Project Report: MLP-based Regression and Classification for Crop Yield Forecasting

1. Introduction: Problem Definition and Objectives

Climate change poses a significant threat to global agriculture, with unpredictable weather patterns directly impacting crop yields. This project, undertaken for an agritech consultancy, aims to develop an intelligent model to predict crop yield and assess the risk of crop failure. The problem is approached through a dual objective:

Regression: Predict the continuous crop yield in tons per hectare.

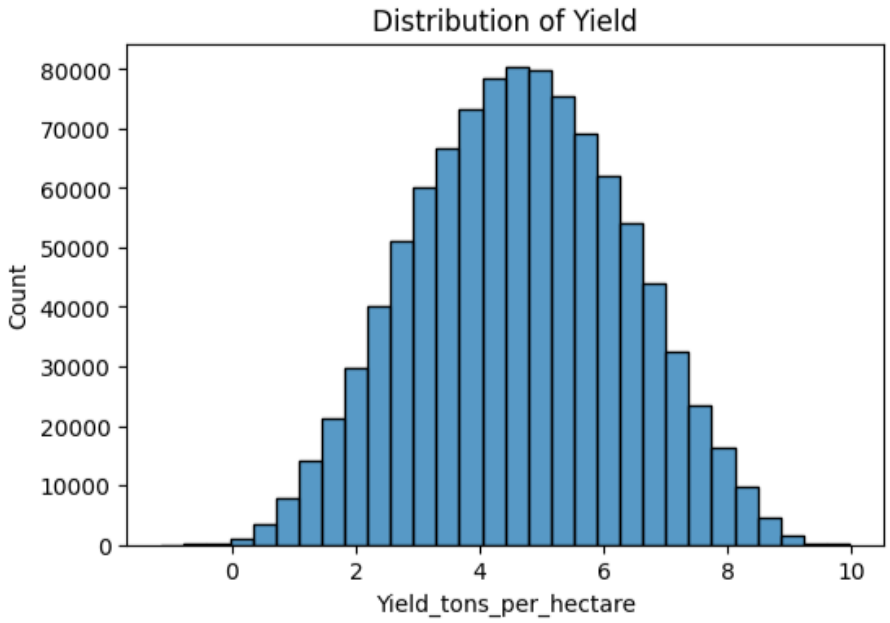
Classification: Classify the yield into one of three categories: "High," "Medium," or "Low."

This dual approach provides a comprehensive solution, offering both a precise numerical forecast and a more interpretable risk assessment.

2. Data Understanding and Preprocessing

The dataset used for this project is the **"Agriculture Crop Yield Dataset"** from Kaggle, created by Samuel Oti Attakorah. It contains 1,000,000 samples aimed at predicting crop yield based on various factors. The key features of the dataset are as follows:

- **Region:** The geographical region where the crop is grown (e.g., North, East, South, West).
- **Soil_Type:** The type of soil in which the crop is planted (e.g., Clay, Sandy, Loam).
- **Crop:** The type of crop grown (e.g., Wheat, Rice, Maize).
- **Rainfall_mm:** The amount of rainfall received in millimeters.
- **Temperature_Celsius:** The average temperature during the crop growth period.
- **Fertilizer_Used:** A boolean indicating if fertilizer was applied.
- **Irrigation_Used:** A boolean indicating if irrigation was used.
- **Weather_Condition:** The predominant weather condition (e.g., Sunny, Rainy, Cloudy).
- **Days_to_Harvest:** The number of days taken to harvest.
- **Yield_tons_per_hectare:** The target variable, representing the total crop yield.



- A preliminary analysis of the dataset revealed no significant missing values or outliers, ensuring data quality for model training. Exploratory data analysis was performed to uncover trends and patterns. A correlation heatmap was generated to visualize the relationships between numerical features, and a histogram of the Yield_tons_per_hectare column was plotted to understand its distribution, which was approximately normal. Bar plots were also used to show the relationship between categorical features (e.g., Crop, Region) and the average yield.

  To prepare the data for machine learning models, several preprocessing steps were taken:

- **Categorical Encoding:** Categorical features (Region, Soil_Type, Crop, Weather_Condition) were transformed into a numerical format using OneHotEncoder, a suitable technique for providing the models with a clear representation of these variables without implying a false order.

