UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE CIÊNCIAS APLICADAS

FINAL REPORT SCIENTIFIC INITIATION

An Investigation into Local Interpretability Methods for Medical Diagnosis Predictions Based on Machine Learning.

Student: Lucas Tramonte

Course: Engenharia de Produção

Responsible: Cristiano Torezzan (FCA - UNICAMP)

Limeira, SP

August 30, 2023

Abstract

This report presents a synthesis of the main results obtained in this scientific initiation project. The description of activities and results demonstrates that the proposed project timeline was fully adhered to and the objectives were completely met. Publicly available databases were used, with a particular focus on COVID-19 diagnoses based on symptoms. Notably, it is highlighted that when analyzing the symptom “Sore Throat,” the interpretability results converged not only among the applied interpretable methods but also with the medical literature relating to the prevalence of symptoms during the first wave of COVID-19.

Palavras-chave : Aprendizado de máquina; Interpretabilidade; Diagnóstico médico.

Contents

1 Introduction

Machine Learning (ML) models have been widely used in various fields, driving significant paradigm shifts in decision-making processes that, until recently, were essentially performed by humans. In the healthcare sector, for example, ML models have been employed to assist professionals in predictions and diagnoses, with a focus on their interpretability.

In this context, since machine decisions can impact our lives, it becomes crucial to understand how algorithms are making certain decisions, which is known as interpretability or Explainable AI (XAI). The more interpretable a system is, the more comprehensible its functioning will be and the easier it will be to identify potential biases. This can provide us with additional insights into the phenomenon under analysis and prevent the “black box” effect in automated decision-making processes. The need for interpretability also arises from the difficulty of formalizing a decision problem with data, which can lead to a misalignment between the real objective of the decision-maker and the output the system is pursuing.

There are two main paradigms of interpretability in machine learning. The first, known as intrinsic interpretability, involves restricting the structural complexity of learning methods, making them interpretable a priori. Examples of intrinsically interpretable methods include decision trees and logistic regression. The second paradigm, known as post hoc interpretability, involves applying methods or heuristics that allow for the interpretation of results from previously trained models. Examples of post hoc interpretation methods include Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP).

Both post hoc interpretable methods mentioned have advantages and disadvantages that deserve attention. The SHAP method provides results whose units are the same as the model’s target variable, making it more comprehensible, and it ensures a fair distribution of the prediction among the features, which is generally not achieved with the LIME method. On the other hand, due to the combinatorial nature of the coalitions in the SHAP method, it has a high computational cost, whereas LIME can be used for applications that require a quicker and less in-depth solution.

In this context, the present project aims to analyze the trade-off between the performance of machine learning models and their interpretability in the healthcare field, with a focus on predicting the SARS-CoV-2 pathogen based on symptom identification. The use of odds ratios in this research field is recurrent, and the goal is to evaluate their positive and negative aspects compared to the SHAP and LIME methods, both in black-box models and in intrinsically interpretable models.

To test the concepts studied in this project, we used real COVID-19 data made available in [1] and implemented the Logistic Regression (LR), XG-Boost, and Random Forest (RF) models in Python, with the support of libraries such as Scikit-Learn [8]. In [1], the authors use ML methods to predict SARS-CoV-2 test results in the Brazilian population based on reported symptoms and sociodemographic characteristics. Although the work provides an understanding of the importance of predictive attributes, the authors do not explore more advanced aspects of interpretability, which is the focus of this study.

2 Materials and Methods

2.1 Linear Regression

Linear models are widely used in the healthcare field because they are interpretable through the weights of each attribute, allowing the determination of the contribution ϕ_j of each attribute to a prediction[5]:

$$\sum_{j=1}^{p} \phi_j(f) = f(x) - E(f(x)), \tag{1}$$

where $f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ is the prediction for the instance x and $E(f(x))$ is the average predicted value for this same instance.

2.1.1 Logistic Regression

Among the main linear models used, Logistic Regression stands out, where the dependent variable Y is binary and there is a set of p independent variables:

$$P(y^{(i)} = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)})}}, \tag{2}$$

where β_0, β_1, . . . , and β_p are the coefficients of the model, determined using the maximum likelihood method..

