CS 6375

Lab \_\_\_\_\_2\_\_\_\_\_\_\_

Names of students in your group:

Xiaohe Yu

Number of free late days used: \_\_\_\_\_\_\_\_\_0\_\_\_\_\_\_\_\_\_\_\_   
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.

**ML methods comparison**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Best Parameters** | **Ave**  **Precision** | **Ave**  **Recall** | **Avg**  **F1** | **Accuracy\_Score** |
| Decision Tree | {'max\_depth': 100, 'max\_features': 'sqrt', 'min\_samples\_leaf': 4, 'min\_samples\_split': 8} | 0.83 | 0.81 | 0.80 | 0.81 |
| Neural Net | {'activation': 'tanh', 'alpha': 0.001, 'hidden\_layer\_sizes': (50, 50), 'learning\_rate': 'adaptive'} | 0.91 | 0.89 | 0.89 | 0.89 |
| SVM | {'C': 10, 'gamma': 0.001, 'kernel': 'rbf', 'max\_iter': 200} | 0.92 | 0.92 | 0.92 | 0.92 |
| Gaussian Naïve Bayes | {'priors': None} | 0.93 | 0.92 | 0.92 | 0.92 |
| Logistic Regression | {'C': 2, 'max\_iter': 100, 'penalty': 'l1', 'tol': 1e-05} | 0.87 | 0.83 | 0.83 | 0.83 |
| K\_NN | {'algorithm': 'ball\_tree', 'n\_neighbors': 8, 'p': 1, 'weights': 'distance'} | 0.89 | 0.89 | 0.89 | 0.89 |
| **Bagging** | **{'max\_features': 8, 'max\_samples': 4, 'n\_estimators': 20, 'random\_state': 1000}** | **0.99** | **0.99** | **0.99** | **0.99** |
| **Random Forest** | **{'criterion': 'gini', 'max\_depth': 100, 'max\_features': 1, 'n\_estimators': 50}** | **0.97** | **0.97** | **0.97** | **0.97** |
| AdaBoost | {'algorithm': 'SAMME', 'learning\_rate': 0.5, 'n\_estimators': 50, 'random\_state': 500} | 0.91 | 0.89 | 0.89 | 0.89 |
| **Gradient Boosting Classifier** | **{'learning\_rate': 1, 'max\_depth': 10, 'max\_features': 4, 'n\_estimators': 200}** | **0.98** | **0.97** | **0.97** | **0.97** |
| XGBoost | {'booster': 'gbtree', 'learning\_rate': 1, 'n\_estimators': 50, 'seed': 100} | 0.93 | 0.92 | 0.92 | 0.92 |

The data used in the lab is from UCI wine data. Use the chemical analysis results to predict the wine class. Data in the program is directly load from <https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data>. Among all methods, Bagging, Gradient Boosting and Random forest performed best. These are ensemble methods which could take advantage of integrating many weak learners to give a low variance prediction results.