HARVESTING BRILLIANCE:

A TAXONOMIC TALE OF PUMPKIN SEED VARIETIES

A Detailed Project Report Submitted in Partial Fulfillment of the Requirements for the Award of Degree

Submitted By:
Aditya Dattatray Dange

Department of Computer Science,
D. Y. Patil Agriculture and Technical University, Talsande
Academic Year: 2025–2026

ABSTRACT

Pumpkin seeds are nutritionally rich and agriculturally important. This project focuses on classifying pumpkin seed varieties using Machine Learning techniques based on morphological characteristics such as area, perimeter, and axis lengths. Multiple supervised learning algorithms were implemented and compared. The best-performing model was deployed using Flask as a web application for real-time prediction. This system supports agricultural decision-making, nutritional research, and food industry innovations.

1. INTRODUCTION

Agriculture plays a crucial role in the global economy. Pumpkin seeds are widely consumed due to their high protein, mineral, and antioxidant content. Different varieties exhibit variation in size, structure, and composition. Manual classification is time-consuming and prone to errors. Machine Learning provides an automated and accurate solution for classification tasks.

This project integrates agricultural science with Artificial Intelligence to classify pumpkin seed varieties efficiently.

2. LITERATURE REVIEW

Several studies have explored the use of Machine Learning in agricultural classification. Algorithms such as Decision Trees and Random Forest have demonstrated high accuracy in seed and crop classification tasks. Feature scaling and proper preprocessing significantly improve model performance.

Research also highlights the importance of Exploratory Data Analysis (EDA) in understanding dataset structure before model building.

3. PROBLEM STATEMENT

To develop a robust Machine Learning model capable of accurately classifying pumpkin seed varieties based on morphological attributes and deploy it through a user-friendly web interface.

4. OBJECTIVES

- To collect and preprocess pumpkin seed dataset.
- To perform detailed Exploratory Data Analysis.
- To train and compare multiple classification algorithms.
- To evaluate model performance using multiple metrics.
- To deploy the best-performing model using Flask framework.

5. DATASET DESCRIPTION

The dataset was obtained from Kaggle in CSV format. It contains morphological features of pumpkin seeds including Area, Perimeter, Major Axis Length, Minor Axis Length, Convex Area, Eccentricity, and Class label.

The dataset contains sufficient samples to train and test classification models effectively.

6. METHODOLOGY

6.1 Data Collection

The dataset was downloaded from Kaggle and imported into the Python environment using Pandas library.

6.2 Data Preprocessing

Data preprocessing involved checking for missing values, detecting outliers using boxplots, scaling numerical features using MinMaxScaler, and removing unnecessary columns.

Outliers were handled carefully to maintain data integrity. Feature scaling ensured that all numerical values were normalized within a fixed range.

6.3 Exploratory Data Analysis

EDA included descriptive statistics and visualization techniques such as histograms, countplots, scatter plots, and pairplots. These analyses helped in understanding data distribution and relationships between variables.

6.4 Train-Test Split

The dataset was split into training (80%) and testing (20%) sets using train_test_split() to evaluate model generalization.

6.5 Model Building

The following algorithms were implemented:

- Logistic Regression

- Decision Tree Classifier

- Random Forest Classifier
- Naive Bayes
- Support Vector Machine
- Gradient Boosting Classifier

Each model was trained using training data and evaluated using test data.

6.6 Performance Evaluation

Models were evaluated using Accuracy Score, Precision, Recall, and F1-Score. Confusion matrix analysis was also performed to understand classification performance.

Hyperparameter tuning improved model accuracy and reduced overfitting.

7. MODEL DEPLOYMENT

The best-performing model was saved using Pickle as model.pkl. A Flask web application was developed to integrate the model with a user interface.

Users enter feature values through HTML forms. The Flask backend processes the input, applies the trained model, and displays predictions on the webpage.

8. APPLICATIONS

Agricultural Sector: Helps farmers select suitable seed varieties for improved yield and pest resistance.

Nutrition and Health: Assists researchers in identifying nutrient-rich seed varieties.

Food Industry: Supports development of innovative pumpkin seed-based products.

9. RESULTS AND DISCUSSION

Among all algorithms tested, Random Forest and Gradient Boosting showed superior performance. The deployed model achieved high accuracy and demonstrated reliability in real-time predictions.

Comparative analysis highlighted the importance of ensemble methods in classification tasks.

10. CONCLUSION

This project successfully demonstrates the application of Machine Learning in agricultural taxonomy. The integration of AI with agriculture enhances productivity, accuracy, and decision-making capabilities.

The deployed web application makes the system accessible and practical for real-world usage.

11. FUTURE SCOPE

- Cloud deployment for global accessibility.
- Image-based seed classification using Deep Learning.
- Integration with IoT devices for automated agricultural analysis.
- Expansion of dataset with genetic and environmental factors.

12. REFERENCES

- Scikit-learn Documentation
- Kaggle Dataset Repository
- Python Official Documentation
- Research Articles on Agricultural Machine Learning Applications