CSE472 (Machine Learning Sessional)
Assignment# 2: Logistic Regression with Bagging and Stacking

Report Submission
Name: MD Sadik Hossain Shanto
ID: 1905101
Lab group: B2

Description(Function parameters):

- path : csv file path
- top_n_features : select top n features(ex: 20) based on correlation. If you want to run the dataset on all the features then select a large number(ex: 10000)
- n_base_classifiers: number of base classifiers used in bagging and stacking
- lr : learning rate for the "Logistic Regression Model"
- num_iteration: number of iteration for the model
- regularization: 'l1', 'l2' or 'none'
- lambda: model's complexity penalty, used in regularization
- val_size: validation set size
- test_size: test set size
-

To run the code on "Dataset:01: Telco Customer Churn": use this function as shown below.

```
run_dataset1_churn(
    path = 'WA_Fn-UseC_-Telco-Customer-Churn.csv',
    top_n_features=20,
    val_size=0.2,
    test_size=0.2,
    n_base_classifiers=9,
    lr=0.01,
    num_iteration=10000,
    regularization='l2',
    lambda_=0.01
)
```

To run the code on "Dataset:02: Adult(Income Census)": use this function as shown below.

```
run_dataset2_adult(
    path_data = 'adult.data',
    path_test = 'adult.test',
    top_n_features=20,
    n_base_classifiers=9,
    lr=0.01,
    num_iteration=10000,
    regularization='l2',
    lambda_=0.01
)
```

To run the code on "Dataset:03: Credit Card Fraud Detection": use this function as shown.

```
run_dataset3_creditcardfraud(
    path='creditcardfraud.csv',
    top_n_features=20,
    val_size=0.2,
    test_size=0.2,
    n_base_classifiers=9,
    lr=0.01,
    num_iteration=10000,
    regularization='l2',
    lambda_=0.01
)
```

