**Practical Assessment (Assignment) Even Semester**

**Course: B.Sc. (H) Computer Science Semester VI (Section B)**

**Paper Name: Data Mining**

**Date: 29th April 2023**

**Name of the Candidate: Vishal Maurya**

**Roll No.: 5750**

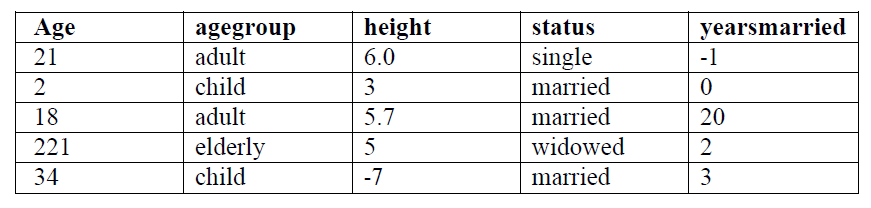
**Examination Roll No.: 20035570117**

**Mobile Number: 9354643132**

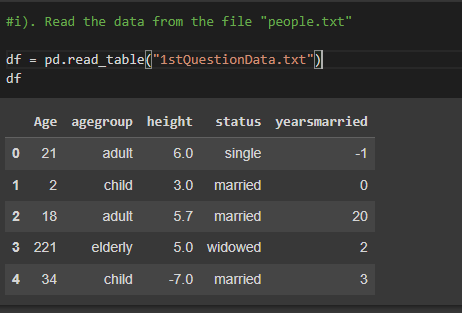
**e-Mail Id:** [**vishal205750@keshav.du.ac.in**](mailto:vishal205750@keshav.du.ac.in)

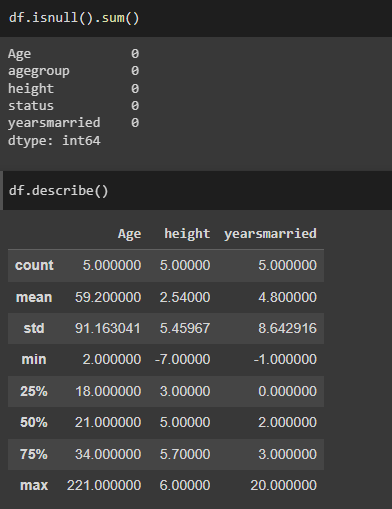
**Submitted To: Nidhi Ma’am**

**Q1. Create a file “people.txt” with the following data:**

****

**i) Read the data from the file “people.txt”.**

****

****

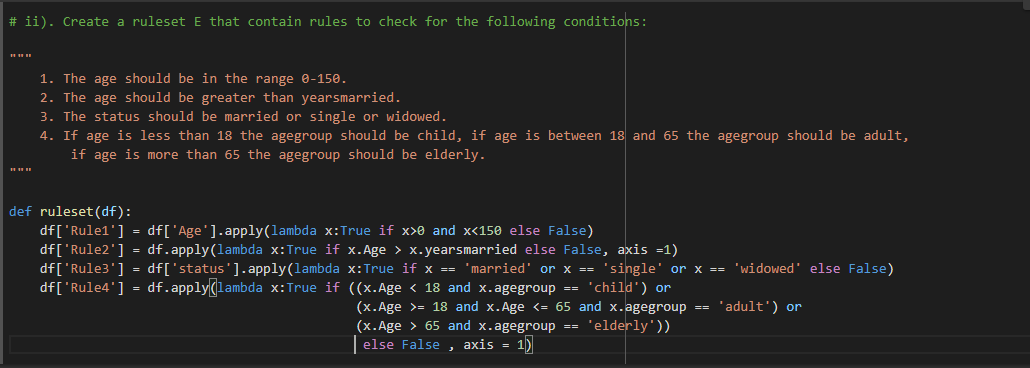
**ii) Create a ruleset E that contain rules to check for the following conditions:**

**1. The age should be in the range 0-150.**

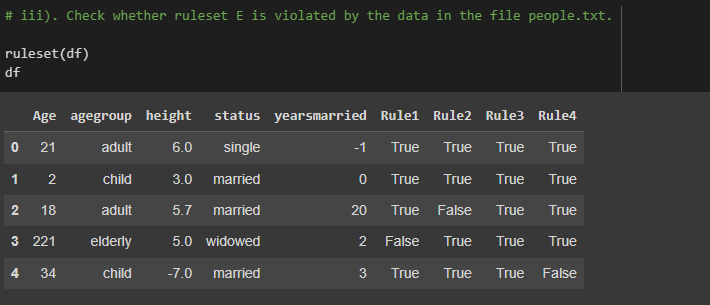
**2. The age should be greater than years married.**

**3. The status should be married or single or widowed.**

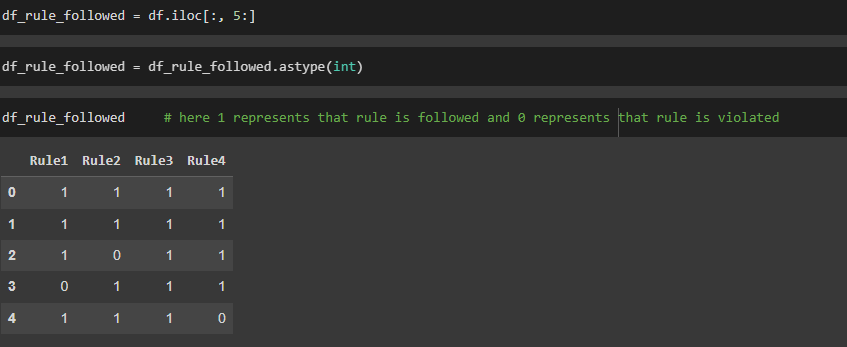
**4. If age is less than 18 the age group should be child, if age is between 18 and 65 the age group should be adult, if age is more than 65 the age group should be elderly.**

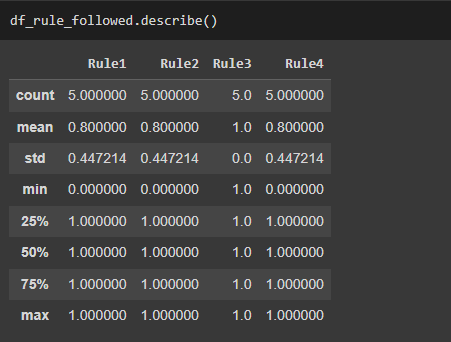
****

**iii) Check whether ruleset E is violated by the data in the file people.txt.**

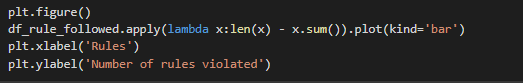


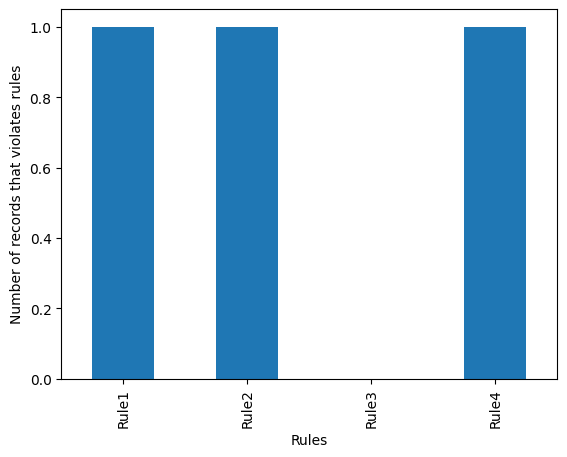
**iv) Summarize the results obtained in part (iii).**

****

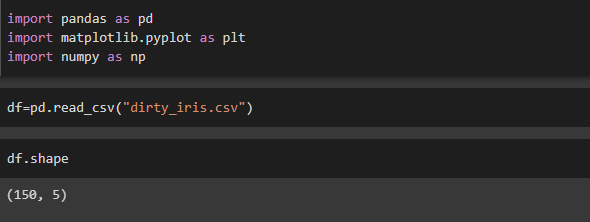
****

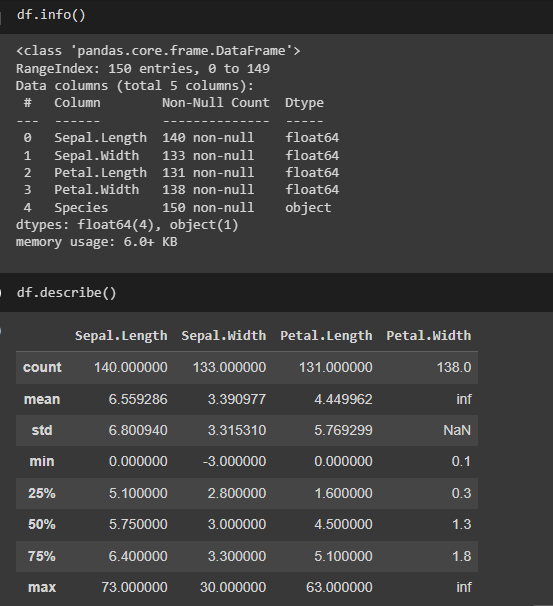
**v) Visualize the results obtained in part (iii)**

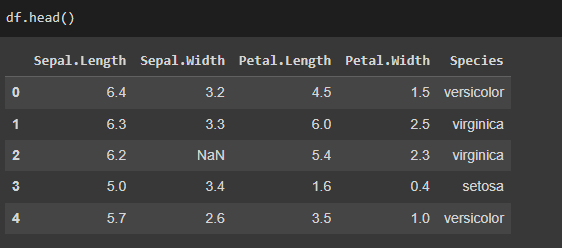
****



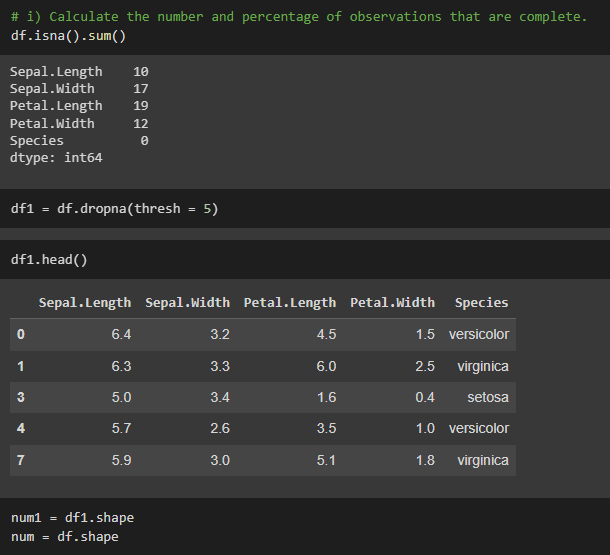
**Q2. Perform the following preprocessing tasks on the dirty\_iris dataset.**

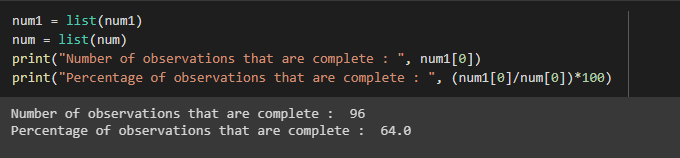
****

****

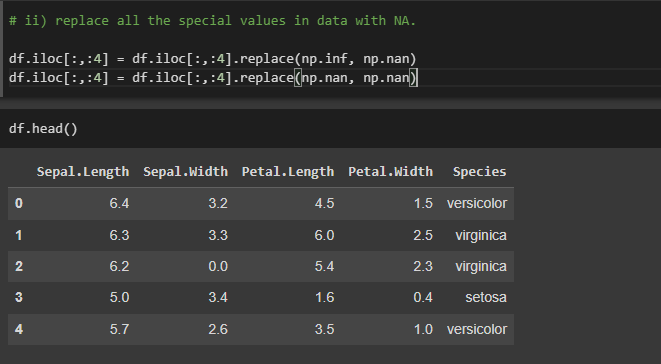
****

**i) Calculate the number and percentage of observations that are complete.**

****

****

**ii) Replace all the special values in data with NA.**

****

**iii) Define these rules in a separate text file and read them. (Use editfile function in R (package editrules). Use similar function in Python).**

**Print the resulting constraint object.**

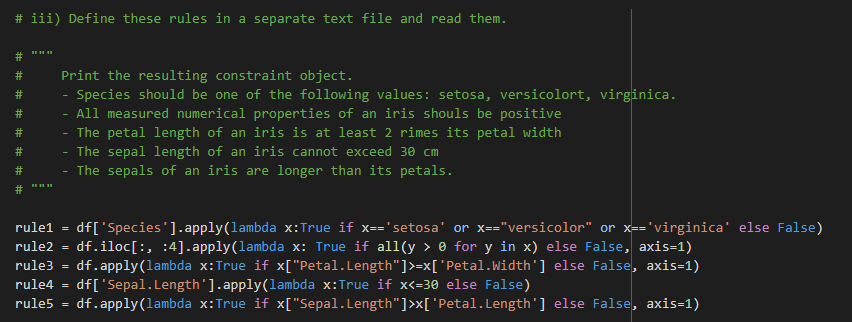
**– Species should be one of the following values: setosa, versicolor or virginica.**

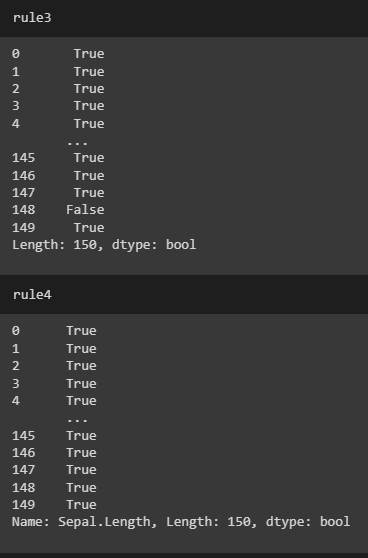
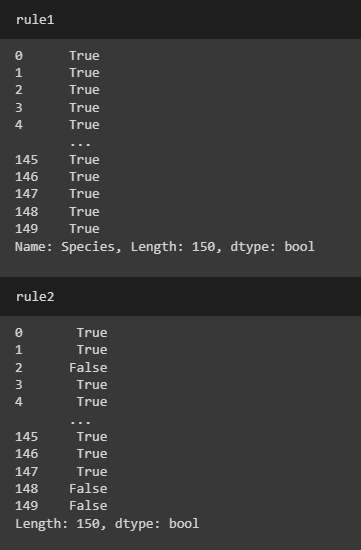
**– All measured numerical properties of an iris should be positive.**

**– The petal length of an iris is at least 2 times its petal width.**

**– The sepal length of an iris cannot exceed 30 cm.**

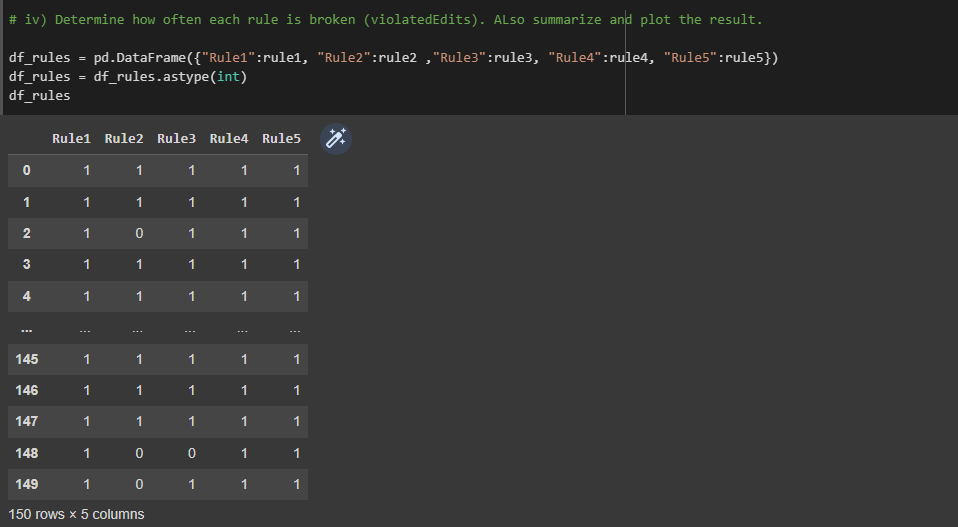
**– The sepals of an iris are longer than its petals.**

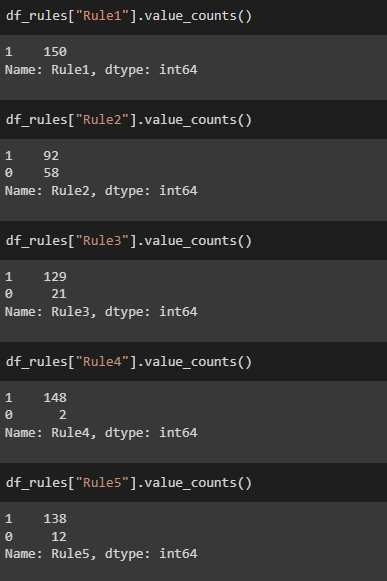
****

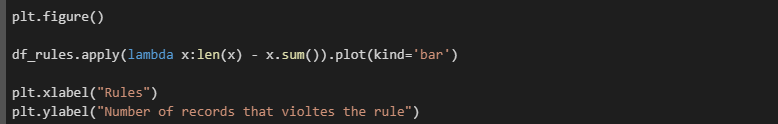
****

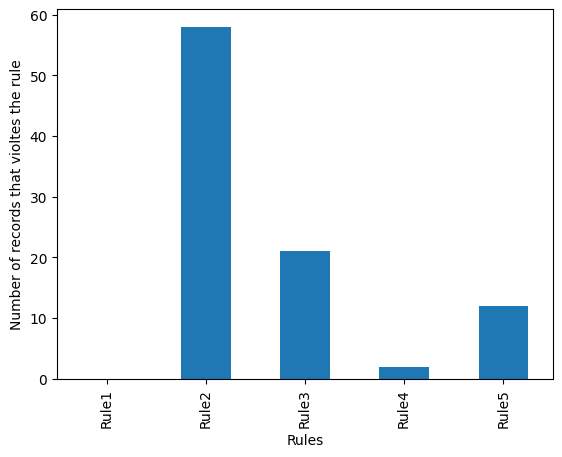
****

**iv) Determine how often each rule is broken (violatedEdits). Also summarize and plot the result.**

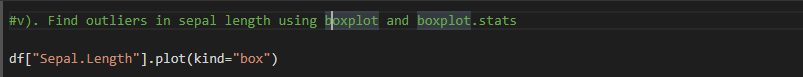
****

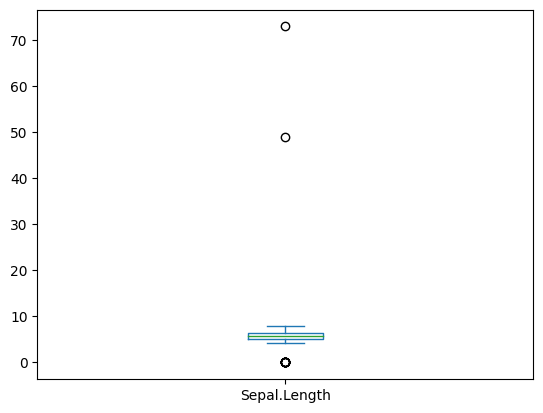
****

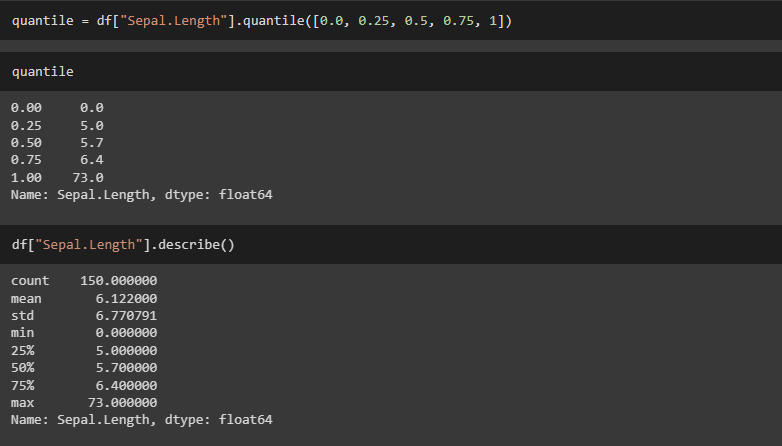
****

****

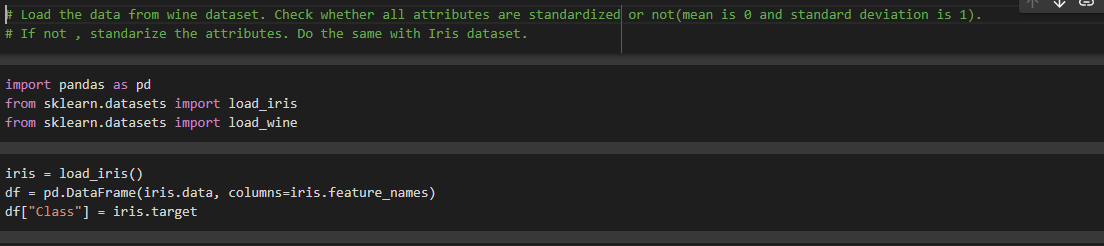
**v) Find outliers in sepal length using boxplot and boxplot.stats**

