CS4038D Data Mining – Assignment 1

1.Downloaded Iris species data set from kaggle:https://www.kaggle.com/uciml/iris

2.

1.There are 6 attributes namely:

COLUMN NAME	DESCRIPTION	TYPE	ATTRIBUTE
Id	It identifies each record with a unique number	int64	Nominal
SepalLengthCm	Length of the sepal (in cm)	float64	Numeric
SepalWidthCm	Width of the sepal (in cm)	float64	Numeric
PetalLengthCm	Length of the petal (in cm)	float64	Numeric
PetalWidthCm	Width of the petal (in cm)	float64	Numeric
Species	Species name	object	Nominal

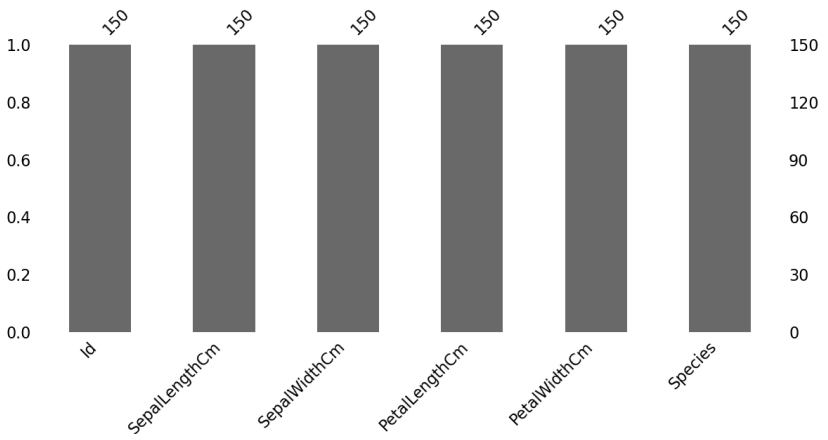
Total Nominal attributes=2

Total Ordinal attributes=0

Total Numeric attributes=4

2.Total number of record = 150

3.To check for missing values we plot a bar graph, to visualize the missing value, missingno package is used.



From the above bar graph we observe that each column show

Same length of bar. Thus this dataset has no missing value.

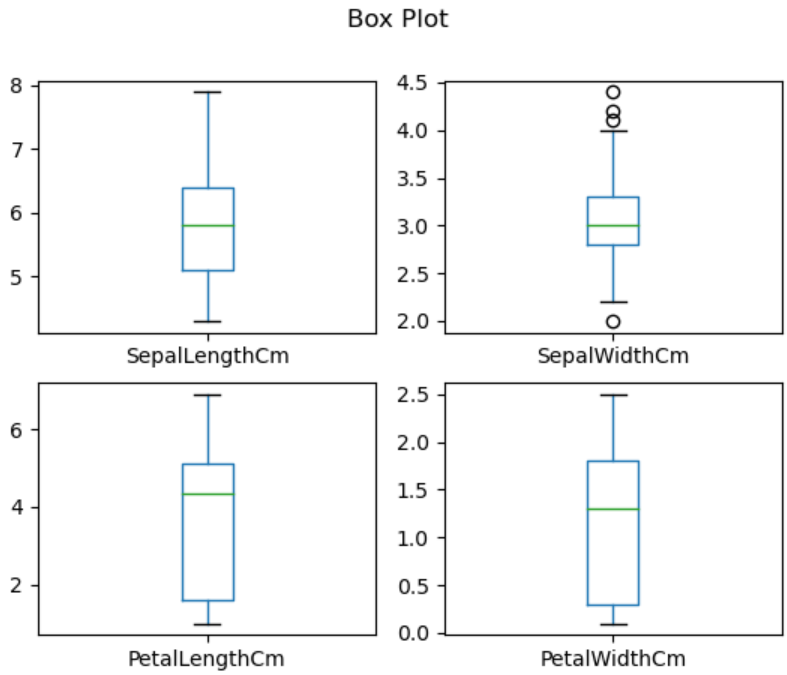
4. Data cleaning is an important step for data analysis which leads to decision making.

Therefore, to improve the quality of data, I did the following.

Now we know that our data has no missing values as each column of the bar has same length.

We have five features - Id,Sepal Length, Sepal Width, Petal Length, and Petal Width and one target field : Species. One column (Id) is of no use for us. So now we will drop that column permanently(using inplace= True).

Secondly, we box plot the features to check for the outliers and obtained the following.



