D.Y. PATIL AGRICULTURAL AND TECHNICAL UNIVERSITY, TALSANDE

PROJECT REPORT

Harvesting Brilliance:

A Taxonomic Tale of Pumpkin Seed Varieties

Submitted By	Sanika Prakash Done
Roll No.	184
PRN	2022011031044
University	D.Y. Patil Agricultural and Technical University, Talsande

1. Project Overview

Pumpkin seeds, though often overlooked, are remarkably diverse in their morphological and genetic characteristics. This project — **Harvesting Brilliance** — is a supervised machine learning pipeline that classifies pumpkin seeds into two distinct botanical varieties: **Çerçevelik** and **Ürgüp Sivrisi**. By leveraging morphological measurements such as area, perimeter, and axis lengths, the model enables accurate taxonomic identification with real-world agricultural, nutritional, and culinary applications.

Attribute	Detail
Problem Type	Binary Classification (Supervised ML)
Target Variable	Class: Çerçevelik (0) / Ürgüp Sivrisi (1)
Tech Stack	Python, Pandas, Scikit-learn, Flask, HTML/CSS
Dataset	Pumpkin_Seeds_Dataset.xlsx — 2,500 samples, 13 features
Best Model	Random Forest (Accuracy: 88.9%)
Deployment	Flask Web Application with real-time prediction

2. Project Description & Real-World Impact

Pumpkin seeds exhibit remarkable diversity across cultivars, with differences in shape, size, color, and biochemical composition. This project explores, analyzes, and classifies these seeds to unravel their hidden potential — from supporting precision agriculture to informing nutritional science and unlocking culinary innovation.

Scenario 1 — Farmers' Conference Presentation

At a national farmers' conference, the research team demonstrated how morphological data can distinguish pumpkin seed varieties with high accuracy. Farmers interacted with the classification results to understand which variety — Çerçevelik or Ürgüp Sivrisi — performs better under specific soil and climate conditions. This empowers data-driven decisions in seed selection, potentially improving crop yields, pest resistance, and post-harvest quality.

Scenario 2 — Nutrition and Health Symposium

At an international nutrition symposium, scientists leveraged the classification framework to cross-reference variety identity with protein, mineral, and vitamin content datasets. The model's ability to reliably distinguish varieties enabled diet researchers and food scientists to formulate evidence-based dietary recommendations and develop functional foods enriched with the most nutritious pumpkin seed variety.

Scenario 3 — Culinary Collaboration with Food Brands

A premier food brand partnered with the project team to explore the culinary differentiation of pumpkin seed varieties. Chefs and food technologists used variety identification to curate recipes, design gourmet snack lines, craft artisanal breads, and develop specialty cold-pressed oils — each tailored to the sensory profile of the classified seed variety. This demonstrates how a machine learning pipeline can directly influence consumer product innovation.

3. Data Collection & Preparation

The dataset — **Pumpkin_Seeds_Dataset.xlsx** — contained 2,500 seed samples across 13 morphological columns. It was loaded via **pandas.read_excel()** and subjected to a rigorous cleaning pipeline before modeling.

- **Null Check:** Verified zero missing values across all columns.
- **Outlier Removal:** Applied the Interquartile Range (IQR) method on the 'Area' column. Records outside [Q1 – 1.5×IQR, Q3 + 1.5×IQR] were dropped.
- **Feature Scaling:** Applied MinMaxScaler to normalize 'Area', 'Perimeter', and 'Major_Axis_Length' to [0, 1].
- **Feature Selection:** Dropped redundant/highly correlated columns: Convex_Area, Equiv_Diameter, Eccentricity, Minor_Axis_Length.
- **Label Encoding:** Converted target Class to binary integers (0 = Çerçevelik, 1 = Ürgüp Sivrisi).

3.1 Dataframe After Feature Engineering

```
Dropping Columns
Columns Dropped. New Dataframe Head:
      Area  Perimeter  Major_Axis_Length  Solidity  Extent  Roundness  Aspect_Ration  Compactness       Class
0  0.119284   0.033709           0.016192    0.9902  0.7453     0.8963         1.4809       0.8207  Çerçevelik
1  0.410519   0.340661           0.294143    0.9916  0.7151     0.8440         1.7811       0.7487  Çerçevelik
2  0.338866   0.365983           0.351048    0.9857  0.7400     0.7674         2.0651       0.6929  Çerçevelik
3  0.264966   0.210828           0.185370    0.9902  0.7396     0.8486         1.7146       0.7624  Çerçevelik
4  0.259944   0.221228           0.192467    0.9850  0.6752     0.8338         1.7413       0.7557  Çerçevelik
```

Figure 1: Dataset head after dropping redundant columns and label encoding.

