amse
école d'économie d'aix-marseille
aix-marseille school of economics

Aix*Marseille
université
Socialement engagée

Classification of Dry Beans Using Grain Features

Master 2 Econometrics, Big Data and Statistics

Authors:

Felicia Sahi, Jean-Baptiste Gomez & Nelly Agossou

Course: Advanced methods in big data

Professors **S. Hué, P. Michel**

Academic year 2023-2024

Summary

Illustration table

1. Introduction

The agricultural sector plays a vital role in the global economy, contributing to food security, economic development and environmental protection. As the main source of food production, it contributes to the gross domestic product (GDP) of many countries and fuels a variety of industries, notably agrifood. The sustainability of agriculture has become a major concern, with growing emphasis on responsible resource use, biodiversity conservation and adaptation to the challenges of climate change. Technological advances, such as biotechnology and artificial intelligence, are increasingly integrated to improve productivity and optimize farming practices. International trade in agricultural products also makes a significant contribution to the global economy. More specifically, agriculture plays a central role in everyday life, ensuring the availability of foodstuffs, promoting economic development and helping to preserve our environment.

Among the most important and widely grown legumes in the world are dried beans. The latter plays a significant role in Turkish agriculture. However, the plant is sensitive to climate change. Determining the best seed is a major challenge for dry bean growers, as the use of inferior seed can result in reduced production. Turkey's dry bean varieties, such as Barbunya, Bombay, Calı, Dermason, Horoz, Seker, and Sira, are classified according to botanical characteristics by the Turkish Standards Institute.

In Turkey, around 10% of dry bean seeds used are certified. The cultivation of dry beans in populations containing mixed seed species poses challenges, as marketing these seeds without separation by species diminishes their market value.

Taxonomic determination of dry bean varieties requires specialized expertise and is a time-consuming process. When morphological similarity between seed lots is predominant, manual classification becomes a delicate undertaking. It should be noted that the task of interpreting or manipulating these seeds by a human operator, lacking specific tools, reaches near impracticality.

In today's world, studies have focused on the use of image processing and machine learning techniques for the classification of dry bean seeds. Various methods, such as color, morphology and shape analysis, have been employed to extract characteristics from seeds. These automatic methods have become essential due to the difficulty and inefficiency of manual sorting, and to guarantee high-quality food products.

The relevance of using artificial intelligence techniques to classify the different varieties of dried beans lies in the fact that they enable:

- Improve the accuracy and speed of the grading process, reducing the errors and costs associated with manual sorting.
- Exploit visual seed data, which is often difficult to analyze using conventional methods.
- Adapt classification to the specific needs of producers, consumers and markets, considering quality, preference and profitability criteria.
- Innovate in the creation of new dry bean varieties, using computer-aided design and biological synthesis techniques.

This sums up that artificial intelligence offers opportunities for adding value to dried beans, which are an important crop for Turkish and global agriculture. Among the classification works that have been tackled using AI, we find: [Vibhute and Bodhe, 2012][1], their paper focuses on the study of image processing applications in agriculture, such as imaging techniques, weed detection, and fruit grading. The authors show that parameter analysis using image processing proves to be more accurate and less time-consuming than traditional methods. They claim that the application of image processing can improve decision-making for vegetation measurement, irrigation, and fruit sorting. Similarly [Murat & Ilker,2020][2], their paper proposes a multi-class classification method for dry beans using computer vision and machine learning techniques. It evaluates four algorithms (MLP, SVM, kNN, and DT) by comparing their performance on 16 features. The results show that the SVM model achieves the best overall classification rate at 93.13%, meeting both growers' and consumers' requirements for uniform bean varieties.

From the work reviewed in the literature, we really understand the impact of classification and the use of AI to guarantee increased traceability and optimal product quality. This is achieved by ensuring continuous, automated control, as well as providing accurate information to consumers. The result is increased data value, fostering knowledge creation, innovation, competitiveness, and encouraging collaboration between players in the agricultural sector. The research problem naturally emerges: How can we optimize the performance of modern machine learning algorithms in the classification of dry beans by effectively taking advantage of the morphological characteristics of the beans? This approach aims to reconcile technological

[1] [Anup Vibhute & S K Bodhe, 2012]. “Applications of Image Processing in Agriculture: A Survey”, International Journal of Computer Applications (0975 – 8887) Volume 52– No.2

[2] [Murat & Ilker,2020]. “Multiclass classification of dry beans using computer vision and machine learning techniques”. Computers and Electronics in agriculture, volume 174.

advances to guarantee accurate and adaptable classification, while meeting the desired traceability and quality imperatives, as identified in the literature.

To answer the problem formulated, we exploited a database extracted from "UCI Machine Learning". This database included 13,611 high-resolution images of 7 dry bean varieties, captured with a camera. A computer vision system was applied to develop a classification model distinguishing the varieties, extracting 16 morphological features such as "Area", "Perimeter" and Shape Factors (SF1-SF4). Through data pre-processing, we ensured data cleanliness before applying various machine learning and deep learning algorithms, including Elastic Net Logistic, Random Forest, GradientBoosting, Neural Network, SVM, KNN Classifier, XGBoost, Decision Tree, Naive Bayes, SGD Classifier, AdaBoost and Ridge Classifier. The results obtained and conclusions drawn from this study will be discussed in the rest of our work.

2. Related works and motivations

In this section, we will draw on the body of work carried out on bean classification using machine learning and artificial intelligence techniques. The results obtained from this work will be of great use in drawing conclusions about the model we will apply in our study.

According to research by [Md S Khana & al, 2023][3], eight popular classifiers, such as logistic regression (LR), naive Bayesian (NB), k-nearest neighbor (KNN), decision tree (DT), random forest (RF), extreme gradient boosting (XGB), support vector machine (SVM) and multilayer perceptual (MLP), were evaluated. Results indicated that the XGB classifier performed well, particularly with balanced and unbalanced class distributions. The KNN and RF approaches, in combination with the XGB technique, also showed increased efficiency in the case of well-balanced classes.

