DATASET

The data depicts Happiness Index has overall 107 Instances and 7 Features which include GDP per capita, Social support, Healthy Life Expectancy, Happiness Score, Freedom to make life choices, Generosity and Perceptions of Corruption.

INSTANCES

The file contains the information about 107 different Countries. Each instance represents the information about different Country.

SOURCE

The source of the dataset is Kaggle and google.

World Happiness Report

IMPUTATION METHODS

1. kNN Method

kNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification. Steps to implement 1NN Method are defined below

Step 1:- First check whether there is any missing value inside dataset. If there is any missing value, then this Imputation method is used to impute the value.

Step 2:- For every missing value, we will conduct the following steps

2.1 Calculating the distance between the missing value instance to other instances of the dataset using different methods (Euclidean, Manhattan, Hamming and so on).

2.2 Now based upon all the distance values, get the minimum one.

2.3 Now get the index value of the nearest instance to impute the missing value

2.4 Impute the value of the attribute into missing value.

2. Weighted kNN Method

Weighted kNN is a modified version of k nearest neighbors. One of the many issues that affect the performance of the kNN algorithm is the choice of the hyperparameter k. If k is too small, the algorithm would be more sensitive to outliers. If k is too large, then the neighborhood may include too many points from other classes.

The intuition behind weighted kNN, is to give more weight to the points which are nearby and less weight to the points which are farther away. Steps to implement 6NN method are defined below.

Step 1:- First check whether there is any missing value inside dataset. If there is any missing value, then this Imputation method is used to impute the value.

Step 2:- For every missing value, I conducted the following steps

2.1 Calculating the distance between the missing value instance to other instances of the dataset using different methods (Euclidean, Manhattan, Hamming and so on).

2.2 Now based upon all the distance values, Sort them in ascending order.

2.3 Since we are using Weighted 6NN, we will take first six minimum distances from the sorted vector.

2.4 Take values from the instances associated with these minimal distances and assign some weights to each value in a way that the value of weight decreases as the distance increases.

2.5 After assigning weight we have to implement a function to calculate an appropriate value using the values and their respective weights.

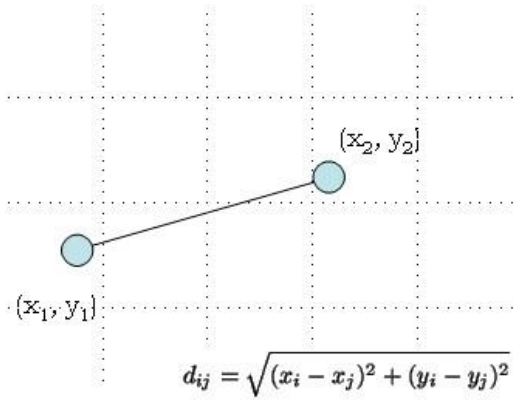
2.6 The final value will get imputed to the missing place.

DISTANCE METHODS

1. Euclidean Distance

Euclidean Distance is the straight line distance between two points in Euclidean Distance. This distance calculation is used while imputing values into continuous attributes. All methods whether kNN or Weighted kNN, both use this method to calculate distance. In, my assignment, I have also used this method to calculate distance between categorical attributes (First converting them into continuous ones.)

The formula used to calculate distance between two data points is shown below: -



Euclidean Distance

FEATURE SCALING METHODS

1. Min-Max Method

An alternative approach to Z-score normalization (or standardization) is the so-called **Min-Max scaling** (often also simply called "normalization"). In this approach, the data is scaled to a fixed range 0-1.
Where X is value, min(x) is the minimum value, max(x) is the maximum value.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

IMPUTATION ACCURACY MEASURES

Accuracy refers to closeness of the measurements to a specific value. Here in this case accuracy means how far the imputed value is different from the original value.

Mean Square Error:

