compared to other models. The **Random Forest** model performed equally well, with an accuracy of **91.03%**, demonstrating its ability to make accurate and interpretable predictions. The **SVM** model performed remarkably well with an accuracy of **88.93%,** standing as our final model to exceed the baseline model, Naive Bayes. The **Naive Bayes** model performed admirably as a baseline, achieving an accuracy of **88.49%**. While it is efficient and simple, its reliance on feature independence limits its ability to handle complex relationships within the dataset. The **Logistic Regression** model performed reasonably well, achieving an accuracy of **88%**. While it is easy to interpret and implement, its linear nature limits its ability to capture complex non-linear relationships, making it less effective for datasets with intricate crop-environment interactions. The **Neural Network** model demonstrated promise, achieving a cross-validation accuracy of **79.62%.** However, its performance was constrained by the dataset size and computational resources. This study recommends deploying the models through a user-friendly application accessible to farmers, enabling them to input soil parameters and receive tailored crop recommendations.

By combining technological innovation with farmer support, this project lays a foundation for transforming agriculture, optimizing resource use, reducing crop failure risks, and contributing to global food security. The adoption of such solutions can significantly improve decision-making in farming and drive the transition toward a sustainable and resilient agricultural future.

To further enhance predictive capabilities, integrating IoT sensors for real-time data collection and providing training programs for farmers are essential steps. Ongoing data collection and model refinement will ensure adaptability to changing conditions, fostering sustainable agricultural practices.

Introduction

In today's fast-paced world, agriculture often goes unrecognized despite being the backbone of human survival and well-being. As lives get busier, people may pay less attention to the nutritional quality of foods, yet consumers still prefer high-quality products when given the choice. This high demand creates challenges for the agricultural industry, which must continuously ensure a steady, reliable, and high-quality supply of crops.

With global population growth and climate change increasingly threatening agriculture by negatively impacting crops and livestock, there is a growing need for sophisticated and efficient food management practices to combat rising food insecurity. According to the *Global Commission on Adaptation* report (2019), "These challenges include a likely 50 percent increase in global demand for food between 2010 and 2050, with even larger increases in the world's most food-insecure regions—about a threefold increase in sub-Saharan Africa and almost twofold in South Asia."

Agricultural productivity growth is therefore essential to alleviate poverty and improve food security. This highlights the importance of responsibly regulating and managing resources to optimize crop performance (García, Zambrano, Alcivar & Romero, 2020). To meet these demands, a critical question arises: Which crops are best suited for specific environments, and when should they be planted to maximize yields, minimize costs, and ensure sustainability?

To address these challenges, this paper employs a machine learning approach to explore the following key questions:

- Which machine learning models have the strongest predictive capacity to predict the most suitable crop for specific soil and environmental conditions?
- What features best determine crop suitability, & how can they guide precision agriculture?

The primary objective of this study was to utilize a machine learning framework to analyze key agronomic features and predict the most suitable crop for given soil and environmental conditions. This approach supports precision agriculture practices and contributes to sustainable farming by optimizing resource use and reducing risks associated with crop failure.

Background

Agriculture plays a critical role in ensuring food security and economic stability worldwide. As global populations grow and climate change alters environmental conditions, precision agriculture has emerged as a vital approach to maximize crop productivity and adaptability. One of the key challenges in precision agriculture is identifying the most suitable crop for a given set of soil and environmental conditions, including soil pH, nutrient levels, moisture, and temperature.

1. *Climate change* is adding stress to the agricultural system, such as extreme weather, soil erosion, water shortages, temperature swings, and many other issues.

2. *Time and productivity* are impacted by the decision-making regarding planting, watering, and crop survival, with each failed crop representing a substantial loss of income and resources.
3. *Competitors* are increasingly adopting advanced technology to compete for market positions and set a higher bar.
4. *Sustainability and nutrition* need to be persistently maintained to produce high-quality crops that achieve consumer expectations and maximize profits.
5. *Market trend fluctuations* tend to increase instability of the market price and consumers' demand which makes it more difficult to plan and manage the production line.

Methodology

This report proposes a smart crop prediction system that uses a comprehensive machine learning methodology to address the agricultural sector's urgent issues. The objective is to predict the crop best suited for given specific soil conditions.

The dataset was chosen from the Kaggle online platform. It contains 10 columns including the target variable, Plant Type, which is categorical and 9 numerical features which include soil pH, electrical conductivity (Soil EC), Phosphorus, Potassium, Urea, Triple Superphosphate (T.S.P), Muriate of Potash (M.O.P), Moisture, and Temperature. The target variable represents the crop that would give the largest yield with the given conditions. The features are critical agronomic parameters that influence crop growth and hence are suitable for predictive modeling in agriculture. The goal is to predict the most suitable crop based on the given soil and environmental conditions.

Column Names

1. pH: Soil pH value.
2. Soil EC: Electrical conductivity of the soil.
3. Phosphorus: Phosphorus content in the soil.
4. Potassium: Potassium content in the soil.
5. Urea: Urea content in the soil.
6. T.S.P: Triple superphosphate content in the soil.
7. M.O.P: Muriate of potash content in the soil.
8. Moisture: Soil moisture level.
9. Temperature: Soil temperature.
10. Plant Type: Target variable, indicating the crop type that would give the largest yield.