By manipulating equation (1), the odds ratios are obtained, which measure the likelihood of the occurrence of the dependent variable given a specific attribute, compared to the likelihood of the same occurrence in the absence of that attribute. In this way, odds ratios produce easily interpretable results, making them a useful predictive model for the healthcare field. [5]:

$$\text{odds ratio} = \frac{P(y^{(i)} = 1)}{1 - P(y^{(i)} = 1)} = e^{-(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)})} \tag{3}$$

In this work, the odds ratio from logistic regression will be used as a baseline reference in the numerical experiments to be conducted in Chapter 3. The main idea is to compare the odds ratio with the values obtained using the SHAP and LIME methods, which will be presented subsequently.

2.2 Shapley Values

On the other hand, models such as XG-Boost and Neural Networks are considered "black box" models because, despite their high performance in terms of accuracy, they are not easily interpretable and do not have a well-established relationship between the weights of the attributes and the final result. As a solution, post hoc interpretability methods can be used to present the importance of the main attributes for classification to healthcare professionals.

In this context, the SHAP method proposes to explain the local predictions of a given model by randomly generating various coalitions of attributes and determining their respective effects. The theoretical foundation of SHAP is based on game theory, assigning marginal costs to each attribute and analyzing its contribution across all possible coalition combinations. This contribution is quantified by a value known as the Shapley value. [6]:

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!(p - |S| - 1)!}{p!} \cdot (val_x(S \cup \{j\}) - val_x(S)) \tag{4}$$

where S is the subset of attributes used in the model, x is the vector of attribute values for the instance to be explained, and val_x is the prediction for the attributes in set S after marginalizing over the attributes not included in this set:

$$val_x(S) = \int f(x_1, \ldots, x_p) dP_{x \notin S} - E_X[f(X)]. \tag{5}$$

For example, given a predictive model with attributes x_1, x_2, x_3 and x_4, one can determine the coalition S consisting of only the value of x_1::

$$val_x(S) = val_x(\{1\}) = \int_{\mathbb{R}} \int_{\mathbb{R}} \int_{\mathbb{R}} f(x_1, X_2, X_3, X_4) dP_{X2X3X4} - E_X[f(X)]. \tag{6}$$

In this context, SHAP represents a linear method for additive feature attribution:

$$g(x') = \phi_0 + \sum_{j=1}^{M} \phi_j, \tag{7}$$

where g is the interpretable model and M the maximum size of the collision.

2.2.1 Properties of Shapley Values

The SHAP method demonstrates local accuracy, consistency in coalition effects, and coherence when an attribute is absent from the model [6]. Since this method is based on Shapley values, its properties are grounded in the properties of these values themselves.

Initially, presenting the property of efficiency, it can be highlighted that the model's prediction is efficiently distributed among the attributes, being the sum of all Shapley values and the average predicted value, as presented in equation (1).

The second property to be presented is symmetry, that is, if for all $S \subseteq \{1, \ldots, p\} \setminus \{j, k\}$, attributes j and k contribute equally to all possible coalitions ($val_x(S \cup \{j\}) = val_x(S \cup \{k\})$), then:

$$\phi_j = \phi_k \tag{8}$$

Furthermore, it is known that attributes not used in the predictive model preditivo ($val_x(S \cup \{j\}) = val_x(S)$) assume a zero Shapley value, which constitutes a property known as "Dummy".

Finally, the property of additivity is noteworthy for methods like Random Forest that use ensemble learning, as the final Shapley value can be calculated from the average of the respective Shapley values of each model used in the ensemble.

2.2.2 Case study for the SHAP method

To exemplify the calculation of each feature's contribution, consider a function f(x) from a predictive model that includes two binary attributes: x_1, representing the symptom "Loss of Smell," and x_2, representing the symptom "Sore Throat":

$$f(x) = a_1 \cdot x_1 + a_2 \cdot x_2, \tag{9}$$

where a_1 e a_2 are the weights of each attribute.

The features have a certain distribution in a database, as represented in Table 1 for illustrative purposes.

