****

**Assessment Report**

on

**“Predict Crop Yield Category”**

submitted as partial fulfillment for the award of

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE(AIML)**

By

Name : Abhishek Kumar Singh

Roll Number : 202401100400009

Section: ‘A’

**Under the supervision of**

“BIKKI GUPTA”

**KIET Group of Institutions, Ghaziabad**

**April, 2025**

**1. Introduction**

**With the rise of precision agriculture, data-driven approaches have become essential for improving crop management and productivity. This project focuses on predicting crop yield categories using supervised machine learning. Using a dataset that includes soil quality, rainfall, and seed type, a classification model is developed to predict whether the crop yield will be low, medium, or high.**

**2. Problem Statement**

**To predict the category of crop yield (low, medium, or high) based on environmental and agricultural features. This prediction can help farmers and agronomists make informed decisions about resource allocation and crop planning.**

### ****3. Objectives****

● Preprocess the agricultural dataset for machine learning.  
● Train a Logistic Regression model to classify crop yield into categories.  
● Evaluate model performance using standard classification metrics.  
● Visualize the confusion matrix using a heatmap for better interpretability.



### ****4. Methodology****

● **Data Collection:** The dataset is uploaded as a CSV file (crop\_yield.csv).  
● **Data Preprocessing:**  
 ○ Missing numerical values are handled with mean imputation.  
 ○ Categorical variables (seed\_type) are encoded using one-hot encoding.  
 ○ Features are normalized using StandardScaler.  
● **Model Building:**  
 ○ The data is split into training (80%) and testing (20%) sets.  
 ○ Logistic Regression is used for multiclass classification.  
● **Model Evaluation:**  
 ○ Evaluation metrics include accuracy, precision, recall, and F1-score.  
 ○ A confusion matrix is visualized using a Seaborn heatmap.

### ****5. Data Preprocessing****

● Missing values in numerical features (soil\_quality, rainfall) are filled using mean values.  
● The categorical feature seed\_type is converted using one-hot encoding.  
● All features are scaled to ensure uniformity using StandardScaler.  
● Data is split into training and testing subsets in an 80:20 ratio.



**6. Model Implementation**

**Logistic Regression is used as it is a strong baseline classifier for multiclass problems. The model is trained on the processed dataset and used to predict the yield\_category.**

### ****7. Evaluation Metrics****

● **Accuracy:** Overall correctness of the model.  
● **Precision:** How many predicted yields in a class were actually correct.  
● **Recall:** How many actual yields in a class were correctly identified.  
● **F1 Score:** Balance between precision and recall.  
● **Confusion Matrix:** Helps assess model performance across each class (low, medium, high).

### ****8. Results and Analysis****

● The model showed acceptable performance on the test set.  
● The confusion matrix illustrated how well the model differentiated between yield categories.  
● Precision and recall metrics revealed strengths and weaknesses in predicting specific classes.

**9. Conclusion**

**The Logistic Regression model provides a foundational approach for crop yield prediction using agricultural features. While effective, future work could involve experimenting with more complex models like Random Forest or XGBoost, and addressing any class imbalance in the dataset.**

### ****10. References****

● scikit-learn documentation  
● pandas documentation  
● Seaborn visualization library  
● Research articles on crop yield prediction and agricultural analytics



 

 

