

DryBeans dataset visualization and classification

Report



June 17, 2022

submited by imran khan (051-19-0010)

SUBMITTED TO SIR DR. GULAM MURTAZA

**1-Download public dataset from UCI or Google dataset or any other**

Downloaded the dataset from UCI link is given blow:

<https://archive.ics.uci.edu/ml/datasets/Dry+Bean+Dataset>

**2-Briefly explain dataset metadata and its requirements.**

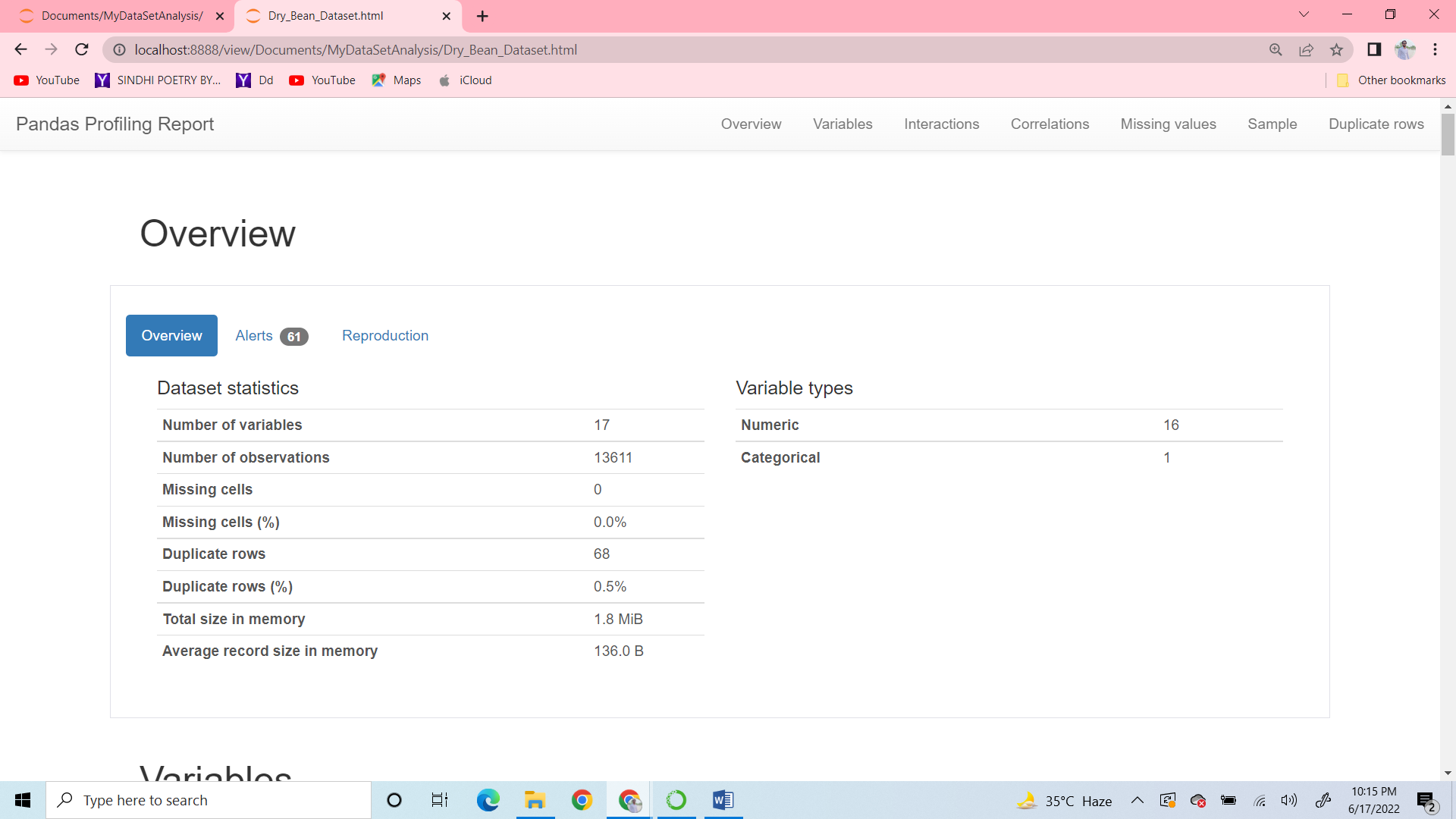
Dry Beans dataset is a collection of different dry beans. The dataset is consist of 17 features **'Area', 'Perimeter',MajorAxisLength', 'MinorAxisLength','AspectRation', 'Eccentricity','ConvexArea', 'EquivDiameter', 'Extent', 'Solidity','roundness','Compactness', 'ShapeFactor1', 'ShapeFactor2',**