Dataset-1: Churn Dataset

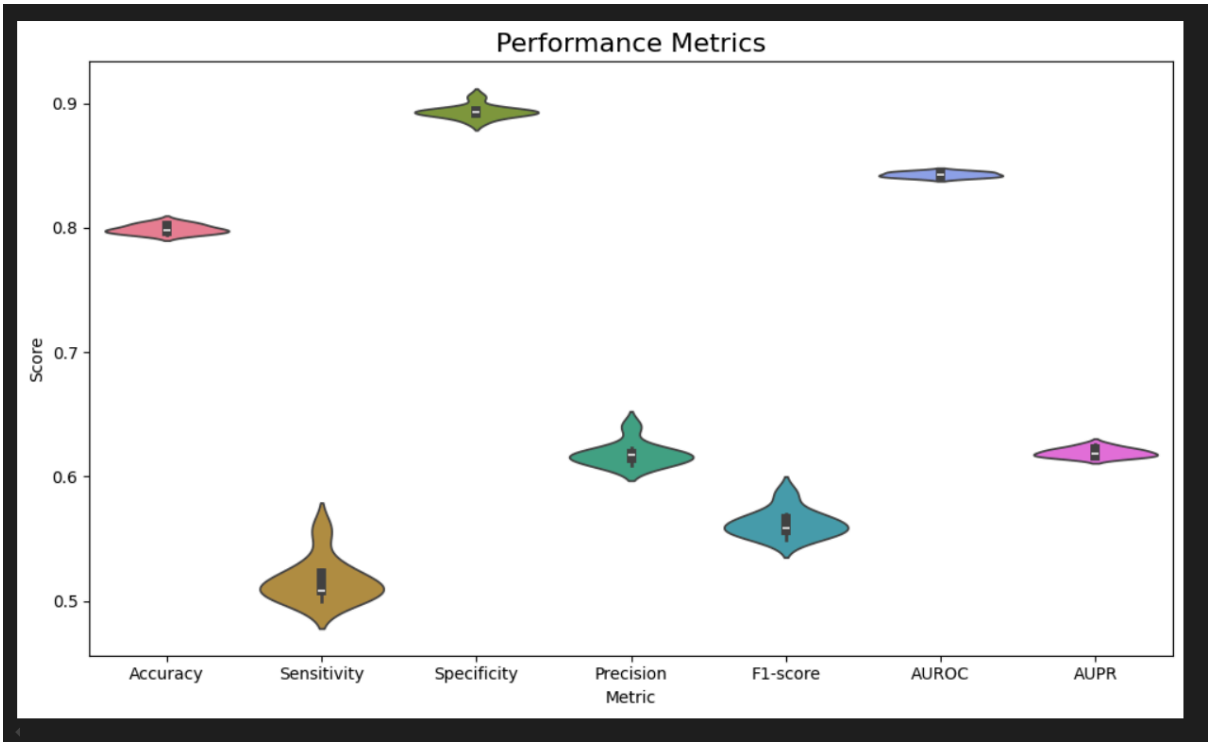
Number of iterations: 10000

Learning Rate: 0.01

Regularization: l2

Top n features = ALL

Model	Accuracy	Sensitivity	Specificity	Precision	F1	AUROC	AUPR
Logistic Regression	0.798893 ± 0.003374	0.516414 ± 0.016170	0.893321 ± 0.004904	0.618163 ± 0.008870	0.562572 ± 0.010279	0.842721 ± 0.001916	0.619363 ± 0.003330
Majority Voting	0.797153024911032	0.5198863636363636	0.8898385565052231	0.6120401337792643	0.5622119815668203	0.8436771561771562	0.6214123786560468
Stacking Ensemble	0.800711743772242	0.5568181818181818	0.8822412155745489	0.6125	0.5833333333333334	0.8376769835103168	0.6193555850230408



