COMP20008 Classification Report

**Reporting of Task 2a**

Which algorithm (decision trees or k-nn) in Task-2A performed better on this dataset?

Decision trees performed better overall compared to k-nn when the averages of both are compared. The average of k-nn is 0.7 whereas the average of decision tree is 0.709

For k-nn, which value of k performed better? Explain your experiments and the results.

k-nn (k=7) performed much better compared to k-nn (k=3) k-nn(k=7)= 0.727 k-nn(k=3) = 0.673

Experiment involved pre-processing data by firstly splitting into test and training sets, then imputing the data using simple imputer using median, and fitting the data based on the training set to make sure there is no data leakage.

After we have the imputed x\_train and x\_test, we then normalize the data. After the data has been normalized we can then use neighbors.KNeighborsClassifier(n\_neighbors=n\_neighbors)

(k=7) has better accuracy as larger k value is better for classification as it leads to smoothening the decision boundaries. When (k=3) k is small and leads to unstable decision boundaries.

**Reporting of Task-2B**

A description of the precise steps you took to perform the analysis in Task-2B.

Firstly I split the data set into training and testing sets, then I imputed the data using simple imputer using median. I choose simple imputer median strategy as I wanted to compare my results to Task 2a and it was only right to compare using the same imputing strategy. Using the median is also better than the mean as the mean is not heavily affected by outliers whereas the mean is. I also used a random state of 200 as I felt that I wanted to keep the states the same for 2a and 2b in compare results. After imputing I used combinations to form my interaction pairs. I did this to both the training and testing data set naming them feature\_train and feature\_test.

The method you used to select the number of clusters for the clustering label feature generation and a justification for this method.

I then determined to use 3 clusters by using k means clustering. I normalized my original training data set. Then fit my training data to k means from values 1 to 14 and plotted a graph with y axis being sum of squared distance, and x being k values. I used the elbow method based on graph to find 3 points with most bending

I then used calinski and silhouette method on values of k from 2 to 4 and put the scores of both in a csv table. I then used the average of both methods to find the best k. I used best k clusters, and applied cluster label to training set. I combined the interaction pairs and cluster label into one data frame.

The method you used to select four features from the generated dataset of 211 features for your analysis and a justification for this method.

To find the top 4 best features, I used the Univariate Selection. I used Univariate selection ANOVA-F value method as it is appropriate for both numerical and categorical data. I used the top 4 features to fit my data to KNeighborsClassifier and used the top 4 features in feature\_test when predicting.

Which of the three methods investigated in Task-2B produced the best results for classification using 3-NN and why this was the case.

Feature engineering produced the best result as when I implemented feature engineering. Producing 211 features and then taking the top 4. The top 4 features have heavy impact over the accuracy whereas in PCA, reducing all the features down to 4 causes noise as the importance of all the features differ. The first four features are just randomly selected and we do not know how much influence they have on the data. Therefore Feature engineering produced the best result.

What other techniques you could implement to improve classification accuracy with this data.

Cross Validation: Separate your train dataset in groups, always separate a group for prediction and change the groups in each execution. Then you will know what data is better to train a more accurate model.

Algorithm tuning: The parameters in machine learning algorithms majorly influence the outcome of learning process. The objective of parameter tuning is to find the optimum value for each parameter to improve the accuracy of the model.

How reliable you consider the classification model to be.

My classification model is somewhat reliable. I used median imputing on the training set, and scaled and transformed based on training set onto the test set but there could be some overfitting issues, If I were to use more data however this has not been tested for.