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

Siddique Sanadi
PRN: [2023011032082]

D. Y. Patil Agriculture and Technical University, Talsande

Academic Year: 2025–2026

Submitted To
Smart-Internz

Abstract

Pumpkin seeds, while frequently underrated, hold considerable nutritional and agricultural significance. They display notable diversity across different cultivars, making their classification a crucial task in agricultural research and biodiversity studies. The project titled *Harvesting Brilliance: A Taxonomic Tale of Pumpkin Seed Varieties* centers on the identification and categorization of pumpkin seed types using machine learning methods.

In this project, a structured dataset comprising morphological traits of pumpkin seeds is gathered and prepared for analysis. Exploratory Data Analysis (EDA) is conducted to comprehend data distributions and detect meaningful patterns. Several machine learning classification algorithms are trained and assessed using performance metrics 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 input seed measurements and promptly receive the predicted seed variety. This project illustrates the practical application of artificial intelligence in taxonomy and agriculture, supporting automated seed classification, advanced crop research, and a deeper understanding of pumpkin seed diversity. It also paves the way for future developments in smart farming, 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 solutions have become indispensable in modern agriculture. Among various agricultural products, pumpkin seeds are nutritionally valuable and commercially significant. Different pumpkin seed varieties exhibit distinct morphological traits, making their classification essential for research, seed quality control, and crop enhancement.

Traditionally, seed classification is done manually by experts relying on visual inspection and experience. However, manual classification is time-intensive, susceptible to human error, and inefficient for large datasets. To address these limitations, machine learning offers automated, precise, and rapid solutions for seed variety identification.

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

This project showcases the integration of data science and agriculture, demonstrating how artificial intelligence can contribute to smart farming, improved seed analysis, and biodiversity conservation. It also offers a user-friendly web interface where users can input seed measurements and obtain classification results instantly, making the system practical for real-world applications.

1.1 Project Overview

The Pumpkin Seed Variety Classification project aims to build an intelligent prediction system that identifies different pumpkin seed varieties using numerical morphological features. A structured dataset containing various seed measurements is collected and preprocessed before applying machine learning techniques.

Multiple classification models are trained and evaluated to identify the best-performing algorithm. The final trained model is deployed through a web-based application that accepts user inputs for seed measurements and returns the predicted seed variety.

The system workflow includes data collection, preprocessing, exploratory data analysis, model training, validation, optimization, and deployment. This end-to-end approach ensures that the project not only focuses on theoretical model development but also delivers a functional and interactive application for practical use.

The project effectively bridges the gap between agricultural research and artificial intelligence by offering an automated solution for pumpkin seed classification.

2.Objectives

The main objective of the project *Harvesting Brilliance: A Taxonomic Tale of Pumpkin Seed Varieties* is to develop an intelligent system that can accurately classify different pumpkin seed varieties using machine learning techniques based on their morphological characteristics.

Specific Objectives:

1. To collect and analyze a dataset containing morphological features of pumpkin seeds.

2. To perform data preprocessing and exploratory data analysis to understand feature distributions and patterns.

3. To apply various machine learning classification algorithms for seed variety prediction.

4. To evaluate model performance using appropriate metrics such as accuracy and classification reports.

5. To optimize the best-performing model to improve prediction reliability.
6. To develop a user-friendly web-based interface for entering seed measurements.

7. To provide instant and accurate prediction of pumpkin seed categories.
8. To demonstrate the application of artificial intelligence in agricultural and taxonomic research.

3.1 Define Problem Statement

Pumpkin seeds are extensively used in food, nutrition, and agriculture due to their rich nutritional content and economic value. Different pumpkin seed varieties possess unique morphological traits important for seed quality evaluation, breeding initiatives, and biodiversity studies. However, manually identifying and classifying pumpkin seed varieties demands expert knowledge, is time-consuming, and may lead to inconsistencies due to human error.

With increasing availability of agricultural datasets and progress in artificial intelligence, there is a need for an automated and accurate system capable of classifying pumpkin seed varieties based on measurable morphological features. Traditional classification methods lack scalability for large datasets and often produce inconsistent predictions.

Hence, the problem addressed in this project is:
To design and develop a machine learning-based system that automatically classifies pumpkin seed varieties using their morphological characteristics and delivers instant prediction results through a user-friendly interface.