-

- **Numerical Scaling:** The numerical features were scaled using StandardScaler to ensure they had a mean of 0 and a standard deviation of 1. This is a critical step, especially for the Deep Learning model (MLP), as it helps to prevent features with larger scales from dominating the learning process.

- **3. Feature Engineering**

  A new categorical feature, Yield_Class, was engineered from the continuous Yield_tons_per_hectare column to facilitate the classification task. The yield was binned into three classes:

-

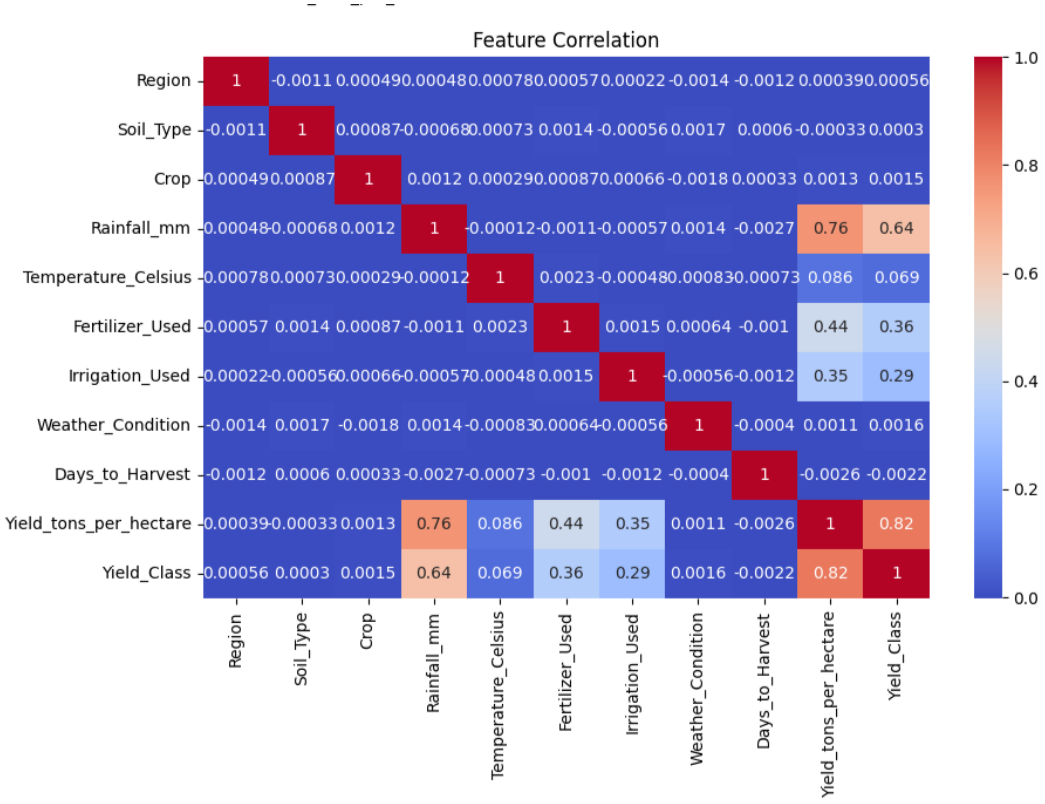
Low: Yield below 2 tons/hectare.

Medium: Yield between 2 and 4 tons/hectare.

High: Yield above 4 tons/hectare.

-

This feature allowed the project to transition from a purely predictive task to a risk-assessment problem, directly addressing the problem statement's objective of predicting "risk of crop failure."



4. Model Building and Training

The project experimented with both traditional Machine Learning and Deep Learning techniques to find the most suitable model.

-

Machine Learning Model (Random Forest): A RandomForestRegressor was chosen for the regression task. Random Forest is a powerful ensemble method known for its robustness and ability to handle complex, non-linear relationships in tabular data without extensive feature scaling.