Dataset-1: Churn Dataset

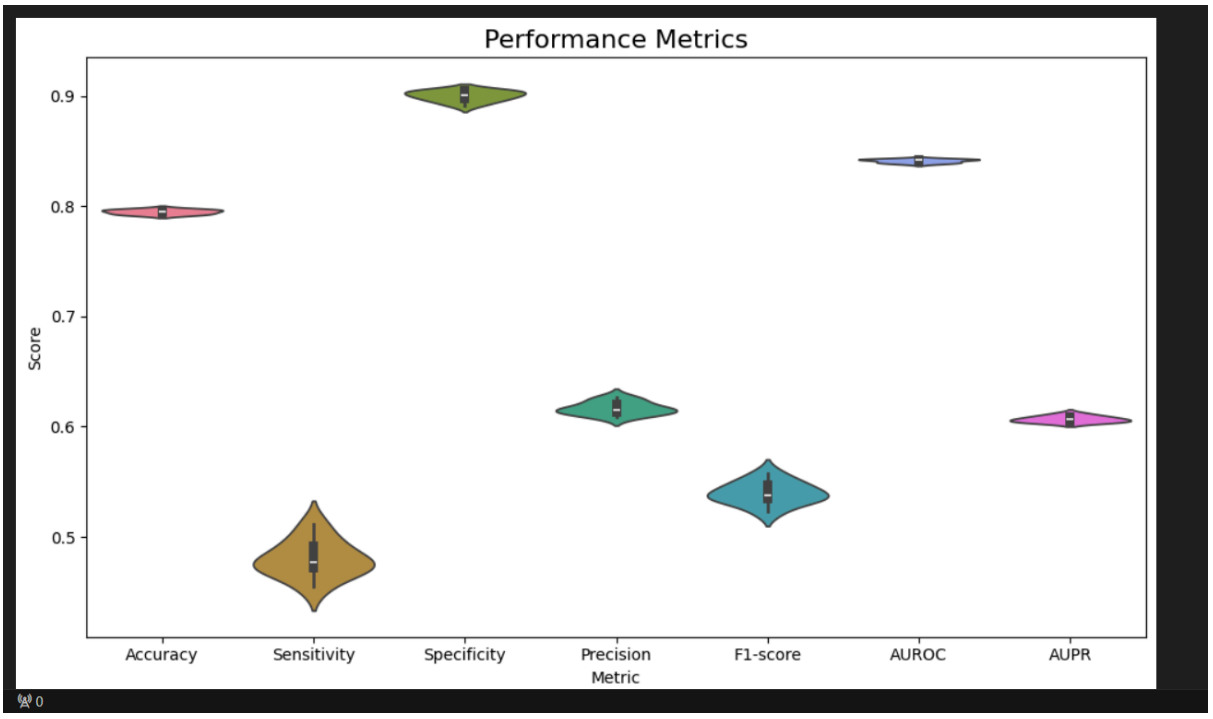
Number of iterations: 10000

Learning Rate: 0.01

Regularization: l2

Top n features = 20

Model	Accuracy	Sensitivity	Specificity	Precision	F1	AUROC	AUPR
Logistic Regression	0.795018 ± 0.001985	0.480114 ± 0.015618	0.900285 ± 0.004432	0.616879 ± 0.005519	0.539803 ± 0.009523	0.841291 ± 0.001526	0.606956 ± 0.002665
Majority Voting	0.797153024911032	0.48011363636363635	0.9031339031339032	0.6236162361623616	0.5425361155698234	0.8418789929206596	0.6062176097914601
Stacking Ensemble	0.795017793594306	0.5085227272727273	0.8907882241215574	0.608843537414966	0.5541795665634675	0.8347699753949754	0.5950082249338955



Dataset-2: Adult Data(Income Census)

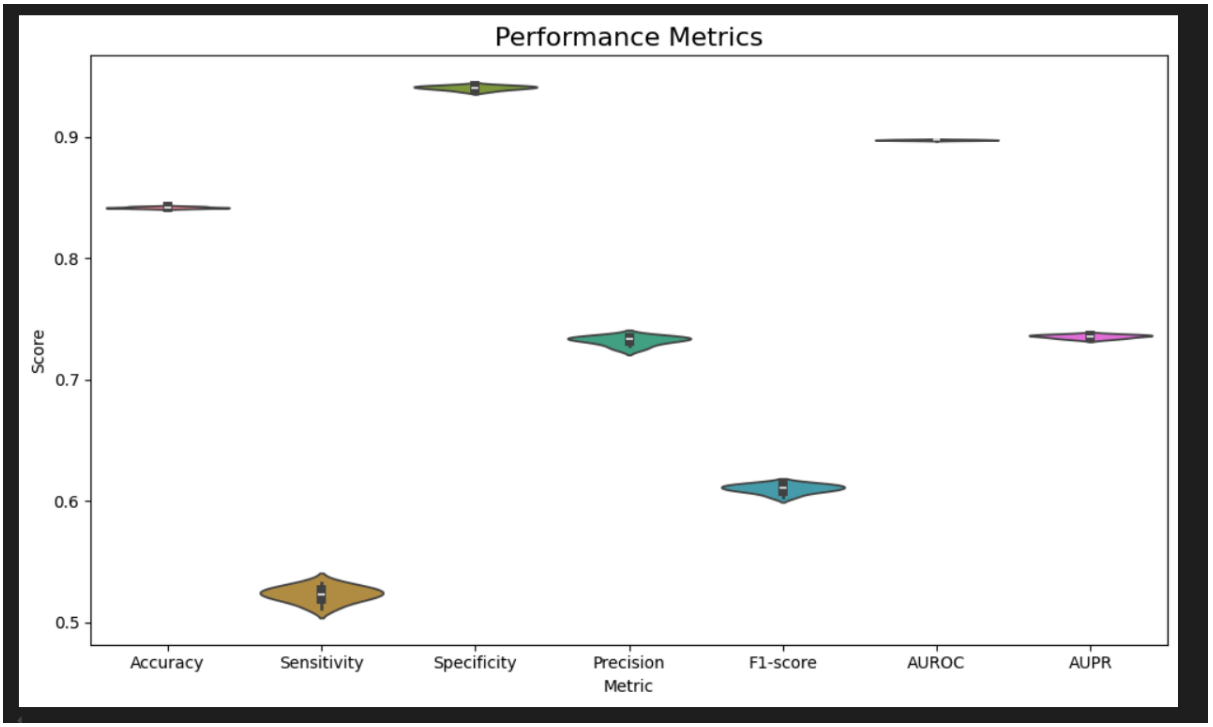
Number of iterations: 10000

Learning Rate: 0.01

Regularization: l2

Top n features = ALL

Model	Accuracy	Sensitivity	Specificity	Precision	F1	AUROC	AUPR
Logistic Regression	0.842003 ± 0.000620	0.523046 ± 0.005901	0.940754 ± 0.001600	0.732186 ± 0.003419	0.610159 ± 0.003251	0.897702 ± 0.000299	0.735650 ± 0.001367
Majority Voting	0.8421052631578947	0.5222366710013004	0.9411385779853451	0.73311427528295	0.6099635479951397	0.8978535133935784	0.735992328152544
Stacking Ensemble	0.8422897196261683	0.5271781534460338	0.9398502294870763	0.7307137707281903	0.6124792264692551	0.8985736021958048	0.7378290873939668