In addition, [Jaime C & al, 2023][4] adopted a data mining approach, integrating principal component analysis, hyperparameter optimization, and sample balancing with the SMOTE technique. The KNN model, with adjusted hyper-parameters and the application of SMOTE,

[3] [Md S Khana , Tushar Deb Nathb , Md Murad Hossainc , Arnab Mukherjee d , Hafiz Bin Hasnatha , Tahera Manhaz Meema , Umama Khane, 2023]. « Comparison of multiclass classification techniques using dry bean dataset », International Journal of Cognitive Computing in Engineering 4 (2023) 6–20.

[4] [Jaime Carlos Macuácua, Jorge António Silva Centeno, Caísse Amisse, 2023]. "Data mining approach for dry bean seeds classification », Smart Agricultural Technology 5.

achieved 95% accuracy. For the RF and SVM classifiers, principal component analysis and SMOTE balancing were more appropriate, significantly improving accuracy.

In addition, [Md. Mahadi H & al, 2021][5], explored a deep neural network-based approach to automatically classify dry beans, achieving 93.44% accuracy and a 94.57% F-1 score. Their results demonstrate a significant improvement over the traditional machine learning approach.

In the study by [Murat & Ilker,2020][6], entitled "Multiclass classification of dried beans using computer vision and machine learning techniques", multilayer perceptron (MLP), support vector machine (SVM), k-Nearest Neighbors (kNN) and decision tree (DT) models were developed with 10-fold cross-validation, and performance was compared. Correct classification rates were evaluated at 91.73%, 93.13%, 87.92% and 92.52% for MLP, SVM, kNN and DT respectively. The SVM model, with the most accurate results, achieved classification rates of 92.36%, 100.00%, 95.03%, 94.36%, 94.92%, 94.67% and 86.84% for the Barbunya, Bombay, Cali, Dermason, Horoz, Seker and Sira bean varieties respectively. These results demonstrate considerable satisfaction of growers' and customers' demands for uniform bean varieties.

Another study conducted by [Islam et al. in 2017][7], titled “Potato disease detection using image segmentation and multiclass support vector machine”, the researchers presented an innovative approach combining image processing and machine learning to diagnose diseases from images of potato leaves. This automated method successfully classified diseases, or determined the absence of diseases, on potato plants using a plant image database called “PlantVillage”. Thanks to their segmentation approach and the use of a support vector machine, the researchers obtained impressive classification results, achieving an accuracy of 95% on more than 300 images. These results demonstrate the feasibility of an automated diagnosis of plant diseases on a large scale.

[5] [Md. Mahadi Hasan ; Muhammad Oussama Islam; Muhammad Jafar Sadeq, 2021], “A Deep Neural Network for Multi-class Dry Beans Classification”. 24th International Conference on Computer and Information Technology (ICCIT).

[6] [Murat & Ilker,2020]. “Multiclass classification of dry beans using computer vision and machine learning techniques”. Computers and Electronics in agriculture, volume 174.

[7] [Monzurul Islam,Anh Dinh, Khan Wahid, Pankaj Bhowmik, 2017]. “Détection de maladies de la pomme de terre à l’aide d’une segmentation d’images et d’une machine à vecteurs de support multiclasses”. IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCGECE).

The empirical study conducted by [Ajay et al.][8] in 2022, entitled "Wheat seed classification: using the ensemble machine learning approach", implemented machine learning and computer vision techniques to classify wheat seeds. Various models, such as multilayer perceptron (MLP), support vector machine (SVM), k-Nearest Neighbors (kNN) and decision tree (DT), were developed and evaluated. Correct classification rates reached 91.73%, 93.13%, 87.92% and 92.52% respectively for MLP, SVM, kNN and DT. In particular, the SVM model stood out with an accuracy of 93.13%, demonstrating the effectiveness of integrating computer vision and machine learning for the accurate classification of wheat seed varieties.

The empirical evaluation of [Kumari et al.2021][9], an ensemble approach to diabetes classification used a variety of state-of-the-art classifiers. Results showed that the ensemble approach outperformed other methodologies, achieving 79.04% accuracy, 73.48% precision, 71.45% recall and an F1 score of 80.6% on the PIMA dataset. In addition, this methodology was effectively applied to breast cancer, achieving 97.02% accuracy. These findings underline the effectiveness of the ensemble approach in the classification and prediction of diabetes.

3. Some statistical data on dry bean production

According to a 2021 report[10], the dry beans market is expected to register a compound annual growth rate (CAGR) of 4.5% over the period 2023-2028. Dry beans are an ideal source of protein, particularly in regions where meat and dairy products are not accessible for geographical or economic reasons. In 2020, average consumption of dried beans worldwide was estimated at 2.4 kg per capita per year, with wide variations between continents: Latin America, 10.1 kg, North America, 5.9 kg, Africa, 2.4 kg, Asia, 1.4 kg, Europe, 0.8 kg3. India, Brazil, the United States and China are the main producers and consumers of dried beans.

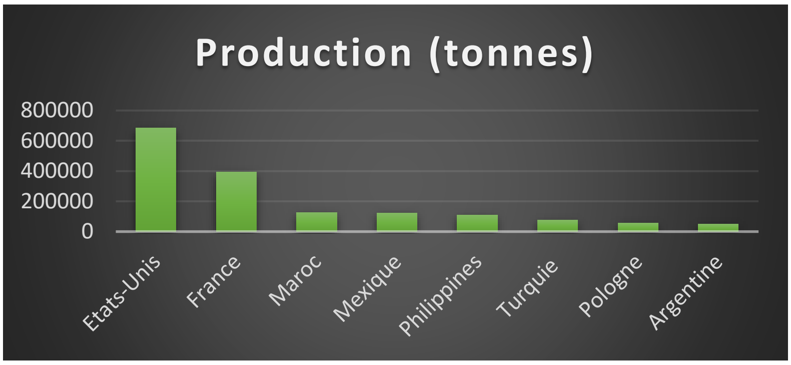
Figure 1 Dry beans production in the world[11]

[8] [Ajay Khatri,Shweta Agrawal, & Jyotir M. Chatterjee, 2022]. “Wheat Seed Classification: Utilizing Ensemble Machine Learning Approach”. Scientific Programming, Article ID 2626868, 9 p.

