Applying machine learning to assist the diagnosis of COVID-19 from blood and urine exams

Jessica Almeida dos Santos[1] **, Lilian Berton**[1]

[1]Institute of Science and Technology – Federal University of Sao Paulo (UNIFESP)
São José dos Campos 12247-014 – SP – Brazil

{jessica,lberton}@unifesp.br

Abstract. *The COVID-19 pandemic declared in March 2020 by the World Health Organization (WHO) challenged the health system of several countries with the growing number of infected people. During the pandemic's peak in Europe, the low incidence of infection in South Korea drew the international community's attention, since not long ago that country was considered the epicenter of the pandemic outside its origin, in China. The mass testing protocol and tracing policies were pointed out as the formula for South Korean success, however, in view of the high demand and little supply of diagnostic tests for COVID-19 in the market, this strategy proved to be unfeasible to be implemented mainly in countries with large populations and with few financial resources, such as Brazil. There is also the aggravating factor regarding the effectiveness of the tests currently available, especially the rapid serology test with a high rate of false negatives. In order to offer a screening method for the application of tests, this work aims to develop a predictive model for assisting the identification of COVID-19 infection in suspected patients based on data from clinical laboratory examinations, such as blood count and urine tests. The data used comes from three sources in Sao Paulo and are hosted in the COVID-19 Data Sharing/BR Repository, a shared database of Sao Paulo Research Foundation (FAPESP). The work also proposes a comparison between balanced × imbalanced dataset and traditional × ensemble algorithms for this problem.*

1. Introduction

The year 2020 brought an unprecedented challenge for humanity with the emergence of the COVID-19 pandemic. A new type of coronavirus, Sars-Cov-2, discovered in December of the previous year in Wuhan (China), created a true scientific race at the global level while also forcing several nations to impose mitigating measures in order to curb the contagion in their countries. Such measures, from the simplest, covering personal hygiene and wearing a mask, to the most radical such as lockdown, have opened up the seriousness of the pandemic scenario and its economic and social consequences [WHO 2020a].

During the pandemic's peak in Europe, the decrease in the contagion and death curve in Asian countries stood out among the international community, especially with South Korea as an example in the fight against COVID-19. Further, the tracking policies implemented, the mass testing protocol adopted by the South Korean government ensured greater control of the reality of the disease in its territory [Dighe et al. 2020], being considered a determining factor in changing the situation of the nation. With this case of success, several countries that were at the beginning of their pandemic curves also decided to adopt the testing protocol, aggravating the global dispute for tests and other inputs for coping with the disease.

With the high demand for tests, Personal Protective Equipment (PPE) and other medical and hospital equipment, some countries with such missing resources had to create their own protocols to deal with suspected and confirmed coronavirus cases, such as Brazil. In the case of diagnoses, the medical indication was to test only critically ill patients in the Intensive Care Unit (ICU) [da Saúde 2020], which generated high rates of underreporting cases of COVID-19 in the country. In April 2020, the Brazilian testing rate has stood out as one of the lowest in the world, being only higher than countries like Pakistan, India and Mexico. Such scenario of uncontrolled tracking of Sars-Cov-2 in the Brazilian territory contributed even more to the aggravation of the pandemic in the country, which made Brazil the third largest nation in the number of cases in May 2020 and the second in deaths from this disease in June of the same year [WHO 2020b].

With the advancement of the Artificial Intelligence area, more specifically the Machine Learning (ML) and Data Mining (DM) subareas, health and technology academics saw the potential of this advent to modernize many manual or high-rate processes of human error in medicine. In view of this, the health sector is increasingly employing the use of ML for research in the area of radiology [Stephen et al. 2019], prediction of epidemic outbreaks and even for diagnoses of the most varied types [Gagliano et al. 2017]. Regarding diagnoses, it is known that currently, most research in the field of ML involves the classification and analysis of medical images, such as radiographs, ultrasound and X-rays, most of which are aimed at identifying various types of cancers. Despite this, there is an increasing number of projects involving other types of data sources, such as laboratory tests [Delafiori et al. 2020]. In the case of infectious diseases, ML has been gaining prominence as a way to mitigate transmission and provide improved diagnostics [Agrebi and Larbi 2020]. The development of a tracking system based on vital sign data to detect likely infected with Influenza [Sun et al. 2014] as well as the diagnosis of infected with Dengue [Mello-Román et al. 2019], exemplifies the expansion of research on the diagnosis of these types of diseases with the use of Artificial Intelligence.

