Quang Tri Le – 1118563

Report for Task 2

Task 2A:

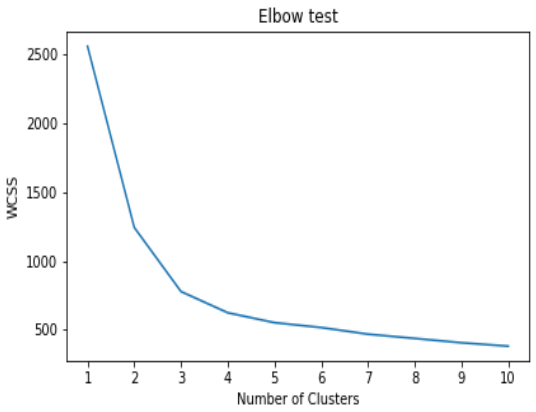
Since *world.csv* and *life.csv* do not have the same number of samples, inner merge must be performed (based on **Country Code**) in order to achieve a consistent record that include features associated with their correct class label. The record was then sorted by Country Code and randomly divided into training set (7/10 ratio) and testing set (3/10 ratio). Imputing median value for each column was conducted in order to fill up missing data and Standard Scale was applied for the consistency in the scale between features. The pre-processed record was then used for testing the accuracy of Decision Tree algorithm and K-Nearest Neighbors algorithm for two different k values (3 and 7).
The experiment yielded **0.673** and **0.727** for **3-nn** and **7-nn** algorithms, respectively. Meanwhile, decision tree produces the accuracy of **0.709**. In KNN algorithm, Euclidean distance was extensively used to find k nearest neighbor. However, in high dimensional data such as this record, such distance metrics are less efficient in identifying the similarity between two data points. Therefore, by considering more nearest neighbors, there is slightly more chance that a data point would fall into the correct class label. This could be observed by the improvement in 7-nn's accuracy compared to that of 3-nn. Decision tree also suffers from high dimensional data since there would be numerous features to be considered for splitting. Therefore, it requires more samples to improve splitting quality in order to boost the accuracy. Thus, in this experiment, KNN with higher value of k (k = 7) will yield the highest accuracy.

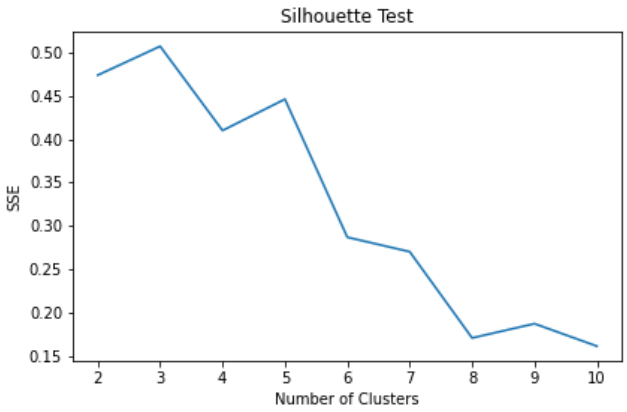
Task 2B:

Like task 2a, inner merge between *world.csv* and *life.csv* must be performed to achieve a consistent record. The record is then sorted on Country Code and divided to train set and test set with proportion of 7/3. Imputing median value for missing fields and Standard Scale are then performed to complete the pre-processing procedure.

Interaction term pairs:
Original data set has 20 features. The number of new features created by interaction term pairs can be calculated as the total combination of 2 out of 20 features which can be mathematically expressed as C(20,2) = 190. In order to achieve such number of new features, *PolynomialFeatures(degree = 2, include_bias = False, interaction_only = True)* was executed. This model is then fitted with train data set and make a new data set with 210 number of features in total.

Clustering Label:
In order to generate clustering label for the data, the number of clusters (k) must be determined by Elbow test and Silhouette test using original data set with 20 features. In Elbow test, Within-Cluster-Sum of Square is calculated for multiple values of K. The change of slope is clearly shown at k = 2 and k = 3. Therefore, these two values of k are the candidates for the optimal number of clusters. The final number of clusters can then be determined through Silhouette test which measure SSE for a range of k. Silhouette Score (SSE) measures the similarity of a point to its own cluster compared to other clusters. Therefore, the number of clusters that gives the global maximum value for SSE is chosen. In this case, k = 3 gives global maximum value for SSE. Hence, the number of clusters for the K-means algorithm is 3.





