Harvesting Brilliance

A Taxonomic Tale of Pumpkin Seed Varieties

A Project Report

Submitted to Smart-Internz

for the completion of

Artificial Intelligence Internship Program

Prepared By

Mohseen Riyaj Attar

PRN: 2022011031004

D. Y. Patil Agriculture and Technical University, Talsande

Academic Year: 2025–2026

Submitted To

Smart-Internz

Abstract

Pumpkin seeds, despite being frequently overlooked, hold notable nutritional and agricultural significance. They display considerable variation across different cultivars, rendering their classification a vital task in agricultural research and biodiversity conservation. This project, titled “Harvesting Brilliance: A Taxonomic Tale of Pumpkin Seed Varieties,” centres on identifying and categorizing pumpkin seed types through machine learning methods.

A structured dataset comprising morphological traits of pumpkin seeds is gathered and prepared for examination. Exploratory Data Analysis (EDA) is conducted to comprehend data distributions and detect meaningful trends. Several machine learning classification models are trained and assessed using performance measures such as accuracy and classification reports. The top-performing model is chosen and further refined to enhance prediction dependability.

To ensure accessibility and ease of use, the trained model is deployed via an interactive web-based interface where users can enter seed measurements and receive instant predictions of the seed variety. This initiative illustrates the successful implementation of artificial intelligence in taxonomy and agriculture, aiding automated seed classification, advancing crop research, and enriching the understanding of pumpkin seed diversity. It also paves the way for future developments in smart agriculture, nutritional studies, and biodiversity preservation.

Index

1. Introduction

Agriculture is fundamental to food production and economic growth. With rising demand for high-quality crops and efficient farming methods, technology-driven approaches have become indispensable in contemporary agricultural systems. Among various farm products, pumpkin seeds are nutritionally and commercially valuable. Different pumpkin seed varieties possess distinct morphological traits, making their classification an essential activity for research, seed quality assurance, and crop enhancement.

Traditionally, seed classification is done manually by specialists relying on visual examination and expertise. However, this method is time-consuming, susceptible to human error, and inefficient for large datasets. Machine learning offers automated, precise, and rapid alternatives for seed variety recognition.

This project, “Harvesting Brilliance: A Taxonomic Tale of Pumpkin Seed Varieties,” applies machine learning algorithms to classify pumpkin seed types using morphological attributes such as area, perimeter, axis lengths, roundness, and other geometric properties. By analysing these features, the system predicts the pumpkin seed category with high accuracy.

The project showcases the merger of data science and agriculture, illustrating how artificial intelligence can foster smart farming, better seed analysis, and biodiversity protection. It also offers a user-friendly web interface where users can input seed measurements and obtain classification outcomes promptly, making the system viable for practical use.

1.1 Project Overview

This project builds an intelligent system to classify pumpkin seed varieties using numerical morphological data. The workflow includes data collection, cleaning, exploratory analysis, model training, evaluation, tuning, and deployment via a web app.

The end-to-end pipeline ensures not just model building but also a functional tool that bridges AI and agriculture.

2. Objectives

The main goal of the project “Harvesting Brilliance: A Taxonomic Tale of Pumpkin Seed Varieties” is to create an intelligent system capable of accurately classifying different pumpkin seed varieties using machine learning methods based on their morphological features.

Specific Objectives :

1. To gather and examine a dataset containing morphological attributes of pumpkin seeds.
2. To perform data pre-processing and exploratory data analysis to comprehend feature distributions and trends.
3. To implement various machine learning classification algorithms for predicting seed varieties.
4. To evaluate model performance using suitable metrics such as accuracy and classification reports.
5. To refine the best-performing model to improve prediction reliability.
6. To develop a user-friendly web-based interface for inputting seed measurements.
7. To deliver immediate and accurate predictions of pumpkin seed categories.
8. To illustrate the application of artificial intelligence in agricultural and taxonomic research.

