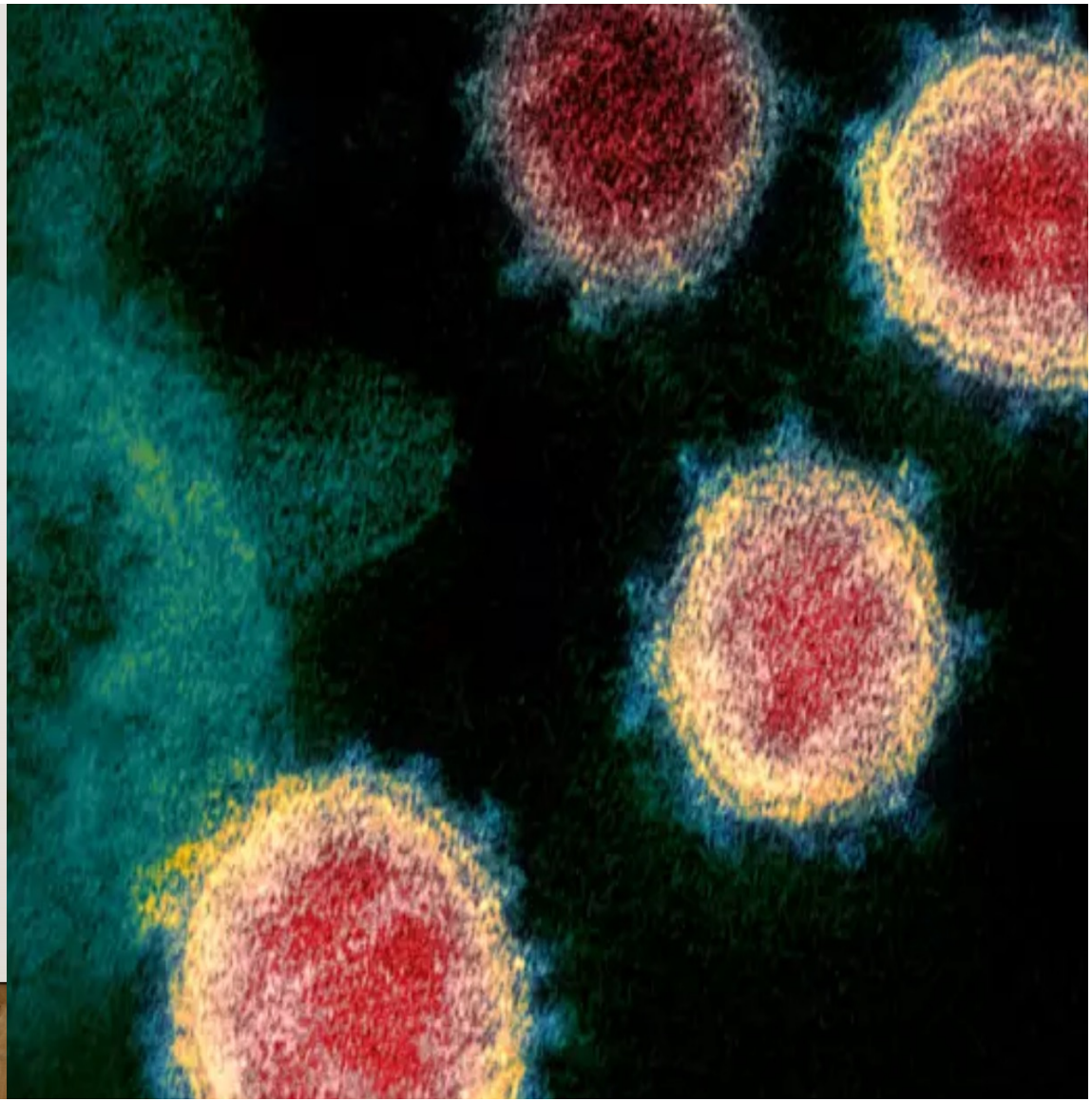
COVID-19 Predictive Analysis

Murat K. Osman

10/08/2022



Outline

- Introduction
- Analysis
- Results
- Conclusions and Recommendations

Introduction

A. Project background and context

1. The exponential rise in the number of COVID cases threatened to overwhelm health systems worldwide with a demand for hospitalization and for ICU beds far above the capacity.