3.2 Outlier Detection & Removal

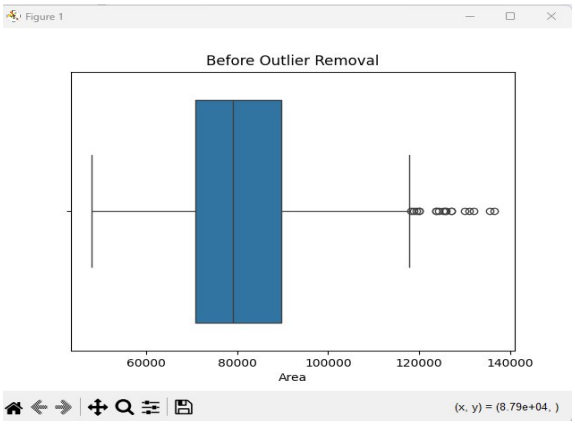

Figure 2a: Boxplot of 'Area' — Before Outlier Removal

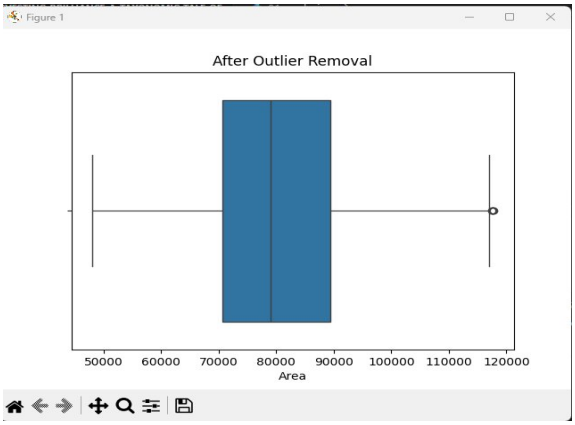

Figure 2b: Boxplot of 'Area' — After Outlier Removal

4. Exploratory Data Analysis (EDA)

Comprehensive EDA was performed to understand the data distribution, class balance, feature relationships, and multicollinearity before training the models.

4.1 Descriptive Statistics

	Area	Perimeter	Major_Axis_Length	Solidity	Extent	Roundness	Aspect_Ration	Compactness
count	2482.000000	2482.000000	2482.000000	2482.000000	2482.000000	2482.000000	2482.000000	2482.000000
mean	0.463459	0.442677	0.411127	0.989479	0.693502	0.791838	2.039858	0.704435
std	0.188186	0.180938	0.167586	0.003499	0.060676	0.055916	0.315819	0.053053
min	0.000000	0.000000	0.000000	0.918600	0.468000	0.554600	1.148700	0.560800
25%	0.325145	0.307016	0.285968	0.988300	0.659300	0.752325	1.800325	0.663900
50%	0.443448	0.433898	0.391094	0.990300	0.713250	0.798200	1.982850	0.707900
75%	0.592092	0.567014	0.522056	0.991500	0.740275	0.834575	2.258775	0.743700
max	1.000000	1.000000	1.000000	0.994400	0.829600	0.939600	3.144400	0.904900

Figure 3: Statistical summary (count, mean, std, min, quartiles, max) for all features.

4.2 Class Distribution

A countplot confirmed a balanced dataset with approximately 1,300 samples of Çerçevelik and 1,182 samples of Ürgüp Sivrisi — minimizing class imbalance bias in model training.