3. Project Initialization and Planning Phase

3.1 Define Problem Statement

Pumpkin seeds are extensively utilized in food, nutrition, and agriculture due to their rich nutritional profile and economic worth. Different pumpkin seed varieties exhibit unique morphological characteristics, which are crucial for seed quality evaluation, breeding initiatives, and biodiversity investigations. However, manually identifying and classifying these seeds demands expert knowledge, is lengthy, and may yield inconsistencies owing to human error.

With increasing availability of agricultural datasets and progress in artificial intelligence, there is a necessity for an automated and precise system that can classify pumpkin seed varieties based on measurable morphological attributes. Conventional classification approaches are not scalable for large datasets and lack consistency in outcomes.

Hence, the problem tackled in this project is:

To design and develop a machine learning-based system that automatically classifies pumpkin seed varieties using their morphological characteristics and provides instant prediction results through a user-friendly interface.

This system intends to lessen manual lab or, enhance classification accuracy, and aid agricultural research and smart farming practices.

3.2 Project Proposal (Proposed Solution)

To address the challenges of manual classification, this project proposes an automated machine learning-driven classification system. The solution involves analysing key morphological features of pumpkin seeds—such as area, perimeter, axis lengths, roundness, and compactness—to predict seed varieties using trained classification models.

The system workflow includes:

- Sourcing a structured dataset of seed measurements.
- Pre-processing data to ensure quality and consistency.
- Training and evaluating multiple classification algorithms.
- Selecting and optimizing the best-performing model.

- Deploying the model via a web application for real-time predictions.

This approach ensures a fast, accurate, and user-friendly solution that minimizes reliance on manual inspection and supports agricultural automation.

3.3 Initial Project Planning

A systematic project plan was formulated to guide development and ensure timely completion. The planning phase involved defining the project architecture, selecting appropriate tools and technologies, identifying resource requirements, and establishing a phased timeline.

Key activities included:

- Outlining major stages: data collection, pre-processing, model development, evaluation, optimization, and deployment.
- Selecting software tools: Python, Pandas, Scikit-learn, Flask, and visualization libraries.
- Assessing hardware and system requirements for efficient data processing and model training.
- Creating a timeline with milestones for dataset acquisition, analysis, model training, interface development, testing, and final deployment.
- Identifying potential risks such as data quality issues and model overfitting, along with contingency strategies.

This structured planning provided a clear roadmap, ensuring organized progression from conception to implementation.

4. Data Collection and Pre-processing Phase

4.1 Data Collection Plan and Raw Data Sources Identified

Data collection is a critical step in constructing an effective machine learning model. For this project, an open-source dataset containing morphological characteristics of pumpkin seeds was used. The dataset is available in CSV format and includes numerical feature measurements essential for classification.

The dataset was sourced from a publicly accessible online data repository. Specifically, it was downloaded from Kaggle, a recognized platform that provides open datasets for research and educational purposes.

Dataset Source: Kaggle Dataset – Pumpkin Seed Classification

After downloading, the dataset was imported into the development environment for further analysis. The raw data contains various seed measurements such as area, perimeter, axis lengths, roundness, compactness, and other shape-based attributes.

Once collected, the next step involved reading and understanding the data using data visualization and analytical techniques. These methods assisted in identifying data distribution patterns, comprehending feature relationships, and preparing the data for pre-processing and model training.

4.2 Data Quality Report

Before applying machine learning algorithms, it is important to inspect the quality of the collected dataset. A data quality report helps identify issues like missing values, duplicate records, inconsistent data types, and outliers that could impact model performance.

In this project, the pumpkin seed dataset was thoroughly examined after loading. The dataset consists of numerical morphological features representing different physical traits of pumpkin seeds. Initial data exploration was performed to check dataset structure, record count, attribute count, and data types of each feature.

The dataset was inspected for missing or null values to ensure completeness. No significant missing values were found, indicating the dataset was well-structured

and ready for further processing. Duplicate records were also checked to avoid biased training outcomes, and none were present.

Basic statistical analysis, such as mean, minimum, maximum, and standard deviation, was conducted to understand the range and distribution of each feature. Visualization techniques like histograms and box plots were used to detect outliers and grasp data spread. Minor variations in feature values were observed, which is natural in real-world biological data.