Table 1: Probability distribution for each attribute.

i	$P(x_1 = i)$	$P(x_2 = i)$
0	0.3	0.6
1	0.7	0.4

To calculate the Shapley values of the attributes, one must obtain the values of all possible coalitions among them. In this sense, by fixing the values of the attributes as x_1 = 0 and x_2 = 1, for example, one obtains:

$$E_X[f(X)] = 0.7 \cdot a_1 + 0.4 \cdot a_2. \tag{10}$$

$$val_x(\{1, 2\}) = f(0, 1) - E_X[f(X)] = 0.6 \cdot a_2 - 0.7 \cdot a_1. \tag{11}$$

Continuing with the calculations for the remaining coalitions using equation (5), we have:

$$val_x(\{1\}) = \int f(0, x_2) dP_{X2} - E_X[f(X)], \tag{12}$$

$$val_x(\{1\}) = \sum_{i=0}^{1} f(0, i) \cdot P(x_2 = i) - E_X[f(X)] = -0.7 \cdot a_1. \tag{13}$$

$$val_x(\{2\}) = \int f(x_1, 1) dP_{X1} - E_X[f(X)], \tag{14}$$

$$val_x(\{2\}) = \sum_{i=0}^{1} f(i, 1) \cdot P(x_1 = i) - E_X[f(X)] = 0.6 \cdot a_2. \tag{15}$$

$$val_x(\{\}) = \int \int f(x_1, x_2) dP_{X1X2} - E_X[f(X)], \tag{16}$$

$$val_x(\{\}) = \sum_{j=0}^{1} \sum_{i=0}^{1} f(i, j) \cdot P(x_1 = i) \cdot P(x_2 = j) - E_X[f(X)]. \tag{17}$$

$$val_x(\{\}) = a_2 \cdot 0.3 \cdot 0.4 + a_1 \cdot 0.7 \cdot 0.6 + (a_1 + a_2) \cdot 0.7 \cdot 0.4 - (0.7 \cdot a_1 + 0.4 \cdot a_2) = 0. \tag{18}$$

Thus, since $|S| = 1$ and $p = 2$ for $S = \{1\}$;and $|S| = 0$ and $p = 2$ for $S = \{\}$, we can obtain the Shapley value for x_1 and x_2 from equation (4):

$$\phi_1 = \frac{1}{2} \cdot (val_x(\{1, 2\}) - val_x(\{2\})) + \frac{1}{2} \cdot (val_x(\{1\}) - val_x(\{\})) = -0.7 \cdot a_1 = val_x(\{1\}). \tag{19}$$

$$\phi_2 = \frac{1}{2} \cdot (val_x(\{1, 2\}) - val_x(\{1\})) + \frac{1}{2} \cdot (val_x(\{2\}) - val_x(\{\})) = 0.6 \cdot a_2 = val_x(\{2\}). \tag{20}$$

From equations 19 and 20, we can observe that the results are consistent with the fact that the interpretable SHAP method is local, as the function f adopted is linear. Moreover, the obtained results align with the property of efficiency highlighted in the equation. 1, applied to the instance (x_1 = 0 e x_2 = 1):

$$f(x) - E(f(x)) = a_2 - (0.7 \cdot a_1 + 0.4 \cdot a_2) = 0.6 \cdot a_2 - 0.7 \cdot a_1 = \phi_1 + \phi_2 = \sum_{j=1}^{p} \phi_j(f). \tag{21}$$

2.3 LIME

The LIME method involves an interpretable model g that aims to explain individual predictions by approximating them to the original model f, which exhibits the black-box effect. In this sense, LIME seeks to understand the behavior of this model by generating a new dataset, which contains both the original samples and samples that have undergone a certain perturbation. This allows for observing the proximity between these instances in a weighted manner based on the training of the data with the model g:

$$\underset{g \in G}{\operatorname{argmin}}(L(f, g, \pi_x) + \Omega(g)) = \zeta(x), \tag{22}$$

where $\Omega(g)$ is the complexity of the interpretable model, G is the family of all possible interpretable models such that $g(z') = w_g \cdot z'$ and L is the *loss function* which must be minimized:

$$L(f, g, \pi_x) = \sum_{z, z' \in Z} \pi_x(z) \cdot [f(z) - g(z')]^2, \tag{23}$$

with $(z', f(z))$ the dataset of perturbed samples and $\pi_x(z) = e^{\frac{-D(x,z)^2}{\sigma^2}}$ the proximity measure that determines the size of the neighborhood around the instance x [7].

