CSE 343: Machine Learning Project Proposal

Naman Jindal, naman22311@iiitd.ac.in Nishchay Sharma, nishchay22331@iiitd.ac.in

Harshil Handoo, harshil22206@iiitd.ac.in Himanshu Kumar, himanshu22215@iiitd.ac.in

AgriPredict: Machine Learning-Based Crop Yield Prediction

In an era where agricultural efficiency and sustainability are more critical than ever, accurate crop yield prediction has become a key factor in ensuring food security and optimizing resource management. **AgriPredict leverages machine learning, by analyzing a range of influential factors such as area, rainfall, wind, crop type, and soil condition,it aims to deliver precise and actionable insights, predicting the quantity of crop yield.** This empowers farmers and agricultural stakeholders to make informed decisions and maximize productivity.

1. Motivation

1.1. Problem Statement

Agriculture plays a crucial role in global food security, yet predicting crop yield remains challenging due to complex interactions between environmental factors. Traditional methods often rely on historical data and intuition, which may not fully capture modern agriculture's dynamic nature. Advanced, data-driven approaches are essential for improving crop yield predictions, especially in the face of climate change and resource constraints. This project aims to address this gap by applying machine learning techniques to enhance agricultural productivity and sustainability.

1.2. Motivation behind this project

The idea for AgriPredict was driven by the need to ensure food security amid climate challenges and population growth. We recognized that accurate crop yield prediction can significantly improve farming practices by enabling informed decision-making. Leveraging machine learning to analyze various factors affecting yield, our goal is to provide insights that help farmers optimize their practices and enhance productivity, contributing to a more secure and sustainable food supply.

2. Related Work

2.1. Crop Yield Prediction Using Machine Learning: A Pragmatic Approach

This study focuses on improving crop yield prediction in India's agricultural sector using machine learning techniques such as Random Forest, Adaboost, Gradient Boost, and SVM. It finds that Random Forest achieves the highest accuracy, helping farmers enhance their yields.

2.2. Progress in Research on Deep Learning-Based Crop Yield Prediction

This paper provides a comprehensive review of deep learning-based crop yield prediction methods, analyzing the advantages and challenges of various algorithms.

2.3. Yield Prediction for Crops by Gradient-Based Algorithms

This study evaluates machine learning algorithms for predicting crop yields, finding that CatBoost, LightGBM, and XGBoost outperform others.

3. Dataset: Dataset Details with Data Preprocessing Techniques:

In this project, we utilized a comprehensive dataset that includes records from various countries over several decades and multiple crop types. We collected data on pesticides and yield from the FAO, while rainfall and average temperature data came from the World Data Bank. After cleaning and merging the data, we created a final dataset with 28,242 rows and 7 columns.

- **Area**: The country or region where the data is recorded.
- **Item**: The type of crop (e.g., Maize, Potatoes, Wheat).
- **Year**: The year of observation.
- **hg/ha_yield**: Crop yield measured in hectograms per hectare.
- **average_rain_fall_mm_per_year**: The average annual rainfall measured in millimeters.
- **pesticides_tonnes**: The amount of pesticide use measured in tonnes.
- **avg_temp**: The average temperature recorded during the year.

3.1. Data Preprocessing

- **Filtering Countries**: Countries with fewer than 100 entries were removed to ensure that the dataset only includes regions with sufficient data for meaningful analysis. This filtering helps reduce noise and improve model training.

- **Normalization and Scaling**: Continuous variables such as `hg/ha_yield`, `average_rain_fall_mm_per_year`, and `pesticides_tonnes` were normalized using Min-Max scaling to ensure that each feature contributes equally during model training.

- **Missing Values**: We didn't find any missing values in the dataset, as each column contains 28,242 non-null entries.