**'ShapeFactor3', 'ShapeFactor4', 'Class'** from which 16 are numerical and 1 is categorical that is class feature and there are 13611 tuples. In this dataset dry beans are classified in seven different classes **BOMBAY, DERMASON, SIRA, SEKER, HOROZ, CALI,BARBUNYA** according to their features.

**3-Visualize dataset: Use any SPSS/Python/Weka/Excel, Show boxplot for five measures, histograms for distribution, scatter plot for correlation. Discuss the issues with dataset.**

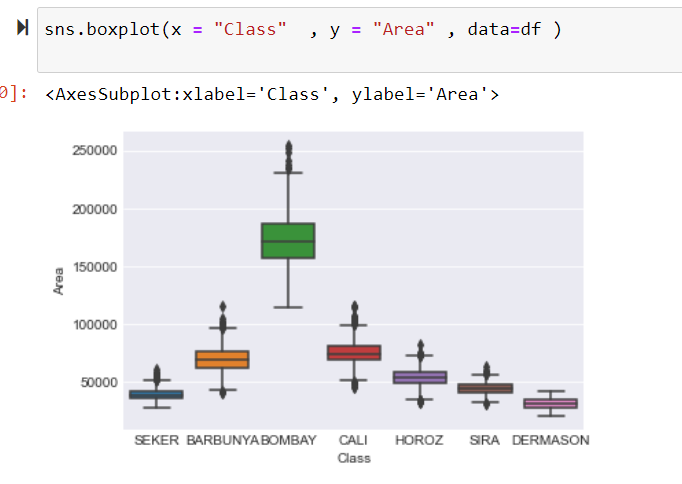
1. **One go visualization**

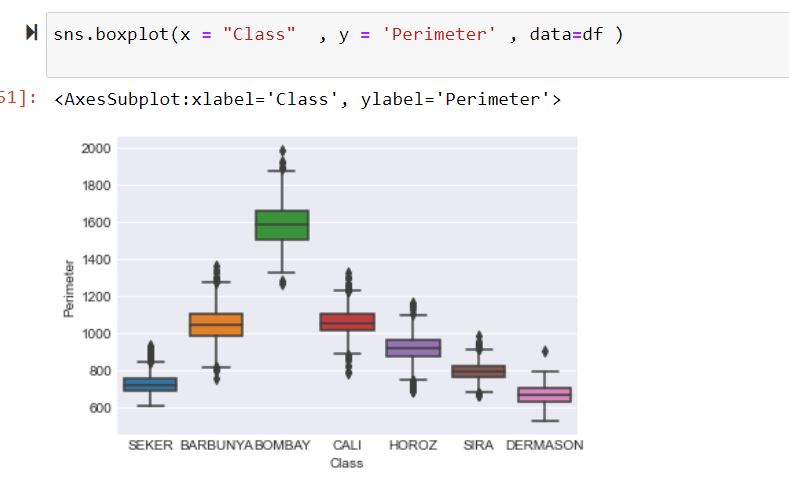
First of all I have used a method of pandas profiling to generate the report of whole dataset for visualization in one go. That file is placed in this project folder as Dry\_Bean\_Dataset.html that contains whole report for dataset. If you run the project first cell a file generated in the project folder with .html extension at local host you can click and view the whole report.

