CS 7641 – Supervised Learning
mgupta318

The datasets chosen for the classification problems are:

1. White wine quality – This dataset has multiple features, which help in predicting the wine quality. This is an interesting classification problem where the last column 'quality' contains the values from 0-10. This column was transformed to a classification problem into multi-classification "High", "Medium" and "Low" which was transformed into 0, 1 and 2 using a label encoder. Using this classification we can predict the quality of wine based on its attributes
2. Adult Income Dataset – This dataset has multiple columns related to an individual like age, state. The target column is income, which is classified as a binary classification problem. This dataset had both categorical and integer data which was transformed using encoding into numerical data. This is useful in predicting the income of an individual based on its different attributes

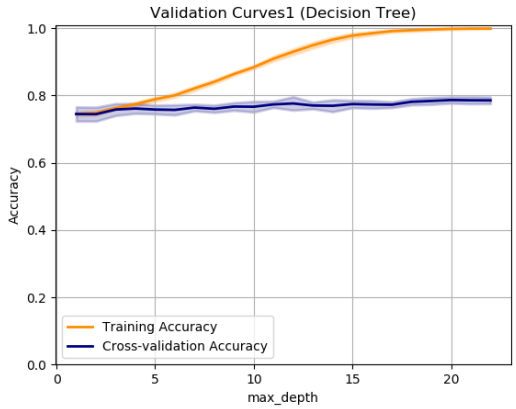
Both datasets were non-trivial and run on multiple classifiers including Decision Trees, Neural Networks, Boosting, Support Vector Machines and *k*-Nearest Neighbors.

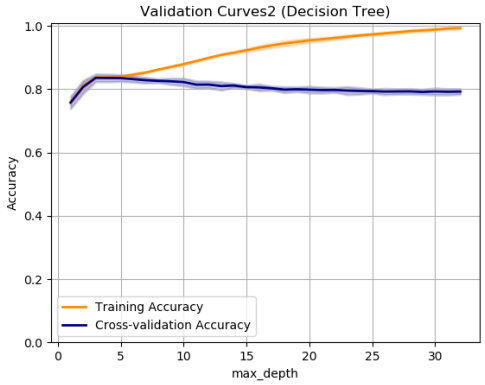
The results of the analysis for all the classifiers are outlined below.

Decision Trees

Both datasets were run using split criteria of 'Gini index' and 'Entropy' and accuracy were compared. Dataset1 had a higher accuracy with Entropy and Dataset2 had a higher accuracy with Gini index.

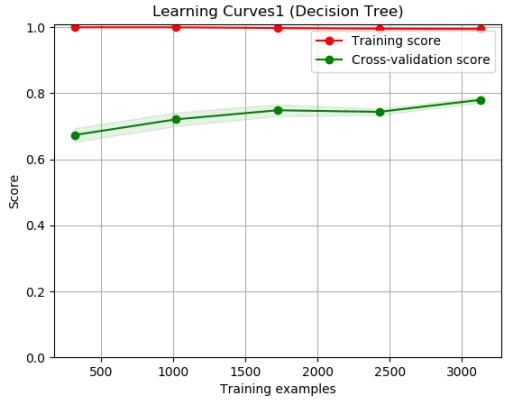
Validation curve – for dataset 1 and dataset2

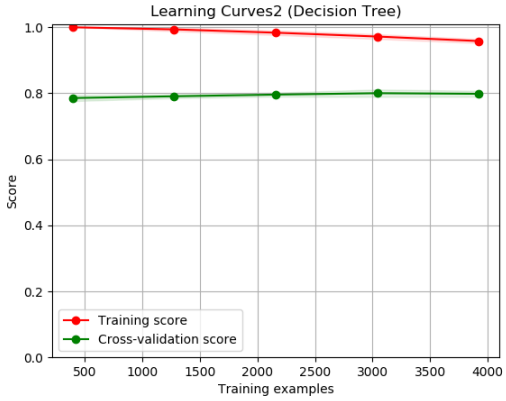




Validation curves were plotted for both datasets for 'max_depth' parameter. As we see in the plots above the cross validation accuracy for dataset 1 became stable around max_depth = 18 and dataset 2 became stable around max_depth = 21 which were chosen as hyper parameters for learning.

