Machine Learning - Assignment 1

Maor Sagi

Introduction

In this assignment we requested to detect Covid-19 by using regular blood tests by using different classifiers, experience feature extraction and explainability tools.

Data and Preprocessing

The data provided is Blood Test dataset, with 5645 samples and 111 features. I commit the preprocessing as described on the paper provided. First I filtered the 23 most significant features. Later, filter the samples to 608 rows by filter null values with 95% threshold. Because of imbalanced data, I used Iterative Imputer to fill empty records.

** I notice the "Lactic Dehydrogenase" (LDH) missing ratio on the dataset provided in the assignment and in the paper itself is significantly different from the reported missing rate on the peper 83% instead of 0.98%. It is not clear how they fill this missing data and it might affect the scores significantly. Details for missing ratio appear in the Appendices.*

Hyperparameters Tuning

The next step was choosing the best combination of hyperparameter and creating the best model from each kind. The method I used is Grid Search and the search made by nested cross validation on the specified grid parameters. The grid parameters for the CATBoost and LGBM models were identical to the XGBoost, except the max_depth for CATBoost that was eliminated because of time considerations. Best model choosing criteria was F1-Score.

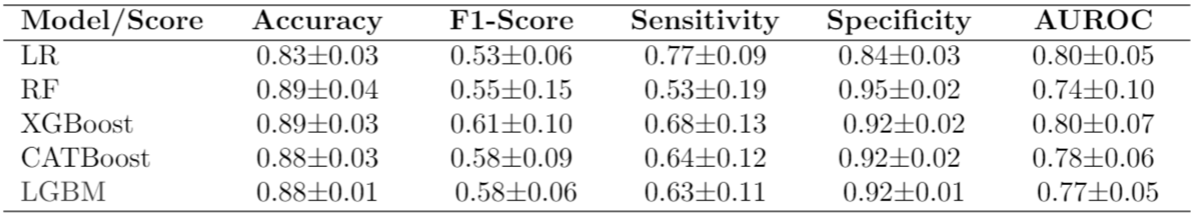
Models Training

The models trained with 10 iterations and splited - 80% train set and 20% test set. The partitions were different for each iteration.

Evaluation Results and Discussion

The evaluation metrics are according to the paper metrics - Accuracy, F1-score, Sensitivity, Specificity and AUROC. F1-score first chosen as the negative class f1-score (the default one).

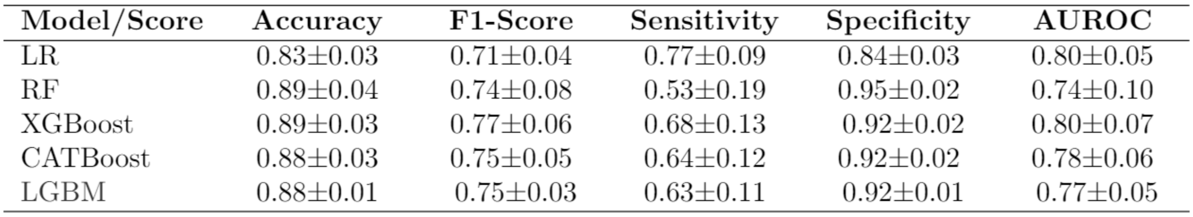
Results with F1-Score for Negative:



Model/Score	Accuracy	F1-Score	Sensitivity	Specificity	AUROC
LR	0.83±0.03	0.53±0.06	0.77±0.09	0.84±0.03	0.80±0.05
RF	0.89±0.04	0.55±0.15	0.53±0.19	0.95±0.02	0.74±0.10
XGBoost	0.89±0.03	0.61±0.10	0.68±0.13	0.92±0.02	0.80±0.07
CATBoost	0.88±0.03	0.58±0.09	0.64±0.12	0.92±0.02	0.78±0.06
LGBM	0.88±0.01	0.58±0.06	0.63±0.11	0.92±0.01	0.77±0.05

But after short research it seems by the paper result, the metric chosen by the paper authors to be shown is the Macro F1-Score.

Results with Macro F1-Score:



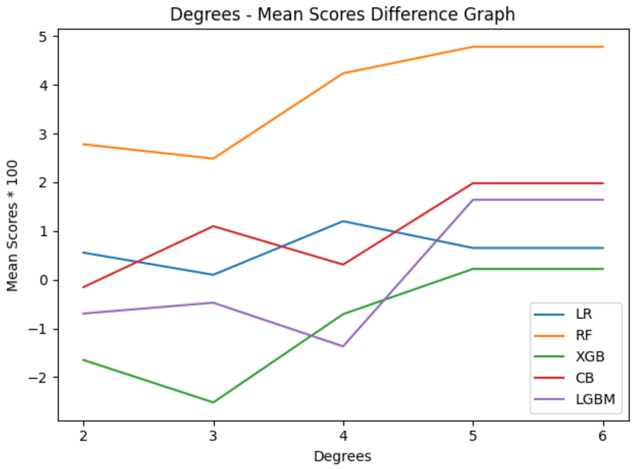
Model/Score	Accuracy	F1-Score	Sensitivity	Specificity	AUROC
LR	0.83±0.03	0.71±0.04	0.77±0.09	0.84±0.03	0.80±0.05
RF	0.89±0.04	0.74±0.08	0.53±0.19	0.95±0.02	0.74±0.10
XGBoost	0.89±0.03	0.77±0.06	0.68±0.13	0.92±0.02	0.80±0.07
CATBoost	0.88±0.03	0.75±0.05	0.64±0.12	0.92±0.02	0.78±0.06
LGBM	0.88±0.01	0.75±0.03	0.63±0.11	0.92±0.01	0.77±0.05

** Full detailed results appear in the Appendices.*

Feature Extraction

The feature extraction method I chose is extracting by polynomial combinations of the features with degree less than or equal to the specified degree. I used Polynomial Features Model for feature extraction and then selected the best 5 by using SelectKBest model with "mutual_info_classif" who provided the best scores.

The degree selection made by few experiments. First I calculated the contribution of the new features by calculating the difference between the mean of all metric scores, till 6 degrees (chosen by resources considerations).

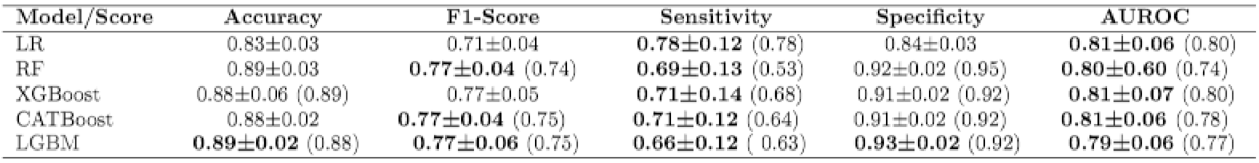


If we take a look at 2,3,4 degrees, there is no contribution at all for part of the models, and the scores get lower sometimes. It looks like 5,6 degrees provide better results, I chose to show the 5 results because the selection of 5 best features is the same for both of them.

The features selected when degree=5:

1. HCT^2 PLT WBC^2 (=HCT^2 * PLT * WBC^2)
2. HCT^2 WBC^3 (=HCT^2 * WBC^3)
3. PLT WBC^4 (=PLT * WBC^4)
4. WBC^5 (=WBC^5)
5. WBC^3 MONO^2 (=WBC^3 * MONO^2)

Results after adding the new features:



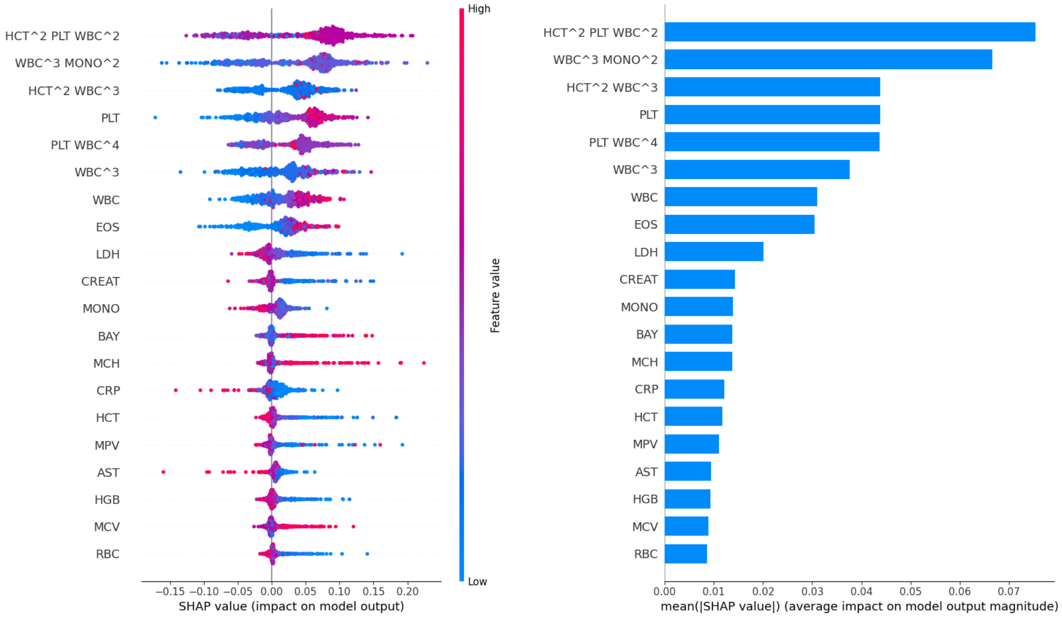
Model/Score	Accuracy	F1-Score	Sensitivity	Specificity	AUROC
LR	0.83±0.03	0.71±0.04	**0.78±0.12** (0.78)	0.84±0.03	**0.81±0.06** (0.80)
RF	0.89±0.03	**0.77±0.04** (0.74)	**0.69±0.13** (0.53)	0.92±0.02 (0.95)	**0.80±0.60** (0.74)
XGBoost	0.88±0.06 (0.89)	0.77±0.05	**0.71±0.14** (0.68)	0.91±0.02 (0.92)	**0.81±0.07** (0.80)
CATBoost	0.88±0.02	**0.77±0.04** (0.75)	**0.71±0.12** (0.64)	0.91±0.02 (0.92)	**0.81±0.06** (0.78)
LGBM	**0.89±0.02** (0.88)	**0.77±0.06** (0.75)	**0.66±0.12** (0.63)	**0.93±0.02** (0.92)	**0.79±0.06** (0.77)

The Bold results are the results who showed improvement by adding the features, and the results that were before are the ones on the brackets next to the current (if there are no parenthesis, the results were identical).
As you can see, the F1-Score that monitored the best model, improved for RF, CATBoost and LGBM, or didn't change for the rest. The Sensitivity and the AUROC improved for all models, and even though the Specificity got lower for most of the models, its score is still high. Note the LGBM results were better for all metrics.

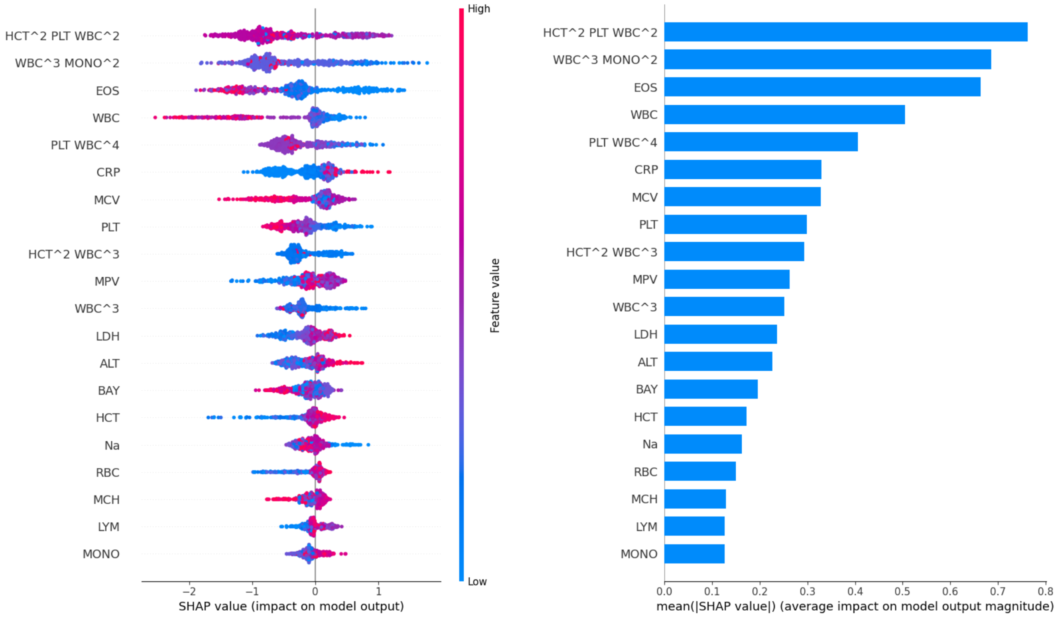
Explainability

The method to measure features importance is SHAP method. Here there are easy to read figures that present the features importance. The left one gives an overview of the most important features for each model, the plot sorts the sum of SHAP values over all samples and shows the distribution of the impacts each feature has on the model output. The colors represent the feature value, low or high. The right one gives the mean absolute value of the SHAP values for each feature.

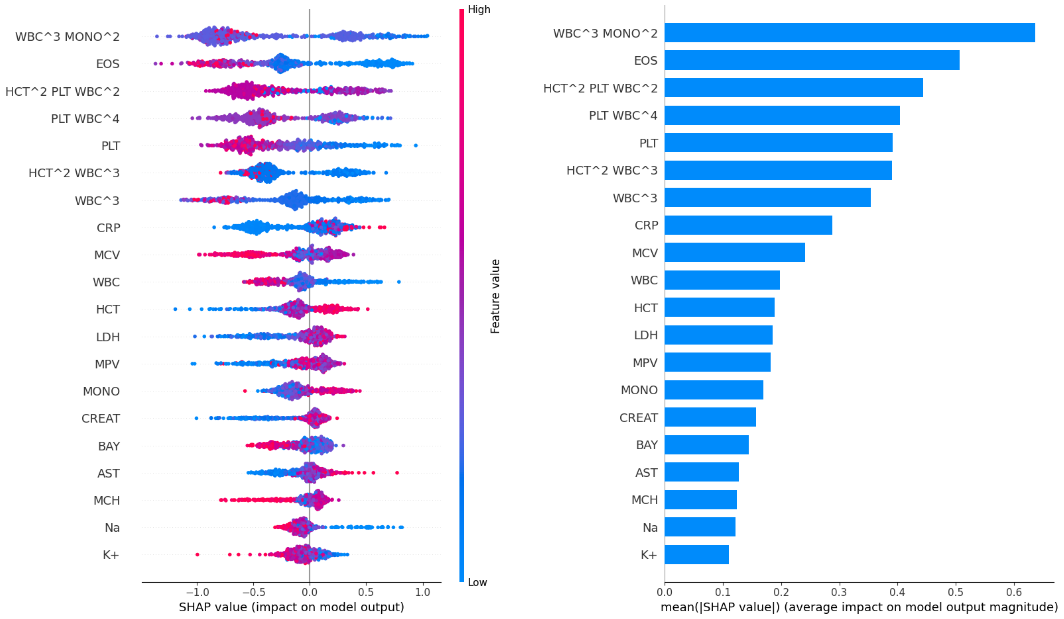
Random Forest:



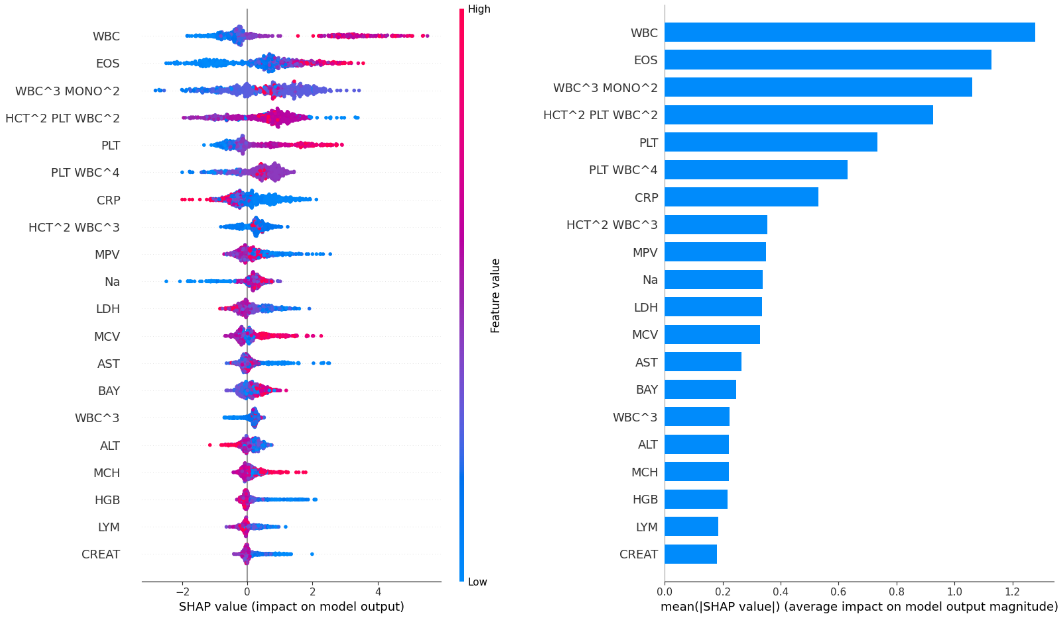
XGBoost:



CATBoost:



LGBM:



As you can see, the new features seem to be pretty high on the sorted features by feature importance for all models, notice that only 20 out of 28 features appear on the plots.
PLT and EOS also highlight features for all the models.

** The high missing rate feature - “Lactic Dehydrogenase” (LDH) appears to be in the middle of all plots, it is not very good for us, It was much more reliable for the authors to emit this feature for better results.*

Appendices

Missing Ratio after elimination of columns-

Hematocrit	0.822368
Hemoglobin	0.822368
Platelets	0.986842
Red blood Cells	0.986842
Lymphocytes	0.986842
Mean corpuscular hemoglobin (MCH)	0.986842
Mean corpuscular hemoglobin concentration (MCHC)	0.986842
Leukocytes	0.986842
Basophils	0.986842
Eosinophils	0.986842
Lactic Dehydrogenase	83.388158
Mean corpuscular volume (MCV)	0.986842
Red blood cell distribution width (RDW)	0.986842
Monocytes	1.151316
Mean platelet volume	1.480263
Neutrophils	15.625000
Proteina C reativa mg/dL	16.776316
Creatinine	30.263158
Urea	34.703947
Potassium	38.980263
Sodium	39.144737
Aspartate transaminase	62.828947
Alanine transaminase	62.993421

Full and accurate results of the experiments for each best model -

Results with F1-Score for Negative:

LR

accuracy: 0.827049180327869±0.030767744883808095

f1: 0.5301798617750586±0.06314256311679649

sensitivity: 0.7707456990893524±0.09738903336769619

specificity: 0.8361507563967763±0.03590981273385927

roc_auc: 0.8034482277430645±0.048753566453777444

RF

accuracy: 0.8909836065573769±0.03914738995494936

f1: 0.5479082565964609±0.1558471047399993

sensitivity: 0.53196951655156±0.19869120444365362

specificity: 0.9473352489492399±0.019298155245542545

roc_auc: 0.7396523827503999±0.10412711544377996

XGB

accuracy: 0.8918032786885247±0.03191462677693736

f1: 0.6139073020187658±0.10371063217510759

sensitivity: 0.6840758003606301±0.1394121639075899

specificity: 0.9250641429076548±0.023884681432260794

roc_auc: 0.8045699716341426±0.0746474578425453

CB

accuracy: 0.8844262295081966±0.030767744883808963

f1: 0.5855184722025499±0.09473605420038249

sensitivity: 0.643330271379807±0.12471770995879325

specificity: 0.9220581890487061±0.02460848657234504

roc_auc: 0.7826942302142564±0.06760678126296636

LGBM

accuracy: 0.8868852459016392±0.01861937162557469

f1: 0.5816676037601334±0.060887113132666836

sensitivity: 0.6309942843534174±0.11091975416625698

specificity: 0.9268219185866083±0.016936892764397946

roc_auc: 0.778908101470013±0.05300581228198788

Results with Macro F1-Score:

LR

accuracy: 0.827049180327869±0.030767744883808950

f1: 0.7117537393201088±0.0378981632604014

sensitivity: 0.7707456990893524±0.09738903336769619

specificity: 0.8361507563967763±0.03590981273385927

roc_auc: 0.8034482277430645±0.048753566453777444

RF

accuracy: 0.8909836065573769±0.03914738995494936

f1: 0.7428814051372312±0.08835957587392151

sensitivity: 0.53196951655156±0.19869120444365362

specificity: 0.9473352489492399±0.019298155245542545

roc_auc: 0.7396523827503999±0.10412711544377996

XGB

accuracy: 0.8918032786885247±0.03191462677693736

f1: 0.7754398032529416±0.06079608910007273

sensitivity: 0.6840758003606301±0.1394121639075899

specificity: 0.9250641429076548±0.023884681432260794

roc_auc: 0.8045699716341426±0.0746474578425453

CB

accuracy: 0.8844262295081966±0.030767744883808963

f1: 0.7590867241851356±0.05561368825762681

sensitivity: 0.643330271379807±0.12471770995879325

specificity: 0.9220581890487061±0.02460848657234504

roc_auc: 0.7826942302142564±0.06760678126296636

LGBM

accuracy: 0.8868852459016392±0.01861937162557469

f1: 0.7580347763318542±0.034139529664532174

sensitivity: 0.6309942843534174±0.11091975416625698

specificity: 0.9268219185866083±0.016936892764397946

roc_auc: 0.778908101470013±0.05300581228198788

Results with the new features:

LR

Accuracy: 0.8295081967213115±0.02786885245901639

F1: 0.7157859458259789±0.04226127802452814

Sensitivity: 0.7869111126231869±0.12727361196210954

Specificity: 0.8373737604194311±0.030918795571958094

AUROC: 0.8121424365213091±0.06288465464743931

RF

Accuracy: 0.8909836065573771±0.02540983606557377

F1: 0.7756742399511974±0.04683181819374808

Sensitivity: 0.6935015706687533±0.13079180233150747

Specificity: 0.9232784710588833±0.025132110128980544

AUROC: 0.8083900208638182±0.06478156161476578

XGB

Accuracy: 0.8844262295081966±0.028735324824521485

F1: 0.7698888558309195±0.051098263229783586

Sensitivity: 0.7123393891969744±0.13944546617553394

Specificity: 0.9127919241906188±0.022842972149545244

AUROC: 0.8125656566937967±0.07202494816825036

CB

Accuracy: 0.8819672131147541±0.0226561884607955

F1: 0.767322360511124±0.03961089202383436

Sensitivity: 0.7184797400741674±0.12660795518484347

Specificity: 0.9090616568141006±0.02314530323850899

AUROC: 0.8137706984441339±0.061240380992327544

LGBM

Accuracy: 0.8959016393442623±0.028406104018300755

F1: 0.7773987439137973±0.057855529472125966

Sensitivity: 0.6610756529332381±0.12392351046827829

Specificity: 0.9324980180759377±0.020484338319116294

AUROC: 0.796786835504588±0.06625439515440291