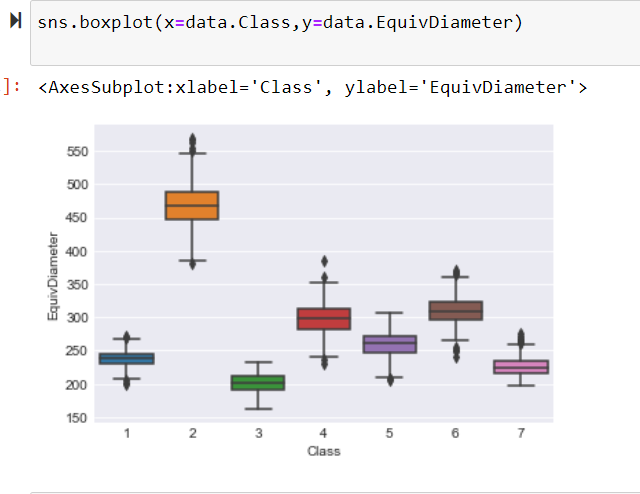


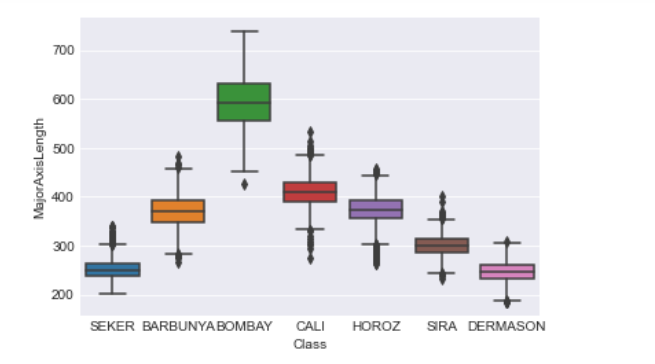
1. **Show box plot for five measures**

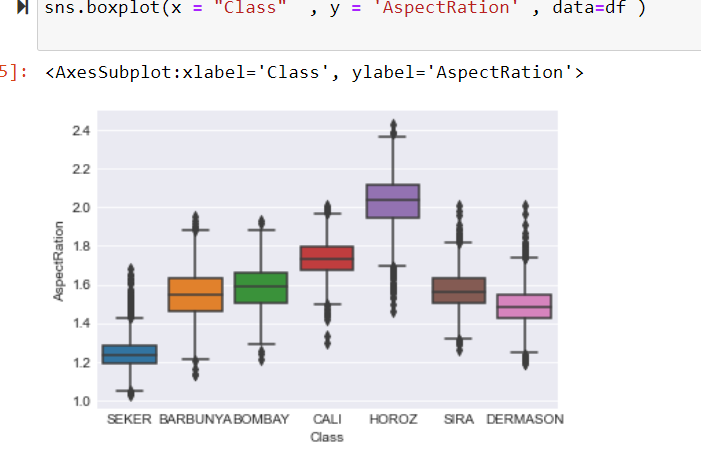
Plot boxplot for five measures using class as x label and y label as a feature.