2. Testing every case with some mild symptoms, would be impractical in the context of an overburdened health system with the potential limitation to performing tests for the detection of SARS-CoV-2

B. Problems you want to find answers

1. Based on the laboratory tests collected from the suspected cases, improve a classification model that predicts the chances of being positive/negative for covid19

2. What features could influence in the acceptance of patient into hospital?

Data Collection

- **Describe how data sets were collected.**

The data was gathered using a combination of API queries from Hospital Israelita Albert Einstein. There are 5644 rows and 111 columns. The summary of the dataframe for nine column is shown below.

Hospital Israelita Albert Einstein API

Patient ID	Patient age quantile	SARS-Cov-2 exam result	Patient addmited to regular ward (1=yes, 0=no)	Patient addmited to semi-intensive unit (1=yes, 0=no)	Patient addmited to intensive care unit (1=yes, 0=no)	Hematocrit	Hemoglobin	Platelets

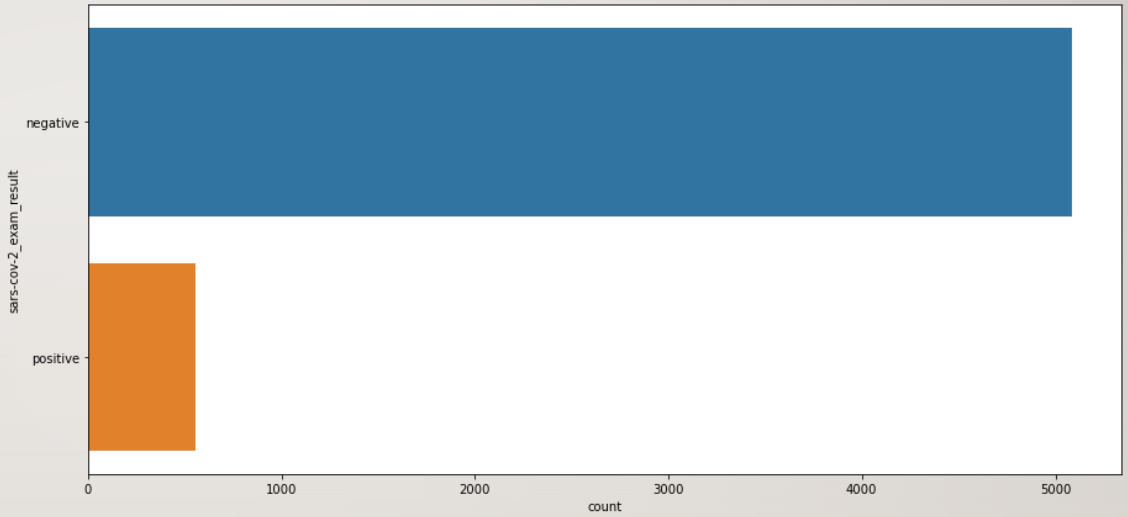
Analysis: Data pre-processing (Discrete Features)

- There are **42 categorical** values with variable 5 and less than 5.
- There are several missing values, ranging from **%76 to %100.**
- Missing values of more than **90%** were removed.

As a result, 22 columns were left.

Distribution of the Variables

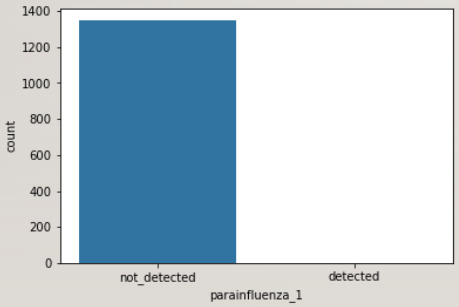
- The number of cases of COVID reported as both positive and negative were analyzed.

