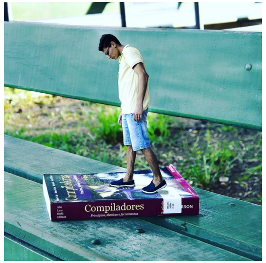
HOSPITAL ISRAELITA
ALBERT EINSTEIN

kaggle

# Brazil against the advance of Covid-19



07th May

Crislânio Macêdo

# Diagnosis of COVID-19 and its clinical spectrum

AI and Data Science supporting clinical decisions(from 28th Mar to 1st Apr)

source

# Task Details

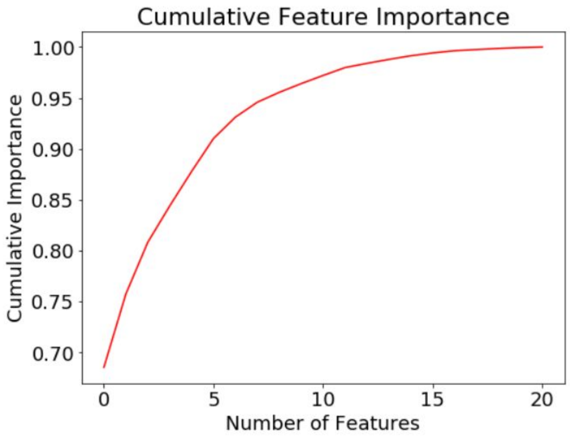
Predict confirmed **COVID-19** cases among suspected cases. Based on the results of laboratory tests commonly collected for a suspected **COVID-19** case during a visit to the emergency room, would it be possible to predict the test result for **SARS-Cov-2** (positive/negative)?

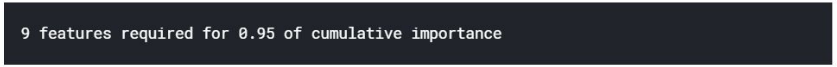
# Content

1. Glimpse of Data
   1.1 Imports 1.2 Read in Data
2. Exploratory Data Analysis 2.1 Examine the Distribution of the Target Column 2.2 Examine Missing Values 2.3 Handling missing values 2.4 Column Types 2.5 Correlations
3. Visualizations
4. Encoding Variables 4.1 Label Encoding
5. Feature Selection 5.1 Remove Collinear features 5.2 Remove features with greater than a threshold percentage of missing values 5.3 Feature Selection through Feature Importances 5.4 LGB Shap Values
6. Modeling 6.1 Confusion Matrix function 6.2 Gradient Boosting Model function 6.3 XGB
7. Ensembling With StackNet 7.1 StackNet Modeling 7.1 Evaluation StackNet ROC_AUC
8. General Findings
9. End Notebooks