This system aims to reduce manual effort, enhance classification accuracy, and support agricultural research and smart farming practices.

3.2 Project Proposal (Proposed Solution)

To tackle the issue of manual and inconsistent classification of pumpkin seed varieties, this project proposes an automated machine learning-based classification system. The proposed solution focuses on analyzing morphological features of pumpkin seeds and predicting their variety using trained classification models.

The system starts with collecting a structured dataset containing various physical measurements of pumpkin seeds such as area, perimeter, major axis length, minor axis length, roundness, and other shape-related attributes. This dataset is then preprocessed to handle missing values, normalize data, and prepare it for model training.

Multiple machine learning classification algorithms are applied to the processed dataset to build predictive models. These models are trained and evaluated using performance metrics such as accuracy and classification reports to determine the most effective algorithm. The best-performing model is selected and optimized to improve prediction accuracy.

To ensure practical usability, the final trained model is integrated into a web-based application. This interface allows users to enter seed measurement values and receive instant predictions of pumpkin seed variety. The proposed system delivers a fast, accurate, and user-friendly solution for automated seed classification, reducing reliance on manual inspection.

This project demonstrates how artificial intelligence can be effectively employed in agriculture to support seed analysis, enhance research efficiency, and promote smart farming technologies.

3.3 Initial Project Planning

Effective project planning is crucial for smooth execution and successful system completion. The initial project planning phase involved defining the workflow, selecting suitable tools and technologies, identifying required resources, and establishing a structured timeline for project development.

During this phase, the overall project architecture was designed, outlining key steps such as data collection, preprocessing, model development, evaluation, optimization, and deployment. Required software tools, including Python, machine learning libraries, and web development frameworks, were chosen based on project needs. Hardware and system requirements were also identified to ensure seamless data processing and model training.

A timeline was created to divide the project into distinct phases, assigning specific tasks such as dataset collection, data analysis, model training, interface development, testing, and final deployment. Potential risks, such as data quality issues and model performance challenges, were considered, and alternative strategies were planned to address them effectively.

Overall, this planning phase provided a clear roadmap for project execution, ensuring systematic progress from data collection to final implementation and documentation.

4.1 Data Collection Plan and Raw Data Sources Identified

Data collection is a vital step in building 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 (Comma Separated Values) format and includes numerical feature measurements required for classification.

The dataset was obtained from a publicly accessible online data repository. In this project, the dataset was downloaded from Kaggle, a well-known platform providing open datasets for research and educational purposes.

DatasetSource:
Kaggle Dataset – Pumpkin Seed Classification

After downloading the dataset, it 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 visualization and analytical techniques. These methods helped identify data distribution patterns, understand feature relationships, and prepare the data for preprocessing and model training.

4.2 Data Quality Report

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

In this project, the pumpkin seed dataset was thoroughly inspected after loading it into the system. The dataset contains numerical morphological features representing different physical characteristics of pumpkin seeds. Initial data exploration was performed to examine the dataset structure, number of records, number of attributes, and data types of each feature.The dataset was checked for missing or null values to ensure completeness. No significant missing values were found, indicating that the dataset was well-structured and ready for further processing. Duplicate records were also examined to avoid biased training results, and none were detected in the dataset.

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

Overall, the dataset was found to be clean, consistent, and suitable for building machine learning classification models. Only standard preprocessing techniques such as normalization and feature scaling were required before model training.

4.3 Data Preprocessing

Data preprocessing is an essential step in machine learning projects, as raw data is rarely in a directly usable form. Proper preprocessing enhances model performance and ensures reliable predictions. In this project, several preprocessing techniques were applied to prepare the pumpkin seed dataset for training classification models.

First, the dataset was loaded into the system using Python libraries such as Pandas. The data was inspected for missing or null values. Since the dataset contained no significant missing data, no major imputation was required. Duplicate records were also checked and removed if necessary to prevent biased learning.

Next, irrelevant attributes were examined, and only meaningful morphological features were selected for model training. The target column representing the pumpkin seed variety was separated from the input features.

Since machine learning algorithms perform better when numerical features are on a similar scale, feature scaling and normalization techniques were applied to standardize the data. This improved model convergence and prediction accuracy.