Learning curve – for dataset 1 and dataset2



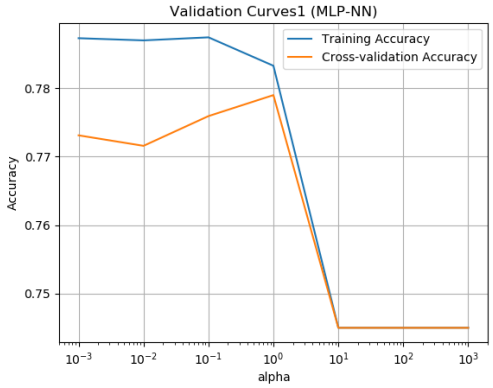


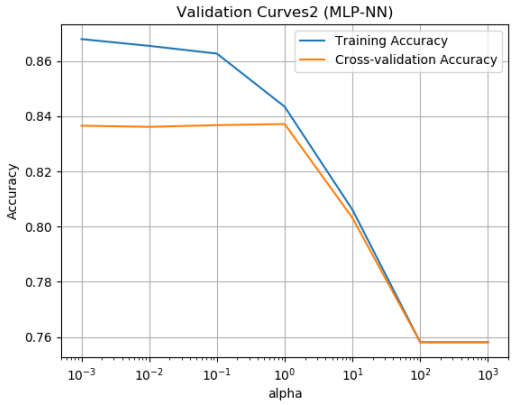
The learning curve for dataset2 performed better since the cross validation curve is going in a smooth line till the end. For dataset1 it is trying to converge after more training examples.

Neural Networks

MLP classifier was used for this analysis. The model was trained on 5 input layers and 2 hidden layers.

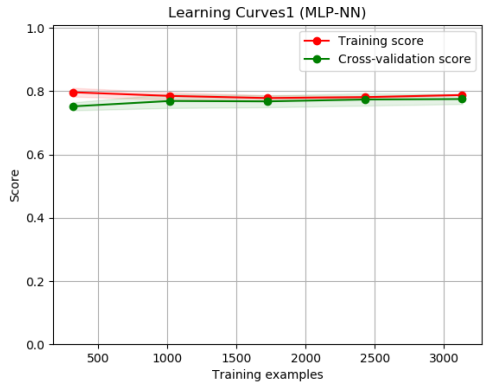
Validation curve – for dataset 1 and dataset2

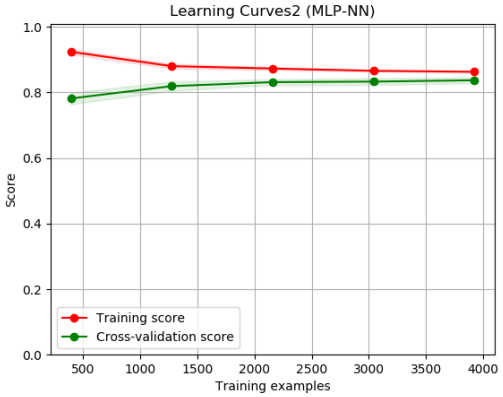




The validation curves were plotted to find the best value of the regularization parameter 'alpha' for the MLP classifier.

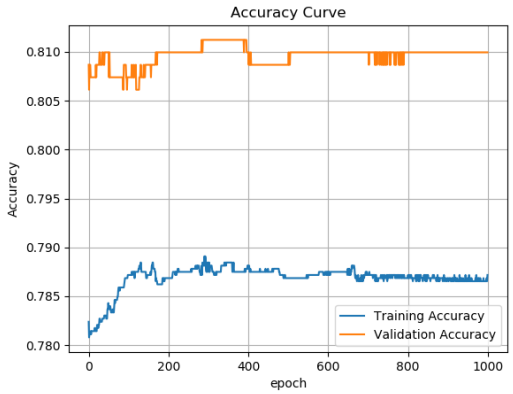
Learning curve – for dataset 1 and dataset2

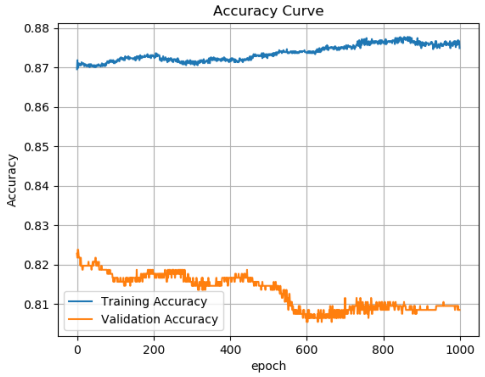




The model was fitted with the value of alpha=0.1 found from validation. Learning curve for both dataset1 and dataset2 indicates overfitting indicating further need for tuning.

