Big Data Assignment 1

Data:

Source:

Abstract: Data taken from the Blood Transfusion Service Center in Hsin-Chu City in Taiwan -- this is a classification problem.

https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center (UCI Repository)

Features:

Continuous Feature 1: Monetary quantitative c.c. blood Input

Continuous Feature 2: Time quantitative Months Input

Categorical Feature: Whether he/she donated blood in March 2007 binary 1=yes 0=no

Instances: 600

Methods:

Imputation Methods:

KNN:

In *k-NN classification*, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small).

In *k-NN regression*, the output is the property value for the object. This value is the average of the values of *k* nearest neighbors.

1NN:

This algorithm is same as K-NN algorithm but If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.

Weighted KNN:

In this algorithm the neighbors are found for the imputing point and the inverse of each distance is taken. The sum of each distance is considered and then divided by each inverse value, the result is the distance from the imputing point. The one with lowest distance is then imputed as the original value.

Big Data Assignment 1

Feature Scaling Methods:

Standard Scaler:

Standardize features by removing the mean and scaling to unit variance.
The standard score of a sample x is calculated as:
z = (x - u) / s
Where u is the mean of the training samples or zero if with mean=False, and s is the standard deviation of the training samples or one if with_std=False.

MinMax Scaler:

Transforms features by scaling each feature to a given range.
This estimator scales and translates each feature individually such that it is in the given range on the training set, e.g. between zero and one.
X_std= (X -X.min (axis=0)) / (X.max (axis=0) -X.min (axis=0))
X_scaled=X_std* (max-min) +min

Imputation Accuracy Methods:

The Accuracy is measured after comparing the original value and the predicted value. If the difference between the two is zero then the values match and accuracy is 100%. If the values don't match then the count of each matched value is divided by total number of values and the accuracy percentage is calculated.

In some of the cases of continuous features the accuracy was found to be zero as the mean imputed value did not exactly match with the original integer.

Distance Methods:

Euclidean Distance:

The Euclidean distance between two items is the square root of the sum of the squared differences of coordinates.

Manhattan Distance:

The Manhattan distance between two items is the sum of absolute differences of coordinates.

Tools:

Python (version 3.7) language along with Spyder IDE(Anaconda 4.7.11)
Import scikit using command pip install sklearn
Import numpy using pip install numpy

