Final Project

The purpose of this project is to compare the performance of different classifiers on the same dataset. Your task is to apply three of the following approaches:

1. K Nearest Neighbors
2. Decision Trees
3. Naïve Bayes
4. Ensemble methods: AdaBoost, Random Forest
5. Logistic Regression
6. Support Vector Machine
7. K-Means

You must use Python code to compare the performance of your chosen classifiers.

Dataset

You are free to choose any dataset. Your dataset must be big enough so that you won't have problems with overfitting. Also, make sure you don't have imbalance in the classes. If there is imbalance, you need to use data augmentation.
The following websites will give you access to datasets:
https://archive.ics.uci.edu
https://www.kaggle.com/datasets
https://paperswithcode.com/datasets

Choosing a classifier

As we discussed each classifier during the course of the semester, we focused on how some of those classifiers are more amenable to numeric data than to categorical data.

Numeric data

Support Vector Machine, Logistic Regression, AdaBoost

Categorical

Decision trees

Numeric and Categorical

K Nearest Neighbors, K-means, Naïve Bayes

If your dataset has both numeric and categorical and let's say you want to use Support Vector Machine on that dataset, one way to deal with it is to ignore the categorical data.

Training the classifier

A split 70-30 or higher is better. 70% of the data or above should be used to train the data and 30% of the data or a lower percentage should be used to test the model.

You should use an ROC curve where necessary and a confusion matrix to test the performance of classifiers.

It’s always a good idea to preprocess the data first. You might want to check for correlation, attributes you can ignore, null values, or duplicates.

Report

Your report should include:

A. Executive Summary
    a. You should capture the questions you addressed, your key results and any insight or recommendations for the future. This shouldn’t be more than half a page.
B. Main report
    Your report should answer the following questions:
    a. What problems did you specifically address?
    b. what dataset did you use?
    c. how did you preprocess the data?
    d. how did you organize the data?
    e. What kind of results did you get across the 3 different classifiers?
    f. What metrics did you use to compare the three classifiers?
    g. a conclusion section where you summarize your results and offer recommendations for future research