After scaling, the dataset was divided into training and testing sets. The training set was used to build the machine learning models, while the testing set was used to evaluate model performance on unseen data.

Through these preprocessing steps, the dataset was transformed into a clean and structured format suitable for efficient model training and evaluation.

5.1 Model Selection Report

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

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 during model selection:

- Logistic Regression
- K-Nearest Neighbors (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 that 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 preprocessed 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 evaluating model performance. 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 used was classification accuracy, which measures the percentage of correctly predicted seed varieties. In addition to accuracy, confusion matrices and classification reports were generated to analyze precision, recall, and F1-score for each seed category.

During initial training, different models exhibited varying performance levels. Tree-based models such as 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 Neighbors performed reasonably well but required careful selection of the value of K. Logistic Regression provided baseline performance for comparison.

Based on evaluation results, Random Forest Classifier achieved the highest accuracy among the tested models and showed stable and consistent prediction results. Therefore, it was selected 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.

6.1 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 underfitting.

In this project, the Random Forest Classifier was selected for optimization due to its high initial accuracy and stable performance. Hyperparameter 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 such as Grid Search and cross-validation were used to test different parameter values systematically. Cross-validation ensured that the model performance was consistent across different subsets of the dataset and not dependent on a single train-test split.

Through repeated experiments and evaluation, the best hyperparameter configuration was selected, which improved classification accuracy and reduced prediction errors. The tuned model showed 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 usage.

6.2 Final Model Selection Justification

After conducting initial training, evaluation, and hyperparameter tuning, the Random Forest Classifier was selected 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 demonstrated 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 Neighbors, 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

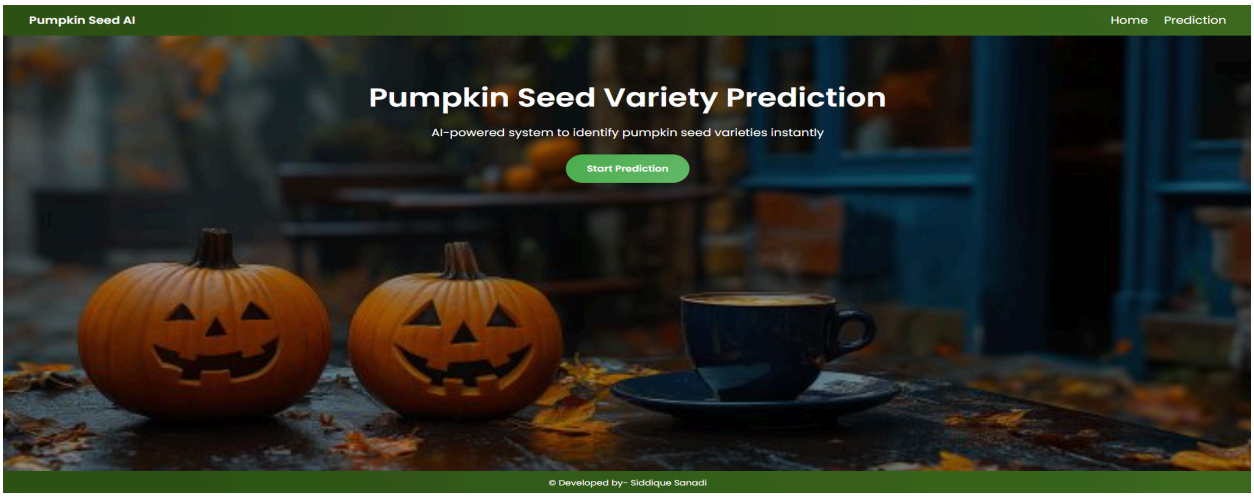
After successful model training, evaluation, and optimization, the final tuned Random Forest model was integrated into a web-based application. The system was tested using different input values representing pumpkin seed morphological features, and the model accurately predicted the corresponding seed variety.

The developed application provides a simple and interactive user interface where users can enter seed measurements such as area, perimeter, axis lengths, and shape-related attributes. Upon clicking the Predict Category button, the system processes the input data through the trained model and instantly displays the predicted pumpkin seed variety.

