Week 6 (22.11 - 28.11.2019)

Part 1

Check metrics with hand-picked features

File: hand_picked_features.py

1. nAcid, nBase, nAtom, nH, nB, nC, nN, nO, nS, nP, nF, nCl, nBr, nI, nX, BalabanJ, nHBAcc, nHBDon, TopoPSA, Radius, Vabc, MW

	Random Forest	Gradient Boosting	SVM Classifier
Accuracy	0.90	0.90	0.89
Precision	0.83	0.73	0.65
Recall	0.31	0.36	0.21
F1 Score	0.45	0.48	0.32
ROC-AUC Score	0.65	0.67	0.60

2. VE3_A, nAromAtom, nAtom, nHeavyAtom, nC, AATS7d, GATS8i, BalabanJ, VE3_DzZ, nBondsO, nBondsM, C2SP2, C3SP2, C4SP3, RNCG, SpMAD_Dt, StCH, SaaCH, AETA_dBeta, fMF, GhoseFilter, PEOE_VSA7, MDEC-23, AMID_C, piPC10, n5Ring, n6Ring, nHRing, RotRatio, SLogP, JGI9, Radius, VAdjMat, MWC05

	Random Forest	Gradient Boosting	SVM Classifier
Accuracy	0.90	0.90	0.89
Precision	0.94	0.70	0.76
Recall	0.30	0.41	0.25
F1 Score	0.46	0.51	0.37
ROC-AUC Score	0.65	0.69	0.62

Part 2

Stratifying dataset

Performance after setting ‘stratify’ parameter in train_test_split()

	Random Forest	Gradient Boosting	SVM Classifier
Accuracy	0.90	0.90	0.89
Precision	0.91	0.69	0.72
Recall	0.24	0.36	0.22
F1 Score	0.38	0.48	0.34
ROC-AUC Score	0.61	0.67	0.60

Part 3

Running Feature Selection on Stratified Folds of dataset

File: stratifiedKFold.py

I set number of folds to 5. I ran SelectKBest() and SelectPercentile() feature selection methods from sklearn to see the differences between features selected in various folds. On an average, 85% of the features were same between all folds.

File: ensemble_feature_selection_stratifiedKFold.py

Here I run the ensemble feature selection methods from the previous weeks on stratified folds. On an average, 47% of the features were same between all folds.

Part 4

Running the Evolution Algorithm for Feature Selection

I am running the algorithm on a stratified subsample of 200 datapoints from the original dataset. As of yet, I have these results:

	Best roc_auc	**Number of selected features**	**Test roc_auc**
BGA	0.66	60	0.60
BPSO	0.69	640	0.64
BPS	0.68	164	0.60
BFFA	0.70	157	0.71
BBA	0.68	487	0.58
BGSA	0.64	901	0.49
BDFA	0.64	722	0.52