International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Optimizing Crop Selection Using Machine Learning: Analysing Soil NPK and Climate Data for Precision Agriculture.

P Gopala Raju[1], M Sowmya[2], N Chakradhar[3], K Nithin Kumar[4]

[1, 2, 3, 4] GMR Institute of Technology, Rajam, India

ABSTRACT:

In the realm of precision agriculture, optimizing crop selection based on soil and climate conditions is crucial for enhancing productivity and sustainability. This study investigates the application of machine learning techniques to predict the most suitable crops for specific soil and climate conditions using a dataset comprising Nitrogen (N), Phosphorus (P), Potassium (K), Humidity, pH, Temperature, and Rainfall as features. We employed a variety of machine learning algorithms, including Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Decision Tree, Random Forest, Gradient Boosting, Bagging Classifier, Naïve Bayes, and AdaBoost to build predictive models. The primary objective was to determine the model that offers the highest accuracy and reliability in predicting crop types based on the given features. Initial results indicate that ensemble methods, particularly Random Forest and Gradient Boosting, outperform other models in terms of accuracy and robustness. By leveraging these machine learning models, the study demonstrates the potential for data-driven approaches to significantly contribute to precision agriculture. The findings can guide future research and development of more sophisticated tools for agricultural planning and resource management, ultimately leading to increased agricultural efficiency and sustainability.

Keywords: Precision Agriculture, Machine Learning, Crop Selection, Soil and Climate Data, Predictive Modelling.

Introduction:

Precision agriculture is a modern farming approach that uses technology to improve crop production. By collecting data about soil and climate, farmers can make better decisions about planting, watering, and fertilizing their crops. This method helps increase yields, reduce waste, and promote sustainable farming practices. The use of machine learning (ML) in agriculture is a key advancement, as it enables farmers to analyze complex data and make more accurate predictions about crop growth. Machine learning is a type of artificial intelligence that allows computers to learn from data and make predictions. In agriculture, ML can analyze soil nutrients and climate conditions to recommend the best crops for specific areas. This is especially useful because farming conditions can vary widely, even within a single field. This project, titled "Optimizing Crop Selection Using Machine Learning: Analyzing Soil NPK and Climate Data for Precision Agriculture," aims to use ML to help farmers choose the best crops for their land. The dataset used in this project includes important soil nutrients (Nitrogen, Phosphorus, and Potassium) and climate factors (Humidity, pH, Temperature, and Rainfall). These variables are crucial for determining which crops will thrive in a particular environment. The main goal of this project is to develop and evaluate ML models that can accurately predict the best crop to plant based on soil and climate data. The steps to achieve this include: Data Preprocessing and Exploration: Cleaning the data, handling any missing values, and exploring the relationships between variables. Model Training: Implementing and training various ML algorithms to predict crop types. Model Evaluation: Assessing the performance of each model to determine the best one. Hyperparameter Tuning: Optimizing the settings of the models to improve accuracy. Several ML algorithms will be used in this project to find the most effective one for predicting crops like Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Gradient Boosting, AdaBoost, Bagging Classifier, and Naïve Bayes. This project shows how ML can transform traditional farming by providing precise recommendations for crop selection. Accurate predictions help farmers make informed decisions, leading to better yields and more sustainable practices. By using resources more efficiently, farmers can reduce their environmental impact and increase profitability This project highlights the potential of combining technology with agriculture to create a more productive and sustainable future. As the global population grows and the demand for food increases, innovations like these are essential for meeting the challenges of modern agriculture and ensuring food security.

Methodology:

The methodology for this project involves several critical steps, including data preprocessing and exploration, model training, model evaluation, and hyperparameter tuning. Each step is essential to ensure the accuracy and effectiveness of the machine learning models used for crop prediction.

