**Name: Dev Bhardwaj**

**Enrollment Number: 03419011621**

**Email ID: dev.bhardwaj2103@gmail.com**

**Contact Number: 8168410771**

**Google Drive link:** [**https://drive.google.com/drive/folders/1vfHeDrsdxWhswHjw-BqxxdlGIqbNPMyw?usp=drive\_link**](https://drive.google.com/drive/folders/1vfHeDrsdxWhswHjw-BqxxdlGIqbNPMyw?usp=drive_link)

**Dry Beans Classification Using Machine Learning**

**Abstract:**

The dry bean is the most widely grown edible legume crop in the world, and it has a diverse genetic makeup. Crop productivity is undoubtedly influenced by seed quality. As a result, seed classification is crucial for marketing as well as production in order to supply the fundamentals of sustainable agricultural systems. Since crop production is in the form of population rather than a single variation, the main goal of this study is to offer a technique for getting homogeneous seed variants. To recognise seven different registered types of dry beans with comparable traits and achieve uniform seed categorization, a computer vision system was created. A high-resolution camera was utilised to capture photos of 13,611 grains from 7 different classified dry bean varieties for the classification model. The MATLAB GUI, or graphical user interface, was used establishing a user-friendly interface. Following segmentation and feature extraction processes on bean pictures obtained using a computer vision system (CVS), a total of 16 features—12 dimension-based and 4 shape-based—were extracted from the grains. Tenfold cross validation was used to develop classification models for Support Vector Machine (SVM), Logistic Regression (LR)and Decision Tree (DT), Random Forest (RF)and performance metrics were compared. For SVM, LR, DT and RF, the overall correct classification rates have been found to be 92.61%, 92.50%, 90.32%, and 92.82%, respectively. The Barbunya, Bombay, Cali, Dermason, Horoz, Seker, and Sira bean types were categorised by the SVM classification model, which has the highest accuracy results, with 92.36%, 100.00%, 95.03%, 94.36%, 94.92%, 94.67%, and 86.84%, respectively. With these results, the producers' and consumers' needs for consistent bean types are largely achieved.

**Keywords**

Dry bean ,Interquartile range , ADASYN ,Multiclass classification techniques ,Performance measures.

* **Introduction**

Classification of dry beans is of some economic importance. It has become economically significant to classify dry beans. Manual classification requires a lot of labour, etc. More than 13 k samples of dry beans from 7 different species were photographed, and using computer vision techniques, their geometry was evaluated. The set was then examined using a variety of deep learning (also known as artificial neural networks) and machine learning (also known as data science) techniques. Depending on the technique utilised, the overall accuracy was between 87.92 and 93.13%. The UCI machine learning repository has published the dataset used in [1]. In this study, the same dataset is analysed using somewhat different methods. The dimensionality of the data has decreased. Accuracy improvements have been made. Discussion and comparisons have already taken place.

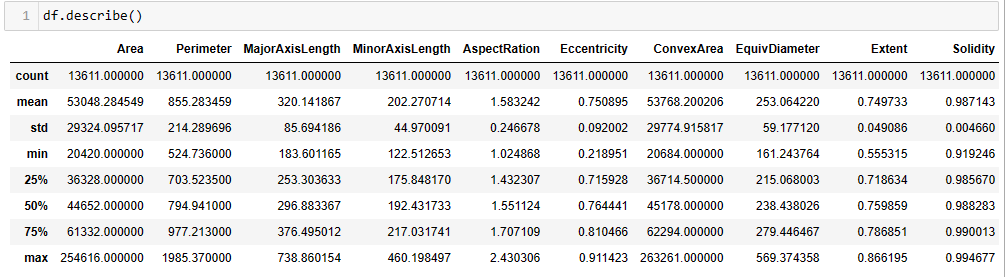
* **Preliminary analysis and visualisation of the dataset**

There are 13611 samples in the dataset being examined. A sample consists of 16 geometrical components along with a label that specifies the bean's species. Barbunya, Bombay, Cali, Dermason, Horoz, Seker, and Sira are the different species. Area, perimeter, major axis length, and ConvexArea, Eccentricity, EquivDiameter, Extent, Solidity, Roundness, Compactness, MinorAxisLength, Aspect Ratio, Eccentricity, ShapeFactors 1, 2, 3, and 4 make up the shape factor system. The method by which the features were determined is given in great detail.

