**Task 2C**

*Vinay Pinjani | 1151832*

*Reporting of Task-2A*

This task relies on classification algorithms provided by the “sklearn” library. The given datasets were merged together based on the “Country Code” attribute and any mismatches were dropped. The merged dataset was sorted in alphabetical order by “Country Code”.

The dataset was split into training and testing set by the ratio of 7:3 respectively. The “Life Expectancy” attribute was set as the target variable. The missing values of any features were filled with the medians and scaling was applied to standardize the values. The median, mean and the variance used for the preprocessing was stored in a csv file.

Using the preprocessing was trained on two different algorithms, k-NN (k=3 and k=7) and Decision Tree. The accuracy obtained for the particular dataset was:

* Accuracy of decision tree: 0.709
* Accuracy of k-nn (k=3): 0.673
* Accuracy of k-nn (k=7): 0.727

The performance for all algorithms is in a close range. K-NN with k=7 did have a higher accuracy than k = 3, this may be due higher noise when using smaller number of neighbours. The decision tree algorithms also perform fairly well in comparison to the k-NN.

*Reporting of Task-2B*

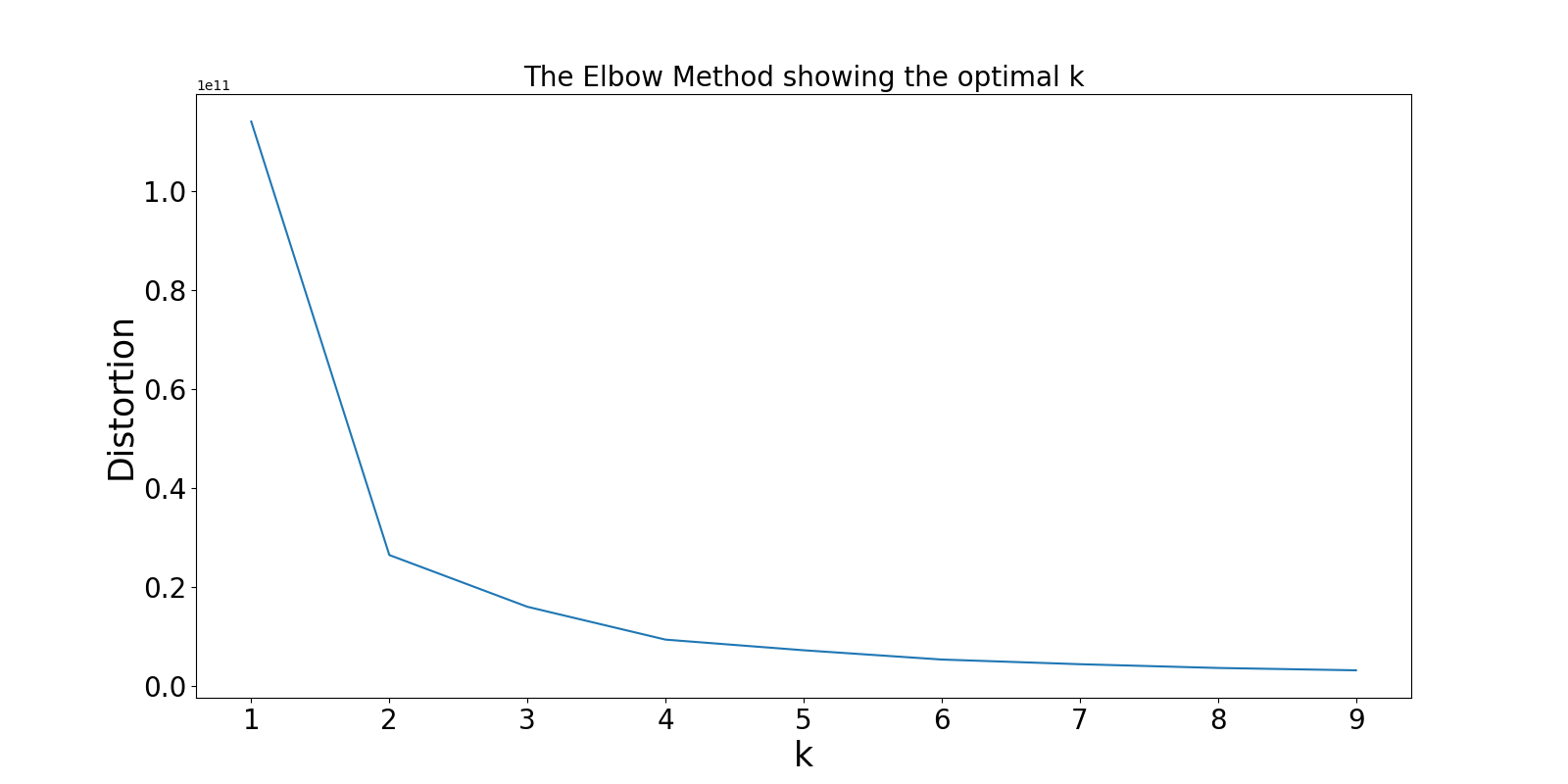
This task involved k-NN (k=3) on three different modifications of the “world.csv” dataset, each of which required significant processing before the model could be trained. Each time the data was split the same way as the earlier task with the exception of the “random\_state” parameter. The median of the column was used to impute the missing values.

Feature Engineering:

Interaction pairs were generated using the “PolynomialFeatures” algorithm fitted on the training dataset. This created an additional 190 features over the already existing.

To generate the cluster labels, the elbow method was used to compare the number of clusters with distortion for the original dataset. A graph was produced that outlined a clear pivot at k=2 hence, 2 clusters were used using the “KMeans” algorithm fitted on the training dataset. The labels gave values for the training set and the prediction gave values for the test data set.

The interaction pairs and cluster labels were concatenated to produce 211 features that were scaled and trained under k-NN (k=3) after scaling.



Principal Component Analysis (PCA):

The PCA features were generated using the “PCA” algorithm fitted on the training dataset with the number of components set to 4. This transformation was applied to both the training and testing dataset from the original data. The model was trained under k-NN (k=3) after scaling.

First Four Features:

The original features from the data were split and only the first four columns were taken before splitting the data. The data was split and trained under k-NN (k=3) after scaling.

*Conclusions for Task-2B*

The following accuracy was obtained using a random state of 4 each of my algorithms:

* Accuracy of feature engineering: 0.545
* Accuracy of PCA: 0.618
* Accuracy of first four features: 0.719

The accuracy of the first four features was the highest, this may just be because the first four features are most important to assume life expectancy or just very lucky. The PCA method did perform better than the custom feature engineering, this may be because the PCA algorithm is more optimised.

A way to improve the accuracy is by taking more sample of 4 features from the data and see which one performs best and use feature engineering on the best four.

The classification model varies highly according to the random state that is chosen, so that can give us different accuracies and the model can be inconsistent.