Feature Selection from 211 features:
To select 4 features from 211 features, *SelectKBest(score_func=mutual_info_classif, k=4)* was executed. These four features are selected because they have the highest Mutual Information Score with the class label among all other features. This feature selection strategy is also known as **Feature Filtering** which selects features that are most well-correlated with class. Therefore, Mutual Information can be utilized in this experiment.

Results:
Feature engineering yields the highest accuracy as expected which is **0.782**. Accuracy for PCA and naive features selection are **0.727** and **0.709,** respectively. In task 2b, the final object is to reduce the number of dimensions by applying features selection. Therefore, the method that can select the features which are best correlated with class will yield the highest accuracy.

In naïve features selection, there is no proof that first four features (columns) are well-correlated with the class. Therefore, this method yields the worst result which is 0.709. However, comparing **0.709** to **0.673** in task 2a, there is brief improvement. This can be explained by the fact that k-nn algorithm (using distance measurement) works more efficiently in lower dimensions.

In theory, applying PCA does not necessarily mean better accuracy since this is just a tool for dimensional reduction that would be algebraically beneficial. However, since PCA helps decrease the number of dimensions by considering a smaller number of principal components (the ones that show most variance across the classes), it indirectly improves the quality for k-nn model which works better in lower dimensions. Therefore, there is a slight improvement from **0.673** to **0.727** for 3-nn algorithm when using PCA.

Feature engineering, as expected, yields the highest accuracy boost from **0.673** to **0.782**. Since features generation was done at this state, there are new-made features that are more well-correlated with class label than the original 20 features.

```
Best 4 features with highest Mutual Information Values
                                      Feature Name  MI score
0  Lifetime risk of maternal death (%) [SH.MMR.RI...  0.604985
1  Cause of death, by communicable diseases and m...  0.589226
2                                    Cluster Label  0.532111
3  Prevalence of anemia among children (% of chil...  0.530016
```

The image above shows four most well-correlated features in the set of 211 features (extracted from running task2b.py). Cluster Label, which is one of many new generated, lies among the best features based on Mutual Information Score. Therefore, by creating new features and using proper technique to select some of the best, it is not wrong to say this is actually "mining gold".

Feature engineering is surely the best technique for this experiment since it does not only choose the best features from original data set but also from highly potential new-generated features.

Reliability and Limitations:
The purpose of K-NN classification model is to identify whether life expectancy of a country is low, medium or high based on metrics about the country's economy (including employability), health care system, access to information and access to daily essentials. In general, although K-NN algorithm is simple to implement, it could be un-reliable in case where the problem has large number of features to assess. Therefore, feature selection methods must be performed, and their reliability should be assessed.

The most unreliable feature selection method is naïve feature selection which chooses first four columns of the data set. This method is extremely risky considering how the features in original dataset are arranged. In the worst-case scenario, if the features in dataset are arranged in a way that least-correlated features with class label are set first to the left, by choosing D-G columns of features, the classification model will perform poorly because there is not much variance among class labels showed in these features. In such case, the accuracy could be much lower than **0.709**. Therefore, applying K-NN classification model using naïve feature selection is extremely unreliable.

In the model where PCA is applied, it would be more reliable since principal components are made based on the original data set. PCA helps reduce overfitting by reducing the number of features, thus the model can reach higher level of generalization. However, using PCA for transforming the original dataset will ignore some important features that are highly corelated with class label. For example, *Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total) [SH.DTH.COMM.ZS]*. Intuitively, we can guess this feature would lie among best features that directly determine life expectancy of a country. As a matter of fact, it lies at the second most important feature based on Mutual Information Score (shown above). Therefore, if applying PCA to K-NN classification model, we must accept the risk of overlooking significant feature from the original datasets.

Feature Engineering is doubtlessly the most reliable feature selection method for K-NN classification model among the three experimented. As discussed in the *Result* section, choosing best relevant features using Mutual Information resolve the issue of high dimensional data and overcome the disadvantage when applying PCA (overlooking significant features).

Recommendation:
From the experiment, K-NN classification method with feature engineering could be further discovered to increase the reliability of the model. Instead of using Mutual Information, Chi-2 can be used so that sample size could be considered which increases the

generalization of the model. In order to do that, *MinMaxScaler* should be used instead of *StandardScaler* so that data field would be in positive range.

Moreover, wrapper approach (decremental) and embedded method for feature selection can be experimented. These two methods would consider the interaction between features which leads to higher generalization and overfit reduction. Therefore, it is promising to use these two methods for feature selection in K-NN classification model to achieve higher accuracy.