2.4 Materials

The primary supporting materials used for the development of this project will be scientific articles that address the applications of machine learning methods to problems in the healthcare field. (ex. [3], [7], [2]).

In terms of methodology, the tests will be conducted through computational implementations in the Python programming language, using specific machine learning libraries such as Scikit-learn [8] and Imodels [9]. This project will delve into the analysis of local interpretability methods, particularly the LIME and SHAP methods, and their applicability to clinical diagnostic prediction problems using classical methods such as Logistic Regression, XGBoost, and Random Forests.

3 Results

In the tests conducted using the data provided in [1], the aim was to understand the relationship between the symptoms Cough, Fever, Sore Throat, Runny Nose, Myalgia, Nausea, Diarrhea, Loss of Smell, and Shortness of Breath, and the binary result of the COVID-19 test.

The main statistical information about the population sample is presented in Table 2, which shows the percentage values for categorical variables and the median for continuous variables.

It is evident that there is an imbalance in the response variable, with only 9,101 (14.25%) of the individuals diagnosed with the SARS-CoV-2 pathogen. Consequently, Logistic Regression, XGBoost, and Random Forest models were applied to the dataset, using an undersampling technique called "Edited Nearest Neighbors." Grid-search strategies were employed for hyperparameter optimization of each model, with 5-fold cross-validation, and the area under the ROC curve (AUC) was used as the optimization metric.

Figure 1 presents a comparison of the ROC curves for the three models used. It can be observed that the RF and XG-Boost models show superior performance compared to RL. However, these models produce results that are not as easily interpretable as those from RL, where the parameters are directly related to the importance of the attributes and allow for the calculation of odds ratios.

In this context, we used the SHAP method to calculate the distribution of Shapley values per attribute for the RF, XGBoost, and RL models, which are represented in Figures [1] 2, 3, and 4, respectively. Figure 5 presents the odds ratios obtained from the RL model along with their corresponding 95% confidence intervals.

[1] Each point depicted in the *Summary Plots* represents a Shapley value for each sample and for each attribute.

Table 2: Summary Statistics of Test Results

	Total	Positive Test	Negative Test
Participants, n (%)	63,872 (100.0)	9,105 (14.26)	54,767 (85.74)
Female, n (%)	38,970 (61.01)	5,437 (59.71)	33,533 (61.23)
Age (years), median [IQR]	41.0 [33.0 - 52.0]	43.0 [34.0 - 54.0]	41.0 [33.0 - 52.0]
Health Professional, n (%)	30,653 (47.99)	3,686 (40.48)	26,967 (49.24)
Fever	18,747 (29.35)	4,423 (48.58)	14,324 (26.15)
Cough	32,056 (50.19)	5,734 (62.98)	26,322 (48.06)
Sore Throat	27,671 (43.32)	4,037 (44.34)	23,634 (43.15)
Runny Nose	34,558 (54.11)	5,287 (58.07)	29,271 (53.45)
Myalgia	28,335 (44.36)	5,459 (59.96)	22,876 (41.77)
Nausea	9,261 (14.5)	1,865 (20.48)	7,396 (13.5)
Diarrhea	17,005 (26.62)	2,916 (32.03)	14,089 (25.73)
Loss of Smell	17,111 (26.79)	5,176 (56.85)	11,935 (21.79)
Shortness of Breath	1,136 (1.78)	312 (3.43)	824 (1.5)
NA Symptom	12,212 (19.12)	1,022 (11.22)	11,190 (20.43)

