AI 221: Machine Exercise 7

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.

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. In order 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?

Predicting High School Student Performance

Download the Student Performance data set from the UCI Repository:

https://archive.ics.uci.edu/ml/datasets/Student+Performance

The data set contains 30 descriptors on 600+ students from two Portuguese schools. The goal is to predict G1, G2, G3, which are the 1st, 2nd, and 3rd period grades, respectively—a regression problem. The accompanying paper to this data set can be found in:

http://www3.dsi.uminho.pt/pcortez/student.pdf

For this problem, we will only deal with the Math scores: student.zip >> student-mat

a. **[10 pts]** Make the necessary encoding for categorical inputs. Split the data into 80% Training and 20% Testing.

b. **[40 pts]** Run any AutoML procedure (either LazyPredict, Optuna, TPOT, or Auto-sklearn) to predict the G3 score using the 30 descriptors and the G1 and G2 scores as input features (32 features all in all). You may limit your search to only a few ML models, especially if you use Optuna. Report the R^2 metric on all models that were tried.

c. **[10 pts]** Based on your result in item (b), perform an explainability analysis on the best model using *Shapley values*. Report the summary plot of the most influential descriptors, then write a discussion on your analysis.

END OF EXERCISE