In view of this already well-established scenario of the use of ML techniques for the diagnosis of diseases, this work seeks to replicate this idea for the prediction of the diagnosis of COVID-19 based on data from laboratory tests of people with suspected pathology, being carried out from blood and urine samples. Another justification for this work covers the importance of testing as a way of tracking the contagion of coronavirus in a given location. Countries of great territorial and population extension, such as Brazil, tend to have difficulties in implementing an efficient testing policy, either due to the scarcity of tests generated by the high international demand or the lack of financial resources for the purchase of this input. In addition, the quality of the test is also a relevant factor, since the accuracy of the most widely applied type of test until June 2020, the rapid serology test [Government 2020], has a high probability for false negatives. Even when it comes to the use of the gold standard diagnostic test for COVID-19, the RT-PCR, there are problems, all related to the sample collected to be subjected to analysis. The way in which the sample is treated, the place where the sample was taken and the moment of collection of the material directly interfere with the quality of the result obtained [Sinha and Balayla 2020].

This way, the existence of alternative methods for diagnosing infection with the Sars-Cov-2 virus is extremely valuable during this pandemic period and can serve as a screening for testing or counter-testing, as well as being applied to other diseases. In addition, laboratory tests of blood and urine, which will serve as the database for this work, in general, are much cheaper and faster to obtain than the most commonly used best COVID test (RT-PCR), making it even more feasible to obtain the data for prediction. The data to be used in this work comes from three different sources in the Sao Paulo city (*Instituto Fleury*, *Hospital Sírio-Libanês* and *Hospital Israelita Albert Einstein*), characterizing the need for the datasets integration process. This integration process brings some difficulties since each hospital organizes information in different formats. In addition, missing data problems and imbalanced datasets will also be addressed in the data pre-processing stage. Both traditional classification algorithms and ensemble algorithms were used for experimentation.

The main contributions of this work are threefold: i) an empirical analysis of the potential of ML assist the diagnosis of COVID-19 by blood and urine exams; ii) a comparison among many traditional ML classifiers and ensemble; iii) the provision of a pre-processed dataset for further researchers explore other types of analysis regarding COVID-19.

The remaining of this paper is organized as follows: Section 2 summarizes some works that employed ML techniques for COVID-19 analysis from blood/urine tests. Section 3 presents the materials and methods employed in this work. Section 4 presents the results obtained in the classification. Section 5 presents the concluding remarks.

2. Related work

This section presents other papers that considered blood and urine tests for the prediction of COVID-19 using ML. All these studies have shown the importance of blood tests for diagnosis or indicative of the degree of severity of COVID-19. Two of these works used the same dataset COVID-19 Data Sharing/BR, but they did not perform a broad analysis as we did, comparing traditional and ensemble classifiers on a large amount of data, exploring undersampling and oversampling.

[Yao et al. 2020] investigated the detection of severely ill patients with COVID-19 from those with mild symptoms using the clinical information and the blood/urine test data. They consider 137 clinically confirmed cases of COVID-19, which were collected from a hospital in China. Each sample has 100 features, consisting of 8 clinical, 76 blood tests, and 16 urine test values. They consider a binary classification problem between 75 severe/deceased cases and 62 mild/moderate ones. They used Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), k-Nearest Neighbor (kNN) and AdaBoost. SVM achieved the best prediction accuracy of 81% using only 28 features.

[Brinati et al. 2020] developed an ML classification from hematochemical values from routine blood exams drawn from 279 patients admitted to the San Raffaele Hospital (Milan, Italy). Of these patients, 177 resulted positive, whereas 102 received a negative response. They used Decision Tree (DT), Extremely Randomized Trees, kNN, LR, Naive Bayes, RF and SVM to discriminate between patients who are either positive or negative to the SARS-CoV-2. Random Forest was selected as reference best performing model with accuracy equal 86%. The models mentioned above have been trained, and evaluated, through nested cross-validation.

[de Freitas Barbosa et al. 2021] proposed a system to support COVID-19 diagnosis based on blood testing. They used the database provided by *Hospital Albert Einstein*, Sao Paulo, Brazil. The dataset consists of 5,644 patients where 559 were COVID-19 positive, so they used *Synthetic Minority Oversampling TEchnique* (SMOTE), an oversampling technique for generating synthetic samples from the minority class. Finally, they used 9,155 training instances and 1,017 validation instances but it's not clear if the validation set is contaminated by SMOTE synthetic samples. They employed Multilayer Perceptron (MLP), SVM, Random Trees, RF, Naive Bayes and Bayes Network being the best of them for accuracy, which was greater than 95%. RF also showed good performance, with accuracies around 95%. In addition, only 24 blood tests were needed.

