AI 221: Machine Exercise 3

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.

Wine Quality Data Set

Go to https://archive.ics.uci.edu/ml/datasets/wine+quality

Download the Wine Quality dataset. For this problem, we will only use the Red Portuguese "Vinho Verde" wine data. This data set contains information about 1599 wine samples with 11 attributes coming from physicochemical tests. The attributes are as follows:

1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol

The goal is to build a classifier that predicts the wine quality from the wine attributes. The wine quality is a sensory preference score from 0 (poor) to 10 (excellent), graded by experts. However, in the data set, some score values are actually empty.

Based on the original paper where the data set first appeared:

https://www.sciencedirect.com/science/article/pii/S0167923609001377

SVM was found to be the best predictor of wine quality, but the SVM parameters were only tuned by heuristics in that paper. Also, other models were not tried. Can you do better?

a. **[50 pts]** Split the samples into 70% Training and 30% Testing data at random with stratification (stratify=y). Train **SVM**, **MLP**, **Random Forest**, and **Gradient Boosting classifiers** on the data. You are free to do your own hyper-parameter tuning. Report the accuracy, precision, recall, F1-score, and confusion matrix of all the models on test data.

b. **[50 pts]** Actually, the paper treated the problem as regression rather than classification. Repeat the prediction task in (a), but this time, let's treat it as a **regression problem**. This means that we are allowing the model to predict continuous values from 0 to 10. Again, use **SVR, MLP, Random Forest, and Gradient Boosting regressors**. Report the MSE, R^2, and MAD (mean absolute deviation) on the test data. Were you able to improve against the result in the paper?

Based on your results, discuss the difference of treating this problem as classification or regression. How will this decision impact the users of your model?

END OF EXERCISE