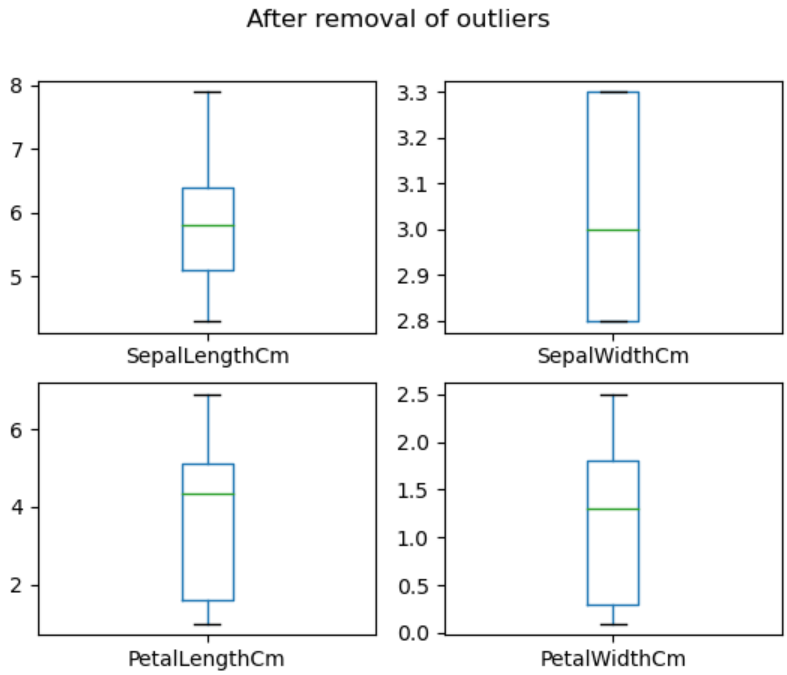
We realized SepalWidthCm has outliers. To remove the outliers, there are two solutions.

A). Remove the corresponding record.

B). Replacing the outliers with suitable value.

Choosing option A may result incur loss of valuable data. Therefore we will go for option B.

In option B, by suitable value we mean replacing the outliers greater than Q3 by Q3 and replacing the outliers smaller than Q1 by Q1. By doing so we obtained the following box plot.

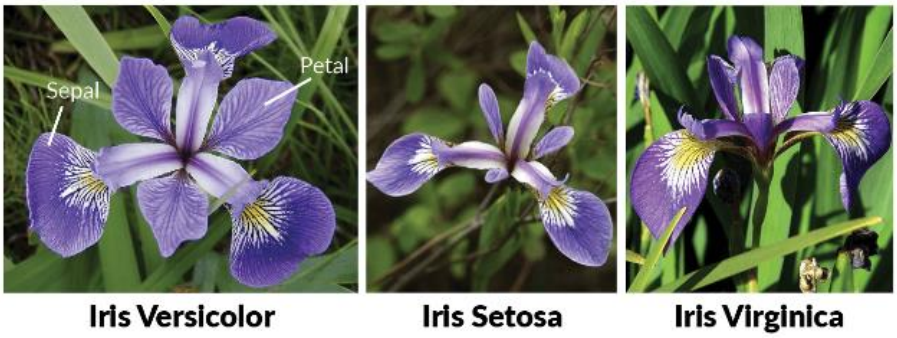


Thus, we can see there are no outliers in any of the feature.

So, our data is cleaned.

5. The Iris data set contain 3 classes, where each class refers to the type of Iris flower i.e. Iris Versicolor, Iris Setosa, Iris Verginica.

Let's explore our data(using classification algorithm) to predict type of Iris flower based on the sepal length, sepal width, petal length and petal width.



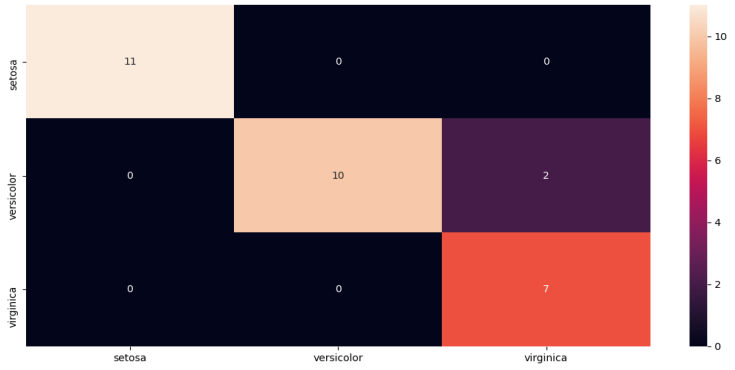
3.

3.1.We obtained the following result by dividing our dataset into train and test set using the 80:20 method.