Data Preprocessing and Exploration

1. **Data Collection:** The dataset consists of 2200 samples, each with seven predictor variables (Nitrogen, Phosphorus, Potassium, Humidity, pH, Temperature, and Rainfall) and one target variable (crop name).
2. **Data Cleaning:** Handling missing values, outliers, and inconsistencies in the dataset. This involves removing or imputing missing values and ensuring data integrity.
3. **Feature Scaling:** Normalizing the numerical features to ensure that they have a consistent scale, which is particularly important for algorithms sensitive to feature magnitude (e.g., SVM, KNN).
4. **Exploratory Data Analysis (EDA):** Visualizing data distributions, identifying patterns and correlations, and understanding the relationships between variables. Techniques such as histograms, box plots, scatter plots, and correlation matrices are used in this step.
5. **Dimensionality Reduction:** If necessary, applying techniques like Principal Component Analysis (PCA) to reduce the number of features while retaining most of the data's variance, thus simplifying the model training process.

Machine Learning Algorithms

A diverse array of machine learning algorithms will be employed in this project to capture different aspects of the data and identify the most effective model for crop prediction:

1. Logistic Regression: A statistical method for binary classification problems, extended here to multiclass classification for predicting crop types.
2. Support Vector Machine (SVM): A powerful classifier that finds the hyperplane which best separates the data into different classes.
3. K-Nearest Neighbours (KNN): A simple, instance-based learning algorithm that classifies samples based on the majority class among their nearest neighbours.
4. Decision Tree: A model that splits the data into branches to make predictions based on decision rules inferred from the features.
5. Random Forest: An ensemble method that builds multiple decision trees and merges their predictions for improved accuracy and robustness.
6. Gradient Boosting: An advanced ensemble technique that builds models sequentially, each correcting the errors of its predecessor.
7. Bagging Classifier: Another ensemble method that combines the predictions of multiple models to reduce variance and avoid overfitting.
8. Naïve Bayes: A probabilistic classifier based on Bayes' theorem, assuming independence between features.
9. AdaBoost: An ensemble technique that combines weak classifiers to form a strong classifier, focusing on samples that are difficult to classify.

Results & Discussion:

Exploratory Data Analysis (EDA) Hist Plot

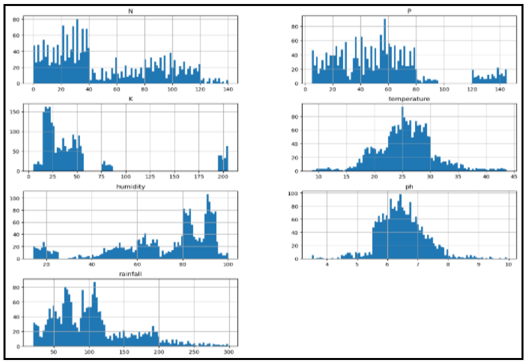

Figure 1 Hist Plot

Interpretations: Image show histograms for N, P, K, temperature, humidity, pH, and rainfall. Histograms are a way of visualizing the distribution of data. They show the number of data points that fall within a certain range of values (bins) on the x-axis. The y-axis shows the frequency or density of the data points.

The distribution patterns of these variables provide valuable insights into the soil and climate conditions present in the dataset. The bimodal distributions in potassium and humidity suggest that the dataset includes samples from diverse agricultural regions with varying soil fertility and climatic conditions. The normal distributions of temperature and pH indicate that these variables are relatively stable and within the optimal range for crop growth. These insights are crucial for developing accurate predictive models for crop selection. The variability in soil nutrients and climatic conditions must be accounted for to ensure that the model can generalize well to different regions. The presence of distinct peaks in certain variables highlights the need for robust data preprocessing and feature engineering to capture the underlying patterns effectively

Boxplots

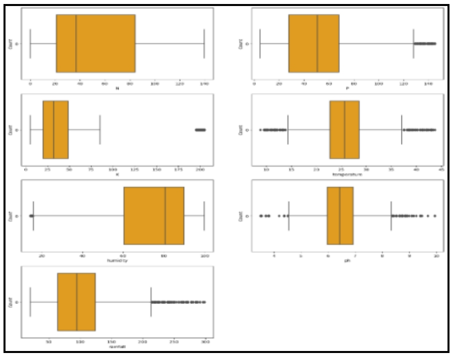

Figure 2 BoxPlots

Interpretation: The box plots reveal significant variability in soil and environmental conditions. Nitrogen (N) and phosphorus (P) have wide distributions with high outliers, while potassium (K) levels are mostly low with significant outliers. Temperature shows a consistent range with minimal outliers, and humidity values are generally high with a few extreme values. pH values are concentrated between 5 and 7, indicating varied soil acidity. Rainfall has a broad distribution with many high outliers, indicating diverse rainfall conditions.

