Pumpkin Seed Classification System Using Machine Learning

Project Report

Submitted By:

Amit Suresh Powar

B.Tech (Computer Science & Engineering)

D Y Patil College of Engineering and Technology

1. Abstract

This project presents a Machine Learning-based web application for classifying pumpkin seeds into different classes using morphological features. The system utilizes supervised learning techniques to train a classification model and integrates it with a Flask-based web interface for real-time predictions. A Random Forest classifier is used for achieving high accuracy and robustness. The project demonstrates practical implementation of data preprocessing, model training, evaluation, and deployment in a web environment.

2. Introduction

Agricultural product classification plays a crucial role in food processing, quality control, and grading systems. Manual classification of seeds is time-consuming and prone to human error. This project automates the classification of pumpkin seeds using machine learning algorithms based on measurable physical attributes.

The objective of this project is:

To develop a machine learning model for pumpkin seed classification.

To deploy the model as a user-friendly web application.

To ensure accurate and fast prediction of seed class.

3. Problem Statement

To develop a machine learning-based system capable of accurately classifying pumpkin seeds using morphological features such as area, perimeter, axis lengths, eccentricity, solidity, etc., and deploy it as a web application for real-time predictions.

4. Dataset Description

The dataset used is:

Pumpkin Seeds Dataset (Excel format)

The dataset contains various morphological measurements of pumpkin seeds, including:

- Area
- Perimeter
- Major Axis Length
- Minor Axis Length
- Convex Area
- Equivalent Diameter
- Eccentricity
- Solidity
- Extent
- Roundness
- Aspect Ratio
- Compactness
- Class (Target Variable)

The target variable represents the category of pumpkin seed.

5. Technologies Used

Programming Language

- Python

Libraries

- NumPy
- Pandas
- Scikit-learn
- Pickle
- Flask

Frontend

- HTML
- CSS

Machine Learning Algorithm

- Random Forest Classifier

6. System Architecture

The system consists of the following modules:

1. Data Loading Module
2. Data Preprocessing Module
3. Model Training Module
4. Model Evaluation Module
5. Model Saving Module
6. Web Application Module

Workflow:

Dataset → Preprocessing → Train-Test Split → Scaling → Model Training → Evaluation → Model Saving → Flask Integration → User Input → Prediction

7. Methodology

7.1 Data Preprocessing

- Dataset loaded using Pandas.
- Missing values checked.
- Target variable encoded using LabelEncoder.
- Features separated from target.
- Dataset split into training (80%) and testing (20%) sets.
- Feature scaling performed using MinMaxScaler.

7.2 Model Training

- A Random Forest Classifier was used because:
- It reduces overfitting.
- It works well with structured/tabular data.
- It provides high accuracy.
- It handles nonlinear relationships effectively.

7.3 Model Evaluation

Evaluation metrics used:

- Accuracy Score
- Classification Report (Precision, Recall, F1-score)

The model achieved high classification accuracy, demonstrating effective performance.

7.4 Model Deployment

After training:

Model saved as **model.pkl**

Scaler saved as **scaler.pkl**

Label Encoder saved as **label_encoder.pkl**

A Flask web application loads these files and performs real-time predictions based on user input.

8. Working of the Web Application

Step 1: User Interface

The user enters seed feature values in the form.

Step 2: Data Processing

- Input values are converted to a DataFrame.
- Scaling is applied using the saved MinMaxScaler.

Step 3: Prediction

- The trained Random Forest model predicts the seed class.
- The predicted class is decoded using LabelEncoder.
- Result is displayed on the webpage.

9. Key Features of the System

- Accurate seed classification
- Real-time prediction
- User-friendly interface
- Scalable and extendable system
- Efficient preprocessing pipeline

10. Advantages

- Reduces manual effort
- High prediction accuracy
- Fast and automated
- Easily deployable
- Reusable model architecture

11. Limitations

- Depends on quality of dataset

- Requires proper feature measurements
- Limited to trained seed classes only

12. Future Enhancements

- Add more seed varieties
- Use Deep Learning models
- Deploy on cloud platform
- Add database integration
- Implement graphical analytics dashboard

13. Conclusion

The Pumpkin Seed Classification System successfully demonstrates the application of machine learning in agricultural product classification. The Random Forest model provides accurate and reliable predictions. Integration with Flask enables real-time classification through a web interface, making the system practical and user-friendly.

This project highlights the practical implementation of data science concepts including preprocessing, supervised learning, model evaluation, and deployment.