CSE 472

Machine Learning Sessional

Offline 2 Report

Submitted by:

Name: Tanhiat Fatema Afnan

Section: A

Roll: 1905014

Running the file:

Run all cells. The last cell will ask for input to run on a dataset.

In prompt for input:

For dataset 1: input 1 (The csv file should be in the same folder)

For dataset 2: input 2 (The files should be under the folder ‘adult’)

For dataset 3: input 3 (The csv file should be in the same folder)

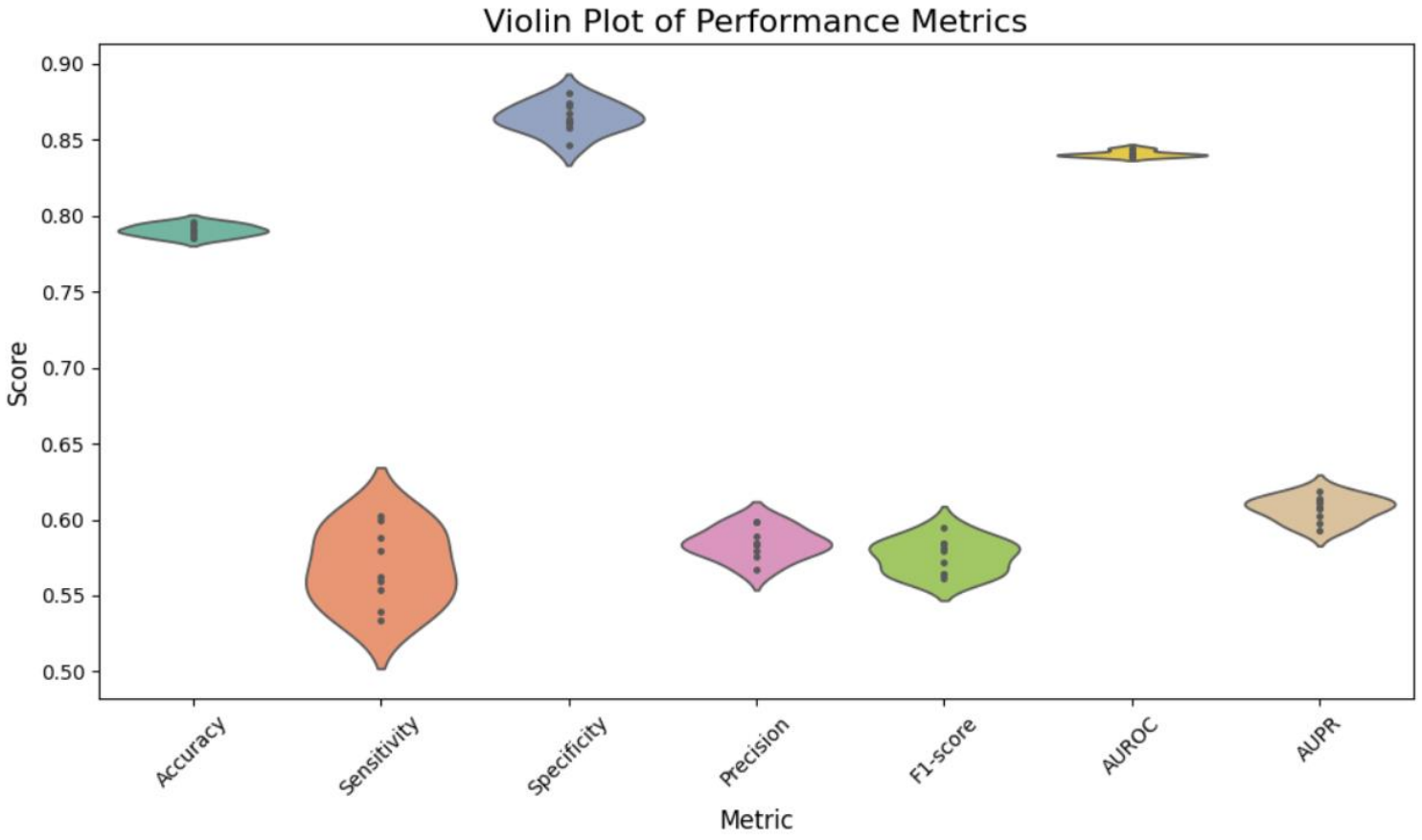
The outputs are at the end of the file

Tables and Plots:

1. Dataset 1

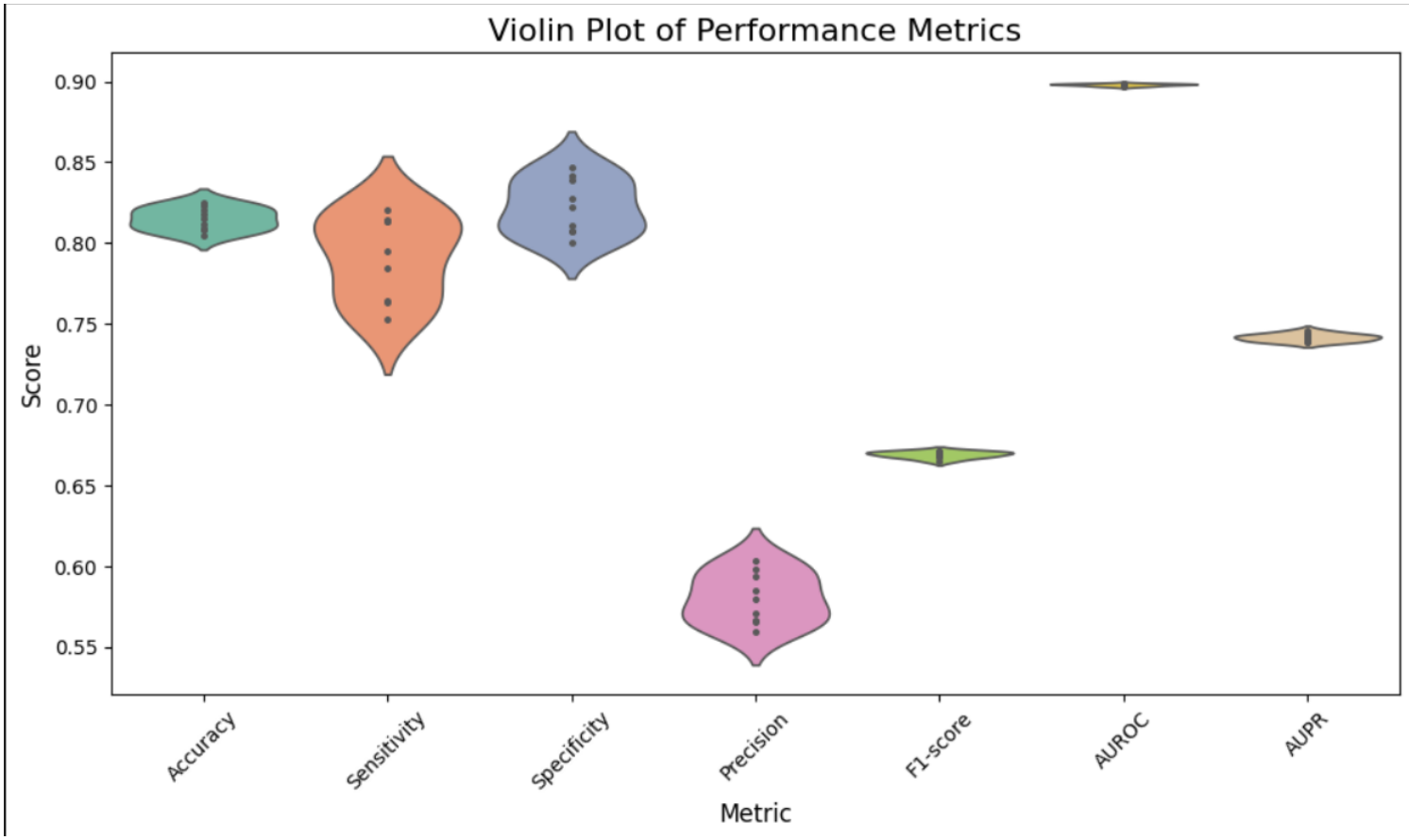
Performance on Test set

	Accuracy	Sensitivity	Specificity	Precision	F1-score	AUROC	AUPR
LR	0.7905100 8303 ± 0.0035508 0179	0.5688131 3131± 0.0234182 2852	0.8646196 053± 0.0094767 23908	0.5844710 280 ± 0.0096372 564519	0.5761564 381 ± 0.0105636 41513	0.8408161 6257 ± 0.0018946 8530037	0.6072517 3098± 0.0076842 979355
Voting Ensemble	0.7935943 060498221	0.5681818 181818182	0.8689458 689458689	0.5917159 76331361	0.5797101 449275363	0.8440467 711301044	0.6173342 686564606
Stacking Ensemble	0.7316725 978647687	0.8267045 454545454	0.6999050 332383666	0.4794069 192751236	0.6068821 689259646	0.8447940 947940948	0.6124116 488106599