The hue of the bean is not disclosed in the geometrical data. It's terrible from a practical standpoint because the colour of many dry bean species varies frequently. On the other hand, if we only want to use the dry bean classification problem as a learning exercise for creating and comparing machine learning models, it is of no significance.

Prediction is based on these beans attributes:

* 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)



* **Correlation analysis**

Correlation analysis has shown that several of the features are strongly (positively or negatively) correlated. This is due to the fact that basically all of them are kind of geometric measures. The decision was taken to drop some features to avoid correlations over 0.9 (or negative correlation below -0.9) between them.

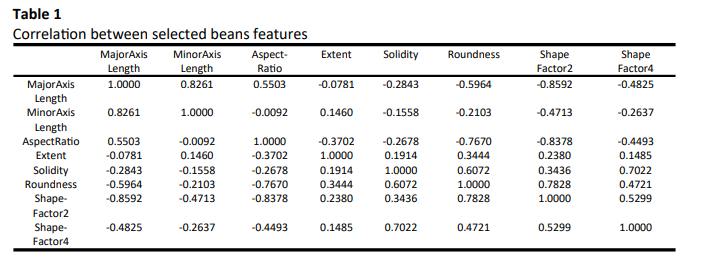
The benefits of such a decision should be:

1) a significant reduction of the computational complexity

2) a lower risk of overfitting

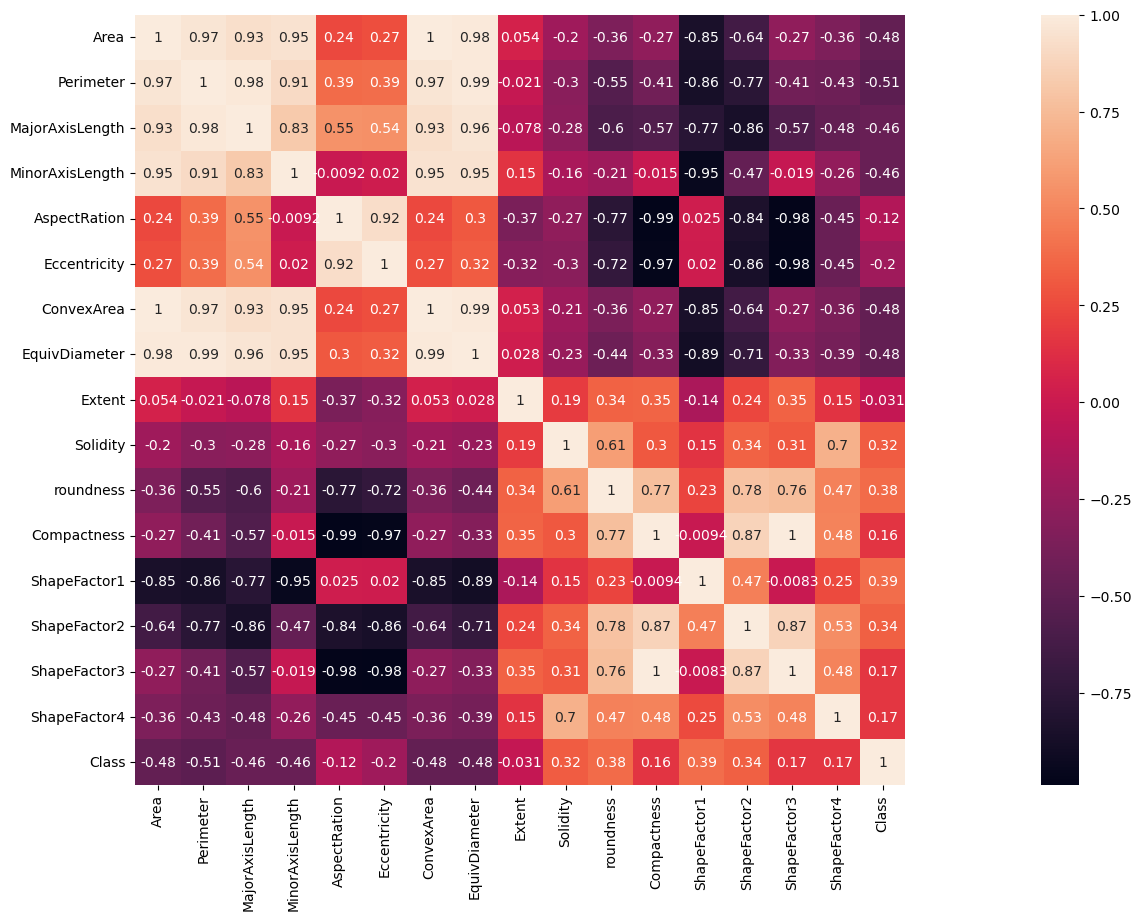
3) ease of visualisation.