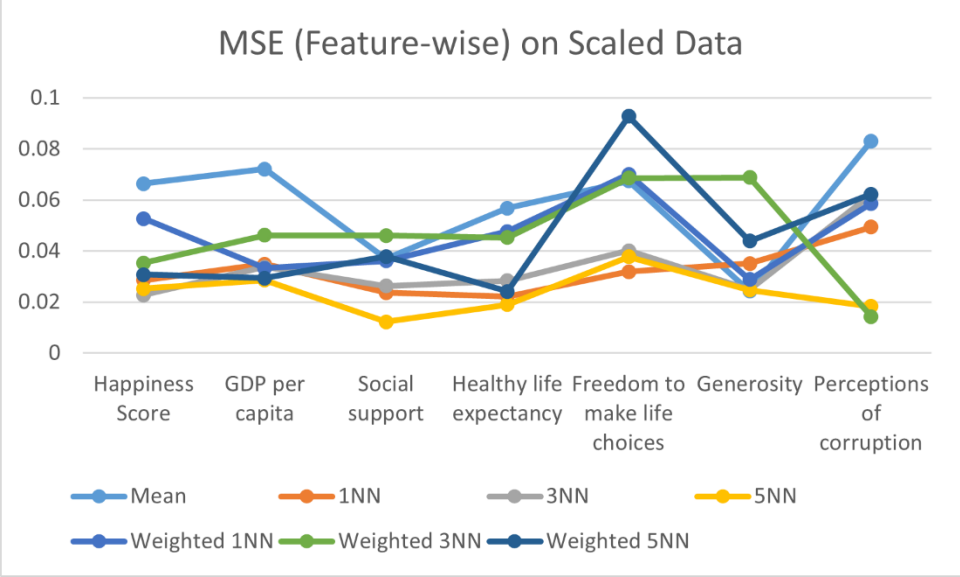
The mean squared error or mean squared deviation of an estimator measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

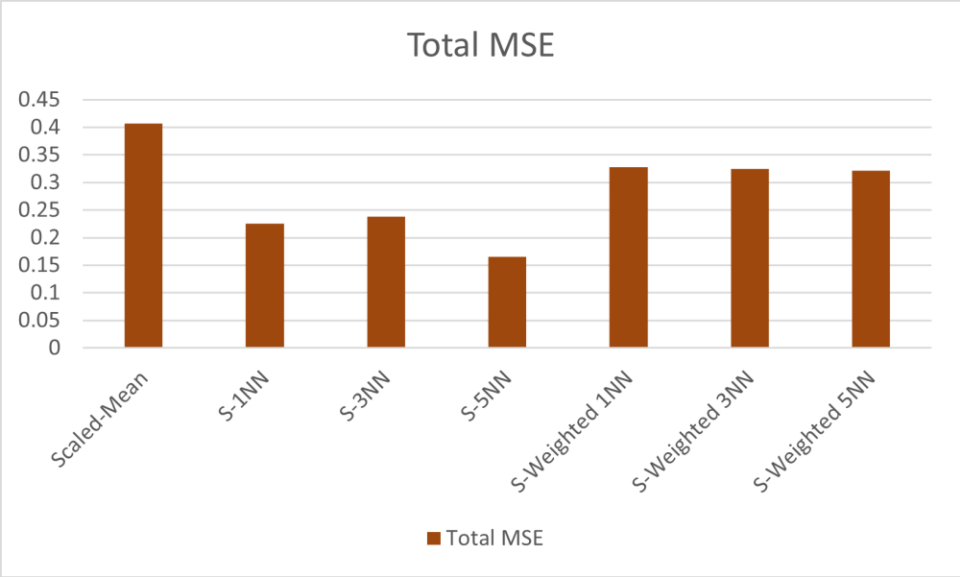
$\mathbf{MSE}$ = mean squared error
n = number of data points
Y_i = observed values
$\hat{Y}_i$ = predicted values

RESULTS

Scaled Data:

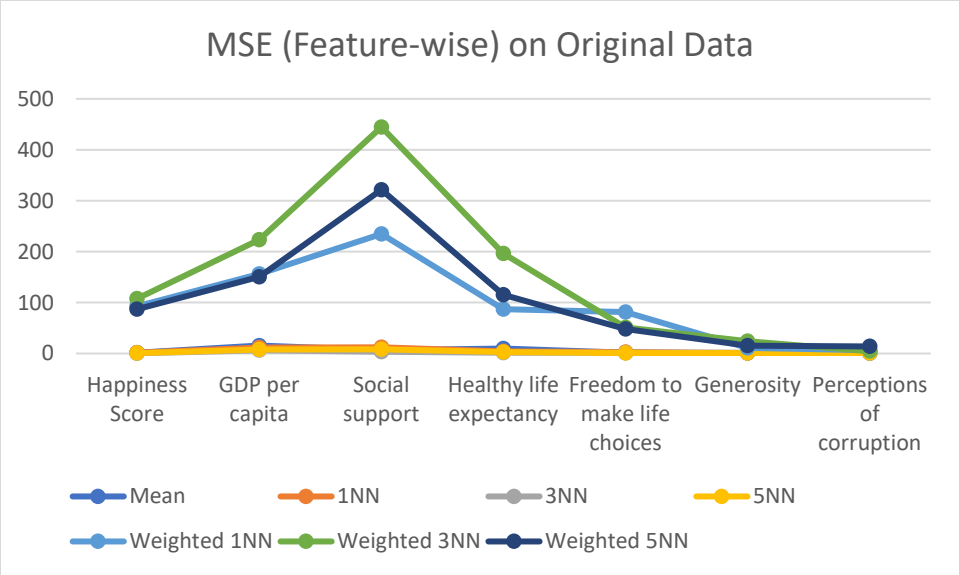


Overall, all 7 imputation methods performed equally well on scaled data when compared to Original data.