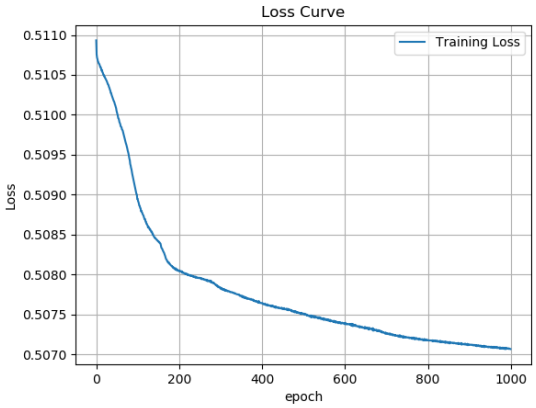
Accuracy curve – for dataset 1 and dataset2

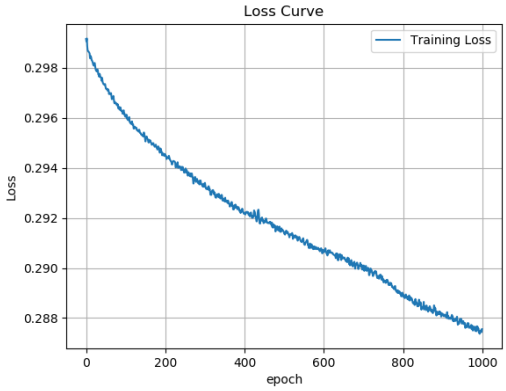




The validation and accuracy curves were run for 1000 epochs. Training accuracy for dataset2 is better and validation accuracy is better for dataset1

Loss curve – for dataset 1 and dataset2



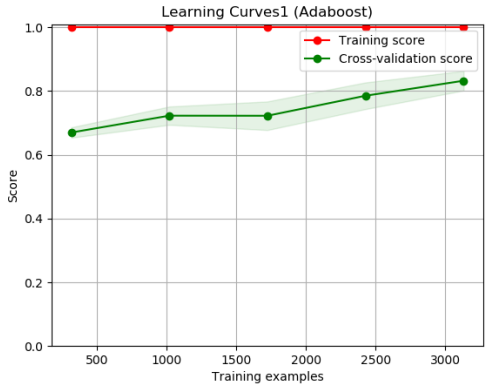


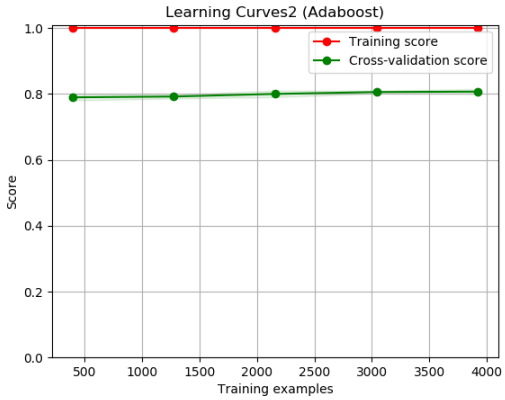
Both the datasets loss curve indicate that with increasing number of epochs the training loss is decreasing which is helping in tuning the model

Boosting

Boosting was performed using the Adaboost classifier in which decision tree was used as a weak learner with the same hyperparamters of the above tuned Decision Tree. As observed in the classifiers accuracy graph towards end of this paper, Adaboost performed better than stand-alone Decision Tree

Learning curve – for dataset1 and dataset2





Tried with multiple num_learners = 500, 1000, 2000. Weak learner Decision tree already pruned for max_depth. Final parameters chosen -> learning_rate=1.0, num_learners=1000

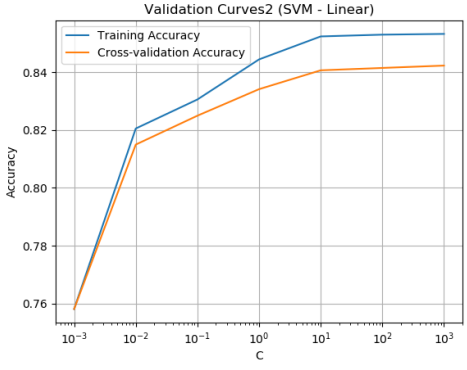
Dataset2- Tried with multiple num_learners = 500, 1000, 2000. Weak Learner Decision tree already pruned for max_depth. Learning rate of 0.01 gave the best accuracy. Tried with learning rate 1, 0.01 and 0.05. Final parameters chosen -> learning_rate=0.01, num_learners=1000