[9] [Saloni Kumari, Deepika Kumar & Mamta Mittal,2021]. “An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier”, international journal of cognitive computing in engineering. Vol 2, pages 40-46.

[10] Analyse de la taille et de la part du marché des haricots secs - Rapport de recherche de lindustrie - Tendances de croissance (mordorintelligence.com).

[11] Production mondiale de haricots verts par pays - AtlasBig.com

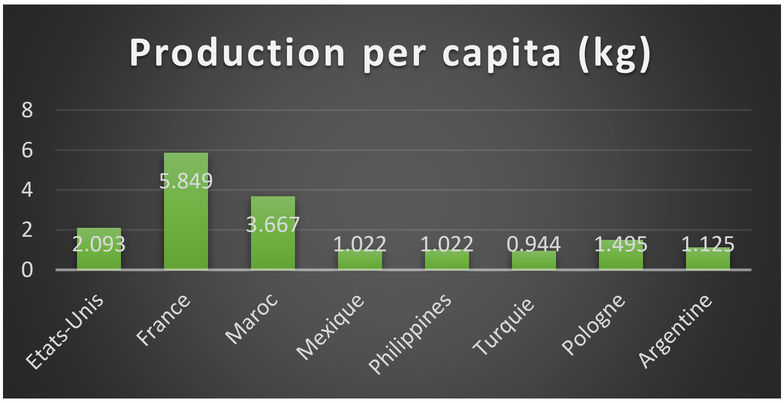


From the article "Turkey - Dry Bean - Market Analysis, Forecast, Size"[12], we concluded that in 2022, dry bean production fell slightly by -4.2%, reaching a volume of 292,000 tonnes. However, over the period from 2017 to 2022, production showed positive average annual growth of +4.1%. This upward progression was marked by notable fluctuations, including a remarkable 24% growth in 2020.

Production peaked at 305,000 tonnes in 2021, followed by a slight decline the following year. This dynamic was influenced by a significant expansion in the cultivated area, while yields showed a relatively stable trend.

In terms of value, dry bean production expanded significantly, reaching $297 million in 2022, estimated at export price. The total value of production recorded an average annual growth of +3.1% from 2017 to 2022, with notable fluctuations over this period. The most marked growth was observed in 2020, with an increase of 47% on the previous year. These statistics underline the importance of this data in the analysis of the dry bean sector, and can be relevantly integrated into our project to enrich our understanding of the market.

Figure 2: production per capita



[12] Production of Dry Bean in Turkey - 2022 - Charts and Tables - IndexBox

4. Base de données

In this project, our objective is to predict the type of dry beans by leveraging characteristics such as shape and structure, while considering market conditions. So, our dataset from Dry Bean Dataset   UCI Machine Learning Repository in UCI machine learning Website. The classification model was developed utilizing a high-resolution camera to capture images of 13,611 grains representing seven registered dry bean varieties. Bean images, acquired through a computer vision system, underwent segmentation and feature extraction stages, resulting in a total of 16 features (12 dimensions and 4 shape forms) extracted from the grains. Various machine learning and deep learning algorithms were employed to classify the seven most well-known types of beans in Turkey: Barbunya, Bombay, Cali, Dermason, Horoz, Seker, and Sira.

List and description of variables:

Name	Type	Values	Description
Area (A)	Integer	From 20420 to 254616	The area of a bean zone and the number of pixels within its boundaries.
Perimeter (P)	Continuous	From 524.74 to 1985.37	Bean circumference is defined as the length of its border.
MajorAxisLength (L)	Continuous	From 183.60 to 738.86	The distance between the ends of the longest line that can be drawn from a bean.
MinorAxisLength (l)	Continuous	From 122.51 to 460.19	The longest line that can be drawn from the bean while standing perpendicular to the main axis.
AspectRatio (K)	Continuous	From 1.02 to 2.43	Defines the relationship between MajorAxisLength(L) and MinorAxisLength(l).
Eccentricity (Ec)	Continuous	From 0.21 to 0.91	Eccentricity of the ellipse having the same moments as the region.
ConvexArea (C)	Integer	From 20684 to 263261	Number of pixels in the smallest convex polygon that can contain the area of a bean seed.
EquivDiameter (Ed)	Continuous	From 161.24 to 569.37	Equivalent diameter: The diameter of a circle having the same area as a bean seed area.
Extent (Ex)	Continuous	From 0.55 to 0.86	The ratio of the pixels in the bounding box to the bean area.
Solidity (S)	Continuous	From 0.91 to 0.99	Also known as convexity. The ratio of the pixels in the convex shell to those found in beans.
Roundness (R)	Continuous	From 0.48 to 0.99	Calculated with the following formula: (4* pi * A)/(P^2)
Compactness (CO)	Continuous	From 0.64 to 0.98	Measures the roundness of an object: Ed/L
ShapeFactor1 (SF1)	Continuous	From 0.002 to 0.010	L/d
ShapeFactor2 (SF2)	Continuous	From 0.000 to 0.003	l/d

ShapeFactor3 (SF3)	Continuous	From 0.41 to 0.97	4A/(L^2 * pi)
ShapeFactor4 (SF4)	Continuous	From 0.94 to 0.99	4A/(L* l * pi)
Class	Nominal	It can be any of BARBUNYA, SIRA, HOROZ, DERMASON, CALI, BOMBAY, and SEKER.	The class of the bean.

5. Méthodologie

The objective of our project is classified based on several criteria: dry beans grains. To do this, we will use several classification methods/models that we will compare to choose the best one, i.e. the one that allows us to obtain the best forecast.

