Machine Learning

Term Project

You task is to perform multiclass classification on the dataset provided:

Minimum requirements:

- Dataset analysis and report on important statistics
- Correlation analysis
- Dealing with missing values (if applicable)
- Dealing with imbalanced data (if applicable)
- Feature transformation/engineering
- List of appropriate evaluation measures with justifications
- Three Classifiers:
    - One of linear classifier (logistic regression or SVMs)
    - One of: KNN, Decision Trees, Neural Networks
    - One of Ensemble Learning (Random Forest, Adaboost,…)
- Proper hyper-parameter tuning based on separate validation set (or cross-validation)
- Error analysis and possible improvements
- Final results on the test set

Other possible ideas to try (as examples):

- More than 3 classifiers and comparison
- Investigate the concept of margins
- Feature importance, feature selection, dimensionality reduction
- Investigate different feature scaling techniques
- Investigate different techniques for encoding categorical values.
- Clustering the data in K clusters (K= number of classes) and compare the labels
- Interpreting the learned models (for example by examining the weights of a linear model or by constructing decision rules from the learnt decision tree)
- …

Important Notes:

1. All the documents (code and report) should be submitted in Jupyter notebooks.
2. Your work will be checked with appropriate plagiarism detection tools like iThenticate.
3. You can work as a team of 2 members.
4. Best **performing** team gets a bonus.