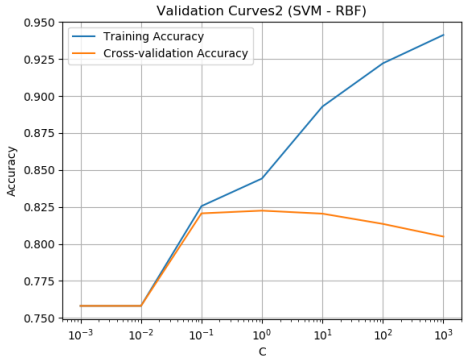
Support Vector Machines

Support Vector Machines was run using the below two kernels:

a) kernel='linear'
b) kernel='rbf'

Validation curves -> SVM – Linear and SVM - RBF

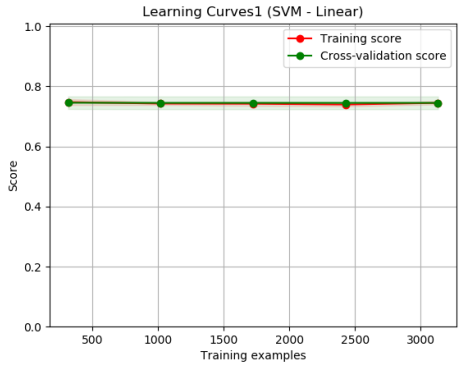


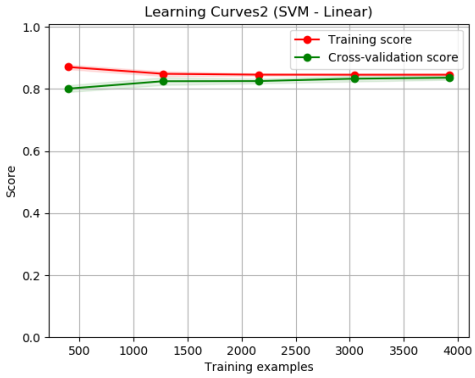


The validation curve for dataset1 and dataset2 was found for regularization parameter 'C'. Based on the validation curves the optimal

value chosen for dataset1 → C=1.2 (SVC-Linear and SVC-RBF) and dataset1 → C=1.2 (SVC-Linear) and C=1.5(SVC-RBF)

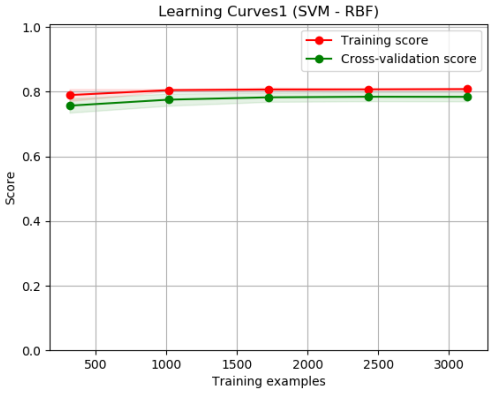
Learning curve -> SVM – Linear (dataset1) and SVM – RBF (dataset1)

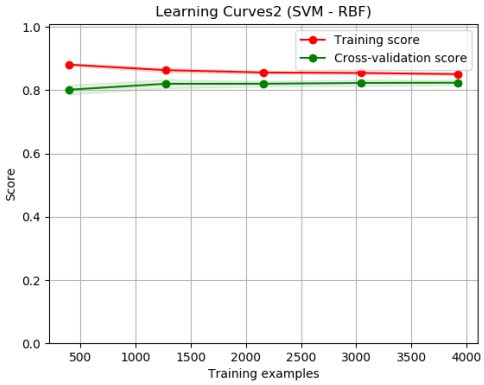




Dataset1 and Dataset2 denote overfitting drawing a conclusion that the points are not in linear plane.