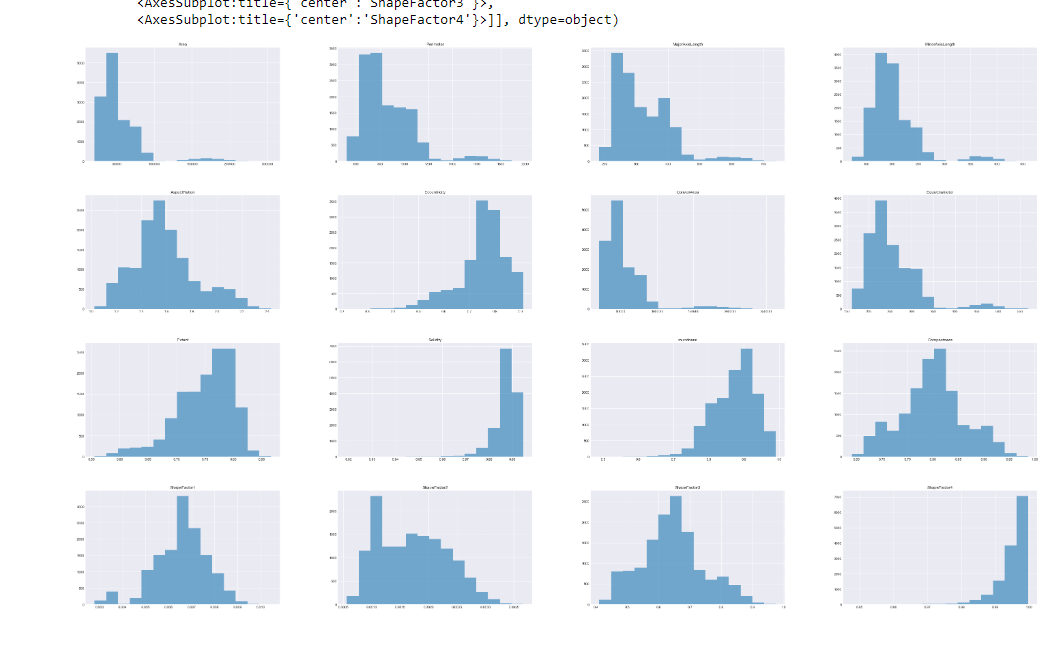




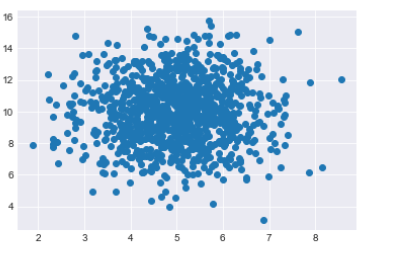


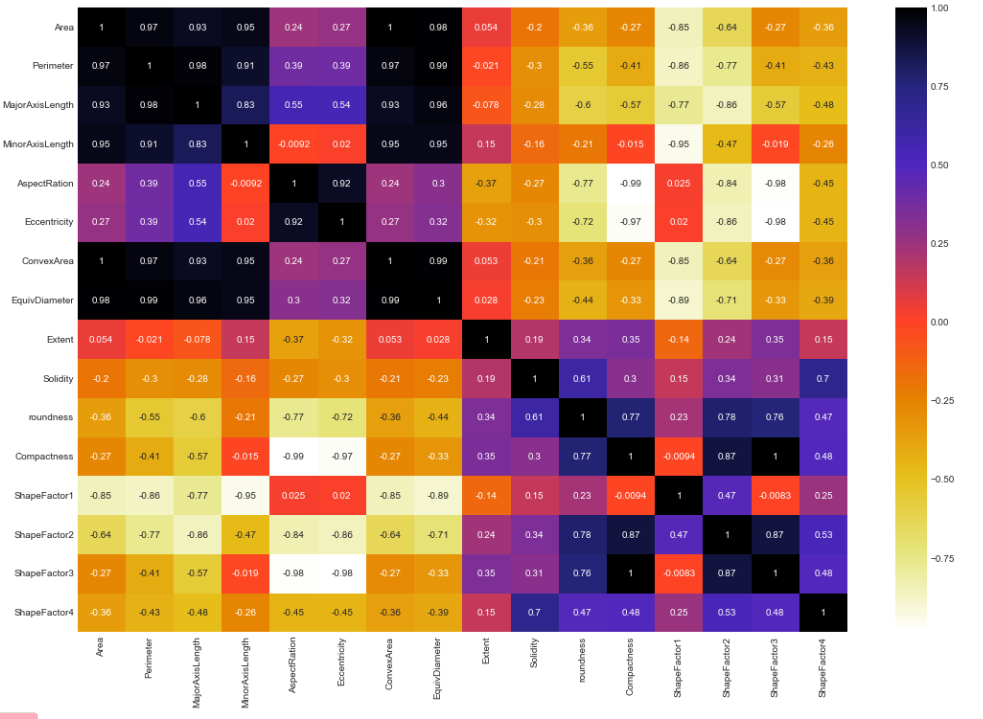
1. **Histograms for distribution**

Plotted histograms for distribution in one chart



1. **Used scatter plot and heatmap to visualized the correlation**





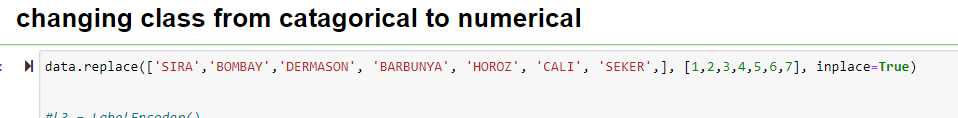
1. **Discuss the issue with dataset**
2. **Duplicate tuples**
3. **Highly correlated features**
4. **Class imbalance**
5. **One categorical feature (class)**
6. **Outliers**

**4-Perform Preprocessing like normalization n etc to resolve issues observed in visualization**

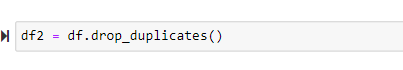
1. **Dropped highly correlated features**



