Support Vector Machines

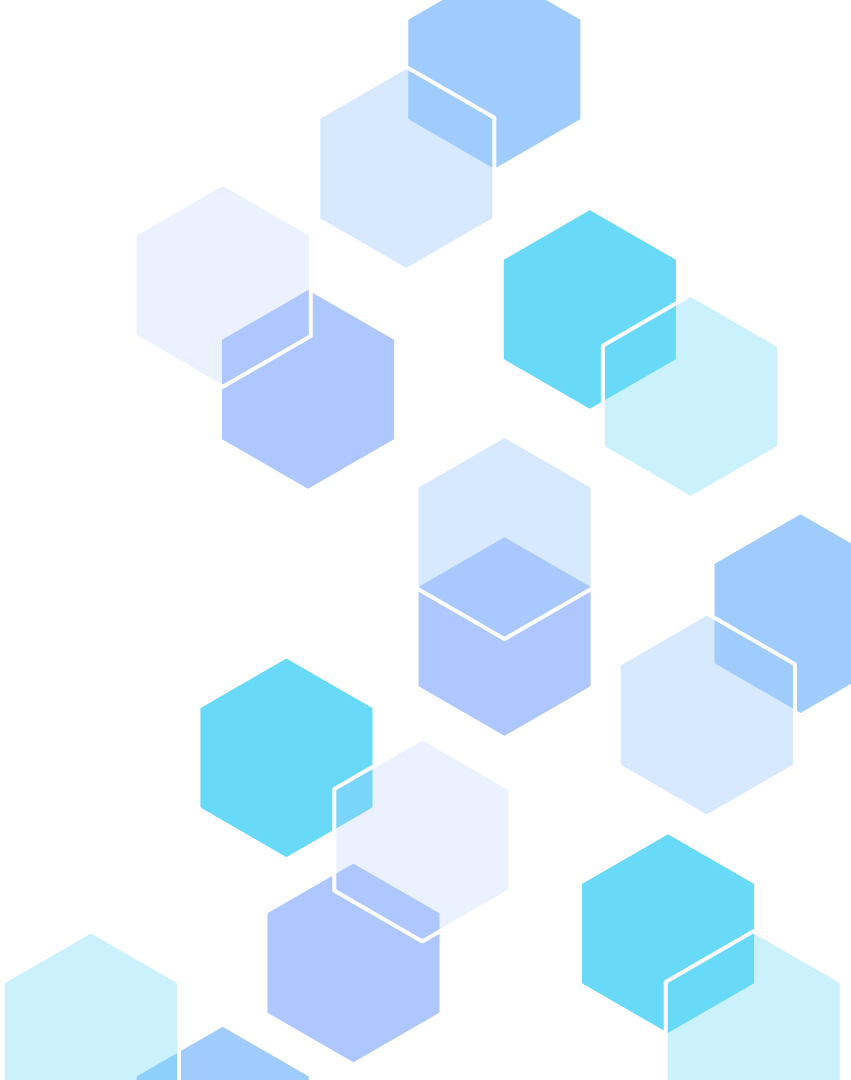
Applied on Dry Beans Dataset

Presented by:
Fatemeh Gol Pour
Kawtar Ezzati
Marieme Asselman Tafristan

Table of Contents

01

Introduction



Introduction

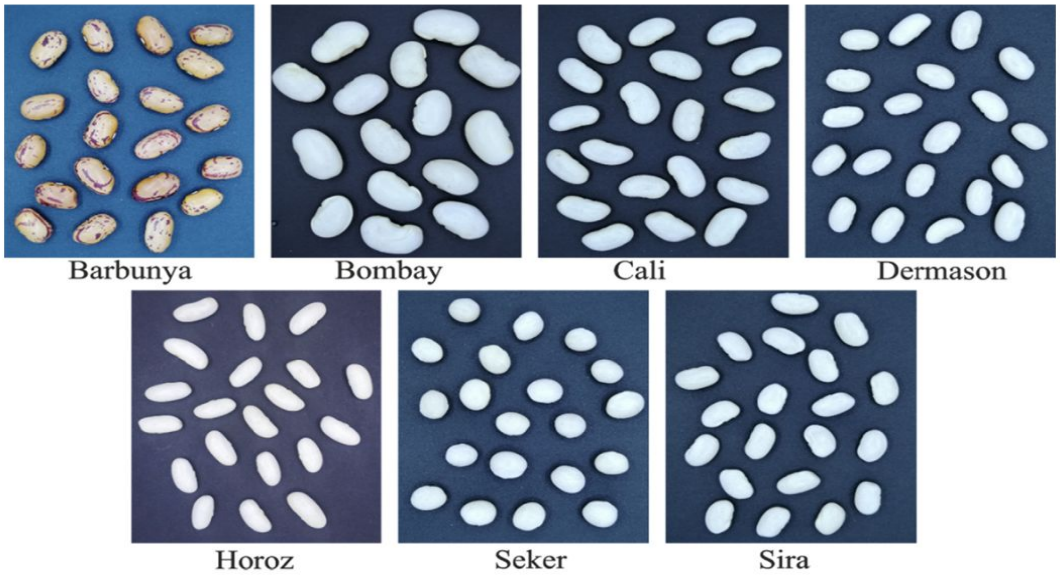
This project focuses on enhancing the performance of SVMs through a systematic approach involving data normalization, hyperparameter tuning, and rigorous evaluation. The project begins with acquiring a dataset relevant to the classification task .

02

Dataset Description

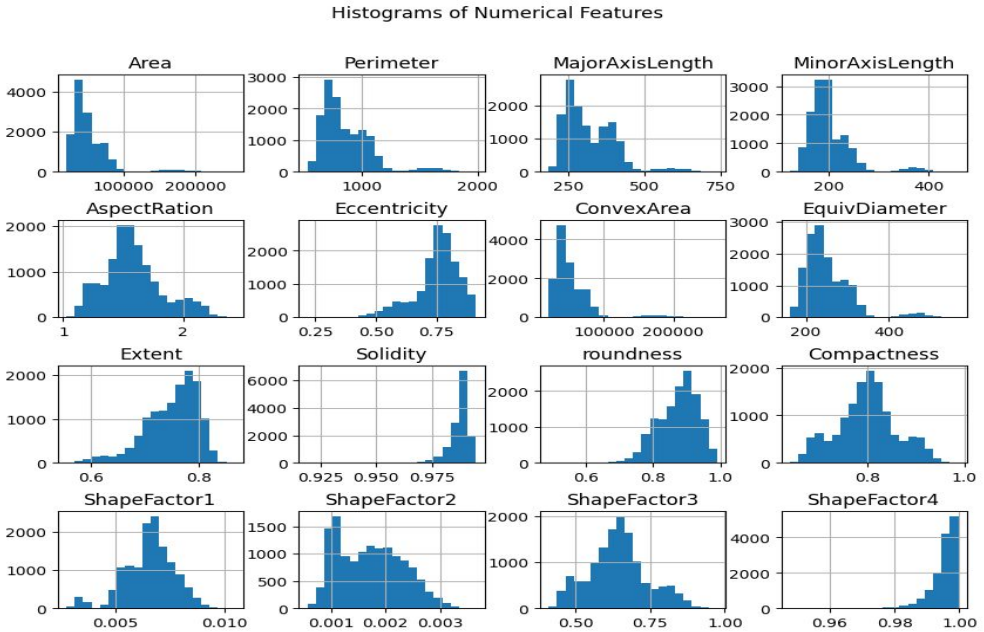
Dataset Description

This seven different types of dry beans were used to collect this data, taking into account the features such as form, shape, type, and structure.



Data Distribution

This histogram shows the distribution of numerical features



03

Tuning Hyperparameters

SVM Hyperparameters

Kernel Parameters(gamma): which controls the influence of individual training samples on the decision boundary.
Regularization Parameter (C): This hyperparameter controls the trade-off between maximizing the margin and minimizing the classification error.
Kernel Trick: handle non-linear decision boundaries.

Tuning Hyperparameters

To find the best combination of hyperparameters, we use Grid Search, which is a brute-force approach used in machine learning to find the optimal set of hyperparameters for a model.

