

# Dry Beans Classification

IART G03 - Checkpoint

Ana Inês Oliveira de Barros - up201806593

João de Jesus Costa - up201806560

João Lucas Silva Martins - up201806436

# Problem specification

- **Task**

Classify between seven different registered varieties of dry beans with similar features, based on the features collected. The beans can be of any of the following classes: Seker, Barbunya, Bombay, Cali, Dermosan, Horoz, and Sira.

- **Experience**

A dataset with information collected about dry beans and their classification, with the following attributes: Area, Perimeter, Major axis length, Minor axis length, Aspect ratio, Eccentricity, Convex area, Equivalent diameter, Extent, Solidity, Roundness, Compactness, ShapeFactor1, ShapeFactor2, ShapeFactor3, ShapeFactor4.

- **Performance**

The classification accuracy, the classification precision, and the training/classification times.

# Related work and bibliographic search

- Previous solutions of the same problem
  - <https://github.com/NaitikJ/DryBean--Dataset>
  - [https://github.com/HimankSehgal/DSGRecruitmentTask\\_DryBeanDataset](https://github.com/HimankSehgal/DSGRecruitmentTask_DryBeanDataset)
- Data Analysis and Machine Learning Projects
  - <https://github.com/rhiever/Data-Analysis-and-Machine-Learning-Projects/blob/master/example-data-science-notebook/Example%20Machine%20Learning%20Notebook.ipynb>
- Performance metrics to classification problems
  - <https://www.kaggle.com/usengecoder/performance-metrics-for-classification-problems>
- Feature Selection Techniques
  - <https://pierpaolo28.github.io/blog/blog27/>
  - <https://www.kaggle.com/rxsraghavagrawal/feature-selection-techniques>
  - <https://www.kaggle.com/prashant111/comprehensive-guide-on-feature-selection>
- Select k best: feature selection example in python
  - <https://www.datatechnotes.com/2021/02/selection-best-feature-selection-example-in-python.html>
- Remove outliers in python
  - <https://www.statology.org/remove-outliers-python>

# Tools & algorithms


## Libraries & tools

Programming language: Python 3.9.4



Programming environment: Jupyter Lab



- matplotlib 3.4.1-2 
- numpy 1.20.2-1 
- pandas 1.2.3-1 
- scikit-learn 0.24.1-1 
- scipy 1.6.3-1 
- seaborn 0.11.1-1 

