U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Supervised Learning

IART 2nd Assignment

Work specification

The selected theme for our project was the Dry Beans Dataset.

We are given a dataset which contains various features regarding dry beans.

Our goal is compare different models which will take in features of data beans and in turn will predict a bean's type, taking into account the features such as form, shape, type, and structure by the market situation.

For the classification model, images of 13,611 grains of 7 different registered dry beans were taken with a high-resolution camera. Bean images obtained by computer vision system were subjected to segmentation and feature extraction stages, and a total of 16 features, 12 dimensions and 4 shape forms, were obtained from the grains.

Attribute information

- Area (A): The area of a bean zone and the number of pixels within its boundaries.
- Perimeter (P): Bean circumference is defined as the length of its border.
- Major axis length (L): The distance between the ends of the longest line that can be drawn from a bean.
- Minor axis length (l): The longest line that can be drawn from the bean while standing perpendicular to the main axis.
- Aspect ratio (K): Defines the relationship between L and l.
- Eccentricity (Ec): Eccentricity of the ellipse having the same moments as the region.
- Convex area (C): Number of pixels in the smallest convex polygon that can contain the area of a bean seed.
- Equivalent diameter (Ed): The diameter of a circle having the same area as a bean seed area.
- Extent (Ex): The ratio of the pixels in the bounding box to the bean area.
- Solidity (S): Also known as convexity. The ratio of the pixels in the convex shell to those found in beans.
- Roundness (R): Calculated with the following formula: (4piA)/(P^2)
- Compactness (CO): Measures the roundness of an object: Ed/L
- ShapeFactor1 (SF1)
- ShapeFactor2 (SF2)
- ShapeFactor3 (SF3)
- ShapeFactor4 (SF4)
- Class (Seker, Barbunya, Bombay, Cali, Dermosan, Horoz and Sira):

Implementation work

The selected language to implement our solution was Python 3.

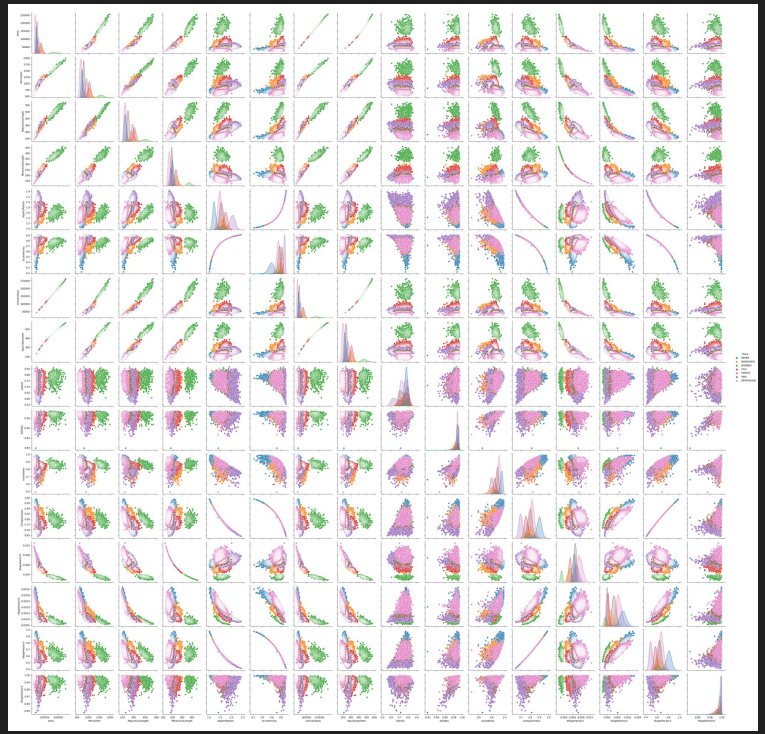
The development environment that was used was Jupyter Notebook.

So far, we have implemented Decision Tree, Support Vector Machine, K-nearest Neighbors, Naive Bayes and Random Forest models.

Both Support Vector Machines and K-Nearest Neighbours need the data to be scaled to give decent results so we used *Scikit-Learn*'s StandardScaler.