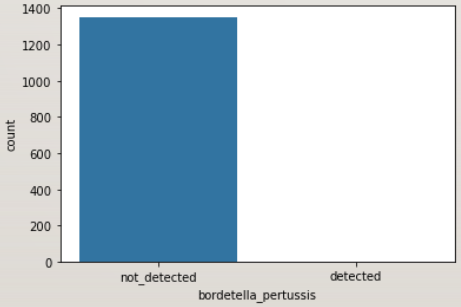


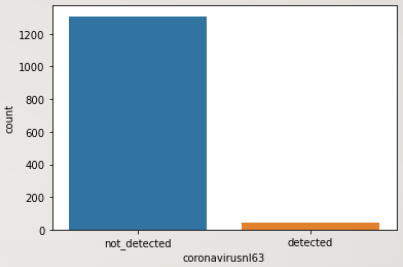
Negative	5086	91%
Positive	558	9%

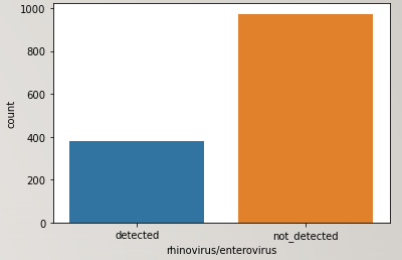
Distribution of the Variables

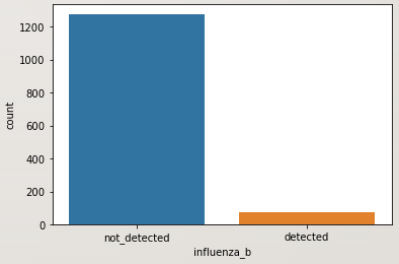
- All the discrete features were analyzed.
- Bordetella and parafiluenza_1 were never detected.
- The distribution of rhinoviruses and enteroviruses between the two groups is "relatively" more equal.
- Negative cases dominate significantly for other categories.

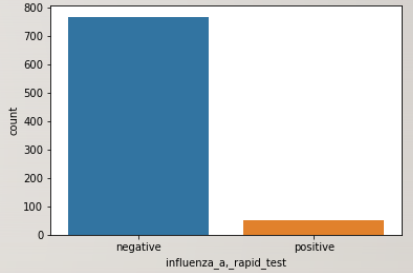


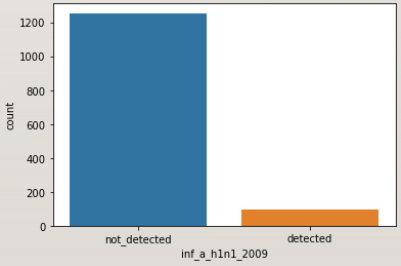


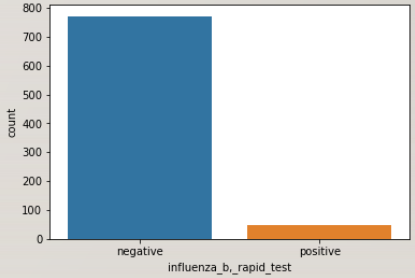












Bivariate Distributions

- **A stripplot is a fantastic way to demonstrate which attributes are associated with positive sars-cov-2_exam_result.**
- **COVID results were invariably negative when some attributes such as the influenza_a, parainfluenza_1, coronavirus_hku1, rhinovirus and parainfluenza_3 were detected. These characteristics may serve as excellent indicators of a patient's absence of a COVID positive case.**

Data pre-processing (Continuous Features)

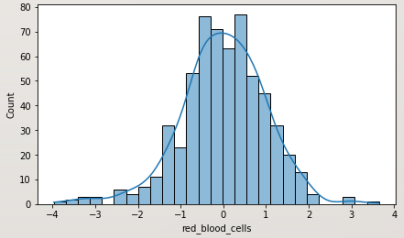
- There are several missing values, ranging from **%89 to %100.**
- Missing values of more than **90%** were removed.

