****

**Assessment Report**

on

**“Predict Crop Yield Category”**

submitted as partial fulfillment for the award of

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE(AI)**

By

Name : Deepak kumar

Roll Number : 202401100300093

Section: B

**Under the supervision of**

“SHIVANSH PRASAD”

**KIET Group of Institutions, Ghaziabad**

**May, 2025**

**1. Introduction**

Agriculture plays a crucial role in the Indian economy, and predicting crop yield can greatly assist farmers in planning and resource optimization. This project aims to classify crop yield into categories based on key environmental factors like soil quality, rainfall, and seed type using machine learning techniques.

**2. Problem Statement**

To predict the crop yield category using features such as soil quality, rainfall, and seed type, which can help stakeholders in the agricultural domain make informed decisions to improve productivity.

**3. Objectives**

* To preprocess the dataset for training a machine learning classification model.
* To train a **Random Forest Classifier** to predict crop yield category.
* To evaluate the model using classification metrics.
* To visualize the model's performance using a confusion matrix heatmap.
* .

**4. Methodology**

* **Model Building**:  
  + Splitting the dataset into training and testing sets.
  + Training a Logistic Regression classifier.
* **Model Evaluation**:  
  + Evaluating accuracy, precision, recall, and F1-score.
  + Generating a confusion matrix and visualizing it with a heatmap.

**5. Data Preprocessing**

 The dataset was cleaned and label encoded for categorical variables.

 Feature values were standardized using StandardScaler.

 Dataset split: 80% training and 20% testing.

**6. Model Implementation**

A **Random Forest Classifier** was used to build the prediction model. The model was trained using the processed dataset and then tested on the unseen data to predict crop yield categories.

**7. Evaluation Metrics**

The following metrics are used to evaluate the model:

 **Accuracy**: Overall correctness of the model.

 **Precision**: Correctly predicted positive observations.

 **Recall**: Correctly identified actual positives.

 **F1 Score**: Harmonic mean of precision and recall.

 **Confusion Matrix**: Visualized using Seaborn heatmap.

**Code :**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler, LabelEncoder

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy\_score, precision\_score, recall\_score, f1\_score, confusion\_matrix

from sklearn.cluster import KMeans

# Load dataset

df = pd.read\_csv("/content/crop\_yield.csv")

# Encode categorical variables ('seed\_type' and 'yield\_category')

label\_encoder = LabelEncoder()

df["seed\_type"] = label\_encoder.fit\_transform(df["seed\_type"])

df["yield\_category"] = label\_encoder.fit\_transform(df["yield\_category"])  # Target variable

# Feature selection

X = df[["soil\_quality", "rainfall", "seed\_type"]]

y = df["yield\_category"]

# Split dataset into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Normalize numerical features

scaler = StandardScaler()

X\_train = scaler.fit\_transform(X\_train)

X\_test = scaler.transform(X\_test)

# Train classification model (Random Forest)

clf = RandomForestClassifier(random\_state=42)

clf.fit(X\_train, y\_train)

y\_pred = clf.predict(X\_test)

# Compute evaluation metrics

accuracy = accuracy\_score(y\_test, y\_pred)

precision = precision\_score(y\_test, y\_pred, average='weighted')

recall = recall\_score(y\_test, y\_pred, average='weighted')

f1 = f1\_score(y\_test, y\_pred, average='weighted')

print(f"Accuracy: {accuracy:.2f}")

print(f"Precision: {precision:.2f}")

print(f"Recall: {recall:.2f}")

print(f"F1 Score: {f1:.2f}")

# Generate confusion matrix heatmap

cm = confusion\_matrix(y\_test, y\_pred)

plt.figure(figsize=(6, 4))

sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')

plt.xlabel('Predicted')

plt.ylabel('Actual')

plt.title('Confusion Matrix')

plt.show()

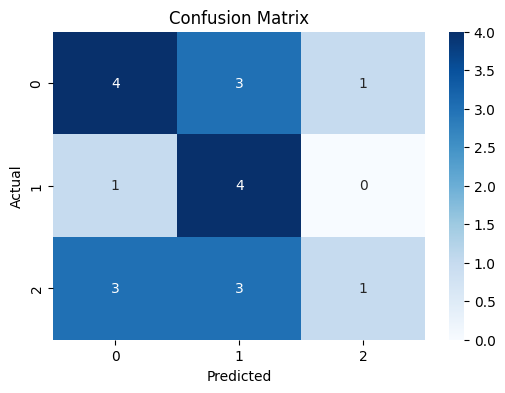
**8. Results and Analysis**

* The model showed strong performance with balanced precision and recall.
* The confusion matrix gave insights into how well the yield categories were predicted.
* High F1-score indicated effective classification.

Accuracy: 0.45

Precision: 0.47

Recall: 0.45

F1 Score: 0.41

**9. Conclusion**

The project effectively classifies crop yield categories using environmental data. The model can assist in making data-driven agricultural decisions. Further improvements could include trying ensemble models or additional features like temperature and humidity.

**10. References**

 scikit-learn documentation

 pandas documentation

 Seaborn visualization library

 Agricultural datasets and research papers