- **Elastic net for logistic regression**

Elastic Net is a regularization technique implemented by Hui Zou and Trevor Hastie (2005) and used in machine learning. It combines lasso regression penalties ($L_1 = \lambda \sum_{i=1}^{n} |w_i|$) and ridge ($L_2 = \lambda \sum_{i=1}^{n} |w_i^2|$) to overcome their respective limitations. The term L_1 favors the parsimony of the coefficients by forcing some of them to be exactly equal to zero, while the term L_2 penalizes the overall magnitude of the coefficients. This approach is particularly useful when we are working with many characteristics, some of which are potentially correlated. The goal is to perform both the selection of variables and the regularization of the coefficients.

- **Random forest**

The random forest algorithm propose by Leo Breiman (2001) excels at manipulating multiple classes and can provide accurate and robust predictions even in complex datasets. Using a multitude of decision trees, he can understand the nonlinear relationships between characteristics and target classes. Each tree is trained on a random subset of the data, and at each decision node, only certain characteristics are considered, thus avoiding overcorrelation between trees. We chose this model because it provides a measure of feature importance, which can be particularly useful in our context to identify the variables that contribute the most to prediction.

- **Gradient Boosting**

This method combines several weak models to create a more performant model. It is an iterative algorithm that, at each step, builds a new model to correct errors in the previous model.

The principle is as follows: we start by building a weak model, for example a decision tree. This model is then used to predict the values of the target variable. The prediction errors are then calculated and used to build a new model. This new model is trained to correct errors from the previous model. The whole process is repeated until a satisfactory model is obtained. Gradient boosting is a very effective method for solving classification problems and, as far as we are concerned, multivariate classification.

- **Neural Network**

By using interconnected neural layers, neural networks can capture complex, nonlinear relationships between input features and output classes. During training, the weights of the connections between neurons are iteratively adjusted to minimize the discrepancy between network predictions and actual values. We will use them by configuring them with multiple layers and activation functions (RELU and Linear) to efficiently process multiple classes. In summary, neural networks represent a powerful and adaptable approach to multivariate classification, harnessing flexibility, and deep learning capability to deliver accurate results.

- **Support Vector Machine (SVM)**

The SVM is a supervised machine learning method developed by Vladimir Vapnik (1995) and used for many tasks, including classification. The basic idea of SVM is to find a hyperplane that separates data from multiple classes. The hyperplane is defined by a direction vector and a constant. We chose to use this model because it is very efficient, robust to noise and outliers.

- **eXtreme Gradient Boosting (XGBoost)**

XGBoost is a method that combines multiple decision trees to create a more performant model. It is a variant of gradient boosting, which uses a different approach than traditional gradient boosting to calculate prediction errors. It uses a second-order approximation of the loss function. This allows XGBoost to be more accurate and efficient than traditional gradient boosting.

- **K-Nearest Neighbours**

KNN (k-nearest neighbours) is a classification method based on spatial proximity in space of features. It used to solve classification and regression problems (Mukherjee et al., 2022). In multivariate classification, KNN classifies a point based on most of the labels of its k-nearest neighbors. Neighbors are determined by measuring the distance between points. KNN is simple, robust, but can be sensitive to the high dimensionality of the data. The judicious choice of k is crucial to obtain accurate results in a multivariate context.

- **Decision Tree**

The decision tree model is classification algorithms that segment feature space based on feature thresholds, creating a tree structure for making decisions. These models are particularly suitable for multivariate classification, as they can handle multiple output classes and capture complex relationships between features.

- **Naive Bayes**

The Naive Bayes model is based on Bayes' theorem and assumes the conditional independence of caractéristiques.il is based on the calculation of conditional probabilities to assign classes to observations based on their characteristics. We chose this model because it is particularly effective for high-dimensional datasets and can be quick to train.

- **Stochastic Gradient Descent (SGD)**

SGD is an optimization algorithm used to minimize a cost function and adjust the parameters of a model. We chose to use it because it treats training examples one by one randomly (stochastic) rather than deterministically. Unlike classic Gradient Descent, which uses the full set of training data for each parameter update, SGD uses only one training example at a time randomly. This makes the algorithm faster, especially with large data sets.

- **Ridge Classifier**

The Ridge classifier, also known as Ridge logistic regression, is a variant of the classical logistic regression algorithm that incorporates Ridge-like regularization (L_2). This regularization is added to the cost function of the logistic regression to penalize the coefficients of the least important characteristics, thus helping to avoid overfitting the model. In other words, the term regularization L_2 acts as a filter that can help improve the generalizability of the model to unseen data.

- **Adaptative Boosting (AdaBoost)**

AdaBoost is a supervised learning meta-algorithm that combines multiple weak classifiers to create a strong classifier. The algorithm works by iteratively training a set of weak classifiers and assigning weights to each data point based on its performance. The weights are then used to train the next classifier, which focuses on the data points that were misclassified by the previous classifier. This process is repeated until a desired number of classifiers have been trained or the error rate converges. We chose this model because it is relatively robust to noise in the data, which can make it a good choice for problems where there is some uncertainty in the data.

However, except for the random forest and net elastic models, we chose to reduce dimensionality before training the data. This is mainly due to the presence of several variables, some of which are highly correlated. With this in mind, we used **principal component analysis (PCA)** to narrow down our dataset.

Indeed, PCA is a statistical technique that allows you to reduce the dimensionality of data while preserving as much variance as possible. This method transforms a set of correlated variables into a new set of uncorrelated variables called principal components. These principal components capture most of the variability in the data, making them easier to interpret and analyze. The process of PCA involves calculating the eigenvectors and eigenvalues of the covariance matrix of the data. Eigenvectors define the directions of the new variables, while eigenvalues measure the importance of these directions in terms of data variability.