Heat Map

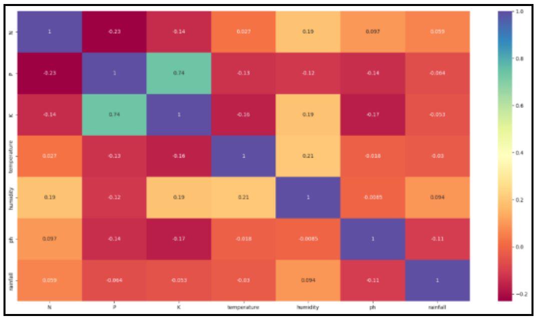

Figure 3 Heat Map

Interpretation:

The heatmap illustrates the correlation between various factors in the dataset. Phosphorus (P) and potassium (K) exhibit a strong positive correlation (0.74), indicating they often increase together. Nitrogen (N) shows weak negative correlations with phosphorus (-0.23) and potassium (-0.14). Temperature and humidity have a modest positive correlation (0.21). Other correlations are relatively weak, suggesting limited direct relationships among these variables. These insights highlight the interplay between soil nutrients and environmental factors, which is crucial for agricultural management.

Pair Plot

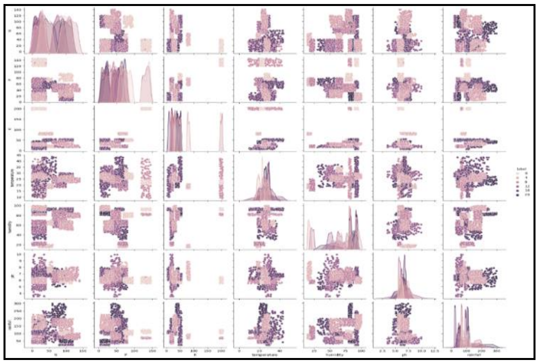

Figure 4 Pair Plot

The image is the output of a crop recommendation system, likely using machine learning. The system appears to be designed to analyze environmental factors to suggest ideal crops.

The left side of the image shows three graphs. The graphs plot temperature, humidity, and rainfall against an unspecified label. The right side of the image shows a table with four columns labeled "rainfall", "ph", "humidity", and "temperature". It appears that the table contains values from 0 to 300 for each of these factors. Without more information, it is difficult to say exactly what the results mean. However, we can make some general observations. The graphs show that the system is capable of considering multiple environmental factors at the same time. The table on the right likely shows ideal ranges for these factors for a particular crop. Crop recommendation systems are a type of precision agriculture technology. Precision agriculture uses technology to collect data about crops and their environment. This data can then be used to make decisions about how to manage crops more effectively. Crop recommendation systems can help farmers to improve crop yields, reduce water use, and minimize the use of pesticides and fertilizers. These systems are still under development, but they have the potential to revolutionize agriculture. By using machine learning to analyze complex data sets, crop recommendation systems can help farmers to make better decisions about their crops. This could lead to more sustainable and productive agriculture.

Confusion Matrix

Interpretation:

The confusion matrix demonstrates perfect model performance with 31 true negatives and 28 true positives, and no errors (false positives or false negatives). This indicates the model's 100% accuracy in correctly classifying both classes within the dataset, reflecting exceptional predictive capability.

