**Introduction to Machine Learning**

**Final**

**Fall 2024**

Q1. In this question, you are given details on an SVM binary classifier (with radial basis kernel) trained on a data set with two features x1 and x2. The file *SVM\_model3.csv* has 25 Support Vectors (SVs) of the trained model. The first two columns are the coordinates/feature values of the SVs. The third column represents the corresponding Lagrange multipliers (also called dual coefficients) for the SVs multiplied by the class label the SV.

belongs to. The estimated intercept (b0) in this model is 0.0462.

A test data is also provided in file *SV3\_test\_data.csv*. The third column in this file is the actual class label. Your goal is to write python code to compute the SVM score on the

samples in the test data and predict if samples belong to class 0 or 1. Report the overall accuracy of prediction. **[30 points]**

Q2. You are given training data *train\_data.csv* and test data *test\_data.csv* with class labels in the last column. You are required to use at **least four** of the following classification models to train on the training data. The testing data MUST not be used in training.

* Random Forest
* Decision Trees
* Naïve Bayes
* Multilayer Perceptron
* Non-Linear Support Vector Machines with Radial Basis Kernel
* Non-Linear Support Vector Machines with Polynomial Kernel
* Logistic regression
* Gradient Boosting
* XGBoost
* Adaboost
* Find one single feature that can yields the best accuracy. For this, you have to

tune **at least four of the above models** above using 5-fold cross validation on the training set. For feature selection, you can adopt any method. Note that the identification of the feature cannot be a guess. You need to justify why the feature was chosen. **Report the accuracy on the testing set for each model in a table.** Note that the training process should only involve training data, the testing data should not be involved in the training process. **[35 points].**

Your points for this Q = 35\*(a/M), where a = your best accuracy, M = highest accuracy from all students.

* Use all the features and tune at **least four of the above models** above using 5-fold cross validation on the training set. Report the accuracy on the testing set. Note that the training process should only involve training data, the testing data should not be involved in the training process. **Report the accuracy on the testing set for each model in a table. [35 points]**

Your points for this Q = 35\*(a/M), where a = your best accuracy, M = highest accuracy from all students.