Big Data Assignment 1

a. Data:

The data comprises of 5 features: sepal length, sepal width, petal length, petal width and the class. The continuous features i.e., sepal and the petal length is measured in cms. The feature ‘Class’ is categorical. There are 150 instances, with no missing values in the dataset. Each class is distributed evenly i.e., 33% for each class. The dataset is taken from the (https://archive.ics.uci.edu/ml/datasets/iris).

	Min	Max	Mean	SD	Class Correlation
Sepal Length	4.3	7.9	5.84	0.83	0.7826
Sepal width	2.0	4.4	3.05	0.43	-0.4194
Petal Length	1.0	6.9	3.76	1.76	0.9490
Petal Width	0.1	2.5	1.20	0.76	0.9565

b. Methods

1. Imputation methods

Imputation is done by using **mean** wherein mean of the observed values is calculated which are non-missing.

The other technique used is **substitution** wherein the value of instance is selected which is not used in the sample.

The third imputation technique used for **categorical** is just adding a new term called not available where the data is missing.

2. Distance Methods:

Continuous Values:

The first method used is the **Euclidean measure**. The distance of each instance is calculated by taking the square root of the sum of all square of required records subtracted with eachother. i.e., Sqrt(((a-b)^2) + ((a’ -b’)^2))

The second method for the other continuous feature is the **manhattan distance**. It is measured as the modulus of the sum of difference of each instance. i.e., |(a-a’) + (b’ -b’)|

Categorical Value

Distance for Categorical feature is calculated using the **Euclidean measure** as well. The instances for the categorical feature is converted into integers first and the same formula is used to measure the distance.

c. Feature Scaling Methods:

Z Score method: is the number of standard deviations a given data point lies away from the mean. **Z**-**score** is calculated by subtracting the mean from each data point (of feature "petal_length") and divide the result by the standard deviation

Min Max method: In this technique, the instance is subtracted from its min value from the respective feature divided by the difference of its min and max value.

d. Imputation accuracy measures

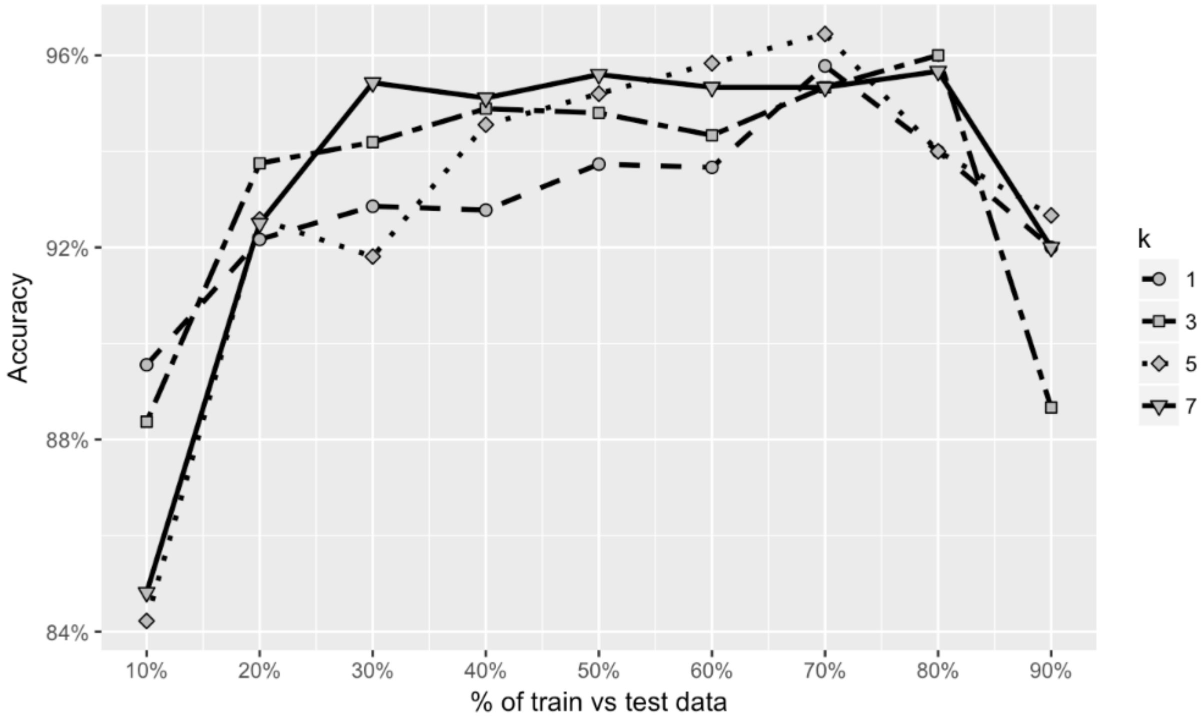
Knn is used to impute the missing values because a point value can be approximated by the values of the points that are close to itbased on other variables. The other method that's used is mean wherein the mean of the summary statistic with the missing value.