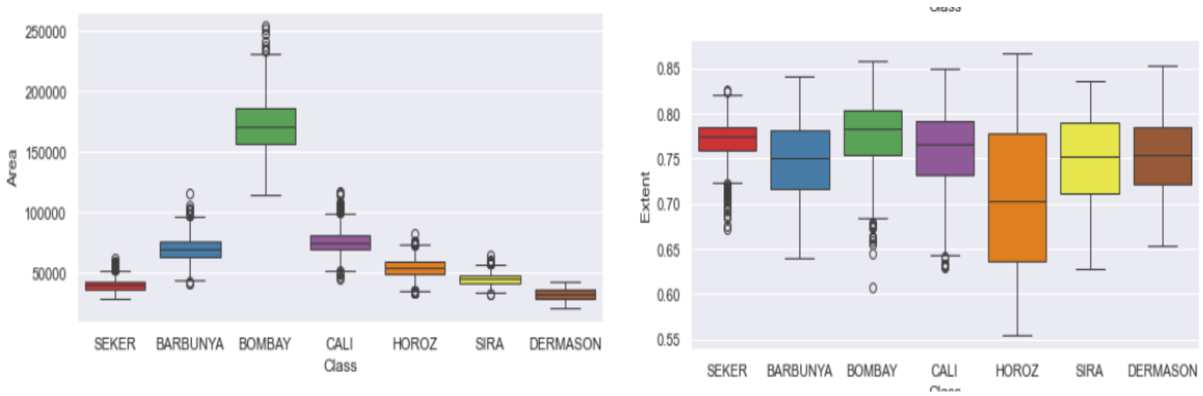
6. Results

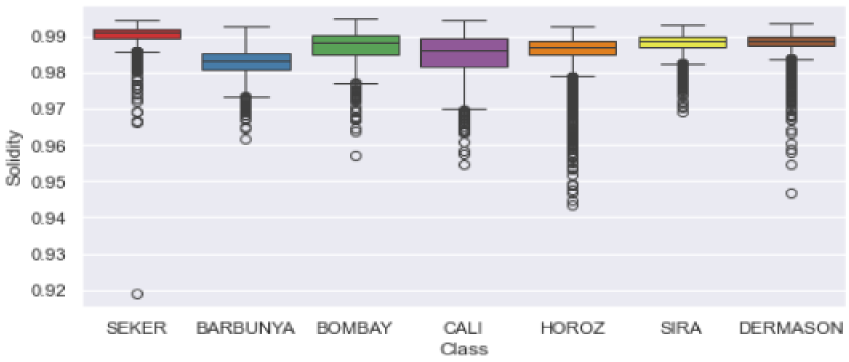
Recall that we worked on a database of 13611 observations and 17 variables. To begin, we will present the results of the Exploratory Data Analysis (EDA).

a. Explanatory Data Analysis (EDA)

Several operations were carried out on the basis as part of this analysis, namely: checking for missing data, removing duplications, descriptive analysis and checking outliers with a boxplot of numerical variables for each class of dry beans.

Figure 3: Boxplots of some variables for each class

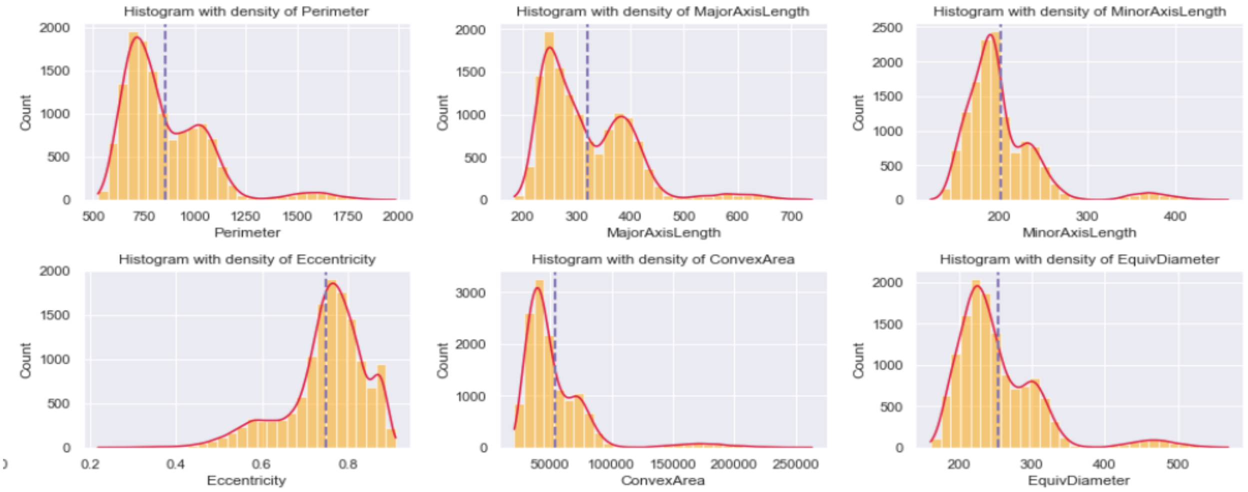




As shown in the figure above, The **Bombay** and **Horoz** classes show notable differences that set them apart from the others. Boxplot analysis reveals a considerable number of outliers in attributes such as **Solidity**, **Roundness** and **ShapeFactor4**. On the other hand, the **Extent** attribute shows a significantly lower frequency of outliers.

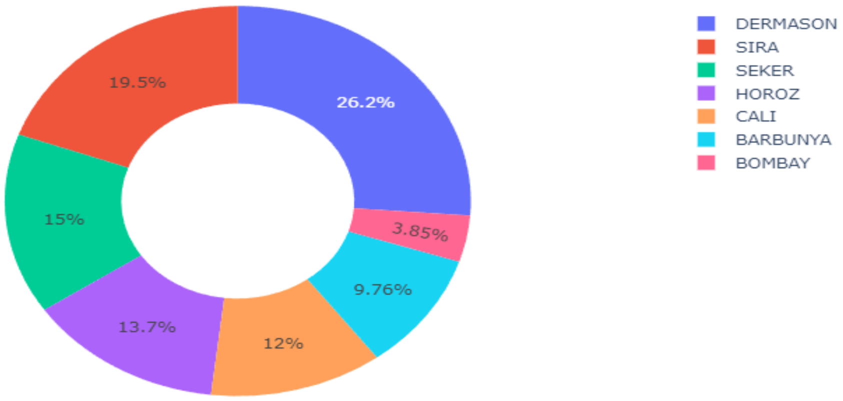
In addition, by doing the univariate analysis with the cumulative distribution of variables for each bean class, we notice that several distributions have elongated tails, and a majority has a **bi-modal** pattern, indicating distinct characteristics between the different bean classes. The figure below depicts the univariate distribution of a few variables within individual classes.