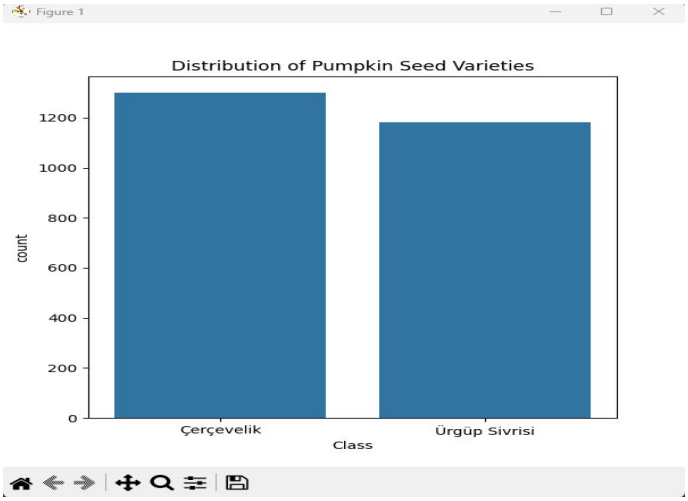

Figure 4: Distribution of pumpkin seed varieties — near-balanced classes.

4.3 Bivariate Analysis — Area vs. Perimeter

A scatter plot of normalized Area vs. Perimeter, color-coded by class, revealed a strong positive linear correlation with partial but imperfect class separation. Both varieties overlap in the mid-range, underscoring the need for multi-feature classification models.

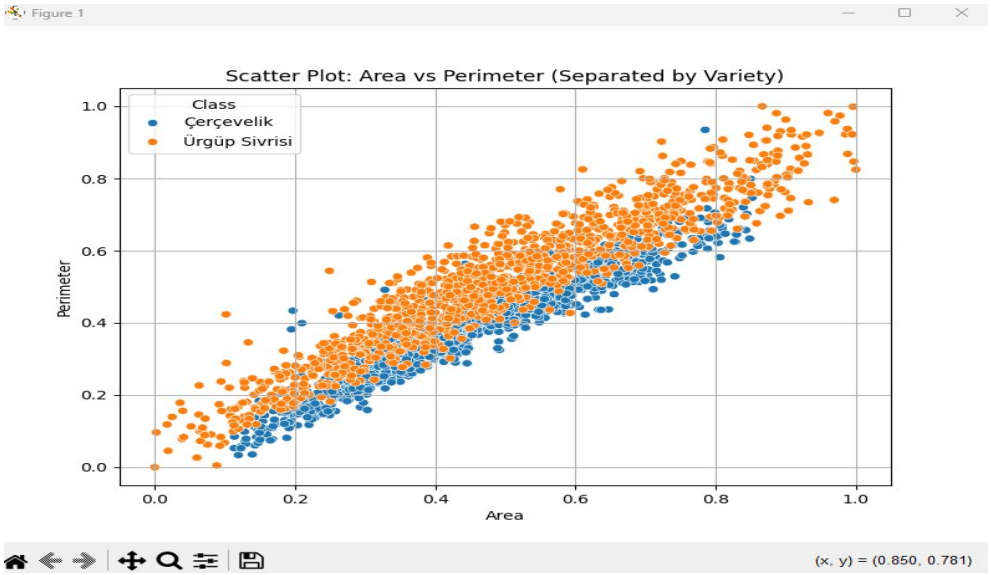

Figure 5: Scatter plot of Area vs. Perimeter, separated by seed variety.

4.4 Multivariate Analysis — Correlation Heatmap

The correlation matrix highlighted key relationships: Area–Perimeter (0.92), Roundness–Aspect_Ratio (–0.93), and Compactness–Aspect_Ratio (–0.99). These strong correlations informed the feature engineering step, where redundant features were eliminated.