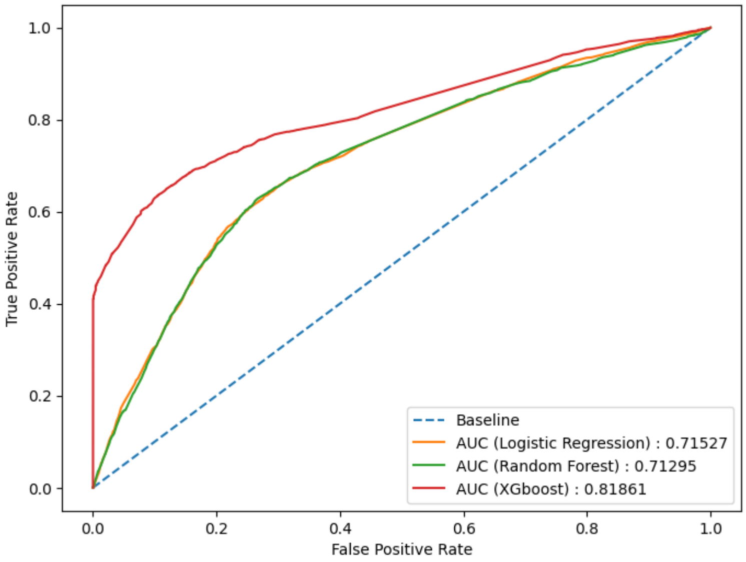

Figure 1: ROC Curve for Each Model Used.

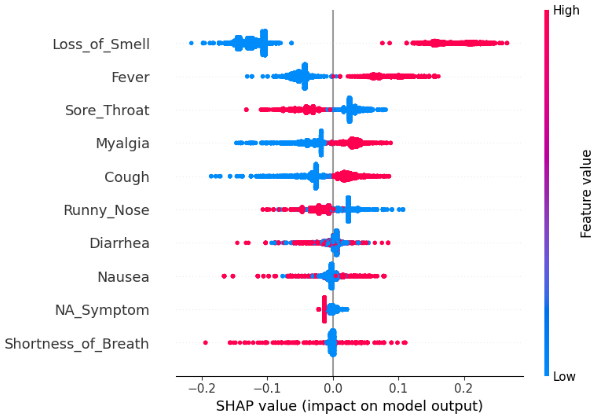

Figure 2: Summary Plot Random Forest.

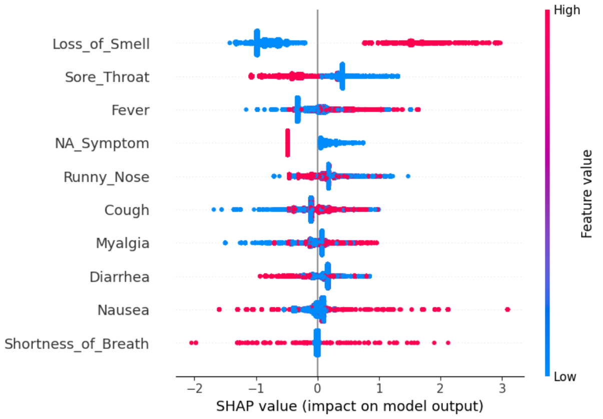

Figure 3: Summary Plot Xgboost.

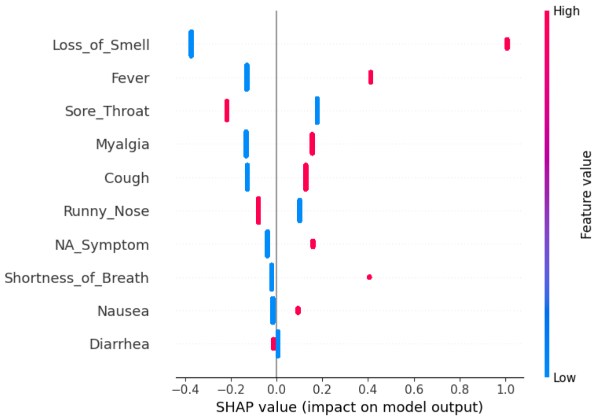

Figure 4: Summary Plot Logistic Regression.

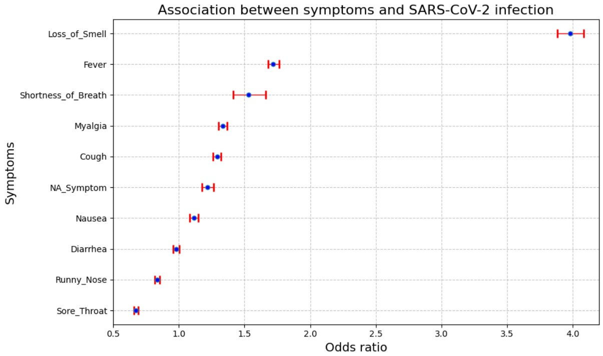

Figure 5: Odds Ratios for Each Symptom, with Their Corresponding Confidence Intervals.