## Classifiers used

1. Decision trees
2. K-nearest neighbors
3. Support vector
4. Naive bayes
5. Random forest

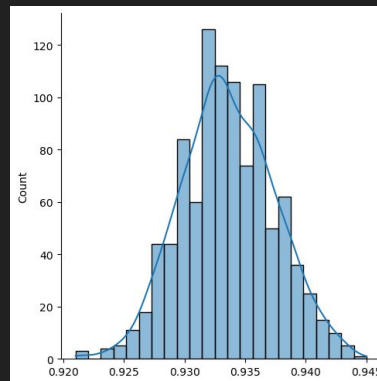


Figure 1- Decision tree results

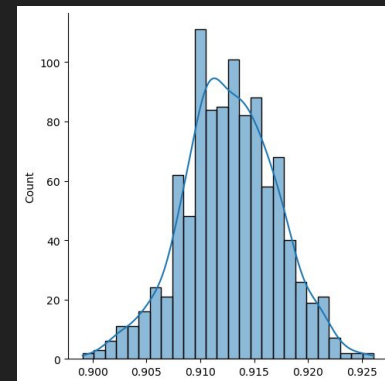


Figure 2- K-nearest neighbors results

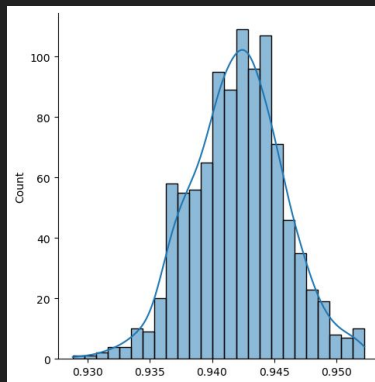


Figure 3- Support vector results

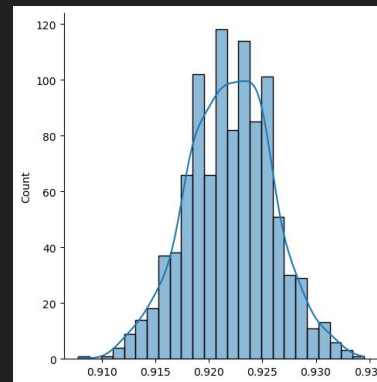


Figure 4- Naive bayes results

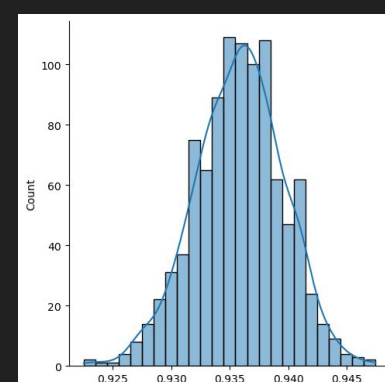


Figure 5- Random forest results

# Implemented Work

## 1. Data analysis

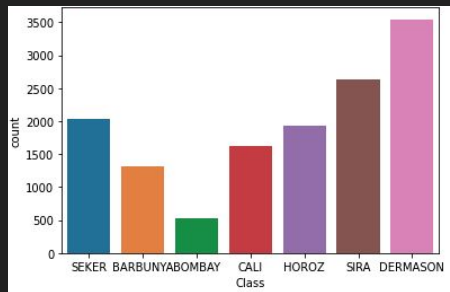


Figure 6 - Imbalance of bean classes in the data.

## 2. Outlier removal

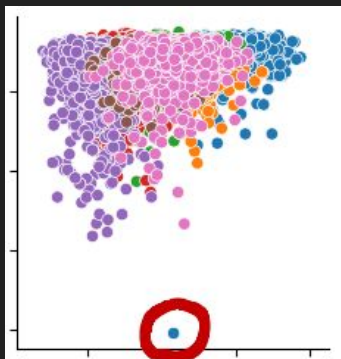


Figure 7 - Example of outlier.

## 3. Attribute selection

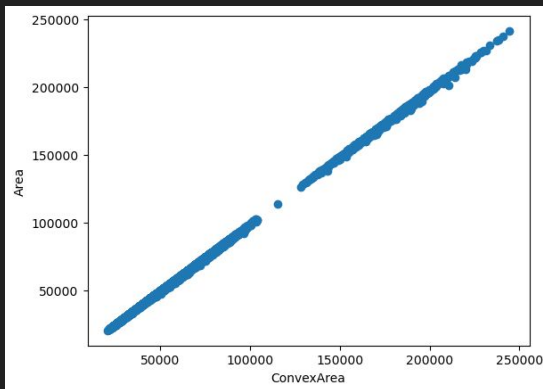


Figure 8 - Correlation between Area and ConvexArea.

## 4. Model comparison

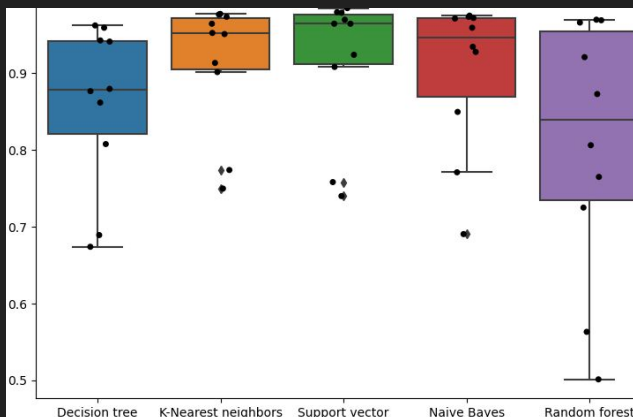


Figure 10 - Model accuracy comparison.

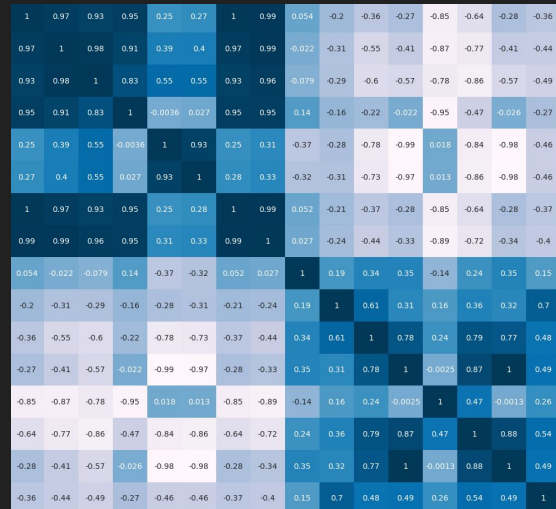


Figure 9 - Attribute correlation heatmap.

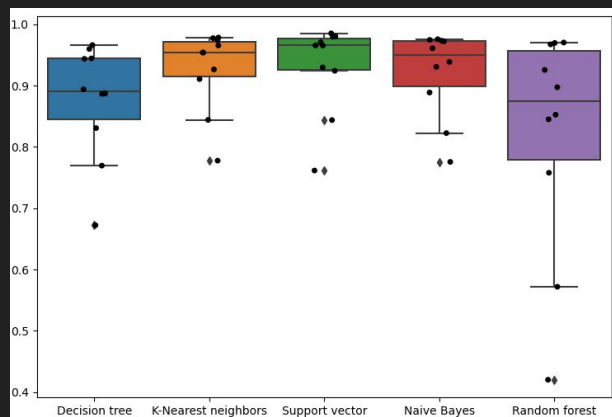


Figure 11 - Model precision comparison.