Dataset-2: Adult Data(Income Census)

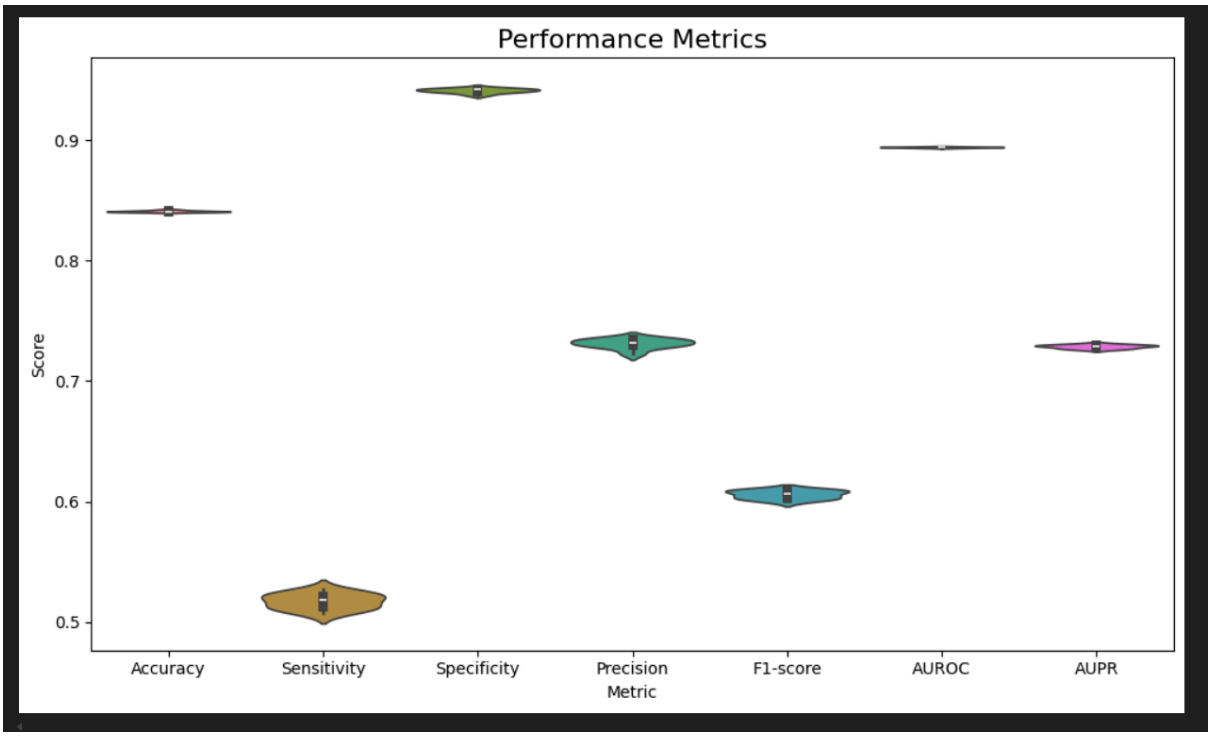
Number of iterations: 10000

Learning Rate: 0.01

Regularization: l2

Top n features = 20

Model	Accuracy	Sensitivity	Specificity	Precision	F1	AUROC	AUPR
Logistic Regression	0.840882 ± 0.000615	0.517237 ± 0.006093	0.941085 ± 0.001730	0.731102 ± 0.003764	0.605812 ± 0.003298	0.894100 ± 0.000355	0.728698 ± 0.001430
Majority Voting	0.8413674372848008	0.5183355006501951	0.941380143328704	0.7324513046674017	0.6070667072799268	0.8942115624461073	0.7289773244059258
Stacking Ensemble	0.8418593212001968	0.5219765929778933	0.9408970126419196	0.7322145202480846	0.6094746431825083	0.8942817074931643	0.7296685359992228