Figure 4 Cumulative Distribution of numerical attributes for the complete set of classes



However, we also experienced an uneven distribution between dry bean classes as shown in the following figure:

Figure 5: Comparison of class frequencies

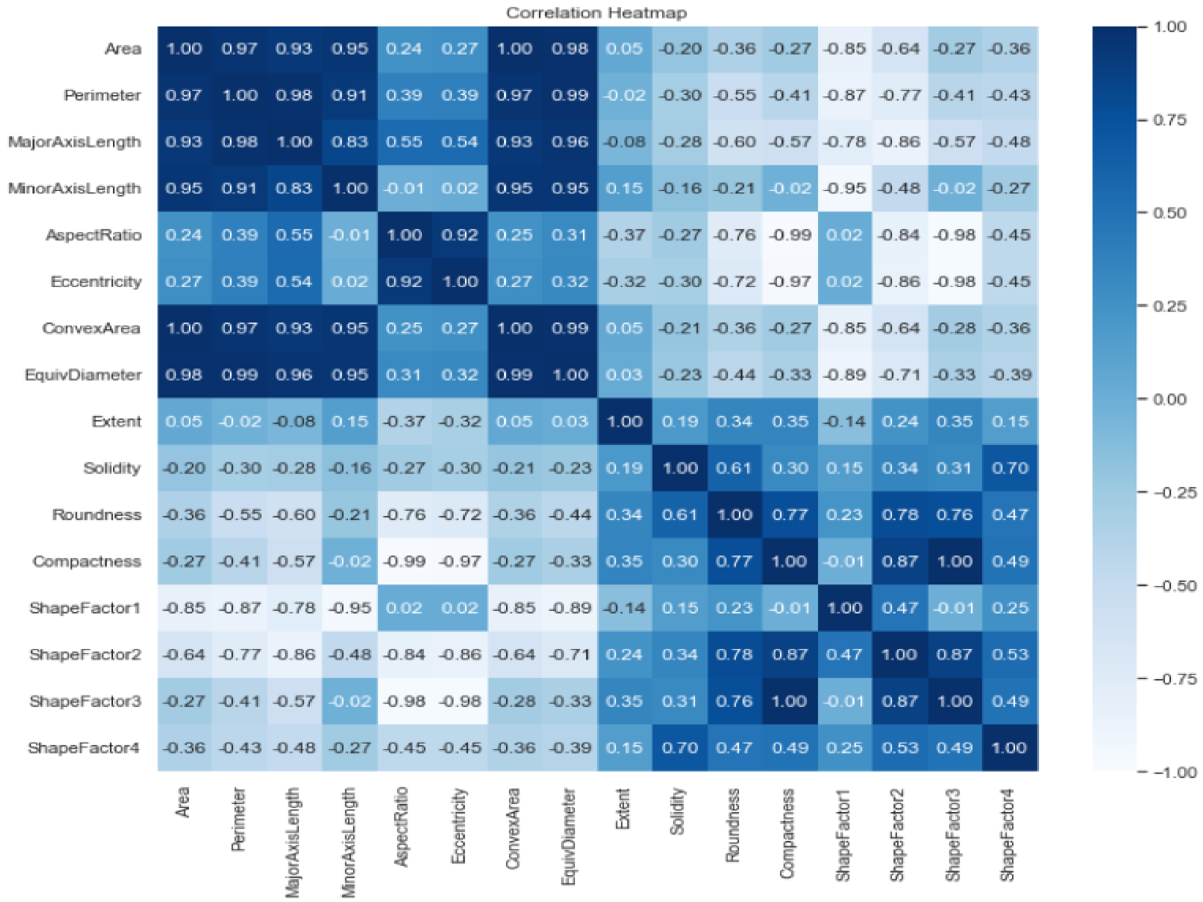


To overcome this problem of data imbalance, we used the over-sampling technique. Specifically leveraging **SMOTE** (Synthetic Minority Oversampling Technique).

Over-sampling involves increasing the number of observations from the minority class to mitigate the imbalance.

As for multivariate analysis, a high correlation appears between several attributes, as shown in the correlation matrix below. Therefore, the use of techniques such as **Principal Component Analysis (PCA)** becomes advantageous for dimensionality reduction and allows for uncorrelated features. Some of its variables are **Area** & **ConvexArea, ShapeFactor3 & Compactness, AspectRatio & Compactness, Area & Perimeter, Perimeter & ShapeFactor1, AspectRatio** & **Eccentricity etc.**

Figure 6: Correlation heatmap



To summarize, the **BOMBAY** class stands out distinctly in several attributes, such as **Area**, **MinorAxisLength**, **ConvexArea**, **MinorAxisLength**, and **ShapeFactor1**, showcasing a well-separated distribution from other classes. Notably, **Area** and **ConvexArea** exhibit remarkably similar distributions across all classes, suggesting a very high correlation between these attributes. The **Extent** attribute displays a consistent range of values across all bean classes, indicating that the bounding box's representation of the bean area (measured by the extent attribute) is uniformly effective for all classes. Attributes like **ShapeFactor2** and **Solidity** exhibit highly skewed distributions with long tails, implying non-uniform patterns. Boxplot analysis reveals a substantial number of outliers in attributes such as **Solidity**, **Roundness**, and **ShapeFactor4**. Boxplot clearly depicts the **Bombay** class's distinctiveness from other classes across various attributes. A prevalent observation from the correlation matrix is the high correlation among many attributes, emphasizing the suitability of **Principal Component Analysis (PCA)** for reducing dimensionality and generating uncorrelated features.