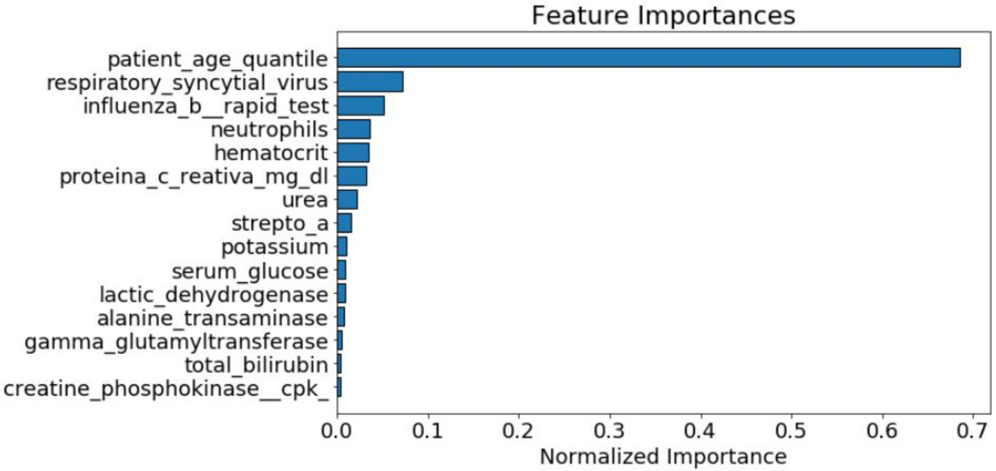
# Demo

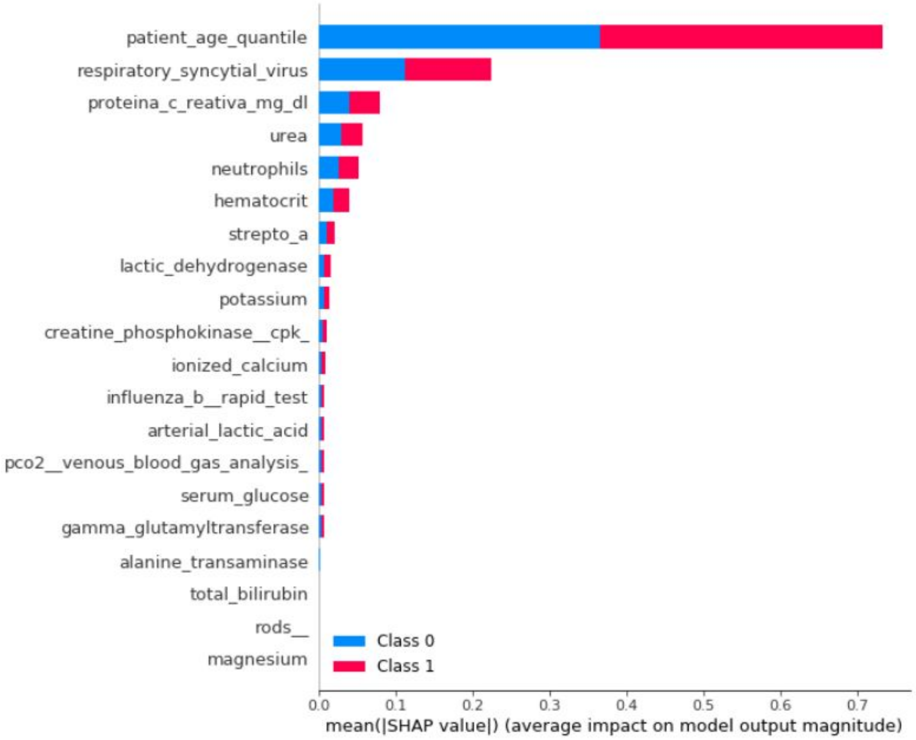
Original features 111, after reduction 21 features.





```
9 features required for 0.95 of cumulative importance
```




patient_age_quantile
respiratory_syncytial_virus
proteina_c_reativa_mg_dl
urea
neutrophils
hematocrit
strepto_a
lactic_dehydrogenase
potassium
creatine_phosphokinase__cpk_
ionized_calcium
influenza_b__rapid_test
arterial_lactic_acid
pco2__venous_blood_gas_analysis_
serum_glucose
gamma_glutamyltransferase
alanine_transaminase
total_bilirubin
rods__
magnesium
🟦 Class 0
🟥 Class 1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
mean(|SHAP value|) (average impact on model output magnitude)

# Research Question

What is the best way to predict positive results?

Focus on reducing false negatives. VS

Focus on reducing false positives. VS

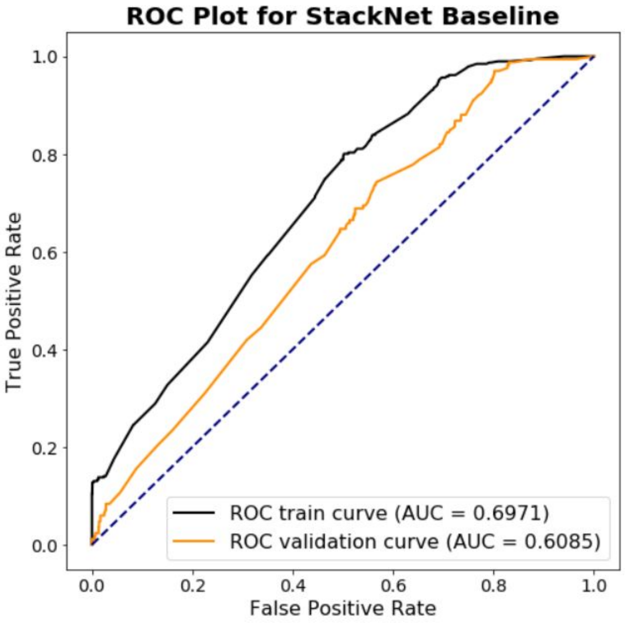
Focus on a custom balance?