****

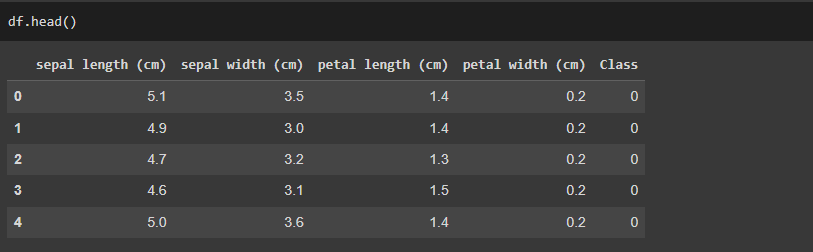
****

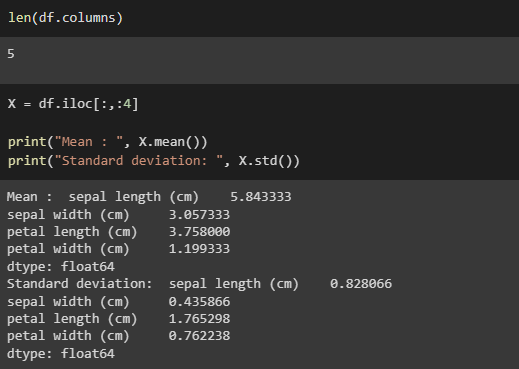
****

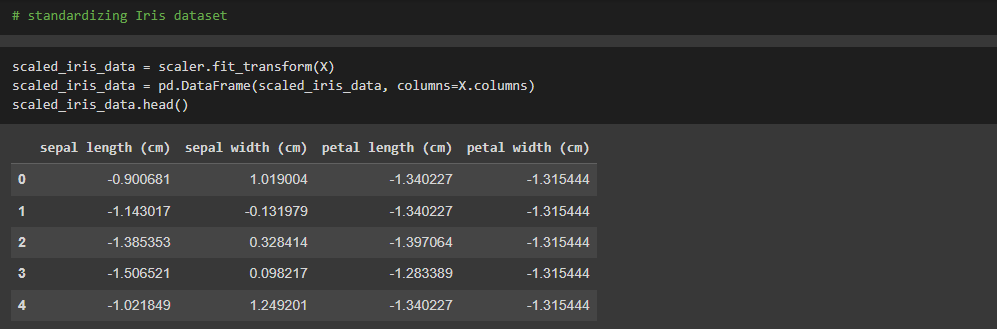
**Q3. Load the data from wine dataset. Check whether all attributes are standardized or not (mean is 0 and standard deviation is 1). If not, standardize the attributes. Do the same with Iris dataset.**

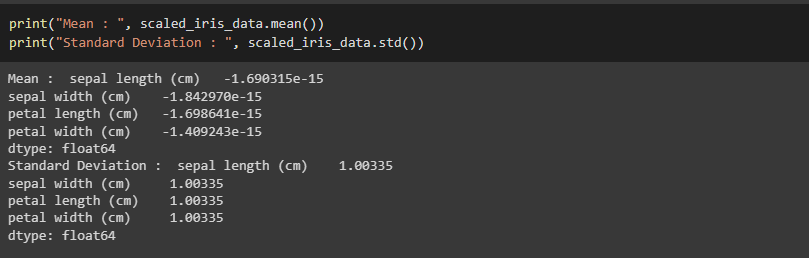
****

**For Iris Dataset:**

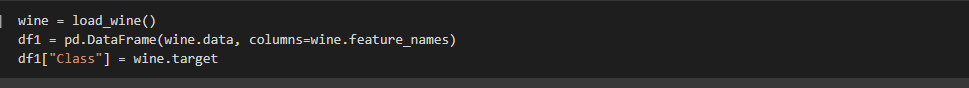
****

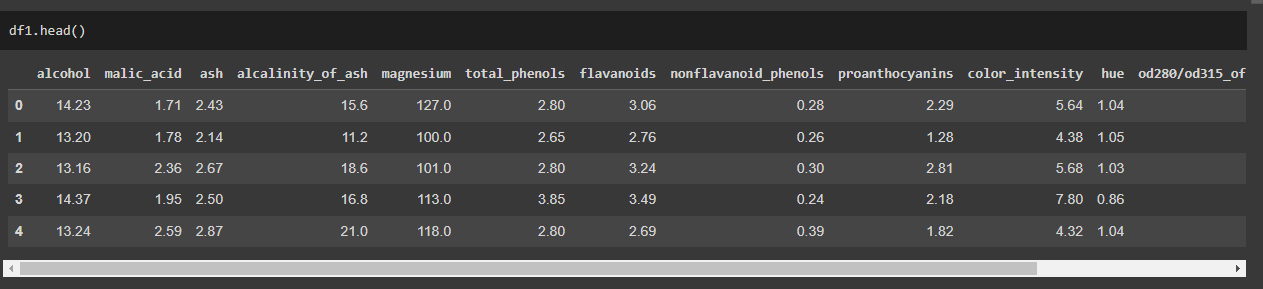
****

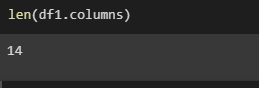
****

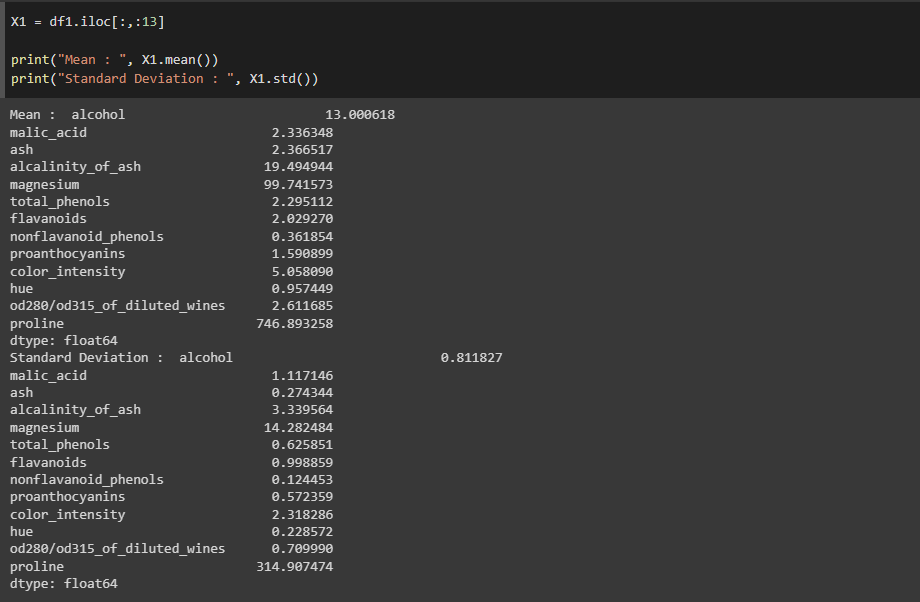
****

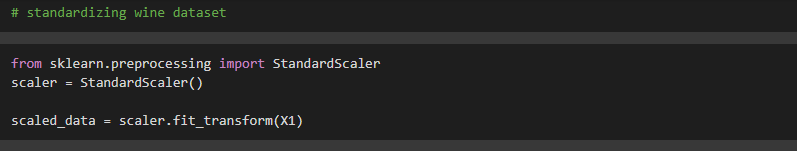
**For Wine Dataset**

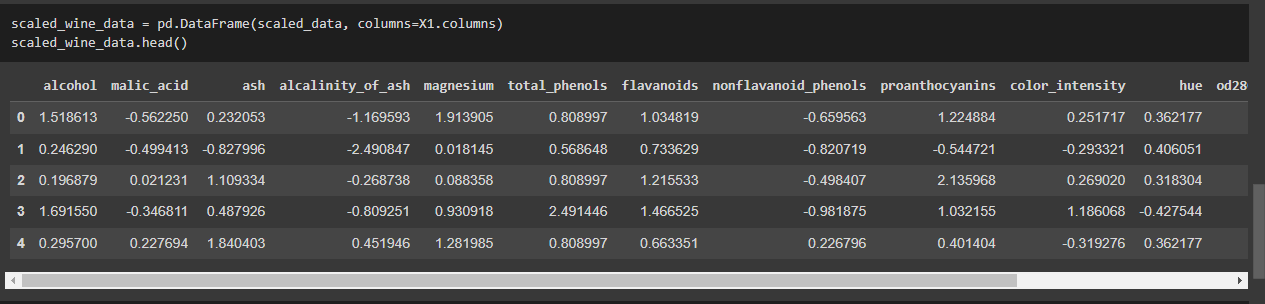
****

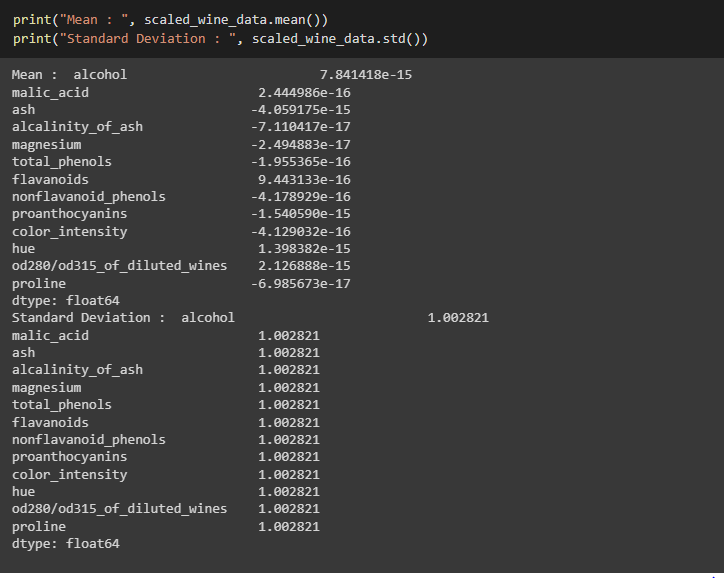
****

****

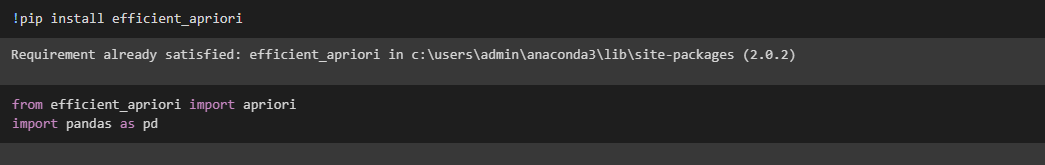
****

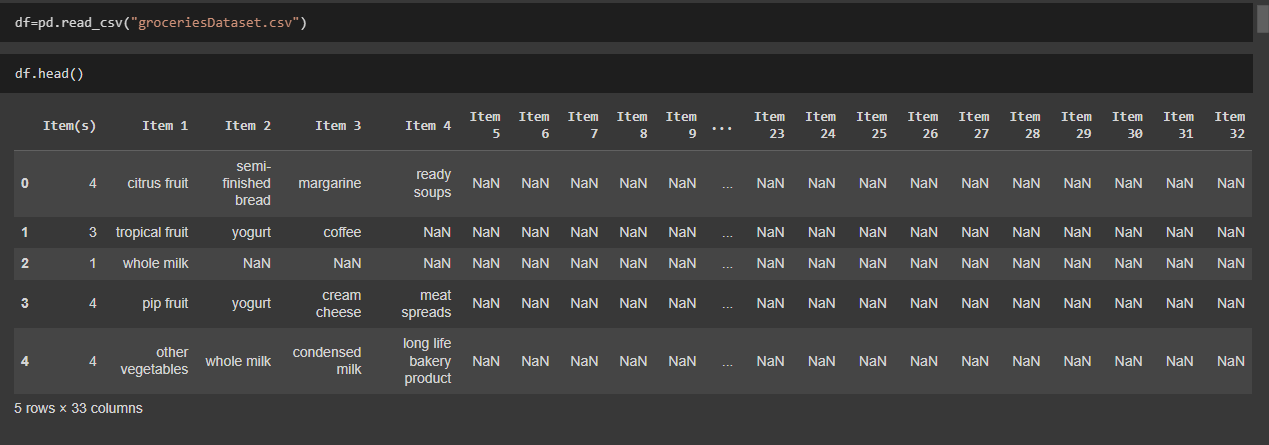
****

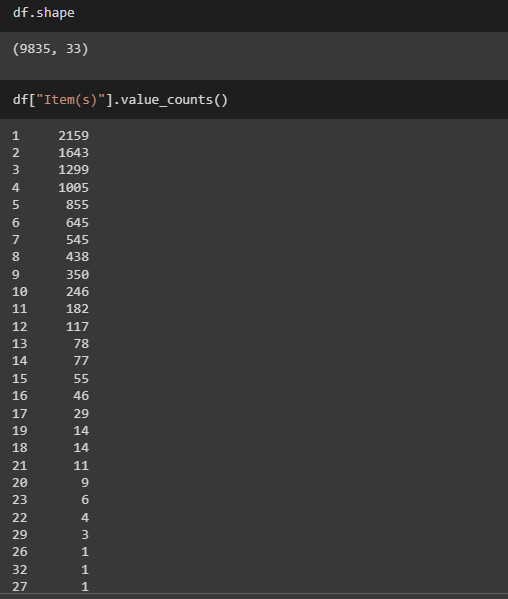
****

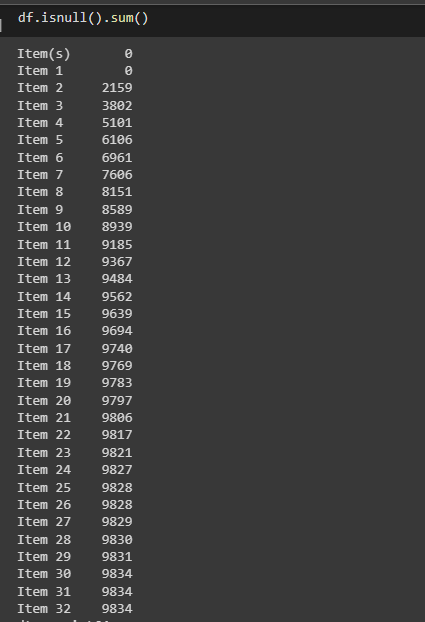
****

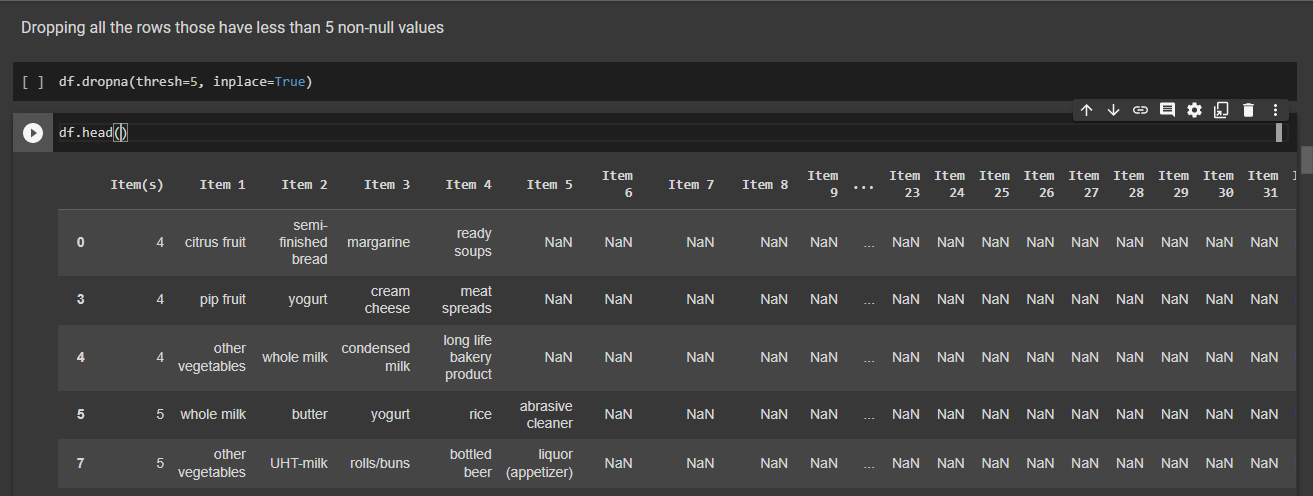
**Q4. Run Apriori algorithm to find frequent itemsets and association rules**