This is a multi-class classification problem and will include using cross-validation to ensure the confidence of each model, constructing the prediction models, training on a 72% train set, testing on a 28% test set, performing analysis with evaluation metrics, and providing recommendations. The Naïve Bayes model will be used as the baseline alongside models such as Decision Tree, Logistic Regression, K-Nearest Neighbors, Random Forest, and Neural Network. Through model implementation and analysis, we aim to assist farmers, and our business in predicting different Plant Types with the highest accuracy. Some of the metrics used for model

evaluation include accuracy, precision, recall, F_1 score, AUC and ROC Curve.

Analysis & Findings

Naïve Bayes Model Analysis

The Gaussian Naive Bayes (GNB) model was used for its simplicity, efficiency, and suitability for continuous data like pH and temperature, assuming a Gaussian distribution. It handles multi-class classification effectively and serves as a robust baseline for comparison with more complex models. Overall, it is an accessible and interpretable model that offers a balance between performance and simplicity, making it a practical initial choice for the crop recommendation task. The mean cross-validation score was **0.8883**, or **88.8%**, indicating that the model consistently performs well across different subsets of the data.

After training the GNB model, its performance on the test data was evaluated using several key metrics:

- The model achieved an **accuracy** of **0.8849** or **88.49%**, indicating that approximately 88% of predictions matched the actual crop type.
- The **precision** score was **0.8925**, reflecting the proportion of correctly predicted crop types out of all predictions made for a given crop.
- The **recall** score, which measures the model's ability to identify the correct crop type among all actual instances, was **0.8854**.
- The **F1-score**, which balances precision and recall, was calculated to be **0.8889**.

These metrics suggest that the Gaussian Naive Bayes model performs well overall, with strong classification capabilities for the crop types, especially for Carrots, Rice, Tomato, and Wheat. However, some areas for improvement were noted in certain crop types, such as Chili, Eggplant, and Sunflowers, where misclassifications were observed.

Support Vector Machine Model Analysis

A support vector machine model was developed to predict Plant Type best suited for given soil and environmental conditions. The model was validated using 5-fold cross-validation to ensure robust evaluation and mitigate the effects of data variability. The model had a cross-validation score of **0.8893** ± 0.0028. After training the SVM model, its performance on the test data was evaluated using several key metrics.

- The model achieved an **accuracy** of **0.8869** or **88.69%** indicating that approximately 89% of predictions matched the actual crop type.
- The **precision score** was **0.8879**, reflecting the proportion of correctly predicted crop types out of all predictions made for a given crop.
- The model's ability to identify the correct crop type among all actual instances, the **recall score**, was **0.8863**.
- The model had an **F1-score** of **0.8860**.

These results demonstrate that the SVM model is extremely effective in predicting crop types based on the given dataset. The findings from this analysis indicate that the model performed better than our base model, Naïve Bayes, and had been chosen as one of our preferred models.

Logistic Regression Model Analysis

To predict the crop type, we developed a Logistic Regression model based on soil and environmental factors. We applied a 5-fold cross-validation to assess its consistency, yielding an average cross-validation accuracy of **0.89 ± 0.0023**, which means that the model has an average accuracy of 89% across different cross-validation folds, with a standard deviation of 0.0023. This indicates that the model's performance is quite consistent across different subsets of the data, and the variation in accuracy is minimal.

After training the Logistic Regression model, its performance was evaluated on the test data using key metrics:

- **Accuracy**: The model achieved an accuracy of **0.88** or **88%**, indicating that the model correctly identified the crop type in 88% of the instances.
- **Precision**: The precision score was **0.88**, reflecting the proportion of correctly predicted crop types out of all predictions made for a given crop.
- **Recall**: The recall score was **0.88**, which means the model was able to correctly identify **88%** of all instances of a given crop type.
- **F1-Score**: The F1-score was **0.88**, which means the model achieved a balanced performance between precision and recall. This indicates a good overall performance.

The Logistic Regression model demonstrated strong performance in predicting crop types. While simpler models like this may not outperform more complex algorithms like Random Forests, they offer several advantages such as simplicity (the model is easy to understand and interpret), efficiency (it requires less computational resources to train and predict), and reliability (despite its simplicity, the model consistently delivers accurate results).

Neural Network Model Analysis

A neural network model using Stochastic Gradient Descent solver was designed for the predictions. It was found that the 'sgd' solver performed better than the Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm. This model had a cross-validation training accuracy of **0.7962 ± 0.0394.**

On the test data, the model's performance across various evaluation metrics is as follows:

- **Accuracy:** The model achieved an accuracy of **0.8061** or **80.61%**
- **Precision:** The precision score was **0.8169** or **81.69%**, reflecting the proportion of correctly predicted crop types out of all predictions made for a given crop.
- **Recall:** The recall score, which measures the model's ability to identify the correct crop type among all actual instances, was **0.8055** or **80.55%**.
- **F1-Score:** The F1-score, which balances precision and recall, was calculated to be **0.8051** or **80.51%**.

Although this model averaged about **80%** across all the metrics, the model underperformed when compared to Naïve Bayes, and hence, this model was not chosen as one of our preferred models.

Decision Tree Model Analysis

A Decision Tree Classifier was designed to predict the most suitable Plant Type based on soil and environmental factors. A 5-fold cross-validation was implemented to ensure a thorough and reliable assessment of the model. The cross-validation accuracy was **0.9123 ± 0.0023**

After training the Decision Tree model, its performance on the test data was evaluated using several key metrics.