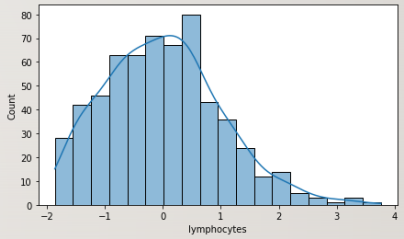
As a result, 15 columns were left.

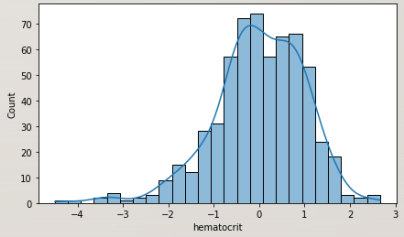
Distribution for continuous features

- More of the continuous features are close to normal distribution. For example, mean_platet, red blood_cells, monocytes etc.

- However, some of them are right, and some of them are left skewed. As an example, lymphocytes is right skewed, and hematocrit is left-skewed

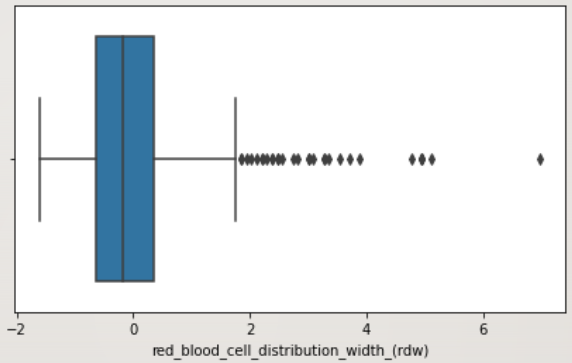


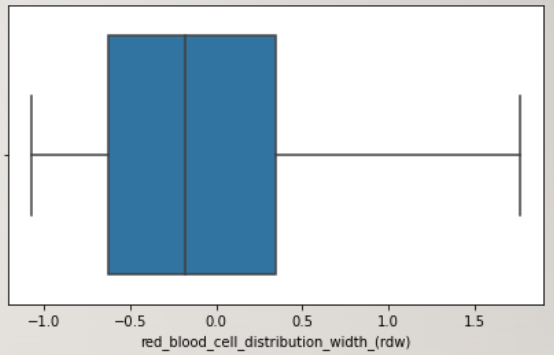




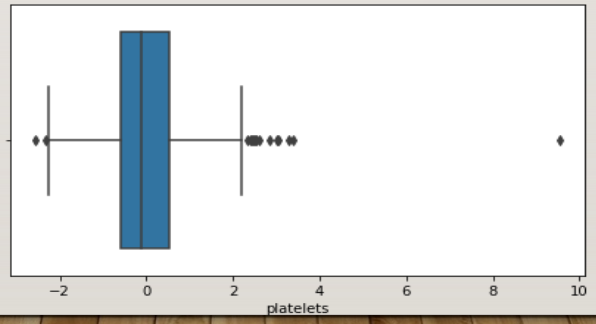
Outlier Treatment

- There are a few outliers in each numerical features, except for patient age. I decided the remove the data with 95% confidence level.

