Machine Learning

Term Project

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

Part-A (70%)

Minimum requirements:

- [5 points] Dataset analysis and report on important statistics
- [15 points] Feature selection/transformation/engineering
- [10 points] Dealing with missing values (if applicable)
- [10 points] Dealing with imbalanced data (if applicable)
- [40 points] 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 validation set (or cross-validation)…**Note: you can use part of train set to take our validation set or perform cross validation.**
  - List of appropriate evaluation measures with justifications (**we will use GMean between Sensitivity and Specificity as the main metric**)
- [10 points] Error analysis and possible improvements
- [10 points] Final results on the test set

Other possible ideas to try (as examples):

- More classifiers and comparison
- Investigate the concept of margins
- Dimensionality reduction as preprocessing before classification
- Investigate different feature scaling techniques
- 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 (30%)-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 50 samples (initially), i.e., you start with 50 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. **Note: Use logistic regression classifier.**

Important Notes:

1. All the documents (code and report) should be submitted in Jupyter notebooks.
2. You work as a team of 2 members. You need to decide your team member ASAP. A link in Blackboard has been provided for self-enrollment.