[Silveira 2020] performed the prediction of the diagnosis of COVID-19 based on blood count results and age of patients using data from 1,157 patients made available by the COVID-19 Data Sharing/BR repository (the same dataset used in this work). They used 349 patients from the positive group, and 808 from the negative group. The patients were grouped into positive and negative for the COVID-19 training set (which corresponded to 75% of the data) and the independent test set (which corresponded to the remaining 25%). Only the GBoost classifier was used and reached an accuracy of 80%.

[Tem-Caten et al. 2020] also used the data from the repository COVID-19 Data Sharing/BR to measure the influence of age and sex in the clinical profile of infected people from COVID-19. Through bioinformatics methods, more than 200 laboratory parameters of the thousands of patients in the dataset were analyzed. Among the most relevant findings in the context of this work, we can mention that individuals with the Sars-Cov-2 virus have a lower level of basophils, eosinophils and platelets.

3. Methodology

In this section, each activity related to the stages of preparing this work are detailed, involving the entire cycle of a data modeling project with ML and the dataset used. We chose to treat the data relating to exams first since such information was the most significant in the context of the work. Then, patient data was processed and integrated with information from the previous step. From the resulting dataset, which was imbalanced in relation to the COVID-19 diagnostic classes, two other balanced sets of data were created: one by the *random undersampling* technique and the other by SMOTE *oversampling*. In the following sections, all the transformations made to the raw data will be detailed until the final three datasets are obtained.

3.1. Pre-processing of exams dataset

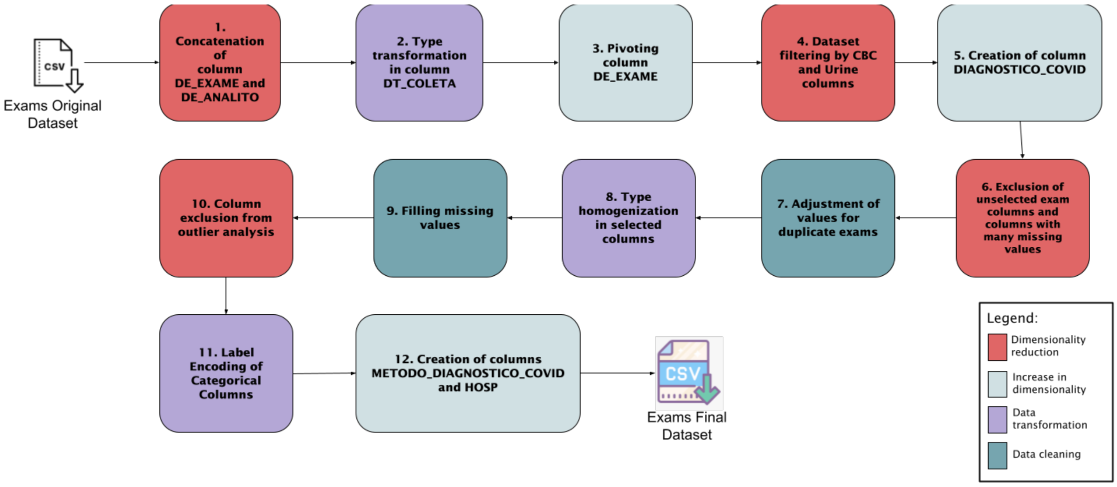
For the exams datasets, the features are shown in Table 1. The steps employed to process the data are shown in Figure 1. Each of the procedures are briefly described below:

1. Concatenation of the columns DE_EXAME and DE_ANALITO: This procedure was necessary since the same exam, such as the blood count, adds several analytes of interest, such as basophils, eosinophils, etc. After the concatenation, the column DE_ANALITO was excluded to avoid redundancies.
2. Type transformation of column DT_COLETA: When investigating the types of data in the raw dataset, it was observed that the date collection column of the exam

Table 1. Features from exam dataset. Source: [FAPESP 2020].