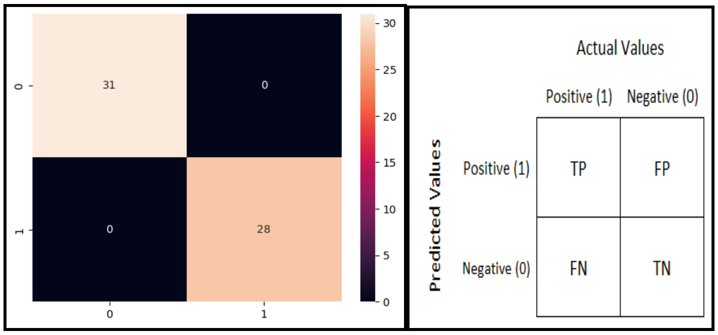

Figure 5 Confusion Matrix

Model Accuracy Comparisons

Comparing a classification algorithm's test and training accuracy reveals its ability to generalize. A high gap suggests overfitting, where the model memorizes training data but performs poorly on unseen data. Logistic Regression, SVM, and Naive Bayes are generally less prone to this, while Decision Trees and KNN can be more susceptible. Ensemble methods like Random Forest and Gradient Boosting can mitigate this by combining multiple models, often leading to better generalization performance.

Table 1 Accuracy Comparison

Model	Training Score	Testing Score
Logistic Regression	**95.25%**	**95%**
SVM	**99.1%**	**99%**
KNN	**100%**	**98.6%**
Decision tree	**100%**	**98.5%**
Random forest	**100%**	**99.5%**
Gradient Boost	**100%**	**99.39%**
Bagging classifier	**100%**	**99.24%**
Naïve Bayes	**96.1%**	**95.1%**

The table displays the training and testing accuracy scores for various classification models. Logistic Regression, SVM, and Naïve Bayes show high testing accuracies of 95%, 99%, and 95.1% respectively, with slight overfitting in SVM. Models like KNN, Decision Tree, Random Forest, Gradient Boost, and Bagging Classifier exhibit perfect training scores of 100%, indicating overfitting. However, their testing scores remain high, with Random Forest achieving the highest at 99.5%. Gradient Boost and Bagging Classifier follow closely with testing accuracies of 99.39% and 99.24%. The Decision Tree and KNN also show excellent testing performance at 98.5% and 98.6%, respectively. The results suggest that ensemble methods (Random Forest, Gradient Boost, Bagging Classifier) and SVM offer robust performance. Despite overfitting in training, these models generalize well to the test data, providing high testing accuracies. In contrast, simpler models like Logistic Regression and Naïve Bayes, though slightly lower in accuracy, show less overfitting

Conclusions:

- Advanced ensemble methods like Random Forest, Gradient Boost, and Bagging Classifier, along with SVM, achieve exceptional generalization, with Random Forest attaining 99.5% testing accuracy, and other algorithms falling from 95% to 99.5%.
- Logistic Regression and Naïve Bayes, while slightly less accurate, exhibit reduced overfitting, making them robust for simpler datasets where model interpretability is important.
- The dataset shows bimodal distributions in potassium and humidity, and normal distributions in temperature and pH, with significant outliers in nitrogen, phosphorus, and rainfall, necessitating robust data preprocessing.
- Strong positive correlation between phosphorus and potassium (0.74) and modest correlation between temperature and humidity (0.21) highlight interdependencies; the confusion matrix confirms the model's exceptional predictive accuracy.

REFERENCES:

1. Zhang, C., & Kovacs, J. M, The application of small unmanned aerial systems for precision agriculture: a review, 2012.
2. Khan, M. A., Rehman, S., Zaman, U. K., & Tahir, M., Predicting crop yields using SVM and Random Forests: A case study of wheat crops in Faisalabad, Pakistan, 2018.
3. Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L., & Mouazen, A. M., Wheat yield prediction using machine learning and advanced sensing techniques, 2016.
4. Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D., Machine learning in agriculture: A review, 2018.
5. Jeong, J., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., ... & Kim, S.-H., Random forests for global and regional crop yield predictions, 2016.
6. Probst, P., Wright, M. N., & Boulesteix, A.-L., Hyperparameters and tuning strategies for random forest, 2019
7. Bergstrom, J., & Bengio, Y., Random search for hyper-parameter optimization, 2012.