Overall, the dataset was clean, consistent, and appropriate for building machine learning classification models. Only standard pre-processing techniques such as normalization and feature scaling were needed before model training.

4.3 Data Pre-processing

Data pre-processing transforms raw data into a format suitable for machine learning. The following steps were applied:

1. **Data Loading:** The dataset was loaded using Pandas.
2. **Handling Missing Values:** Since no major missing data existed, imputation was not required.
3. **Duplicate Removal:** Duplicate entries were checked and removed to prevent skewed learning.
4. **Feature-Target Separation:** The target column (seed variety) was separated from input features.
5. **Feature Scaling:** Normalization and standardization were applied to bring all numerical features to a common scale, improving model convergence and accuracy.
6. **Train-Test Split:** The data was divided into training and testing sets to evaluate model performance on unseen data.

These pre-processing steps ensured a refined and structured dataset ready for effective model training and evaluation.

5. Model Development Phase

5.1 Model Selection Report

After pre-processing the dataset, the next step was to select suitable machine learning algorithms for classifying pumpkin seed varieties. Since the problem involves predicting discrete seed categories based on numerical features, supervised classification techniques were deemed appropriate.

Multiple classification algorithms were explored and tested to determine the best model for accurate seed variety prediction. Model selection was based on their ability to handle numerical data, interpret feature relationships, and deliver high classification accuracy.

The following machine learning algorithms were considered:

- Logistic Regression
- K-Nearest Neighbours (KNN)
- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Machine (SVM)

Each model was trained using the training dataset and evaluated on the testing dataset. Performance metrics such as accuracy score, confusion matrix, and classification report were used to compare model results.

Initial experiments indicated that ensemble-based models like Random Forest and tree-based classifiers performed better due to their ability to handle non-linear feature relationships effectively. Support Vector Machine also performed well in distinguishing between seed categories.

Based on comparative analysis of accuracy and evaluation metrics, the best-performing model was shortlisted for further training and optimization in the next phase. This systematic model selection process ensured the final chosen model was both reliable and efficient for pumpkin seed variety classification.

5.2 Initial Model Training, Validation and Evaluation Report

After selecting suitable machine learning algorithms, the next phase involved training, validating, and evaluating the models using the pre-processed pumpkin seed dataset. This phase aimed to assess how well each model learned from the training data and how accurately it predicted seed varieties on unseen test data.

The dataset was split into training and testing sets. The training set was used to train the selected classification models, while the testing set was reserved for evaluation. Each model was trained using the training data, and predictions were generated for the testing data.

To validate model effectiveness, standard evaluation techniques were applied. The primary performance metric was classification accuracy, measuring the percentage of correctly predicted seed varieties. Additionally, confusion matrices and classification reports were generated to analyse precision, recall, and F1-score for each seed category.

During initial training, different models exhibited varying performance levels. Tree-based models like Decision Tree and Random Forest demonstrated strong predictive capability due to their ability to capture complex relationships among morphological features. Support Vector Machine also achieved good classification results with clear decision boundaries. K-Nearest Neighbours performed reasonably well but required careful selection of the K value. Logistic Regression provided baseline performance for comparison.

Based on evaluation results, Random Forest Classifier attained the highest accuracy among tested models and displayed stable and consistent prediction outcomes. Therefore, it was chosen for further optimization in the next phase. This training and evaluation phase confirmed that machine learning algorithms can effectively classify pumpkin seed varieties using morphological feature data.

6. Model Optimization and Tuning Phase

6.1 Tuning Documentation

After selecting the best-performing model in the initial training phase, the next step was to optimize the model to achieve better accuracy and reliable predictions. Model tuning is an important process that involves adjusting model parameters to improve performance and reduce overfitting or under fitting.

In this project, the Random Forest Classifier was selected for optimization due to its high initial accuracy and stable performance. Hyper parameter tuning was performed to find the optimal combination of parameters such as the number of trees (n_estimators), maximum tree depth (max_depth), minimum samples required to split a node, and minimum samples required at leaf nodes.

Techniques like Grid Search and cross-validation were used to systematically test different parameter values. Cross-validation ensured model performance was consistent across different dataset subsets and not dependent on a single train-test split.