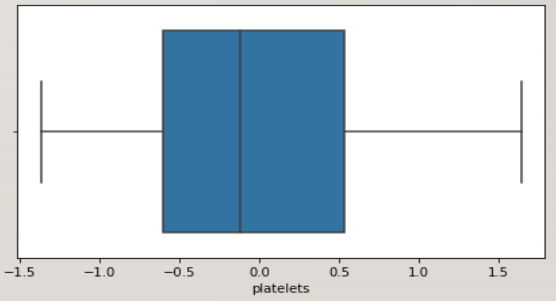




Before



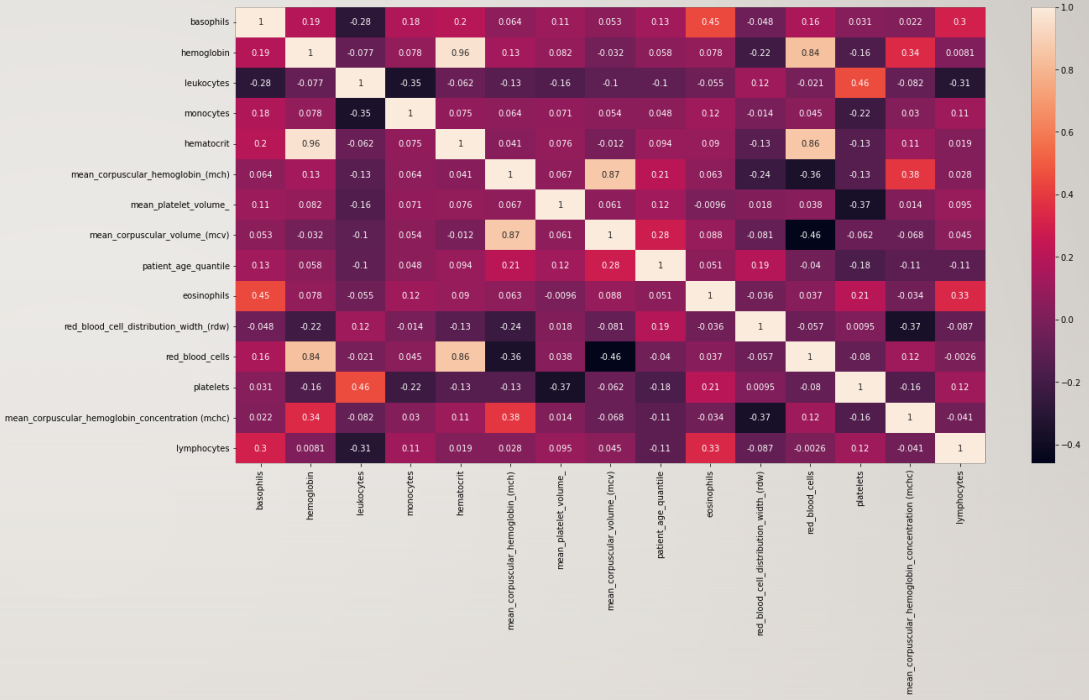
After



Correlation between features

- Red blood cells and haemoglobin have a positive correlation of 0.86. Haemoglobin, a protein rich in iron that gives blood its red color, is found in red blood cells.
- Hematocrit and haemoglobin have a positive correlation of 0.96. Red blood cell measures like haemoglobin and hematocrit identify dietary deficits, acute diseases, and long-term medical disorders.

Not : We shall be careful about the collinearity in statistics. When predictor variables in the same regression model are highly-correlated, they cannot independently predict the value of the dependent variable.

