UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

GUILHERME B. B. O. MALTA

# An interpretable machine learning approach for predicting sleep quality in three temporal waves throughout the COVID-19 pandemic

Work presented in partial fulfillment of the requirements for the degree of Bachelor in Computer Science

Advisor: Profª. Drª. Mariana Recamonde Mendoza
Coadvisor: Prof. Dr. Ives C. Passos

Porto Alegre
May 2022

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitor: Prof. Carlos André Bulhões
Vice-Reitora: Prof$^{a}$. Patricia Pranke
Pró-Reitora de Graduação: Prof$^{a}$. Cíntia Inês Boll
Diretora do Instituto de Informática: Prof$^{a}$. Carla Maria Dal Sasso Freitas
Coordenador do Curso de Ciência de Computação: Prof. Rodrigo Machado
Bibliotecária-chefe do Instituto de Informática: Beatriz Regina Bastos Haro

## AGRADECIMENTOS

Agradeço à Profª Dra. Mariana, minha orientadora neste trabalho e na minha trajetória até aqui, por todo o suporte e mentoria. Muitos dos aprendizados que levarei deste bacharelado são fruto do seu constante empenho como professora e orientadora. Agradeço também ao Prof. Dr. Ives C. Passos e ao Dr. Thyago Antonelli por terem possibilitado este trabalho e me apoiado com sua experiência e dedicação à área. Agradeço à minha família, pelos incentivos e apoio que sempre me deram e que me possibilitaram chegar até aqui, obrigado por sempre confiarem em mim e estarem por perto - mesmo longe. A todos os meus amigos que partilharam da mesma trilha que eu nestes últimos anos, em especial ao Guilherme Bazzo e Nícolas Duranti, o convívio, aprendizados e apoio de vocês foi essencial para a construção da minha formação - como profissional e pessoa. Ao meu amigo Pedro Gyrao, pelo papel fundamental na minha construção profissional e pelas inúmeras conversas sempre muito enriquecedoras. À Rafaela, por termos dividido tantos momentos de dificuldades, de conquistas e tantas térmicas de café desde o começo dessa jornada. Obrigado por estar por perto e me ensinar a força do coração e da mente caminhando juntos, seu amor e parceria foram essenciais nessa história. A todos que não citei mas conviveram comigo nessa trajetória, deixo meus sinceros agradecimentos por terem feito a diferença nessa construção.

# ABSTRACT

The COVID-19 pandemic has changed life quality globally, impacting aspects such as mental health and sleep quality. Although it is known that sleep quality can be associated with traumatic experiences, anxiety and depression symptoms, physical activities, and social and economic struggles, studies reported a non-uniform effect of these factors in the population during the COVID-19 pandemic. Additionally, most of the related studies used classical statistical analysis to investigate the association between sleep quality and covariates. Using a machine learning (ML) approach, this work aims to assess the most relevant variables to describe sleep quality in three different waves during the first six months of social distancing in Brazil. Our sample is composed of 1559 volunteers that filled the three phases of a web survey with questions divided into several subgroups (sociodemographics, COVID-19 exposure, information vehicles, social distancing, mental health protection, mental health variables, anxiety, depression, and suicidal ideation), originating 111 variables. We trained classifiers by testing different balancing methods (downsampling, SMOTE, and no resampling) and different classification algorithms (Naïve Bayes, Random Forest, and Gradient Boosting Machine) within a cross-validation process. Models' explainability was explored using the SHAP framework. The best classifiers for each wave were fitted using Naïve Bayes and downsampling. The results for wave 1 (W1) were PR-AUC: 0.589, Sensitivity (Sens): 0.726, Specificity (Spec): 0.660; for wave 2 (W2) were PR-AUC: 0.586, Sens: 0.771, Spec: 0.628; and, for wave 3 (W2) were PR-AUC: 0.531, Sens: 0.836, Spec: 0.636. The most important variables for the three waves were overall related to anxiety disorder symptoms (GAD) and depression symptoms (PHQ). In W1, leisure activities and family relationships were also relevant for predicting sleep quality. The results from SHAP analysis suggested that in W1, a period closer to the beginning of social distancing measures, the relationship between variables was complex and varied significantly among the individuals, except for more extreme cases where GAD and PHQ symptoms held higher importance in predictions. For W2 and W3, bad and good sleep quality were more directly related to the high and low prevalence of anxiety and depressive symptoms. Thus, our results assist in identifying the most relevant variables for predicting sleep quality during the COVID-19 pandemic and highlight how the variables' associations evolved over a social distancing period, indicating a much more unstable scenario in W1 compared to W3.

**Keywords:** COVID-19. machine learning. interpretability. sleep quality. pandemic.

**Uma abordagem de aprendizado de máquina interpretável para prever a qualidade do sono em três ondas durante a pandemia de COVID-19**

## RESUMO

A pandemia de COVID-19 mudou a qualidade de vida globalmente, impactando aspectos como a saúde mental e qualidade do sono. Apesar de já se saber que a qualidade do sono pode ser associada com experiências traumáticas, sintomas de ansiedade e depressão, prática de atividades físicas e problemas econômicos e sociais, estudos mostram um efeito não-uniforme destes fatores na população durante a pandemia de COVID-19. Além disso, a maioria dos estudos abordou a temática através de análise estatística clássica para investigar a associação entre qualidade do sono e covariáveis. Através de uma abordagem de aprendizado de máquina, este trabalho tem como objetivo avaliar as variáveis mais importantes para descrever qualidade do sono em três ondas de coleta de dados durante os primeiros seis meses de distanciamento social no Brasil. A amostra usada neste estudo é composta de 1559 voluntários que preencheram as três etapas de um questionário online, com questões divididas entre subgrupos (sociodemográficas, exposição ao COVID-19, veículos de informação, distanciamento social, proteção à saúde mental, variáveis de saúde mental, ansiedade, depressão e ideação suicida), originando 111 variáveis. Treinamos classificadores testando diferentes algoritmos de balanceamento (*downsampling*, *SMOTE* e sem balanceamento) e diferentes algoritmos de classificação (*Naïve Bayes*, *Random Forest* e *Gradient Boosting Machine*) através de um processo de validação cruzada. A explicabilidade dos modelos foi explorada usando o *framework* SHAP. Os melhores classificadores para cada onda foram treinados usando o algoritmo *Naïve Bayes* e o método de balanceamento *downsampling*. Os resultados para a onda 1 (W1) foram PR-AUC: 0.589, Sensibilidade (Sens): 0.726, Especificidade (Espec): 0.660; para a onda 2 (W2) foram PR-AUC: 0.586, Sens: 0.771, Espec: 0.628; para a onda 3 (W3) foram PR-AUC: 0.531, Sens: 0.836, Espec: 0.636. As variáveis mais importantes para as três ondas foram, de forma geral, relacionadas a sintomas de distúrbios de ansiedade (GAD) e sintomas de depressão (PHQ). Na W1, atividades de lazer e relacionamento familiar também foram relevantes para a predição de qualidade de sono. Os resultados da análise SHAP sugerem que na W1, um período próximo ao início de medidas de distanciamento social, a relação entre as variáveis foi mais complexa e variou significativamente entre os indivíduos, exceto para casos mais extremos onde sintomas de GAD e PHQ possuíram uma importância maior nas predições. Para W2 e W3, uma qualidade de sono boa e ruim foram mais diretamente relacionadas à baixa e alta prevalência, respectivamente, de sintomas de GAD e PHQ. Portanto, nossos resultados contribuem para a identificação das variáveis mais relevantes para predição de qualidade do sono durante a pandemia de COVID-19 e destacam como as associações entre as variáveis evoluíram durante um período de distanciamento social, indicando um cenário muito mais instável na W1 em comparação à W3.

