- Which algorithm (decision trees or k-nn) in Task-2A performed better on this dataset? For k-nn, which value of k performed better? Explain your experiments and the results.

```
Accuracy of decision tree: 77.049%
Accuracy of k-nn (k=5):81.967%
Accuracy of k-nn (k=10):83.607%
```

  By looking at the results of accuracy of each algorithm, k-nn has the best accuracy. For k-nn, k = 10 performed better.

  I personally think that the reason of why k = 10 perform better than k = 5 is that when k = 10, even there is one neighbor has a big difference by comparing with the centroid, other neighbors can pull up the accuracy but when k = 5, since the number of neighbors decreases, the neighbor which has a big difference with the centroid is likely to have a bigger inflection on the centroid and the accuracy. The reason why decision tree does not have the best accuracy is probably that we cannot find the optimal classification features.

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

  First, I extended the table for feature engineering and f_clusterlabel. Then, since the question says that select four features to calculate the accuracy, therefore, PCA(n_components) should equal to 4 and change it to a dataframe and named by X_reduced. Therefore, when calculating the accuracy of PCA, the X in the train_test_split bracket should be changed to X_reduced. After this step, I just simply copy the code from Task2 – A , the code that calculate the accuracy of k-nn = 5(because the question asks to use k = 5) to find the accuracy of the PCA.

  For the third part, which is calculating the accuracy of first four features, I replaced X_reduced by X.iloc[:, 0:4] in the train_test_split bracket so that the dataset is changed from X_reduced to the first four columns from X, and copy the code from the last question so that I can calculate the accuracy of first four features.

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

  I used Kmean(n_cluster) to choose the number of the clusters for the clustering label feature generation. The number I selected for the clustering label feature generation is 3 because () only includes one of the three descriptions which is High or Medium or Low. Therefore, I chose 3 to be the f_clusterlabel.

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

  The method I used to select four features from the generated dataset of 211 features for my analysis is that I compute the value of mutual information of each feature and list out the four features that have the biggest value of mutual information by comparing with other features because MI obtained about one random variable through observing the other random variable. Therefore, the larger the value, the better the MI. therefore, I chose the four columns with the four biggest MI.

- Which of the three methods investigated in Task-2B produced the best results for classification using 5-NN and why this was the case

```
Train Accuracy(MI): 85.246%
Test Accuracy(MI): 78.689%
Accuracy of k-nn (k=5) by using MI:78.689%
========================================
Train Accuracy: 82.787%
Test Accuracy: 80.328%
Accuracy of k-nn (k=5) by using pca:80.328%
========================================
Train Accuracy: 81.967%
Test Accuracy: 75.41%
Accuracy of k-nn (k=5) by selecting the first four columns:75.41%
```

  By looking at the result, the result is difference with my expectation. I expect that MI has the biggest accuracy.
  I think the reason of why the result of accuracy after applying PCA to the dataset performs better than the result of accuracy by choosing the first four features from the dataset and the four columns that are selected from the dataset is that since PCA is used to dimensionality reduction is to retain the most important features of high-dimensional data, remove noise and unimportant features, therefore, I think the result of using PCA is the important part from the dataset and which can present most of the information from the dataset. Moreover, the choose of the first four features cannot represent the whole dataset, therefore, we cannot achieve too much useful information by only choosing the first four columns. More than that, the way that I chose the four columns is based on the value of MI, if the choice is made without considering the value of MI, than the accuracy might be less since the four column with the biggest MI cannot represent the whole dataset. Therefore, the result of accuracy after applying PCA to the performs better than the result of accuracy by selecting the first four features from the dataset.

- What other techniques you could implement to improve classification accuracy with this data.
  I think the use of Random Forest Classifier might improve the classification accuracy with this data.

- How reliable you consider the classification model to be.

```
Train Accuracy(MI): 85.246%
Test Accuracy(MI): 78.689%
Accuracy of k-nn (k=5) by using MI:78.689%
========================================
Train Accuracy: 73.77%
Test Accuracy: 72.131%
Accuracy of k-nn (k=5) by using AMI:72.131%
========================================
Train Accuracy: 85.246%
Test Accuracy: 78.689%
Accuracy of k-nn (k=5) by using NMI:78.689%
```

  I think my classification is reliable since I tried to use Normalized_mutual_info_score and Adjusted_mutual_info_score to find the four columns with the biggest NMI and AMI and then calculate their accuracies. The results I achieved from NMI and AMI are very close to the result from MI, therefore, I think the classification using interaction term pairs and clustering labels is reliable.