- The model achieved an accuracy of **0.9107**, indicating that approximately **91.07%** of the predictions made by the model were correct.
- The precision score was **0.9362**, demonstrating that **93.62%** of the crop types predicted by the model were accurate.
- A recall score of **0.9098** reflects the model's ability to correctly identify **90.98%** of the actual crop types in the dataset.
- The F1-score was **0.9055**, representing a balanced trade-off between precision and recall.

Random Forest Model Analysis

To end with, we worked with Random Forest Model to predict the crop type. We used 200 estimators, a maximum depth of 8 and entropy as our criterion for this model. The cross-validation scores resulted in a training accuracy of **0.9128 ± 0.0026**, indicating high consistency and stability in the model's performance during training.

After training the Random Forest model, its performance on the test data was evaluated using several key metrics.

- **Accuracy:** The model achieved an accuracy of **0.9103** or **91.03%**, indicating that approximately 91% of predictions matched the actual crop type.
- **Precision:** The precision score was **0.9360** or **93.6%**, indicating the proportion of correctly predicted crop types out of all predictions made for a given crop.
- **Recall:** The recall score, which measures the model's ability to identify the correct crop type among all actual instances, was **0.9094** or **90.94%**.
- **F1-Score:** The F1-score, which balances precision and recall, was calculated to be **0.9051** or **90.51%**.

The results of Decision Tree and Random Forest models had the highest effectiveness in predicting the crops. These models topped the charts across all the evaluation metrics with little differences between them and would serve as the most preferred models in using machine learning in agriculture.

Implementation Plan

The goal of this implementation plan is to empower farmers with actionable insights through machine learning-powered crop prediction, fostering sustainable agricultural practices. A user-friendly mobile or web application will be developed to serve as the primary interface for farmers. The application will allow farmers to input soil and environmental parameters to receive tailored crop recommendations and related resource needs, such as water, fertilizer, and temperature management. To ensure accessibility, the application will be multilingual, compatible with various devices, and capable of generating visualizations and reports for enhanced understanding of the analytics.

To automate and enhance data collection, IoT (Internet of Things) sensors will be integrated into the system, bridging the gap between physical and digital environments. These sensors will collect and transmit real-time soil and environmental data directly to the platform, minimizing manual effort and improving accuracy. This continuous data flow will enable the machine learning models to adapt to changing conditions and refine predictions over time. Additionally, periodic surveys will gather insights on crop yields, further supporting model evaluation and improvement.

Farmer training and support programs will be key to ensuring effective use of the technology. Workshops and hands-on training will educate farmers on soil testing, data input, and interpreting application outputs. Local agricultural advisory centers will be established to provide on-demand assistance, enabling farmers to resolve issues, maximize benefits, and trust the reliability of the system.

In the long term, the system will be continuously updated with new data to maintain accuracy and relevance. This will not only address immediate farming challenges but also promote sustainable practices, reduce waste, and optimize resource use. By leveraging advanced technology and offering consistent support, this plan aims to transform agriculture, fostering resilience and innovation that will benefit both current and future generations.

Conclusion

This study aimed to address the pressing challenges in agriculture by developing a Smart Crop Prediction System using various machine learning models. The models were evaluated based on their performance in predicting the most suitable crops for specific soil and environmental conditions. The findings from this study aim to provide valuable insights into the potential of machine learning in transforming agricultural practices.

Our findings reveal that the Decision Tree and Random Forest models emerged as the most effective in the predictions. Random Forest achieved an accuracy of 91.03%, precision of 0.9360, recall of 0.9094, and an F1-score of 0.9051. The Decision Tree performed with an accuracy of 91.07%, precision of 0.9362, recall of 0.9098, and an F1-score of 0.9055, demonstrating its ability to make accurate and interpretable predictions.

The SVM model demonstrated strong performance with an accuracy of 88.69%, precision of 88.79%, recall of 88.63%, and an F1-score of 88.60%. These results indicate that the SVM model is a robust choice for crop prediction, offering high accuracy and balanced precision and recall, making it a valuable addition to our Smart Crop Prediction System.

Another strong candidate for practical applications in agriculture, especially when computational resources are limited, is the Logistic Regression model, which achieved an accuracy of 88%, and its precision, recall, and F1-scores consistently remained steady at around 88%, demonstrating balanced classification. Another model with similar performance was Gaussian Naive Bayes with an accuracy of 88.49%, precision of 0.8925, recall of 0.8854, and an F1-score of 0.8889. Its simplicity and efficiency make it a solid baseline model, though it struggles with complex relationships in the data. This model is particularly effective for crops like Carrots, Rice, Tomato, and Wheat, but less so for Chili, Eggplant, and Sunflowers.

The Neural Network model showed promise, achieving a cross-validation accuracy of 79.62%, but its performance was constrained by the dataset size and computational resources. It underperformed compared to simpler models like Naive Bayes, making it less preferable for this application. We also worked on another model, the KNN model. This model showed an accuracy of only 72.5%. While it provides reasonable predictions, its performance is limited, indicating the need for further optimization and leading us to exclude its analysis from our final report.

In conclusion, the study underscores the importance of integrating advanced machine learning models into agricultural practices to optimize crop selection and improve productivity. The Decision Tree model, in particular, stands out as the most promising tool for practical deployment. By leveraging these technological advancements, we can support farmers in making informed decisions. Future work should focus on integrating climate related data, IoT sensors for real-time data collection, providing training programs for farmers, and continuously refining the models to adapt to changing conditions. With the future scope indicating an increase in the dimensionality of the data, Random Forest could prove to be the better performing model and should be switched with Decision Tree as the primary model. These steps will ensure the ongoing success and scalability of the Smart Crop Prediction System, contributing to global food security and sustainable agricultural practices.