# Results - StackNet

StackNet allows you to define all kinds of models. For example, Sklearn models, **LightGBM**, **RandomForest** and **CatBoost** can all be used with StackNet.

The **StackNetClassifier** will perform cross-validation (CV) and will output the CV scores for each model. To make sure we can output a probability of pacient result for sarscov2 exam result we specify "use_proba=True".

# Results - StackNet



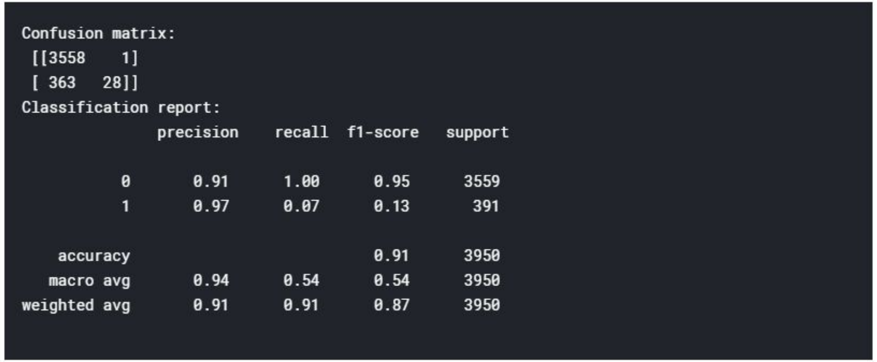
A model with high recall achieves good results in finding positive patients among those true positive patients. But, the hospital may not have enough resources to apply the necessary procedures for all patients assigned with a positive label if that number is too high.

Hence, an ideal model is one that is well-balanced, i.e., one that has high recall but it does not over-assign patients with positive labels.

# Results - StackNet

Classification report train

Code

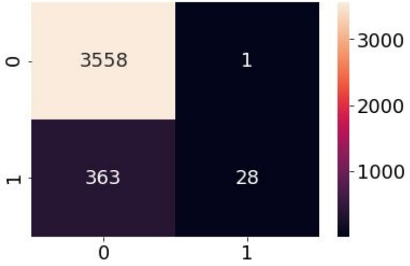


```
Confusion matrix:
 [[3558    1]
 [ 363   28]]
Classification report:
               precision    recall  f1-score   support

           0       0.91      1.00      0.95      3559
           1       0.97      0.07      0.13       391

    accuracy                           0.91      3950
   macro avg       0.94      0.54      0.54      3950
weighted avg       0.91      0.91      0.87      3950
```

Code



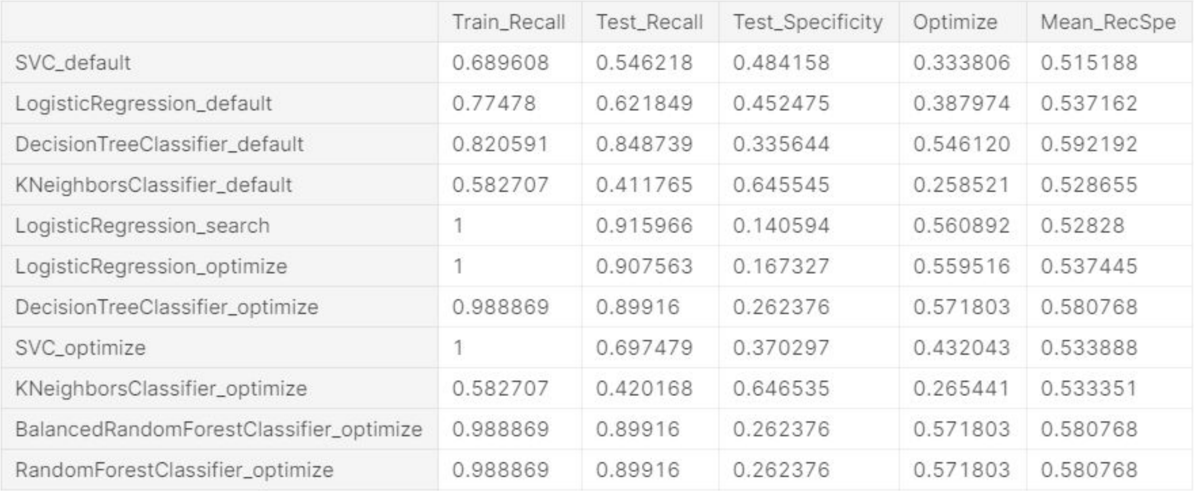
What proportion of actual positives was identified correctly?