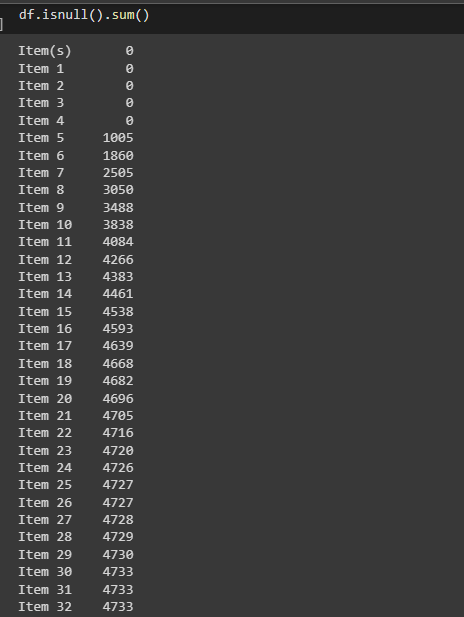
****

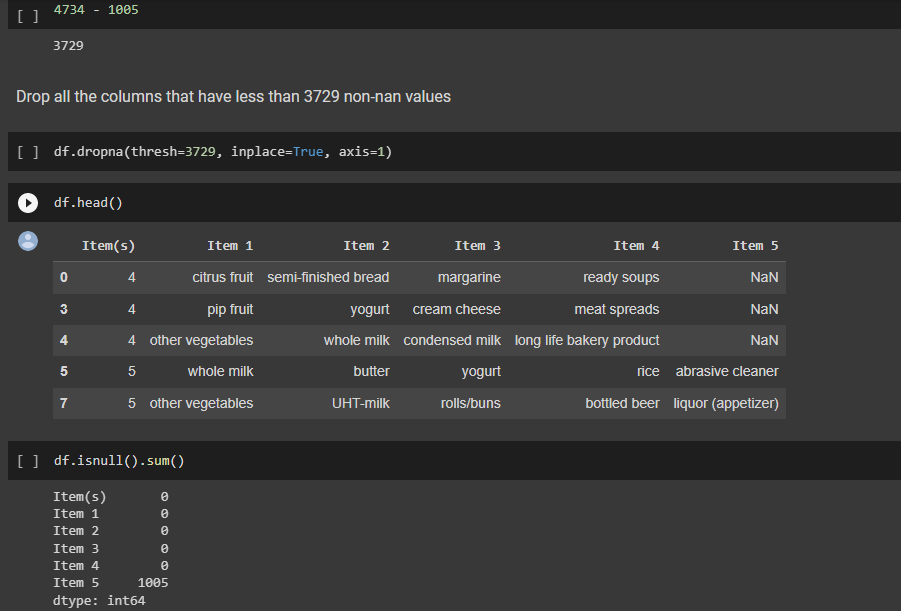
****

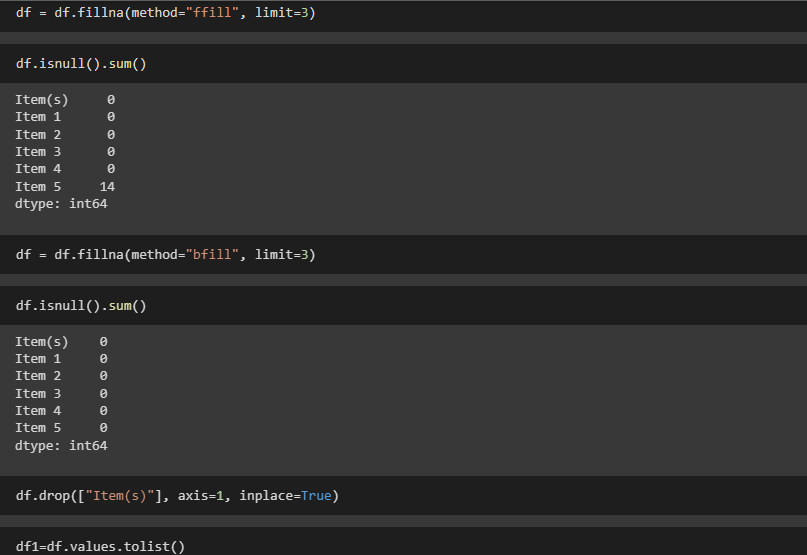
****

****

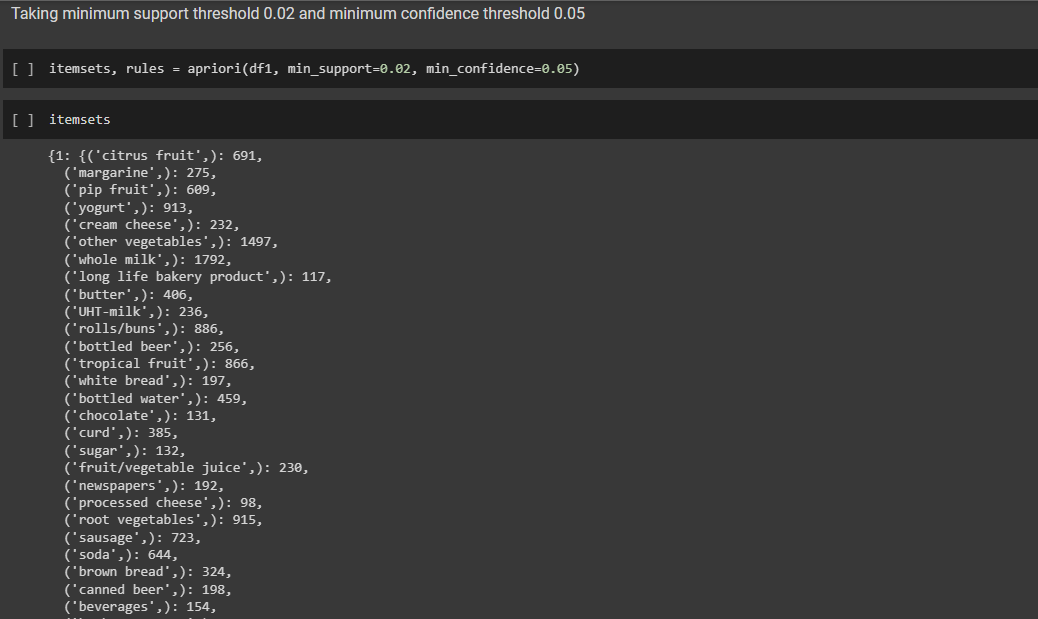
****

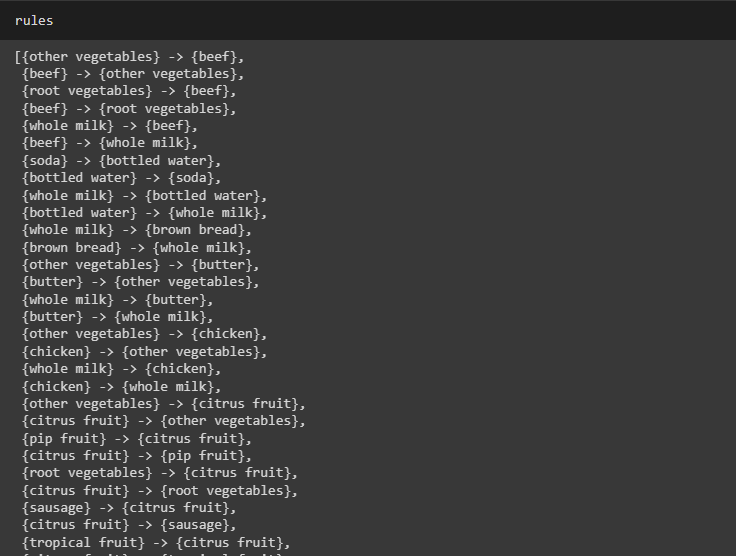
****

****

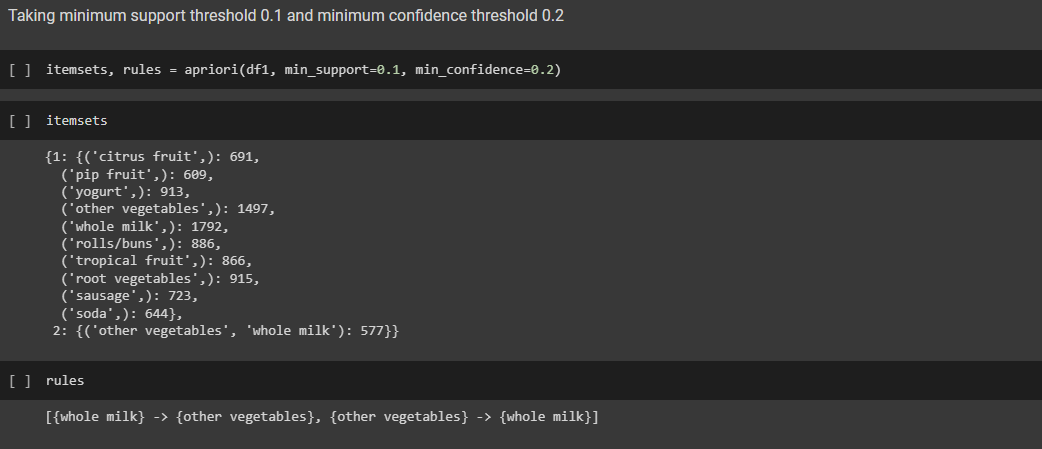
****

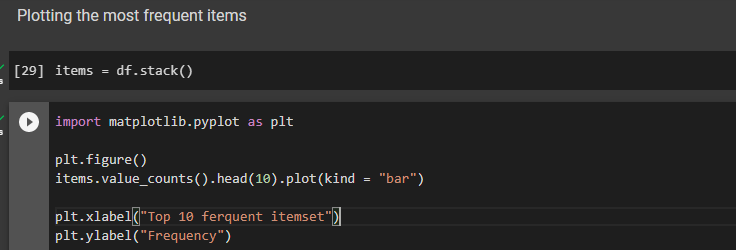
* 1. **Use minimum support as 2% and minimum confidence as 5%**

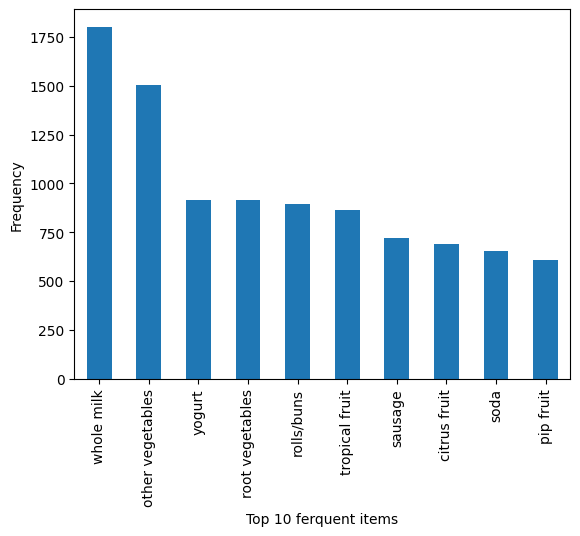
****

****

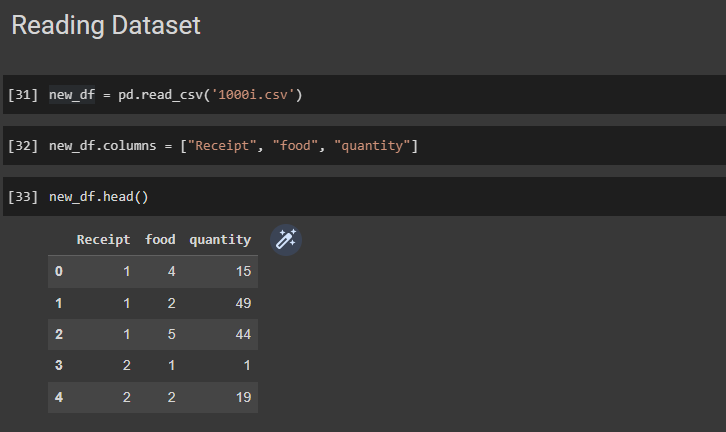
**4.2 Use minimum support as 10% and minimum confidence as 20 %**

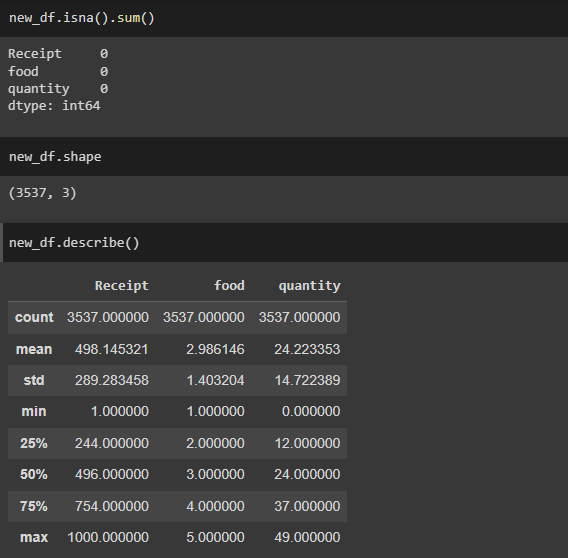
****

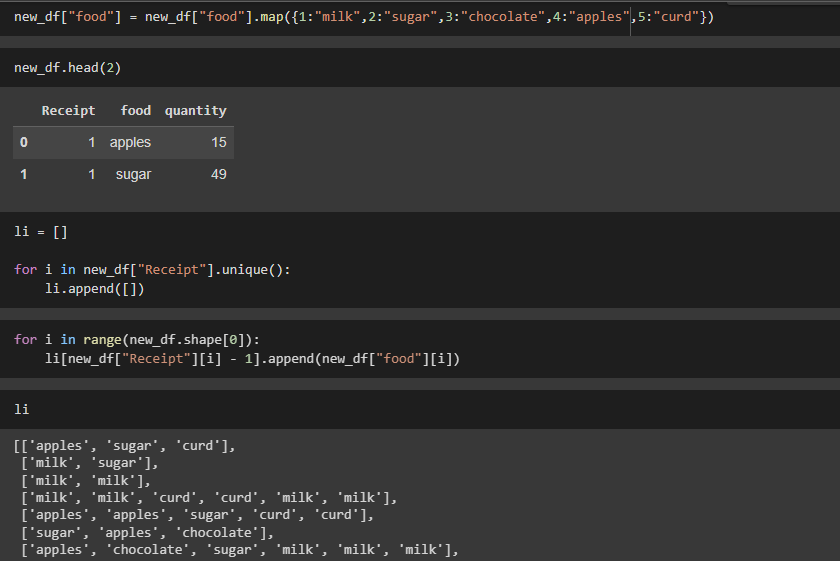
****

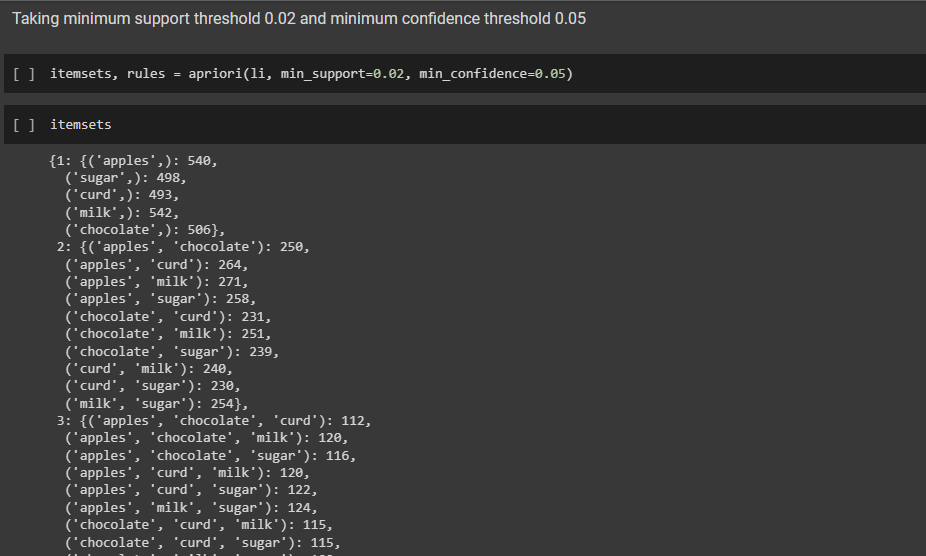
****

**Another Dataset**

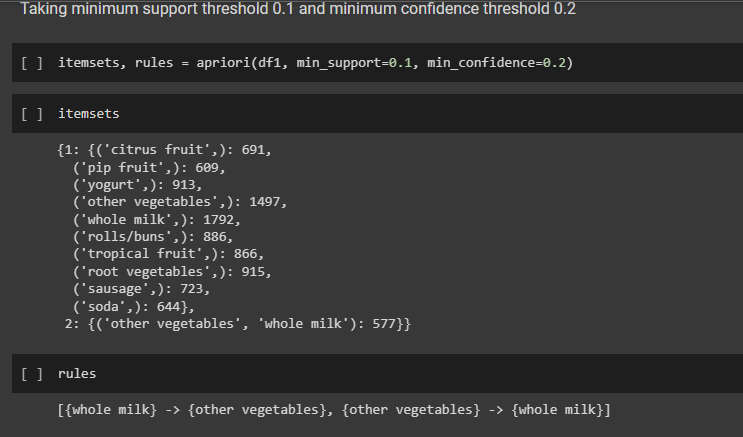
****

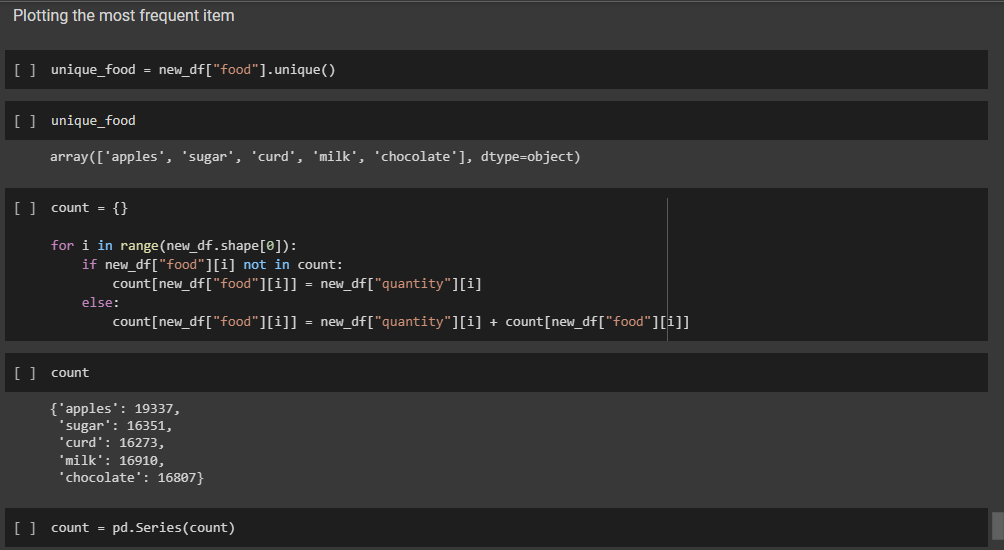
****

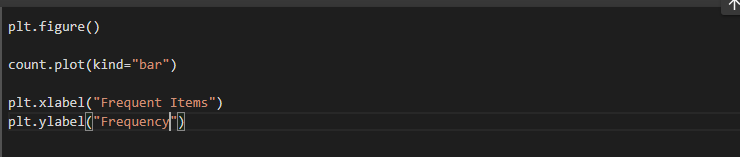
****

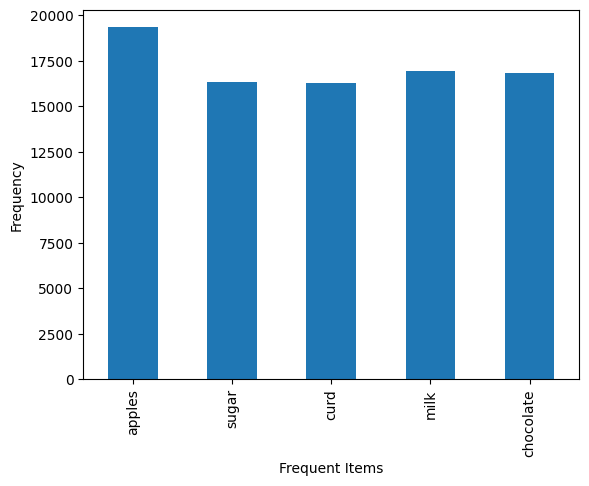
****

****

****

****

****

****

**Q5. Use Naive bayes, K-nearest, and Decision tree classification algorithms and build classifiers. Divide the data set into training and test set. Compare the accuracy of the different classifiers under the following situations:**

