CS 6375 ASSIGNMENT
SciKit Learn Lab - 2

Names of students in your group:

Akhila Kancharana (AXK180025)

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.

1.Code

Algorithm	Code
Decision Tree	https://colab.research.google.com/drive/1kYihfnpvRXERP-Z2GJnRnmbP3SrsLdFl
Neural Net	https://colab.research.google.com/drive/1N3FWmLTUP4zyJJHmoMc0HeaSRK1kyBNa
SVM	https://colab.research.google.com/drive/1V7HYGU3upEymHEovqaDswCkHOOFbCeeH
Logistic Regression	https://colab.research.google.com/drive/1v7E4DJVwJHSHr9isjpkGJiYiMSLVejpT
Random Forest	https://colab.research.google.com/drive/1BuEhN2f4EM83PPC1rbbauQRgqNuekLCp
Bagging	https://colab.research.google.com/drive/1OFIPvRK2Z6MupQIHRuEiLsb__Kil2zsk
K Nearest Neighbours	https://colab.research.google.com/drive/1cofnqIQL4WeIEov1WqwjJInuRMyd0da1
XGBoost	https://colab.research.google.com/drive/1KjNfZlk2fZK_c-ArWvrvV9qafz9LeyLP
Gradient Boosting	https://colab.research.google.com/drive/10YYVFvGUzE16MF_GnAzRhICQEaMyBawm
Adaboost	https://colab.research.google.com/drive/111oy9Lpe7j9wcVs5AzH1S6-1qp_4NO9Y
Gaussian Naive Bayes	https://colab.research.google.com/drive/1_ueJ-MjyrpiyruEjxlijK2xlI9wd2EgN

2. Evaluation Metrics:

Algorithm	Best Parameters	Avg Precision	Avg Recall	Avg F1	Accuracy Score
Decision Tree	{'criterion': 'entropy', 'max_depth': 20, 'max_features': 25, 'max_leaf_nodes': 80, 'min_impurity_decrease': 0.001}	0.84	0.83	0.83	0.83
Neural Net	{'activation': 'tanh', 'alpha': 1.3717421124828532e-15, 'hidden_layer_sizes': 11, 'learning_rate_init': 0.06}	0.92	0.91	0.91	0.914
SVM	{'C': 10, 'gamma': 0.001, 'kernel': 'rbf', 'max_iter': 10}	0.99	0.99	0.99	0.99
Logistic Regression	{'C': 1.0, 'fit_intercept': 1, 'max_iter': 50, 'penalty': 'l2'}	0.95	0.95	0.95	0.95
Random Forest	{'criterion': 'entropy', 'max_depth': None, 'max_features': 'auto', 'min_samples_split': 5}	0.94	0.94	0.94	0.94
Bagging	{'max_features': 0.5, 'max_samples': 0.5, 'n_estimators': 10, 'random_state': 1}	0.95	0.94	0.94	0.94
K Nearest Neighbours	{'algorithm': 'auto', 'n_neighbors': 8, 'p': 3, 'weights': 'distance'}	0.99	0.99	0.99	0.986
XGBoost	{'booster': 'gblinear', 'min_child_weight': 1, 'seed': 7}	0.98	0.97	0.98	0.975
Gradient Boosting	{'learning_rate': 0.5, 'max_depth': 4, 'min_samples_leaf': 8, 'n_estimators': 10}	0.95	0.95	0.95	0.95
Adaboost	{'algorithm': 'SAMME.R', 'learning_rate': 0.01, 'n_estimators': 100,'random_state': 1}	0.71	0.65	0.65	0.65
Gaussian Naive Bayes	{'priors': None}	0.87	0.82	0.83	0.825

3. Analysis

In the table above, we deduce that SVM, K Nearest neighbour classifiers performed the best among the classifiers given. This might be because, we are checking on a single dataset digits and the classifiers SVM,KNN fit well for this kind of dataset. As the data does not undergo pre-processing or scaling it might result in a significant increase or decrease on accuracy. I would suggest on preprocessing the data sets, testing on other dataset and implementation of various other parameters to improve and get rid of false positives increasing the accuracies and obtain optimum results.