Feature name	Description	Value type
ID_PACIENTE	Unique identification	*String*
DT_COLETA	Date of collection	*String* DD/MM/AA
DE_ORIGEM	Patient origin	LAB or HOSP
DE_EXAME	Exam description	*String*
DE_ANALITO	Analyte description, such as leukocytes, etc.	*String*
DE_RESULTADO	Exam result, associated with DE_ANALITO	Integer or String
CD_UNIDADE	Unit of measure used in the Fleury Group	*String*
DE_VALOR_REFERENCIA	Reference values for DE_RESULTADO	*String*

Figure 1. Pre-processing flowchart for data exams


CSV
Exams Original Dataset
1. Concatenation of column DE_EXAME and DE_ANALITO
2. Type transformation in column DT_COLETA
3. Pivoting column DE_EXAME
4. Dataset filtering by CBC and Urine columns
5. Creation of column DIAGNOSTICO_COVID
6. Exclusion of unselected exam columns and columns with many missing values
7. Adjustment of values for duplicate exams
8. Type homogenization in selected columns
9. Filling missing values
10. Column exclusion from outlier analysis
11. Label Encoding of Categorical Columns
12. Creation of columns METODO_DIAGNOSTICO_COVID and HOSP
CSV
Exams Final Dataset
Legend:
Dimensionality reduction
Increase in dimensionality
Data transformation
Data cleaning

was erroneously in the format of the string. With that, the data of that column was transformed into DateTime.

3. Pivoting of the DE_EXAME column: In order to obtain a dataset in the appropriate format for the performance of ML algorithms, it was essential to transform each unique value of DE_EXAME in a new column in the dataset, which would be filled with their corresponding value in the DE_RESULTADO column. Therefore, it was necessary initially to index the columns ID_PACIENTE and DT_COLETA, since some patients often performed the same exam more than once and this behavior, without indexed collection date, could be considered a duplicate instance in the raw dataset. However, in the first pivoting attempt, it was found that, for the same patient, the same exam was performed more than once on the same date. Although this behavior can characterize an inconsistency, it was decided to add with the character ';' (semicolon) the multiple values for the same collection, being able to conclude the pivoting successfully. It is important to note that after this operation we obtained the dimensionality problem.
4. Filtering of the dataset by CBC (Complete Blood Count test) and Urine columns:

With the successful pivot, we get a real idea of the percentage of missing data in the dataset. Since the vast majority of columns had a lot of missing data and considering the proposal of the work to use the results of urine and blood tests as the main data, a filter was then applied to return only the instances with the non-empty column of the first analyte with less missing values from the tests of interest. Specifically, lines that were filled in the "Hemogram - Basophils" and "Urine - Aspect" columns were returned. With that, it was possible to obtain instances with very few missing values and to identify columns that were still predominantly empty.

5. Creation of the column DIAGNOSTICO_COVID: This is the attribute to be predicted by the classification algorithms, and, consequently, determinant in the permanence of an instance in the dataset, there was a need to centralize this information in the dataset in a single column. For this purpose, initially, only the results of the RT-PCR exam were copied to the new attribute, since, for the first data source analyzed, this method was the most used for the diagnosis of COVID. However, for the other sources, it was found that PCR was not always the most applied diagnostic tool, which implied the addition of IgM serology data in the new column. The choice was based on the information that IgM reagent may mean that the patient is or has been in the acute phase of infection, i.e., he is still infected.
6. Exclusion of unusual exam columns with many missing values: After the pivoting of the DE_EXAME column, it was possible to identify the level of diversity of exams performed in each of the data sources, highlighting the dataset of the Fleury Institute as the most diverse. From this, it was possible to conclude that there would be exams carried out only in a specific data source, which would make it impossible to remain in order to integrate all sources later. In addition, common tests that showed a large percentage of missing values in at least one hospital were also deleted in all datasets.
7. Adjustment of values for duplicate exams: In instances that, after pivoting, had multiple values divided by ';' (semicolon) in the same column, a function was then applied to return a unique value for this attribute. In the case of numerical values, the function returned the average of them if there was no repetition of the same value. Otherwise, the mode of numbers is returned. For categorical values, the mode was the default value returned.
8. Homogenization of type in selected columns: For some exams, it was observed that, although its corresponding result is predominantly numerical, there were values of the type "Less than *x*" or "Greater than *x*", with *x* being a numerical reference value for the interpretation of this exam. Given this situation, values of this type have been replaced by the reference value in its numerical form.
9. Filling missing values: Even after excluding predominantly empty columns in previous steps, some of the remaining attributes still had a few missing values. In order to avoid excluding instances of this type, it was decided to replace the missing value with the average of the attribute in the corresponding dataset.
10. Exclusion of columns with many outliers: We notice that most of the blood count analytes had redundant information located in two different columns with different units of measurement: one in cubic millimeter (mm^3) and the other in percentage (%). To decide which of the measurement units would be chosen to remain in the dataset, we visualize the distribution of the values of these columns by means of

a boxplot. Since for all data sources, the columns in mm^3 had a large number of observed outliers, it was decided to exclude them and keep the information only in the columns in percentage.