**5.1 a) Training set = 75% Test set = 25%**

**b) Training set = 66.6% (2/3rd of total), Test set = 33.3%**

**5.2 Training set is chosen by**

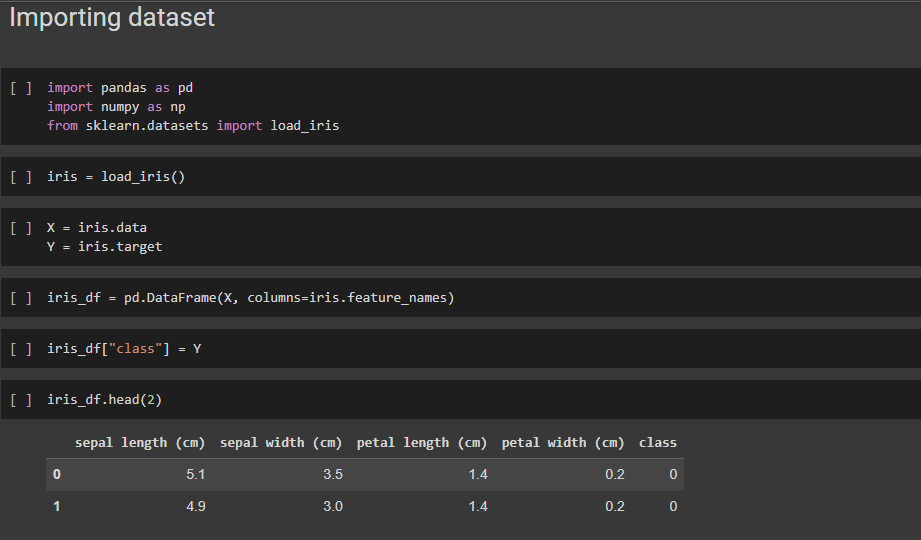
**i) hold out method**

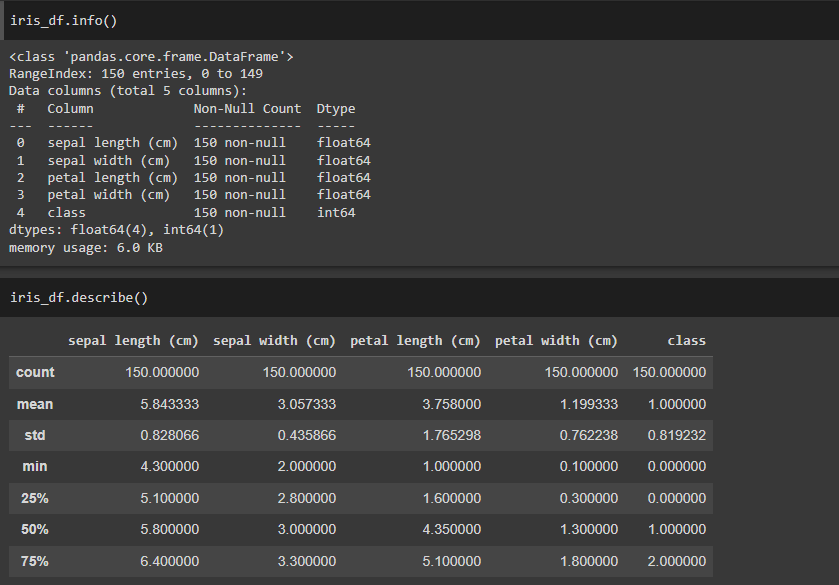
**ii) Random subsampling**

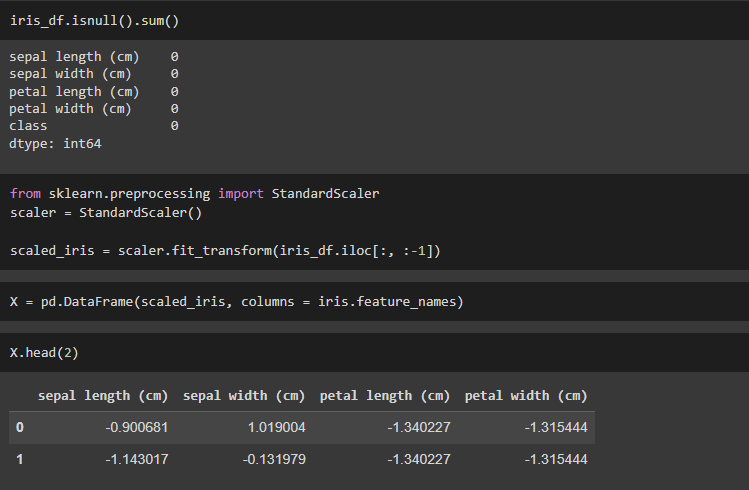
**iii) Cross-Validation. Compare the accuracy of the classifiers obtained.**

**5.3 Data is scaled to standard format.**

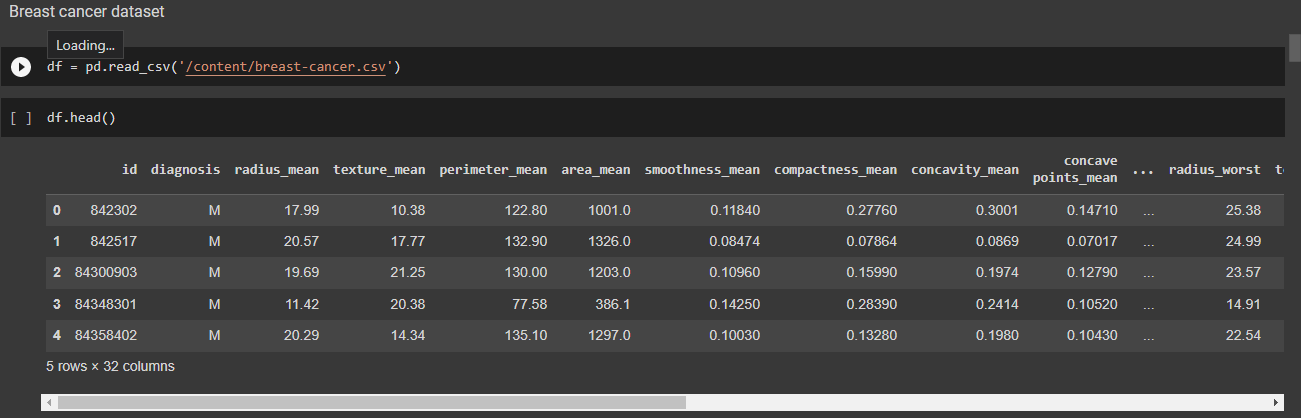
**For Iris Dataset**

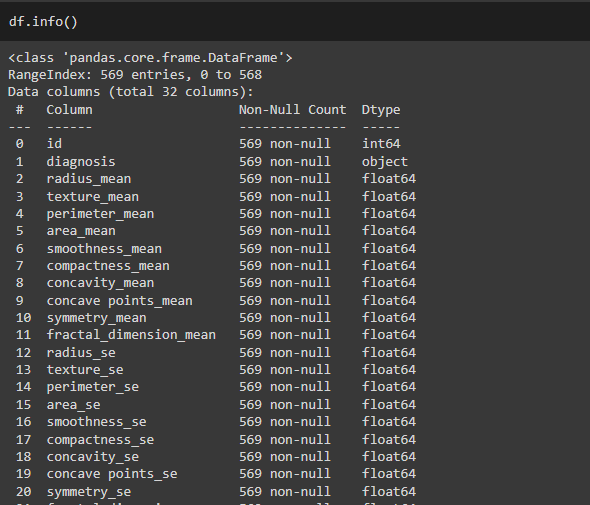
****

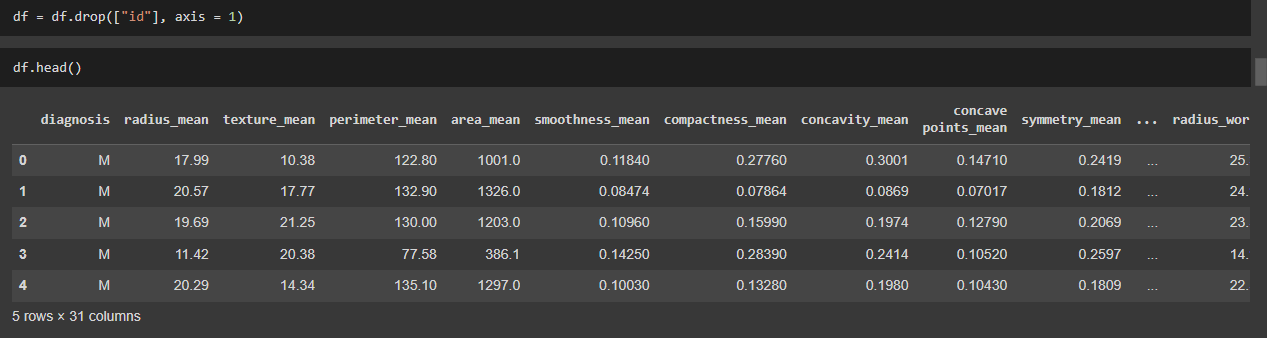
****

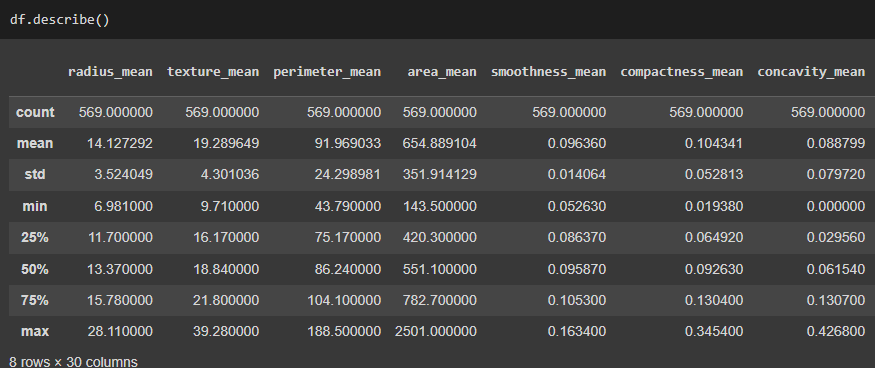
****

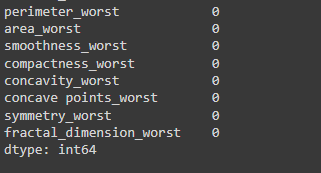
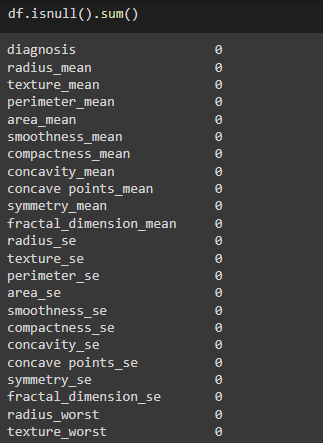
**For Breast Cancer Dataset:**