**Palavras-chave:** COVID-19. aprendizado de máquina. interpretabilidade. qualidade do sono. pandemia.

# LIST OF FIGURES

# LIST OF TABLES

## LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| ML | Machine Learning |
| W1 | Wave 1 |
| W2 | Wave 2 |
| W3 | Wave 3 |
| GAD | Machine |
| PHQ | Machine |
| PSQI | Pittsburgh Sleep Quality Index |
| RF | Random Forest |
| GBM | Gradient Boosting Machine |
| SHAP | SHapley Additive exPlanations |
| WHO | World Health Organization |
| PTSD | Post-traumatic stress disorder |
| CV | Cross-validation |
| TPR | True Positive Rate |
| TP | True Positive |
| TN | True Negative |
| FPR | False Positive Rate |
| RFE | Recursive Feature Elimination |
| AUC | Area Under Curve |
| ROC | Receiver operating characteristic |
| PR | Precision-Recall |
| KNN | K-Nearest Neighbors |

## CONTENTS

# 1 INTRODUCTION

The COVID-19 outbreak had its first known case in December 2019, and the World Health Organization (WHO) declared it a pandemic on March 11th, 2020 (CUCINOTTA; VANELLI, 2020). Since then, the WHO reported an increase in cases and deaths in several countries worldwide (WHO, 2021), especially in Brazil (BRASIL, 2021), where in July 2020, the number of cases was already the second largest in the world (CANDIDO et al., 2020). In order to reduce the spread and mortality rate related to the pandemic, several countries adopted social distancing and lockdown strategies. Although this is a necessary measure to contain the spread of the disease, together with the unpredictability of the pandemic's evolution, lockdown and physical distancing can lead to several risk factors that can affect mental health.

Most surveys conducted on the general public showed increased symptoms of depression, anxiety, and stress related to COVID-19 due to psychosocial stressors such as life disruption, fear of illness, or fear of adverse economic effects (MORENO et al., 2020). Given the extensive period that lockdown and social distancing measures were made necessary during the COVID-19 pandemic, identifying risk factors for mental health in people's daily lives becomes even more critical to prevent the evolution of severe psychiatric disorders. One key aspect in the preservation of mental health is sleep quality. Studies reported that poor sleep quality is associated with several mental health difficulties, such as post-traumatic stress, eating disorders, psychosis spectrum experiences, and that people with insomnia are much more likely to experience clinical depression and anxiety (SCOTT et al., 2021). On the contrary, sleep quality improvements can significantly reduce depression, anxiety, rumination, stress, and psychosis symptoms (TANG et al., 2017).

However, while the social distancing provided some flexibility of social schedules for a vast part of the population, which could have contributed to a slight increase in sleep time, some reports highlight a worsening sleep quality in population groups during the COVID-19 pandemic. This decrease in sleeping quality can be related to social jetlag, a mismatch between external social time and internal biological time, and sleep restrictions (BLUME; SCHMIDT; CAJOCHEN, 2020), but it appears to be a non-uniform effect over different population subgroups during the pandemic. Researchers found profound differences in the effect of lockdown on perceived sleep quality related to pre-pandemic sleep profile, environment, social behavior, socioeconomic status, marital status, race, social loneliness, anxiety symptoms, depressive symptoms, and others (KOCEVSKA et al., 2020). This non-uniformity enforces the importance of understanding which factors contribute to a change in sleep quality during the COVID-19 pandemic and, more importantly, how these factors contribute.

To further investigate how the COVID-19 pandemic and social distancing or quarantine affected the general population's mental health, this work aims to analyze which factors can explain different levels of sleep quality during the COVID-19 pandemic and how exactly they affect sleep quality using a machine learning approach. We perform a secondary analysis of data collected through a longitudinal online research in Brazil composed of three waves of evaluation, in a study led by Prof. Dr. Ives Cavalcante Passos. The three waves correspond simply to three different temporal windows and have no relationship with the COVID-19 reported waves. The data collected provides information about suicidal ideation, the prevalence and impact of depression, anxiety, post-traumatic stress disorder (PTSD) symptoms, and substances use. It also provides information on the impact of protective factors on mental health, such as physical activities, leisure activities, meditation, and sleep quality.

Machine learning has proven itself a valuable tool in the COVID-19 pandemic, aiding in various perspectives associated with the pandemic, from a better understanding of the pattern of viral spread to a more precise diagnostic or identification of risk factors. (LALMUANAWMA; HUSSAIN; CHHAKCHHUAK, 2020; SCHAAR et al., 2021). Although machine learning has increased as a mining strategy in healthcare, there is still a justified concern because of the "black box" nature of various algorithms (PETCH; DI; NELSON, 2021). In machine learning, the term "black box" refers to complex models that are not sufficiently interpretable by humans on themselves. Since many decisions on healthcare can profoundly impact human lives, it is critical to interpreting which factors led to the automated decision provided by machine learning models. This concept is called explainability in machine learning.

In this work, our goal is to achieve an explainable binary classification model to identify poor or standard sleep quality. We compare the performance of three classic machine learning algorithms and select the best to apply the Shapley Additive Explanations (SHAP) framework for predictions interpretation (LUNDBERG; LEE, 2017). With the results found in this work, we aim to provide more information on (i) how to identify poor sleep quality during a health crisis such as the COVID-19 pandemic; (ii) and explain how the most relevant variables on sleep quality prediction affect the classifiers' decision.

Particularly, the second objective can provide relevant information to the non-uniformity on sleeping patterns observed during the COVID-19 pandemic.

This work is structured as follows. In Chapter 2, we explain the theoretical background needed to understand the proposal. In Chapter 3, we analyze the related works, discussing the most common approaches to study sleep quality during the COVID-19 pandemic and highlighting the relevance of the present study. In Chapter 4, we describe our work's methodology. In chapter 5, we present the results for our experiments. In Chapter 6, we present the conclusion from our findings in this study, and, in Chapter 7 we briefly discuss opportunities for futures works.

# 2 COMPUTATIONAL BACKGROUND

This work approaches the presented problem through known machine learning (ML) algorithms and data pre-processing techniques used for supervised learning problems. With the increase of data volume and computational processing capabilities, ML has gained popularity as a data mining methodology. ML is a subfield of artificial intelligence that uses data sets to build a representation model of an approximate function or set of rules. Although several learning paradigms exist, such as supervised, unsupervised, and semi-supervised learning, in this chapter we will focus on supervised learning algorithms, which is the paradigm explored in our work. Specifically, we review the ML algorithms applied in this work: Naïve Bayes, Random Forest, and Gradient Boosting Machine. Lastly, we discuss the Shapley Additive Explanations unified framework to support the interpretation of the model's predictions on sleep quality.

## 2.1 Supervised Learning

Supervised learning is the subfield of ML that makes use of labeled data to develop predictive models. A supervised algorithm expects a set of instances in the format $(X_i, y_i)$, where $X$ is a subset of predictors variables (features) and $y$ is the target variable. The target variable defines whether the problem is a classification or regression one. If the target variable has a numerical nature, it is a regression problem. If the target variable has a categorical nature, it is a classification problem.

In this work, we use three ML algorithms to train an ML classifier for sleep quality prediction:

(i) *Naïve Bayes*: a probabilistic ML algorithm exploring Bayesian Statistics concepts to create a classification model. Probabilistic methods such as Naïve Bayes can present advantages when data is incomplete or imprecise. The Naïve Bayes classifier assumes that the feature's values for an instance are independent among each other given a class $y_i$. This independence assumption is the reason the algorithm is called naive.

