RESULTS OF THE ASSIGNMENT GIVEN BY SILVERSPARRO FOR SUMMER INTERNSHIP TEST. **Submission Date** : *24th March'18*

Submission By : *Rakshit Jain (IIT B.H.U)*

Bonus Task:

Create a model to output 1 if all 3 fields are predicted correctly and 0 if 2 or fewer fields are predicted correctly with above 90% precision and recall.

Conclusion : Random Forest classifier model is used to predict correctness index.

	DATE			TIME		PREDICTION_GT	CONFIDENCE SCORE		
Independent variables(Features)	Pred_day	pred_month	Pred_year	Pred_hour	Pred_min	Pred_gt	gt_confidence	dt_confidence	tm_confidence
Feature importance	0.11311374	0.3746187	0.19034065	0.04416201	0.06803098	# 0.07144354		0.13829037	
	✓	✓	✓	✓	✓	✓	×	✓	×

.

Comparison matrix :

	0	1
0	355	15
1	20	110

Accuracy = (355+110)/(335+110+20+15) = 0.96875

Note: 1) At the very first attempt Support Vector Machine(SVM) classification model was used so as to avoid influence of outliners and avoid overfitting. But due to numerous independent variables and complex kernel requirement accuracy was not high.

2) On second attempt Naive Bayes classification model was used which gave an accuracy equal to 0.886

3) Finally Random Forest classification model is used to get an accuracy of 0.92 which was increased by recalling and removing some independent variables(gt_confidence and tm_confidence) to 0.96875

Task1:

Train a model (ML/DL or anything else) to output 1 if the prediction is correct and 0 if the prediction is incorrect.

a) List down features created and used (if using ML techniques)

b) Get recall and precision above 90% on validation set provided. We shall check accuracy on test set available separately with us

Conclusion : Random Forest classifier model is used to predict correctness index.

	DATE			TIME		PREDICTION_GT	CONFIDENCE SCORE		
Independent variables(Features)	Pred_day	pred_month	Pred_year	Pred_hour	Pred_min	Pred_gt	gt_confidence	dt_confidence	tm_confidence
Feature importance	0.04499998	0.04050885		0.08073076	0.09136551	0.07764796	0.31959038	0.21363613	0.13152043
	✓	✓	×	✓	✓	✓	✓	✓	✓

Comparison Matrix :

	0	1
0	9	8
1	0	483

Accuracy = (9+483)/(9+483+0+8) = 0.984

Note: 1) Random Forest classifier is directly used as the nature of data for Bonus Task and Task 1 is same. Initially when all the features (independent variables) were used, the accuracy was 0.978.

2) On recalling and removing the least important feature i.e. Pred_year, the accuracy was increased to 0.984.

3) Since the number of independent variables are more so Random Forest is powerful and accurate.