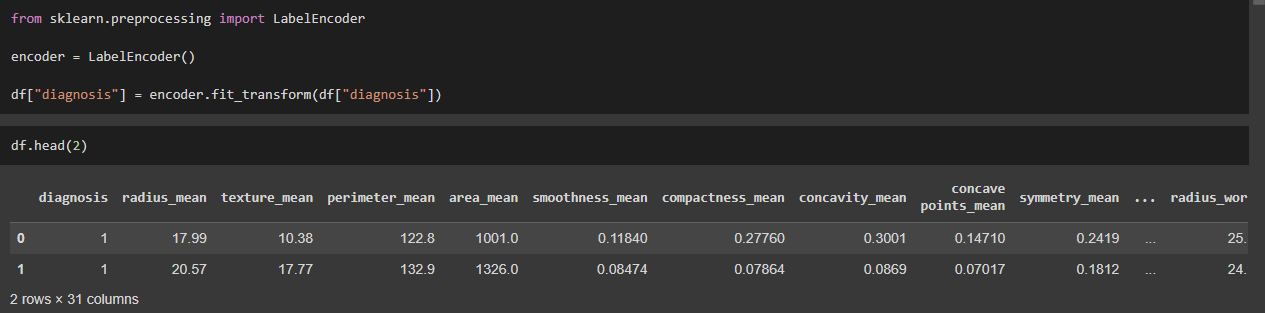
****

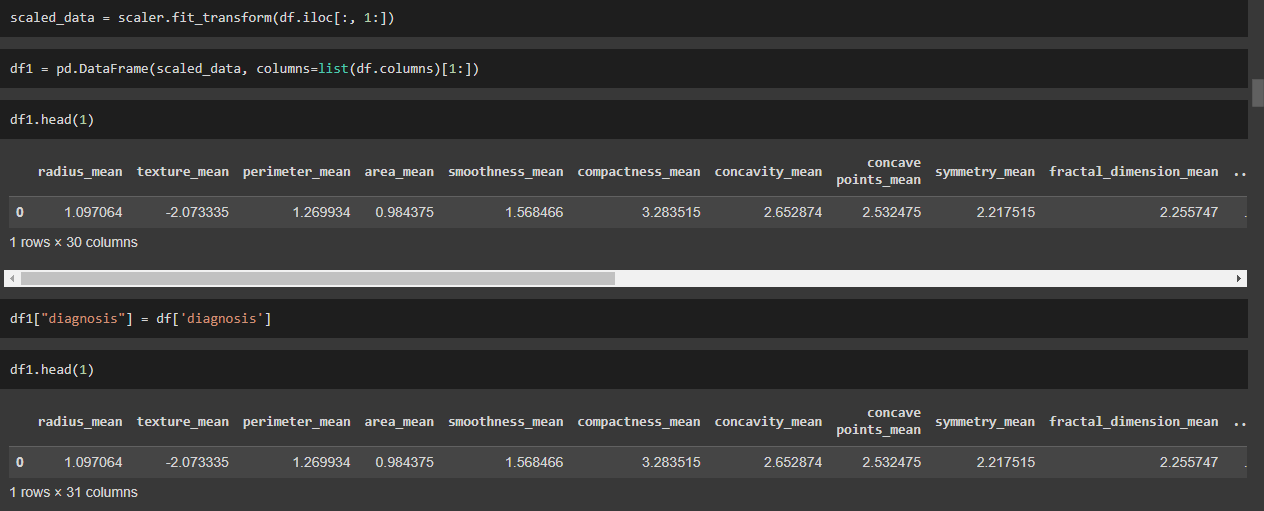
****

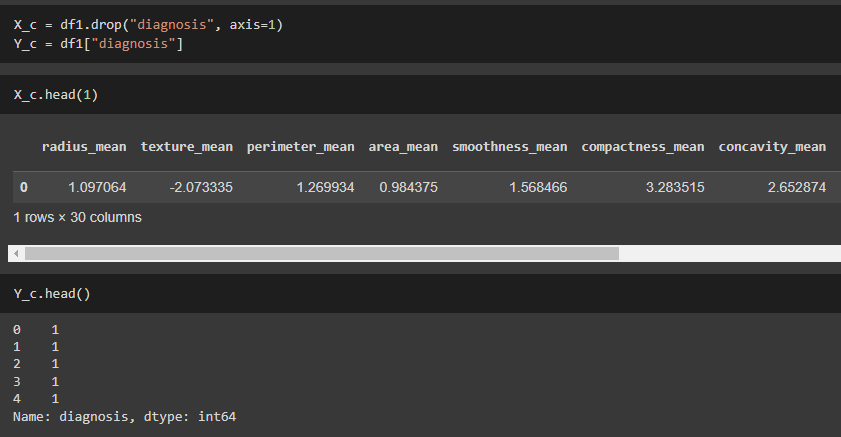
****

****

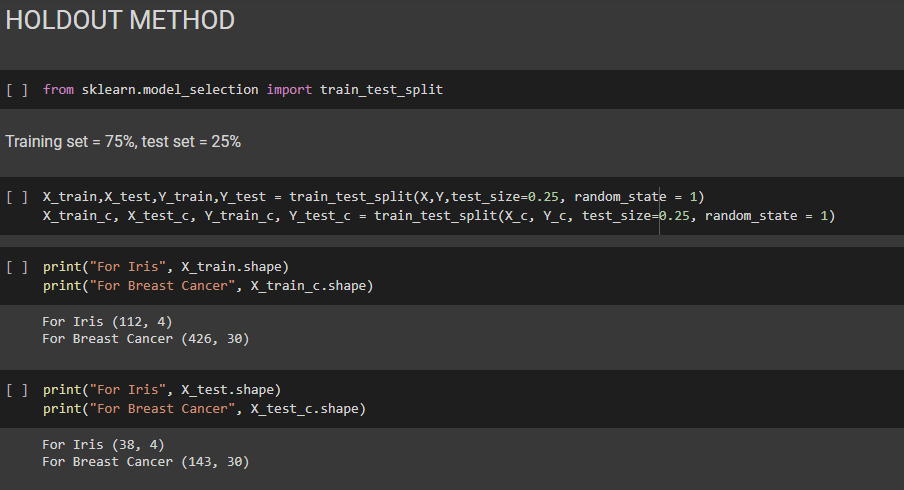
****

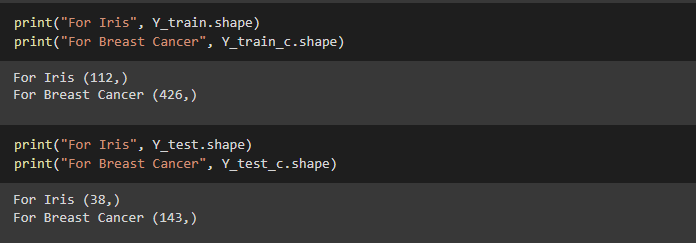
****

****

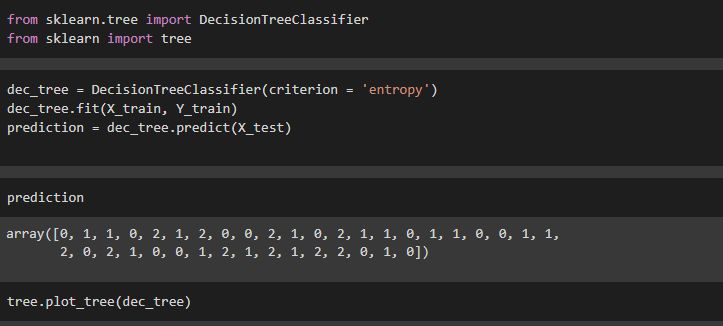
****

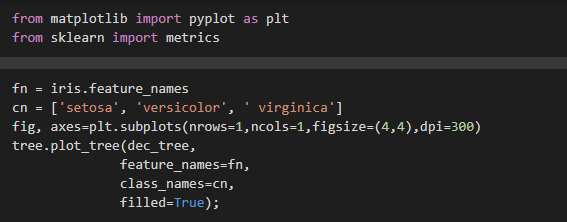
**Holdout Method**

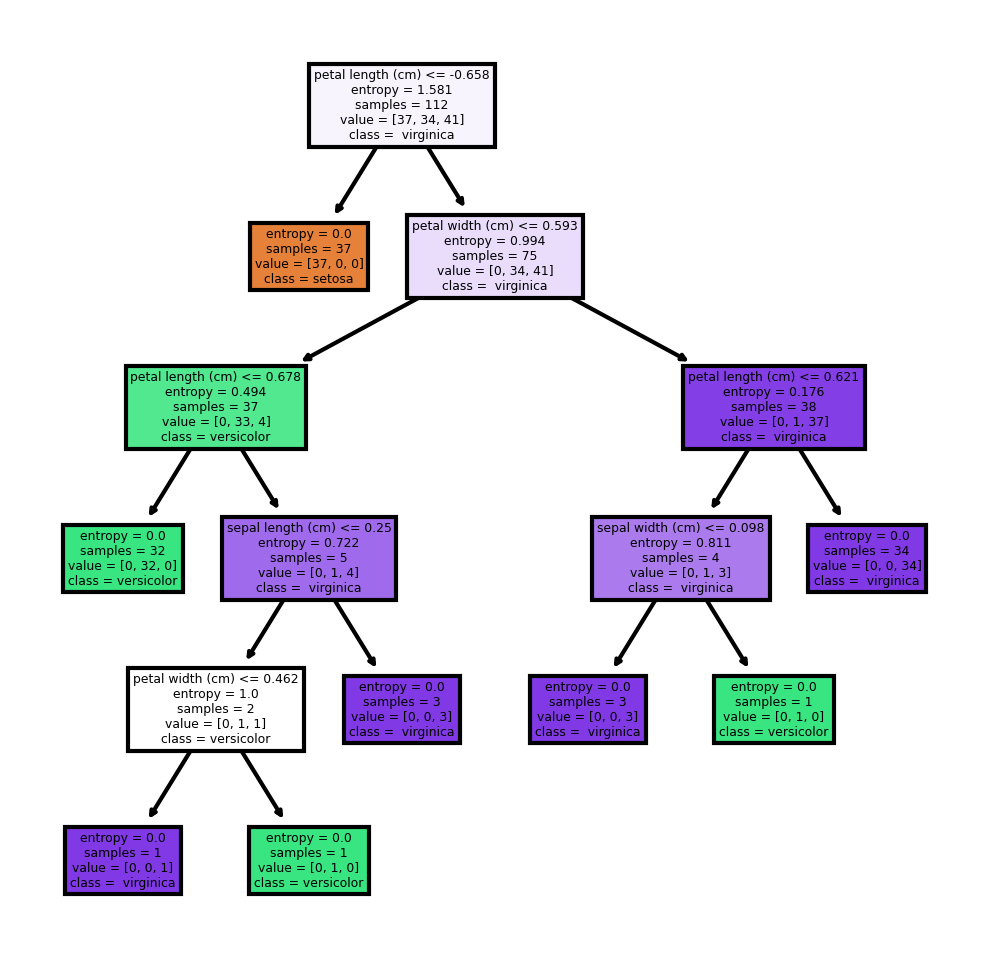
****

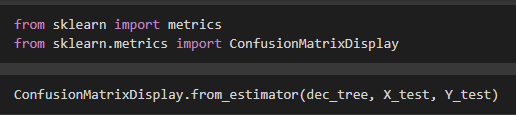
****

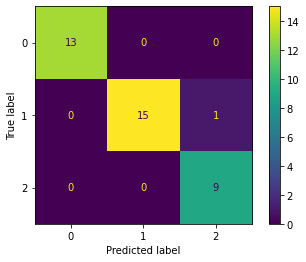
**Decision Tree (On Iris dataset).**

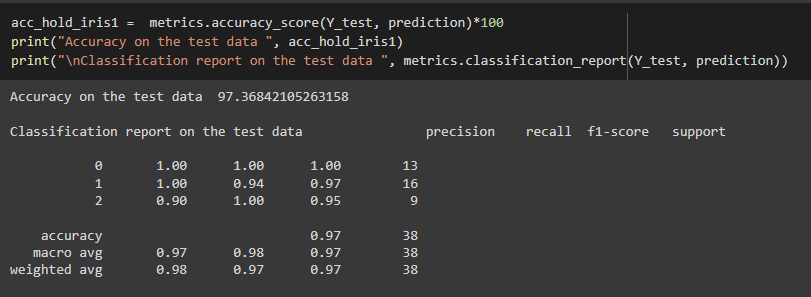
****

****

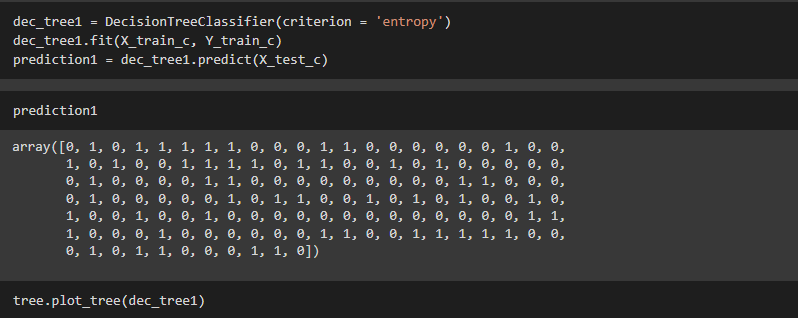
****

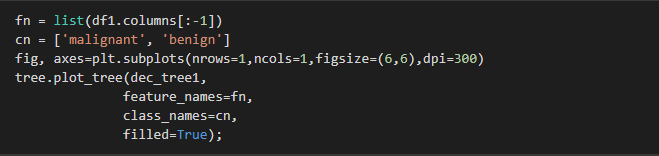
****

****

****

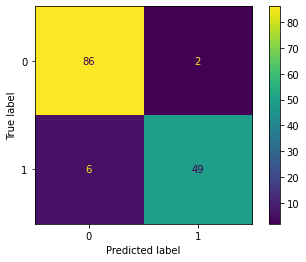
**Decision Tree (On Breast Cancer Dataset)**

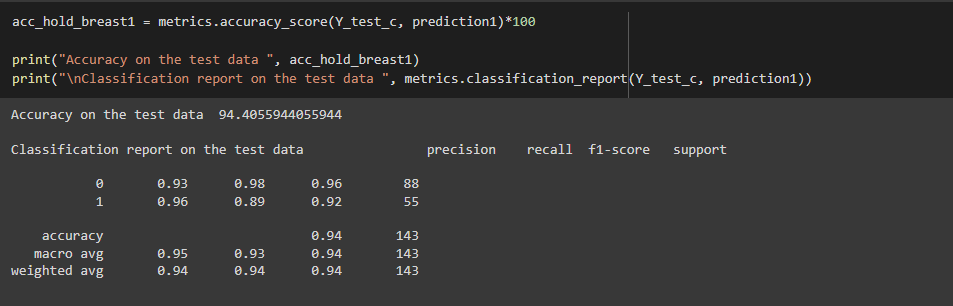
****

****

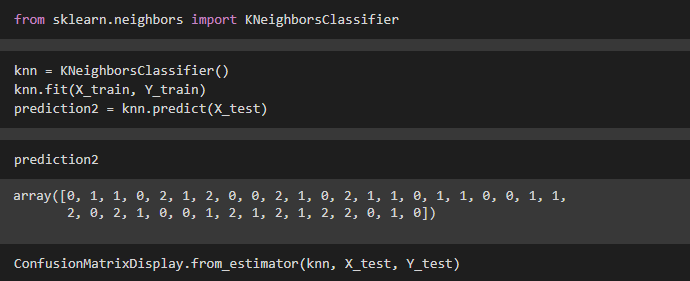
****

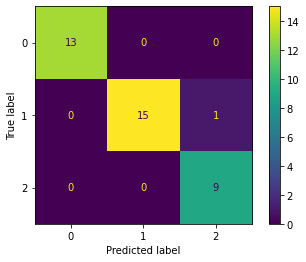
****

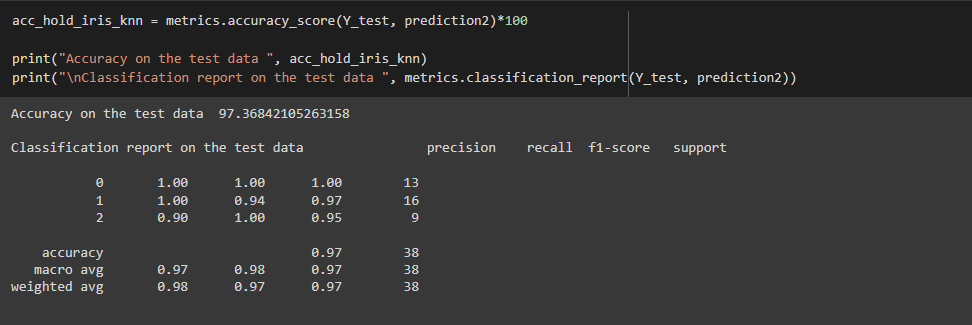
****

****

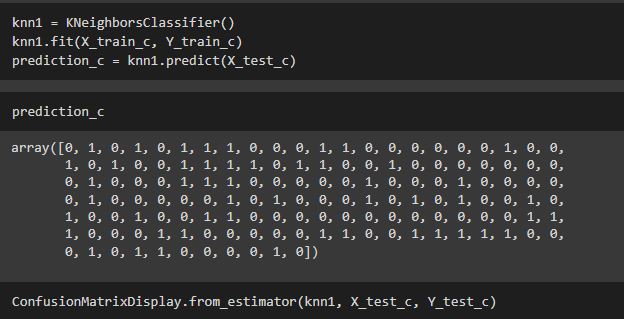
**KNN (On Iris Dataset)**

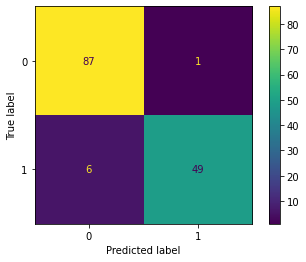
****

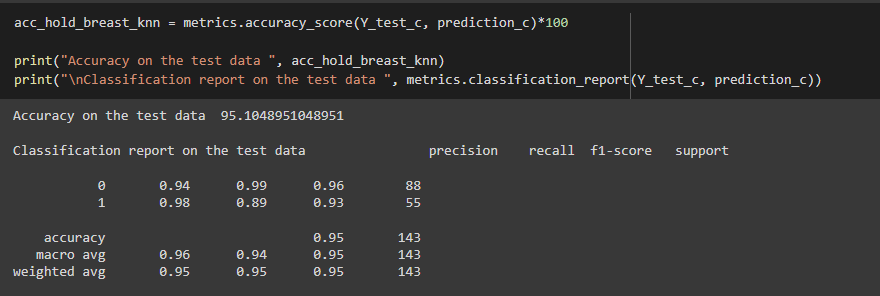
****

****

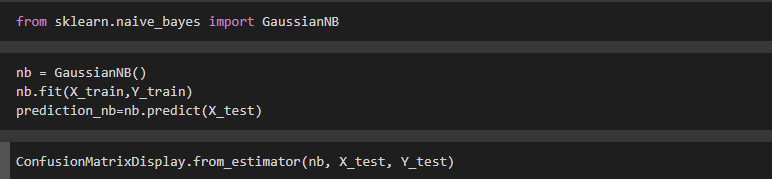
**KNN (On Breast Cancer Dataset)**

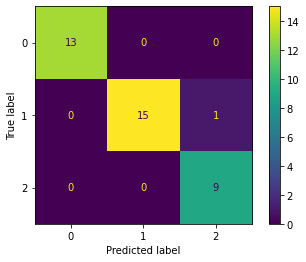
****

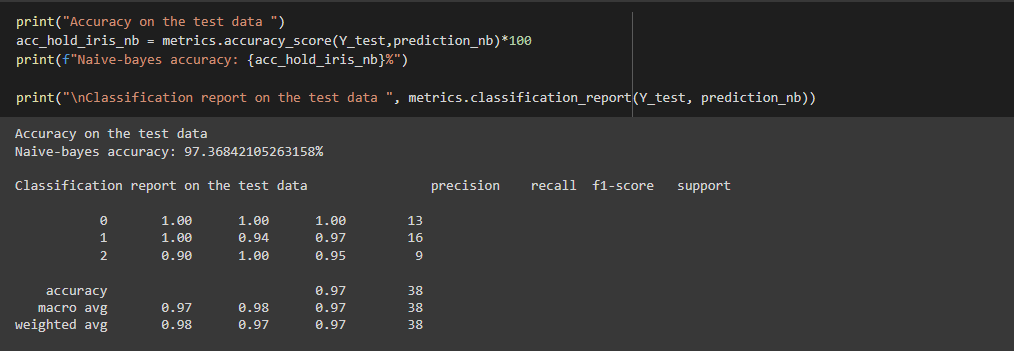
****

****

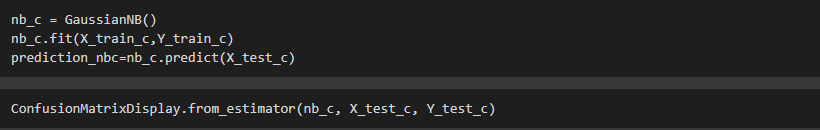
**Naïve Bayes (On Iris Dataset)**

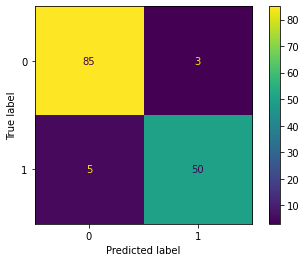
****

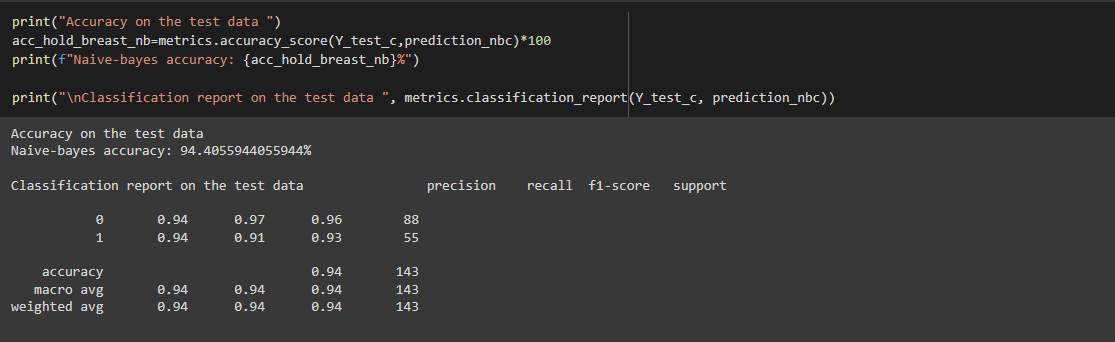
****

****

**Naïve Bayes (On Breast Cancer Dataset)**

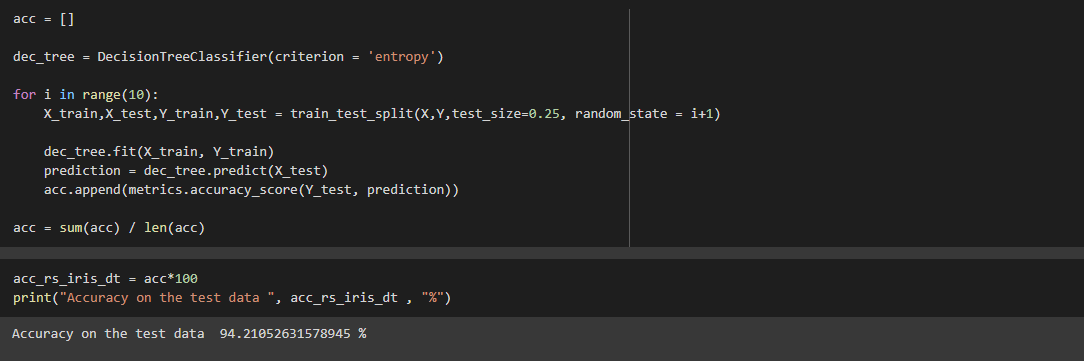
****

****

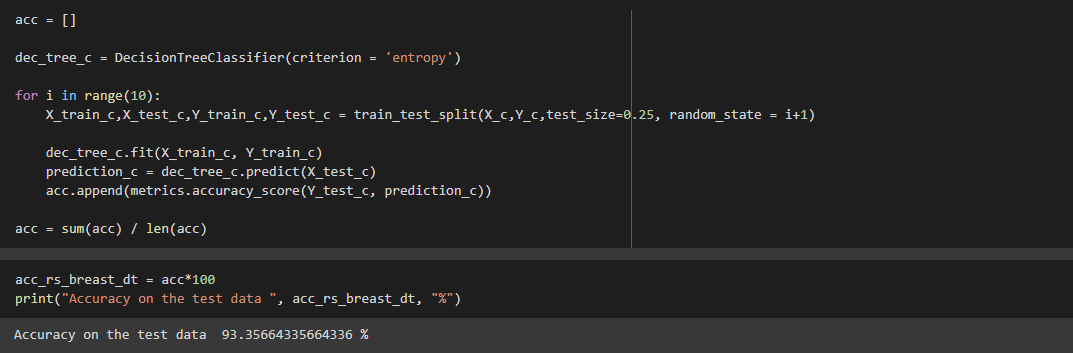
****

**Random Subsampling**

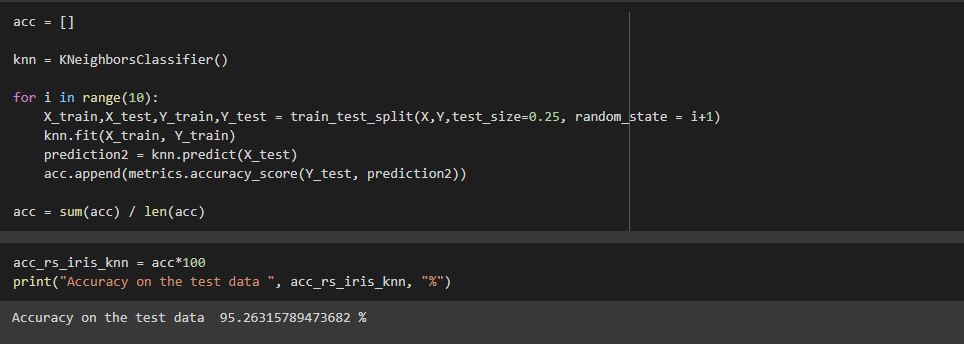
**Decision Tree (On Iris Dataset)**

****

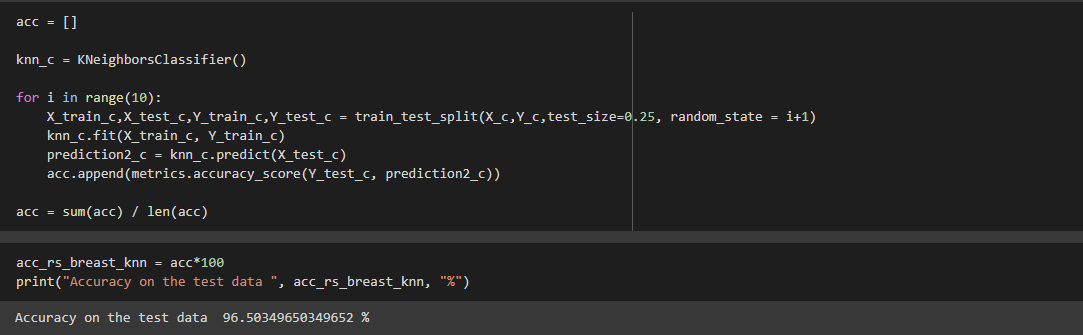
**Decision Tree (Breast Cancer Dataset)**