The disadvantage is a limited risk of loosing some valuable information and, as a result, a decrease in accuracy

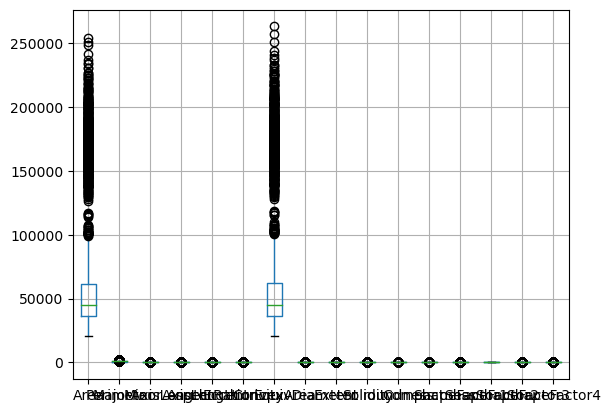


Thus, in this work it was decided to limit the set of features list to these 8 members: MajorAxisLength, MinorAxisLength, AspectRatio, Extent, Solidity, Roundness, ShapeFactor2, ShapeFactor4, and to exclude: Area, Perimeter, Eccentricity, ConvexArea, EquivDiameter, Compactness, ShapeFactor1, ShapeFactor3. The issue of high correlations among some features was not addressed in [1]. The visualisation of the data was done by pair-plot, and is presented in figure 1. It shows that the Bombay species is trivial to classify as its beans are significantly bigger than others. the classification of other species seems to be much more difficult, and we can expect more errors. The correlations between pairs of the selected features are listed in Table 1.