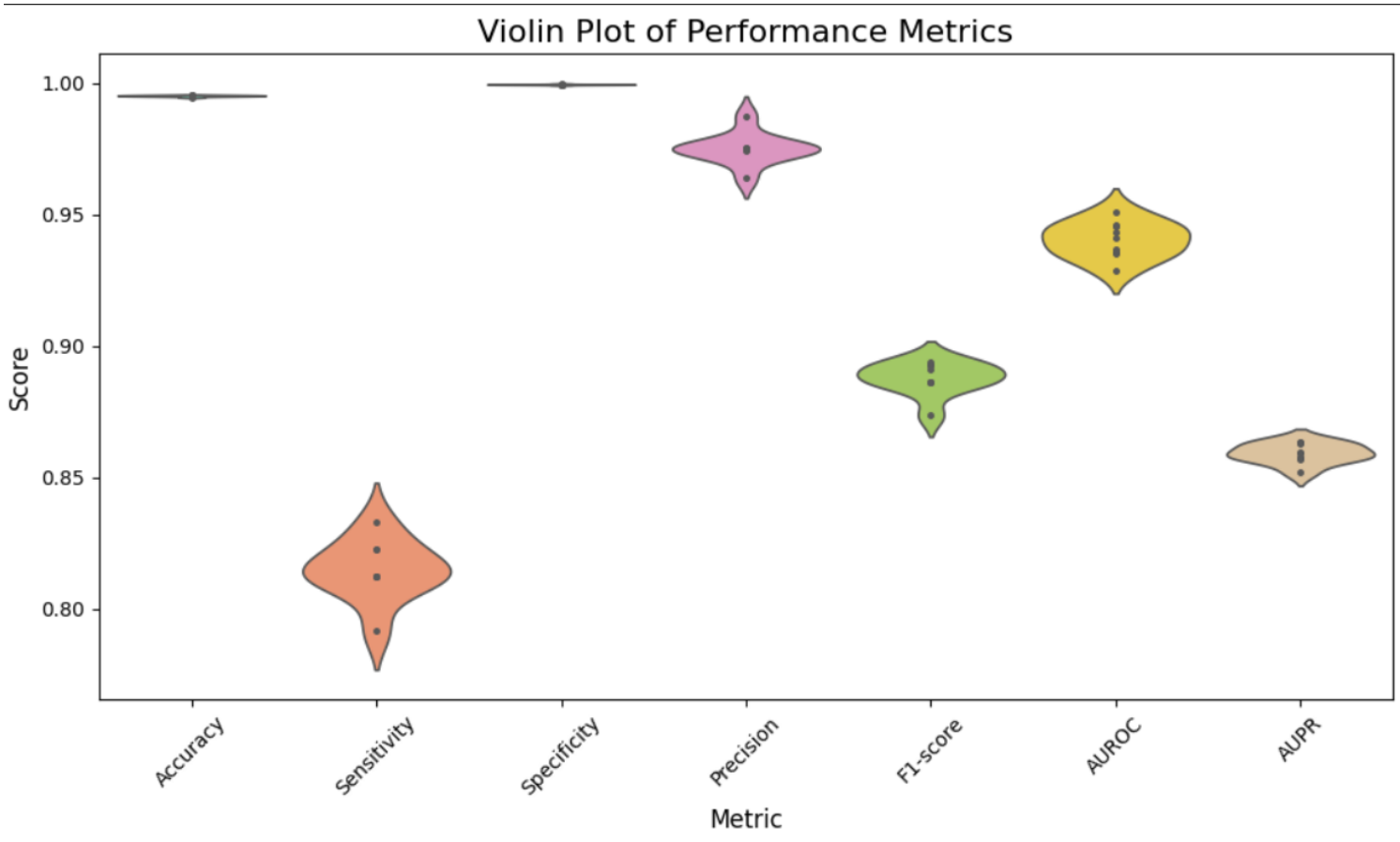
2. Dataset 2

	Accuracy	Sensitivity	Specificity	Precision	F1-score	AUROC	AUPR
LR	0.8151765 1795± 0.0065940 421473	0.7912694 28 ± 0.0245176 24383	0.8225707 009 ± 0.0161129 16204	0.5805395 5355± 0.0148412 08687	0.6691980 3477± 0.0019270 1599191	0.8980477 13 ± 0.0006319 67490	0.7417152 91 ±0.002232 658314
Voting Ensemble	0.8160432 405871875	0.8000520 020800832	0.8209891 435464415	0.5802376 013577221	0.6726418 187780085	0.8987266 076891568	0.7431844 638400442
Stacking Ensemble	0.7947914 747251398	0.8200728 029121165	0.7869722 557297949	0.5435119 765638463	0.6537465 022282102	0.8945648 626105881	0.7104785 495890729



3. Dataset 3

	Accuracy	Sensitivity	Specificity	Precision	F1-score	AUROC	AUPR
LR	0.9951749 74926 ± 0.0002235 29078941	0.8148148 1481± 0.0107333 54740	0.9995003 7471 ± 0.0001177 6280809	0.9751303 835 ± 0.0055474 6589	0.8877346 4648 ± 0.0058487 96324	0.9404225 7658 ±0.006459 0005692	0.8591954 5637 ± 0.0036011 224968
Voting Ensemble	0.9953647 23103195	0.8229166 666666666	0.9995003 747189608	0.9753086 419753086	0.8926553 672316384	0.9525538 13806312	0.8646400 948383735
Stacking Ensemble	0.9958526 4698707	0.8333333 333333334	0.9997501 873594804	0.9876543 209876543	0.9039548 02259887	0.9710321 425597469	0.8797344 568388392



Observations:

1. In data-scaling, standard scaling gives higher metrics score in almost all areas than min-max scaling
2. Feature selection by correlation analysis gives better output than by information gain in most cases.
3. A learning rate of 0.001 with 1000 iterations gives a higher score in all metrics compared to learning rate of 0.01 with 1000 iterations