****

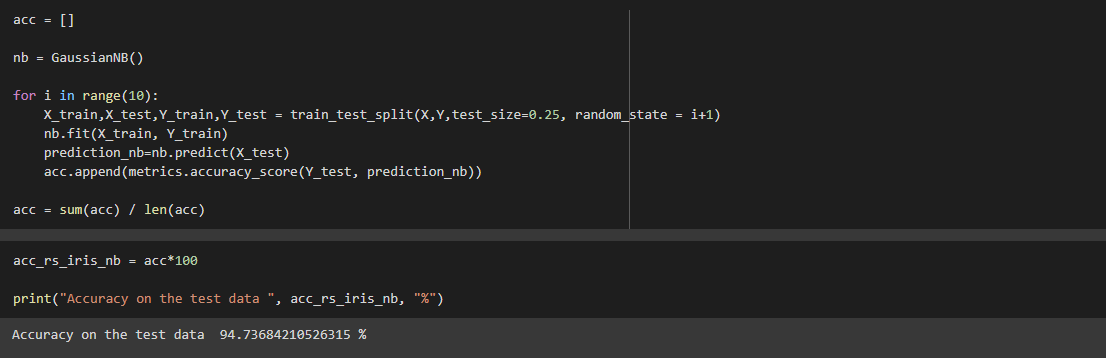
**KNN (Iris Dataset)**

****

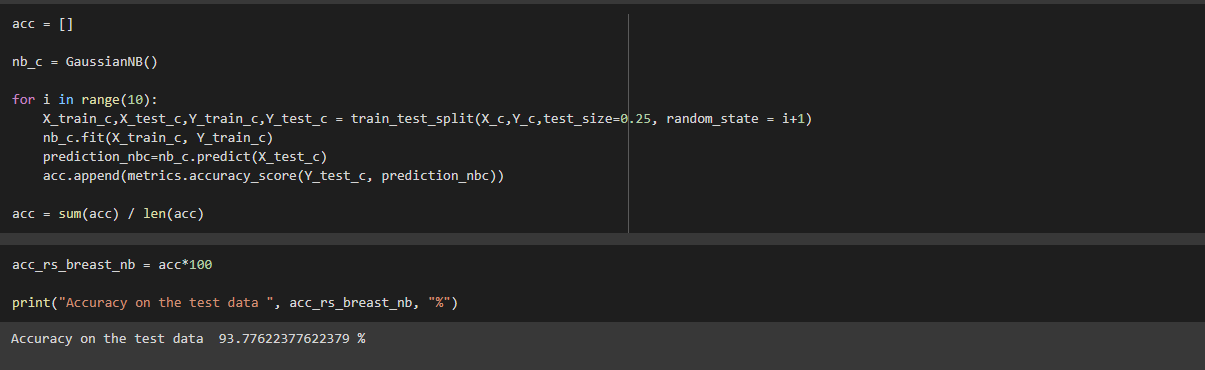
**KNN (Breast Cancer Dataset)**

****

**Naïve Bayes (On Iris Dataset)**

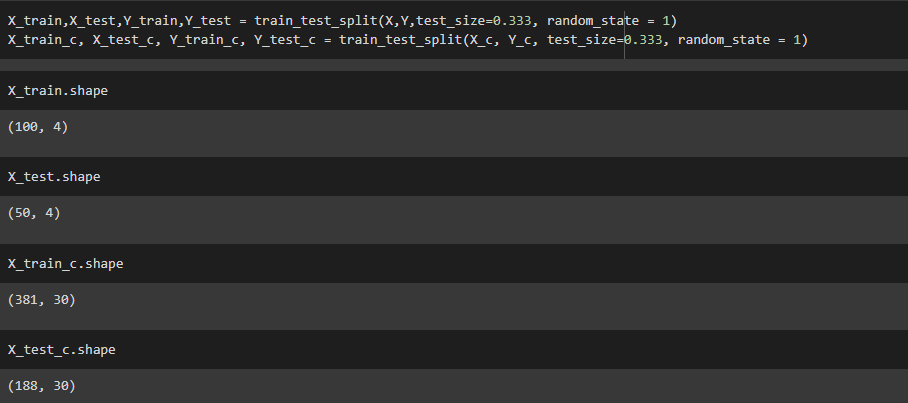
****

**Naïve Bayes (On Breast Cancer Dataset)**

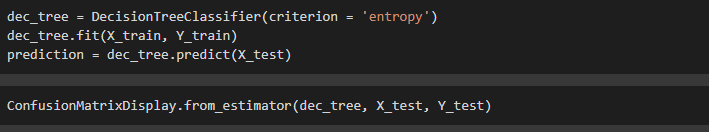
****

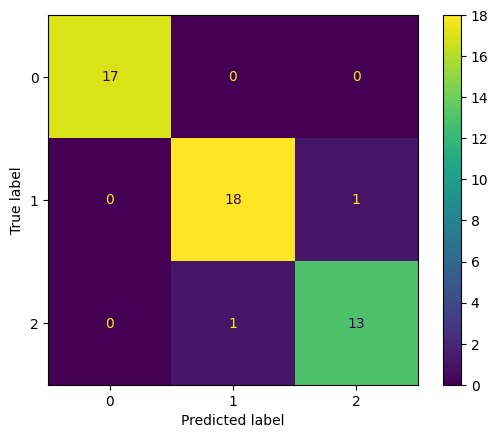
**Training set = 66.6% (2/3rd of total), Test set = 33.3%**

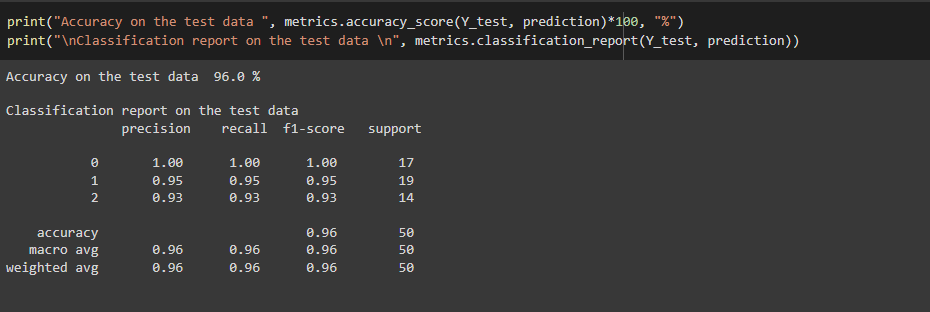
**Holdout Method**

****

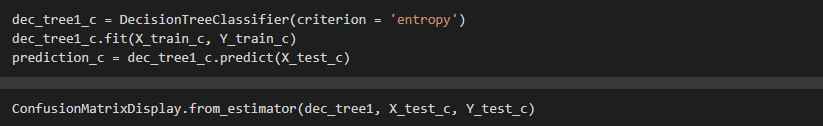
**Decision Tree (On Iris Dataset)**

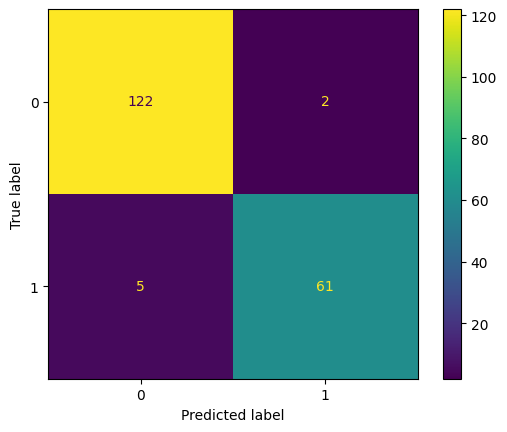
****

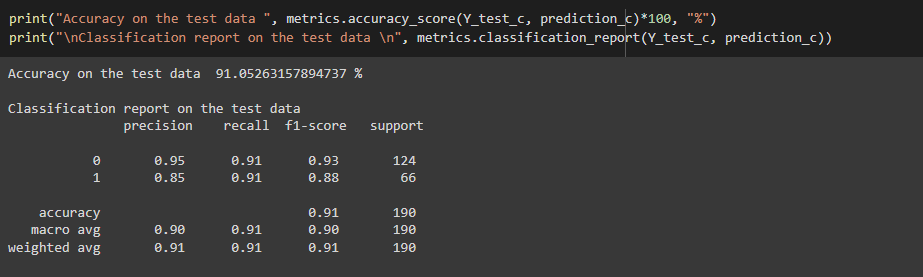
****

****

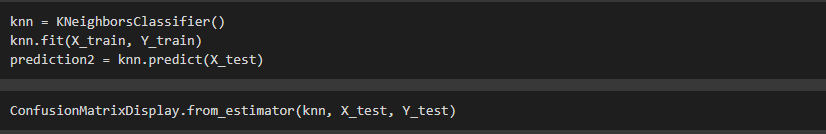
**Decision Tree (On Breast Cancer Dataset)**