Through repeated experiments and evaluation, the best hyper parameter configuration was selected, which improved classification accuracy and reduced prediction errors. The tuned model demonstrated better generalization capability when tested on unseen data.

This tuning process helped achieve a more accurate and robust machine learning model for pumpkin seed variety classification, making the system reliable for real-world use.

6.2 Final Model Selection Justification

After conducting initial training, evaluation, and hyper parameter tuning, the Random Forest Classifier was chosen as the final model for pumpkin seed variety classification. This decision was based on its superior performance compared to other tested algorithms.

The Random Forest model achieved the highest classification accuracy and showed consistent results across training and testing datasets. Its ensemble-based approach effectively handled complex and non-linear relationships among morphological features such as area, perimeter, axis lengths, and shape

descriptors. Additionally, the model exhibited strong resistance to overfitting due to the averaging effect of multiple decision trees.

Compared to single classifiers like Decision Tree or K-Nearest Neighbours, Random Forest provided better generalization and stable predictions. It also required minimal feature engineering and performed well on numerical datasets with multiple correlated features.

Therefore, the Random Forest Classifier was finalized as the most suitable and reliable model for this project. Its accuracy, robustness, and efficiency make it ideal for automated pumpkin seed variety classification.

7. Results

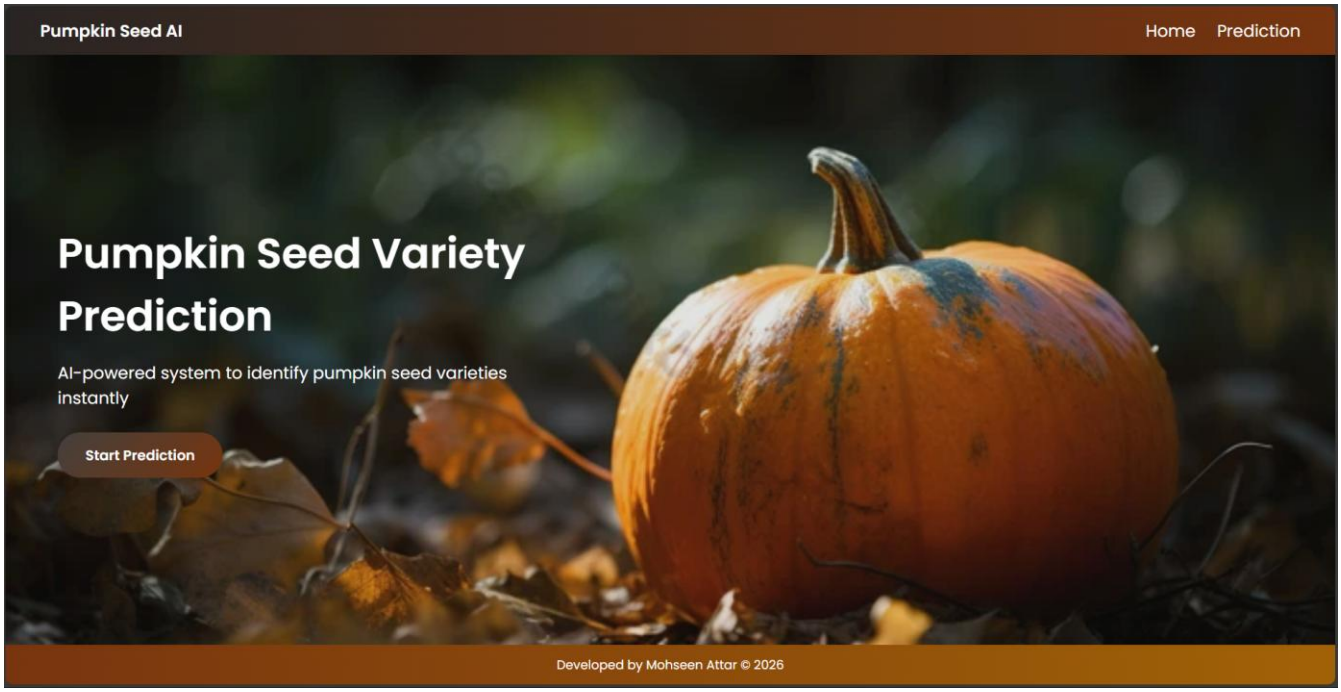
The optimized Random Forest model was integrated into a Flask-based web application. The interface allows users to input seed measurements and receive instant variety predictions.