References

Arteaga, J. J. G., Zambrano, J. J. Z., Cevallos, R. A., & Romero, W. D. Z. (2020). Predicción del rendimiento de cultivos agrícolas usando aprendizaje automático. [Predicting agricultural crop yield using machine learning]. *Revista Arbitrada Interdisciplinaria Koinonía*, *5*(2), 144-160. http://dx.doi.org/10.35381/r.k.v5i2.1013

Global Commission on Adaptation. (2019). *Adapt now: A global call for leadership on climate resilience*. World Resources Institute.

vision, r3tro. (2023, November 3). Soil moisture , temp and nutritions. Kaggle. https://www.kaggle.com/datasets/r3trovision/soil-moisture-temp-and-nutritions/data

Weil, R. R., & Brady, N. C. (2017). The nature and properties of soils. Pearson.

Appendices

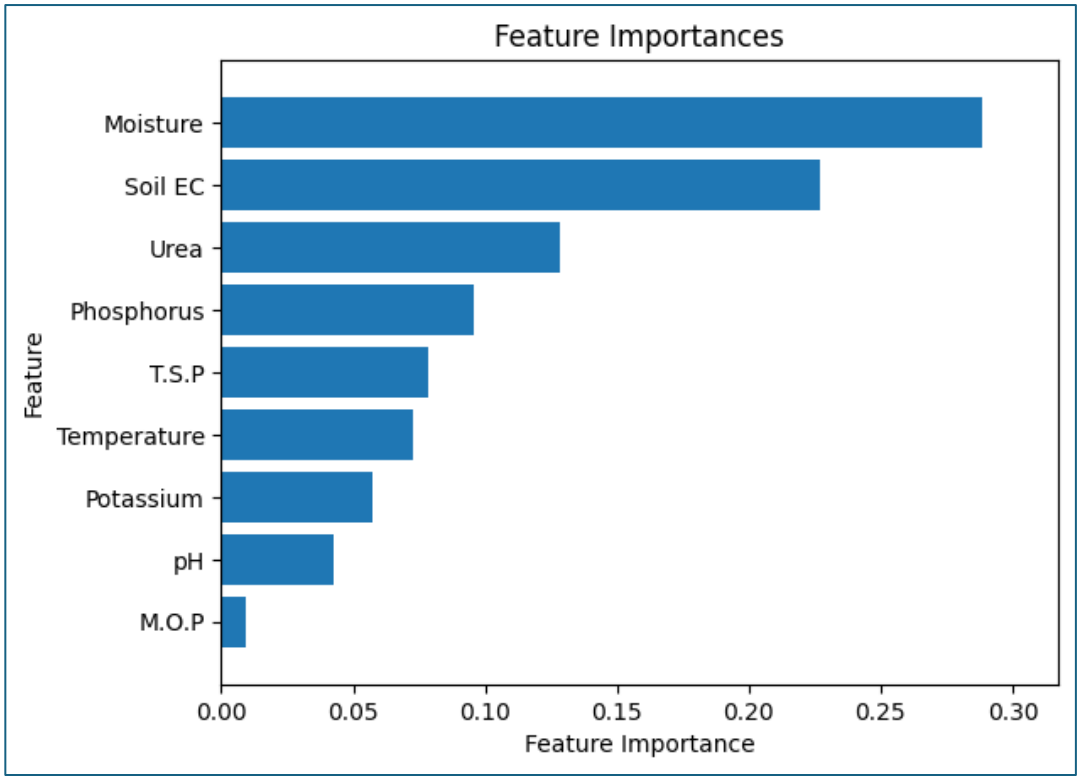

Chart 1 showing the proportion of influence each feature has on the Target Variable, Plant Type.

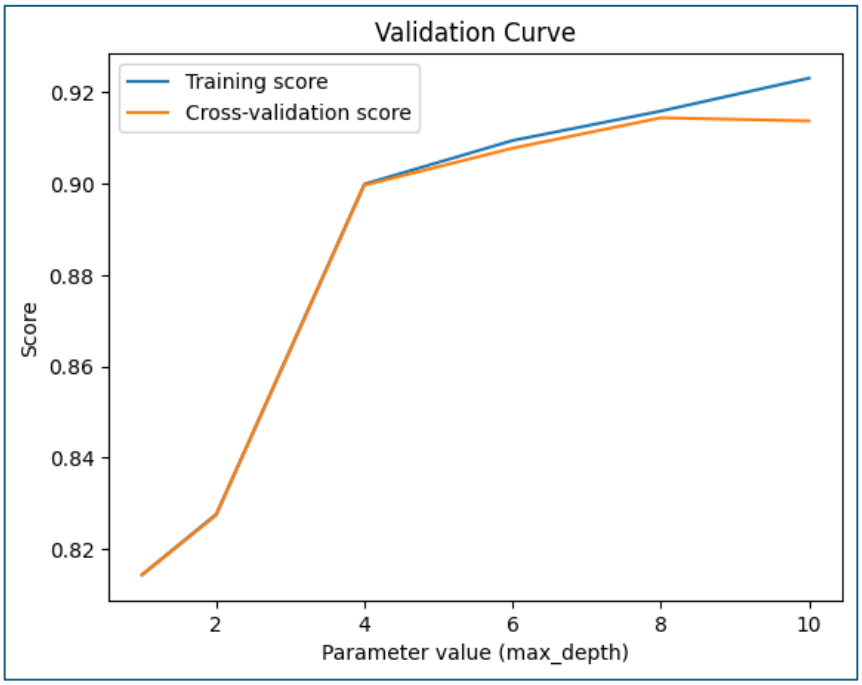

Chart 2 showing Training and Cross-validation scores against max depth.