1. **Changed the categorical feature to numerical**



1. **Removed duplicate rows**

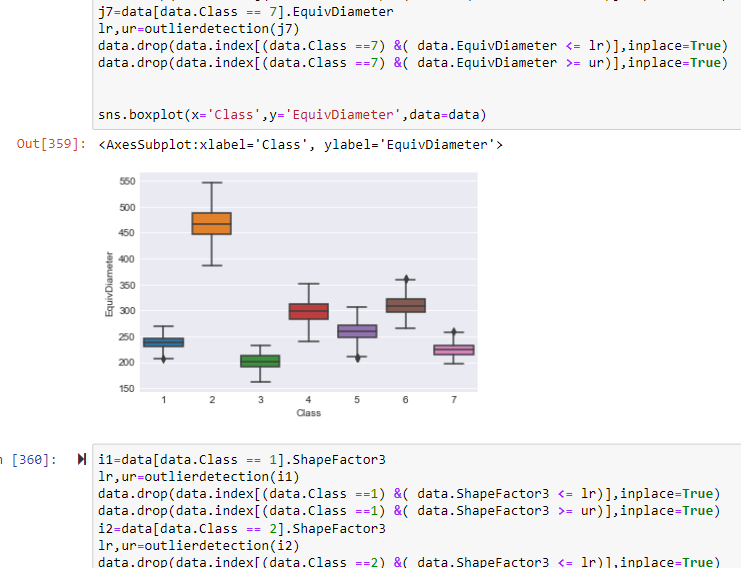


**5-Train model and show accuracy and loss graphs, resolve issues like overfitting, class imbalance or etc.**

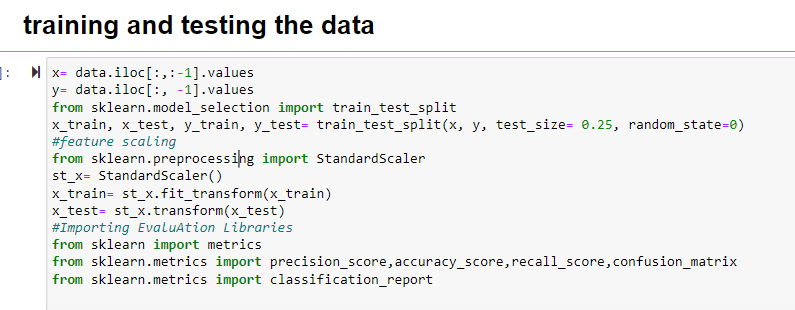
1. **Resolved class imbalance issue**



1. **Removed outliers from all class features**

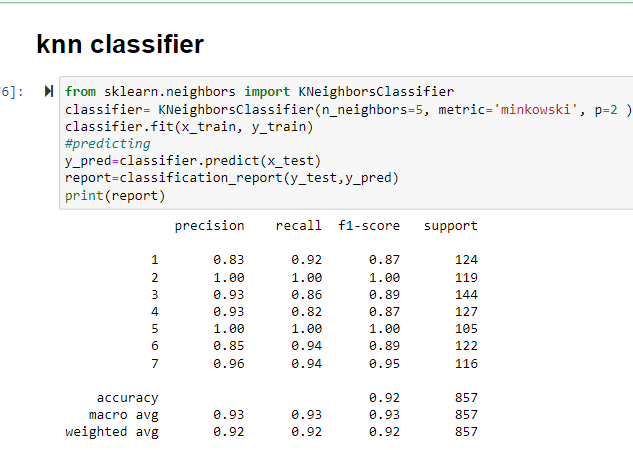


1. **Training and testing**

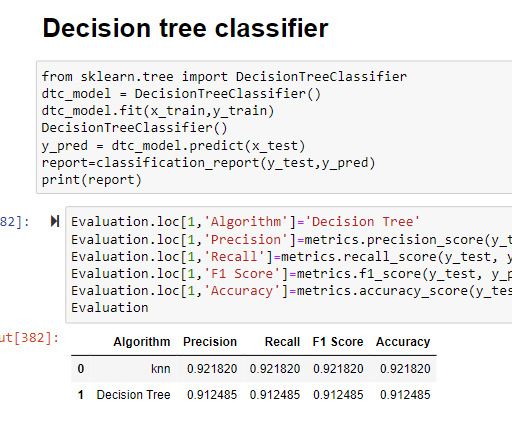


**6-Show various performance evaluation metrics like Accuracy, Sensitivity, Fmeasure, AUROC and etc for test**[**data**](http://elearning.iba-suk.edu.pk/mod/folder/view.php?id=26468)