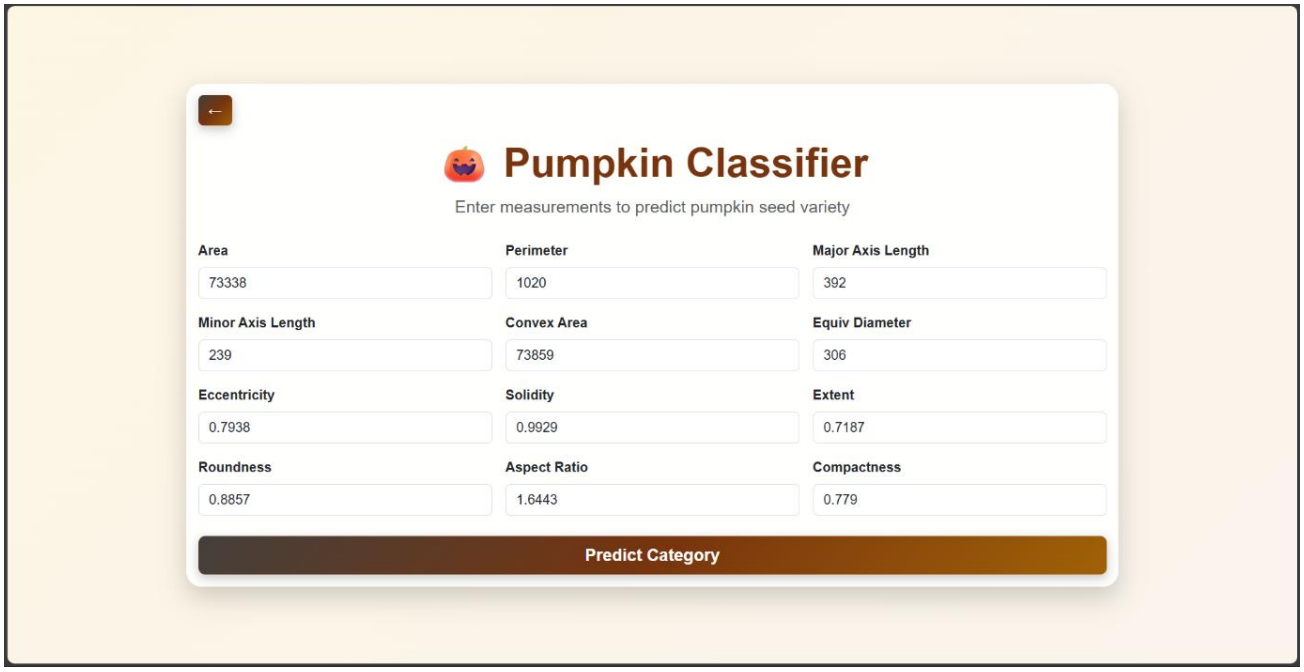
Key Outcomes:

- The model achieved high accuracy on test data.
- Predictions were consistent and reliable across varied inputs.
- The web app delivered real-time results with a clean, intuitive interface.

7.1 Output Screenshots

Figure – Pumpkin Seed Classifier Input Interface




←
Pumpkin Classifier
Enter measurements to predict pumpkin seed variety
Area
73338
Perimeter
1020
Major Axis Length
392
Minor Axis Length
239
Convex Area
73859
Equiv Diameter
306
Eccentricity
0.7938
Solidity
0.9929
Extent
0.7187
Roundness
0.8857
Aspect Ratio
1.6443
Compactness
0.779
Predict Category