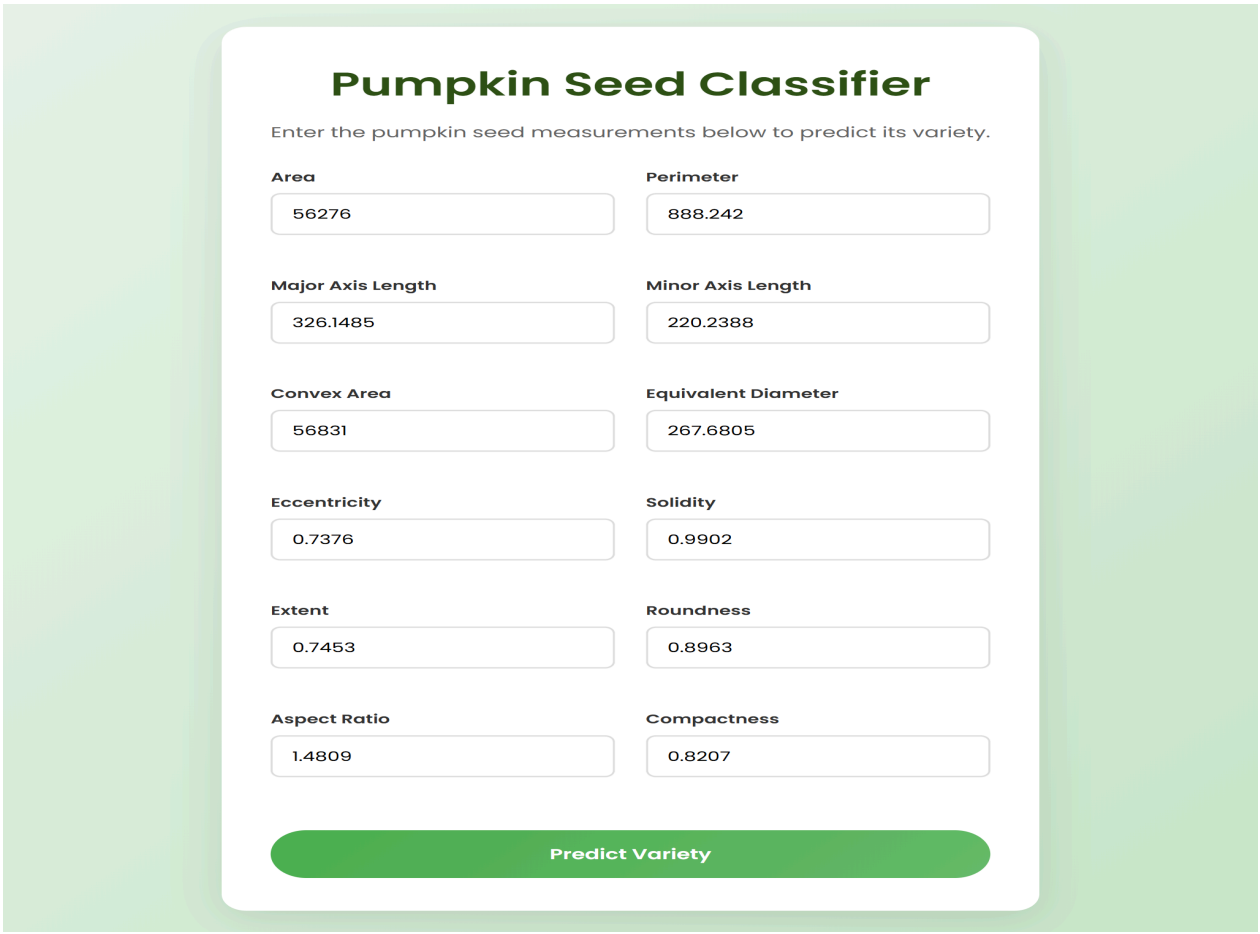
The prediction results were found to be accurate and consistent, demonstrating the effectiveness of the trained machine learning model. The system successfully automates the classification process, reducing manual effort and improving prediction reliability.