11. Label Encoding of categorical columns: Since many ML algorithms cannot execute on datasets with categorical values in the format of the string, an important step in the pre-processing is the replacement of these by corresponding numeric values, the so-called Label Encoding. This procedure was applied to all categorical attributes of all data sources used.
12. Creation of METODO_DIAGNOSTICO_COVID and HOSP columns: Aiming the integration of data sources, the created HOSP column is a way to keep the source of the instance stored. In the same way, it was also decided to create the METODO_DIAGNOSTICO_COVID column in order to maintain the information of what type of diagnostic exam of COVID-19 was used in the patient.

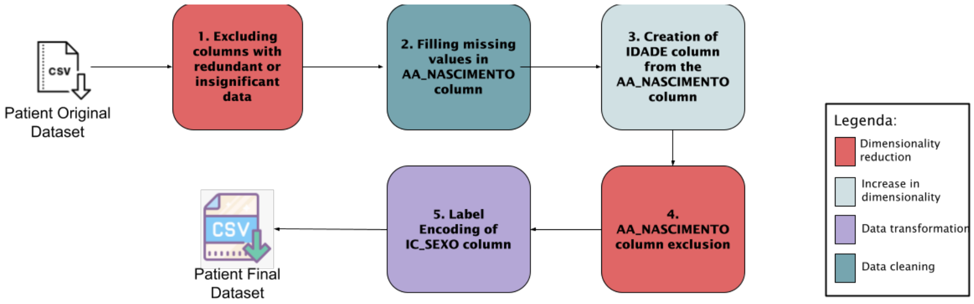
3.2. Pre-processing patients dataset

In the case of datasets with anonymized patient information, the features are shown in Table 2. The flow of data pre-processing is presented in Figure 2, and detailed as follows:

Table 2. Features from patient dataset. Source: [FAPESP 2020].

Feature name	Description	Value type
ID_PACIENTE	Unique identification	*String*
IC_SEXO	Patient Sex	'F' and 'M'
AA_NASCIMENTO	Year of birth	Numeric
CD_PAIS	Country of residence	'BR' or 'XX'
CD_UF	Federation unit of residence	*String*
CD_MUNICIPIO	Municipality of residence	*String*
CD_CEPREDUZIDO	First 5 digits of the ZIP code of residence	*String*

Figure 2. Pre-Processing flowchart for patient data


CSV
Patient Original Dataset
1. Excluding columns with redundant or insignificant data
2. Filling missing values in AA_NASCIMENTO column
3. Creation of IDADE column from the AA_NASCIMENTO column
4. AA_NASCIMENTO column exclusion
5. Label Encoding of IC_SEXO column
CSV
Patient Final Dataset
Legenda:
Dimensionality reduction
Increase in dimensionality
Data transformatio
Data cleaning

1. Exclusion of columns with redundant or insignificant data: Given the three data sources for this problem are located in the state of Sao Paulo, data such as the patient's country and federative unit have little aggregation, thus allowing its exclusion in the dataset. Regarding the patient's city and zip code information, a large number of invalid values were observed in both columns, which would probably indicate a failure in the collection.
2. Filling missing values in column AA_NASCIMENTO: The few missing values detected in the year of the birth column have been replaced by the mode of this attribute in the corresponding dataset.

3&4. Creation of the IDADE column from the AA_NASCIMENTO: This column was created from the year of birth of the patient and the year of collection of the instance. Thus, the attribute AA_NASCIMENTO becomes redundant, so we can exclude it.

5. Label encoding in the IC_SEXO: Values that were previously in the form of strings have been replaced by integers.

3.3. Feature correlation

The Pearson's correlation was calculated for the imbalanced dataset. None attributes have a correlation with the target variable DIAGNOSTICO_COVID, and few attributes have a high correlation with each other. Therefore, it was decided that all attributes will be used for training the models, totaling 23 features.

