Machine Learning for COVID-19 Detection Based on Routine Blood Tests

Cabitza, F., Campagner, A., Ferrari, D., Di Resta, C., Ceriotti, D., Sabetta, E., ... & Carobene, A. (2021). Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests. Clinical Chemistry and Laboratory Medicine (CCLM), 59(2), 421-431.

第十組
M104020029邱承漢 M104020054謝博丞

Known Shortcomings of rRT-PCR

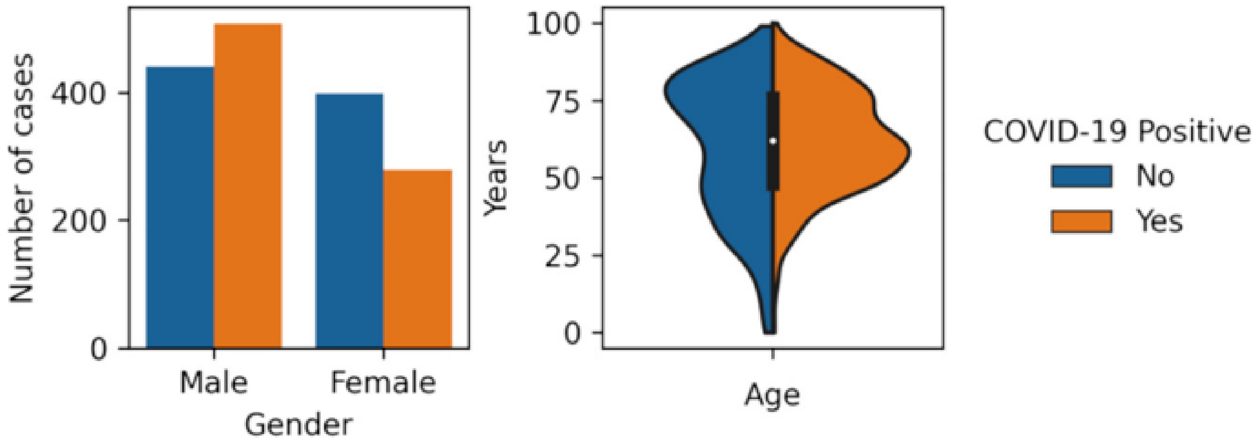
- Long turnaround time
- Potential shortage of reagents
- False-negative rates around 15–20%
- Expensive equipment

OSR dataset

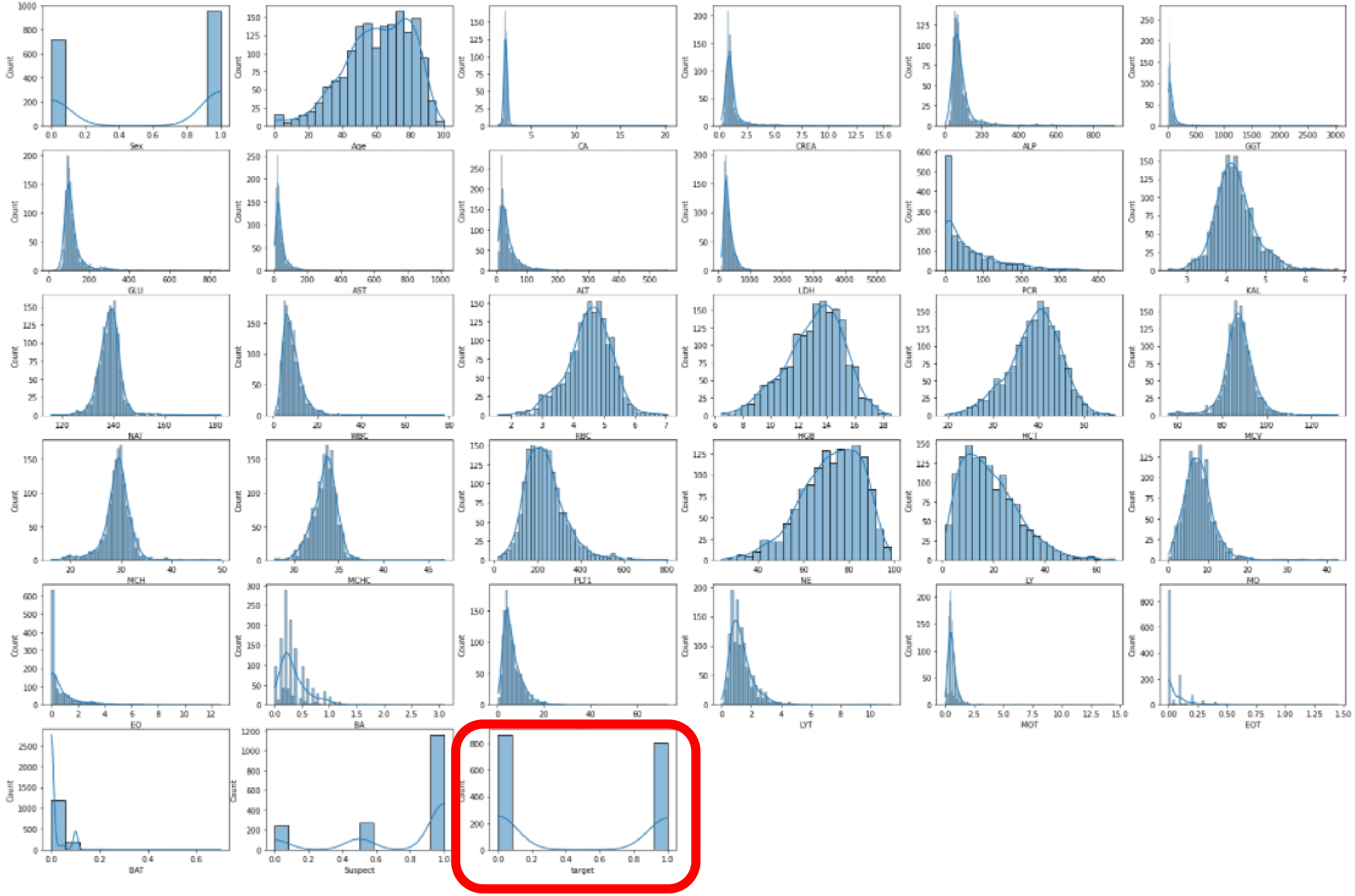
- Routine blood-test results performed on 1,737 patients
- 34 features columns
- 52% COVID-19 positive
- 48% COVID-19 Negative

