Number of Instances in Dataset : 699

Number of Attributes in Dataset : 9

n Fold Cross-Validation : n = 10

Classifier	Parameter1	Parameter2	Parameter3	Parameter4	Parameter5	Parameter6	Parameter7	Accuracy	Precision	Recall
Decision Tree	criterion='entropy'	splitter='random'	max_features='sqrt'					0.9488	0.9608	0.9260
Perceptron	penalty = 'l1'	alpha=0.001	max_iter=200	shuffle=True				0.9590	0.9685	0.9414
Neural Network	hidden_layer_sizes=(8,6)	activation='tanh'	alpha=0.005	learning_rate='adaptive'	max_iter=200	learning_rate_init=0.001	solver='lbfgs'	0.9472	0.9596	0.9241
Deep Learning	hidden_layer_sizes=(20,16,14,10,9,8,7,6,5,4,2)	activation='tanh'	solver='lbfgs'	alpha=0.005	learning_rate='adaptive'	max_iter=60	learning_rate_init=0.001	0.9619	0.9706	0.9461
SVM	C=1.0	kernel='rbf'	max_iter=70					0.9707	0.9773	0.9588
Naïve Bayes								0.9619	0.9703	0.9402
Logistic Regression	penalty='l1'	max_iter=100	solver='liblinear'	C=2				0.9693	0.9764	0.9560
K N N	n_neighbors=20	algorithm='kd_tree'						0.9678	0.9753	0.9538
Bagging	base_estimator=neuralNet	n_estimators=10	bootstrap=True					0.9678	0.9751	0.9546
Random Forest	n_estimators=50	criterion='gini'	max_features='auto'					0.9693	0.9763	0.9563
ADA Boost	base_estimator=SVM	n_estimators=80	learning_rate=0.001	algorithm='SAMME'				0.9663	0.9736	0.9535
Gradient Boosting	loss='deviance'	learning_rate=0.05	n_estimators=100	presort='auto'				0.9678	0.9752	0.9542

Data-Set Info:

Breast Cancer Wisconsin (Original) Data Set

http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

Pre-processing of Dataset:

1. The Dataset contained irregular values like '?'. These were replaced with NaN values. In the Dataset considered, all the attributes are important (X1- X9).
2. All the instances that contained NaN or null values in any of its cells, were removed by using dropna function.
3. The Dataset was scaled using standardscaler function, by importing sklearn preprocessing library.
4. The Data obtained from the previous step was standardized using Transformation function.

Pseudo Code:

1. Read the data (data should be in .xls format)
2. Data Preprocessing – as mentioned above
3. Create the various classifier models with the required parameters.
4. Create 10-fold cross validation for the data set.
5. Create a list of models from step 3.
6. For each model, predict the output and calculate the accuracy and f1-Score.

Analysis:

The processed data is fed to the Classifiers and the corresponding Models are built. 10-fold Cross Validation is performed on the dataset. The Results proved us that, then the accuracy metric, we found the Precision metric as the good evaluation metric to be used along with the Accuracy. Also, the results show the SVM Model performed well than the other Classifiers, because its one of the Complex Model and fits this Dataset perfectly. Bagging is taking time. Deep Learning performed better than Regular Neural Network, as the number of Hidden Layers increased with less number of neurons in each Hidden Layer. Naïve Bayes, Logistic Regression and K-nearest Neighbor performed similar to each other for the Dataset considered. Decision Tree, Perceptron and Neural Network performed decent, as they are Simple Models (Neural Network here has less number of Hidden

Layers, so considered as Simple). Among them, Decision Tree performed very weakly. Overall, the Complex Models performed better compared to Simple Models as the Dataset would require a highly non-linear function to fit perfectly.