By applying Naïve Bayes classification

```
-----According to Naive Bayes Classification----

Accuracy rate = 0.9333333333333333
Error rate =  0.06666666666666665
Sensitivity score =  0.9333333333333333
Specifity Score =  0.9797101449275362
Classification Report =
                precision    recall  f1-score   support

    Iris-setosa       1.00      1.00      1.00        11
Iris-versicolor       1.00      0.83      0.91        12
 Iris-virginica       0.78      1.00      0.88         7

       accuracy                           0.93        30
      macro avg       0.93      0.94      0.93        30
   weighted avg       0.95      0.93      0.93        30

Confusion Matrix =  [[11  0  0]
 [ 0 10  2]
 [ 0  0  7]]
```

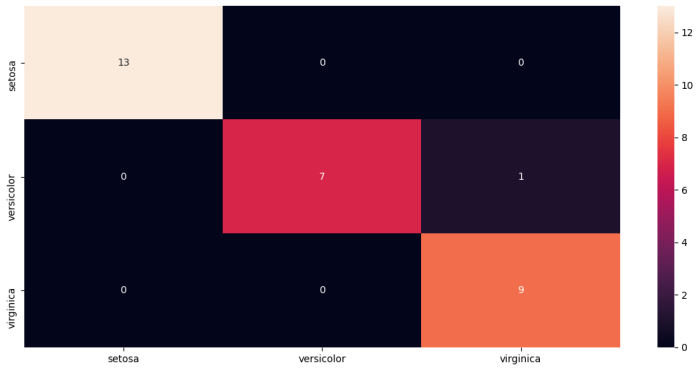
And Confusion matrix as



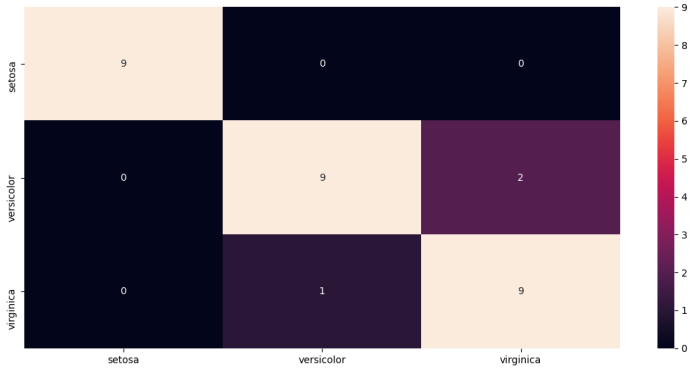
And while applying Decision Tree classification algorithm, using two attribute selection methods i.e. information gain(entropy) and gini index.

Obtained the following confusion matrix result using

1. Information Gain:

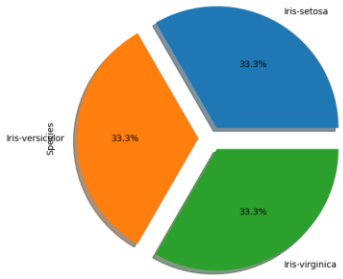


2. Gini Index :



3.2. Iris dataset is balanced dataset, It includes three iris species(setosa, versicolor, viriginica) with 50 samples each as well as some properties about each flower.

To check the same we plot the following pie chart.



This pie graph shows that three classes(setosa, versicolor, virginica) are equally divided 33.3% each. Hence Iris dataset is balanced.

3.3. If we compare on the basis of specificity score than Naïve Bayes classification is better than Decision Tree but, on the basis of accuracy, error rate sensitivity score and classification report then Decision tree (Gini Index) works better than Naïve Bayes algorithm.

We conclude that overall Decision Tree classification algorithm serves better than Naïve Bayes classification algorithm in Iris dataset.

4. For Decision Tree algorithm we have two attribute selection method, as follow:

A). Information Gain(Entropy)

B). Gini Index

Using entropy as an attribute selection method we get the following result.

```
----------Using Entropy-------------------------

Accuracy rate = 0.9666666666666667
Error rate =  0.033333333333333326
Sensitivity score =  0.9666666666666667
Specifity Score =  0.9857142857142857
Classification Report =
                precision    recall  f1-score   support

    Iris-setosa       1.00      1.00      1.00        13
Iris-versicolor       1.00      0.88      0.93         8
 Iris-virginica       0.90      1.00      0.95         9

       accuracy                           0.97        30
      macro avg       0.97      0.96      0.96        30
   weighted avg       0.97      0.97      0.97        30

Confusion Matrix =  [[13  0  0]
 [ 0  7  1]
 [ 0  0  9]]
```

Using Gini Index as an attribute selection method we get the following result.

```
---------Using Gini Index Method----------------

Accuracy rate = 0.9
Error rate =  0.09999999999999998
Sensitivity score =  0.9
Specifity Score =  0.9473684210526315
Classification Report =
                precision    recall  f1-score   support

    Iris-setosa       1.00      1.00      1.00         9
Iris-versicolor       0.90      0.82      0.86        11
 Iris-virginica       0.82      0.90      0.86        10

       accuracy                           0.90        30
      macro avg       0.91      0.91      0.90        30
   weighted avg       0.90      0.90      0.90        30

Confusion Matrix =  [[9 0 0]
 [0 9 2]
 [0 1 9]]
```

From the above result we realize that on the basis of accuracy rate, sensitivity score, specificity score and classification report; in all Gini index performs better for Iris dataset.

5. Using Gini index as an attribute selector we obtained the following decision tree.