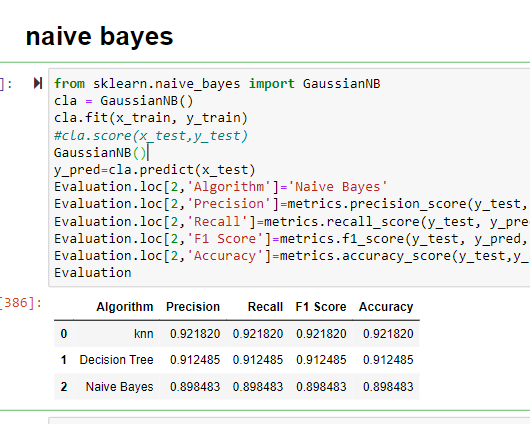
1. **KNN**



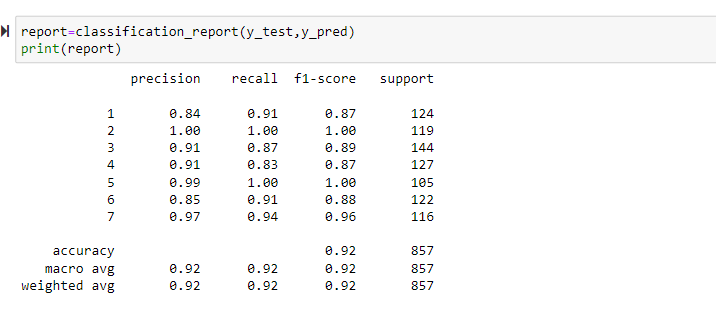
1. **Decision Tree classifier**



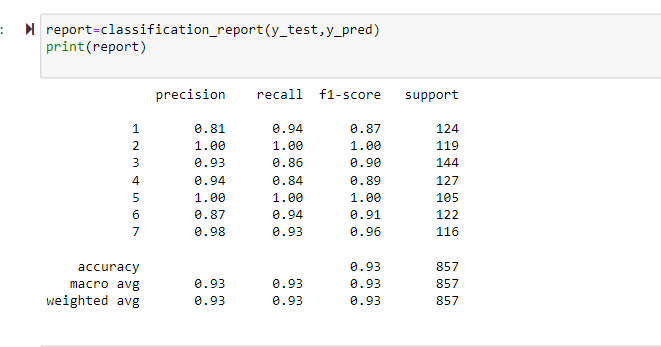
1. **Naïve Bayes Classifier**



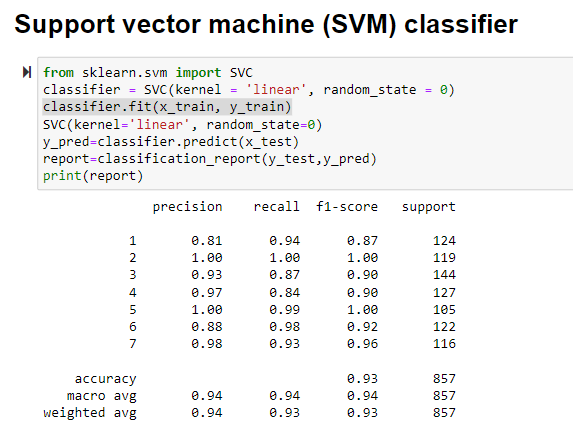
1. **Random forest**



1. **Logistic regression**



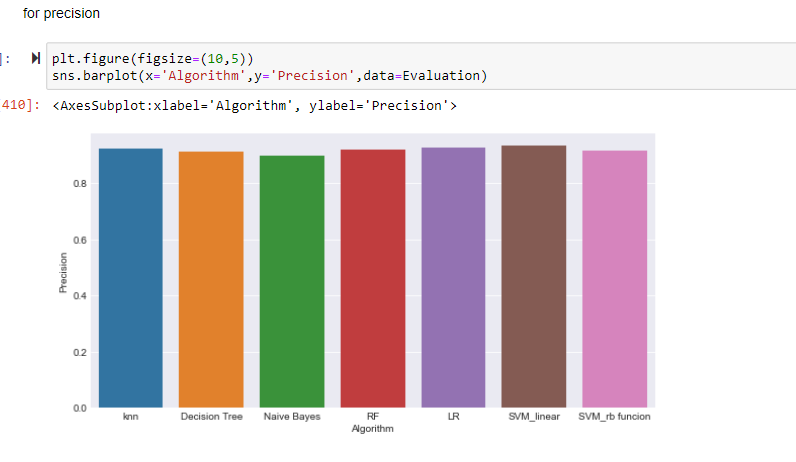
1. **SVM**



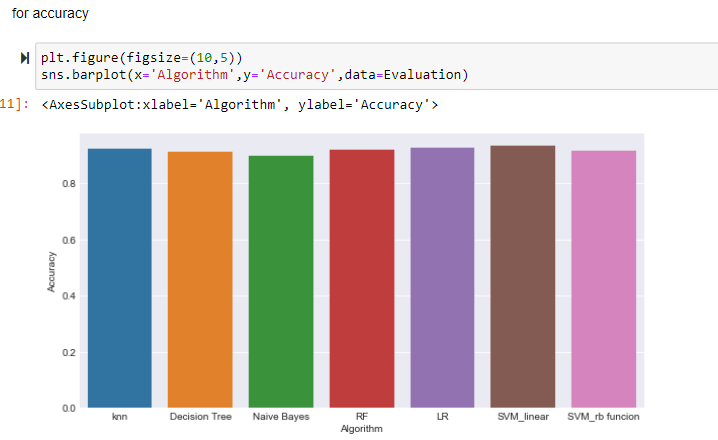
**Conclusion**

**Comparing all classifying models**

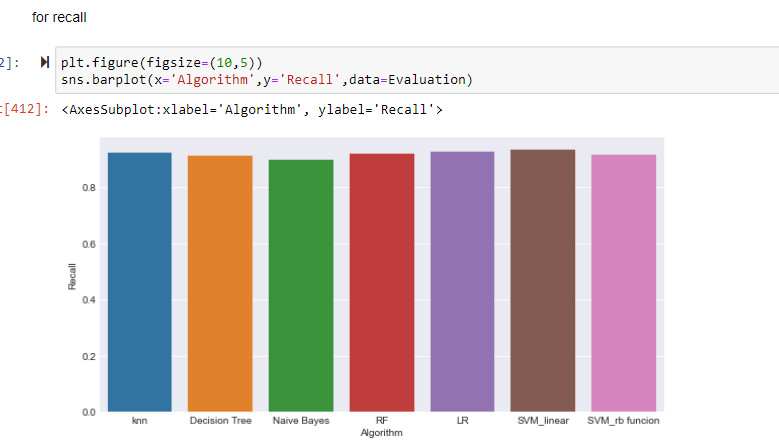
1. **With respect to precision**