Our model has a recall of 7% —in other words, it correctly identifies 7% of all positive tests results for SARS-Cov-2.

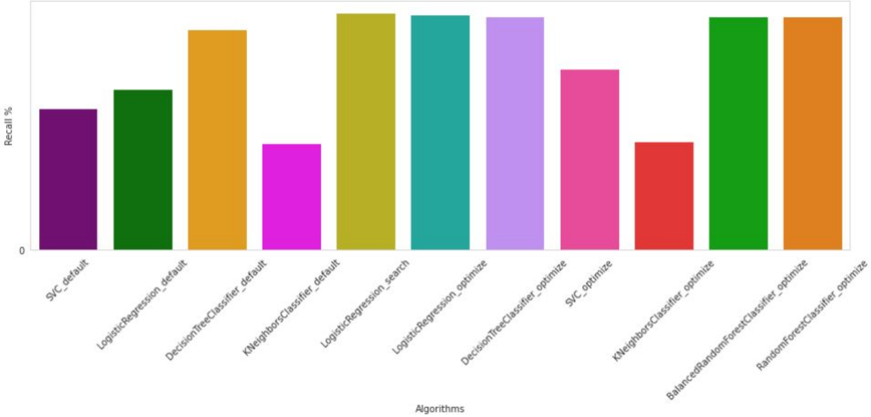
What proportion of positive identifications was actually correct? 97%

Our model has a precision of 97%—in other words, when it predicts a positive test result for SARS-Cov-2, it is correct 97% of the time.

# Results



| | Train_Recall | Test_Recall | Test_Specificity | Optimize | Mean_RecSpe |
|---|---|---|---|---|---|
| SVC_default | 0.689608 | 0.546218 | 0.484158 | 0.333806 | 0.515188 |
| LogisticRegression_default | 0.77478 | 0.621849 | 0.452475 | 0.387974 | 0.537162 |
| DecisionTreeClassifier_default | 0.820591 | 0.848739 | 0.335644 | 0.546120 | 0.592192 |
| KNeighborsClassifier_default | 0.582707 | 0.411765 | 0.645545 | 0.258521 | 0.528655 |
| LogisticRegression_search | 1 | 0.915966 | 0.140594 | 0.560892 | 0.52828 |
| LogisticRegression_optimize | 1 | 0.907563 | 0.167327 | 0.559516 | 0.537445 |
| DecisionTreeClassifier_optimize | 0.988869 | 0.89916 | 0.262376 | 0.571803 | 0.580768 |
| SVC_optimize | 1 | 0.697479 | 0.370297 | 0.432043 | 0.533888 |
| KNeighborsClassifier_optimize | 0.582707 | 0.420168 | 0.646535 | 0.265441 | 0.533351 |
| BalancedRandomForestClassifier_optimize | 0.988869 | 0.89916 | 0.262376 | 0.571803 | 0.580768 |
| RandomForestClassifier_optimize | 0.988869 | 0.89916 | 0.262376 | 0.571803 | 0.580768 |



Recall: the proportion of positive cases that were identified correctly.

Specificity: the proportion of negative cases that have been correctly identified.

Better custom balance model  is Balanced Random Forest with: **recall: 89.9%, specificity: 26%**

Focus on reducing false positives: log. regression **91.5%**

This result helps the hospital to release patients who obtained a pre-analysis diagnosed as NEGATIVE by the artificial intelligence system.

A situation in Brazil will improve #StayHome

# References

- COVID19: Recall
- Brazil against the advance of Covid-19
- EDA - first try / python lgb / shap
- XGB | LGB | CB
- LigthGBM simple FE
- Optimizing Imbalanced Classification
- Feature Selection
- Ensembling With StackNet
- EDA and Prediction

# References

- Kaggle - @CaesarLupum
- Github - @crislanio
- Linkedin - @crislanio
- Medium - @crislanio.ufc
- Twitter - @crs_macedo

# Thanks !