

AI 221: Machine Exercise 5

Instructions:

- Read and answer each problem using computer code. This MEX should be done *individually*.
- Each item should be answered as either a Jupyter Notebook or a MATLAB Live Script, exported as a single PDF file for the entire MEX. Make sure to HIGHLIGHT your final answers.
- When done, submit the PDF file through UVLE.

Country Data from HELP International

Download the Country Data set from the following Kaggle link:

<https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data/data>

This data set contains 167 countries (per row) and 10 columns: Country name, Child mortality, Exports, Health, Imports, Income, Inflation, Life Expectancy, Fertility, and GDP. Do the following:

- a. **[10 pts]** Normalize the features data using Standard Scaler. Then, perform K-means clustering on all features. Display the elbow plot (Inertia vs. no. of clusters) and the silhouette score plot. What number of clusters is recommended?
- b. **[10 pts]** Perform hierarchical clustering on the normalized data set and compare the results of various linkage methods. Which one would you recommend? Why is this recommended clustering informative?
- c. **[10 pts]** This time, perform PCA to reduce the features data into 2D. Then, perform K-means clustering on the reduced data set just as in item (a). Display the elbow plot and silhouette score plot as well. What number of clusters is recommended? You can label some countries in the 2D mapping.
- d. **[5 pts]** Based on the recommended no. of clusters in item (b), make your own descriptions of each cluster. What range of values of features are unique to each cluster?
- e. **[15 pts]** Based on the 2D PCA mapping, perform anomaly detection using any method. Which countries are deemed to be outliers and why? Which features make them outliers?

Early-Stage Diabetes Risk Prediction

Download the following diabetes data set from UCI Repository:

<https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset>.

The data set contains information on 500+ patients from Bangladesh. Your job is to predict whether a patient is Positive / Negative for diabetes—a binary classification problem. The input features are mostly categorical, with only the Age as the numeric feature. See the relevant paper here:

https://link.springer.com/chapter/10.1007/978-981-13-8798-2_12

In the paper, only the Naïve Bayes, Logistic Regression, and Random Forest models were employed. To improve the results, do the following:

- a. **[5 pts]** Make the necessary encoding for categorical inputs. Split the data into 80% Training and 20% Testing with stratification.

- b. **[25 pts]** Using Optuna, find the best model between the MLP Classifier, Random Forest Classifier, XGBoost Classifier, Logistic Regression, Naïve Bayes Classifier, SVM Classifier (SVC), and kNN Classifier. Set Optuna to maximize the 10-fold cross-validation score (cross_val_score). You are free to design the search space for hyper-parameters in these models. What is the Accuracy and F1-score on the Test Data of the best model?
- c. **[10 pts]** In the paper, the best model was found to be Random Forest, having a weighted average F1 score of 0.98. Using your own hyper-parameter search, can you find a better Random Forest model with higher F1 score?
- d. **[10 pts]** Based on the best Random Forest model in item (c), perform any feature importance method to explain the model. What insights can we get from the model?

END OF EXERCISE