By considering the independence over all the features, the probability $P(y_i(X))$ is proportional to

$$P(y_i)\prod_{j=1}^{d} P(X^j|y_i)$$

because $P(X|y_i)$ can be decomposed in the product $P(X^1|y_i) * ... * P(X^d|y_i)$ for a given instance with $d$ independent features.

(ii) *Random Forest*: the Random Forest algorithm is an ensemble method for decision trees that uses randomness selection for features and instances to improve the generalization power of the final model.

This algorithm generates several decision trees using bootstrap aggregation (bagging) and random selection of features for each node split in the decision trees. These methods are responsible for improving the variability of each decision tree and, therefore, the generalization power of the final model. For classification tasks, a majority voting process gives the final model's output from all trees for a given instance $x$.

In bagging, each decision tree is generated using a sample with replacement from the training data with the same size as the complete training data. For the random selection of features in each node split, one feature is selected from a random subset of the features in the training data. The features' subset size for node split is one of the commonly used hyperparameters for the Random Forest algorithm. The second hyperparameter is the number of trees in the forest.

(iii) *Gradient Boosting Machine*: GBM combines decision trees with boosting to improve weak classifiers. Boosting in ML consists in iteratively generating classifiers with different weights for each instance from the training data, associating higher weights for wrongly classified instances in the previous iteration and lower weights for correctly classified instances in the previous iteration.

## 2.2 Feature selection

In supervised ML problems, it often is helpful to remove features from the original dataset. This step is performed because features can be useless or redundant to the model training. Moreover, training an accurate ML model with a smaller subset of variables can improve interpretability and be easier applied in real-life applications since data collection can be a complex process.

Since features can have complex relationships, feature selection is usually implemented through specific algorithms in a ML experiment pipeline. These algorithms are classified into filter, embedded, and wrapper methods. Filter methods rank the features according to the degree of correlation with the target variable and focus on each feature independently. Embedded methods are executed during the model training process and require more computational power than filter methods. Wrapper methods usually require even more computational power than filter and embedded methods and are based on the performance assessment for the classifier of interest over different subsets of features (ARTUR, 2021).

## 2.3 Class Balancing Methods

A common issue for classification tasks in some application areas, such as healthcare, is unbalanced data. Unbalanced data in ML means that a subset of classes appears more frequently in a subset of instances than instances from other classes. In the three data sets used for this work, we can see unbalanced data for all of them. For W1 data, we have only 32.7% from class "Bad Sleep." For W2 data, 31.8% from class "Bad Sleep," and for W3, 24.8% from class "Bad Sleep." Therefore, unbalancing data is a relevant topic for this work.

The problem with unbalanced data is that various ML algorithms tend to favor the classification for new instances in the majority class. Many techniques have been presented over the years to handle this problem, and this work focuses on sampling methods to balance the number of instances from each class, namely downsampling and Synthetic Minority over-sampling TEchnique (SMOTE) (CHAWLA et al., 2002).

Downsampling is a method that randomly samples without replacement data from the majority class until it makes even with the minority class. Although this is an efficient way of balancing classes, removing instances from the data set can harm the model performance. SMOTE (CHAWLA et al., 2002) is a method that artificially generate new instances to balance data by combining both downsampling from the majority class with a special method of oversampling from the minority class. The oversampling is performed by creating artificial instances that assume values based on the k-nearest neighbors from existing instances. The features' values are calculated by taking the difference between the original instance's feature vector and its nearest neighbor, that can be selected randomly depending on the oversampling required. This difference is than multiplied by a random value between 0 and 1 and add to the new feature value. This causes the new artificial instance to be placed near but not exactly in the same coordinate of the decision region and contributes to a more general classifier.

## 2.4 Cross-validation

A popular strategy for assessing a model's performance is the k-fold cross-validation (CV) method. In K-fold CV, the instances from training data are equally distributed over K subsets with a random sample without replacement strategy. Then, in an iterative process that executes for K rounds, the model is trained with data from K-1 folds and has the performance assessed with the one fold left out from training. At the end of the process, it is possible to calculate the average performance of the model.

In this work, we use an also popular variation called stratified K-fold CV, where the number of instances from each class in each fold follows the class distribution in the original data set.

## 2.5 Evaluation Metrics

Evaluation metrics in ML have the purpose of assessing the model's performance regarding incorrect and correct predictions. Various metrics may be used for different purposes, such as understanding not only the overall performance but also the performance per class. In this work, we consider the model's performance as the capability of the model to predict instances for each class correctly. We adopted four metrics for performance assessment:

- Sensitivity: also known as True Positive Rate (TPR) or Recall, refers to the proportion of predicted True Positives (TP) instances over all positive instances in the dataset.
- Specificity: also known as True Negative Rate, refers to the proportion of True Negatives (TN) over Predicted Negative instances.
- The Area Under Receiver Operating Characteristic Curve (AUC ROC): the ROC curve provides information about the TPR versus FPR for various thresholds settings for a classifier. The AUC ROC is the area under this curve and represents the model's capability to correctly predict instances from the positive class over instances from the negative class.
- The Area Under Precision-Recall Curve (AUC PR): follows the same logic of AUC ROC but compares TPR versus Precision (the proportion of TP over Predicted Positive instances).

## 2.6 SHAP Values

Especially in some application areas such as healthcare, interpreting how a ML classifier performs its predictions can be crucial for the practical use and for knowledge extraction in the domain. The Shapley Additive Explanations (SHAP) Values method addresses interpretability on complex ML models. The SHAP framework enables a feature importance assessment for each feature contributing to a given prediction (LUNDBERG; LEE, 2017).

SHAP provides a local explanation for a given model's prediction, attributing a change value for each conditioned feature from the expected model prediction. Combining the local explanations for a set of instances with a global visualization of the instances explanation results makes it possible to assess the average behavior of a given feature's values on impacting the outcome.

The local explanation value is assessed through an explanation model that matches the following equation:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i^{'}$$

where $M$ is the number of input features, usually simplified; $z^{'} \in \{0,1\}^M$, and $\phi_i \in R$ is the local explanation value for the simplified feature value $i$, $i > 0$. $\phi_0$ is the base explanation value. The SHAP Values framework provides a solution to the presented equation.

# 3 RELATED WORKS

In this chapter, we analyze the related works that approach sleep quality during the COVID-19 pandemic. Our goal is to summarise the methodologies and results found by other researchers, as well as highlight the lack of explainable ML-based approaches for the poor sleep quality issue during the COVID-19 pandemic.

The works from Huang and Zhao (2020), Quel et al. (2021), and, Cobb et al. (2022) assessed covariates associated with poor sleep quality. The three works used a statistical approach, combining descriptive analysis, inferential analysis, and, logistic regression to assess how the covariates affected mental health generally, or, sleep quality specifically during the COVID-19 pandemic. The three studies were conducted on samples collected via web surveys, and we review each one of them below.

In Huang and Zhao (2020), authors assessed how generalized anxiety disorder (GAD), depressive symptoms and sleep quality were related to demographic information and COVID-19-related knowledge through a web survey applied to 7236 volunteers in China. The goal, differently from our work, was to find in a more general manner how the COVID-19 pandemic affected the mental health using statistical analyses. Specifically through chi-squared test, uni-variate, and multivariate logistic regression, the study reported a relation between healthcare workers and poor sleep quality. The study also reported an association between respondents of age < 35 years that spent more than 3 hours in average consuming COVID-19-related information with GAD symptoms. Although the regression method can be considered a ML approach for the problem, the study was conducted focusing on classical statistical analysis and only reported classical measures for association between features.