OSR dataset

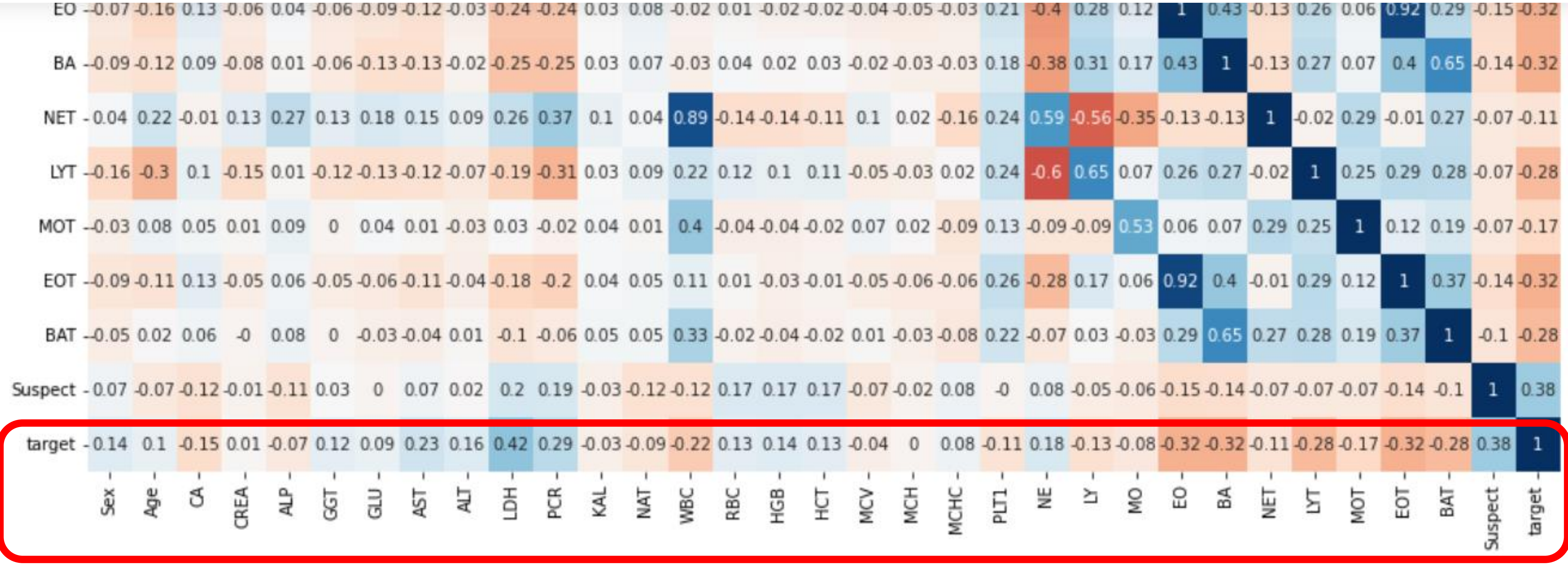
- Hematological
- Coagulation
- Biochemical
- Rapidpoint 500
- Additional information



Exploratory Data Analysis (EDA)



Exploratory Data Analysis (EDA)



Drop Column

Missing Rate	
CK	58.00
UREA	36.75
ALP	24.70
GGT	22.43
EOT	18.02
MO	18.02
EO	18.02
BA	18.02
NET	18.02
LYT	18.02
MOT	18.02
BAT	18.02
LY	18.02
NE	18.02
LDH	14.50

Drop Row

												...										
1530	PSMAY0116_2020-05-04	0	41.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0
1544	PSMAY0061_2020-05-03	1	54.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0
1547	PSMAY0015_2020-05-05	0	61.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	1
1549	PSMAY0027_2020-05-04	1	46.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	1
1554	PSMAY0092_2020-05-18	0	93.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0
1560	PSMAY0209_2020-05-30	0	66.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0
1565	PSMAY0164_2020-05-11	0	35.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0
1569	PSMAY0005_2020-05-05	0	46.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	1
1575	PSMAY0159_2020-05-22	1	60.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0
1598	PSMAY0202_2020-05-01	1	76.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0
1600	PSMAY0017_2020-05-04	1	51.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	1
1620	PSMAY0036_2020-05-19	1	47.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0
1688	6	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0
1711	29	0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0

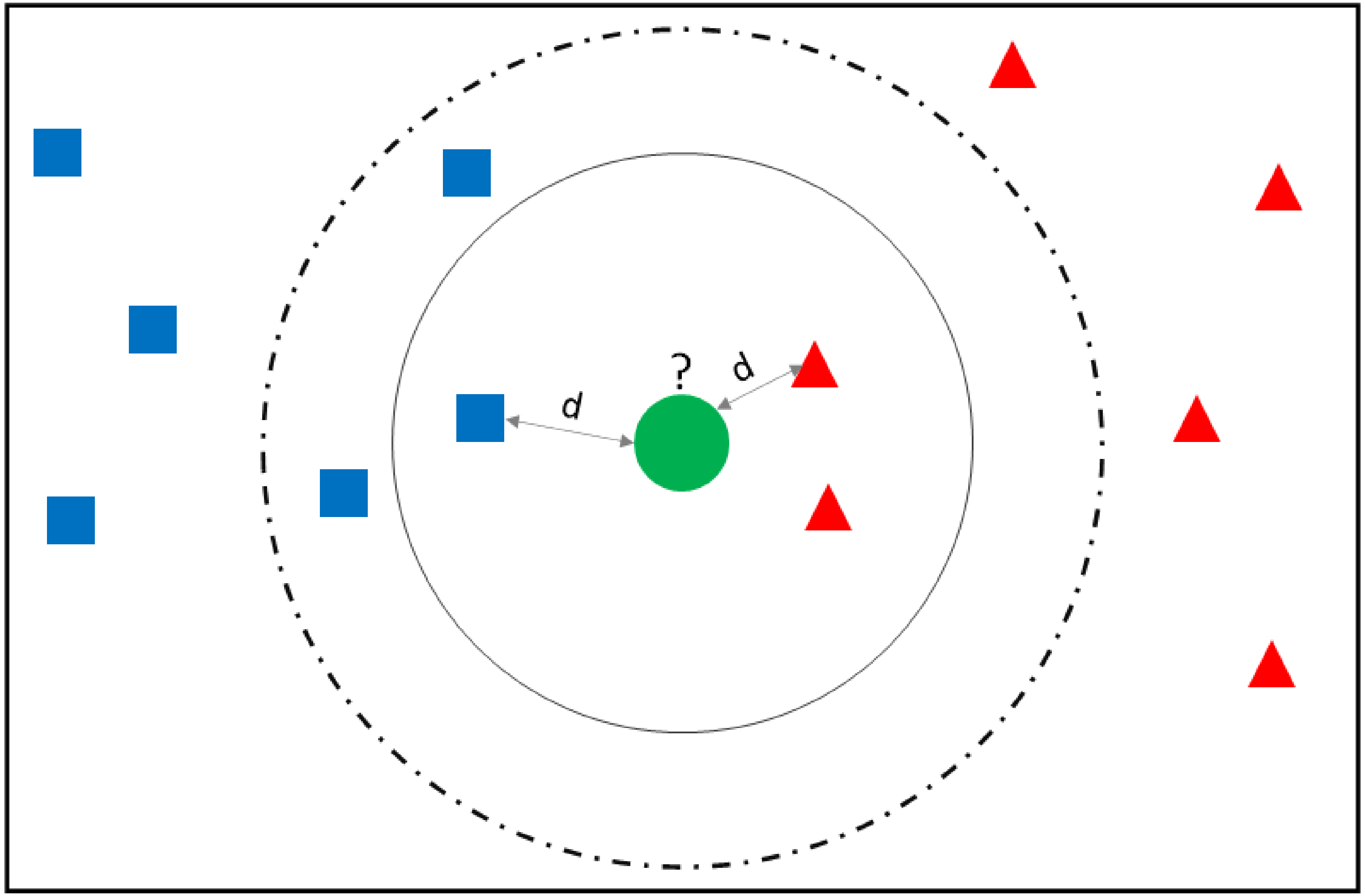
60 rows × 36 columns

Train-Test Split

```
raw data percentage :
0      51.73031
1      48.26969
Name: target, dtype: float64

train percentage :
0      51.716418
1      48.283582
Name: target, dtype: float64

test percentage :
0      51.785714
1      48.214286
Name: target, dtype: float64
```

KNN Imputation

K=5



$$X^* = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Min-Max Normalization

	count	mean	std	min	25%	50%	75%	max
Sex	1340.0	0.57	0.50	0.0	0.00	1.00	1.00	1.0
Age	1340.0	0.61	0.19	0.0	0.48	0.62	0.77	1.0
CA	1340.0	0.05	0.03	0.0	0.04	0.04	0.05	1.0
CREA	1340.0	0.06	0.06	0.0	0.04	0.05	0.06	1.0
ALP	1340.0	0.11	0.08	0.0	0.07	0.09	0.12	1.0
GGT	1340.0	0.05	0.08	0.0	0.02	0.03	0.06	1.0
GLU	1340.0	0.13	0.07	0.0	0.09	0.11	0.13	1.0
AST	1340.0	0.04	0.05	0.0	0.01	0.02	0.04	1.0
ALT	1340.0	0.06	0.08	0.0	0.02	0.04	0.07	1.0
LDH	1340.0	0.14	0.09	0.0	0.08	0.11	0.18	1.0
PCR	1340.0	0.15	0.18	0.0	0.01	0.09	0.23	1.0
KAL	1340.0	0.40	0.12	0.0	0.32	0.39	0.46	1.0
NAT	1340.0	0.34	0.07	0.0	0.30	0.35	0.38	1.0
WBC	1340.0	0.10	0.06	0.0	0.06	0.09	0.13	1.0
RBC	1340.0	0.54	0.13	0.0	0.46	0.55	0.62	1.0
HGB	1340.0	0.57	0.18	0.0	0.46	0.59	0.70	1.0
HCT	1340.0	0.56	0.16	0.0	0.47	0.58	0.67	1.0
MCV	1340.0	0.41	0.09	0.0	0.37	0.41	0.46	1.0

Model

- KNN
- Naive bayes
- Logistic regression
- SVM
- Random forest

```
clf_xgb = XGBClassifier(num_iterations=1000, subsample=0.5,sampling_method='uniform', object='binary:logistic',
                        val_metric='auc', eta=0.3, gamma=10, reg_alpha=0.3, reg_lambda=0.7,
```

Boosting

- AdaBoost
- GradientBoost
- **XGBoost**

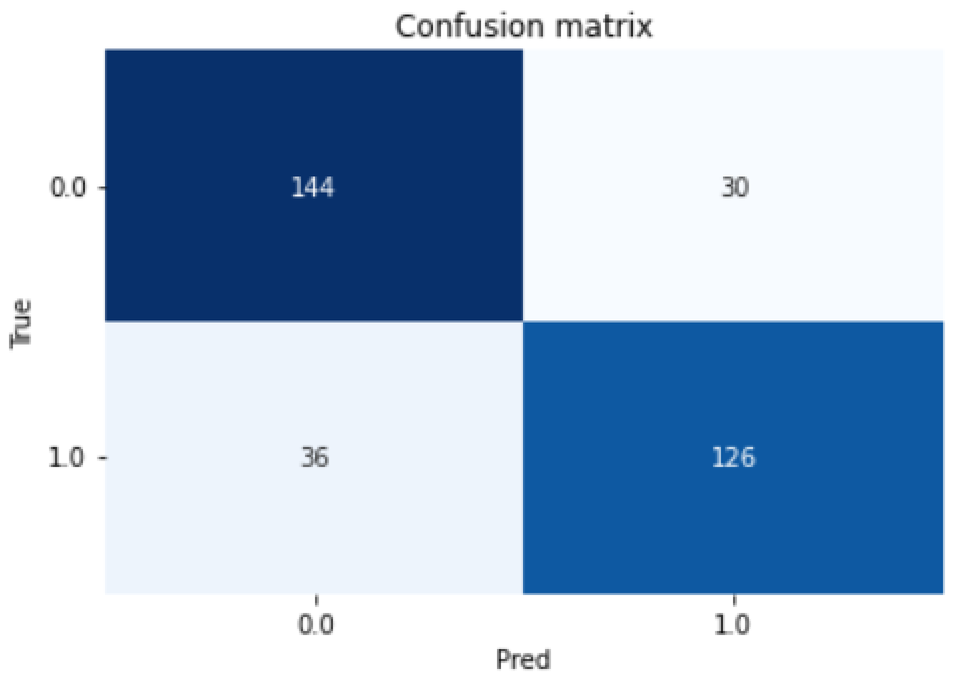
Pros:

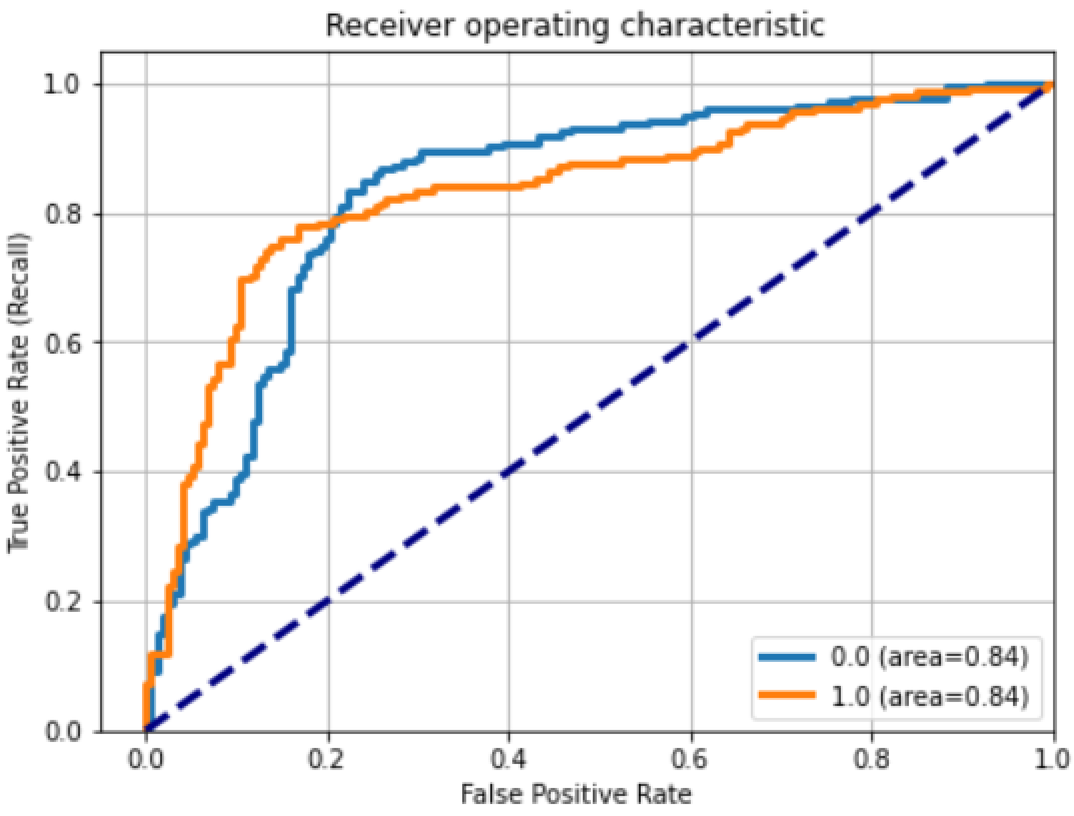
- Continuous Error Correction

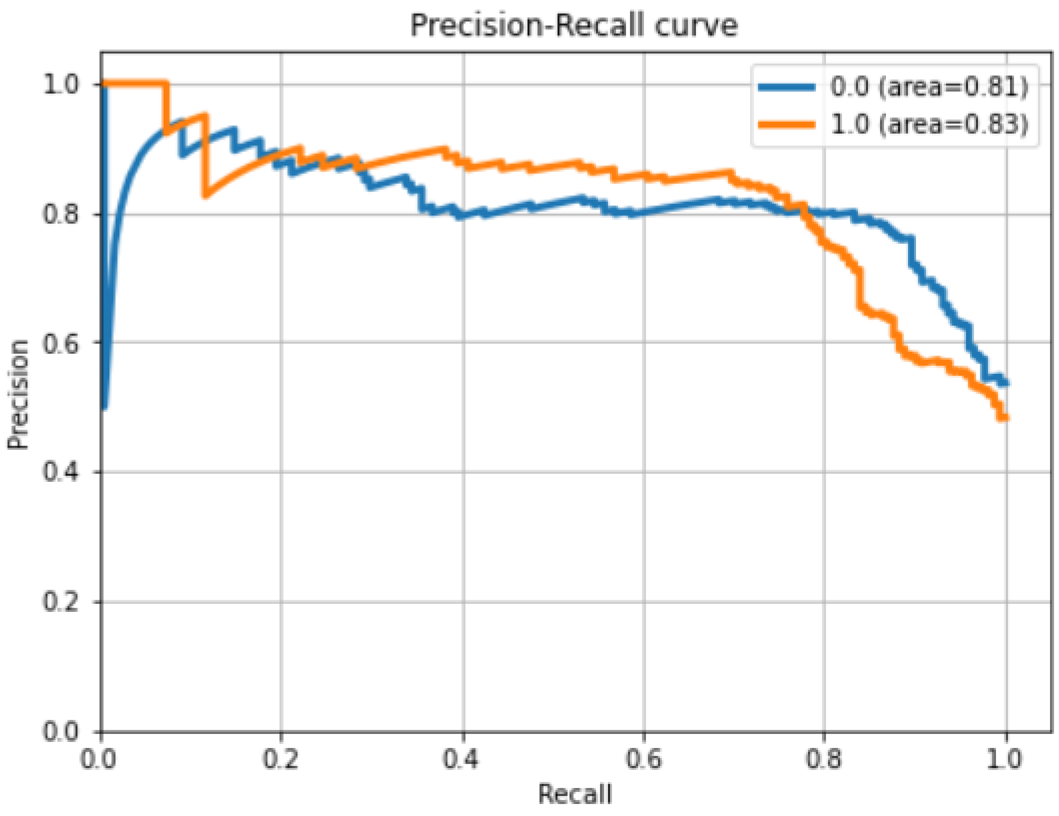
Cons:

- Tree algorithms can over-fit the data

```
model type: XGBoost
time costing: 1.3852753639221191
Accuracy: 0.8
Auc: 0.84
Detail:
              precision    recall  f1-score   support

         0.0       0.80      0.83      0.81       174
         1.0       0.81      0.78      0.79       162

    accuracy                           0.80       336
   macro avg       0.80      0.80      0.80       336
weighted avg       0.80      0.80      0.80       336
```




Receiver operating characteristic
True Positive Rate (Recall)
False Positive Rate
0.0 (area=0.84)
1.0 (area=0.84)
0.0
0.2
0.4
0.6
0.8
1.0


Precision-Recall curve
Precision
Recall
0.0 (area=0.81)
1.0 (area=0.83)
0.0
0.2
0.4
0.6
0.8
1.0

```
clf_mlp = MLPClassifier(hidden_layer_sizes=(128,), activation='relu', solver='adam',
                        alpha=1e-4, max_iter=10000,
```

Neural Network

- **Multi-layer Perceptron (MLP)**

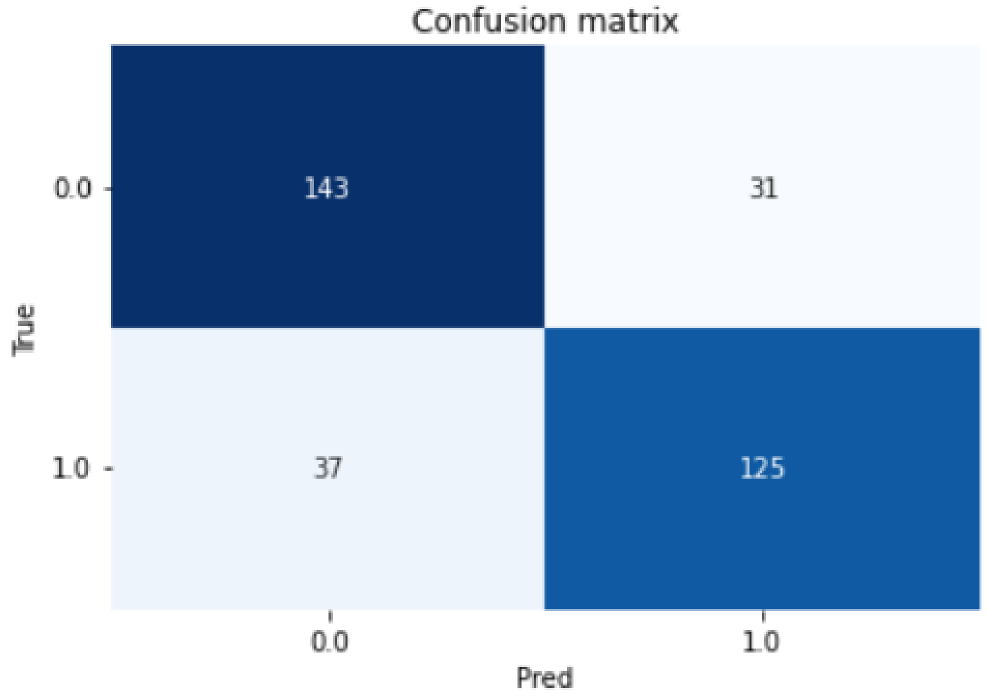
Pros:

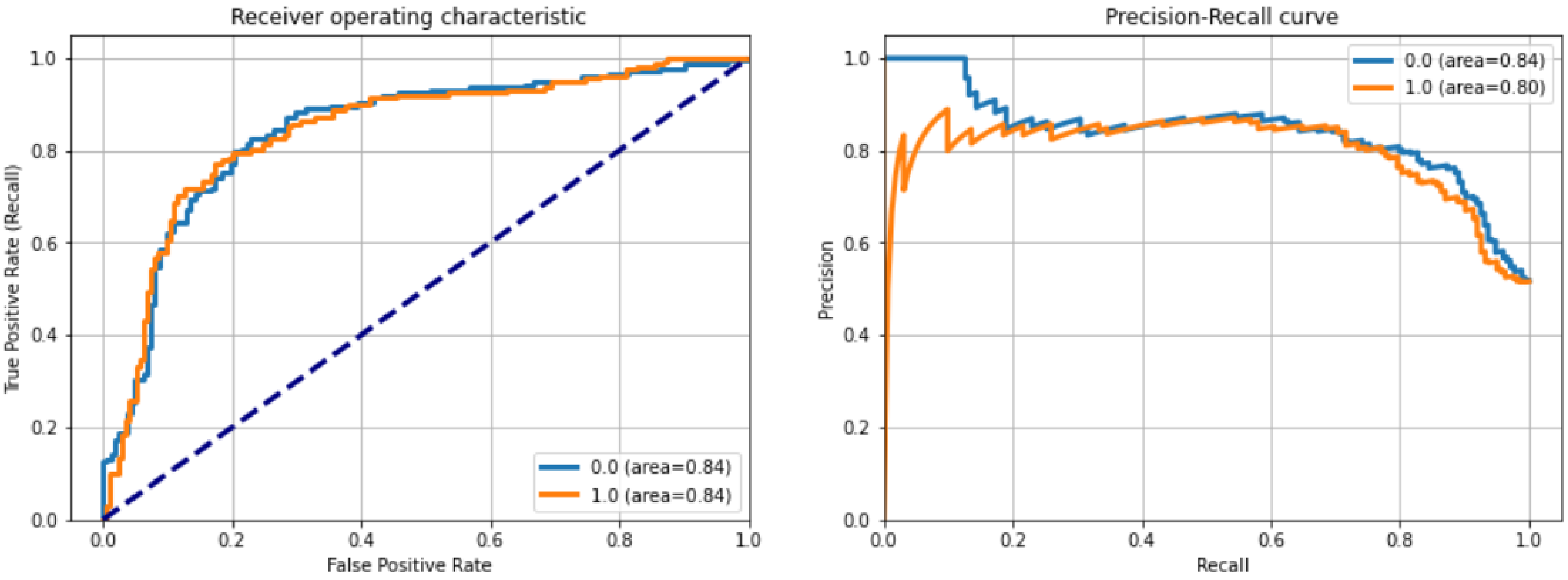
- High Accuracy

Cons:

- Difficult to explain
- High Computaion

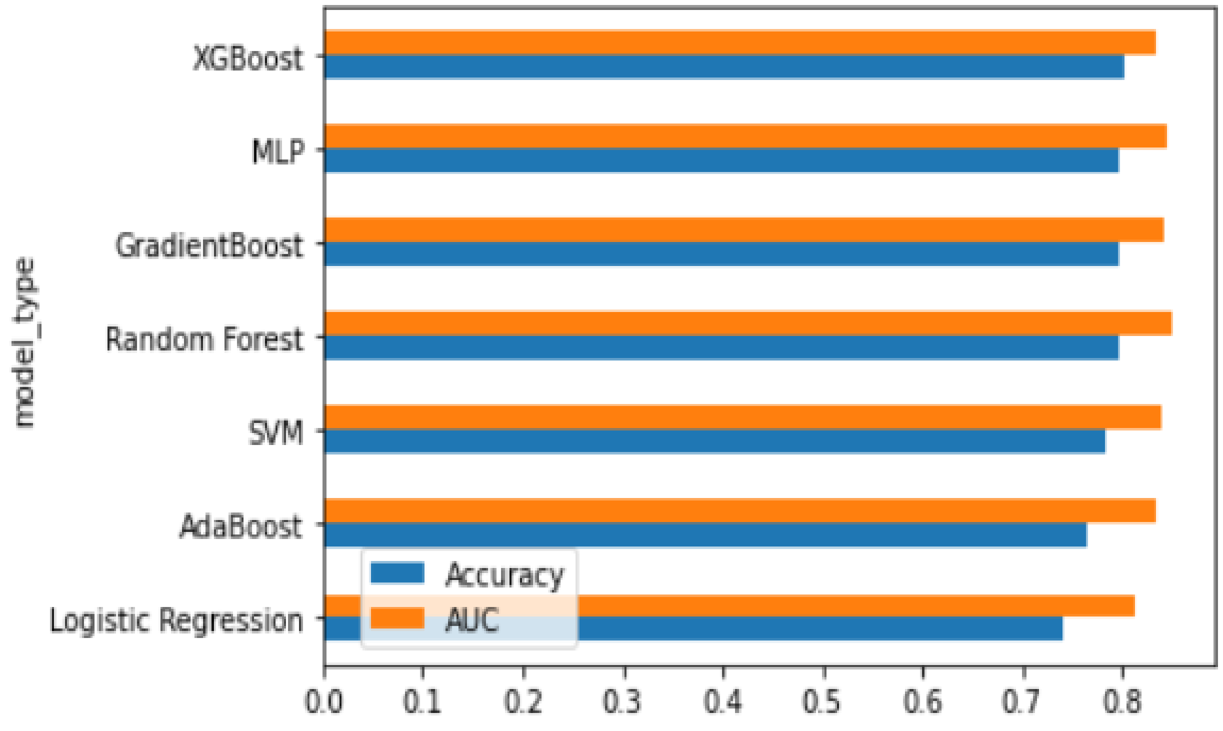
```
Accuracy: 0.8
Auc: 0.84
Detail:
              precision    recall  f1-score   support

         0.0       0.79      0.82      0.81       174
         1.0       0.80      0.77      0.79       162

    accuracy                           0.80       336
   macro avg       0.80      0.80      0.80       336
weighted avg       0.80      0.80      0.80       336
```

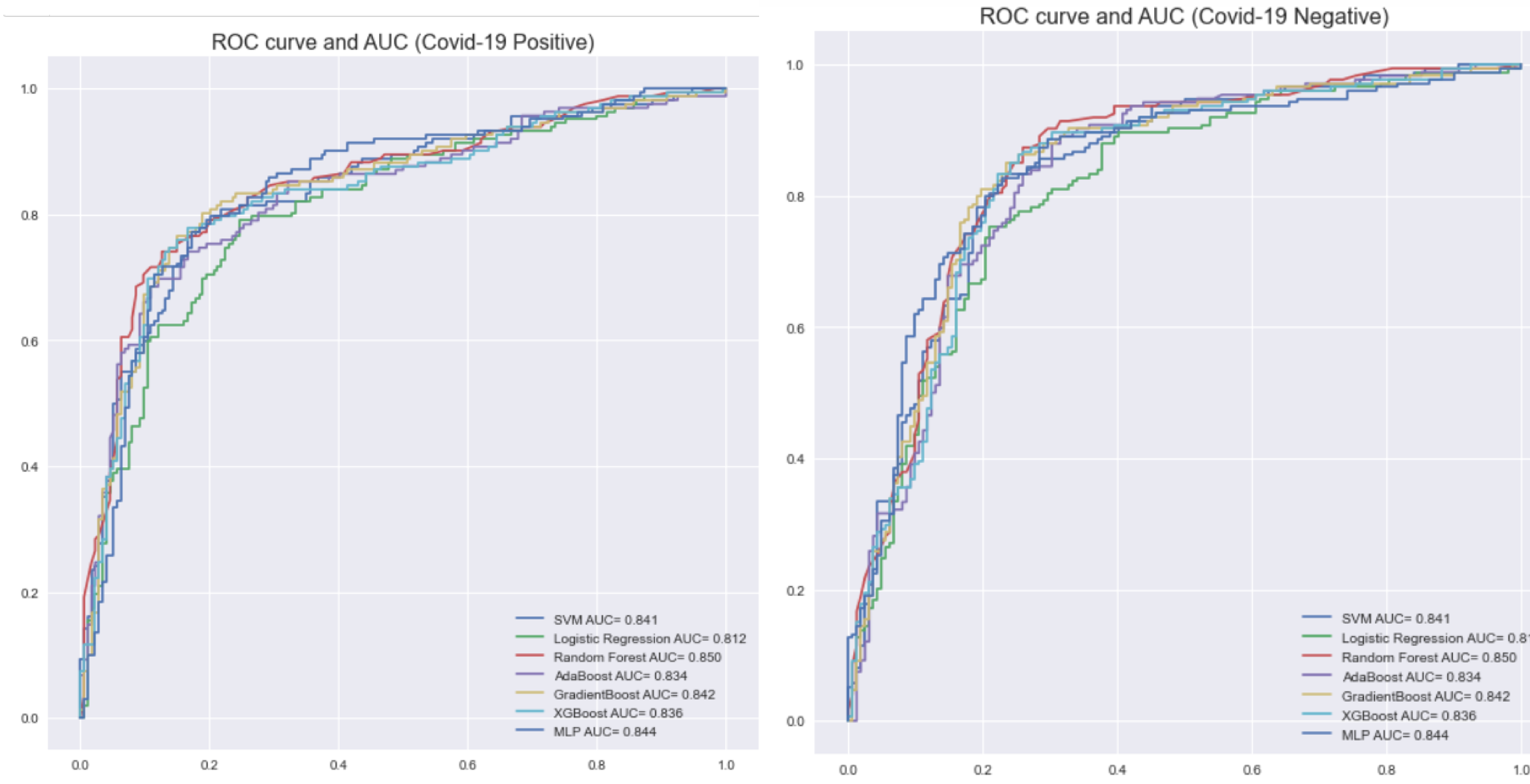



Receiver operating characteristic
True Positive Rate (Recall)
False Positive Rate
0.0 (area=0.84)
1.0 (area=0.84)
Precision-Recall curve
Precision
Recall
0.0 (area=0.84)
1.0 (area=0.80)

Model Comparison

	model_type	Accuracy	AUC
0	XGBoost	0.803571	0.835604
0	Random Forest	0.797619	0.849936
0	GradientBoost	0.797619	0.841812
0	MLP	0.797619	0.84426
0	SVM	0.782738	0.840925
0	AdaBoost	0.764881	0.834007
0	Logistic Regression	0.741071	0.812012




ROC curve and AUC (Covid-19 Positive)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
SVM AUC= 0.841
Logistic Regression AUC= 0.812
Random Forest AUC= 0.850
AdaBoost AUC= 0.834
GradientBoost AUC= 0.842
XGBoost AUC= 0.836
MLP AUC= 0.844
ROC curve and AUC (Covid-19 Negative)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
SVM AUC= 0.841
Logistic Regression AUC= 0.8
Random Forest AUC= 0.850
AdaBoost AUC= 0.834
GradientBoost AUC= 0.842
XGBoost AUC= 0.836
MLP AUC= 0.844

(RFE) Recursive Feature-Elimination

n_features_to_select=20

```
model type: XGBoost
   Features  importances
29      EOT     0.455840
9       LDH     0.310221
13      WBC     0.169918
2        CA     0.024573
15      HGB     0.020856
8       ALT     0.018592
```

(PCA) Principal Component Analysis

n_components=20

Pros:

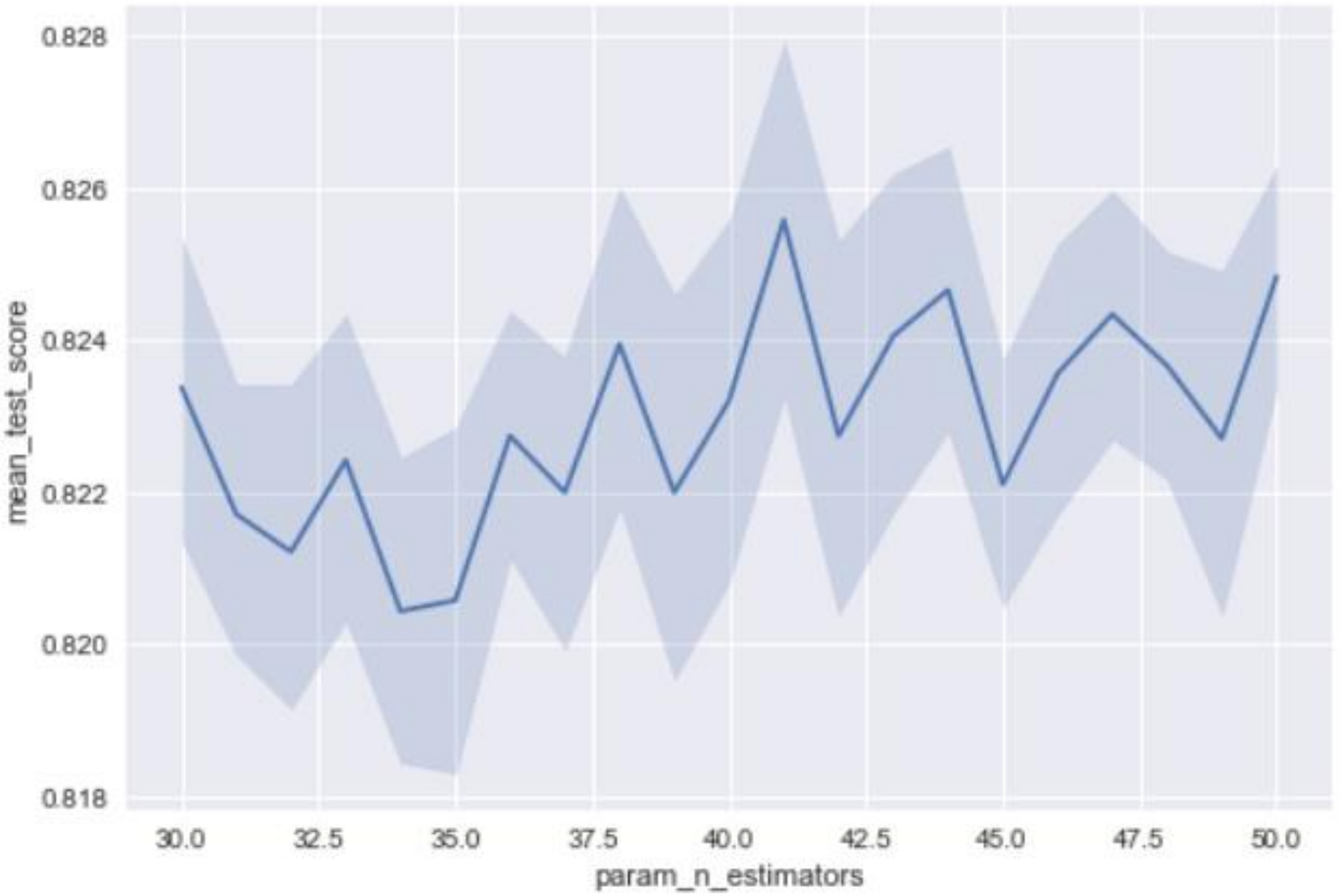
- High versatility

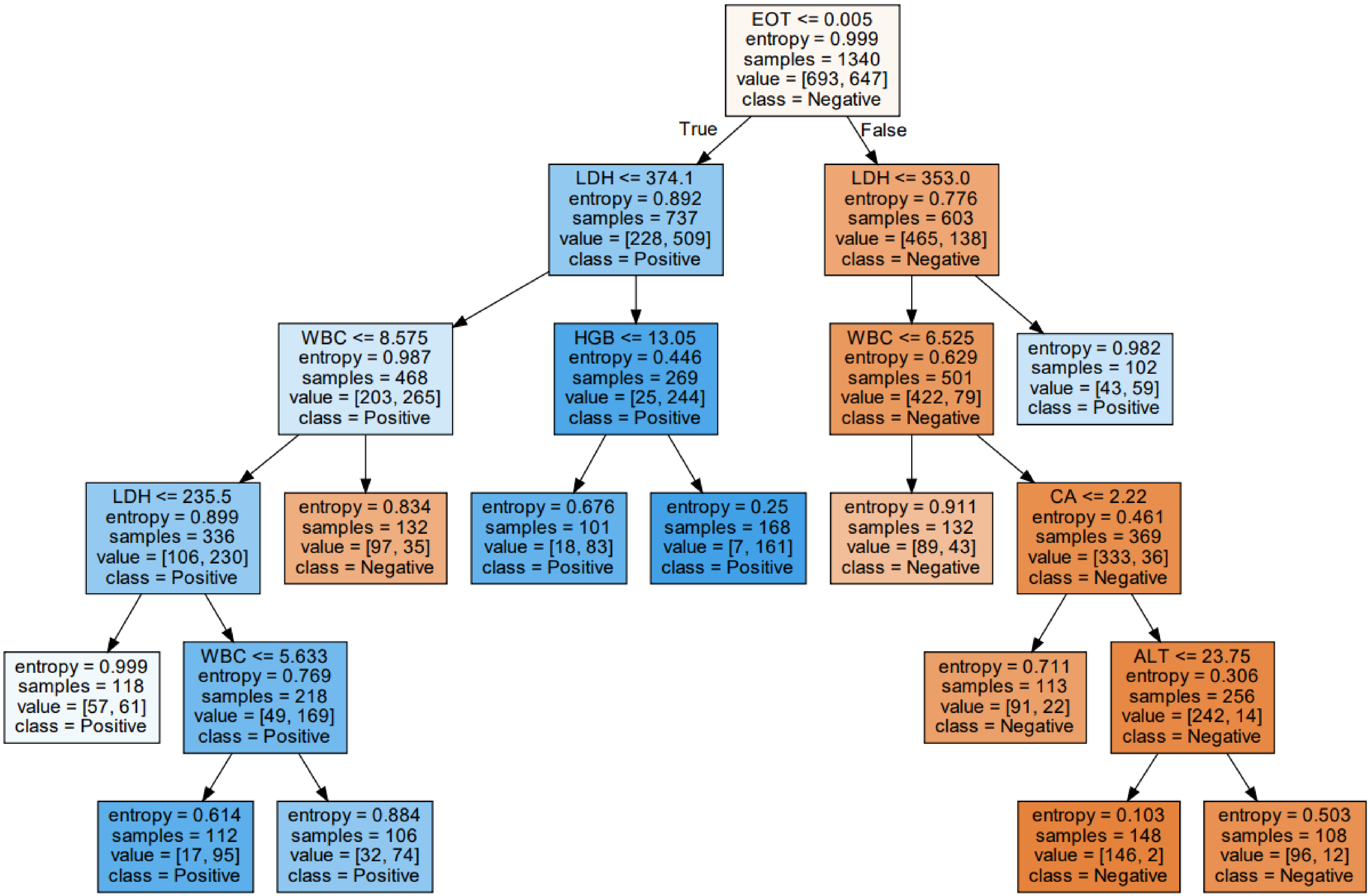
Cons:

- Difficult to explain

Grid Search

```
{'max_depth': 25, 'n_estimators': 39}
0.835820895522381

<AxesSubplot:xlabel='param_n_estimators', ylabel='mean_test_score'>
```





EOT <= 0.005
entropy = 0.999
samples = 1340
value = [693, 647]
class = Negative
True
False
LDH <= 374.1
entropy = 0.892
samples = 737
value = [228, 509]
class = Positive
LDH <= 353.0
entropy = 0.776
samples = 603
value = [465, 138]
class = Negative
WBC <= 8.575
entropy = 0.987
samples = 468
value = [203, 265]
class = Positive
HGB <= 13.05
entropy = 0.446
samples = 269
value = [25, 244]
class = Positive
WBC <= 6.525
entropy = 0.629
samples = 501
value = [422, 79]
class = Negative
entropy = 0.982
samples = 102
value = [43, 59]
class = Positive
LDH <= 235.5
entropy = 0.899
samples = 336
value = [106, 230]
class = Positive
entropy = 0.834
samples = 132
value = [97, 35]
class = Negative
entropy = 0.676
samples = 101
value = [18, 83]
class = Positive
entropy = 0.25
samples = 168
value = [7, 161]
class = Positive
entropy = 0.911
samples = 132
value = [89, 43]
class = Negative
CA <= 2.22
entropy = 0.461
samples = 369
value = [333, 36]
class = Negative
entropy = 0.999
samples = 118
value = [57, 61]
class = Positive
WBC <= 5.633
entropy = 0.769
samples = 218
value = [49, 169]
class = Positive
entropy = 0.711
samples = 113
value = [91, 22]
class = Negative
ALT <= 23.75
entropy = 0.306
samples = 256
value = [242, 14]
class = Negative
entropy = 0.614
samples = 112
value = [17, 95]
class = Positive
entropy = 0.884
samples = 106
value = [32, 74]
class = Positive
entropy = 0.103
samples = 148
value = [146, 2]
class = Negative
entropy = 0.503
samples = 108
value = [96, 12]
class = Negative

Reference

- https://www.degruyter.com/document/doi/10.1515/cclm-2020-1294/html
- https://zenodo.org/record/4081318#.YkwDQS1BxPa
- https://scikit-learn.org/stable/index.html