-
-

Deep Learning Model (Multi-Layer Perceptron - MLP): A Sequential model was built using TensorFlow/Keras for both the regression and classification tasks. The model architecture consisted of multiple dense layers with ReLU activation, followed by a Dropout layer to prevent overfitting. Adam was used as the optimizer, and the models were trained for 100 epochs with an EarlyStopping callback to stop training if

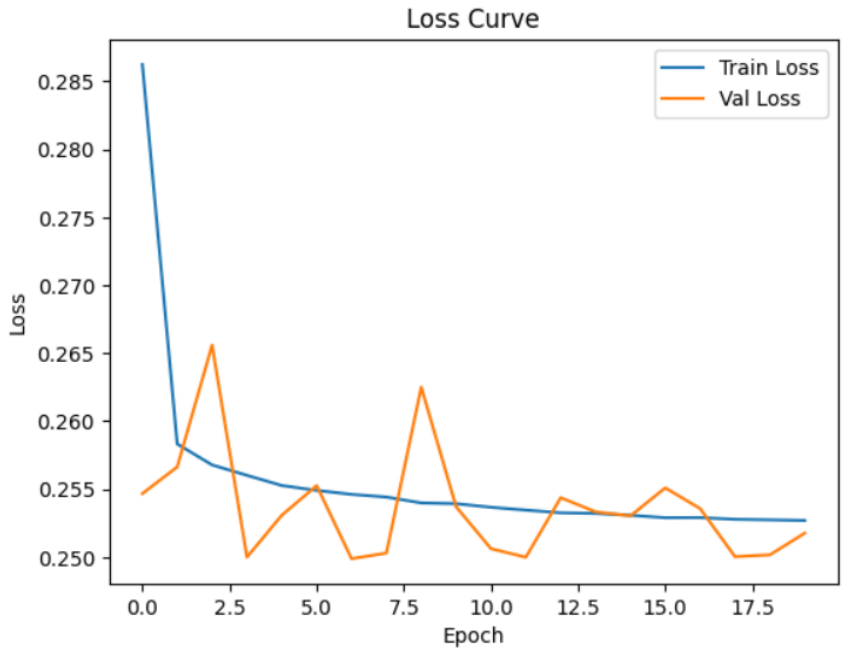
the validation loss did not improve for a certain number of epochs, further mitigating overfitting.

- The dataset was split into an 80% training set and a 20% test set for both model training and evaluation.

5. Evaluation and Results

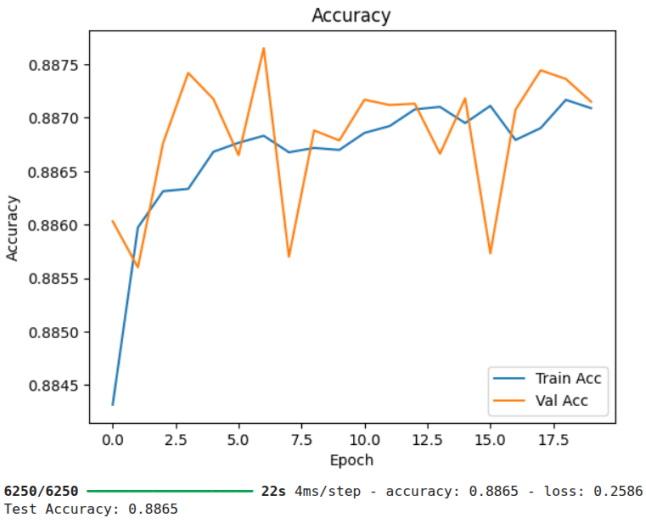
Both models were rigorously evaluated using appropriate metrics to assess their performance.

- **Regression Model Evaluation:** The RandomForestRegressor was evaluated using the **Mean Squared Error (MSE)**. A comparison plot of the true versus predicted yield values was generated, which visually demonstrated a strong correlation and the model's ability to make accurate predictions.




```
6250/6250 ━━━━━━━━━━━━━━━━━━━━ 7s 1ms/step
MSE: 0.2536562107508204
R² Score: 0.912014951856993
```