Quel et al. (2021) studied the impact of mandatory confinement on the physical activity, eating disorders risk, sleep quality, and, well-being on a Spanish sample of university students and general population connected with university students through social networks. The web survey was applied to the respondents in two distinct moments: between March, 15 and March, 31, and between April, 30 and May, 11. Such as our work, the data collection was performed using a longitudinal strategy having the first waves collected within the start of the lockdown or social distancing measures being applied globally. The sleep quality was assessed using the Pittsburgh Sleep Quality Index (PSQI), with a threshold of 5 or higher (from 0 to 21) considered as "poor sleep". The analysis was conducted using several statistical tests, such as t-test, Mann-Whitney U test, Wilcoxon signed-rank test (Z), and, ANOVA (F). Focusing on sleep quality, the study reported a worsening in sleep problems after the lockdown (before: 6.2 ± 3.5 and after: 7.2 ± 3.9, with p-value <0.001; the higher the sleep quality score, the poorer sleep quality).

We also highlight the study of Cobb et al. (2022), in which authors investigated the association between COVID-19 hardships and self-reported sleep problems of an adult sample from US. 8130 volunteers filled a web survey in 3 waves of COVID-19 - mid March 2020, late March 2020, and, late April 2020. The survey assessed sociodemographics, sleep quality, and, COVID-19 hardships variables, including childcare difficulties, job loss, and, pay cuts/hours reduction. The analysis was performed using only descriptive statistics and logistic regression. The researchers reported different covariates associated with sleep trouble for each wave. For the first wave, COVID-19 threat, losing a job, getting a pay cut, and, difficulty with childcare were associated with sleep troubles. For the second wave, only COVID-19 threat and difficulty with childcare were associated with sleep troubles. The metric used for the performance assessment was the Odds Ratio (OR).

Overall, the related studies approach the sleep quality during the COVID-19 pandemic using statistical analysis, including regression models such as logistic regression. For experiments using traditional ML algorithms, we did not find any related work approaching the sleep quality issue. Consequently, there have been no previous efforts to model the relationship between different covariates and sleep quality with models that can describe more complex patterns, such as ML algorithms, and analyze the associations uncovered by them.

# 4 METHODOLOGY

This work has its methodology divided into four sections. The first section describes the research conducted by Prof. Dr. Ives Cavalcante Passos and his group that aimed to collect data from the general population to provide information to study mental health aspects during the COVID-19 pandemic. The second section describes the preprocessing steps applied to the data to prepare it for a supervised ML experiment. The third section describes the ML training and evaluation experiment. Finally, the fourth section describes the SHAP algorithm application and how it assesses the features' contributions to sleep quality classification.

## 4.1 Data Collection

The data collection process happened through a web survey, divided into three temporal waves (0, 1 and 6 months). The first wave (W1) occurred from May 6th, 2020, to June 6th, 2020. The second wave (W2) occurred from June 6th, 2020, to July 6th, 2020, one month after W1. The third wave (W3) occurred from November 6th, 2020, to December 6th, 2020, six months after W1. The complete survey can be found as a supplementary document at (ANTONELLI-SALGADO et al., 2021).

The web survey was part of a larger project, that already published studies investigating other impacts of social distancing, such as in suicidal ideation (ANTONELLI-SALGADO et al., 2021). To be eligible for the study, volunteers had to be at least 18 years old and live in Brazil. The study was approved by the local research ethics committee and all participants signed the informed consent before answering the online questionnaires (ANTONELLI-SALGADO et al., 2021).

The respondents from W1 that wanted to participate in the subsequent waves informed their e-mail addresses and, when the time of the subsequent waves came, they received the following survey. Not all respondents participated in the three waves study, but all the respondents from W3 also filled W1 and W2 forms. The subset of respondents that filled the three waves' surveys is the group used for analysis.

The dataset extracted and processed from the survey has 111 features per wave, plus the target sleep quality. Initially, the survey question assessing sleep quality had four categories: bad, regular, good, and excellent sleep quality. We binarized the variable, considering the answer "bad sleep quality" as our positive class, "bad sleep quality," and the other three categories from the original question as our negative class, "normal sleep quality." Table 4.1 describes the proportion of instances of each class in each of the three temporal waves. Additionally, the features are divided into nine subgroups, namely: a) sociodemographics; b) COVID-19 exposure; c) information vehicles; d) social distancing; e) mental health protection; f) mental health variables; g) anxiety; h) depression; i) suicidal ideation.

Table 4.1 – Class distribution

| Wave | Bad Sleep (Positive) | Normal Sleep (Negative) |
|---|---|---|
| 1 | 510 (32.71%) | 1049 (67.29%) |
| 2 | 497 (31.88%) | 1062 (68.12%) |
| 3 | 388 (24.89%) | 1171 (71.11%) |

This work needs to clarify sample limitations since a critical aspect of sleep quality analysis during the COVID-19 pandemic is related to the population cohort. The final data set comprises Brazilian respondents, where approximately 83% self-identify as female, and 71% are aged 19 to 41 years old. Only 11.3% of respondents total household income are below one brazilian minimum wage.

## 4.2 Data Pre-processing

Although the pre-processing data phase usually consists of several steps to transform and shape data into the expected format from ML algorithms, in this work, we focused our efforts mainly on missing values imputation and transforming textual descriptions into numerical values. The survey design already prevented missing values in nearly all features from our data set, thus, missing values imputation was necessary for features from only two groups from each wave: (i) social distancing and (ii) traumatic and stressful experiences. Since ML algorithms usually do not work appropriately with missing data, missing values imputation is a broadly discussed topic in ML.

Simple techniques such as mean/median/mode imputation or more sophisticated ML-based techniques can work well, especially in cases where the proportion of missing values for each feature is small. We performed this process in two steps in this work. For cases where it was possible to identify that the real value should be Not Applicable, a new categorical value was imputed to indicate this. For cases where we could not define the reason for missing, we used the K-Nearest Neighbors (KNN) algorithm (k=5) for imputing missing values based on the median. This case was found only on PTSD-related features and represented less than 2.25% of the data points for features with missing values. We note that in our work we used a stratified train-test split using 80% of the data for training set and 20% for test set (more details are provided in Section 4.4), and this processing created a missing data imputation model trained using only the training data. Then, we use the imputation model on both training and testing data. This way, we avoid data leakage between training and test sets. Even though the proportion of instances with missing data is low, we conducted a descriptive analysis on this subset to find potential explanations for the missing values. However, the subset does not present any particular difference compared to the rest of the dataset regarding distributions of feature values.

Moreover, we also pre-processed the textual description provided for features "Social Distancing Time" and "Traumatic Event Time". Since these features were created from a textual field in the web survey, we manually converted the phrases into numerical values representing approximately how many days passed since the event, and then applied label encoding depending on the category of time spent (0 days since the event would be labeled as 0, until 7 days since the event: 1; until 30 days: 2; until 90 days: 3; until 180 days: 4; more than 180 days: 5; never experienced: 6).

It is worth noticing that although we intend to minimize data leakage in our ML pipeline, a subtle internal data leakage may still happen since the imputation algorithm has access to all training instances to input missing values. Therefore, although we split the training set into ten folds for cross-validation, as it will be later explained (Section 4.4), the imputed data was based on information not exclusively inside the folds used for training in each iteration of the cross-validation process. We decide to follow this approach because the proportion of missing data and the relevance of the data leakage in this context are low.

### 4.3 Feature selection

As previously discussed, Feature selection is a common dimensionality reduction technique in ML pipelines that aim to remove redundant or irrelevant features. Feature selection is also relevant for interpretability since it is easier to interpret a classifier's outcome based on a smaller subset of features. Besides the theoretical purposes of feature selection, in real-life applications it can be helpful to have a good classifier based on a small set of features since the data collection process can be costly (in time, money, effort). In this work, we implement feature selection using the resampling method for the

Recursive Feature Elimination algorithm (RFE).