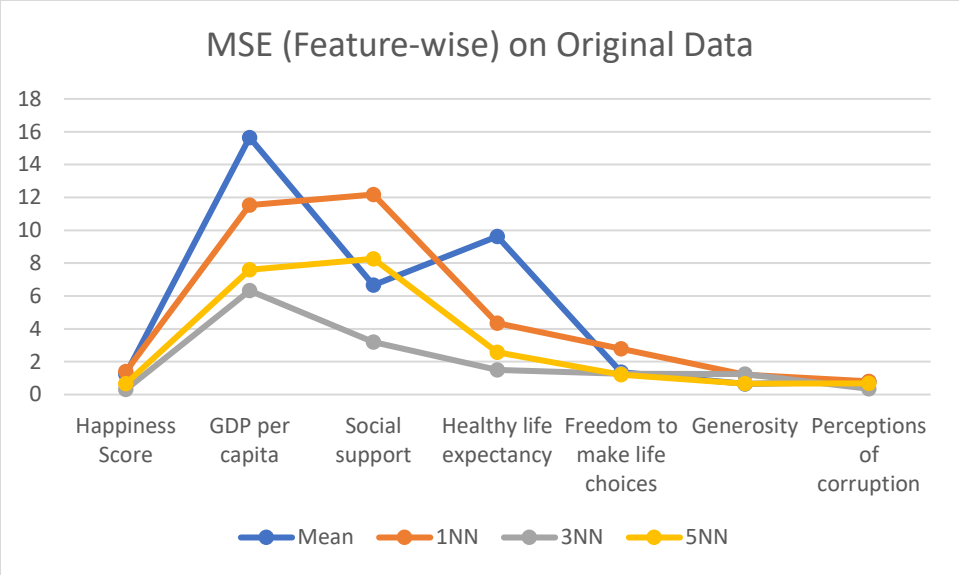


But in case we calculate Total Error across all features, kNN clearly performed slight better in terms of accuracy for all values of k (1,3 & 5) compared to other methods.

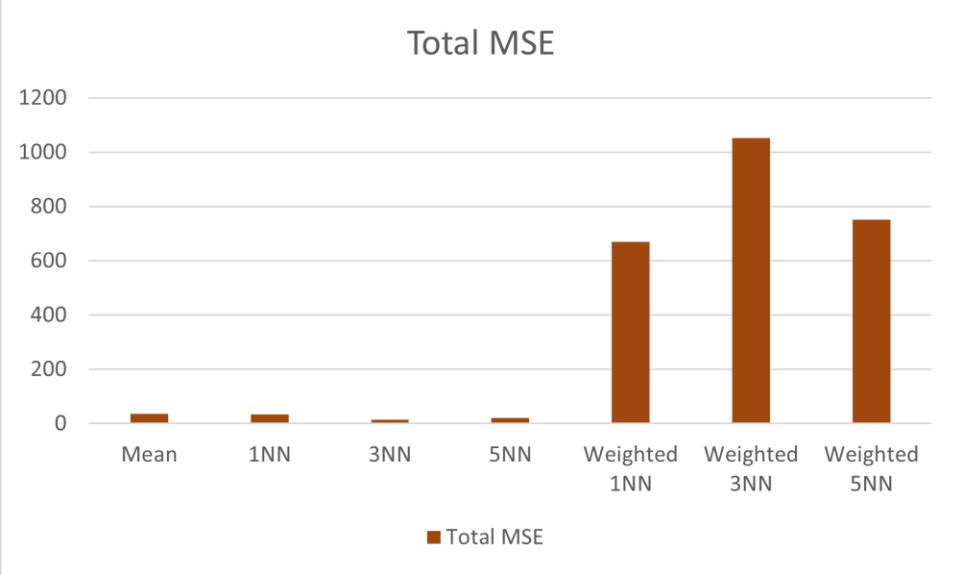
Original Data:



Amongst all the features, the graph shows error was more in weighted kNN at the 'Social Support' feature. One reason for this could be varied and scattered data values in the feature.



Error in Original Data set was much higher when compared with Scaled error.



Even for Original data, kNN performed better compared to other methods.

CONCLUSION:

kNN performed better compared to other imputation methods (weighted kNN and Mean) for both Scaled and Original data.