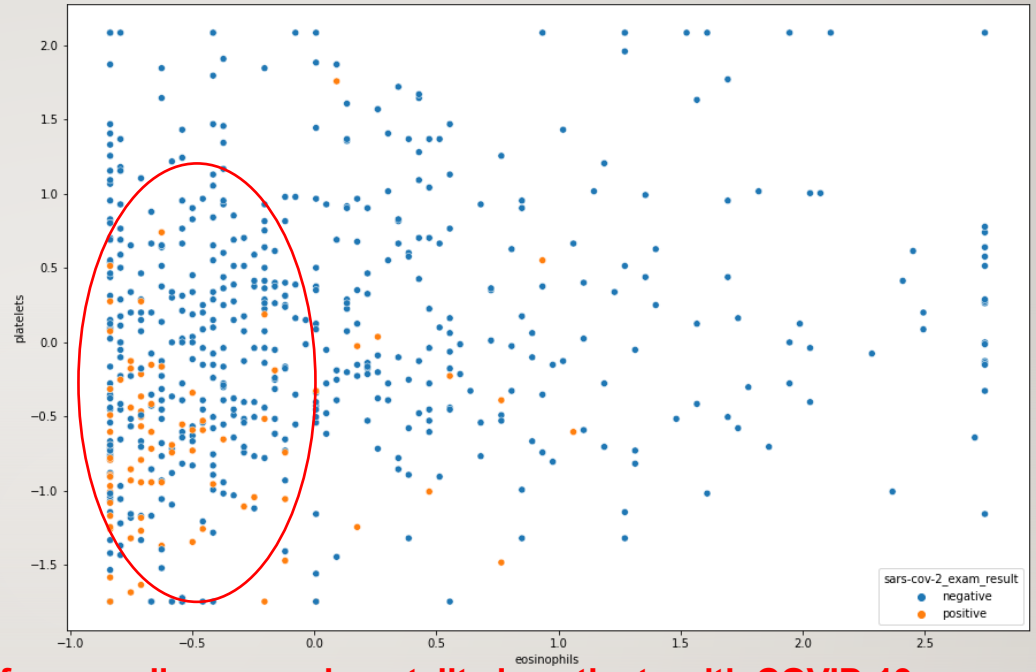


Abs values higher than 0.85 must be taken into account for collinearity.

hemoglobin and hematocrit and red_blood_cells
mean_corpuscular_volume_(mcv) and mean_corpuscular_hemoglobin_(mch)

Platelets and Eosinophils

- Most of the positive cases remain in the area where both platelets and eosinophils are negative.
  - No indication of COVID at certain areas.
  - Red circle stands for the area where low eosinophils and low platelets count exist.