We select the appropriate number of features using a tolerance-based function. Instead of selecting the number of features which resulted in the best classifier, we select the one that differs in a maximum of 5% from the best result value and have the least number of features. For instance, if the best model has a Precision-Recall AUC of 0.60 with 100 features, any other model with fewer features (e.g., 10) has at least (0.95 * 0.60 =) 0.57 Precision-Recall AUC the model with fewer features would be selected. With the presented strategy, we accept the trade-off between a model slightly less accurate but with considerably fewer features.

## 4.4 Model training and evaluation

We built the same experimental setup for the three algorithms used in this work. The complete data for each wave were divided into 80% for training and 20% for testing. For hyperparameter tunning and model assessment, we used a stratified 10-fold cross-validation strategy. We first calculate the most important features using Recursive Feature Elimination for each fold, optimizing the PR-AUC using the previously described tolerance-based selection algorithm. Then, we fit the models optimizing the PR-AUC metric and selecting the best value of the respective hyperparameter for each algorithm. This process repeats for each sampling technique (no sampling, downsampling, and SMOTE).

The setup for the complete experiment consisted of the following variables:

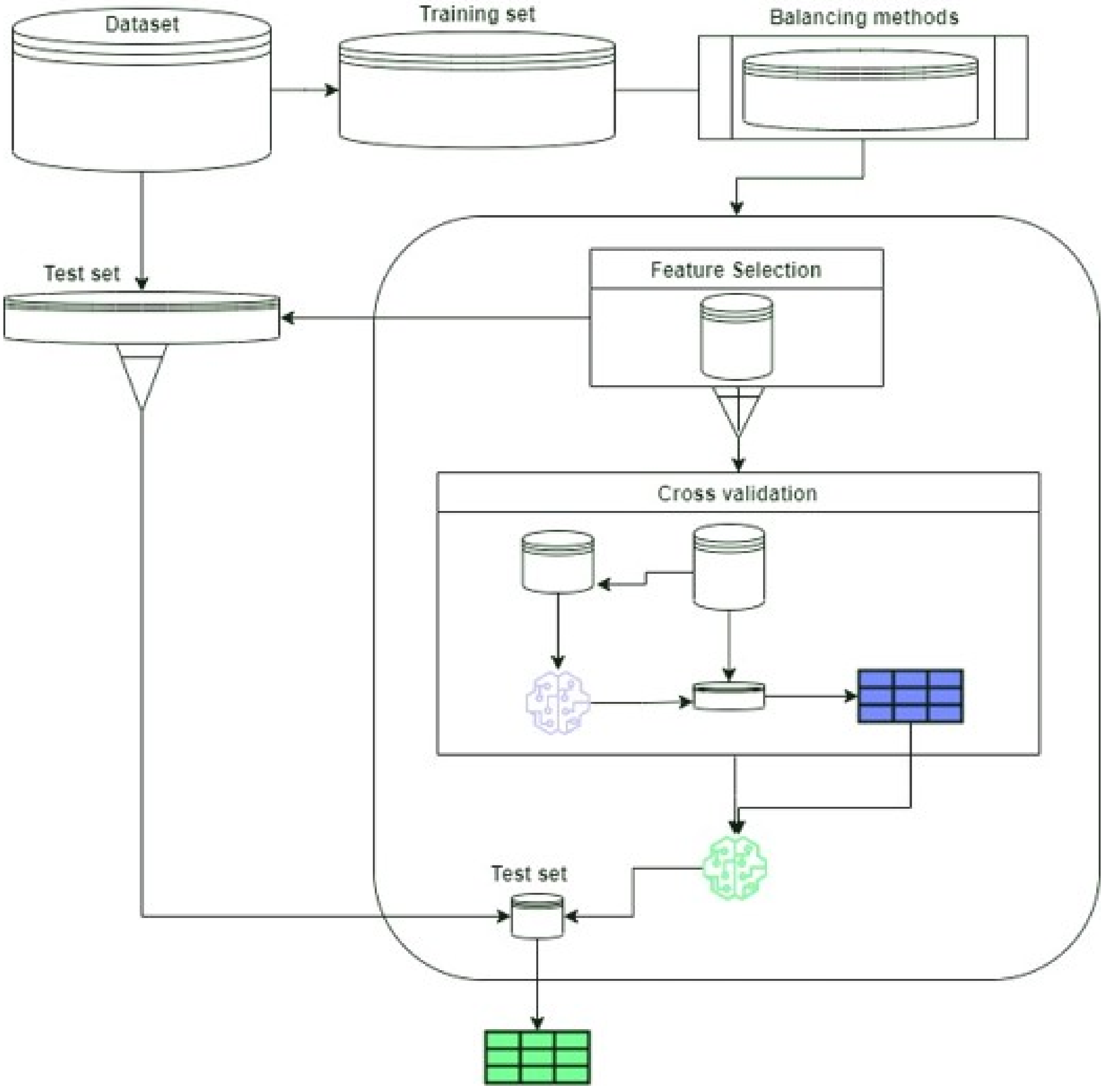
- Number of features tried on RFE: 6, 10, 14, 18, 22, 26, 30, 111
- Sampling methods: No sampling, downsampling, SMOTE
- Classification algorithms: Random Forest (RF), Gradient Boosting Machine (GBM), Naive Bayes (NB)
- Optimization metric: PR-AUC

Each classifier was optimized based on a subset of hyperparameters values. For the Random Forest classifier, the mtry value was optimized considering values in the interval $(2,...,m)$ with an increment of $\sqrt{m}$. For Naïve Bayes, the use of a Kernel density estimator, and for Gradient Boosting Machine, number of trees (10, 50, 100, 500, 1000), interaction depth (1, 2, 3), shrinkage (0.05, 0.1, 0.2), and minimum number of observations in node (5, 10, 25, 50).

After the cross-validation process for hyperparameter optimization and model validation and selection, we use the optimized model fitted from the training data to predict instances from the test set, validating the experiment configuration and confirming that no overfitting occurred during the training phase. All experiments were conducted with the R package caret.

By the end of the cross-validation process, we select the best classifier for each of the three temporal waves, considering the PR-AUC, Sensitivity, and Specificity metrics, resulting in three classifiers to serve as input to the SHAP analysis process. Figure 4.1 describes the model training and evaluation pipeline.

Figure 4.1 – Model training and evaluation pipeline.



## 4.5 SHAP Application

For each classifier selected during the model selection phase for W1, W2, and W3, we created a SHAP explainer using the R package shapper (MAKSYMIUK ALICJA GOSIEWSKA, 2020). To create the SHAP explainer, we used each wave's correspondent pre-processed training set and applied the explainer to each instance from the training set to extract its SHAP value. The individual SHAP values populate a tabular structured data that we use to build the global SHAP plots manually. This methodology allows us to use the same processing pipeline regardless of the classifier algorithm.

We apply one extra normalization step in the features' values to build the global SHAP plots. Although the classification algorithms used in this work deal well with categorical values, the global SHAP plot expects feature values within the same range to allow a fair comparison over different features. Since the features used for this experiment were all ordinal and originally pre-processed using label encoding, applying one-hot-encoding or other data transformation techniques was unnecessary.