Gamma	C	Kernel			
		Linear	Poly	RBF	Sigmoid
Auto	0.1	0.927	0.864	0.923	0.816
	1	0.926	0.906	0.929	0.729
	10	0.928	0.923	0.930	0.720
	100	0.927	0.927	0.930	0.718
Scale	0.1	0.927	0.864	0.923	0.816
	1	0.926	0.907	0.929	0.730
	10	0.928	0.923	0.931	0.719
	100	0.927	0.927	0.930	0.718

04

Model Evaluation

Metrics Used to Evaluate Model's Performance

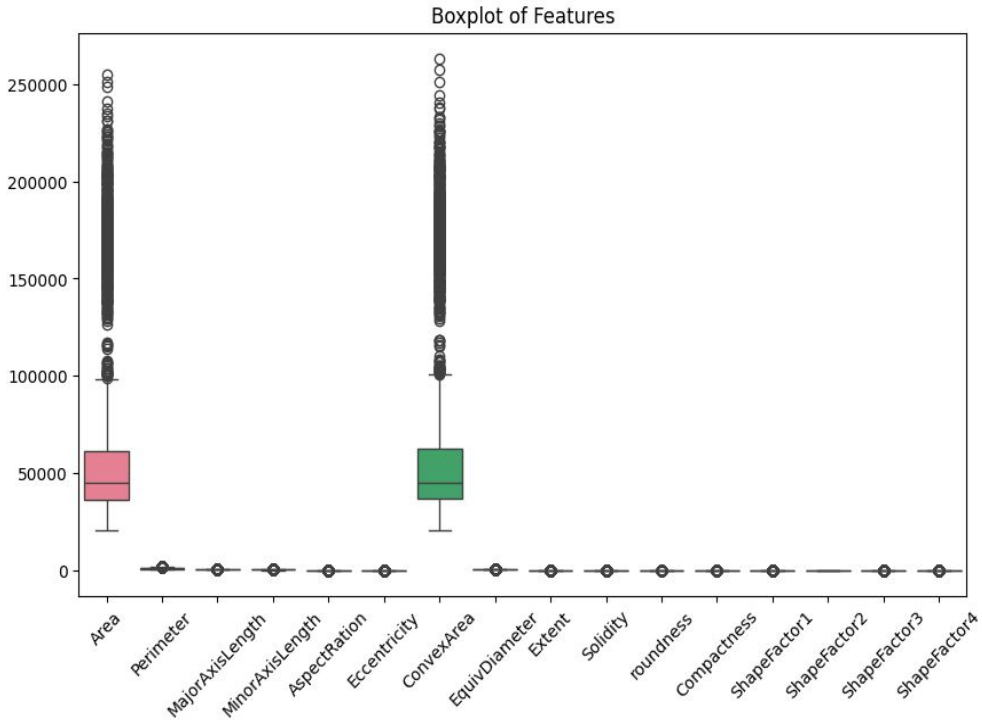
Metric	Formula	Value
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$	93.46 %
Precision	$\frac{TP}{TP+FP}$	93.51 %
Recall	$\frac{TP}{TP+FN}$	93.46 %
F1-score	$\frac{2 \times Precision \times Recall}{Precision + Recall}$	93.48 %

