Machine Learning

Term Project

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

Part-A (65%)

Minimum requirements:

- Dataset analysis and report on important statistics
- Correlation analysis
- Dealing with missing values (if applicable)
- Dealing with imbalanced data (if applicable)
- Feature selection/transformation/engineering
- List of appropriate evaluation measures with justifications (**we will use F1-macro as the main metric**)
- At least 4 Classifiers (each student works on 2 classifier), out of which:
    - One of linear classifier (logistic regression or SVMs)
    - One of: KNN, Decision Trees
    - One 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
- Dimensionality reduction as preprocessing before classification
- 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)
- …

Part-B (35%)-Separate Jupyter Notebook

Implement at least two active learning strategies (e.g., least confidence and entropy, each student works on one) where the dataset is same as part-A EXCEPT the training samples are not labelled except randomly selected 100 samples (initially), i.e., you start with 100 random samples labelled initially and perform active learning on the rest of the training data to achieve comparable results to part-A with minimal number of samples labelled.

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 work as a team of 2 members. It is highly recommended to form teams within the same section. You need to decide your team member ASAP.
4. Best **performing** team gets a bonus.