Analysis of distance measured with original values vs scaled values: The optimal distance was fetched when the data was scaled after imputing the dataset as the original data had variables in large valued units that dominated the computed dissimilarity and the variables measured in small value units contributed very little.

e. Tools

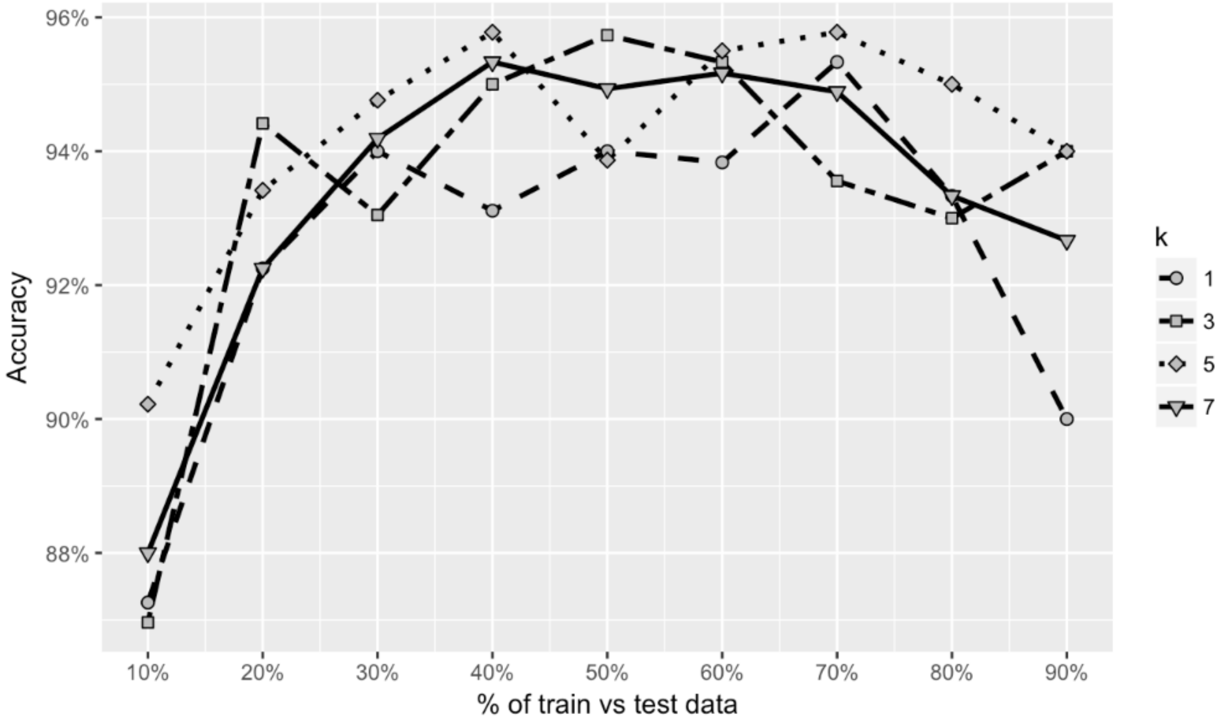
The language used to run this experiment is python, version 3. The code is run and executed in Jupyter Notebook.

Normal K Accuracy



Weighted kNN

Dataset: Iris



```
With 1 Nearest Neighbour


Predicted Class of the datapoint =  Iris-virginica

Nearest Neighbour of the datapoints =  [141]



With 5 Nearest Neighbours



Predicted class of the datapoint =  Iris-virginica

Nearest Neighbours of the datapoints =  [141, 139, 120, 145, 144]
```