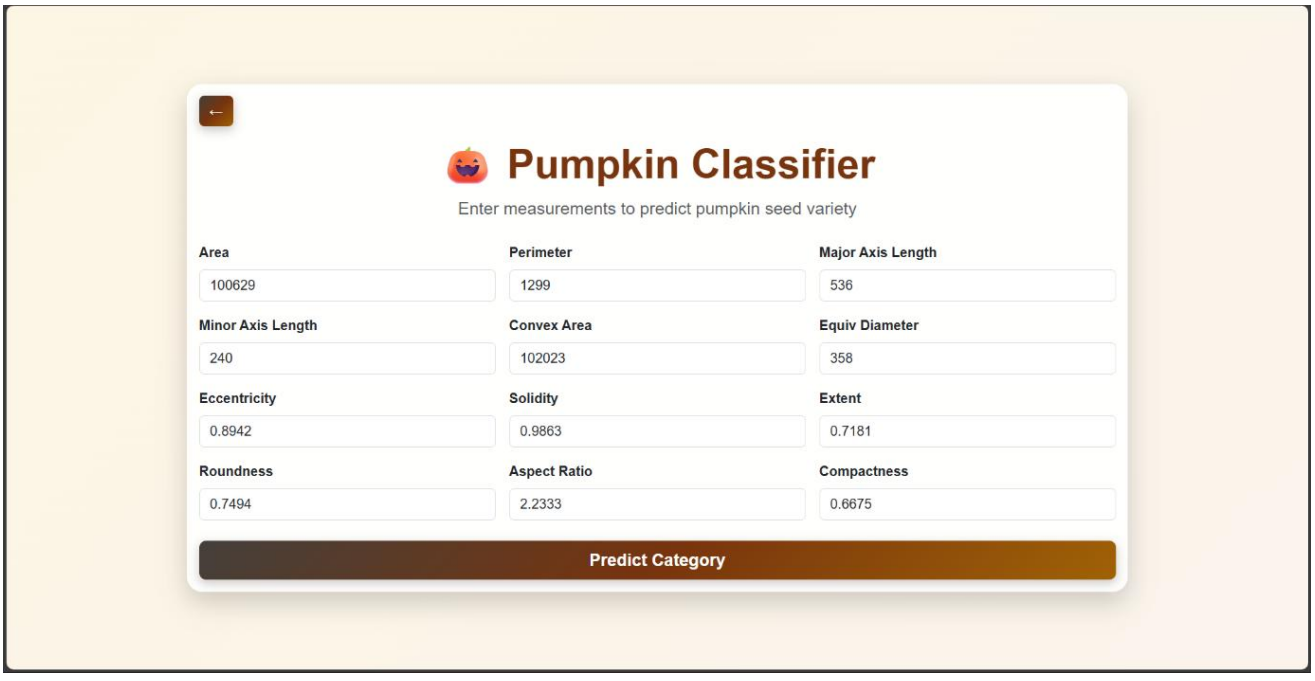
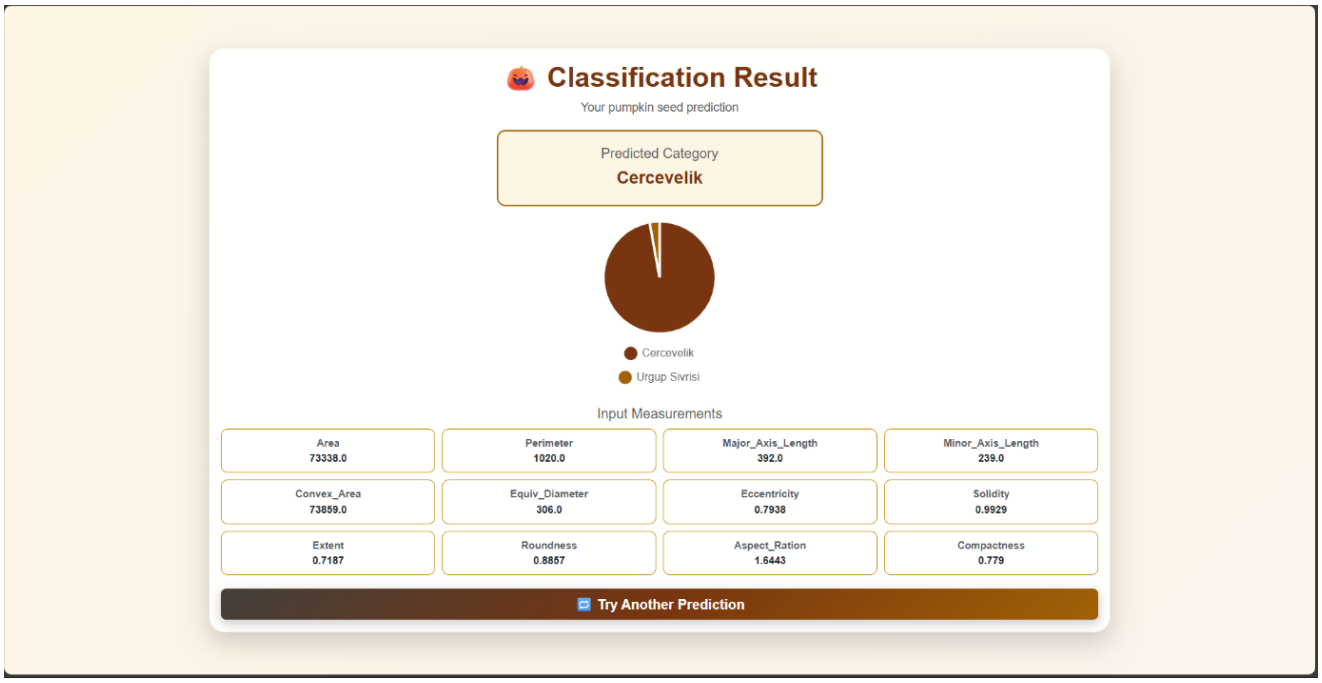
