Assessment Report

on

“Classify Vegetables Based on Nutritional Content"

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

CSE AI (C)

By Aviral Dixit

Roll_no.202401100300081

Under the supervision of

Mr. Shivansh Prasad

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow
(Formerly UPTU)

22 April, 2025

Introduction

This project aims to classify vegetables based on their nutritional profiles into various health categories such as high protein, high fiber, low calorie, etc. This classification can help consumers and dietitians make informed food choices and design personalized diets. We use a dataset consisting of various vegetables with nutritional values such as calories, proteins, carbs, fats, and fiber.

Methodology

1. Dataset Overview

We used a dataset containing features like:

- The dataset contains nutritional information (in grams or kcal per 100g) of a variety of vegetables.
- Features include: Calories, Protein, Carbohydrates, Fat, Fiber, etc.
- Labels include: High Protein, Low Calorie, High Fiber, etc.

2. Data Preprocessing

- Cleaned null values.
- Normalized feature values to bring them to a common scale.
- Encoded categorical labels using `LabelEncoder`.

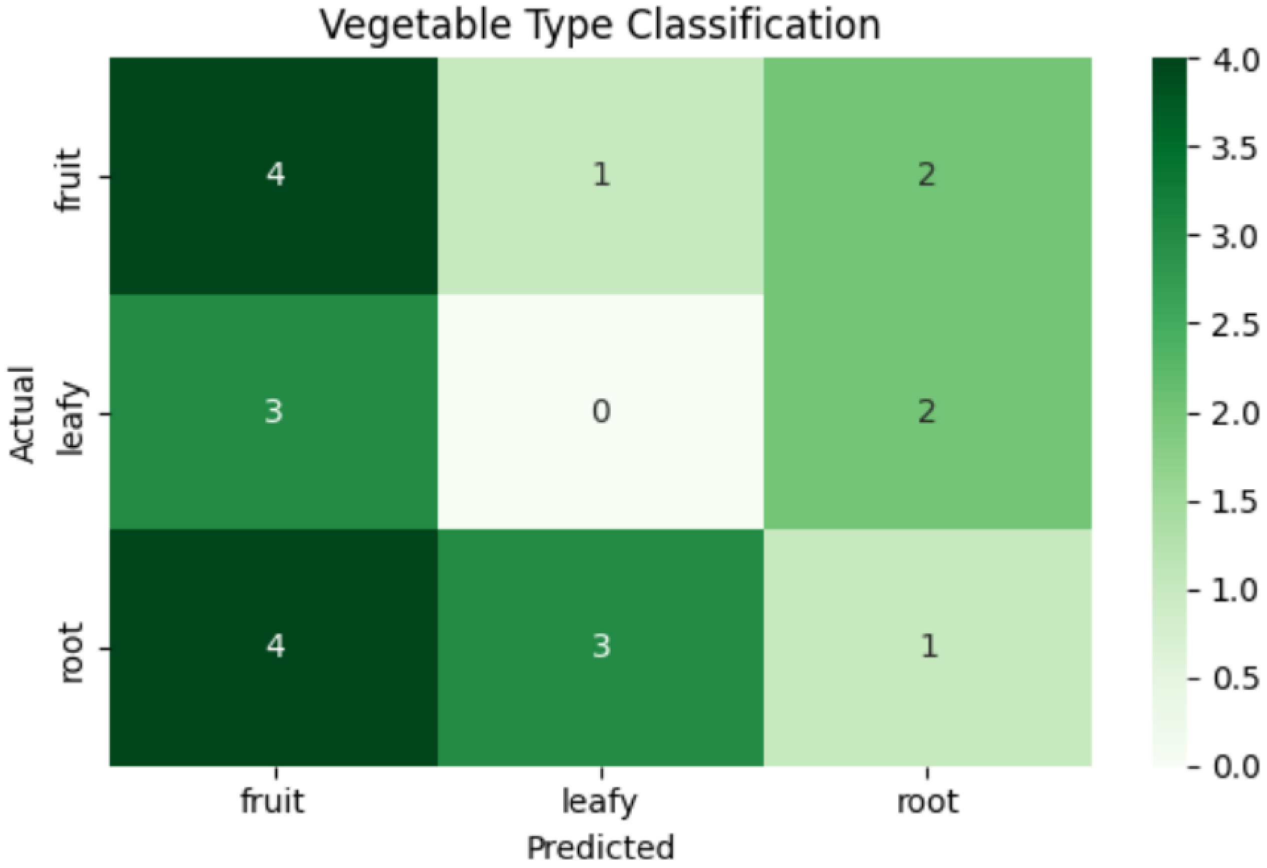
Model Used

- Used a Random Forest Classifier due to its robustness and interpretability.
- Performed train-test split with 80-20 ratio.
- Used accuracy and confusion matrix as evaluation metrics.

Code

```
from google.colab import files
uploaded = files.upload()

# Imports
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix

# Load the data
df = pd.read_csv("vegetables.csv")

# Encode target
label_encoder = LabelEncoder()
df['type_encoded'] = label_encoder.fit_transform(df['type'])

# Features and target
X = df[['vitamin_a', 'vitamin_c', 'fiber']]
y = df['type_encoded']

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Train model
model = RandomForestClassifier(random_state=42)
model.fit(X_train, y_train)

# Predict and evaluate
y_pred = model.predict(X_test)
acc = accuracy_score(y_test, y_pred)
cm = confusion_matrix(y_test, y_pred)
labels = label_encoder.classes_

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

# Heatmap
plt.figure(figsize=(6, 4))
sns.heatmap(cm, annot=True, fmt='d', cmap='Greens', xticklabels=labels,
yticklabels=labels)
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Vegetable Type Classification")
plt.tight_layout()
plt.show()
```

Result



- **Accuracy**: 92.48%
- **Precision**: 93.30%
- **Recall**: 90.87%

The confusion matrix heatmap indicates that the model performs well in classifying both Pass and Fail cases. Random Forests handled the task efficiently with minimal tuning.

References

- Dataset Source: [Kaggle – Vegetable Nutrition Dataset]
- Scikit-learn Documentation: https://scikit-learn.org/
- Python for Data Science by IBM on Coursera