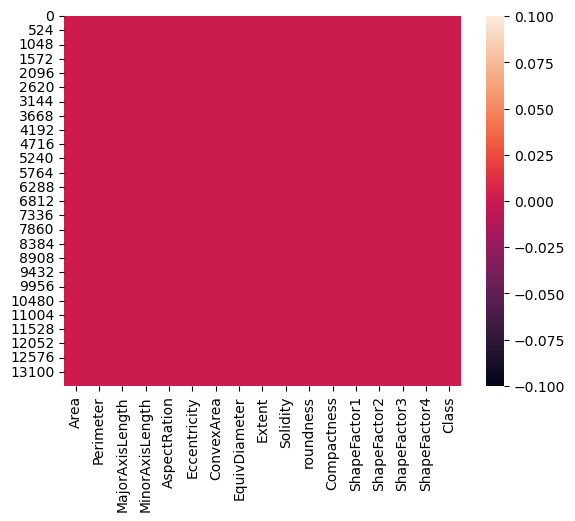
**Corelation Map:**

****

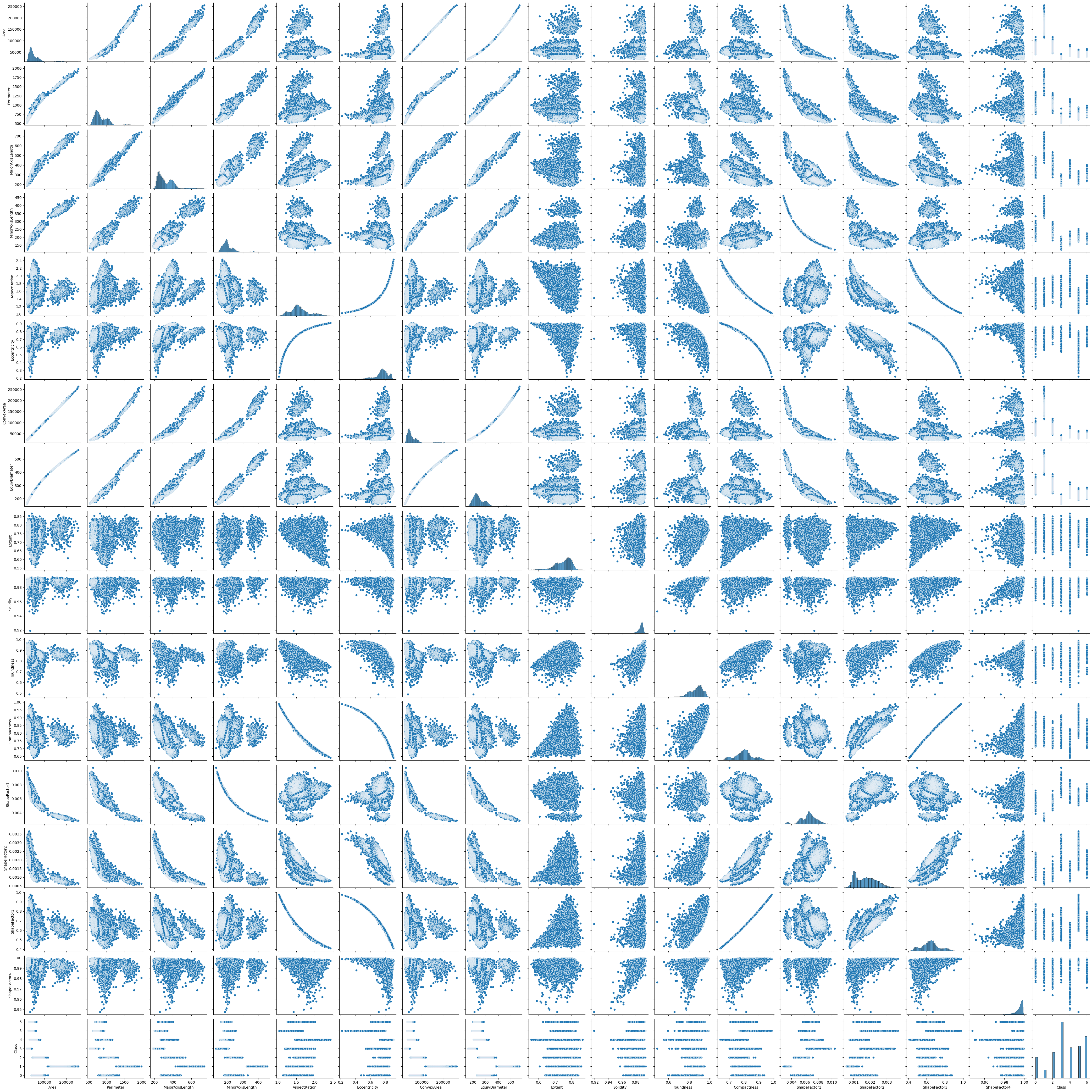
**Boxplot:**

****

**Heatmap:**

****

**PairPlot Graph :**

****

* **Machine Learning techniques used and results**

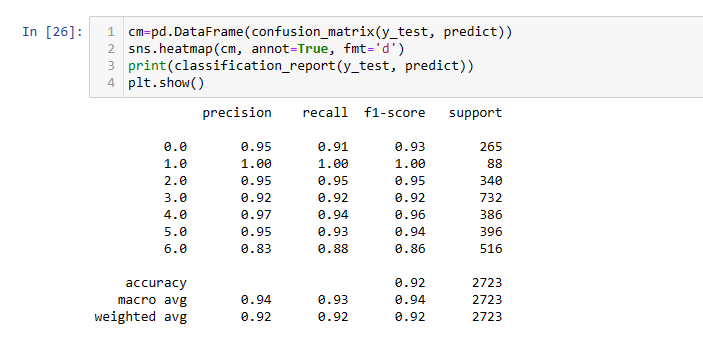
1. **SVM(Support Vector Machine)**
2. **Decision Tree**
3. **Logistic Regression**
4. **Random Forest**
5. **Naïve Bayes Classification**
6. **SVM(Support Vector Machine):**

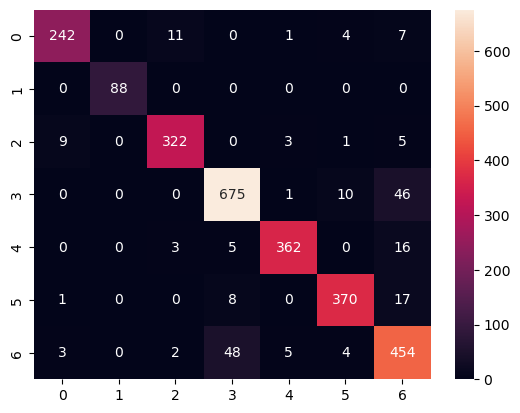
Support vector machines (SVM), which can be used as regressors or classifiers, are considered a very powerful and flexible algorithms. On the other hand. they may need a lot of computing power. The SVM principle is to partition the classes by ”drawing a line” (or plane) in a way that maximises the margin between classes. As straight lines (or planes) do not usually produce the best solution, SVC can apply different kernels (polynomial, radial and others). SVC is wider explained. SVCs with different kernels were tried. The results are quite similar for all kernels. The accuracy can be further improved to some extent (tenths of %, maybe 1%) by increasing C, but this will also significantly increase the training time.