05

Challenges

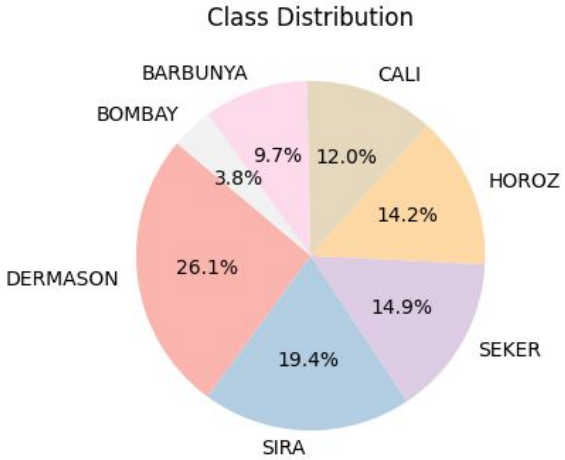
4.1. Outliers
4.2. Imbalance
4.3. Noise

Outliers

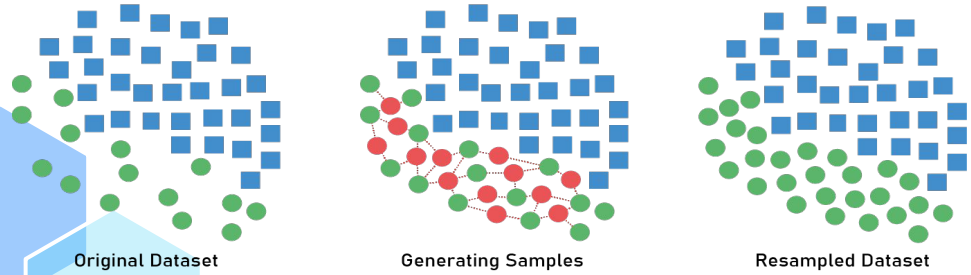


Metric	Value on original data	Value on cleaned data
Accuracy	93.46 %	87.77 %
Precision	93.51 %	84.84 %
Recall	93.46 %	87.77 %
F1-score	93.48 %	86.06 %

Imbalance - SMOTE



Synthetic Minority Oversampling Technique



Metric	Value on original data	Value on cleaned data
Accuracy	93.46 %	93.24 %
Precision	93.51 %	93.32 %
Recall	93.46 %	93.24 %
F1-score	93.48 %	93.26 %

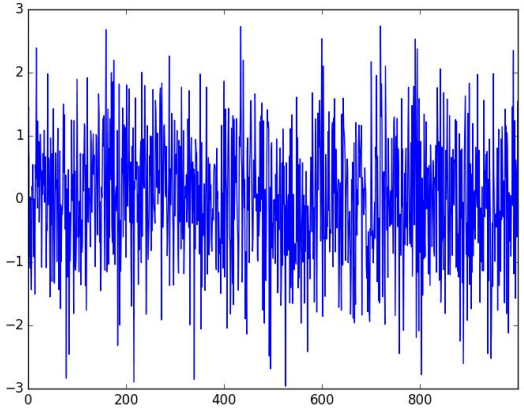
Imbalance - Class Weights

- Using class_weight parameter in SVM
- Assigning higher weights to minority classes
- Using 'balanced' mode automatically adjusts class weights inversely proportional to class frequencies
- Result: the values deteriorate

Metric	Value on original model	Value on weighted model
Accuracy	93.46 %	92.76 %
Precision	93.51 %	92.95 %
Recall	93.46 %	92.76 %
F1-score	93.48 %	92.80 %

Noise

- Dimensionality Reduction:
  - PCA
- Feature Engineering:
  - Polynomial Features



Noise – PCA

- PCA is used to denoise and reduce the dimensionality of the dataset
- Does not eliminate the noise but can reduce it
- Grid search to tune the number of components
- Number of components: 15
- Result : no improvement

Metric	Value on original model	Value on weighted model
Accuracy	93.46 %	93.46 %
Precision	93.51 %	93.51 %
Recall	93.46 %	93.46 %
F1-score	93.48 %	93.48%

Noise - Feature Engineering

- Trying to capture nonlinear relationships in the data generating polynomial features from existing features
- Can help improve the discriminatory power of the model and make it more robust to noise
- Result: deterioration

Metric	Value on original model	Value on weighted model
Accuracy	93.46 %	93.05 %
Precision	93.51 %	93.11 %
Recall	93.46 %	93.05 %
F1-score	93.48 %	93.08%

06

Summary

Summary

- The objective of this project
- Dry Beans dataset with 13,611 instances, 16 features, and
- one categorical target
- Created an SVM model trained and evaluated it
- Applied Grid Search to tune the hyperparameters
- Tried various approaches to address each constraint in our dataset:
  - Removing outliers
  - Removing imbalance in the classes
    - SMOTE
    - Class weights
  - Handling noise
    - PCA
    - Polynomial features
- No improvement in out results