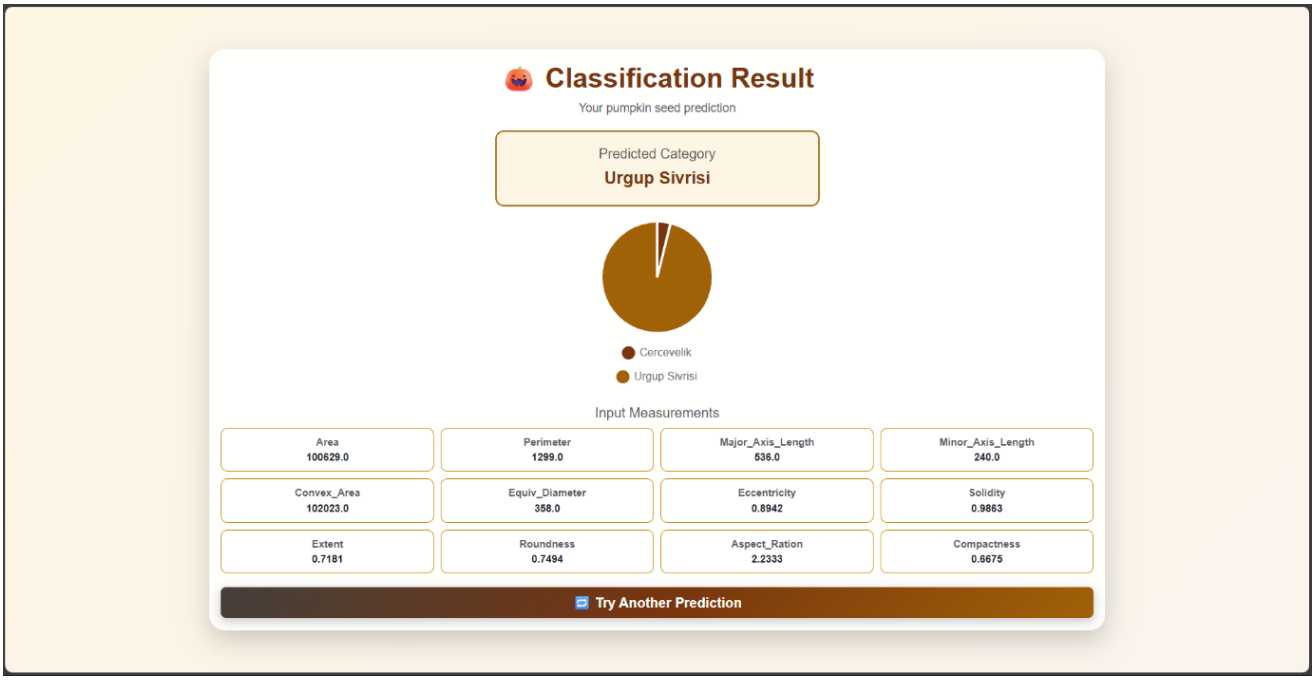