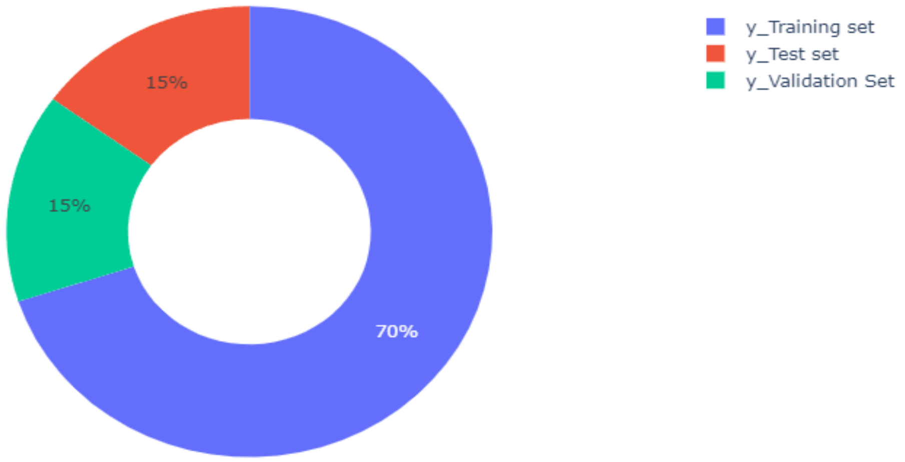
b. Data pre-processing

As we have outliers (cf Boxplot figure) in our explanatory variables we must remove them before bringing our explanatory variables back to Normal distribution (i.e mean 0 and variance 1). **StandardScaler** does not guarantee balanced feature scales, due to the influence of the outliers while computing the empirical mean and standard deviation. This leads to the shrinkage in the range of the feature values. **RobustScaler** is a technique that uses median and quartiles to tackle the biases rooting from outliers. Instead of removing mean, RobustScaler removes median and scales the data according to the quantile range aka IQR (Interquartile Range). By default settings, the IQR is the range between the 1st quartile (25th quantile) and the 3rd quartile (75th quantile) but this can be modified.

$$X_{scale} = \frac{x_i - x_{med}}{x_{75} - x_{25}}$$

By using RobustScaler, we can remove the outliers and then use StandardScaler for the normalization of our covariates.

Figure 7: Comparison of sizes of y-training set, y-validation set and y-test set.



As can be seen in the figure above, we have divided our data into three parts: a part for training our models which represents 70% of the data, a second part called validation which represents 15% of the data and a last part which also represents 15% of the data but used specifically for the evaluation of the final model chosen.

- **Models**

We first performed as a first model the Elastic Net model with logistic regression, which is obviously a model that does not require a dimension reduction method. Following this model, we performed the PCA (Principal Analysis Component) technique for dimension reduction using the Elbow Method for the optimal number of components (4 in our case) and then performed the other classification algorithms of Scikit Learn. A total of 12 classifications algorithms were used.

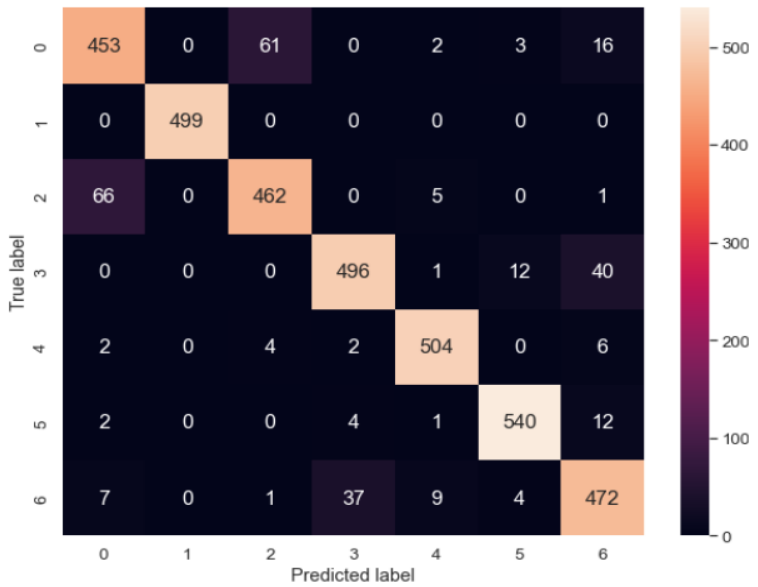
Table *1*: Result

A	Classifier	Training Accuracy	Validation Accuracy
0	Elastic Net Logistic Regression	0.943597	0.947086
1	Random Forest	1.000000	0.918883
2	GradientBoosting	1.000000	0.915928
3	Neural Network	0.919079	0.915391
4	SVM	0.926791	0.914854
5	KNN Classifier	0.929439	0.911093
6	XGBoost	0.929899	0.907601
7	Decision Tree	0.930129	0.897663
8	Naives Bayes	0.867050	0.875369
9	SGD Classifier	0.854043	0.849315
10	AdaBoost	0.813871	0.814128
11	Ridge Classifier	0.592691	0.584743

So, we took the first two models to evaluate it on our test data since the two models are models with different reduction methods. The performance obtained with the Random Forest model is much higher than that obtained with the ElasticNet logistic regression model (**0.919979 vs. 0.907626** for the Elastic Net). In this case, the Random Forest model was used.

