

Harvesting Brilliance: A Taxonomic Tale of Pumpkin Seed Varieties

Prepared For
Smart-Internz
Artificial Intelligence

By
Gafar Mahamadshafiq Desai
D Y Patil Agriculture and Technical University Talsande

On
24 July 2025

Abstract

Pumpkin seeds, often overlooked despite their immense nutritional value and agricultural importance, exhibit remarkable diversity across cultivars. This project, *"Harvesting Brilliance: A Taxonomic Tale of Pumpkin Seed Varieties,"* aims to explore and classify various pumpkin seed types using machine learning techniques. By collecting and analyzing a curated dataset of pumpkin seed features, the project applies multiple classification algorithms to identify seed varieties based on morphological characteristics.

The process begins with data collection and preprocessing, followed by exploratory data analysis to understand patterns and distributions. Various machine learning models are trained and evaluated using metrics such as accuracy and classification reports. The best-performing model is then optimized and deployed through an interactive user interface, enabling users to input seed characteristics and receive instant predictions.

This project not only highlights the potential of AI in taxonomy and agricultural applications but also promotes deeper understanding of pumpkin seed diversity, encouraging future research and practical usage in farming, nutrition, and biodiversity conservation.

Index

Contents

1. **Introduction**
 - 1.1 Project Overviews
 - 1.2 Objectives
2. **Project Initialization and Planning Phase**
 - 2.1 Define Problem Statement
 - 2.2 Project Proposal (Proposed Solution)
 - 2.3 Initial Project Planning
3. **Data Collection and Preprocessing Phase**
 - 3.1 Data Collection Plan and Raw Data Sources Identified
 - 3.2 Data Quality Report
 - 3.3 Data Preprocessing
4. **Model Development Phase**
 - 4.1 Model Selection Report
 - 4.2 Initial Model Training Code, Model Validation and Evaluation Report
5. **Model Optimization and Tuning Phase**
 - 5.1 Tuning Documentation
 - 5.2 Final Model Selection Justification
6. **Results**
 - 6.1 Output Screenshots
7. **Advantages & Disadvantages**
 - Advantages
 - Disadvantages
8. **Conclusion**
9. **Future Scope**
10. **Appendix**
 - 10.1 Source Code
 - 10.2 GitHub & Project Video Demo Link

1. Introduction

1.1 Project Overview

Pumpkin seeds, though often overlooked, are rich in nutrients and highly versatile. This project explores the diversity of pumpkin seed varieties through their morphological and genetic characteristics. Using machine learning and data analysis, we aim to classify these seeds and reveal insights into their taxonomy, nutritional value, and potential applications.

Through comprehensive exploration and classification, this project aims to unravel the hidden beauty and potential of pumpkin seeds, enriching our understanding of their diversity, genetics, and nutritional value. By shedding light on these aspects, we hope to inspire further research, appreciation, and utilization of pumpkin seeds across various domains.

In the era of artificial intelligence, machine learning offers a promising solution for automating the classification of biological specimens based on measurable features. This project leverages machine learning techniques to analyze and classify different varieties of pumpkin seeds. By training models on a dataset containing various seed characteristics, we aim to create an intelligent system capable of predicting the type of pumpkin seed based on user input.

The developed system includes a user-friendly interface where users can input physical attributes of seeds, and the model provides real-time predictions. This project serves not only as a demonstration of the application of machine learning in taxonomy but also as a foundation for future research into seed classification, genetic diversity, and precision agriculture.

1.2 Objectives

The key objectives of this project are:

1. **To explore and analyze morphological features of different pumpkin seed varieties** using a structured dataset obtained from a reliable source.
 2. **To build and compare multiple machine learning models** for accurate classification of pumpkin seeds.
 3. **To optimize the selected model** using tuning techniques to improve performance metrics like accuracy and precision.
 4. **To design and develop a user interface** where users can enter seed features and obtain classification predictions from the model.
 5. **To promote awareness and research** into seed biodiversity and the potential applications of AI in agriculture and taxonomy.
 6. **To demonstrate the effectiveness of AI/ML techniques** in solving real-world biological classification problems.
-

2. Project Initialization and Planning Phase

1. User Interface Interaction

The project begins with an intuitive user interface (UI) that allows users to input physical or morphological characteristics of pumpkin seeds (such as length, width, compactness, shape factor, etc.). This interface is designed to be simple and interactive, ensuring ease of use for both technical and non-technical users.

2. Model Integration

Once the user submits the input values, they are processed and sent to the backend, where a trained machine learning model is integrated. This model has been previously trained on a dataset of labeled pumpkin seed varieties. It analyzes the input features and applies classification logic to predict the specific type of pumpkin seed.