Learning curve -> SVM – Linear (dataset2) and SVM – RBF (dataset2)

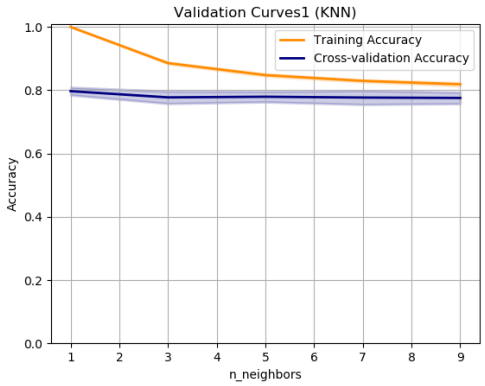


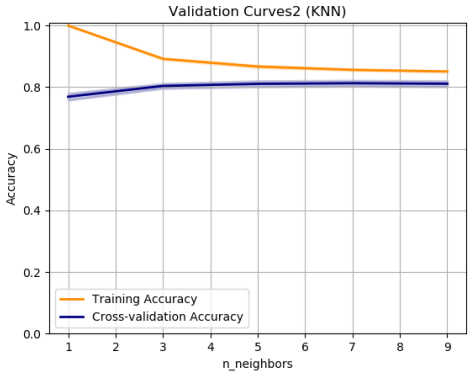


Dataset1 and Dataset2 both perform better with Kernel='RBF' than Kernel='Linear' and cross validation moves parallel to the training score.

k-Nearest Neighbors

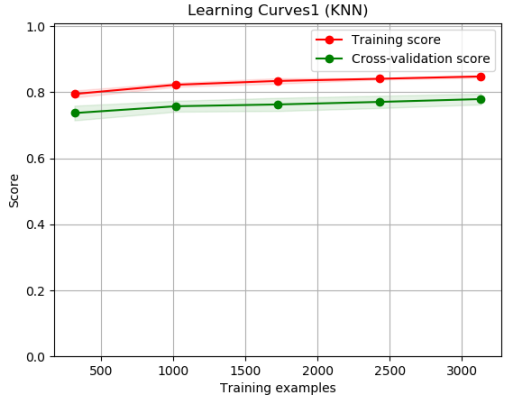
Validation curve – dataset1 and dataset2

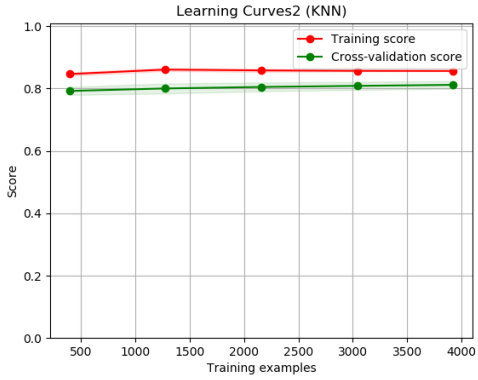




n_neighbors was chosen as a hyper parameter and its optimal value was found using validation curves. The cross validation score becomes stable around n_neighbors=5 and for dataset2 the cross validation score becomes stable around n_neighbors=6

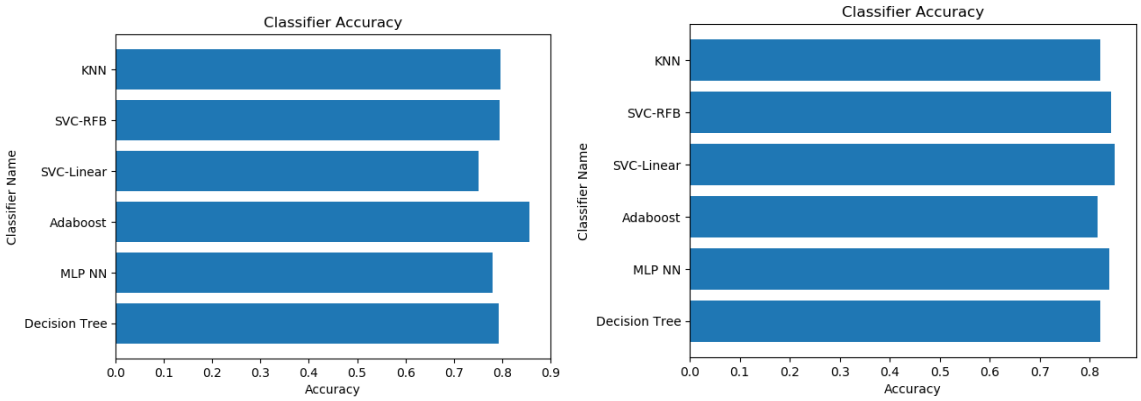
Learning curve – dataset1 and dataset2





Using the n_neighbor value from validation curve, learning curves for both dataset1 and dataset2 moved well along with training score. However, with more training examples for dataset2 it tends to converge which might result in overfitting