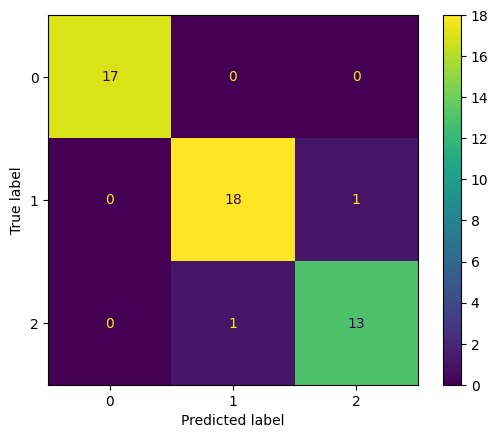
****

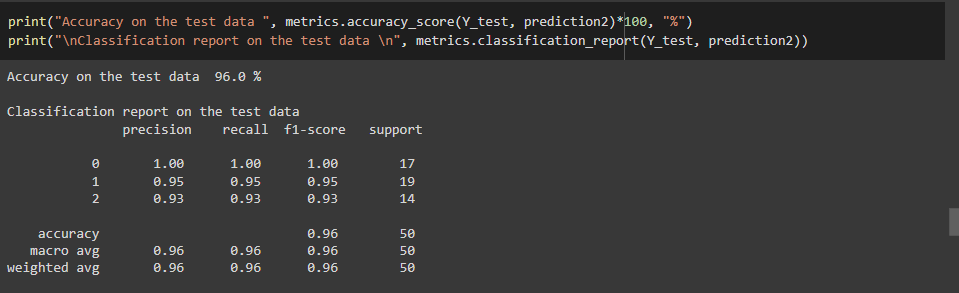
****

****

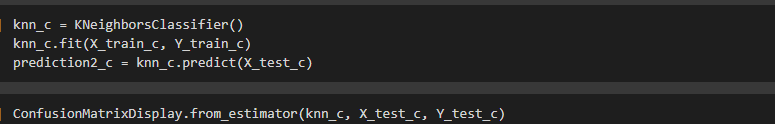
**KNN (On Iris Dataset)**

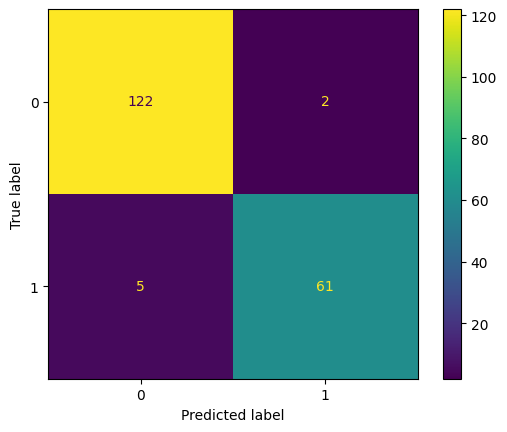
****

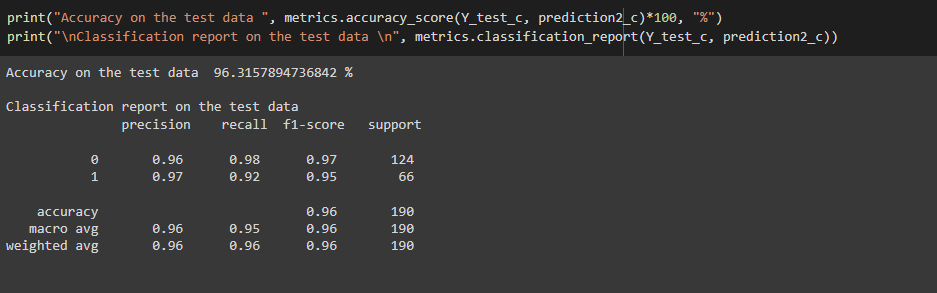
****

****

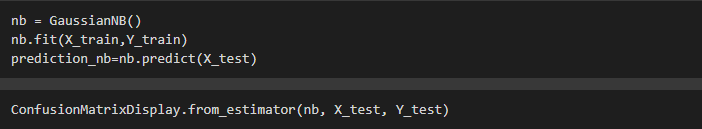
**KNN (On Breast Cancer Dataset)**

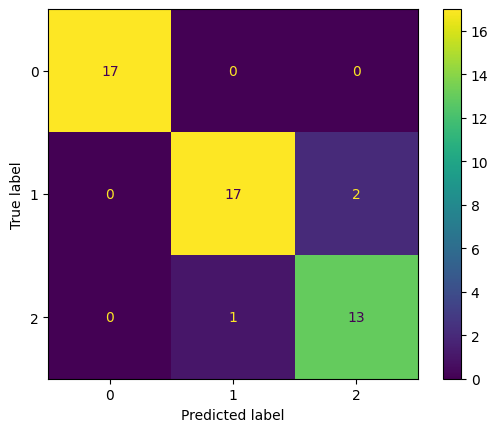
****

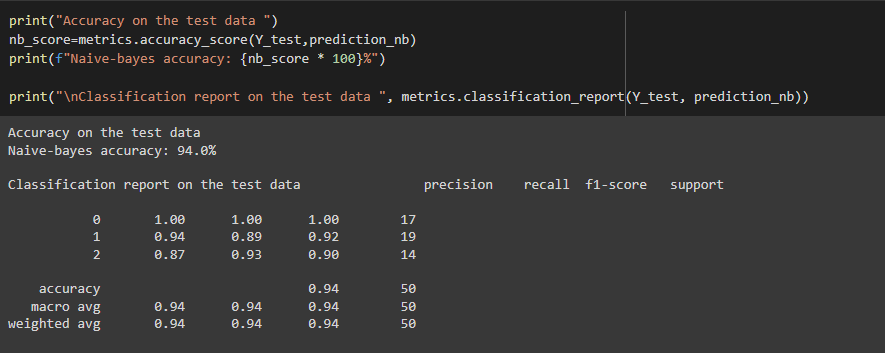
****

****

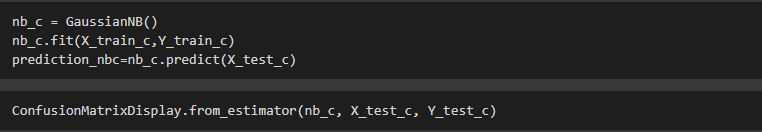
**Naïve Bayes (On Iris Dataset)**

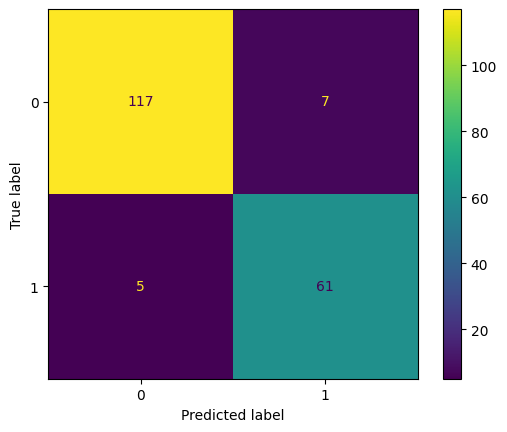
****

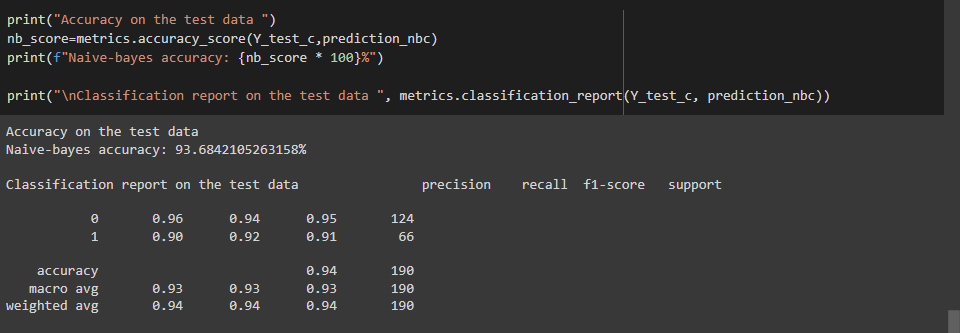
****

****

**Naïve Bayes (On Breast Cancer Dataset)**

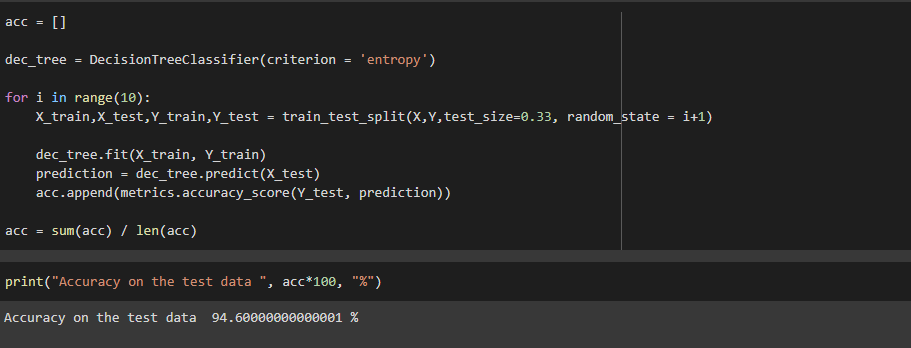
****

****

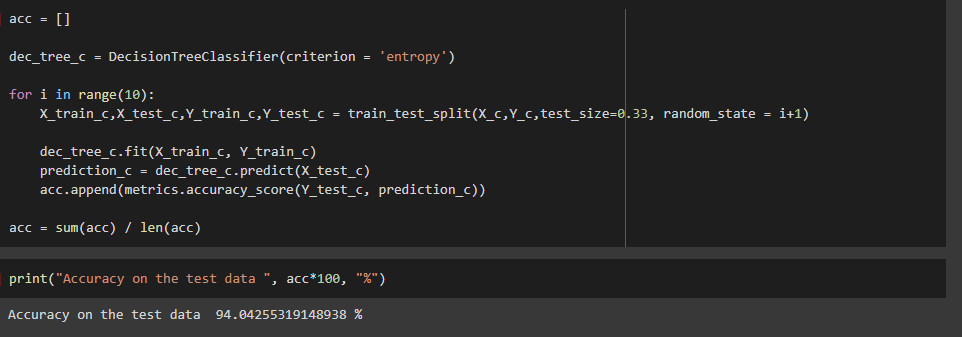
****

**Random Subsampling**

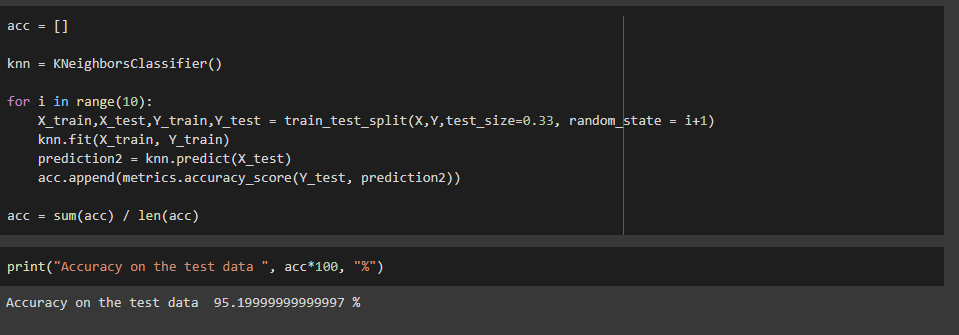
**Decision Tree (On Iris dataset)**

****

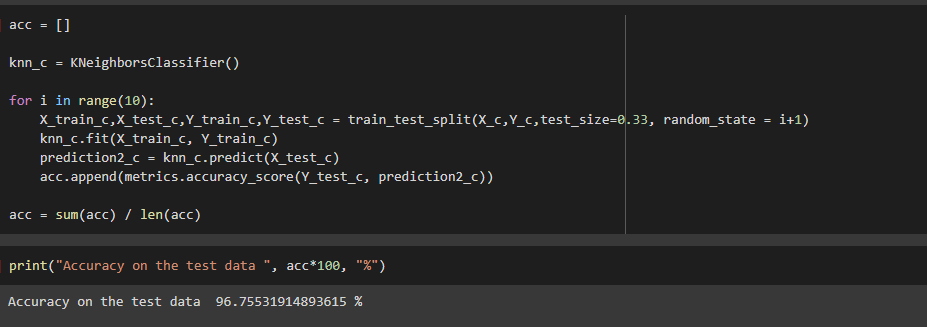
**Decision Tree (On Breast Cancer dataset)**

****

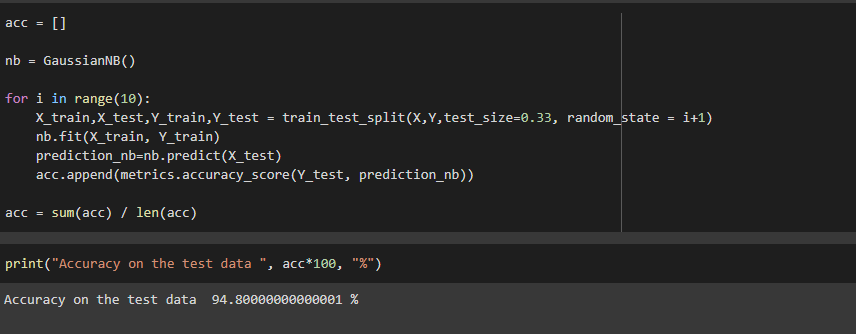
**KNN (On Iris Dataset)**

****

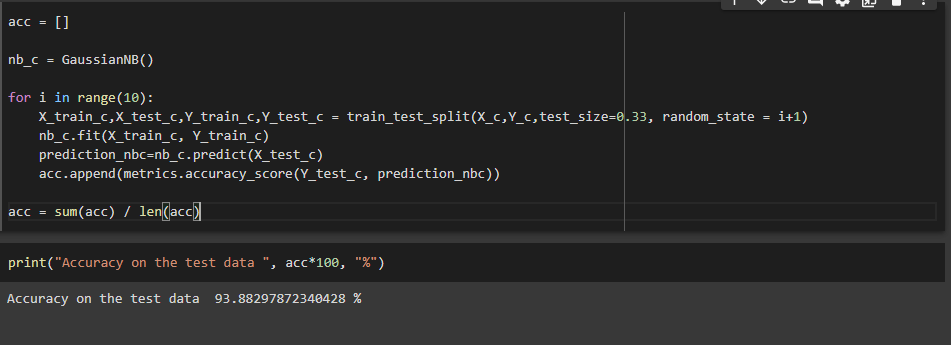
**KNN (On Breast Cancer Dataset)**

****

**Naïve Bayes (On Iris Dataset)**

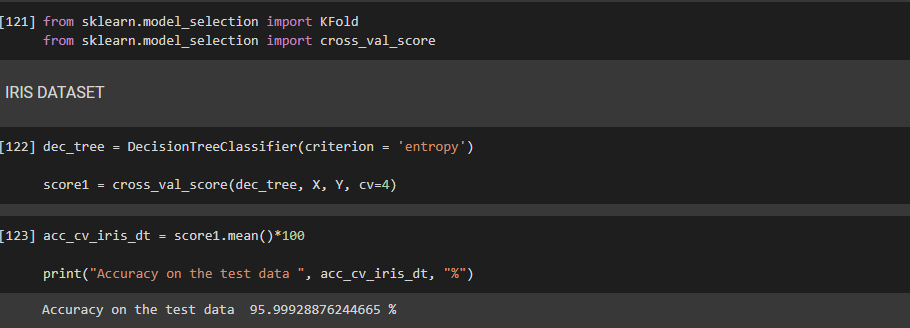
****

**Naïve Bayes (On Breast Cancer Dataset)**

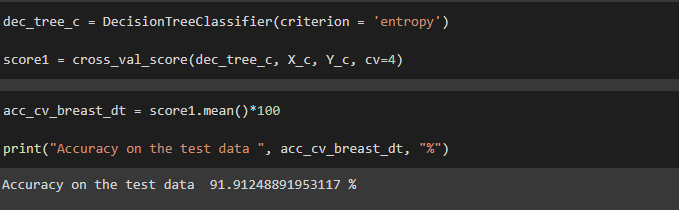
****

**Cross Validation**

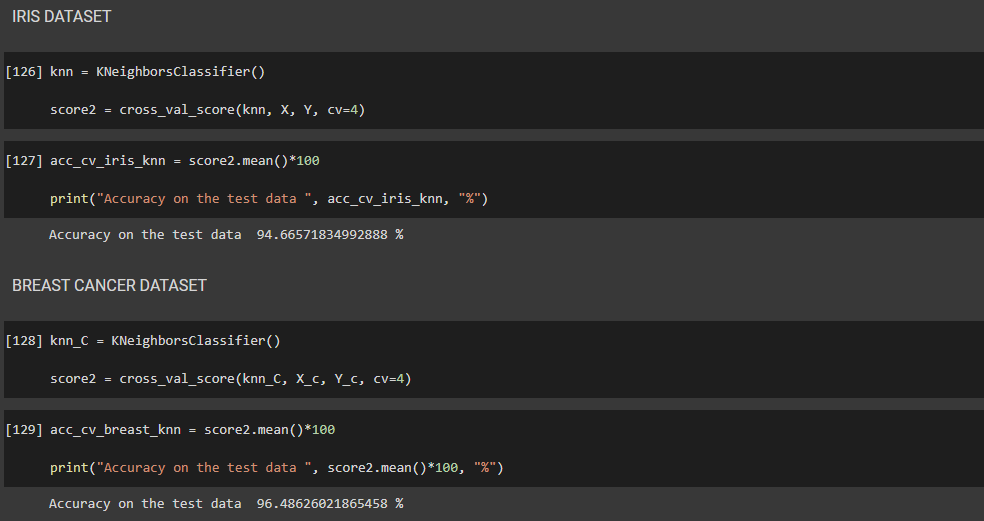
**Decision Tree (On Iris Dataset)**

****

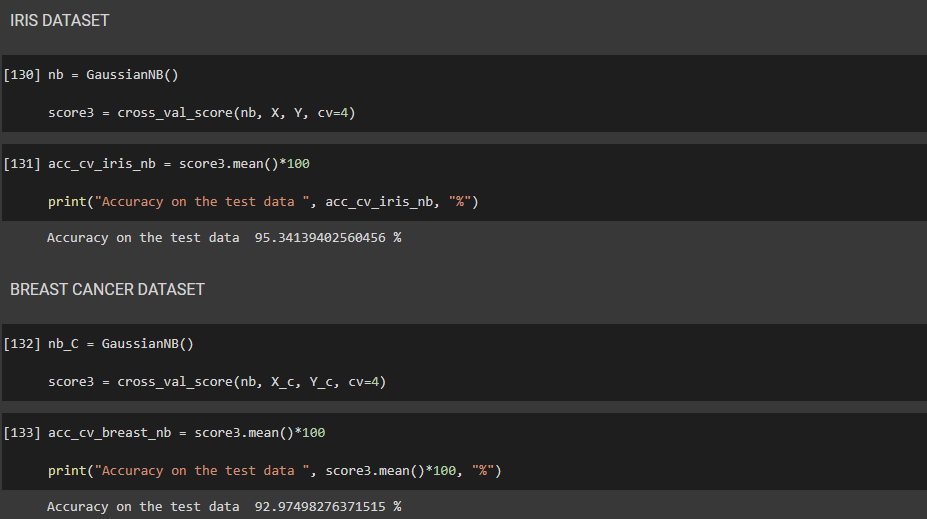
**Decision Tree (On Breast Cancer Dataset)**

****

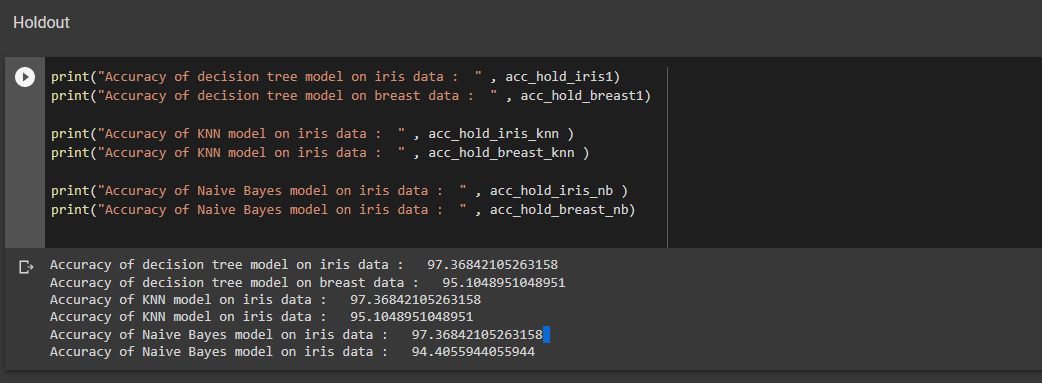
**KNN (On Iris and Breast Cancer Dataset)**

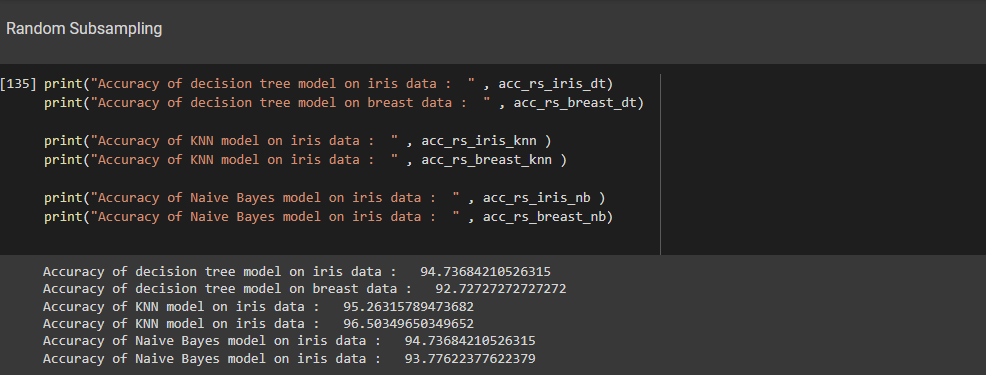
****

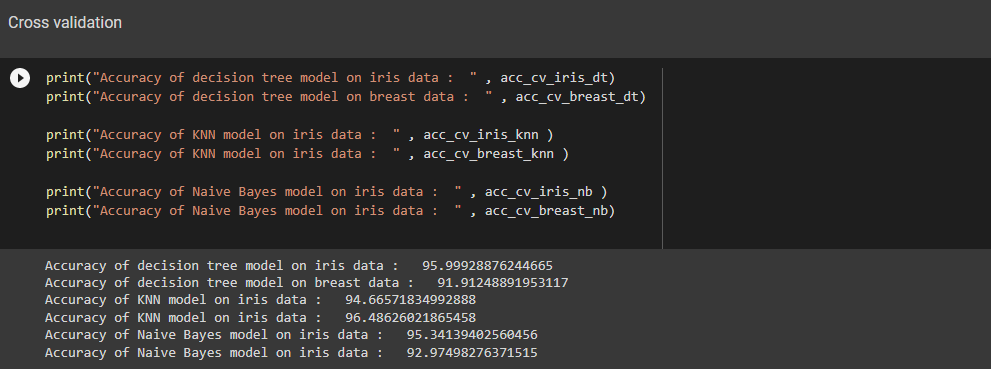
**Naïve bayes (On Iris and Breast Cancer Dataset)**