3.4. Data integration and balancing

After processing the exam and patient data for all sources, the integration step could then be performed in order to obtain a final dataset. The final dataset is available in author GitHub [Santos 2021] for public download. The dataset obtained was highly imbalanced, thus making the data balancing procedure necessary. For this, two balancing techniques were used: *Random Undersampling* and *SMOTE Oversampling*.

3.5. Datasets

All data used in this work come from COVID-19 Data Sharing/BR [FAPESP 2020], a shared database from Sao Paulo Research Foundation (FAPESP). This dataset is an initiative of this foundation in cooperation with the University of Sao Paulo and participation of the *Instituto Fleury*, *Hospital Sírio-Libanês* and *Hospital Israelita Albert Einstein*, with the objective of providing data related to COVID-19. The available data presents three categories of information:

- Demographic Data: Gender, year of birth and region where the patient resides.
- Clinical/laboratory test data: blood count, urine, COVID-19 test, Zika, etc.
- Data on patient movement: Hospitalizations, recovery and death.

The data was downloaded from the repository site on December 4, 2020, containing exam collection information since November 2019. Table 3 shows the amount of information labeled as positive and negative for a COVID-19 infection by the data source. After integrating all data sources, the final dataset distribution of classes is shown in Table 4.

Table 3. Distribution of classes by data source

Hospital	Total instances	Negative class	Positive class
Einstein	1,830	1,553	277
Fleury	4,407	4,083	324
Sírio Libanês	41	30	11
All	6,278	5,647	631

Table 4. Distribution of classes by final dataset

Dataset	**Total instances**	**Negative class**	**Positive class**
Imbalanced	6,278	5,647	631
Random Undersampling	1,262	631	631
SMOTE *Oversampling*	11,294	5,647	5,647

4. Classification results

The default parameters of the library *scikit-learn* for the algorithms were used and the holdout technique was chosen with the classic distribution of 70% for training and 30% for testing. This data split generated an imbalanced test set since no stratified parameter was set.

For the imbalanced dataset, the metrics shown in Table 5 were obtained for each algorithm, where the best ones are highlighted. In this dataset, the classifiers hit only the examples labeled with the majority class, reflecting high accuracy. However, very low values for other relevant metrics, such as sensitivity and precision. The best F1 score was achieved by Naïve Bayes.

Table 5. Evaluation Metrics - Imbalanced dataset

Algorithm	**Accuracy**	**Sensitivity**	**Precision**	**Specificity**	**F1-Score**
Logistic Regression	0.905	0.005	0.333	0.998	0.011
Naïve Bayes	0.831	0.230	0.184	0.894	**0.205**
KNN	0.893	0.0505	0.214	0.980	0.082
SVM	0.905	0.0	0.0	**1.0**	0.0
Decision Tree	0.821	**0.196**	0.153	0.886	0.172
Random Forest	**0.906**	0.034	**0.6**	0.997	0.064
Extra Tree	0.821	0.191	0.149	0.886	0.168
Gradient Boosting	0.904	0.073	0.448	0.990	0.125
XGBC	0.904	0.062	0.44	0.992	0.108

Then, we employed *Random Undersampling* in the dataset. The metrics obtained are shown in Table 6. There is a decrease in accuracy in relation to the previous experiment, however, a significant increase in the other metrics. Such changes suggest that a balanced dataset, even with few examples, guarantees a more assertive model for the task of classifying patients with COVID-19. The best F1-score was achieved by *Gradient Boosting* with a value of 0.602. The True Positive Rate (Sensitivity) was around 0.566 and the True Negative Rate (Specificity) was around 0.713.

For the SMOTE *Oversampling*, the first experiment SMOTE was applied in the training set only (70%) to increase the small class. The evaluation metrics are shown in Table 7. There is an improvement in Sensitivity compared to the experiments with an imbalanced dataset reported in Table 5, however, the results are below the random undersampling. The best F1-score was achieved by Logistic Regression with a value equal 0.263.

Finally, we applied SMOTE *Oversampling* in the training set and the validation set was balanced by *Random Undersampling*. The metrics obtained are shown in Table 8.

Table 6. Evaluation Metrics - *Random Undersampling* dataset

Algorithm	Accuracy	Sensitivity	Precision	Specificity	F1-Score
Logistic Regression	0.630	0.566	0.622	0.688	0.593
Naïve Bayes	0.612	0.377	**0.660**	**0.824**	0.480
KNN	0.519	0.505	0.494	0.532	0.5
SVM	0.577	0.433	0.573	0.708	0.493
Decision Tree	0.593	0.555	0.574	0.628	0.565
Random Forest	0.633	**0.577**	0.623	0.683	0.599
Extra Tree	0.554	0.561	0.528	0.547	0.544
Gradient Boosting	**0.644**	0.566	0.641	0.713	**0.602**
XGBC	0.635	0.566	0.629	0.698	0.596

Table 7. Evaluation Metrics - SMOTE *Oversampling* in training set

Algorithm	Accuracy	Sensitivity	Precision	Specificity	F1-Score
Logistic Regression	0.687	**0.574**	0.170	0.699	**0.263**
Naïve Bayes	0.699	0.513	0.165	0.72	0.249
KNN	0.661	0.415	0.125	0.687	0.192
SVM	0.687	0.415	0.136	0.717	0.205
Decision Tree	0.797	0.207	0.138	0.860	0.166
Random Forest	**0.897**	0.153	**0.424**	**0.977**	0.225
Extra Tree	0.77	0.240	0.130	0.827	0.169
Gradient Boosting	0.888	0.325	0.968	0.929	0.197
XGBC	0.888	0.142	0.329	0.968	0.198

In this case, all the metrics have increased the values achieving the highest results. The best F1-score was achieved by *Extra tree* with a value equal to 0.869, Sensitivity equal to 0.80, and Specificity equal to 0.959. This indicates a complete balanced dataset (train and test set) leads to the best results.

However, the real datasets are imbalanced, so to get a more expressive number of examples for each class is the way to obtain a more assertive final model.

Table 8. Evaluation Metrics - SMOTE *Oversampling* in training set and *Random Undersampling* in validation set

Algorithm	Accuracy	Sensitivity	Precision	Specificity	F1-Score
Logistic Regression	0.643	0.676	0.634	0.610	0.655
Naïve Bayes	0.607	0.508	0.634	0.707	0.564
KNN	0.809	**0.849**	0.787	0.770	0.817
SVM	0.607	0.616	0.605	0.597	0.610
Decision Tree	0.877	0.802	0.944	0.952	0.867
Random Forest	0.877	0.764	**0.987**	**0.990**	0.861
Extra Tree	**0.879**	0.800	0.951	0.959	**0.869**
Gradient Boosting	0.562	0.157	0.832	0.968	0.264
XGBC	0.574	0.184	0.840	0.965	0.301

5. Conclusion

The development of this work included the study and application of all the steps required for the classification task using machine learning techniques, from data preprocessing, exploratory data analysis, application of ML algorithms and evaluation of obtained models. In order to obtain a predictive model capable of correctly classifying a patient infected with the COVID-19 virus, we used data from clinical examinations from a public repository that concentrates information of this type, from three different sources in the state of Sao Paulo. After costly pre-processing of data resulting in 23 features, the final set obtained proved to be extremely imbalanced (about 90% of the negative class), requiring the use of balancing techniques to obtain a more assertive model. We obtained 60% of F1-score using *random undersampling* and *Gradient Boosting* classifier and 86.9% using SMOTE *oversampling* on train set and *random undersampling* on test set and *Extra tree* classifier. The comparison of the classification algorithms reinforces the superiority in terms of the performance of the ensemble methods.

Diagnosing a person, especially considering a new pandemic disease, is challenging since there is not much knowledge about the disease and the symptoms can also indicate other potentially milder diseases. Moreover, it is very difficult for any country to develop testing kits on a large scale. In this way, it is essential to gather accurate test data using cheaper test methods, which helps screen patients. So, the development of intelligent systems based on blood tests is useful and can be an alternative to other tests like RT-PCR that takes some time, potential shortage of reagents, the need for certified laboratories, expensive equipment and trained personnel.

References

Agrebi, S. and Larbi, A. (2020). *Use of artificial intelligence in infectious diseases*, chapter 18, pages 415–423. Elsevier.

Brinati, D., Campagner, A., Ferrari, D., Locatelli, M., Banfi, G., and Cabitza, F. (2020). Detection of covid-19 infection from routine blood exams with machine learning: a feasibility study. *Journal of medical systems*, 44(8):1–12.

da Saúde, M. (2020). Plano de contingência nacional para infecção humana pelo novo coronavírus covid-19. Technical report, Ministério da Saúde, Brasília, Brasil.

de Freitas Barbosa, V. A., Gomes, J. C., de Santana, M. A., Jeniffer, E. d. A., de Souza, R. G., de Souza, R. E., and dos Santos, W. P. (2021). Heg. ia: An intelligent system to support diagnosis of covid-19 based on blood tests. *Research on Biomedical Engineering*, pages 1–18.

Delafiori, J., Navarro, L. C., Siciliano, R. F., de Melo, G. C., Busanello, E. N. B., Nicolau, J. C., Sales, G. M., de Oliveira, A. N., Val, F. F. A., de Oliveira, D. N., Eguti, A., dos Santos, L. A., Dalçóquio, T. F., Bertolin, A. J., Alonso, J. C. C., Abreu-Netto, R. L., Salsoso, R., Baía-da Silva, D., Sampaio, V. S., Judice, C. C., Costa, F. M. T., Durán, N., Perroud, M. W., Sabino, E. C., Lacerda, M. V. G., Reis, L. O., Fávaro, W. J., Monteiro, W. M., Rocha, A. R., and Catharino, R. R. (2020). Covid-19 automated diagnosis and risk assessment through metabolomics and machine-learning. *medRxiv*.

Dighe, A., Cattarino, L., Cuomo-Dannenburg, G., Skarp, J., Imai, N., Bhatia, S., Gaythorpe, K., Ainslie, K., Baguelin, M., Bhatt, S., Boonyasiri, A., Boyd, O., Brazeau, N., Charles, G., Cooper, L., Coupland, H., Cucunubá, Z. M., Djaafara, B., Dorigatti, I., and Riley, S. (2020). Report 25: Response to covid-19 in south korea and implications for lifting stringent interventions.

FAPESP (2020). FAPESP COVID-19 Data Sharing/BR. `https://repositoriodatasharingfapesp.uspdigital.usp.br`.

Gagliano, M., Pham, J., Tang, B., Kashif, H., and Ban, J. (2017). Applications of machine learning in medical diagnosis.

Government, B. (2020). Ministério da saúde amplia possibilidade de testagem para covid-19. Government of Brazil website.

Mello-Román, J., Roman, J., Gomez, S., and Garcia Torres, M. (2019). Predictive models for the medical diagnosis of dengue: A case study in paraguay. *Computational and Mathematical Methods in Medicine*, 2019:1–7.

Santos, J. (2021). Pre-processed data from fapesp covid-19 data sharing/br available on december 2020. `https://github.com/JesssySantos/ENIAC2021`.

Silveira, E. C. (2020). Prediction of covid-19 from hemogram results and age using machine learning. *Frontiers in Health Informatics*, 9(1):39.

Sinha, N. and Balayla, G. (2020). Sequential battery of covid-19 testing to maximize negative predictive value before surgeries. *Revista do Colégio Brasileiro de Cirurgiões*, 47.

Stephen, O., Sain, M., Maduh, U., and Jeong, D. (2019). An efficient deep learning approach to pneumonia classification in healthcare. *Journal of Healthcare Engineering*, 2019:1–7.

Sun, G., Matsui, T., Hakozaki, Y., and Abe, S. (2014). An infectious disease/fever screening radar system which stratifies higher-risk patients within ten seconds using a neural network and the fuzzy grouping method. *The Journal of infection*, 70.

Tem-Caten, F., Gonzalez-Dias, P., Castro, I., Ogava, R., Giddaluru, J., Silva, J., Martins, F., Aquime Gonçalves, A., Costa Martins, A., Araujo, J., Viegas, A., Cunha, F., Farsky, S., Bozza, F., Levin, A., Pannaraj, P., Silva, T., Minoprio, P., Andrade, B., and Nakaya, H. (2020). In-depth analysis of laboratory parameters reveals the interplay between sex, age and systemic inflammation in individuals with covid-19.

WHO, W. H. O. (2020a). Coronavirus disease (covid-19) advice for the public. World Health Organization website.

WHO, W. H. O. (2020b). Who coronavirus disease (covid-19) dashboard. World Health Organization website.

Yao, H., Zhang, N., Zhang, R., Duan, M., Xie, T., Pan, J., Peng, E., Huang, J., Zhang, Y., Xu, X., et al. (2020). Severity detection for the coronavirus disease 2019 (covid-19) patients using a machine learning model based on the blood and urine tests. *Frontiers in cell and developmental biology*, 8:683.