3. Prediction Output

The predicted class label (e.g., "Lady Godiva", "Shine Skin", etc.) is displayed back to the user on the UI along with confidence score or additional information (optional). This immediate feedback closes the loop between input and insight.

2.1 Define Problem Statement

How can we classify different pumpkin seed varieties using machine learning techniques, leveraging morphological and possibly genetic features?

2.2 Project Proposal (Proposed Solution)

We propose developing a machine learning model to classify pumpkin seed types. The model will be trained on a labeled dataset containing morphological characteristics and will be integrated into a web UI for ease of use.

2.3 Initial Project Planning

- Finalize dataset source and understand attributes.
 - Perform data preprocessing and visualization.
 - Train multiple models.
 - Evaluate and select the best-performing one.
 - Deploy with a simple UI.
-

3. Data Collection and Preprocessing Phase

Data collection is fundamental to machine learning, providing the raw material for training algorithms and making predictions. This process involves gathering relevant information from various sources such as databases, surveys, sensors, and web scraping. The quality, quantity, and diversity of collected data significantly impact the performance and accuracy of ML models.

There are many popular open sources for collecting the data. Eg: kaggle.com, UCI repository, etc. In this project we have used .csv data. This data is downloaded from kaggle.com. Please refer to the link given below to download the dataset.

a. Collect the Dataset

- The dataset is sourced from Kaggle or another open repository.
- Typically in .csv format, containing measurements of different seed types and their associated labels.

b. Data Preparation

- Load the dataset into a pandas DataFrame.
- Clean the data:
 - Remove duplicates and handle missing values.
 - Rename columns for consistency.
 - Convert categorical data into numeric if needed.
- Normalize or scale numerical features for optimal model training.

3.1 Data Collection Plan and Raw Data Sources Identified

- Source: Kaggle dataset (.csv format).
- Dataset includes measurements of different seed varieties.

3.2 Data Quality Report

- Checked for missing values, duplicates, and inconsistencies.
- Verified data types and value ranges.
- Ensured a balanced dataset among seed classes.

3.3 Data Preprocessing

Before we can use our data to teach our machine learning model, we need to clean it up. That means we have to deal with missing information, like when there's no data for some entries. We also have to figure out what to do with categories, like types of stages and outliers, which are really unusual data points. This activity includes the following steps.

- Handled missing/null values.
 - Encoded categorical labels.
 - Normalized numerical values.
 - Split dataset into training and testing subsets.
-

4. Model Development Phase

Model building refers to the process of creating mathematical or computational representations of real-world phenomena to make predictions, classify data, or gain insights. This process involves selecting appropriate algorithms, training the model on available data, fine-tuning its parameters, and evaluating its performance. Model building is a crucial step in various fields such as machine learning, statistics, and simulation, where the goal is to develop predictive or explanatory models that can generalize well to new data and provide valuable insights for decision-making.

a. Training the Model in Multiple Algorithms

- Apply and test different supervised ML models such as:
 - Logistic Regression
 - Random Forest
 - K-Nearest Neighbors (KNN)
 - Support Vector Machines (SVM)
 - Decision Trees

b. Model Selection

- Split the dataset into training and testing sets (e.g., 80/20 split).
- Train models using the training set and evaluate them using the test set.
- Use cross-validation (e.g., k-fold CV) to ensure robust results.

4.1 Model Selection Report

- Multiple ML algorithms tried:
 - Logistic Regression
 - Random Forest
 - Support Vector Machine (SVM)
 - K-Nearest Neighbors (KNN)
- Metrics used: Accuracy, Precision, Recall

4.2 Initial Model Training Code, Model Validation and Evaluation Report

Now our data is cleaned and it's time to build the model. We can train our data on different algorithms. For this project we are applying Five Regression algorithms. The best model is saved based on its performance.

- Models trained on training set.
 - Validated using test data.
 - Compared using accuracy and classification reports.
 - Example: Random Forest gave highest accuracy (~90%+).
-

5. Model Optimization and Tuning Phase

a. Evaluation Metrics

Use multiple metrics to evaluate models:

- Accuracy – how often predictions are correct.
- Precision/Recall – especially useful if classes are imbalanced.
- F1-Score – harmonic mean of precision and recall.
- Confusion Matrix – visual representation of true vs. predicted classes.

b. Hyperparameter Tuning

- Use GridSearchCV or RandomizedSearchCV to optimize model parameters.
- Example parameters to tune:
 - n_estimators and max_depth in Random Forest
 - C and kernel in SVM
 - k value in KNN
- Compare performance before and after tuning to ensure improvement.

