Elements Of Data Processing

Assignment 2 Report - [redacted]

Question 1.

- K - NN algorithm has performed better for task 2A with accuracies of 82.0 and 86.9 %.
- Amongst K-NN (K = 10) had a greater accuracy to predict the class feature life expectancy at birth(years) of the data .
- For Task2A, firstly the missing values “..” Have been replaced by nan and then imputed with the median of the columns. After imputing the 264 rows of the world.csv data (reduced to 264 because) mean has been removed and data has been brought to unit variance ,then it has been merged with the life.csv data and then the resulting data has been split into training and testing set and then scaled used to calculate accuracy by different methods.
- Since all of the data is numeric and we know that KNN determines neighbourhoods, so there must be distance metrics. This acts as an advantage for KNN here. Also, KNN performs instance-based learning, so it can model complex decision spaces.

Question 2.

- Pre - processing: After changing all the missing values with nan, the data has been imputed with the median. After Imputation 190 features have been created by the Interaction pair method using Polynomial Features. These features have been concatenated and used to calculate the last feature by the help of K-means clustering algorithm.

- Number of clusters needed are selected according to the highest accuracy. After that we have merged the data with data from life.csv and got 183 rows. Then we have split the data.
- For Feature engineering, we have chosen four features by the RandomForest Classifier algorithm and then scaled the four features and then used that to get the accuracy.
- Then we have picked the four features using PCA (from the first 20 features after imputation and scaling and merging) and the first four features and got the accuracy.

Question 3.

- To get the number of clusters, we have used a method which is a bit similar to the elbow method but instead of using distortions we just formulated the graph of the accuracies against the values of number of clusters. We have chosen the value 6 because we got the same accuracy with any number of clusters within the range (1-11) and that can be seen in the graph plot for task2graph1.png.

Question 4.

- To get the four features, we have used RandomForestClassifier because random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

Question 5.

- Feature Engineering provided the best accuracy and this shows that RandomForestClassifier is a better algorithm than PCA to calculate the accuracy as it adds more randomness to the model. This was the best case because random forest fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy whereas PCA is a dimensionality reduction algorithm. Here, a lower-dimensional projection of the data does not preserves the maximal data variance which is the reason why PCA does not have the maximum accuracy.

Question 6.

- To improve classification accuracy we can implement PCA first and get ten features and then use Random Forest to get the best four features from them. This will lead to a lower-dimensional projection of the data and give predictive accuracy to the model.

Question 7.

- This classification model is not fully reliable because of the randomness of the model and we should have more number of features to get a better view of the dataset. Secondly, Classification accuracy alone is typically not enough information to make rely on a classification model.