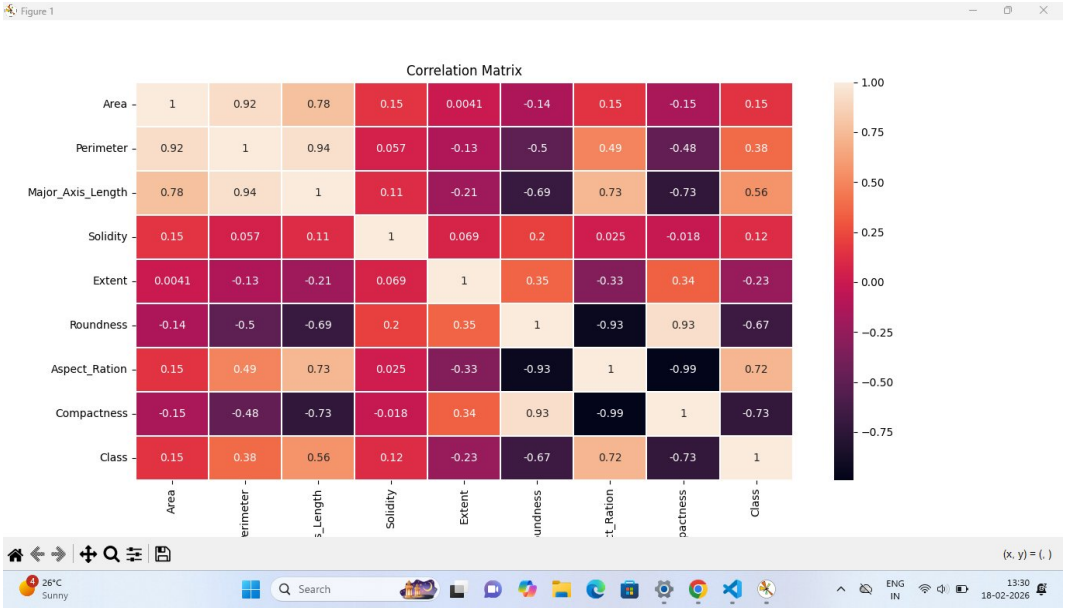

Figure 6: Correlation heatmap showing inter-feature relationships.

5. Model Building

Six supervised learning algorithms were trained on an 80/20 train-test split (random_state=30). Each model was evaluated on accuracy, precision, recall, and F1-score.

Model	Accuracy	Remarks
Gradient Boosting	88.7%	Top performer — ensemble boosting
Random Forest	88.9%	Selected for deployment — robust & interpretable
Support Vector Machine	86.5%	Strong margin-based classifier
Logistic Regression	86.1%	Solid linear baseline
Decision Tree	83.7%	Simple but prone to overfitting
Naive Bayes	70.2%	Weakest — assumes feature independence

Table 1: Model comparison — accuracy scores on test set.

5.1 Classification Reports

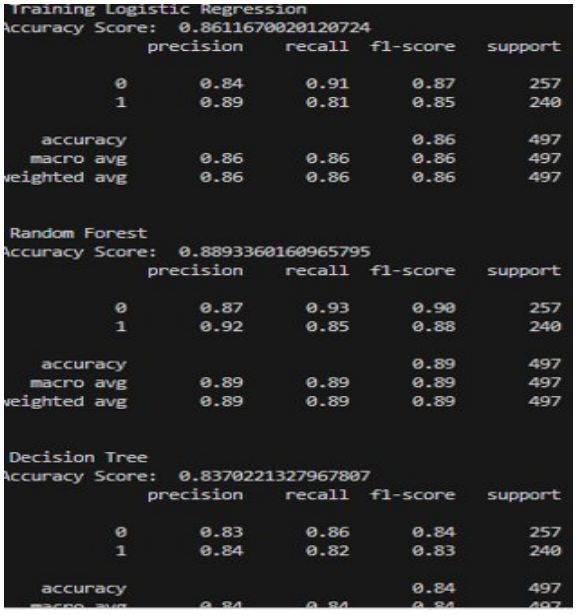

Figure 7a: Classification reports — Logistic Regression, Random Forest, Decision Tree

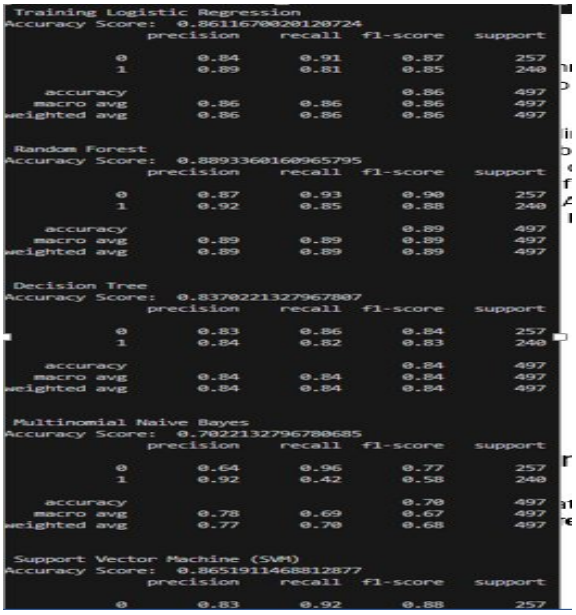

Figure 7b: Classification reports — All six models including SVM, NB, Gradient Boosting