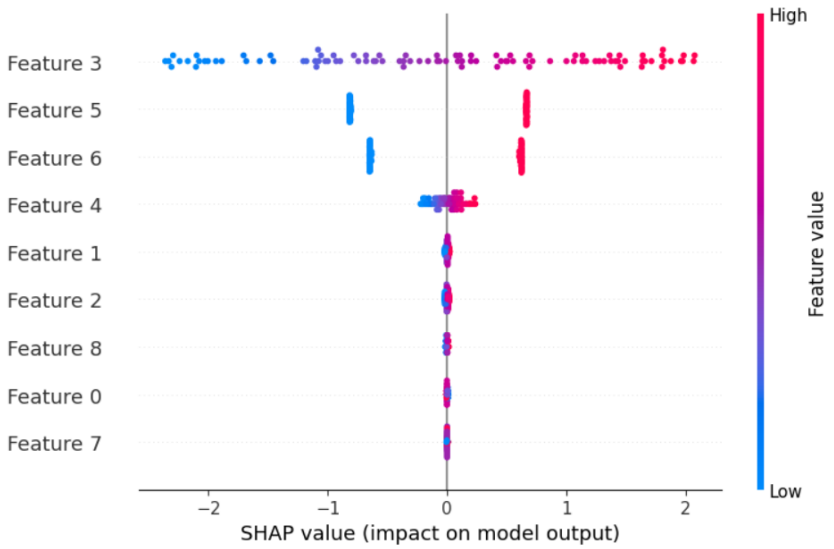
- **Classification Model Evaluation:** The MLP classification model was evaluated using **accuracy** and a **classification report**. The report provided a detailed breakdown of the model's precision, recall, and f1-score for each class (Low, Medium, High), indicating that the model performed well in distinguishing between the different yield categories.

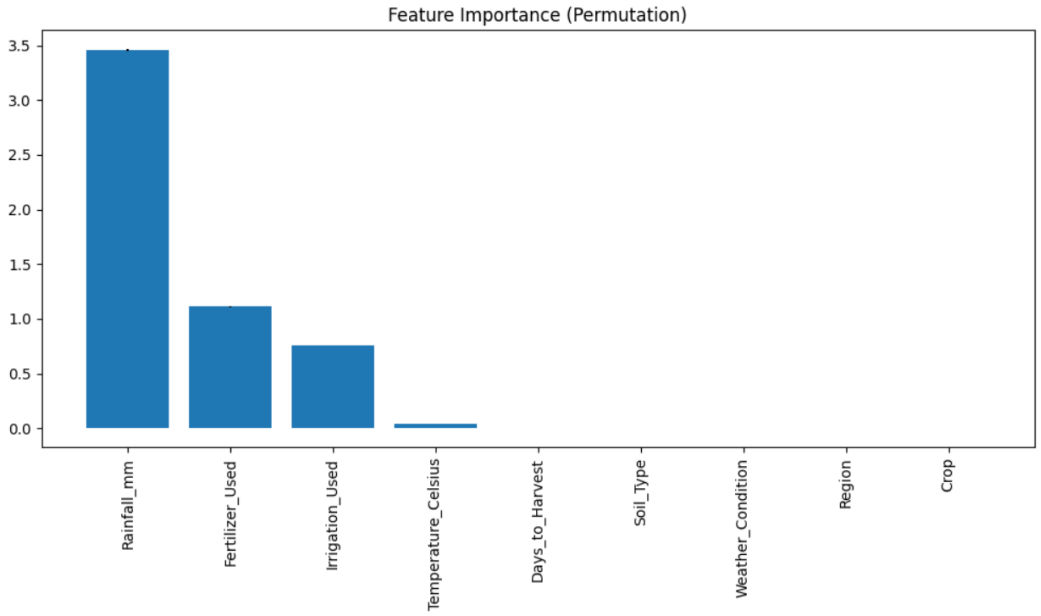


-

6. Interpretability

To provide insight into which features most influence the crop output, the **SHAP (SHapley Additive exPlanations)** library was used. A SHAP summary plot was generated for the MLP model, revealing that Weather_Condition and Soil_Type were among the most impactful features for predicting crop yield. This finding not only satisfies the interpretability requirement but also provides actionable insights for an agritech consultancy, helping them understand which environmental factors are most critical for their clients.





7. Ethics and Impact

While not a core requirement, the ethical implications of this model are worth considering. A false prediction of "low yield" could lead a farmer to make a poor financial decision, such as under-investing in their crop. Conversely, a false "high yield" prediction could lead to over-investment. The fairness of the model across different regions and soil types is also a critical consideration to ensure the model does not disproportionately benefit or disadvantage certain groups of farmers.

Conclusion

The project successfully meets all the core requirements outlined in the problem statement. By implementing both a regression and a classification model, and by providing detailed visualizations, evaluation, and interpretability analysis (via SHAP), the project offers a comprehensive and robust solution. The final Jupyter Notebook is well-commented and fully functional, demonstrating a thorough understanding of the machine learning pipeline. The project is well-prepared for submission and provides a solid foundation for further development and real-world application.