The accuracy score of SVM model is :- 92.6184%

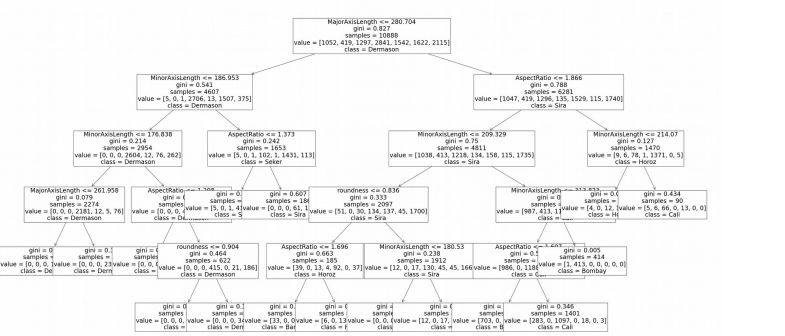


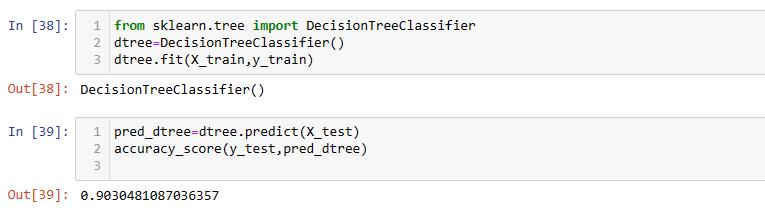


1. **Decision Tree Model :**

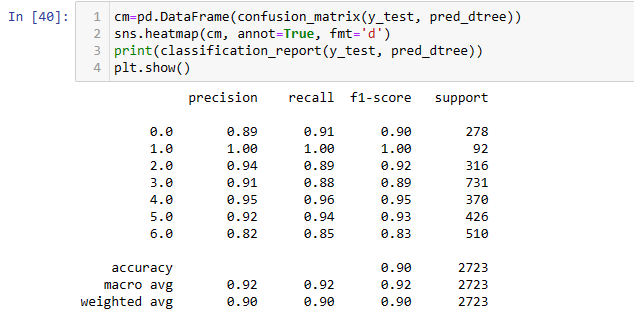
A decision tree (DT) is an example of a non-parametric method. The phrase "nonparametric" can be deceptive. A decision tree does include parameters, although their number varies. A decision tree attempts to determine the optimum questions splitting the dataset during the learning phase in order to reduce information impurity. The fact that decision trees are so intuitive is a significant advantage. A decision tree, on the other hand, has no limited degrees of freedom, making it easy to overfit (if the user is unaware of this). Because decision tree splits are always orthogonal (done on one feature at a time), the decision tree is extremely sensitive to data rotation.

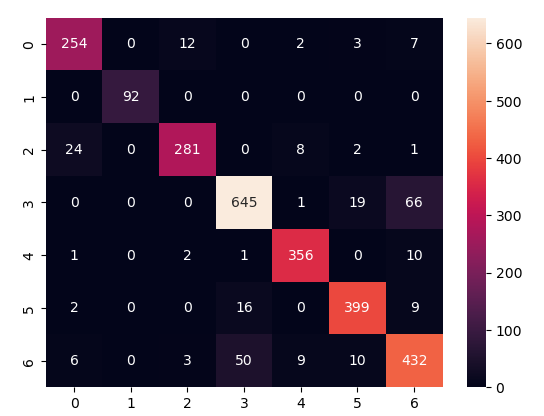
The figure below shows the decision tree of the working model:





The Decision Tree Model accuracy is:- 90.3048%

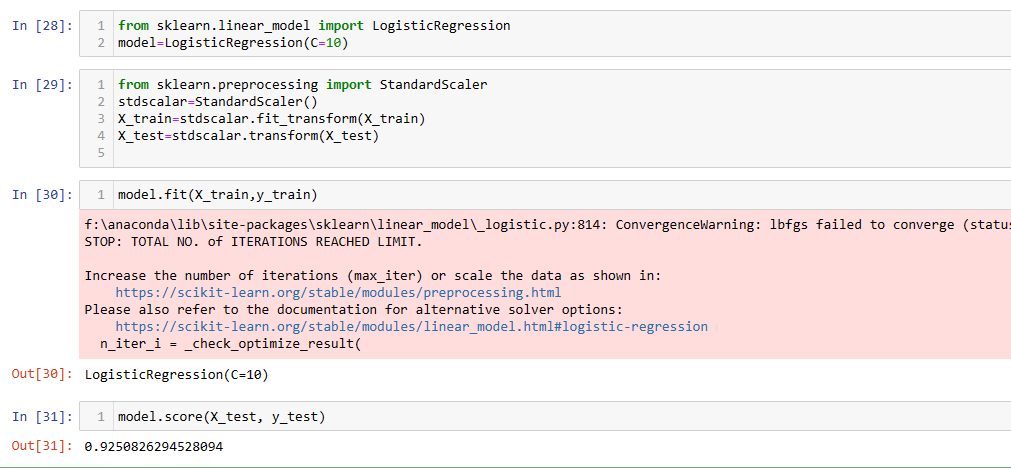




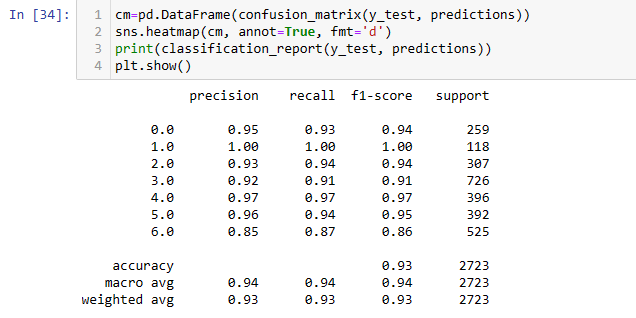
1. **Logistic Regression :**

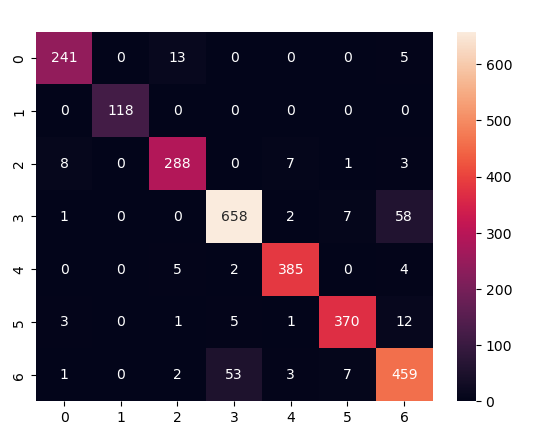
Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.



The accuracy of the Logistic modled is :- 92.5082%

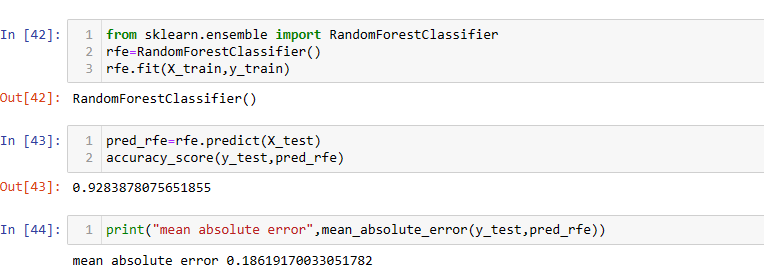




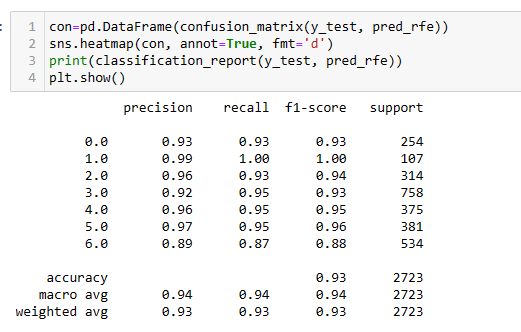
1. **Random Forest:**

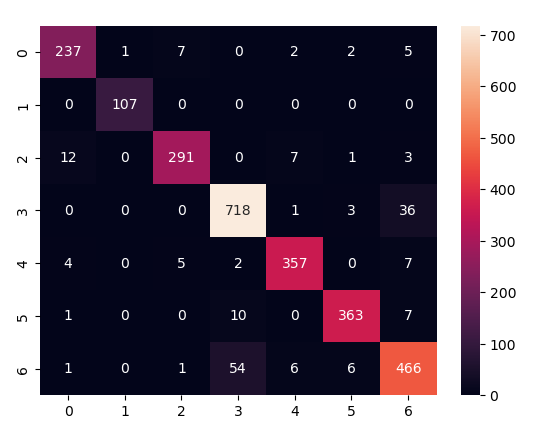
The random forest idea is as follows: take many decision trees (employing some randomness, so the trees differ) and let them vote. So the classification decision taken by a random forest is a decision taken by the most numerous group of decision trees in a random set of trees.

Usually a random forest performs better than a single decision tree. However, a random forest is considered a ”black-box” model being very hard to interpret.



The accuracy of of Random Forest Model is :- 92.8387%





1. **Naïve Bayes Classification:**

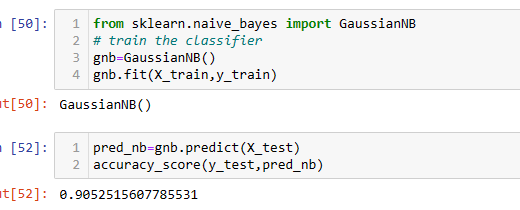
Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems.