# 5 RESULTS

We present our results by dividing our discussion into a feature selection section, model training and testing section, and SHAP analysis section. Since in this work, we have a binary classification problem where errors in one of the classes can be more damaging than in the other one, we optimized the PR AUC value for both feature selection and model training pipelines and considered Sensitivity as well for choosing the best model. This means that if two models have a similar performance considering PR AUC, the best model will be the one with a higher Sensitivity (i.e., the model that results in less errors on identifying Bad Sleep quality).

## 5.1 Feature Selection

The feature selection process assessed the minimum subset with the most relevant features for predicting sleep quality. To select the best subset of features, we considered as the primary metric the PR AUC value and secondary metrics the Sensitivity and Specificity. Therefore, the theoretical best model for our purposes would have a value of 1 for these three metrics.

Table 5.1 – Example of feature selection choice

| Variables | PR | Sensitivity | Specificity | PRSD | SensitivitySD | SpecificitySD | Selected |
|---|---|---|---|---|---|---|---|
| 6 | 0.5278 | 0.6759 | 0.6565 | 0.0916 | 0.0649 | 0.0560 | |
| 10 | 0.5392 | 0.6712 | 0.6482 | 0.0859 | 0.0579 | 0.0671 | * |
| 14 | 0.5590 | 0.6763 | 0.6494 | 0.0760 | 0.0973 | 0.0557 | |
| 18 | 0.5570 | 0.7012 | 0.6447 | 0.0780 | 0.0809 | 0.0592 | |
| 22 | 0.5454 | 0.6837 | 0.6729 | 0.0793 | 0.0750 | 0.0492 | |
| 26 | 0.5641 | 0.6511 | 0.6706 | 0.0856 | 0.0623 | 0.0679 | |
| 30 | 0.5447 | 0.6633 | 0.6800 | 0.0756 | 0.0740 | 0.0562 | |
| 111 | 0.5577 | 0.6709 | 0.6647 | 0.0730 | 0.0809 | 0.0700 | |

Table 5.1 shows an example of feature selection choice during the process, considering the optimization metric and the tolerance-based choice function. We can see that the selected subset is not the one with the highest PR value (0.5641), but the minimum subset with at least $0.95 * 0.5641 = 0.5358$ of PR score. We reduced the subset of variables used to train the models from 111 to 18, 10, and 6 for each of the three waves, respectively, using the selection rule mentioned above.

Table 5.2 – Selected features

| **W1: M=18** | **W2: M=10** | **W3: M=6** |
|---|---|---|
| GAD_1 | GAD_1 | GAD_1 |
| GAD_2 | GAD_2 | |
| GAD_3 | GAD_3 | |
| GAD_4 | GAD_4 | GAD_4 |
| GAD_5 | | |
| GAD_6 | GAD_6 | GAD_6 |
| GAD_7 | GAD_7 | |
| PHQ_1 | PHQ_1 | PHQ_1 |
| PHQ_2 | PHQ_2 | PHQ_2 |
| PHQ_4 | PHQ_4 | PHQ_4 |
| PHQ_5 | PHQ_5 | |
| PHQ_6 | | |
| PHQ_7 | | |
| PHQ_8 | | |
| UCLA_2 | | |
| UCLA_3 | | |
| leisure_activities | | |
| family_relationships | | |

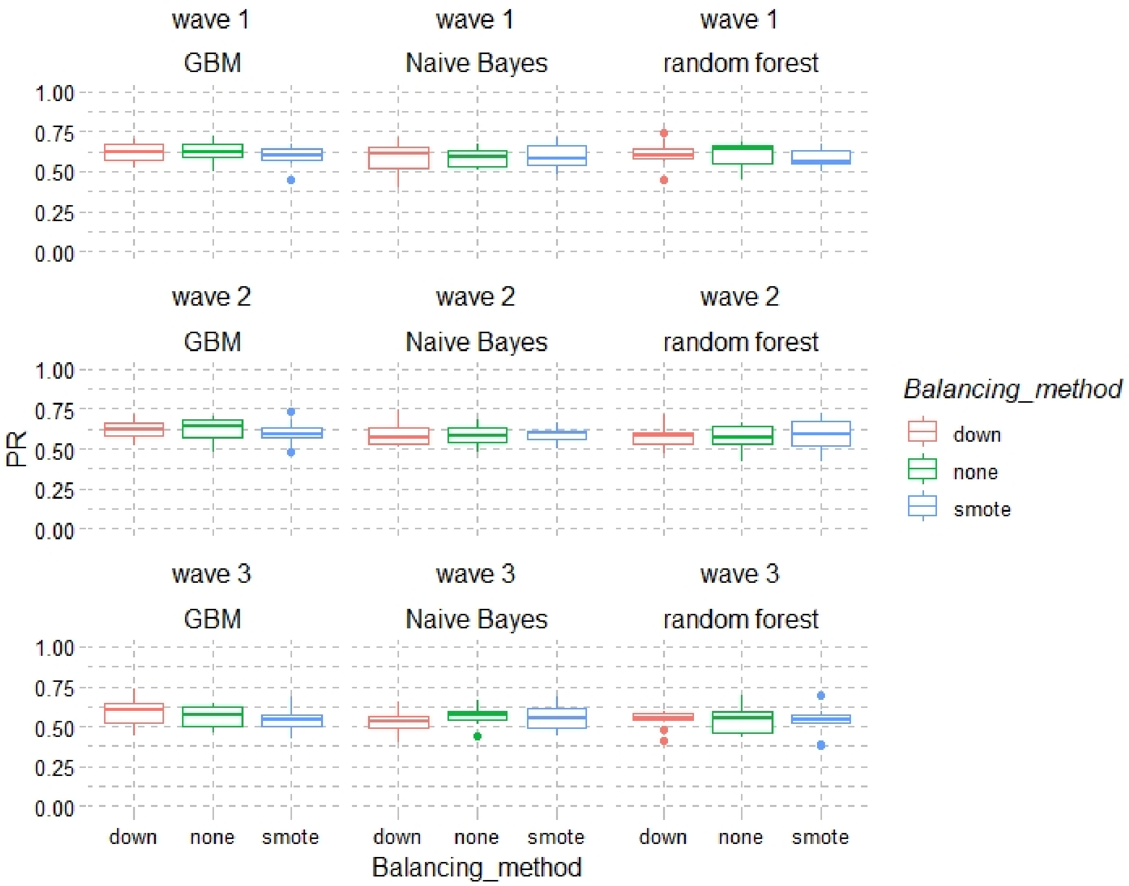
## 5.2 Models performance assessment

A strong classifier, i.e., a classifier that performs much better than a random guess, can provide higher assurance that the selected features used to train this model confidently describe a good enough behavior estimate from the instances in a real-life scenario. This is a crucial characteristic to perform feature importance and interpretability analysis afterward. As presented in Figure 5.1, our final results for the classifiers show no significant difference considering the primary metric PR-AUC. Regardless of the algorithm of balancing method use, median values are between 0.500 and 0.625 for all models in the three waves analyzed.

We present the exact mean and standard deviation from the cross-validation training in Table 5.3. Given the small variation in the PR-AUC values among distinct approaches, to select the best model for each wave, we also considered the secondary metrics, Sensitivity, and Specificity. Since for this domain, accurately predicting a bad sleep quality can have higher importance than predicting normal sleep, we would prefer a model with a Sensitivity closer to 1 in case several models have similar overall performance.

To select the best model for each wave, we used a scatter plot for Sensitivity and

Figure 5.1 – PR-AUC performance for training phase considering the different balancing methods and algorithms used for waves one, two and three.



Specificity and selected the model with the highest combination from these two metrics and considered Sensitivity's importance weight twice the Specificity's importance weight. Figures 5.2, 5.3, and 5.4 describe this logic for each of the three waves. The best model is the one closest to the green lines. This means that the model has the highest combination of Sensitivity and Specificity considering the importance weights we mentioned above.