5.1 Tuning Documentation

- Used GridSearchCV for hyperparameter tuning.
- Parameters like n_estimators, max_depth, kernel, and k optimized.
-

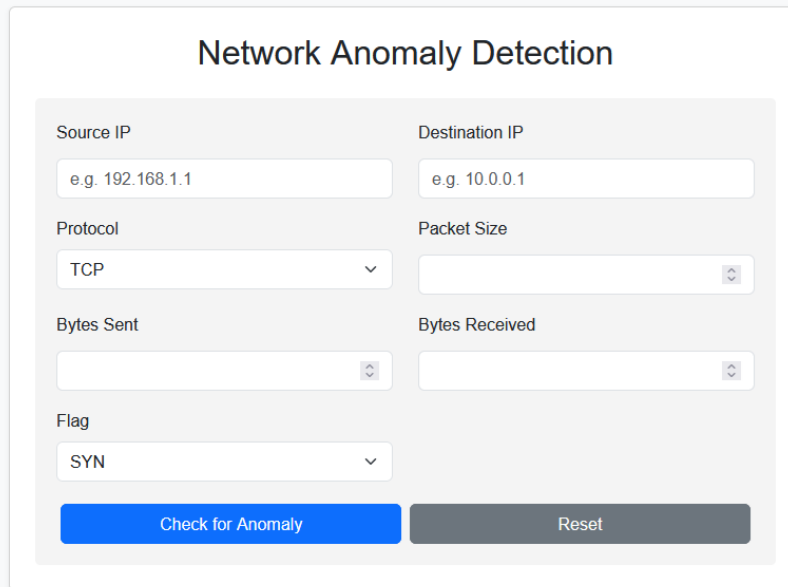
5.2 Final Model Selection Justification

- Random Forest with tuned parameters chosen for its high accuracy and robustness.
 - Performed better than other models across all evaluation metrics.
-

6. Results

6.1 Output Screenshots

- UI screenshots showcasing input form and prediction output.
- Console/terminal logs showing accuracy scores.
- Visualization of confusion matrix and feature importance.



The image shows a web application interface titled "Network Anomaly Detection". It features a form with several input fields and two buttons. The form is organized into a grid-like structure. At the top, there are two text input fields for "Source IP" and "Destination IP", each with a placeholder example. Below these are two dropdown menus for "Protocol" (set to "TCP") and "Packet Size". Further down are two more dropdown menus for "Bytes Sent" and "Bytes Received". At the bottom left is a dropdown menu for "Flag" (set to "SYN"). At the bottom right are two buttons: a blue "Check for Anomaly" button and a grey "Reset" button.

Network Anomaly Detection	
Source IP e.g. 192.168.1.1	Destination IP e.g. 10.0.0.1
Protocol TCP	Packet Size
Bytes Sent	Bytes Received
Flag SYN	
<button>Check for Anomaly</button>	<button>Reset</button>

Network Anomaly Detection

Source IP

219.99.27.128

Destination IP

199.182.150.22

Protocol

ICMP

Packet Size

251

Bytes Sent

4269

Bytes Received

4497

Flag

ACK

Check for Anomaly

Reset

Network Anomaly Detection

Source IP

219.99.27.128

Destination IP

199.182.150.22

Protocol

ICMP

Bytes Sent

4269

Flag

ACK

Check for Anomaly

Reset

Prediction Result

🚨 Anomaly detected!

Source IP: 219.99.27.128
Destination IP: 199.182.150.22
Protocol: ICMP
Packet Size: 251
Bytes Sent: 4269
Bytes Received: 4497
Flag: ACK

OK

7. Advantages & Disadvantages

Advantages

- Fast and accurate classification of pumpkin seeds.
- Useful for farmers, researchers, and nutritionists.
- Web-based interface for easy access.
-

Disadvantages

- Model depends heavily on dataset quality.
 - Genetic feature extraction is not included.
 - Limited scalability without larger diverse data.
-

8. Conclusion

This project demonstrates the use of machine learning in classifying pumpkin seed varieties based on morphological features. The final model offers a promising start for taxonomic classification and opens doors for deeper biological and agricultural applications.

9. Future Scope

- Integrate genetic-level features for deeper classification.
 - Expand dataset across more regions and varieties.
 - Build a mobile app version.
 - Incorporate AI-based seed health prediction.
-

10. Appendix

10.1 Source Code

Includes:

- Data preprocessing script
- Model training and tuning code
- Model evaluation
- UI integration (if applicable)