Dataset-3: Credit Card Fraud Detection

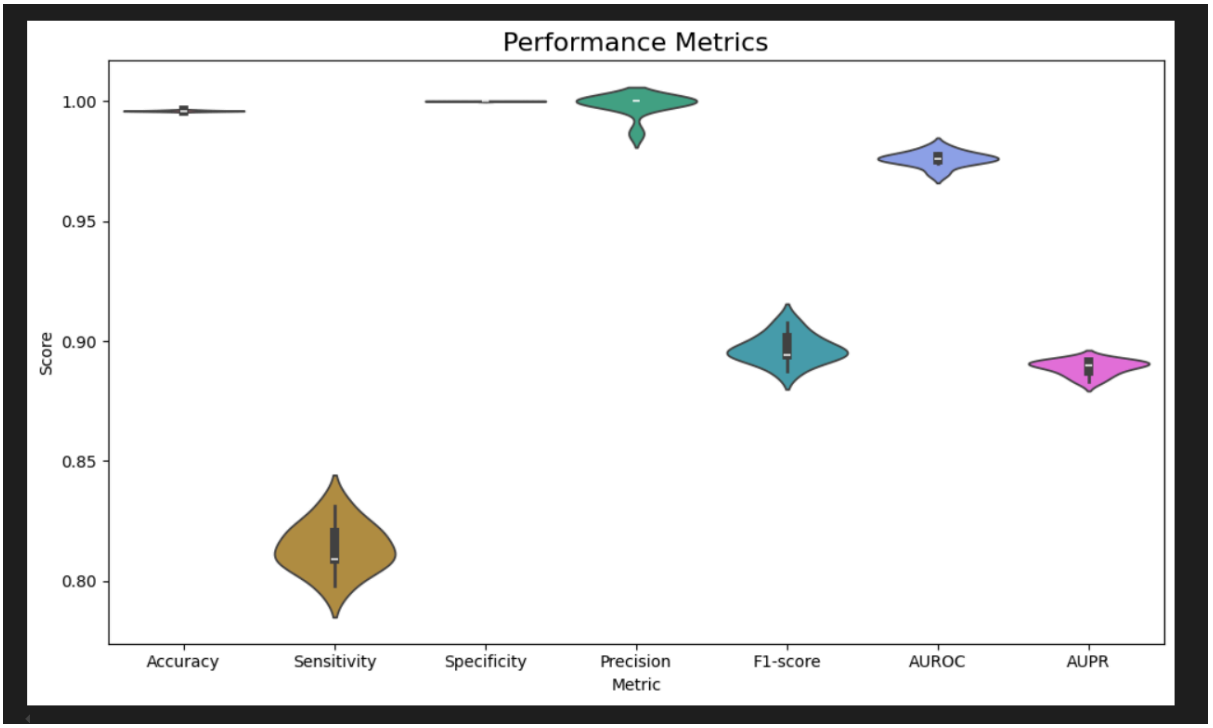
Number of iterations: 10000

Learning Rate: 0.01

Regularization: l2

Top n features = ALL

Model	Accuracy	Sensitivity	Specificity	Precision	F1	AUROC	AUPR
Logistic Regression	0.995929 ± 0.000199	0.813983 ± 0.009342	0.999972 ± 0.000078	0.998498 ± 0.004247	0.896810 ± 0.005525	0.975882 ± 0.002842	0.889034 ± 0.002852
Majority Voting	0.995847581827064	0.8089887640449438	1.0	1.0	0.8944099378881988	0.9769108838670761	0.8903265080942038
Stacking Ensemble	0.9948705422569614	0.7752808988764045	0.9997503121098626	0.9857142857142858	0.8679245283018868	0.9647996184544599	0.8771945725731187



Dataset-3: Credit Card Fraud Detection

Number of iterations: 10000

Learning Rate: 0.01

Regularization: l2

Top n features = 20

Model	Accuracy	Sensitivity	Specificity	Precision	F1	AUROC	AUPR
Logistic Regression	0.994328 ± 0.000192	0.750312 ± 0.013789	0.999750 ± 0.000204	0.985551 ± 0.011722	0.851821 ± 0.006220	0.965076 ± 0.002905	0.880327 ± 0.002704
Majority Voting	0.9943820224719101	0.7528089887640449	0.9997503121098626	0.9852941176470589	0.8535031847133758	0.9658572851351541	0.8815247991180784
Stacking Ensemble	0.9941377625793845	0.7303370786516854	1.0	1.0	0.8441558441558441	0.9636185105696531	0.8705659388865561