For wave 1, the training setup for the selected model involved a GBM algorithm with a downsampling strategy (PR: 0.6167; Sensitivity: 0.7257; Specificity: 0.6774). For wave 2, Naive Bayes with downsampling (PR: 0.5864; Sensitivity: 0.7715; Specificity: 0.6282) and, for wave 3, Naive Bayes with downsampling as well (PR: 0.5317; Sensitivity: 0.8364; Specificity: 0.6360).

Figure 5.2 – Sensitivity vs Specificity scatterplot to assess model performance for Wave 1 data.

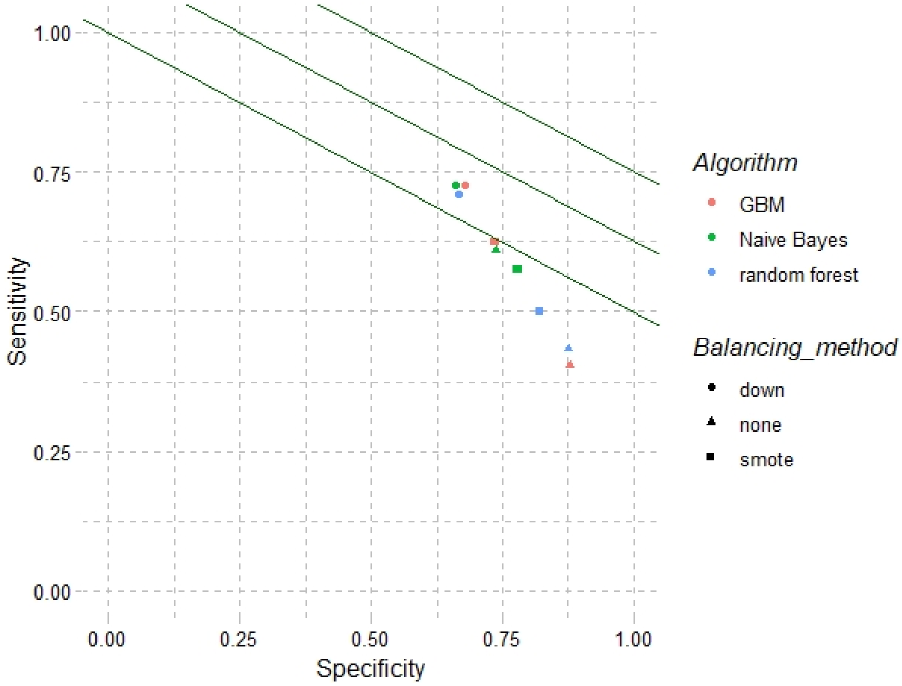


Figure 5.3 – Sensitivity vs Specificity scatterplot to assess model performance for Wave 2 data.

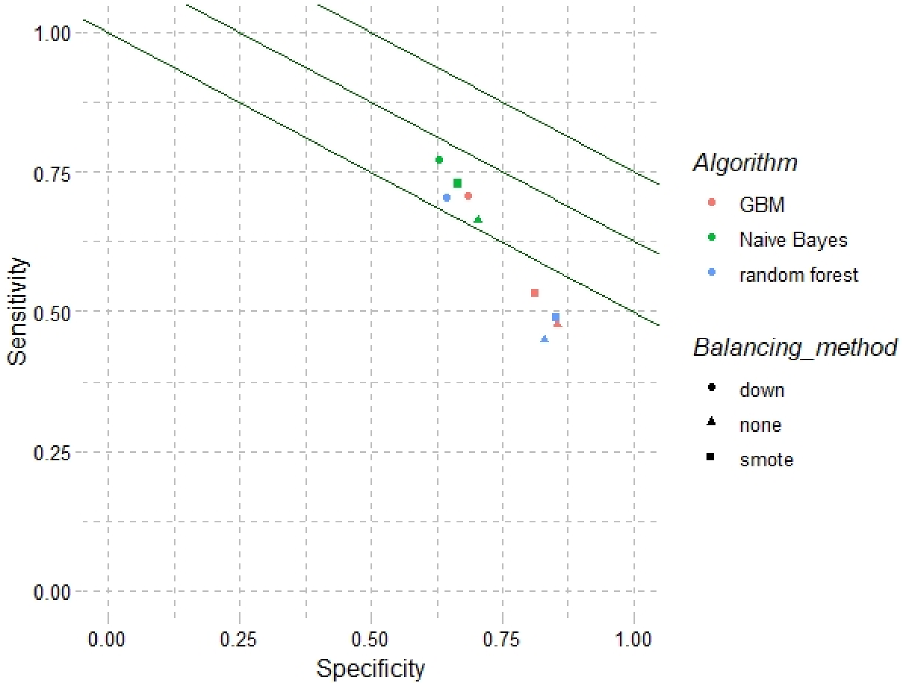


Table 5.3 – Complete results in terms of mean and standard deviation for model training. The best performance for each Wave-Algorithm pair is highlighted in boldface.

| Wave | Algorithm | Balancing method | PR | ROC | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| Wave 1 | GBM | down | **0.6167 (0.0630)** | **0.7745 (0.0351)** | **0.7257 (0.0798)** | **0.6774 (0.0414)** |
| | | none | 0.6265 (0.0671) | 0.7784 (0.0353) | 0.4043 (0.0496) | 0.8786 (0.0504) |
| | | smote | 0.5947 (0.0629) | 0.7614 (0.0316) | 0.6246 (0.0787) | 0.7357 (0.0448) |
| | Naive Bayes | down | **0.5892 (0.1004)** | **0.7601 (0.0600)** | **0.7257 (0.0552)** | **0.6595 (0.0699)** |
| | | none | 0.5847 (0.0581) | 0.7594 (0.0408) | 0.6104 (0.0563) | 0.7381 (0.0561) |
| | | smote | 0.5937 (0.0773) | 0.7651 (0.0370) | 0.5762 (0.0617) | 0.7786 (0.0524) |
| | Random Forest | down | **0.6093 (0.0792)** | **0.7649 (0.0573)** | **0.7110 (0.0690)** | **0.6667 (0.0708)** |
| | | none | 0.6095 (0.0815) | 0.7642 (0.0399) | 0.4335 (0.0652) | 0.8762 (0.0298) |
| | | smote | 0.5791 (0.0580) | 0.7549 (0.0338) | 0.4998 (0.0448) | 0.8202 (0.0382) |
| Wave 2 | GBM | down | **0.6203 (0.0662)** | **0.7723 (0.0444)** | **0.7058 (0.0583)** | **0.6847 (0.0446)** |
| | | none | 0.6249 (0.0773) | 0.7754 (0.0566) | 0.4778 (0.0941) | 0.8541 (0.0347) |
| | | smote | 0.5984 (0.0767) | 0.7605 (0.0470) | 0.5329 (0.0663) | 0.8129 (0.0619) |
| | Naive Bayes | down | **0.5864 (0.0770)** | **0.7673 (0.0423)** | **0.7715 (0.0644)** | **0.6282 (0.0518)** |
| | | none | 0.5850 (0.0707) | 0.7677 (0.0503) | 0.6632 (0.1033) | 0.7035 (0.0683) |
| | | smote | 0.5849 (0.0510) | 0.7665 (0.0301) | 0.7287 (0.0403) | 0.6659 (0.0538) |
| | Random Forest | down | **0.5796 (0.0758)** | **0.7518 (0.0378)** | **0.7035 (0.0644)** | **0.6435 (0.0404)** |
| | | none | 0.5660 (0.0880) | 0.7312 (0.0732) | 0.4497 (0.0669) | 0.8294 (0.0423) |
| | | smote | 0.5792 (0.1063) | 0.7480 (0.0614) | 0.4879 (0.0831) | 0.8529 (0.0643) |
| Wave 3 | GBM | down | **0.5833 (0.0930)** | **0.8043 (0.0405)** | **0.7359 (0.0765)** | **0.7128 (0.0372)** |
| | | none | 0.5640 (0.0672) | 0.8008 (0.0366) | 0.3958 (0.1333) | 0.9007 (0.0371) |
| | | smote | 0.5397 (0.0739) | 0.7985 (0.0475) | 0.6686 (0.0598) | 0.7566 (0.0476) |
| | Naive Bayes | down | **0.5317 (0.0767)** | **0.7992 (0.0493)** | **0.8364 (0.0703)** | **0.6360 (0.0474)** |
| | | none | 0.5615 (0.0610) | 0.8026 (0.0350) | 0.5984 (0.0690) | 0.8026 (0.0458) |
| | | smote | 0.5623 (0.0873) | 0.7994 (0.0457) | 0.7138 (0.0283) | 0.7245 (0.0727) |
| | Random Forest | down | **0.5410 (0.0577)** | **0.7989 (0.0286)** | **0.7778 (0.0544)** | **0.6958 (0.0389)** |
| | | none | 0.5415 (0.0910) | 0.7848 (0.0501) | 0.4471 (0.0774) | 0.8857 (0.0369) |
| | | smote | 0.5311 (0.0918) | 0.7787 (0.0507) | 0.4727 (0.0776) | 0.8495 (0.0372) |