5.2 Final Model Scores

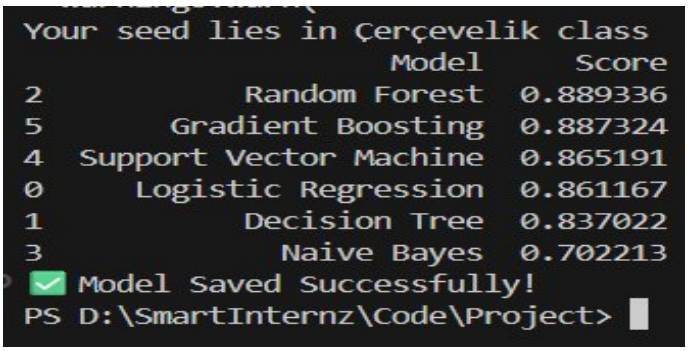

Figure 8: Ranked model scores — Random Forest selected for deployment.

6. Model Deployment

The best-performing model — **Random Forest** — was serialized using Python's **pickle** library and deployed as a **Flask web application**. The application accepts 8 morphological inputs from a browser form and returns the predicted pumpkin seed variety in real time.

- **Serialization:** Trained model saved as model.pkl using pickle.
- **Backend (Flask):** app.py loads model.pkl and exposes a /predict POST route.
- **Frontend — index.html:** User-friendly form accepting: Area, Perimeter, Major Axis Length, Solidity, Extent, Roundness, Aspect Ratio, Compactness.
- **Frontend — predict.html:** Results page displaying the classified seed variety prominently.
- **Integration:** Browser → Flask → model.pkl → prediction → browser response in real time.

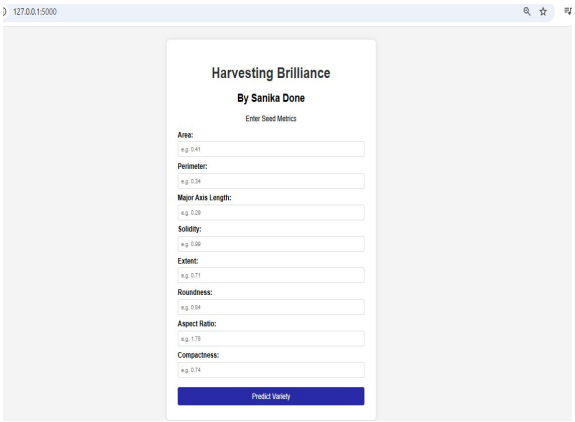

Figure 9: Web application input form — 'Harvesting Brilliance' by Sanika Done

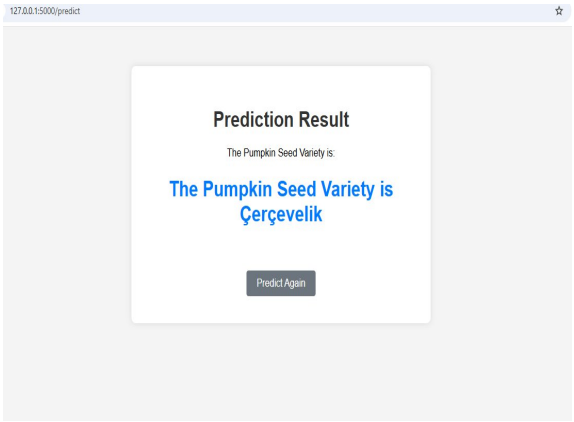

Figure 10: Prediction result page — seed classified as 'Çerçevelik'

7. Conclusion

This project successfully designed, trained, and deployed a machine learning system for taxonomic classification of pumpkin seed varieties using morphological features. The **Random Forest classifier** achieved the highest accuracy of **88.9%**, demonstrating strong generalization on unseen data.

Through rigorous EDA, outlier treatment, feature engineering, and comparative model evaluation, the pipeline showcases best practices in an end-to-end ML workflow. The Flask deployment transforms the model into a usable tool for non-technical stakeholders — farmers, nutritionists, and food technologists alike.

Beyond classification accuracy, this project underscores the transformative potential of data science in agriculture — enabling precision farming, evidence-based nutritional guidance, and product-level culinary innovation through machine intelligence.