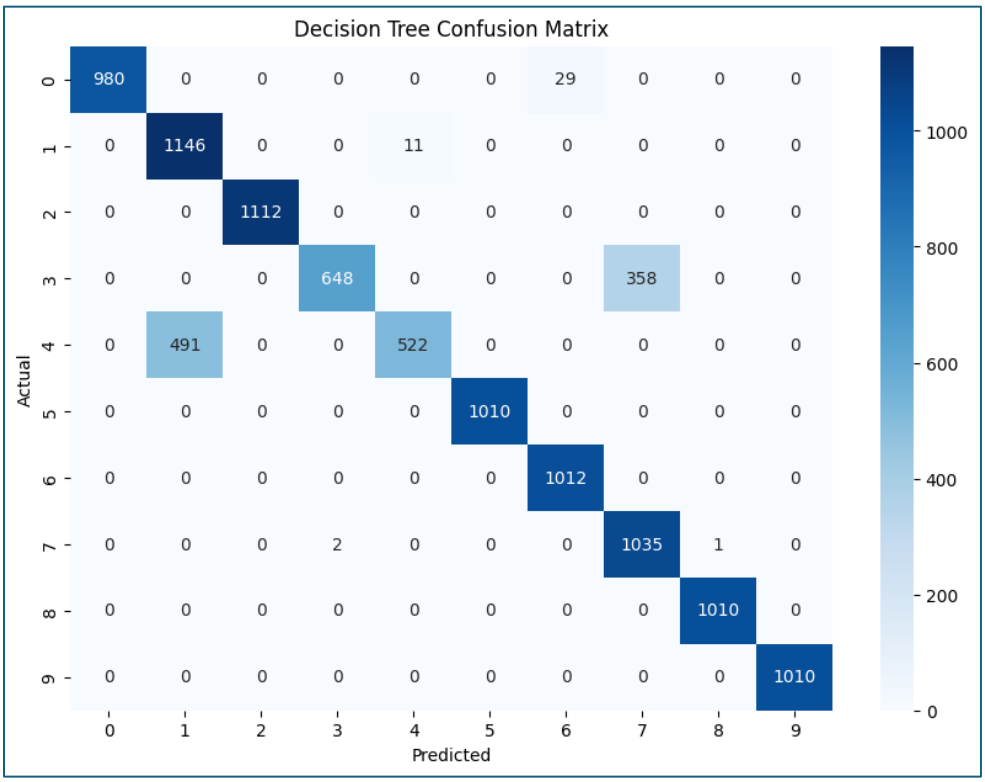

Chart 3 showing the Confusion Matrix for the Decision Tree Model.

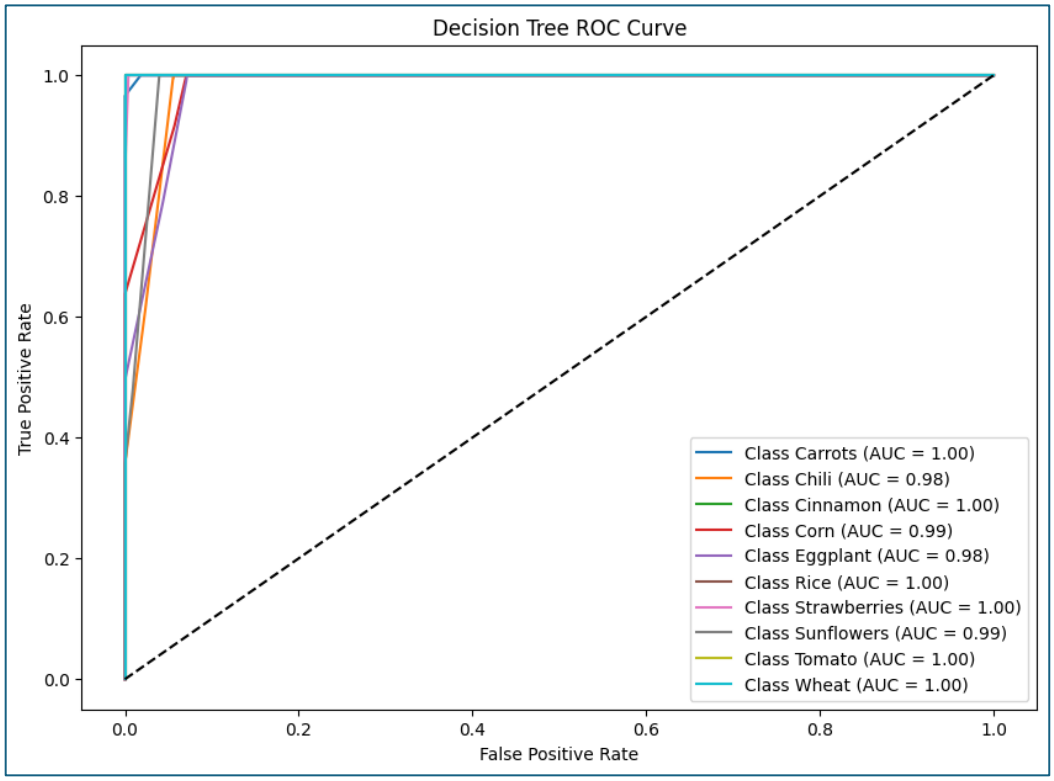

Chart 4 showing the ROC Curve for the Decision Tree Model.