Data analysis and preprocessing

- We detected some outliers, but we were not provided confirmation on whether those outliers are to be expected or not.
- There's a difference factor of about 80% between the least and most common bean type occurrences.
- There are no missing values in the dataset.
- We analysed the correlation of features by constructing a correlation matrix and dropping the most correlated ones. The dropped features are the ones with less correlation with the *Class* column.



https://github.com/Educorreia932/IART-Dry-Beans/blob/master/src/out.png

Data filtering

Outlier detection

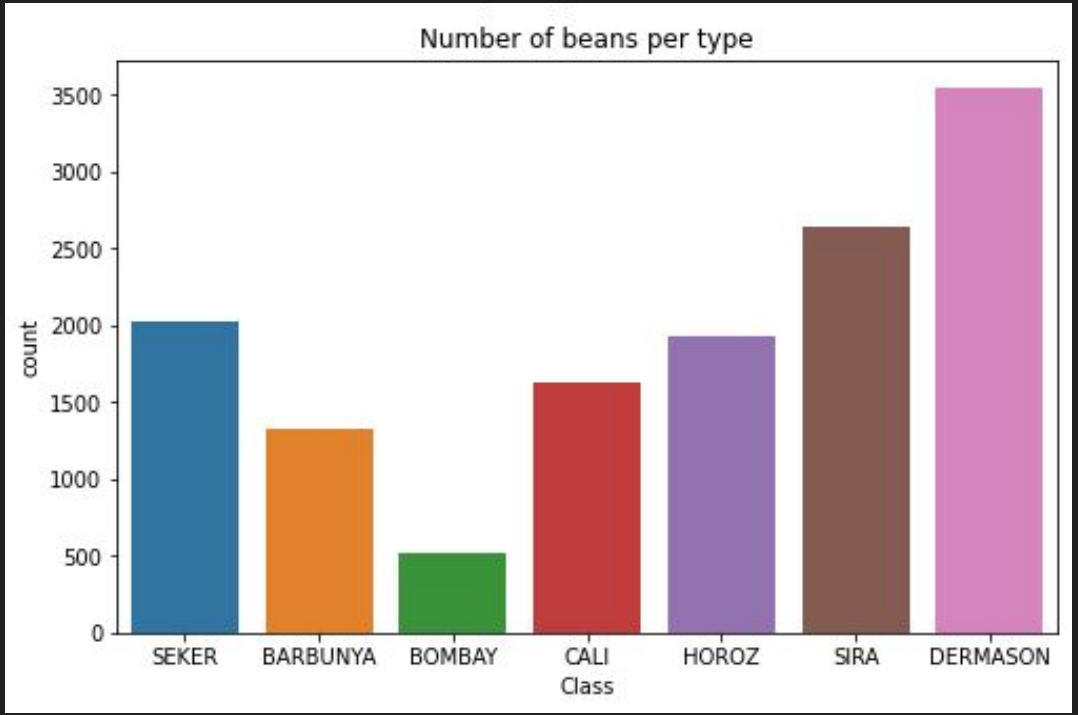
- By looking at the data, we realized there were a few outliers that could interfere with the solution's accuracy. We looked for different ways to detect them, however, each solution had different problems:
- Removing the data manually: We do not have the experience in the field nor confirmation from the data collectors to properly filter and remove outliers.
- Z-score: We cannot guarantee that the data follows a binomial distribution.
- Dbscan: Intended for unsupervised problems.
- Isolation forest: Hard to see data and takes a long time to use.

Null values

- There are no null values in our dataset.

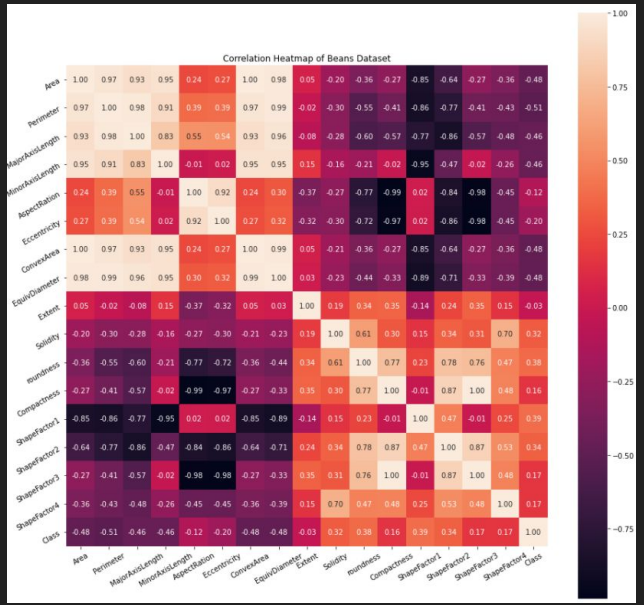
Data distribution

- There's a difference factor of about 6 between the least and most common bean type occurrences. This means that there is (about) 6 times as much data of the most common bean (Dermason) than the least common (Bombay).
- Because of these, we experimented both with and without oversampling on the dataset to balance it and possibly obtain better results.