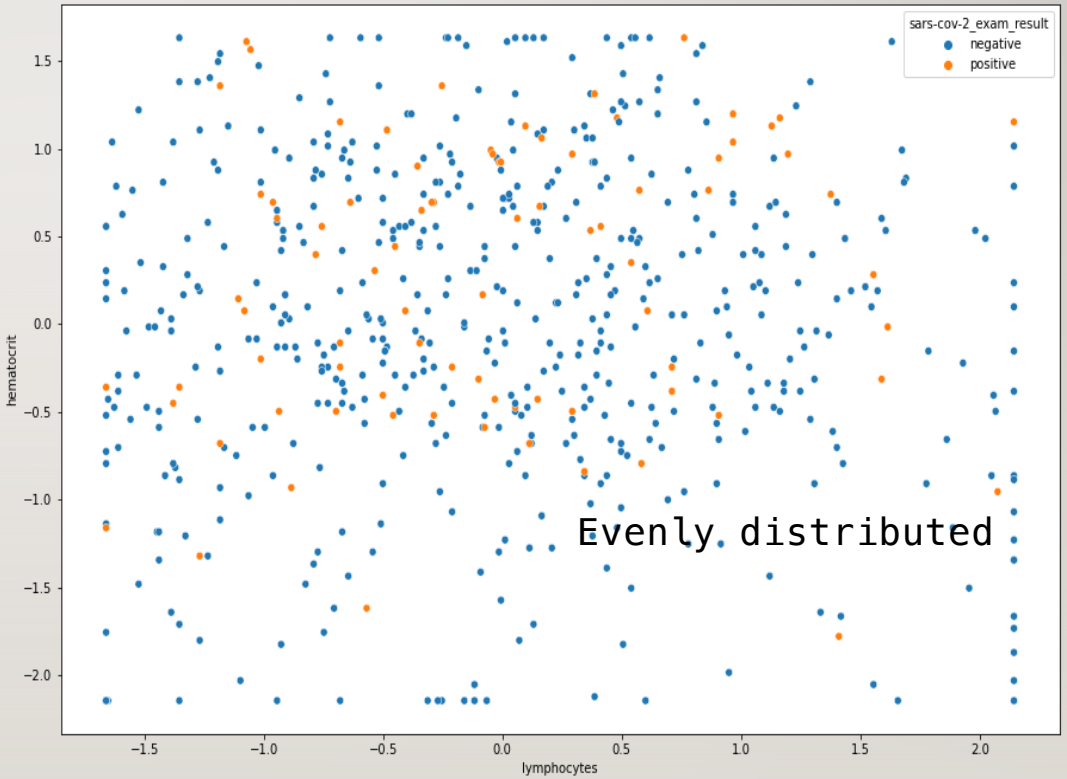


Low platelet count is associated with increased risk of severe disease and mortality in patients with COVID-19. Ref: Delshad, Mahda, et al. "Platelets in the perspective of COVID-19; pathophysiology of thrombocytopenia and its implication as prognostic and therapeutic opportunity." *International immunopharmacology* 99 (2021): 107995.

Eosinophil levels were significantly lower in patients with critical disease, when compared to those with moderate and severe diseases. Yan, Bingdi, et al. "Relationship between blood eosinophil levels and COVID-19 mortality." *World Allergy Organization Journal* 14.3 (2021): 100521.

Lymphocytes and Hematocrit

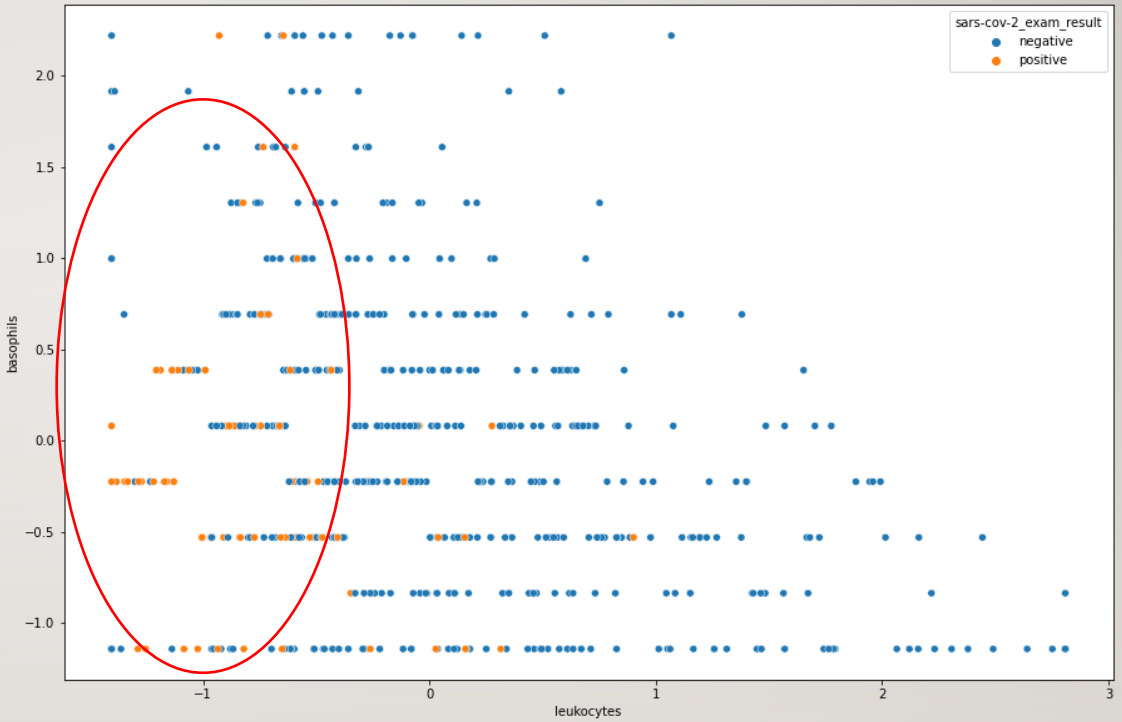
- These are two examples of points where both positive and negative covid instances are evenly distributed along the points.
- No indication of COVID at certain areas.



Basophils and Leukocytes

- Basophils and leukocytes are not well correlated, but positive covid cases exist more in the negative leukocyte's values.

More covid cases



Previous studies in the first analysis showed significantly lower leukocyte, neutrophil and platelet counts in COVID-19 pneumonia.

Ref : Soraya, Gita Vita, and Zulvikar Syambani Ulhaq. "Crucial laboratory parameters in COVID-19 diagnosis and prognosis: an updated meta-analysis." *Medicina clinica* 155.4 (2020): 143-151.

Results: Predictive Analysis (Classification)

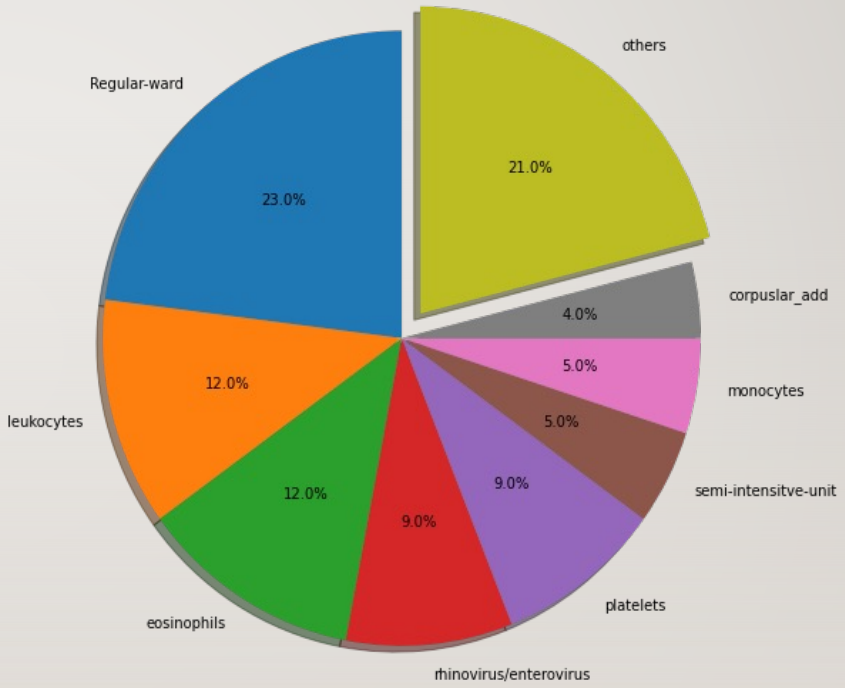
	Tr	Te	Tr	Te
	Recall		Precision	
SVM (RBF Kernel)	0.83	0.82	0.87	0.82
Random Forest	0.80	0.64	0.85	0.88
Logistic Regression	0.63	0.64	0.96	0.88

- Imputation was applied and multicollinearity was removed using "variance inflation factor" from the statsmodels library.
- Precision-Recall Curve was used in SVM and Logistic Regression.
- Tuning with best set of parameters was applied to improve overall performance.

Feature importance

- **Observations:**
- **Corpuslar add = mean_corpuscular_volume_(mcv)**
- **+ mean_corpuscular_hemoglobin_(mch)**

- **8 features cover 79 of total feature importance. Others include the rest of the features such as lymphocytes, basophils, red_blood_cells…**



Conclusions and Recommendations

- SVM with RBF kernel has good recall of 0.82 among all the models with and precision of 0.82. **We recommend the deployment of this model among all other models because hospital will make right predictions as well as spend less time and budget with using this model.**

- We advise better data collection methods, **especially in Leukocytes, eosinophils, rhinovirus, plateles and monocytes,** because only a small number of features overall had missing values of less than 76 %. This caused us to train our model on sparse data and have overfitting issues with trainset frequently.

Thank You ☺

muratkosmanoglu@gmail.com