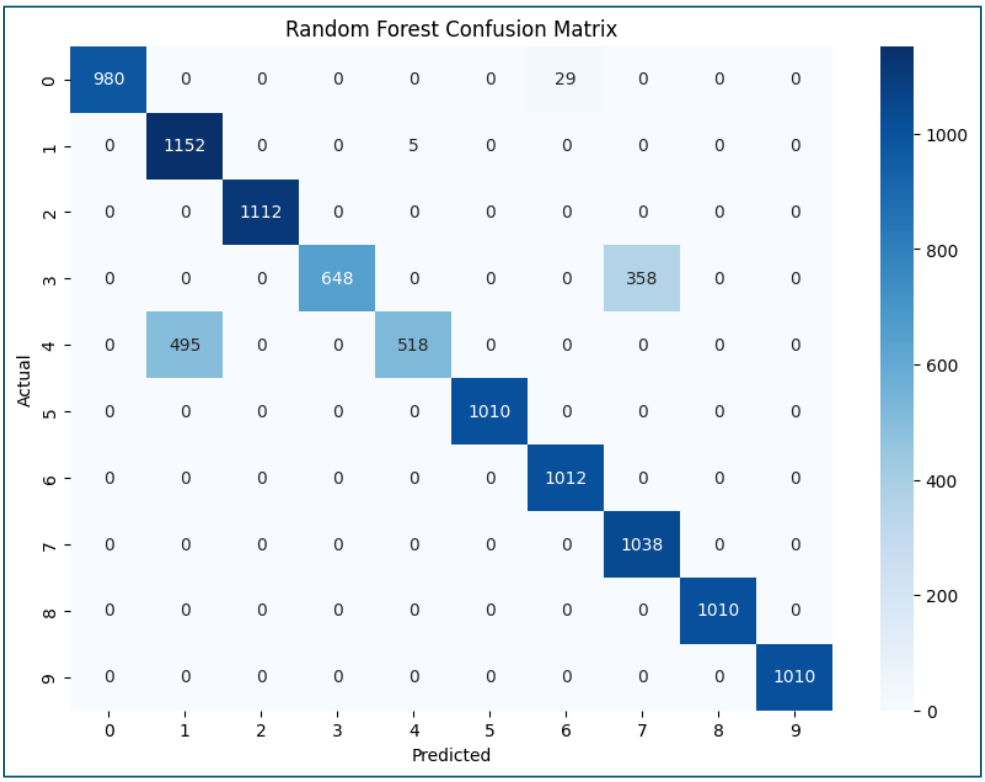

Chart 5 showing the Confusion Matrix for the Random Forest Model.

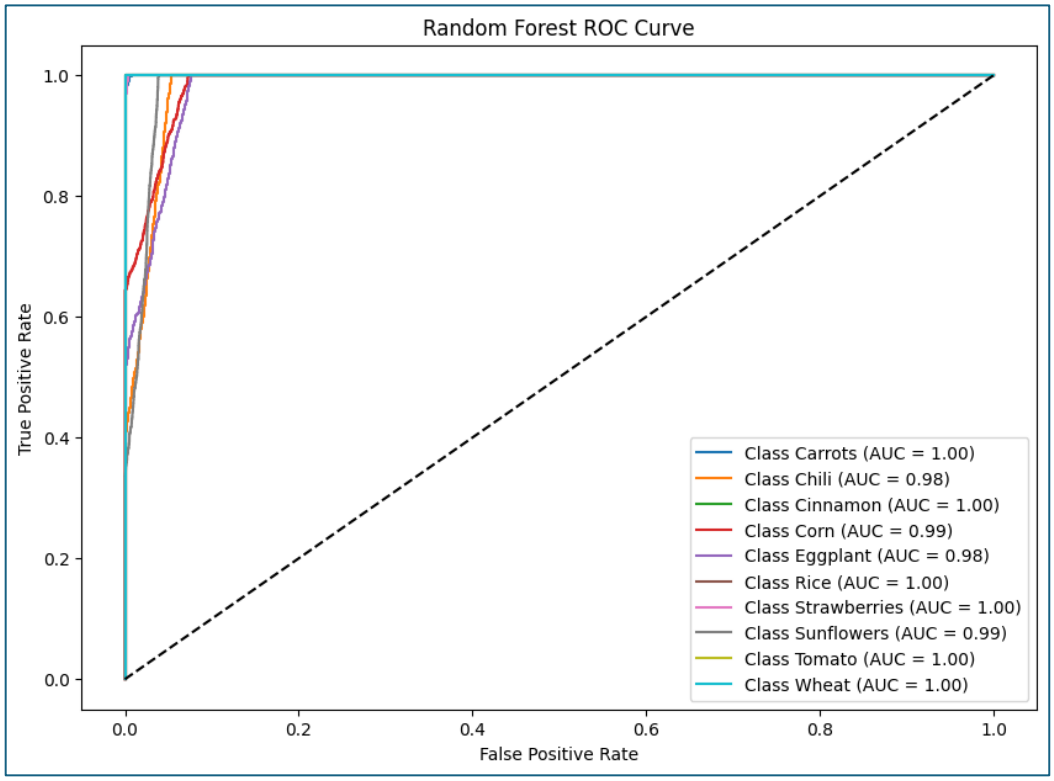

Chart 6 showing the ROC Curve for the Random Forest Model.

