CS 6375

ASSIGNMENT scikitlab2

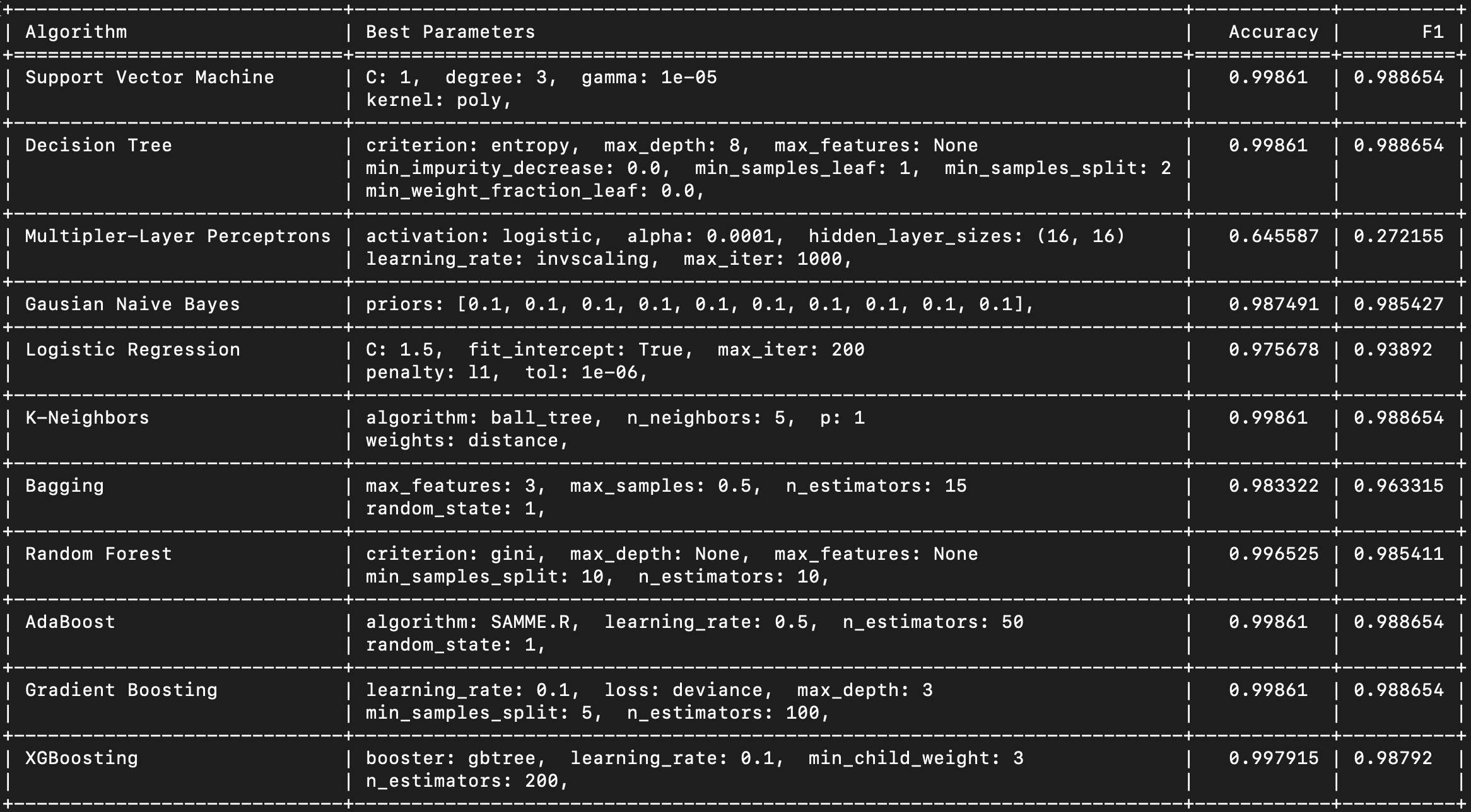
Names of students in your group:

Yiming Zhou

Number of free late days used: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

1. **Output of Code (tabular form shown below):**



1. **Analysis**

According to the experiment output, it shows best parameters for each algorithm, and Decision Tree, KNN, Random Forest, Adaboost, Gradient Goosting and XGBoosting are methods resulting in comparable high performance on test set. Given the dataset (as shown in **Source**) that each sample has 22 features and it is a classification problem, all of those perform-well algorithms except KNN are based on decision tree, while some of them using multiple decisions trees. As decision tree can split classes at each nodes, it is intuitively powerful in this kind of classification problem.

Logistic regression has a bit lower performance. That may be caused by its linearity nature.

To conclude, deep “tree” for decision tree or ensemble learning with a bunch of estimators seem to solve this kind of problem well.

1. **Source**:

Dataset: https://archive.ics.uci.edu/ml/datasets/Anuran+Calls+%28MFCCs%29