It is mainly used in *text classification* that includes a high-dimensional training dataset.

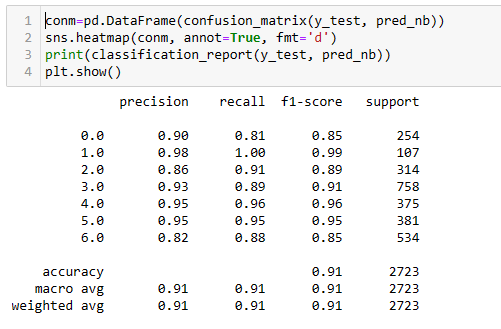
Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

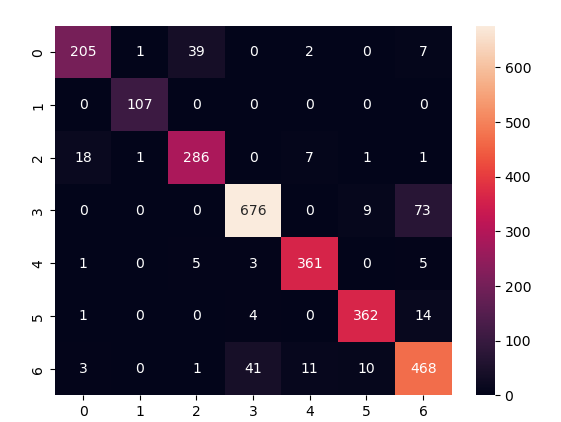
It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.



The Accuracy of Naïve Bayes is : 90.5251%





* **Conclusion**

The dataset used was “Dry Beans Dataset”, it consists of dimensions and shape factors of 13,611 different grains which are then classified into the different bean types. The number of dimensional attributes were 16 and number of shape factor attributes were 4.

The dataset was first filled up with the missing values, as only a few cells were empty and all the values were of float type I did not find it wrong to use the mean to fill the rest of the missing cells. Data analysis was carried out with the different attributes and proper graphs were found using it..

Training of the dataset was carried out. By using multiple classification algorithms testing of the dataset was carried out and the classifier and accuracy score of each of these models were found out and were noted.

It was found out that Random Forest Classification Model has the highest accuracy of value 0.928387.

**6. References**

[1] Murat Koklu, Ilker Ali Ozkan, Multiclass classification of dry beans using computer vision and machine learning techniques, Computers and Electronics in Agriculture 174 (2020) 105507

[2] Dry beans dataset at UCI repository: <https://archive.ics.uci.edu/ml/datasets/Dry+Bean+Dataset>, access 23.06.2021

[4]<https://www.researchgate.net/publication/367131182_Comparison_of_Multiclass_Classification_Techniques_Using_Dry_Bean_Dataset>

[5] Aurelien Geron, Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow, O’Reilly, 2019

[6] [https://ceur-ws.org/Vol-2951/paper3.pdf](%20https:/ceur-ws.org/Vol-2951/paper3.pdf)

[7] Relevant Papers:

KOKLU, M. and OZKAN, I.A., (2020), “Multiclass Classification of Dry Beans Using Computer Vision and Machine Learning Techniques.” Computers and Electronics in Agriculture, 174, 105507.

DOI: [https://doi.org/10.1016/j.compag.2020.105507](%20https:/doi.org/10.1016/j.compag.2020.105507)

[8]Citation Requests / Acknowledgements:

KOKLU, M. and OZKAN, I.A., (2020), “Multiclass Classification of Dry Beans Using Computer Vision and Machine Learning Techniques.” Computers and Electronics in Agriculture, 174, 105507.

DOI: <https://doi.org/10.1016/j.compag.2020.105507>

[9] Murat KOKLU Faculty of Technology, Selcuk University, TURKEY. ORCID : 0000-0002-2737-2360

mkoklu@selcuk.edu.tr

[10]Ilker Ali OZKAN Faculty of Technology, Selcuk University, TURKEY. ORCID : 0000-0002-5715-1040

ilkerozkan@selcuk.edu.tr