CS 6375
ASSIGNMENT _____5(Part II)_____

Names of students in your group:

Muhammad Arsalan Malik (mxm162431)
Md Shihabul Islam (mxi170330)

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.

http://scikit-learn.org/stable/documentation.html
http://scikit-learn.org/stable/modules/grid_search.html
http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

REPORT:

➢ Dataset:

URL = https://archive.ics.uci.edu/ml/datasets/Car+Evaluation (Car Dataset)
Number of instances in datasets = 1728
Number of attributes in datasets = 6
Number of fold cross-validation performed = 10
Evaluation Metric (other than Accuracy) = Recall

➢ Pre-processing:

- The car dataset has no missing values
- Standardized numerical values by using preprocessing function 'Standardscaler()' of scikit-learn, it removes the mean and scales features to unit variance
- Transformed nominal values to numeric by using function 'LabelEncoder()', it encode labels with values between 0 and # of classes-1
- Encoded categorical features using a one of k scheme, input features take on values in the range [0, # of values]. It is needed for feeding categorical data to scikit-learn estimators

➢ Results

#	Classifier	Best Parameters Used	Accuracy (%)	Recall (%)
1	Random Forest	N_estimators = 40, criterion = gini, max_features = log2, random_state = 10, bootstrap = False	97.69	74.0
2	Bagging	base_estimator = Decision Tree, max_samples = 0.9, n_estimators = 30, bootstrap = False	97.45	81.0
3	Adaboost	base_estimator = Decision Tree, n_estimators = 150, learning_rate = 0.5, algorithm = SAMME	97.0	78.0

4	K-Nearest Neighbors	N_neighbors = 10, Weights = distance, Algorithm = ball_tree, Leaf_size = 15, P = 1	95.60	54.0
5	Gradient Boosting	Random_state = 20, N_estimators = 100, Max_features = sqrt, Learning_rate = 1.0, Criterion = friedman_mse	98.84	91.0
6	Support Vector Machine	kernel = poly, gamma = 0.1, cost = 100, degree = 5, coef0 = 0.5	99.80	98.0
7	Neural Net	hidden_layer_sizes = (100,50,20), activation = tanh, solver = lbfgs, alpha = 1.0, max_iter = 300	99.9	98.0
8	Perceptron	penalty = none, random_state = 200, alpha = 0.00001, eta0 = 16, shuffle = False	91.12	54.0
9	Naïve Bayes	alpha = 0.001, binarize = 0.0, fit_prior = True	92.79	62.0
10	Deep Learning	Hidden_layer_sizes = (300, 200, 100, 80, 60, 40, 20), Alpha = 1.0, Solver = lbfgs, Activation = tanh	98.0	85.0

11	Decision Tree	Criterion = entropy, Min_samples_split = 2, Max_depth = 50, Splitter = random	97.37	78.0
12	Logistic Regression	Penalty = l2, C = 10, Solver = lbfgs, Multi_class = multinomial, Max_iter = 400	94.13	76.0

➢ Analysis:

- For parameter tuning, i.e., to find the best parameters, we used RandomizedSeachCV and GridSearchCV functions of the Scikit-learn library. It gave us an idea of what parameter values will give the best results. Different permutations of parameters were then tested (details in the Log File).

- We have used two evaluation metrics to compare the performance of different classifiers on the 'Car' dataset, namely Accuracy and Recall.

  Based on accuracy, all of the classifiers gave pretty good results. Neural Net, SVM, Gradient boosting and Deep learning had the best accuracy, i.e., greater than 98% accuracy.

  Recall of the classifiers was not observed to be as good as the accuracy. The weakest classifiers based on recall were Perceptron, KNN, Naïve Bayes respectively with recall less than 70%. Neural network, SVM, Gradient boosting and Deep learning produced the best results based on recall, i.e., recall greater than 85%.

  Neural Net, SVM and Gradient boosting gave the best results overall, i.e., Both Accuracy and Recall greater than 90%.

- Some attributes do influence the output more than others, like no. of hidden layers in NN has a greater influence on output than the no. of iterations.

- Every classifier we tested gave great accuracy, while the value of recall varied, which shows that the evaluation metric 'recall' gives a better insight into the performance of classifiers than 'accuracy'.