3.2. Dataset Visualization

Heatmap: We created a heatmap and found a strong correlation between `Area` and both `pesticides` and `average_rain_fall_mm_per_year`. Similarly, there is a clear link between `Item` and `hg/ha_yield`, showing that crop type affects yield. **Histograms**: We created histograms and drew a few key inferences. Most rainfall is concentrated between 0-1000mm, with very few areas experiencing around 3000mm. For `pesticides_tonnes`, the majority of areas used little to no pesticides. The `avg_temp` histogram showed most temperatures clustering around 25°C. **Pair Plots**: Pair plots were employed to observe relationships between multiple numerical features. **Scatter Plots**: These plots visualized trends such as yield over the years for each crop type and provided insights into how yield changes with increasing temperature or rainfall. **Violin Plots**: These plots helped us visualize the distribution of yield across different countries, indicating variance in productivity and agricultural practices. **Bar Charts**: Apart from these, we created various bar charts to study pattern of yield over the years, some of them are as follows :

- **Yield for Different Crops over the Years**: Illustrated how crop yields for various items like wheat, rice, and maize have changed over time.

- **Average Annual Rainfall by Country**: Showed regional differences in rainfall and its potential impact on yields.

- **Average Annual Pesticide Usage by Country**: Helped identify countries with higher pesticide use and their corresponding crop yields, contributing to our analysis of pesticide impact.

The Million-Dollar Question: Does Pesticide Use Affect Yield?

We wanted to analyze whether there was any relation between pesticide use and the yield of an area. To explore this, we plotted various visualizations to examine the correlation between pesticide usage and crop yield.

1. We started by plotting the mean yield and mean pesticide usage of various countries.

3.3. Findings

Through our analysis, we discovered an intriguing pattern: while pesticide use initially increases yield production, there is a threshold beyond which yield begins to decline. Our findings suggest the following progression: as we use more pesticides, yield increases up to a maximum point, after which yield gradually decreases, eventually falling below the initial levels, regardless of further pesticide application.

Notably, we observed exceptions to this pattern in countries like France and Japan, which are among the leading agricultural producers. While it is possible to increase production by using more pesticides, this increase is temporary; in the long run, using fewer pesticides proves to be more sustainable.

4. Methodology and Model Details

The prediction task involves selecting suitable machine learning models, performing feature engineering, and evaluating the models to determine the best-performing approach for crop yield prediction.

4.1. Data Splitting

The dataset was split into a training set (70%) and a testing set (30%) to validate the model's performance on unseen data.

4.2. Model Selection

We trained and evaluated several models for predicting crop yields, achieving the following results:

- **Linear Regression**: We first explored the linear relationship between attributes, achieving an accuracy of 65.44%, an R^2 score of 0.65, and a mean squared error (MSE) of 2.35. This indicates moderate performance based on linear assumptions.

- **K-Nearest Neighbors (KNN)**: While this model captures local patterns, it produced a lower accuracy of 34.61%, an R^2 score of 0.35, and a high MSE of 4.44. This indicates challenges with larger datasets and noise sensitivity.

- **Decision Tree Regressor**: We used this model to split the dataset based on feature values, achieving an accuracy of 98.36%, an R^2 score of 0.98, and an MSE of 1.11. While effective, it is prone to overfitting without proper tuning.

- **Random Forest**: Utilizing an ensemble approach, we achieved the highest accuracy of 98.90%, with an R^2

score of 0.99 and a low MSE of 7.49. This demonstrates excellent predictive power through the aggregation of multiple decision trees.

- **Bagging Regressor**: This method improved accuracy by training multiple trees on random samples, achieving 98.91% accuracy, an R^2 score of 0.99, and an MSE of 7.38. It effectively averaged predictions to enhance accuracy.
- **Gradient Boosting**: We implemented this sequential technique, achieving an accuracy of 88.53%, an R^2 score of 0.89, and an MSE of 7.79. It effectively corrected errors by focusing on complex patterns in the data.
- **XGBoost**: This optimized implementation yielded an accuracy of 98.03%, an R^2 score of 0.98, and an MSE of 1.34. Its efficiency in learning intricate patterns made it a strong performer.

4.3. Feature Engineering

One-Hot Encoding: Categorical variables such as `Area` and `Item` were transformed using one-hot encoding to convert them into a numerical format suitable for machine learning models. The transformation ensures that each category is represented by a unique binary vector, enabling the model to process non-numeric data. The code snippet below demonstrates the encoding process.