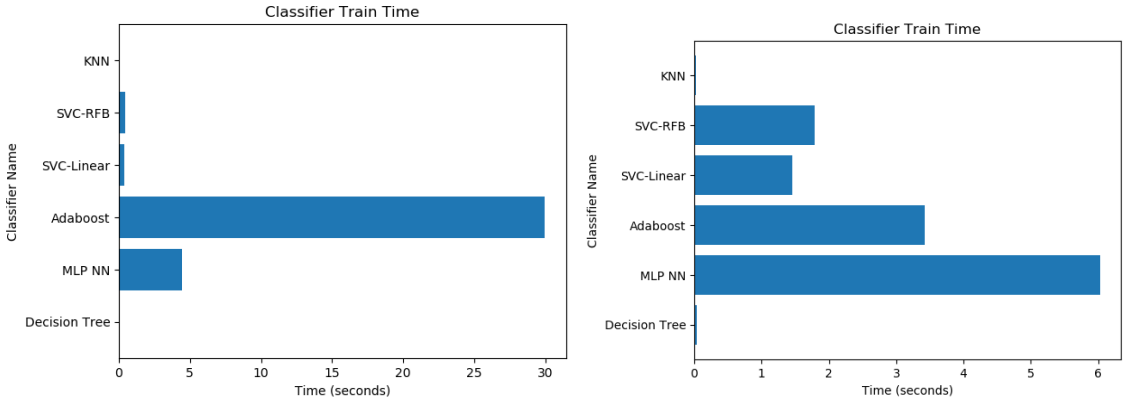
Classifiers Accuracy Score Analysis – dataset1 and dataset2



Dataset1- Adaboost accuracy is highest and SVC linear was lowest. Adaboost improves accuracy by combining weak learners. SVC linear accuracy denotes the samples are in non-linear plane

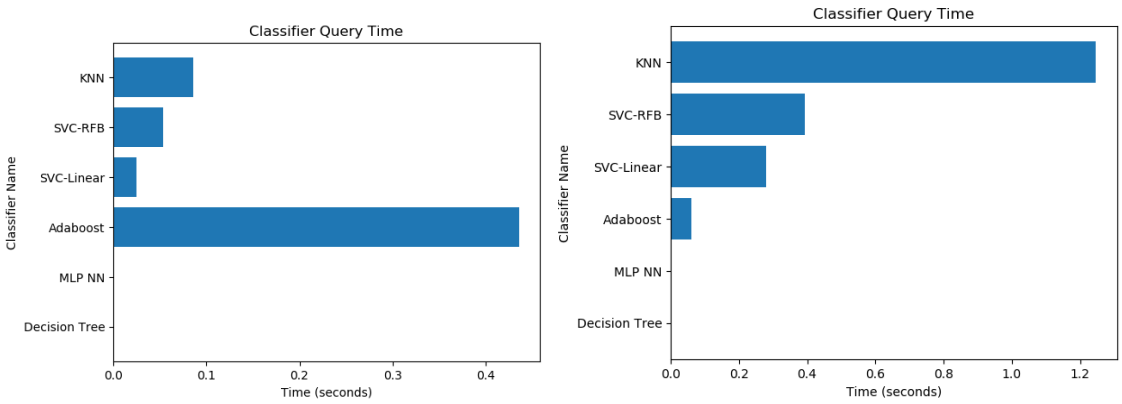
Dataset2 – SVM linear performs the best and Adaboost is worst. This signifies that there is still improvement for tuning

Classifiers Training Time Analysis – dataset1 and dataset2



Due to more number of instances in dataset2 MLP classifier took a long time in training. KNN and Decision tree takes the minimum training time

Classifiers Query (Predict) Time Analysis – dataset1 and dataset2



KNN takes the maximum query time in both datasets. MLP and Decision Tree performs best in the query time.

Citations:

[1] https://scikit-learn.org/stable/modules/neural_networks_supervised.html
[2] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html
[3] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
[4] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
[5] https://scikit-learn.org/stable/auto_examples/tree/plot_cost_complexity_pruning.html
[6] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html
[7] https://scikit-learn.org/stable/auto_examples/model_selection/plot_validation_curve.html
[8] https://datatofish.com/plot-dataframe-pandas/