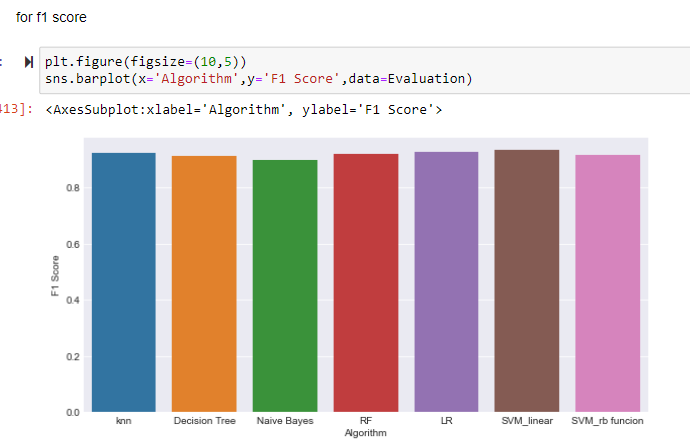
1. With respect to accuracy



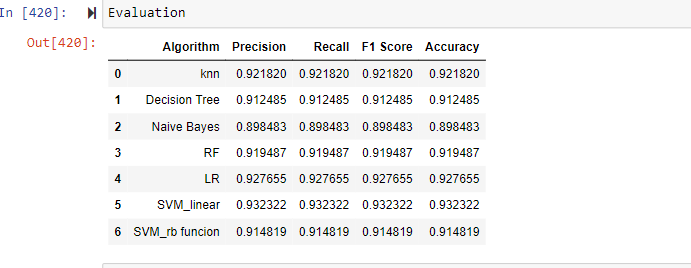
1. With respect to recall



1. With respect to f1 Score



**Over all results for all classification models**



In concluding words Naive bayes, random forest and KNN gives same efficiency which is 92%, Decision tree, SVM\_RBF gives same efficiency which is 93%, Logistic regression and SVM\_Linear gives same efficency which is 94%. All the algorithm gives almost similar Efficiency. Best can **be Logistic or SVM\_linear**

The End