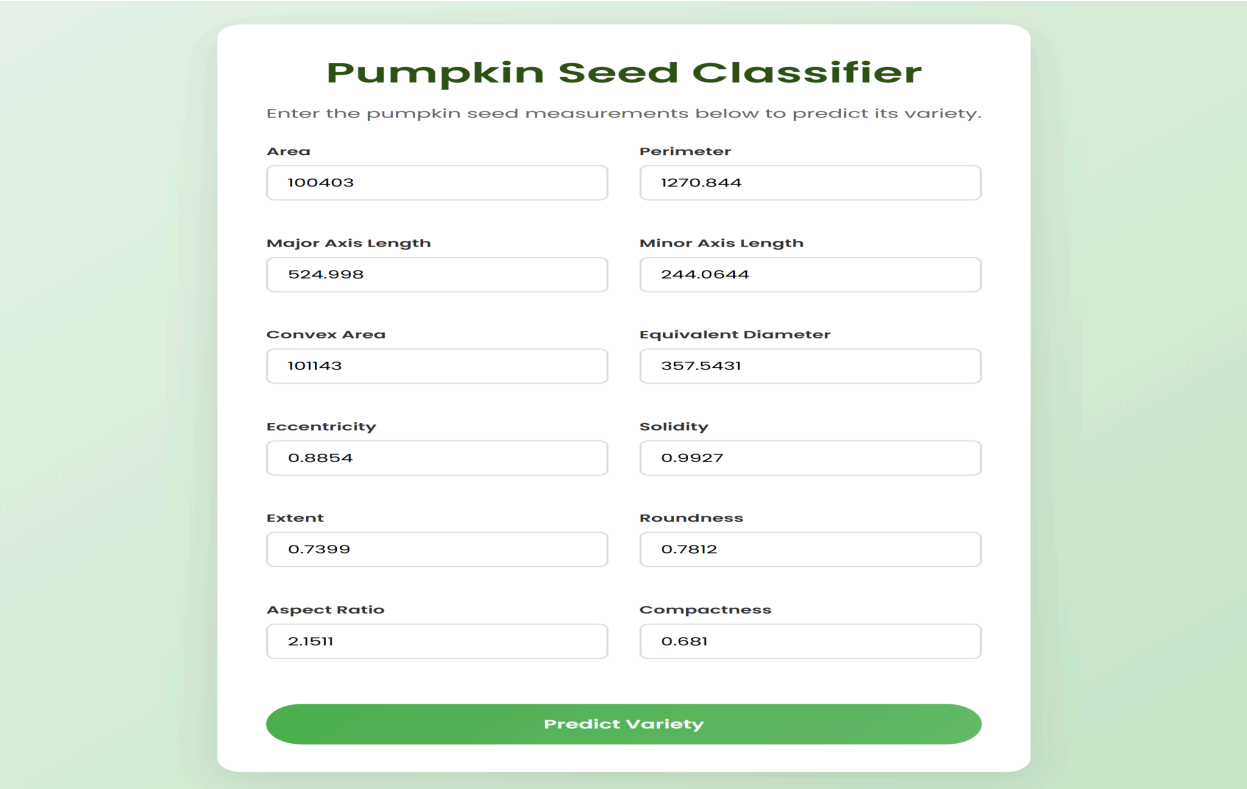
7.1 Output Screenshots

- Figure 1 – Pumpkin Seed Classifier Home Page

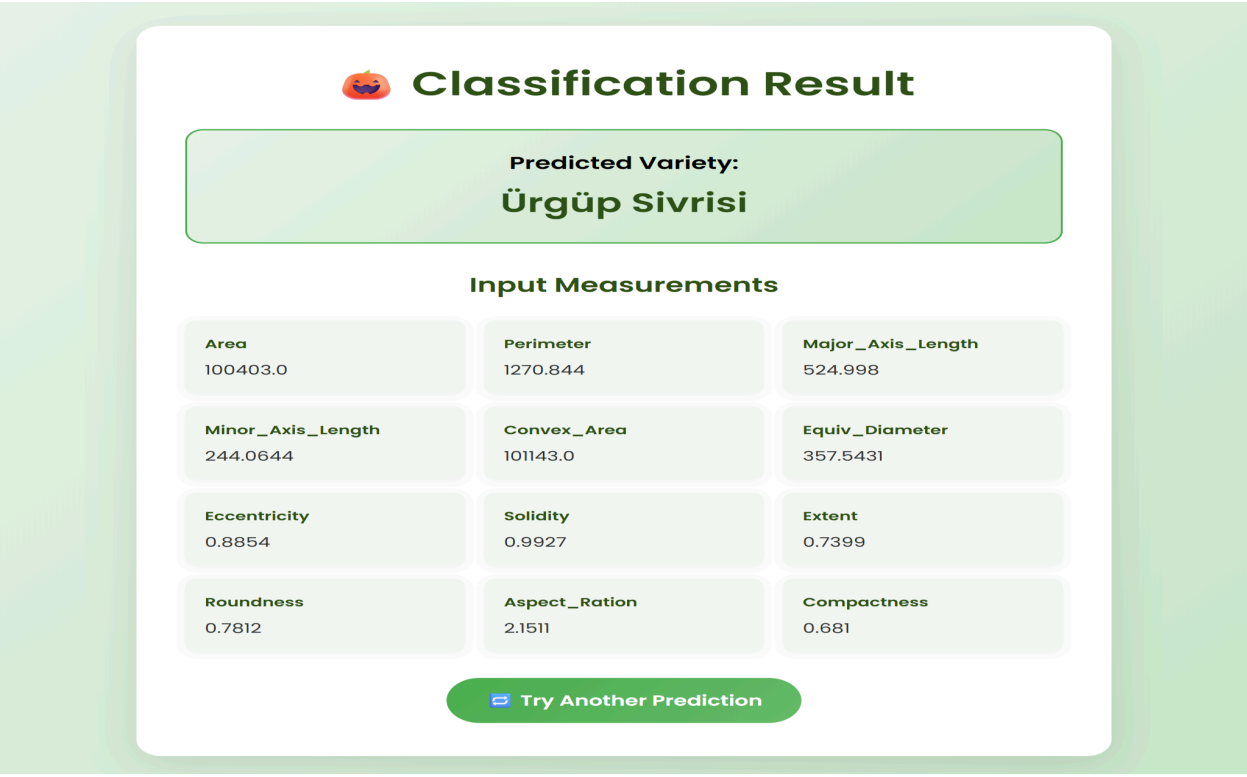


- Figure 2 – Pumpkin Seed Classifier Input Interface





- Figure 3– Pumpkin Seed Classifier Output



🎃 Classification Result

Predicted Variety:

Çerçevelik

Input Measurements

Area 56276.0	**Perimeter** 888.242	**Major_Axis_Length** 326.1485
Minor_Axis_Length 220.2388	**Convex_Area** 56831.0	**Equiv_Diameter** 267.6805
Eccentricity 0.7376	**Solidity** 0.9902	**Extent** 0.7453
Roundness 0.8963	**Aspect_Ration** 1.4809	**Compactness** 0.8207

🔁 Try Another Prediction

8. Advantages and Disadvantages

8.1 Advantages

1. Provides automated and accurate classification of pumpkin seed varieties.
2. Reduces dependency on manual inspection and expert knowledge.
3. Improves speed and efficiency in seed analysis.
4. Machine learning model offers consistent and reliable predictions.
5. User-friendly web interface makes the system easy to use.
6. Supports agricultural research and seed quality assessment.
7. Helps in understanding biodiversity and seed variety distribution.

8.2 Disadvantages

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

scenarios.

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 their morphological characteristics using supervised learning techniques.

A structured dataset was collected, preprocessed, and analyzed 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 show 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, there are several opportunities for further improvement and expansion of the project.

In the future, the dataset can be extended to include more pumpkin seed varieties to improve 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 seed images without requiring manual measurement inputs.

The system can also be enhanced by integrating real-time image capture through mobile or web cameras for automated feature extraction and prediction. Deploying the application on cloud platforms would allow 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 complete intelligent seed analysis platform, contributing significantly to agricultural automation, research, and biodiversity management.

11. Appendix

11.1 Source Code

The complete source code of this project includes data preprocessing scripts, machine learning model training files, and the web application code for seed variety prediction.

All project files such as 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/siddique107/Harvesting-Brilliance-A-Taxonomic-Tale-of-Pumpkin-Seed-Varieties

Project Video Demo Link:
https://drive.google.com/file/d/1S9n7OY8nnOWPOq8PWuvtPLJ6No-v8n1p/view?usp=sharing