As a result, we also implemented a 4-layer deep learning classification model, but the performance of the one is almost equal to that of the Random Forest (**0.9192 vs. 0.9199**). Thus, our final model will be the Random Forest model whose confounding matrix is as follows:

Table 2 : Confusion Matrix for Random Forest model



True label \ Predicted label	0	1	2	3	4	5	6
0	453	0	61	0	2	3	16
1	0	499	0	0	0	0	0
2	66	0	462	0	5	0	1
3	0	0	0	496	1	12	40
4	2	0	4	2	504	0	6
5	2	0	0	4	1	540	12
6	7	0	1	37	9	4	472

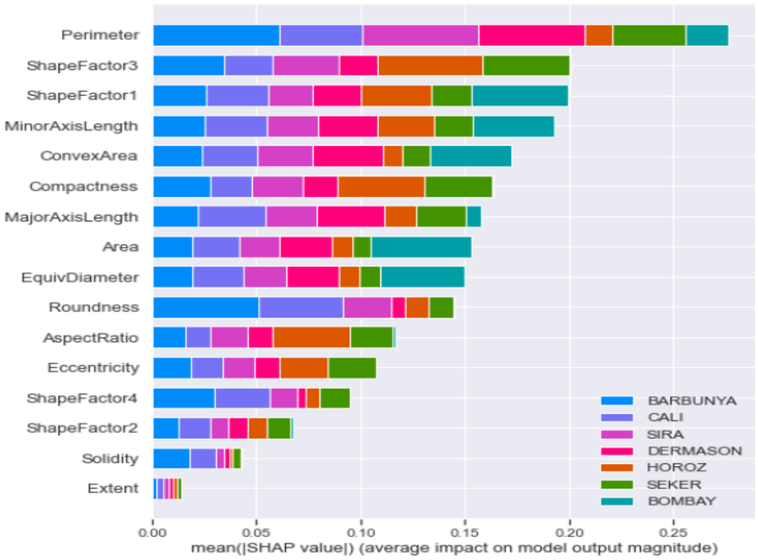
500
400
300
200
100
0

As we can see, we have a lot of correct class predictions for dry bean types and of course less negative false. Especially the **BOMBAY** Class which is class 1 and is correctly predicted.

- **Shapley Additive Explanations - SHAP VALUES with Random Forest Classifier**

Finally, we were a little interested in how the predictions were made in the Random Forest model, which variables participated the most and which ones participated the least.

Figure 8 : Features Importance



As shown in the figure above, we can see that the variables that participated the least in predicting classes in our Random Forest model are **Extent, Solidity,** and **ShapeFactor2**. By removing its variables and performing our model, we obtain a performance of **0.950987** higher than the one obtained previously.

7. Conclusion

In conclusion, meticulous data cleaning and the application of dimension reduction techniques, including principal component analysis (PCA), play a crucial role in obtaining accurate and reliable results, especially in the field of machine learning. Carefully removing errors, inconsistencies, and unusual data from our datasets is critical to ensuring that the models generate intelligent and accurate predictions. Failure to properly clean data can compromise the performance of algorithms, leading to erroneous conclusions and unreliable information. It is therefore imperative to invest time and effort in thorough data cleaning to optimize analysis and take full advantage of the power of machine learning. In addition, it is interesting to note that creating a deep learning classification model can have comparable benefits to that of a Random Forest-based model, although this observation is not necessarily generalizable. The classification of dry bean seed varieties is of crucial importance for seed uniformity and quality assurance, but the complexity of building an accurate model for this task was a major challenge in our project, hence the exploration of several models. In addition, we aim to achieve an accuracy comparable to that obtained by excluding variables considered less important. Although these variables may have a lower weight, they come from the analysis of the bean images, thus justifying our search for an optimization of the final model while keeping the 16 explanatory variables.

References

- [Anup Vibhute & S K Bodhe, 2012]. “Applications of Image Processing in Agriculture: A Survey”, International Journal of Computer Applications (0975 – 8887) Volume 52– No.2
- [Murat & Ilker,2020]. “Multiclass classification of dry beans using computer vision and machine learning techniques”. Computers and Electronics in agriculture, volume 174.
- [Md S Khana , Tushar Deb Nathb , Md Murad Hossainc , Arnab Mukherjee d , Hafiz Bin Hasnatha , Tahera Manhaz Meema , Umama Khane, 2023]. « Comparison of multiclass classification techniques using dry bean dataset », International Journal of Cognitive Computing in Engineering 4 (2023) 6–20.
- [Jaime Carlos Macuácua, Jorge António Silva Centeno, Caísse Amisse, 2023]. “Data mining approach for dry bean seeds classification », Smart Agricultural Technology 5.
- [Md. Mahadi Hasan ; Muhammad Oussama Islam; Muhammad Jafar Sadeq, 2021], “A Deep Neural Network for Multi-class Dry Beans Classification”. 24th International Conference on Computer and Information Technology (ICCIT).
- [Murat & Ilker,2020]. “Multiclass classification of dry beans using computer vision and machine learning techniques”. Computers and Electronics in agriculture, volume 174.
- [Monzurul Islam,Anh Dinh, Khan Wahid, Pankaj Bhowmik, 2017]. “Détection de maladies de la pomme de terre à l’aide d’une segmentation d’images et d’une machine à vecteurs de support multiclasses”. IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCGECE).
- [Ajay Khatri,Shweta Agrawal, & Jyotir M. Chatterjee, 2022]. “Wheat Seed Classification: Utilizing Ensemble Machine Learning Approach”. Scientific Programming, Article ID 2626868, 9 p.
- [Saloni Kumari, Deepika Kumar & Mamta Mittal,2021]. “An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier”, international journal of cognitive computing in engineering. Vol 2, pages 40-46.
- Analyse de la taille et de la part du marché des haricots secs - Rapport de recherche de lindustrie - Tendances de croissance (mordorintelligence.com).
- A. Mukherjee, K.S.N. Ripon, L.E. Ali, et al. “Image gradient based iris recognition for distantly acquired face images using distance classifiers”International Conference on Computational Science and Its Applications, Springer, Cham (2022), pp. 239-252.
- Production mondiale de haricots verts par pays - AtlasBig.com
- Production of Dry Bean in Turkey - 2022 - Charts and Tables - IndexBox
- Hui Zou and Trevor Hastie (2005) in their 2005 paper "Regularization and variable selection via the elastic net"
- Vladimir Vapnik (1995) "The Nature of Statistical Learning Theory,"