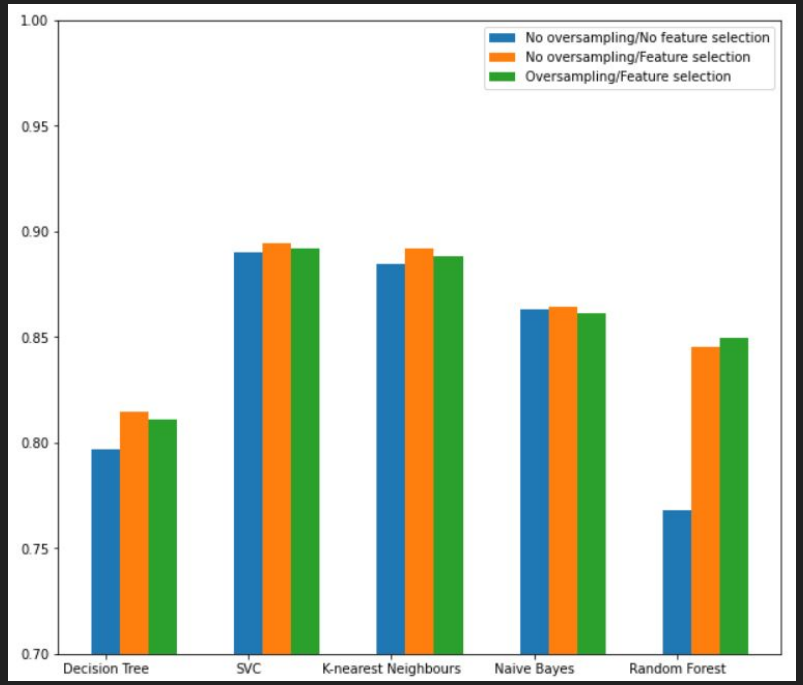
Correlation

- We analysed the correlation of features by constructing a correlation matrix and dropping the most correlated ones. The dropped features are the ones with less correlation with the *Class* column.
- The following features were dropped:
  - ShapeFactor3
  - Compactness
  - Aspect ratio
  - Area
  - ShapeFactor1
  - Major axis length
  - Minor axis length
  - Convex data
  - Equivalent diameter

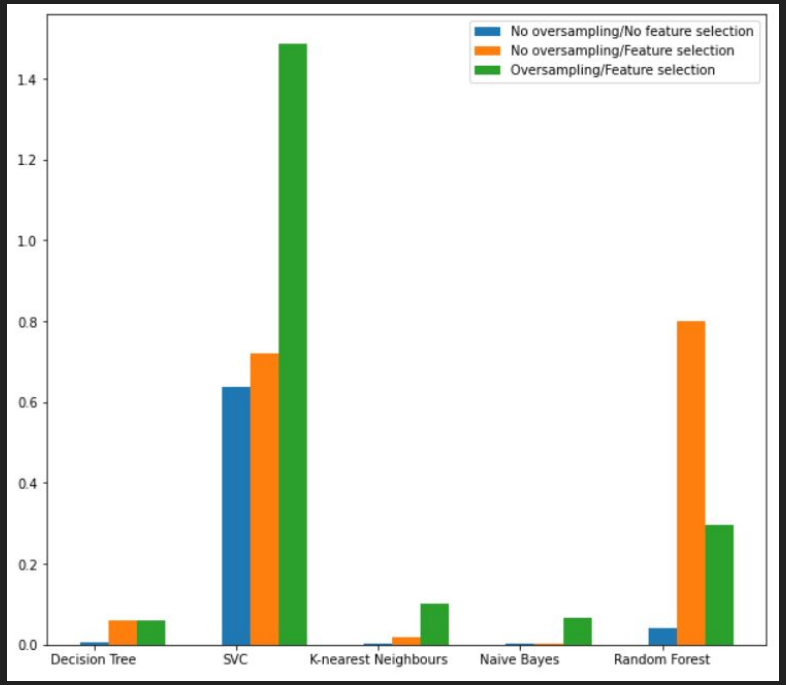


Model comparison

Best score



Time to train



Related work

- HimankSehgal/DSGRecruitmentTask_DryBeanDataset
- NaitikJ/DryBean--Dataset

References

- Improving Classification by Outlier Detection and Removal
- Comprehensive Guide on Feature Selection
- Multiclass classification of dry beans using computer vision and machine learning techniques