←
Pumpkin Classifier
Enter measurements to predict pumpkin seed variety
Area
100629
Perimeter
1299
Major Axis Length
536
Minor Axis Length
240
Convex Area
102023
Equiv Diameter
358
Eccentricity
0.8942
Solidity
0.9863
Extent
0.7181
Roundness
0.7494
Aspect Ratio
2.2333
Compactness
0.6675
Predict Category

Figure – Prediction Result Output





8. Advantages and Disadvantages

8.1 Advantages

1. Delivers automated and accurate classification of pumpkin seed varieties.
2. Decreases reliance on manual inspection and expert knowledge.
3. Enhances speed and efficiency in seed analysis.
4. Machine learning model provides consistent and reliable predictions.
5. User-friendly web interface makes the system easy to operate.
6. Supports agricultural research and seed quality evaluation.
7. Aids in understanding biodiversity and seed variety distribution.

8.2 Disadvantages

1. System performance depends on dataset quality and size.
2. Limited to seed varieties included in the training dataset.
3. Requires accurate numerical input values for correct prediction.
4. Model retraining is necessary when new seed varieties are introduced.
5. Performance may vary if real-world data distribution changes.

9. Conclusion

The project “Harvesting Brilliance: A Taxonomic Tale of Pumpkin Seed Varieties” successfully demonstrates the application of machine learning in agricultural and taxonomic research. The system was designed to classify pumpkin seed varieties based on morphological characteristics using supervised learning techniques.

A structured dataset was collected, pre-processed, and analysed to understand feature patterns. Multiple machine learning models were trained and evaluated, and the Random Forest Classifier was finalized as the best-performing model after optimization. The trained model was then integrated into a web-based application that allows users to input seed measurements and instantly receive classification results.

The obtained results indicate that machine learning can effectively automate seed variety identification with high accuracy and consistency. This system reduces manual effort, minimizes human error, and provides a fast and reliable solution for seed classification.

Overall, the project bridges the gap between artificial intelligence and agriculture, highlighting how smart technologies can support modern farming practices, seed research, and biodiversity conservation.

10. Future Scope

Although the current system successfully classifies pumpkin seed varieties using machine learning, several opportunities exist for further improvement and expansion.

In the future, the dataset can be expanded to include more pumpkin seed varieties to enhance model generalization and classification coverage. Advanced deep learning techniques and image-based classification using Convolutional Neural Networks (CNNs) can be implemented to classify seed varieties directly from images without requiring manual measurements.

The system can also be improved by integrating real-time image capture via mobile or web cameras for automated feature extraction and prediction. Deploying the application on cloud platforms would enable easy access for farmers, researchers, and agricultural organizations. Additionally, integrating the system with IoT-based smart farming tools could further support automated seed quality monitoring.

With these enhancements, the project can evolve into a comprehensive intelligent seed analysis platform, contributing significantly to agricultural automation, research, and biodiversity management.

11. Appendix

11.1 Source Code

The complete source code for this project includes data pre-processing scripts, machine learning model training files, and web application code for seed variety prediction.

All project files—dataset, model, application code, and documentation—are maintained in the GitHub repository.

11.2 GitHub & Project Video Demo Link

GitHub Repository Link:

https://github.com/MohseenAttar/pumpkin-seed-taxonomy

Project Video Demo Link:

https://drive.google.com/file/d/10U2muRPn3b5FqJIa9IRubqPTGNJPSycF/view?usp=sharing