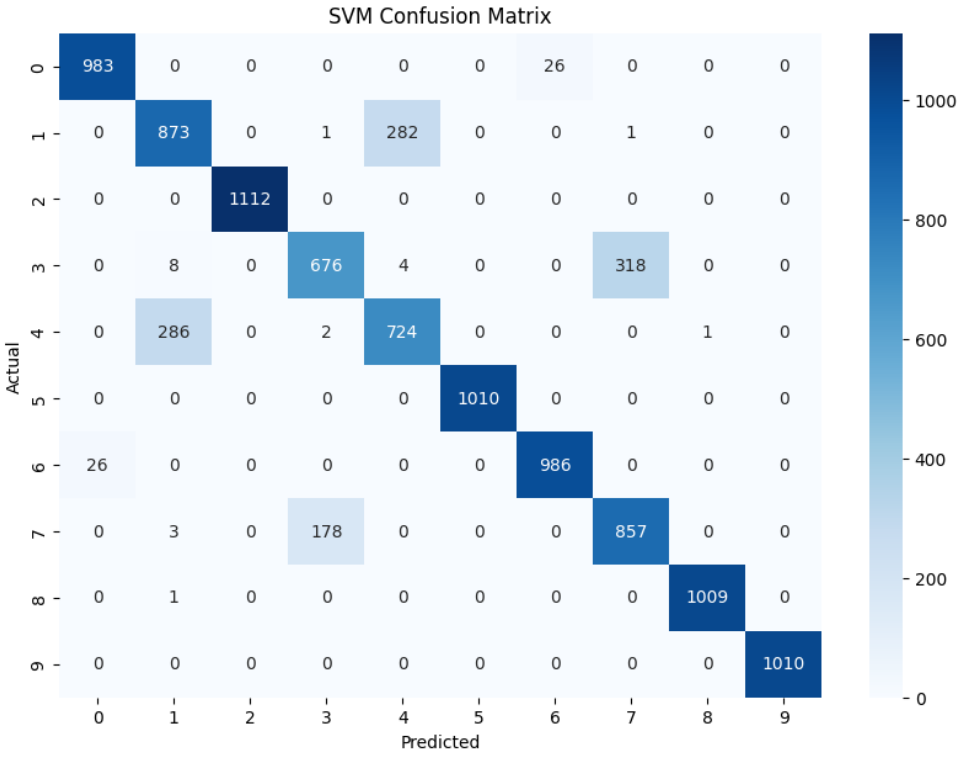

Chart 7 showing the Confusion Matrix for the Support Vector Machine Model.

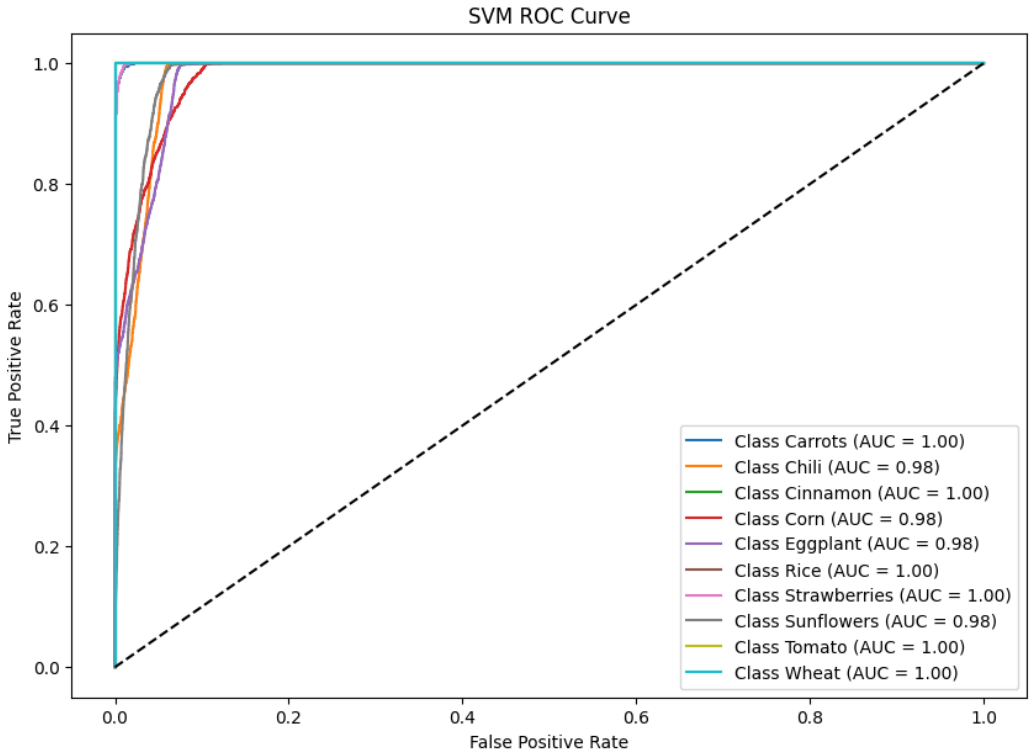

Chart 8 showing the ROC Curve for the Support Vector Machine Model.