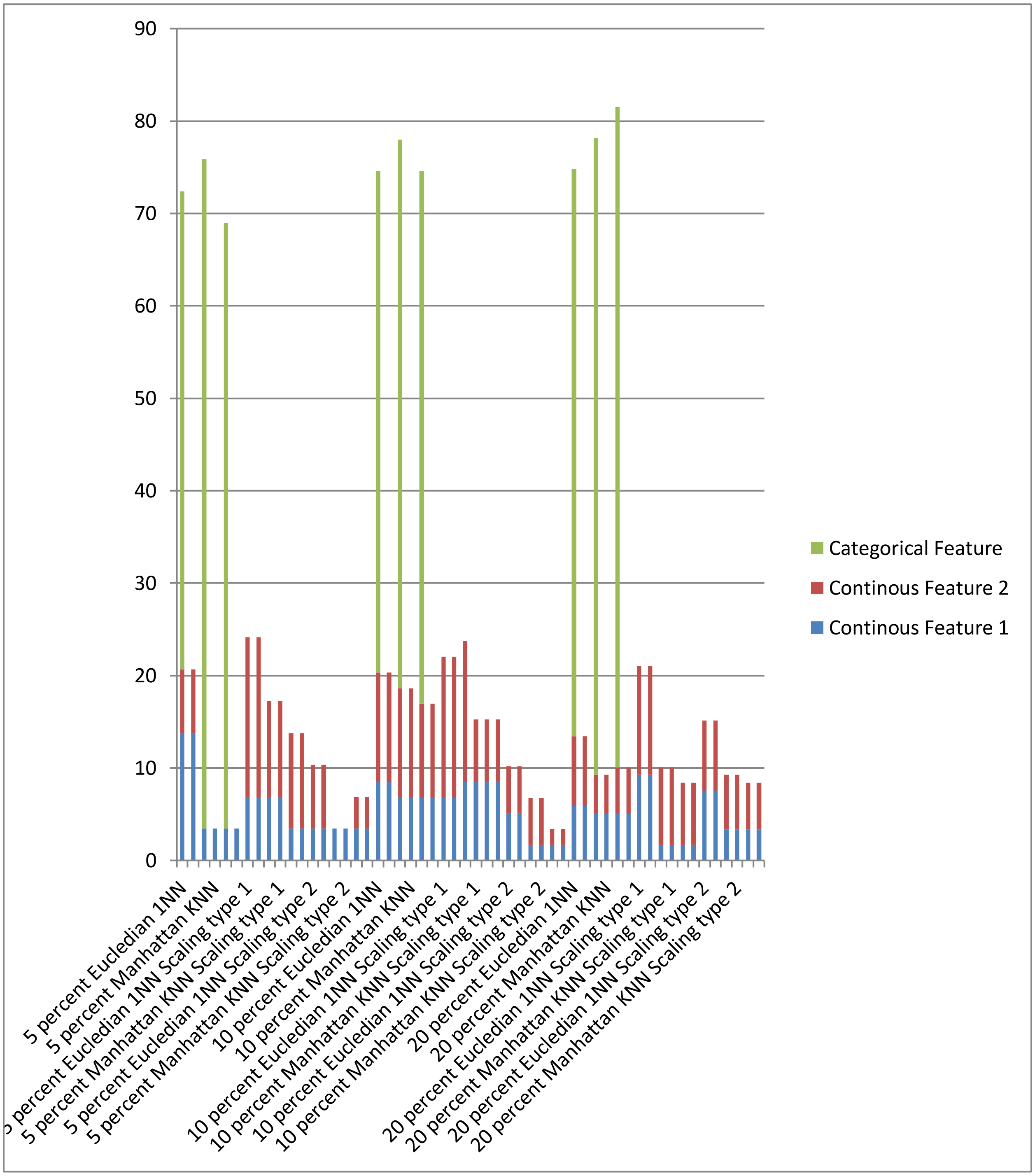
Big Data Assignment 1

Results:

	Continous Feature 1	Continous Feature 2	Categorical Feature
5 percent Eucledian 1NN	13.79310345	6.896551724	51.72413793
5 percent Mahattan 1NN	13.79310345	6.896551724	NA
5 percent Eucledian KNN	3.448275862	0	72.4137931
5 percent Manhattan KNN	3.448275862	0	NA
5 percent Eucledian Weighted KNN	3.448275862	0	65.51724138
5 percent Manhattan Weighted KNN	3.448275862	0	NA
5 percent Eucledian 1NN Scaling type 1	6.896551724	17.24137931	NA
5 percent Manhattan 1NN Scaling type 1	6.896551724	17.24137931	NA
5 percent Eucledian KNN Scaling type 1	6.896551724	10.34482759	NA
5 percent Manhattan KNN Scaling type 1	6.896551724	10.34482759	NA
5 percent Eucledian Weighted KNN Scaling type 1	3.448275862	10.34482759	NA
5 percent Manhattan Weighted KNN Scaling type 1	3.448275862	10.34482759	NA
5 percent Eucledian 1NN Scaling type 2	3.448275862	6.896551724	NA
5 percent Manhattan 1NN Scaling type 2	3.448275862	6.896551724	NA
5 percent Eucledian KNN Scaling type 2	3.448275862	0	NA
5 percent Manhattan KNN Scaling type 2	3.448275862	0	NA
5 percent Eucledian Weighted KNN Scaling type 2	3.448275862	3.448275862	NA
5 percent Manhataan Weighted KNN Scaling type 2	3.448275862	3.448275862	NA
10 percent Eucledian 1NN	8.474576271	11.86440678	54.23728814
10 percent Mahattan 1NN	8.474576271	11.86440678	NA
10 percent Eucledian KNN	6.779661017	11.86440678	59.3220339
10 percent Manhattan KNN	6.779661017	11.86440678	NA
10 percent Eucledian Weighted KNN	6.779661017	10.16949153	57.62711864
10 percent Manhattan Weighted KNN	6.779661017	10.16949153	NA
10 percent Eucledian 1NN Scaling type 1	6.779661017	15.25423729	NA
10 percent Manhattan 1NN Scaling type 1	6.779661017	15.25423729	NA
10 percent Eucledian KNN Scaling type 1	8.474576271	15.25423729	NA
10 percent Manhattan KNN Scaling type 1	8.474576271	6.779661017	NA
10 percent Eucledian Weighted KNN Scaling type 1	8.474576271	6.779661017	NA
10 percent Manhattan Weighted KNN Scaling type 1	8.474576271	6.779661017	NA
10 percent Eucledian 1NN Scaling type 2	5.084745763	5.084745763	NA
10 percent Manhattan 1NN Scaling type 2	5.084745763	5.084745763	NA
10 percent Eucledian KNN Scaling type 2	1.694915254	5.084745763	NA
10 percent Manhattan KNN Scaling type 2	1.694915254	5.084745763	NA
10 percent Eucledian Weighted KNN Scaling type 2	1.694915254	1.694915254	NA
10 percent Manhataan Weighted KNN Scaling type 2	1.694915254	1.694915254	NA
20 percent Eucledian 1NN	5.882352941	7.56302521	61.34453782
20 percent Mahattan 1NN	5.882352941	7.56302521	NA
20 percent Eucledian KNN	5.042016807	4.201680672	68.90756303
20 percent Manhattan KNN	5.042016807	4.201680672	NA
20 percent Eucledian Weighted KNN	5.042016807	5.042016807	71.42857143
20 percent Manhattan Weighted KNN	5.042016807	5.042016807	NA
20 percent Eucledian 1NN Scaling type 1	9.243697479	11.76470588	NA
20 percent Manhattan 1NN Scaling type 1	9.243697479	11.76470588	NA
20 percent Eucledian KNN Scaling type 1	1.680672269	8.403361345	NA
20 percent Manhattan KNN Scaling type 1	1.680672269	8.403361345	NA
20 percent Eucledian Weighted KNN Scaling type 1	1.680672269	6.722689076	NA
20 percent Manhattan Weighted KNN Scaling type 1	1.680672269	6.722689076	NA
20 percent Eucledian 1NN Scaling type 2	7.56302521	7.56302521	NA
20 percent Manhattan 1NN Scaling type 2	7.56302521	7.56302521	NA
20 percent Eucledian KNN Scaling type 2	3.361344538	5.882352941	NA
20 percent Manhattan KNN Scaling type 2	3.361344538	5.882352941	NA
20 percent Eucledian Weighted KNN Scaling type 2	3.361344538	5.042016807	NA
20 percent Manhataan Weighted KNN Scaling type 2	3.361344538	5.042016807	NA

Big Data Assignment 1

Comparative analysis of imputation:


5 percent Eucledian 1NN
5 percent Manhattan KNN
5 percent Eucledian 1NN Scaling type 1
5 percent Manhattan KNN Scaling type 1
5 percent Eucledian 1NN Scaling type 2
5 percent Manhattan KNN Scaling type 2
10 percent Eucledian 1NN
10 percent Manhattan KNN
10 percent Eucledian 1NN Scaling type 1
10 percent Manhattan KNN Scaling type 1
10 percent Eucledian 1NN Scaling type 2
10 percent Manhattan KNN Scaling type 2
20 percent Eucledian 1NN
20 percent Manhattan KNN
20 percent Eucledian 1NN Scaling type 1
20 percent Manhattan KNN Scaling type 1
20 percent Eucledian 1NN Scaling type 2
20 percent Manhattan KNN Scaling type 2
🟩 Categorical Feature
🟥 Continous Feature 2
🟦 Continous Feature 1