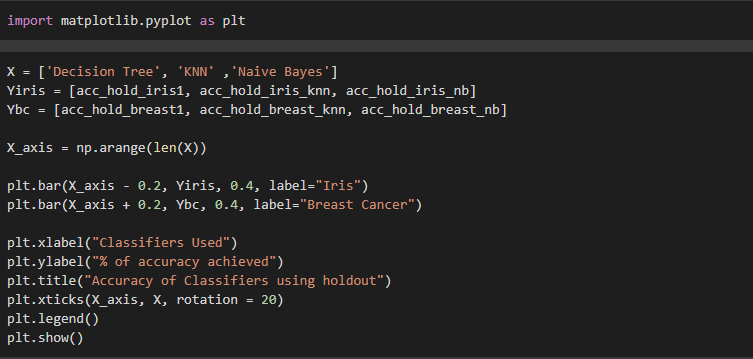
****

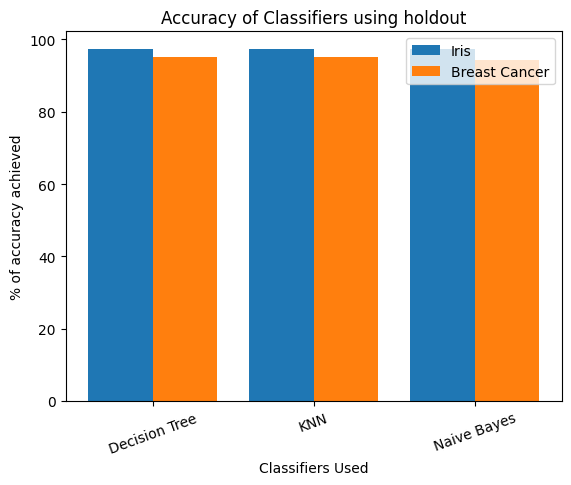
**Result:**

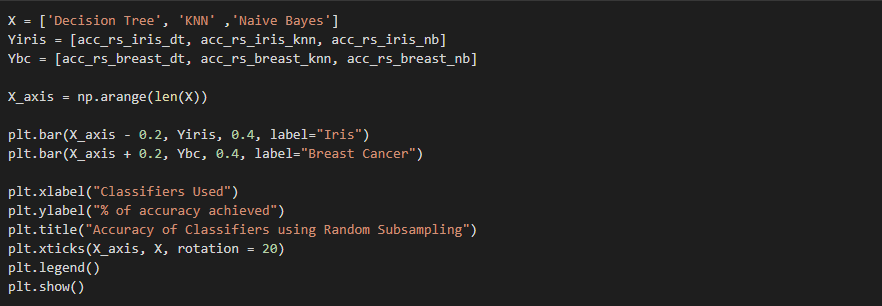
****

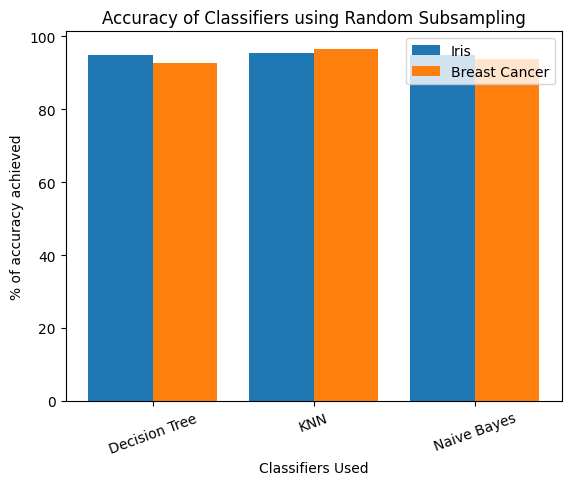
****

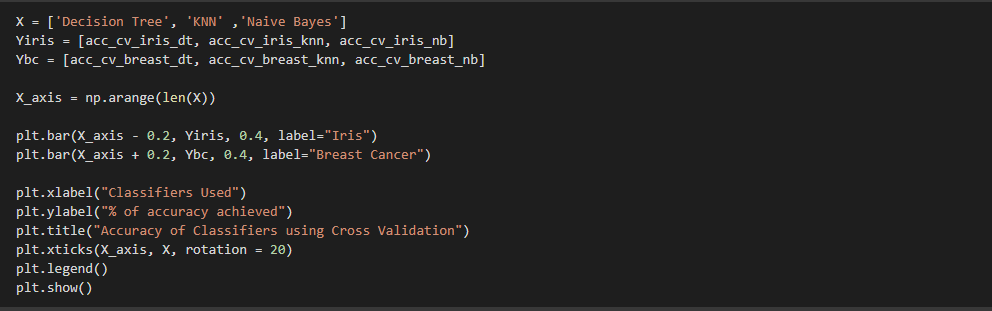
****

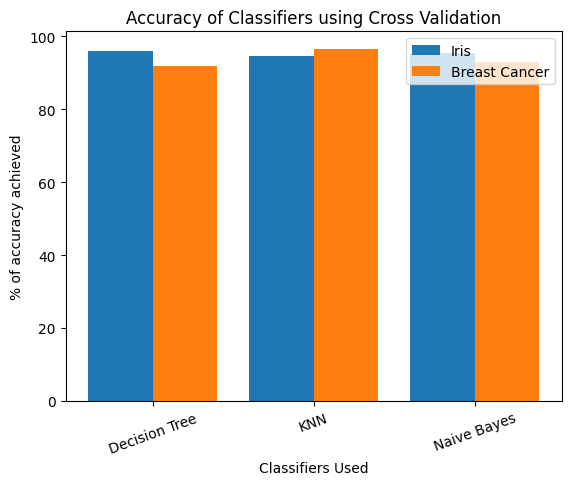
****

****

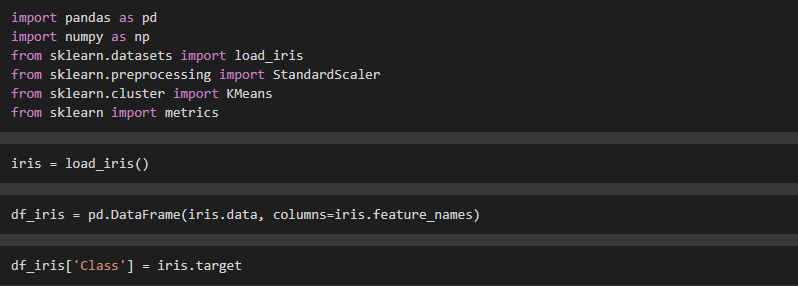
****

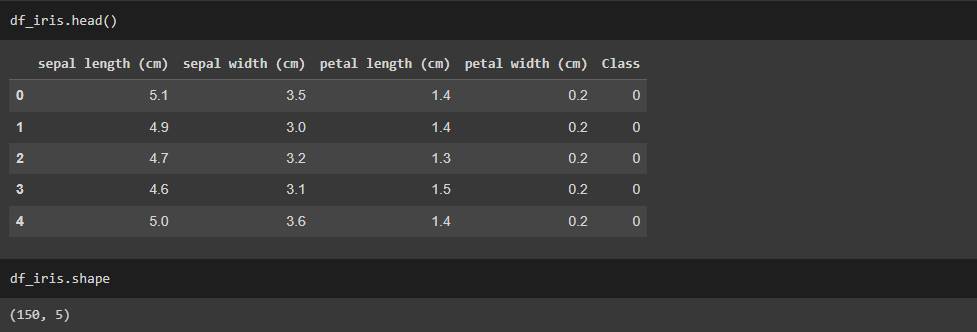
****

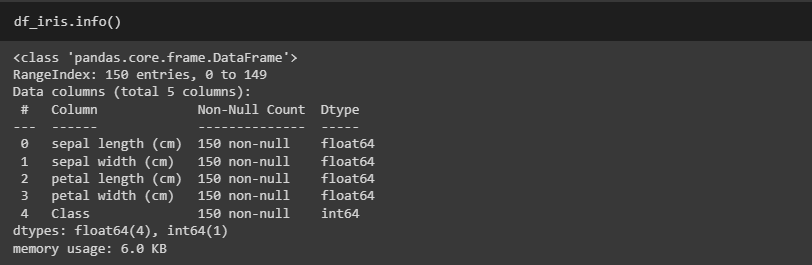
****

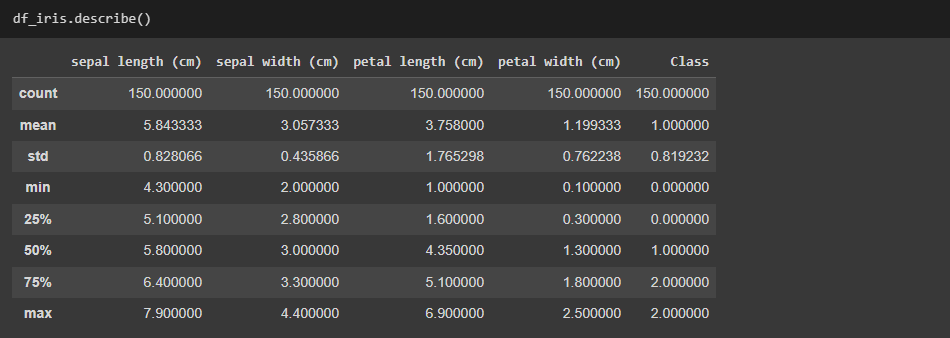
****

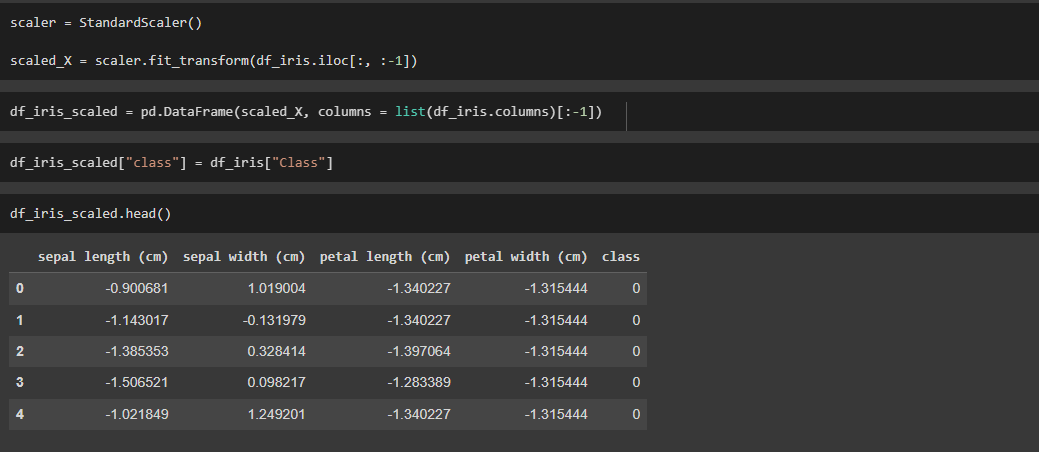
**Q6. Use Simple Kmeans, DBScan, Hierachical clustering algorithms for clustering. Compare the performance of clusters by changing the parameters involved in the algorithms.**

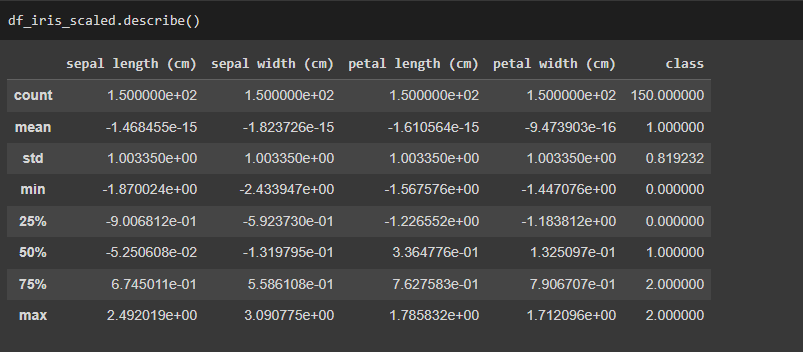
****

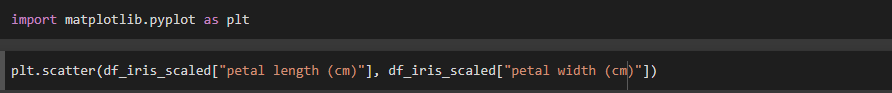
****

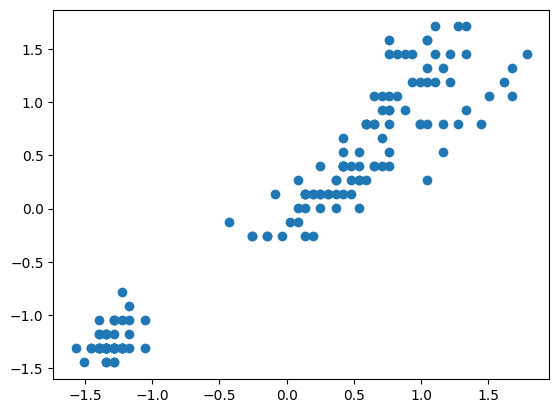
****

****

****

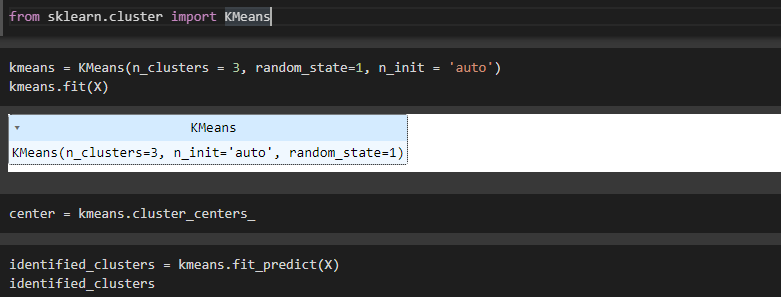
****

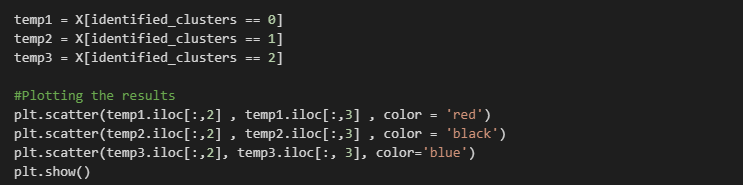
****

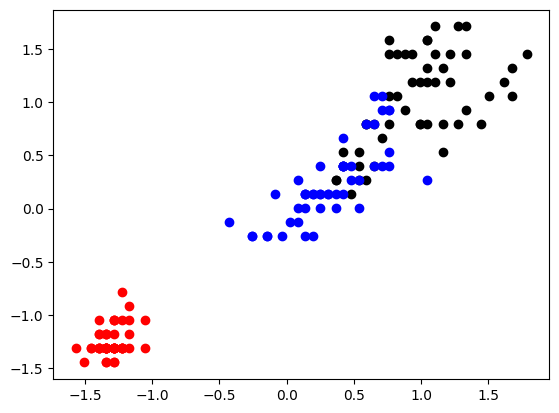




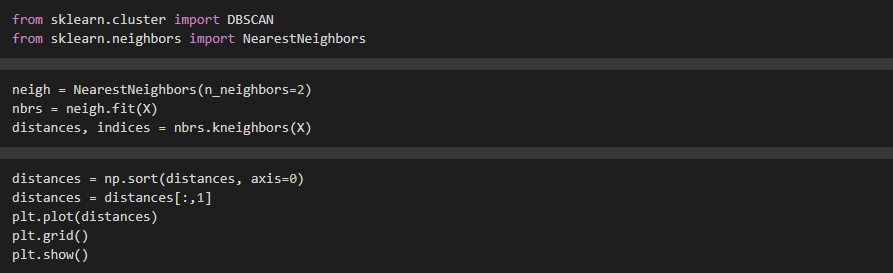
**Kmeans Clustering**

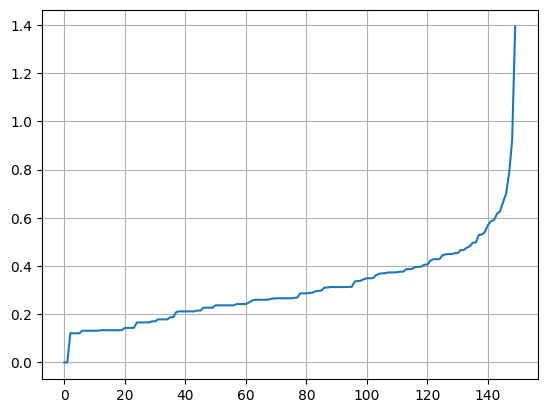
****

****

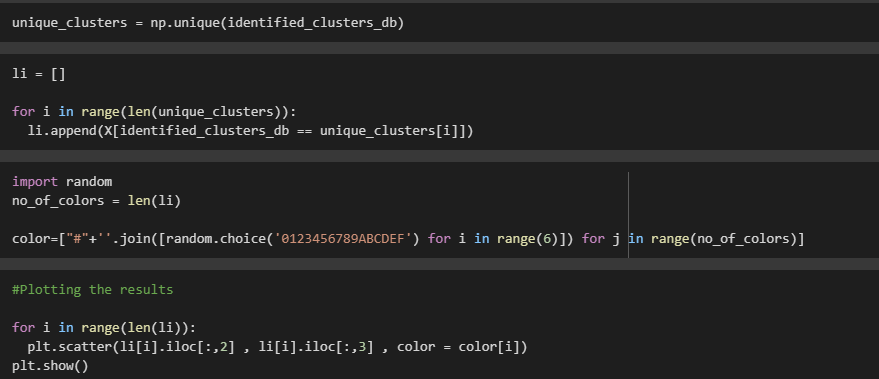
****

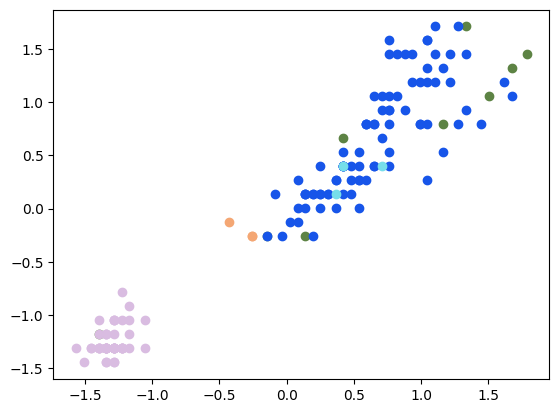
**DBScan**

****

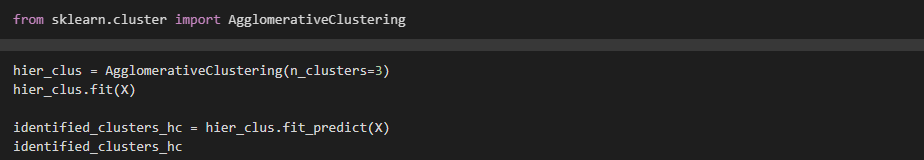
****

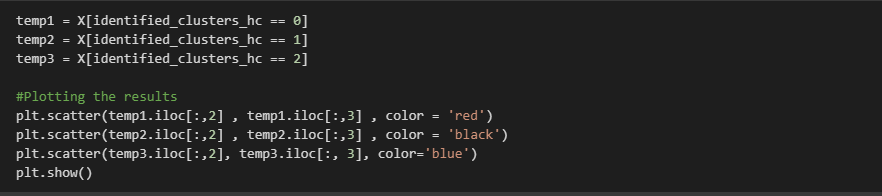
****

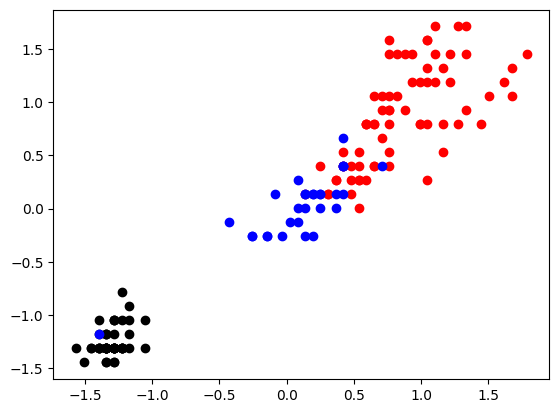
****

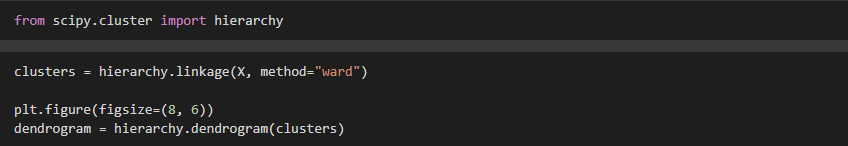
****

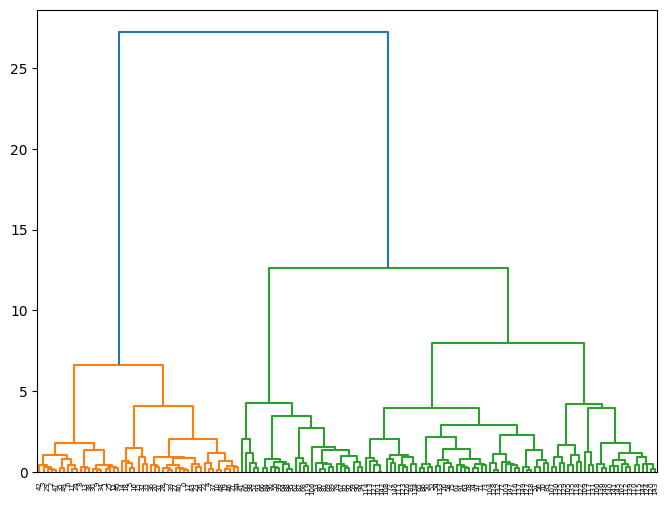
**Hierarchical Clustering**

****

****

****

****

****