Finally, the LIME method was implemented for the three classification models used. Although LIME is a local method, a global aggregation was performed on the first 10,000 samples of the dataset for each model, as shown in the Figures 6, 7 and 8.

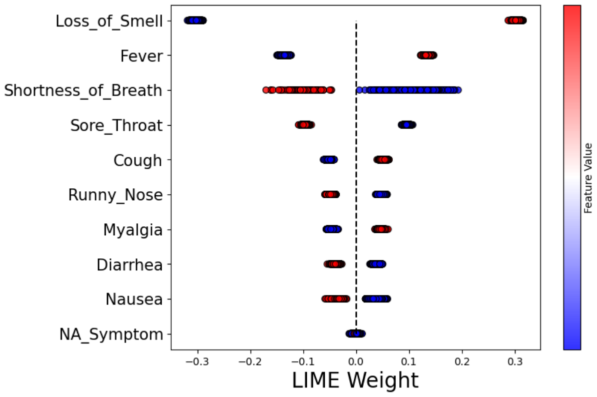

Figure 6: Beeswarm plot Random Forest.

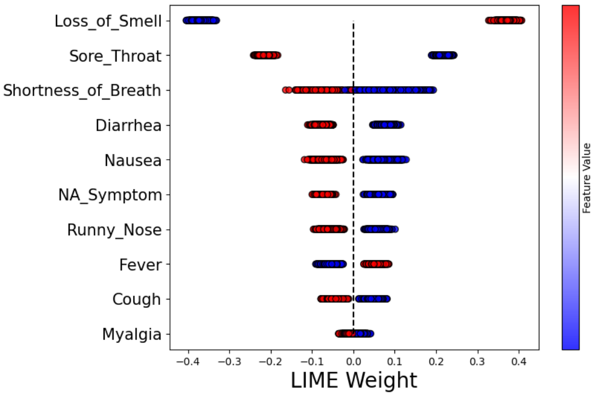

Figure 7: Beeswarm plot Xgboost.

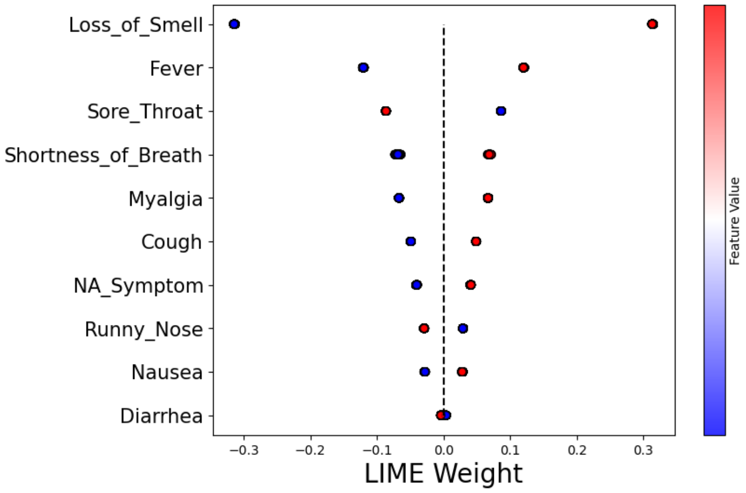

Figure 8: Beeswarm plot Logistic Regression.

4 Discussion of Results

It can be observed that the symptom related to Loss of Smell showed the highest importance in the interpretation across all models used. Additionally, the results for this symptom indicate that the presence of loss of smell in patients significantly increases the predicted probability of them having the SARS-CoV-2 pathogen.

When analyzing the symptom “Sore Throat”, it is also evident that the interpretability results converge. Although this attribute is relevant for both SHAP and LIME methods (ranking high in the ordering), the color distribution along the x-axis in all the SHAP and LIME results indicates that this symptom reduces the probability of patients testing positive for COVID-19. This finding is consistent with the result shown in Figure 5, where the odds ratio for this symptom is approximately 0.67, meaning there is about a 33% reduction in the chance of a patient testing positive for COVID-19 when this symptom is present. Considering that the data for this study were collected in 2020, this finding aligns with medical literature related to the prevalence of symptoms during the first wave of COVID-19 [4].