5. Results and Analysis

The performance of various machine learning models was evaluated based on their accuracy, Mean Squared Error (MSE), and R-squared (R^2) score. The table below summarizes the results:

Model	Accuracy	MSE	R^2 Score
Linear Regression	0.6544	2.3483e+	0.6544
Random Forest	0.9890	7.4885e+	0.9890
Gradient Boost	0.8853	7.7930e+	0.8853
XGBoost	0.9803	1.3366e+	0.9803
KNN	0.3461	4.4431e+	0.3461
Decision Tree	0.9836	1.1121e+	0.9836
Bagging Regressor	0.9891	7.3779e+	0.9891

Table 1. Model performance comparison based on accuracy, MSE, and R^2 score.

5.1. Key Insights

- **Random Forest** and **Bagging Regressor** emerged as the top performers, indicating that ensemble methods can effectively model complex, non-linear interactions.
- **XGBoost** and **Gradient Boosting** performed well but with slightly higher error rates, potentially due to the need for further hyperparameter tuning.
- **Linear Regression** and **KNN** showed significantly lower accuracy, suggesting that these models struggle with the high dimensionality and complexity of the dataset.

5.2. Bias Variance Analysis

As training set increases, bias and variance both are decreasing.

6. Conclusion: Learnings, Work Left, and Contributions

AgriPredict demonstrates the potential of machine learning to transform the agricultural sector by offering accurate, data-driven insights into crop yields. By analyzing a diverse set of features across multiple regions, the model can aid farmers in making informed decisions that enhance productivity. Future work includes refining the model by integrating satellite imagery and other remote sensing data for even more accurate predictions.

6.1. Learnings

- We recognized the significance of data preprocessing, including handling missing values and normalizing features, to enhance model training.
- Visualizations helped us understand the relationships between environmental factors and crop yield, guiding feature selection.
- Utilizing ensemble models like Gradient Boosting and XGBoost enabled us to achieve high prediction accuracy by capturing complex data interactions.

6.2. Work Left

- Model Optimization: Further fine-tuning of the hyperparameters for Gradient Boosting and XGBoost to maximize prediction accuracy.
- Deployment: Building a web-based application to enable farmers to input their data and receive yield predictions.
- K-Fold Cross-Validation: Implement k-fold cross-validation in future iterations to ensure model robustness and mitigate overfitting.

7. Timeline

In Week 1, we reviewed the Kaggle dataset to understand its structure and set clear project goals while establishing the working environment with necessary tools and libraries. Weeks 2 and 3 were dedicated to Exploratory Data Analysis (EDA), where we identified patterns, trends, and correlations through comprehensive visualizations and cleaned the dataset by addressing missing values and outliers.

In Week 4, we analyzed the correlation between pesticide use and crop yield, yielding valuable insights. Weeks 5

and 6 focused on training various models, including KNN, Random Forest, XGBoost, Decision Tree, and Linear Regression. We evaluated model performance using metrics such as accuracy, MSE, and R^2 score.

In the upcoming weeks, we will concentrate on model evaluation and optimization, finalize model saving, and develop a user-friendly GUI interface. We will also make predictions, visualize results, and complete the project with comprehensive documentation and reporting.

8. Individual Contributions

- Harshil Handoo: Conducted data preprocessing, feature engineering, and implemented initial model training.
- Naman Jindal: Focused on data visualization, including heatmaps and scatter plots, and assisted with model evaluation.
- Nishchay Sharma: Handled dataset integration, conducted literature review, and worked on cross-validation techniques.
- Himanshu Kumar: Assisted in hyperparameter tuning, drafting detailed model descriptions, and writing the final report.

9. References

References to research papers:

1. https://www.researchgate.net/publication/381910719_Crop_Yield_Prediction_Using_Machine_Learning_A_Pragmatic_Approach
2. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0291928
3. https://www.sciencedirect.com/science/article/pii/S0168169920302301
4. https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset