Improvements for an interpretable mortality prediction model for COVID-19 patients

Hubert Ruczyński, Dawid Przybyliński, Kinga Ułasik

April 7, 2021

1 Introduction

The main purpose for our study is to analyse and look critically at the paper: An interpretable mortality prediction model for COVID-19 patients. As data analysts we are aware of many flaws of this article and we wanted to show better solutions for the faced problem of predicting death of Covid patients.

2 Discussion about original article

2.1 Replies and other articles

To outline the weak points of mentioned article we're gonna analyse some of the responses to it.

External validation demonstrates limited clinical utility of the interpretable mortality prediction model for patients with COVID-19: The main focus of this article is placed on the poor death prediction which is around 50%. The authors says that the model wasn't tested on any external data and because of this fact it has worse accuracy than expected. Moreover they decided to test it on their own dataset with more records and that confirmed their assumptions. St the end they reached a conclusion that the prediction of proposed model isn't good enough to be used for medical purposes.

Replication of a mortality prediction model in Dutch patients with COVID-19: Authors of this short article also focus on poor death prediction. The problem of low death precision is very visible here, because only 36% of the Dutch patients which were predicted to die actually died. It shows huge inaccuracy of the mentioned model and gave us many thoughts about what should be improved. Moreover authors noticed an important issue of different genetics of Dutch and Chinese people and the difference in admission rules.

Limited applicability of a COVID-19 specific mortality prediction rule to the intensive care setting: The main focus here is put on the target value. The authors are questioning if death is the best and only predicted state that we should consider. They say that we should rebuild models in the ICUs to predict not only death, but also worsening of health or occurrences of severe or critical types of Covid.

Developing a COVID-19 mortality risk prediction model when individual-level data are not available: Authors of this article created a more advanced model which used the knowledge about virus which we gained throughout the pandemic. Analysis of their ideas helped us with improving our models by outlining which variables are actually important to predict patients' death.

2.2 Summary

These articles inspired us and turned our thoughts to focus on a few issues. First of all we decided to increase the precision of death prediction, because that was the main flaw mentioned in this article. Moreover we decided to get more data and build our models with it, which will let us have more accurate environment. We've also used the ideas from the latest article to select the most important variables, which weren't proposed by the authors.

3 Better model for original data using the same parameters

3.1 Analyzing the original model

We posed ourselves a question whether there was a possibility of creating a better model using only original data? In order to answer this question, firstly we recreated the decision tree from the article using $plot_tree$ command.

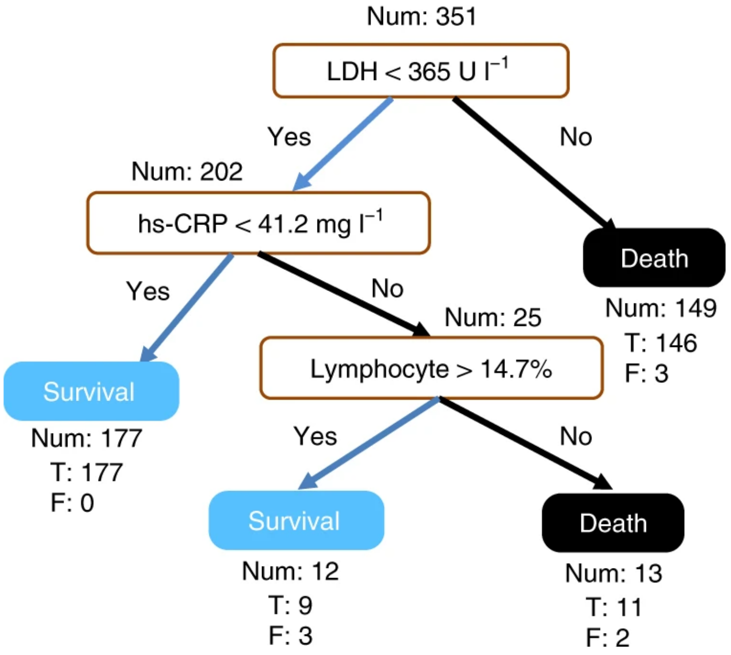


Figure 1: Original decision tree

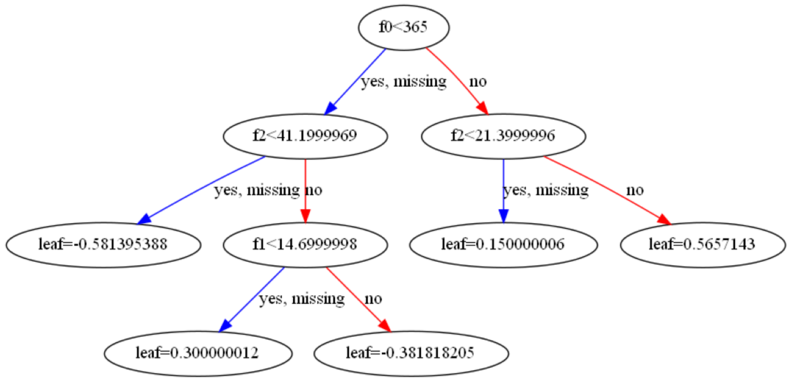


Figure 2: Our recreation of the decision tree

Next we tested both models on the new data to compare them. For original model from the article we received:

	precision	recall	F1-score	support
0	0.94	0.39	0.55	678
1	0.45	0.96	0.61	360
accuracy			0.58	1038
macro avg	0.70	0.67	0.58	1038
weighted avg	0.77	0.58	0.57	1038

We can observe that the precision for the 0 class (survival) is very high, so the proportion between people classified to 0 and the people that actually survived is high enough, but the precision of the 1 class (death) is low, which means that only 0.45 patients classified as 'dead' really died. This can lead to unnecessary burdening the hospitals and more deaths. Additionally, the accuracy is low, which means that general correctness of the model (of no account of the class) isn't high.

3.2 Our model

We created a new model using $AdaBoostClassifier$ which uses the same parameters as the original model. We trained and checked how it performs on a new data:

	precision	recall	F1-score	support
0	0.97	0.37	0.53	678
1	0.45	0.98	0.62	360
accuracy			0.58	1038
macro avg	0.71	0.67	0.58	1038
weighted avg	0.79	0.58	0.56	1038

We can notice that, the the model doesn't really perform better than the original one, some of the parameters are better, but most of them are the same or even worse. But then we discovered something interesting.

3.3 Data distribution analysis

We analyzed the distribution of percentage of Lymphocytes, Lactate dehydrogenase and High sensitivity C-reactive protein by creating histograms for the original and the new data.

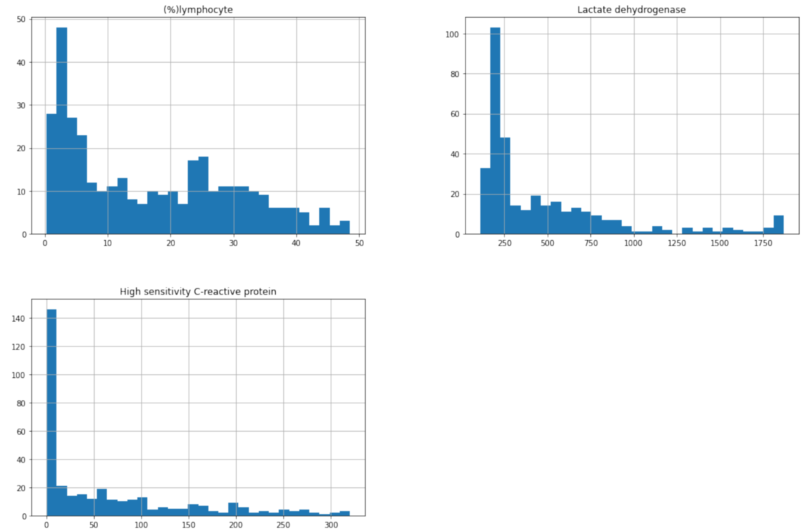


Figure 3: Distribution of the original data

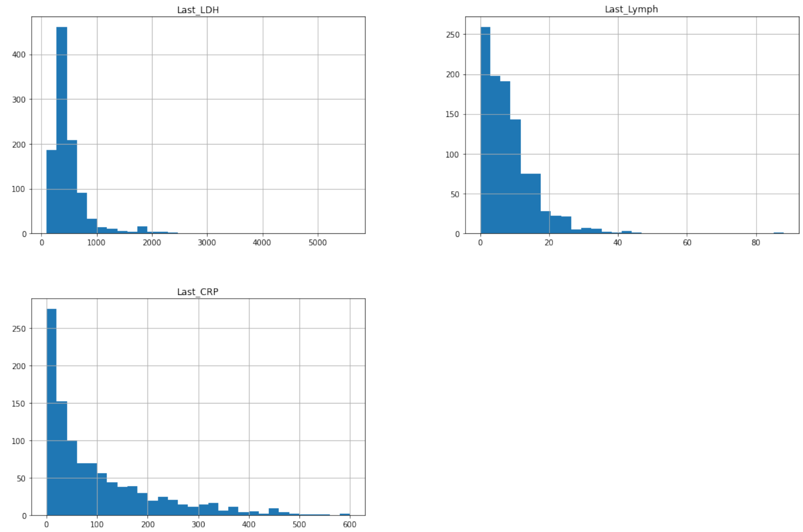


Figure 4: Distribution of the original data

We noticed that all variables are strongly left skewed which is unfavorable for the model because more reliable predictions are made if the predictors and the target variable are normally distributed. Trying to make our model better, we applied square root transformation. Then we trained a new model on them (also using Ada Boost Classifier) and we tested it on the new data:

	precision	recall	F1-score	support
0	0.95	0.42	0.58	678
1	0.47	0.96	0.63	360
accuracy			0.61	1038
macro avg	0.71	0.69	0.61	1038
weighted avg	0.78	0.61	0.60	1038

We can see that the model generally improved (the accuracy is higher) and it performs slightly better.

4 Better model for original data

4.1 Why original model has many flaws?

As it is stated in Answer Article 1 the original model tested on external data has very low level of accurate death prediction (around 50 percent). Even though it tends to learn whilst getting more data that's not good enough for medical use, where prediction models should have much higher standards. Moreover we are also concerned about the low precision score in deaths on original data which counts at the level of 0.81.

Because of these reasons we decided to create new model which won't have such flaws.

4.2 Improved Feature Selection

To create better model we decided to read through more current Covid-19 articles and see which variables are considered important for death prediction now. After all we gained some knowledge about it throughout pandemic.

Inspired by this article, we decided to take a closer look at the following variables: age, C reactive protein, chloride, albumin, lymphocyte count and LDH. Because the data had some missing values we performed an imputation method on them. We also created a correlation heatmap (Figure 5) which shows

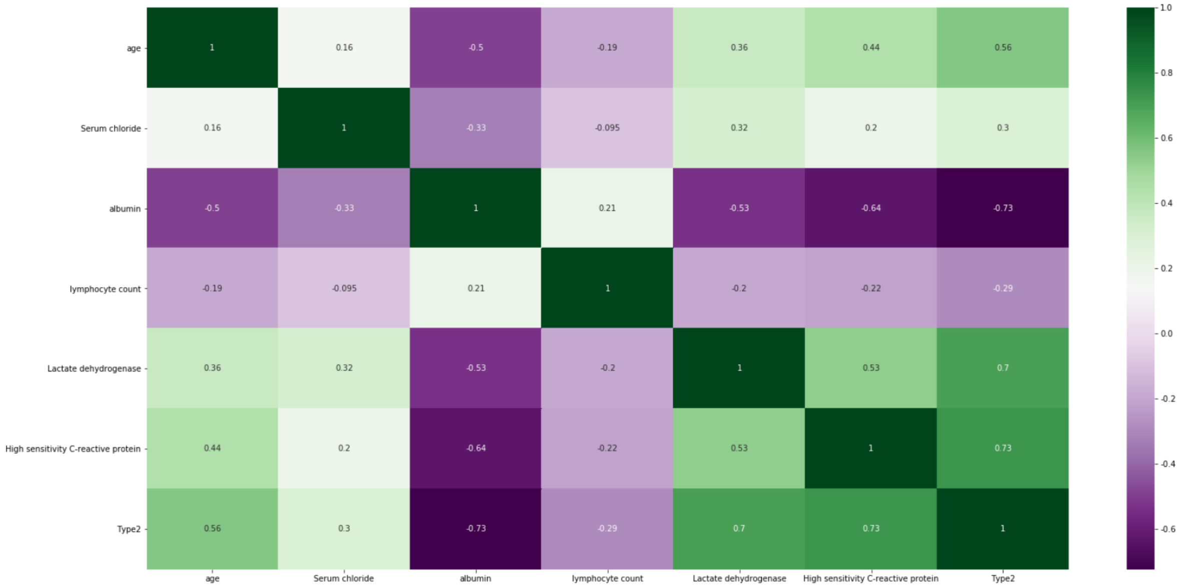
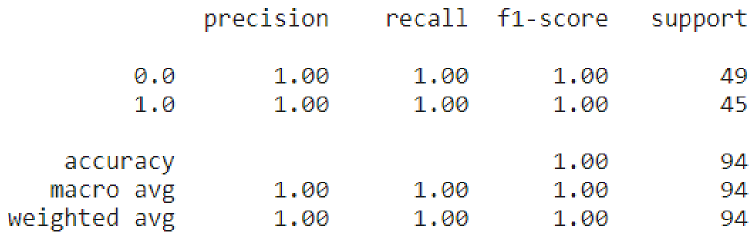


Figure 5: Important data correlation heatmap

From the above analysis we distinguished that the most important features there are age, albumin, LDH and C protein, which we used in our models.

4.3 New models

Unfortunately we had to train our models on smaller datasets cause the variables which we chose were only in original training set. That's why we had to divide it into smaller training and testing sets. After that we created Two models: GradientBoostingClassifier (Confusion Matrix: Figure 6) and AdaBoostClassifier (Confusion Matrix: Figure 7) which gave us following results from classification report and Crossvalidation:



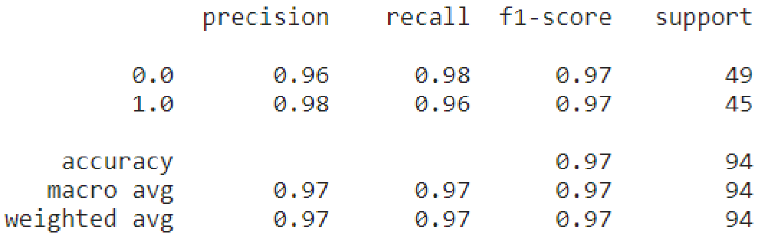
```
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00        49
         1.0       1.00      1.00      1.00        45

    accuracy                           1.00        94
   macro avg       1.00      1.00      1.00        94
weighted avg       1.00      1.00      1.00        94
```

Figure 6: Confusion Matrix for GradientBoosting

GradientBoostingClassifier Crossvalidation precision: 0.9791666666666667
AdaBoostClassifier Crossvalidation precision: 0.9583333333333334



```
              precision    recall  f1-score   support

         0.0       0.96      0.98      0.97        49
         1.0       0.98      0.96      0.97        45

    accuracy                           0.97        94
   macro avg       0.97      0.97      0.97        94
weighted avg       0.97      0.97      0.97        94
```

Figure 7: Confusion Matrix for AdaBoosting

4.4 Summary

Eventually we've managed to create better model with GradientBoostingClassifier than the original one. It's better because the precision and recall are more balanced in data set selected by our team and crossvalidation accuracy score is bigger than the accuracy in original model (around 0.97). We need to keep in mind that in the original example we didn't have the crossvalidation accuracy but from a single example only and as we can see on Figure 4, in single example one tends to have a better effect than with crossvalidation.

5 Principal Component Analysis

In order to analyse data further, we performed Principal Component Analysis. For visualization's simplification we considered only first two, most substantial components. Obtained two-dimensional plot is shown on Figure 8. Explained variance ratios were: 0.226 (for the first component) and 0.063 for the second component), which results in total explained variance ratio of 0.289 for both components together. Taking absolute values of the scores of each feature from the first component might be also used for feature selection. Below are those with highest magnitudes:

feature	PC-1 score
Fibrin degradation products	0.165544
eGFR	-0.166516
D-D dimer	0.170461
Platelet count	-0.172294
(%)lymphocyte	-0.174235
neutrophils(%)	0.176791
Urea	0.181715
albumin	-0.182415
Lactate dehydrogenase	0.188817
Prothrombin activity	-0.200919

There are features that we have already known are important, such as 'albumin', '(%)lymphocyte' or 'Lactate dehydrogenase', but 'age' was around the middle, not among the top ones.

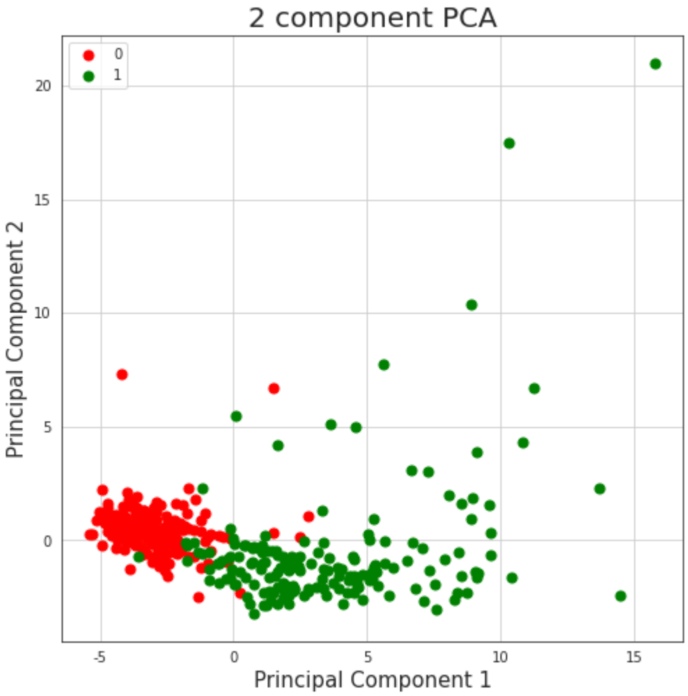


Figure 8: first 2 principal components on scatter-plot

The most noticeable fact is that our two classes are almost separable with just a single line. Even without any sort of complex machine learning or other algorithms, it's possible and not complicated to fit a line that divides cases ended in death from cases followed by patient's recovery. As an example, same visualization with additional function $y = 2.5x+2.5$, created without any sort of optimization techniques, is presented on Figure 9.

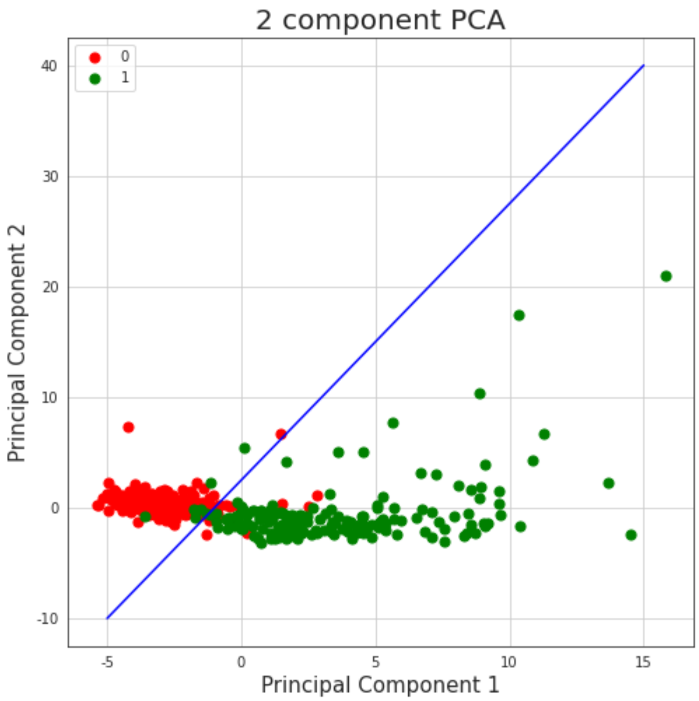


Figure 9: first principal components with simple linear function

Such division achieves (train) accuracy of 0.94, which is almost as good as results received by machine learning algorithms described in the paper, what might encourage to consider given data not authoritative.

	precision	recall	F1-score	support
0	0.96	0.94	0.95	201
1	0.93	0.95	0.94	174
accuracy			0.94	375
macro avg	0.94	0.94	0.94	375
weighted avg	0.94	0.94	0.94	375

6 New data

Thanks to the authors of the paper External validation demonstrates limited clinical utility of the interpretable mortality prediction model for patients with COVID-19, we were given access to additional data with features coincident with those we already had. Dataset contained over 1000 observations and only fourteen features that were selected by owners to fit as best model as possible. Taking it into account we decided to choose two approaches:

- To train the model on original data and test it with the new dataset
- To train and test model using the new dataset

6.1 Models trained with old data

We used the same model as before (AdaBoost Classifier) with features appearing in both datasets. Results were not satisfying:

	precision	recall	F1-score	support
0	0.97	0.37	0.53	678
1	0.45	0.98	0.62	360
accuracy			0.58	1038
macro avg	0.71	0.67	0.58	1038
weighted avg	0.79	0.58	0.56	1038

While analysing each feature's distribution, we noticed that almost all of them are heavily skewed left. We embraced it by transforming data by square root function to make feature's distribution more similar to uniform. Such operation slightly improved the performance, but it is still far from any desired outcome:

	precision	recall	F1-score	support
0	0.95	0.42	0.58	678
1	0.47	0.96	0.63	360
accuracy			0.61	1038
macro avg	0.71	0.69	0.61	1038
weighted avg	0.78	0.61	0.60	1038

6.2 Models trained with new data

Using AdaBoostClassifier we trained model. Some of the features contained entries in a format: '2 days 11:12:59', for simplification we replaced them with only 'days' value. Following results on the test set were received:

	precision	recall	F1-score	support
0	0.86	0.95	0.90	175
1	0.87	0.68	0.76	85
accuracy			0.86	260
macro avg	0.86	0.82	0.83	260
weighted avg	0.86	0.86	0.86	260

Moreover, we performed 20-fold cross-validation for AdaBoost with resulting score of 0.843.

Similarly to the original dataset, column extracted from the new data, are skewed left (Figure 8). However substituting values with their square roots (negative square roots when values were negative all along) did not improve prediction's quality.

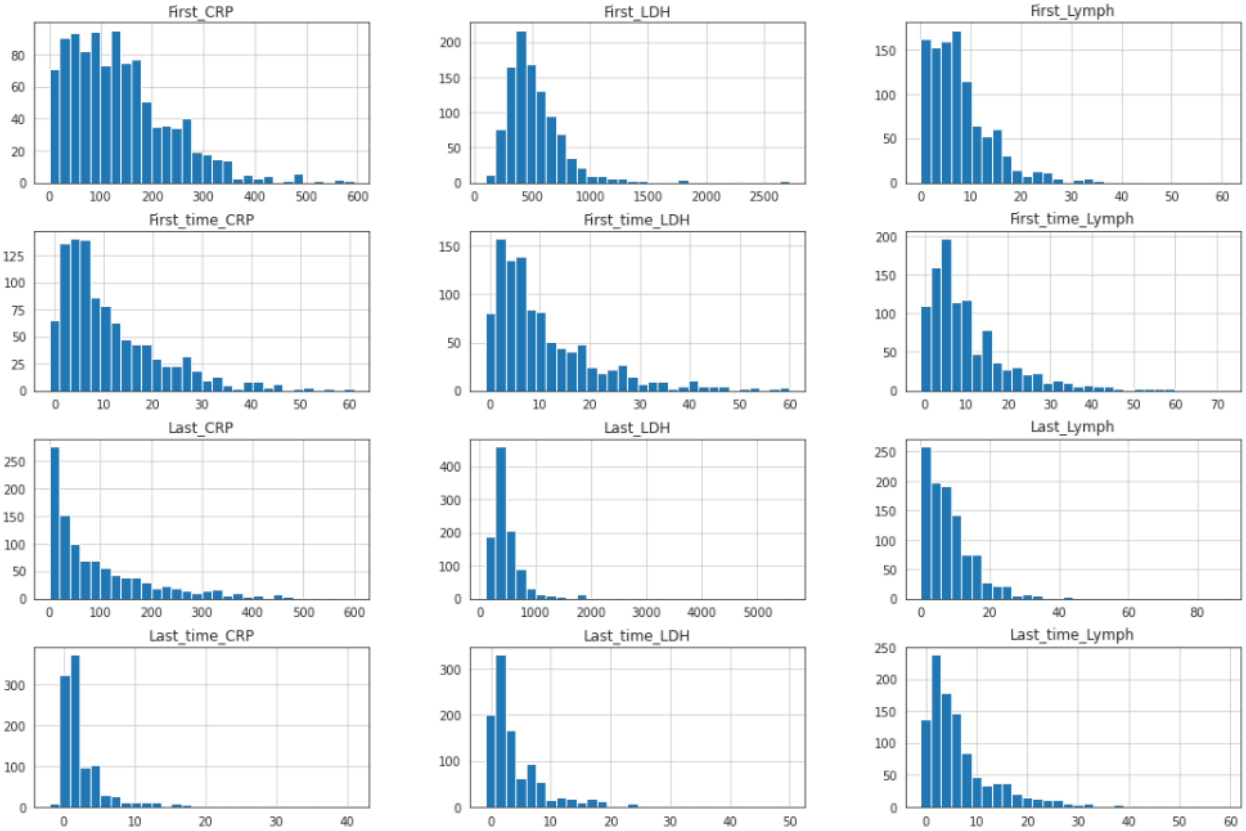


Figure 10: New dataset's features' histograms

7 Conclusions

At the end of this article we want to summarize that we've proposed three different ways and models to predict whether the patients with Covid-19 die or not. First one was achieved by usage of Ada Boost Classifier and unskewing the data on the new data that we achieved. Second one based on the better selection of data and usage of Gradient Boosting Classifier. Our last model was trained with external data with more observations, what makes it more reliable and thus, to some extend, reaches standards, that are obligatory while selecting model for professional, medical purposes, such as defending ourselves at the time of Covid-19 pandemics.