Another symptom of interest is “Shortness of Breath” and the last attribute in Figures 2 and 3. Since this symptom also ranks low in Figure 4, it can be inferred that this discrepancy is related to the interpretability method rather than the classification model used. Additionally, from Table 2, it is observed that this attribute has a small number of samples in the dataset, which may have contributed to its classification as a low-importance symptom by the SHAP method. Finally, all methods indicated that the symptoms “Diarrhea” and “Nausea” reduce the probability of SARS-CoV-2 classification.

5 Conclusions

In this work, we studied the main interpretable methods in machine learning models within the healthcare field, with a special focus on COVID-19 diagnosis based on symptoms. Using a public dataset, the methods LIME, SHAP, and Odds Ratio were implemented in Python for the RF, RL, and XGBoost models.

The results obtained allowed us to interpret black-box models and compare the information extracted with a priori interpretable models, as well as with the medical literature on SARS-CoV-2 classification based on symptoms.

The activities planned for the second semester of the project were fully accomplished, yielding interesting results. Additionally, the scholarship recipient completed 24 credits of their undergraduate course, passing all subjects and finishing the semester with a GPA of 0.9436 (On a scale of 0 to 1), maintaining first place in their class.

6 Material Submitted for Publication

The student presented the preliminary results of this work at the Encontro de Códigos, Reticulados e Informação (EnCoRI) at the Institute of Mathematics, Statistics, and Scientific Computing (IMECC).

References

[1] Leila F Dantas and et al. App-based symptom tracking to optimize sars-cov-2 testing strategy using machine learning. *PloS one*, 16(3):e0248920, 2021.

[2] Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. *Future Hospital Journal*, 6:94–98, 06 2019.

[3] Bert Heinrichs and Simon Eickhoff. Your evidence? machine learning algorithms for medical diagnosis and prediction. *Human Brain Mapping*, 41, 12 2019.

[4] M.F. Kristiansen and et al. Epidemiology and clinical course of first wave coronavirus disease cases, faroe islands. *Emerging Infectious Diseases*, 27(3):749–758, Mar 2021. PMID: 33513332; PMCID: PMC7920693.

[5] Tyler J Loftus, Patrick J Tighe, Tezcan Ozrazgat-Baslanti, J Parker Davis, Mathew M Ruppert, and Yulong. Ren. Ideal algorithms in healthcare: Explainable, dynamic, precise, autonomous, fair, and reproducible. *PLOS Digital Health*, 1(1):e0000006, 2022.

[6] Christoph Molnar. *Interpretable Machine Learning*. Github, 2 edition, 2022.

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “why should i trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[8] Scikit-learn. Machine learning in python. https://scikit-learn.org/stable/. Accessed: 2022-05-17.

[9] Chandan Singh, Keyan Nasseri, Yan Shuo Tan, Tiffany Tang, and Bin Yu. imodels: a python package for fitting interpretable models, 5 2022.

7 Future Perspectives or Continuation of the Work

Analyzing different datasets can yield interesting results for interpretability analysis in healthcare, particularly through a comparative approach between variants of the SARS-CoV-2 pathogen. Additionally, there is room to apply feature selection methods after using SHAP and LIME, with the goal of comparing the performance metrics of the models. Finally, in the context of healthcare applications, sharing the results obtained with medical teams specializing in the problem or those interested in applying machine learning methods to medicine could be valuable for studying the impact of interpretability in this field of research.

8 Other University-Related Activities

The student was selected to receive the BRAFITEC Scholarship for pursuing a Double Degree at CentraleSupélec in France, starting from the second semester of 2023. This program offers the recipient a generalist education during the first year of studies abroad and, during the second year, allows them to continue their studies in statistics, operations research, and machine learning, through the specialization (Master’s) in Artificial Intelligence.

9 Support

In addition to the advisor and the scholarship granted by the Institutional Scientific Initiation Scholarship Program (PIBIC), this Scientific Initiation project benefited from the collaboration of other faculty members and graduate students from Unicamp who are involved in research related to the topic and are affiliated with BI0S and the Data Analysis and Decision Support Laboratory (LAD2).

10 Acknowledgments

I would like to thank Prof. Dr. Cristiano Torezzan for his availability and interest in the guidance provided, the Faculty of Applied Sciences at UNICAMP, LAD2 for the infrastructure, and PIBIC for the support offered.