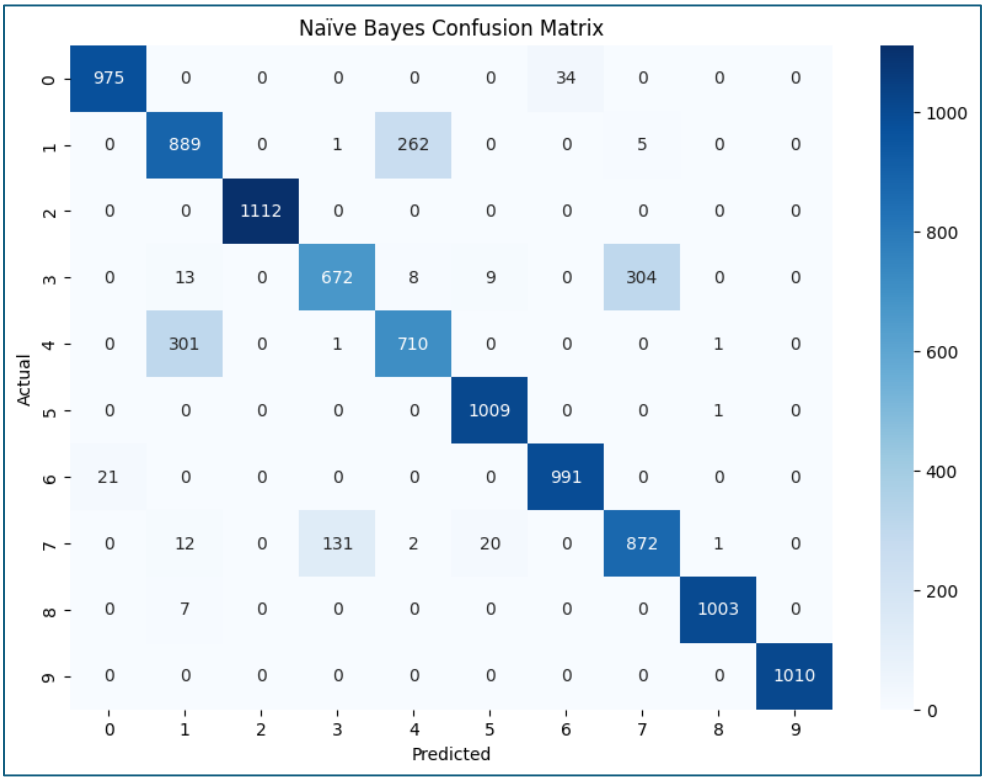

Chart 9 showing the Confusion Matrix for the Naive Bayes Model.

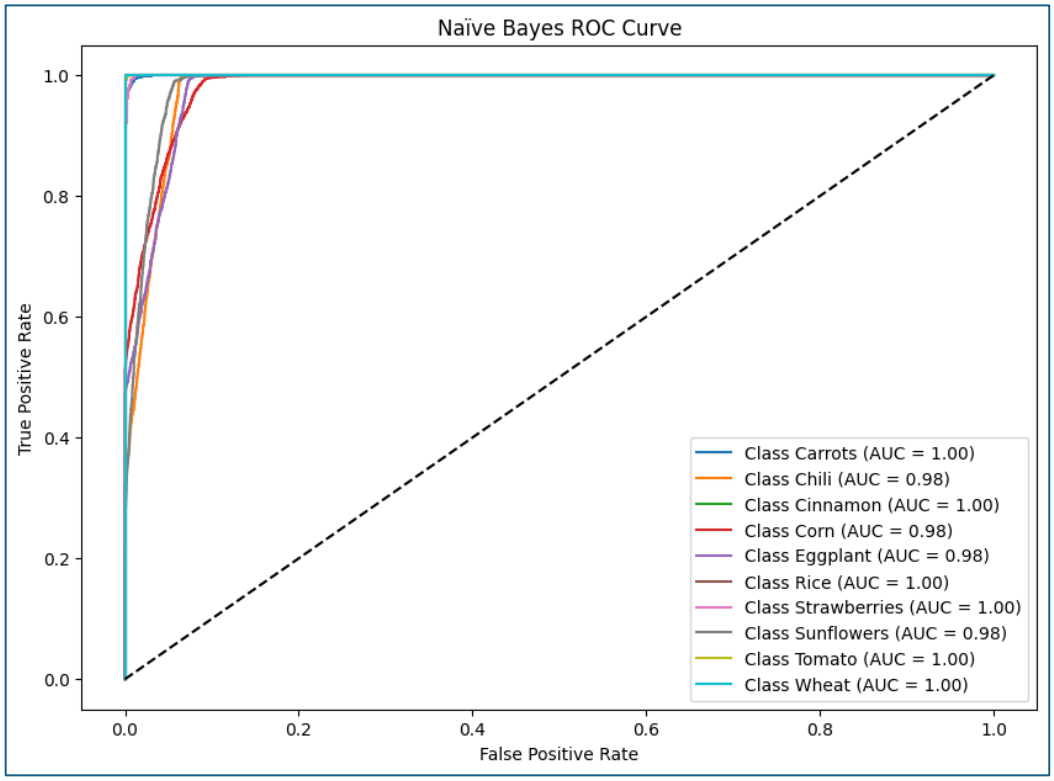

Chart 10 showing the ROC Curve for the Naïve Bayes Model.