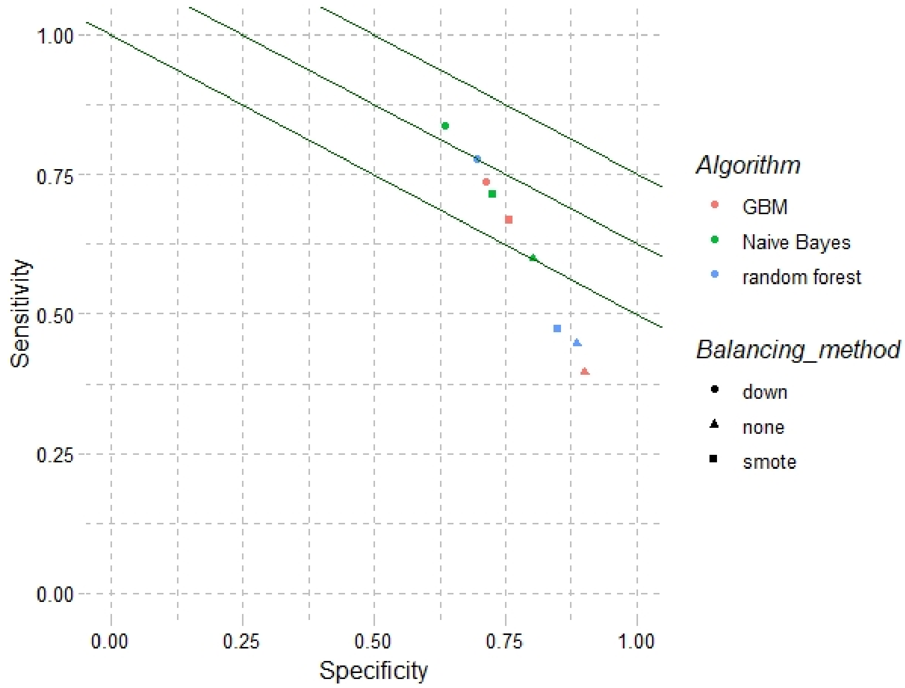
## 5.3 SHAP Values Analysis

We applied the SHAP Values framework in the training sets for the three waves to analyze how each selected feature for the final models impacts the outcome. We emphasize that while a GBM had a slightly better performance according to the Sensitivity vs Specificity scatterplot for W1, we selected the Naïve Bayes classifier as best model for this wave. This choice was based on both classifiers having roughly the same performance and the best models for W2 and W3 were based on the Naïve Bayes algorithm, therefore we can perform the SHAP analysis upon the same ML algorithm and balancing methods for the three waves.

Calculating the SHAP Value for a given instance is beneficial for explainability because it is transparent how each feature's value contributed to the class prediction. Although it is helpful to check the local interpretability, we focused our analysis on the global interpretability plot, calculating the SHAP Values for each instance and analyzing globally how each features' values contributed to the outcome - bad or normal sleep quality.

All the features reported in Figures 5.5, 5.6, and 5.7 are ordinals, meaning that a

Figure 5.4 – Sensitivity vs Specificity scatterplot to assess model performance for Wave 3 data.

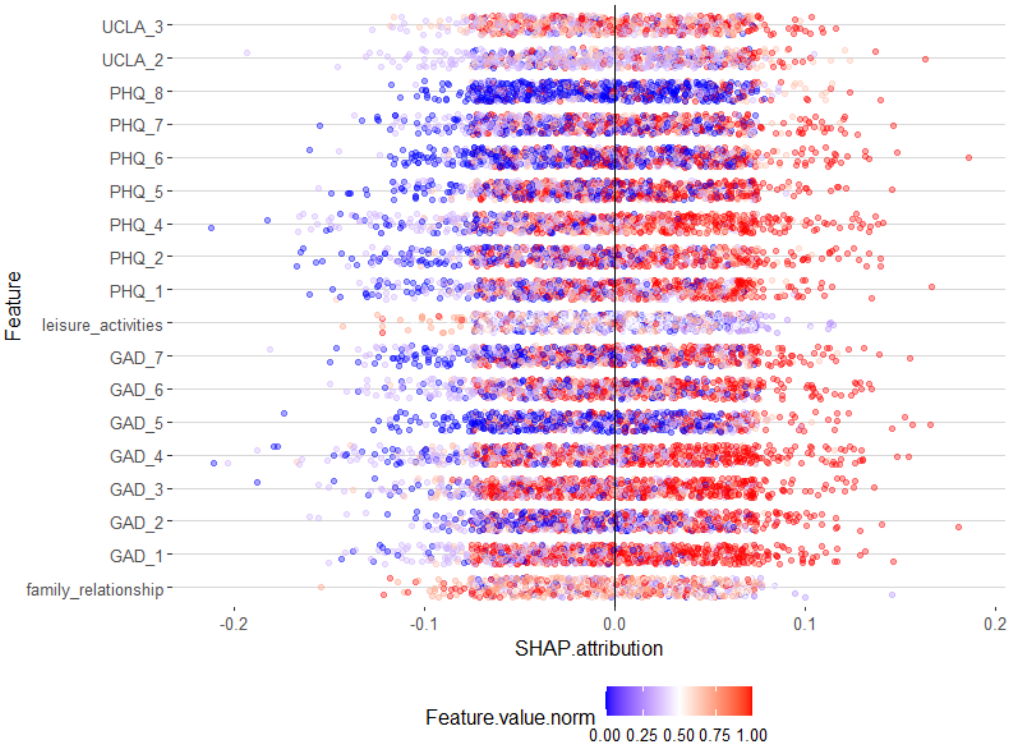


value of 0.0 (solid blue) represents a rare event or a bad relationship, and a value of 1.0 (solid red) represents a frequent event or good relationship.

At W1 (Figure 5.5), the most important features (m=18) were related to social distancing and quarantine (UCLA), anxiety (GAD), depression (PHQ), family relationship, and leisure activities. For the UCLA, GAD, and PHQ features, a lesser frequency of the feelings being questioned contributed to a lower probability of bad sleep. In counterpart, a higher frequency of the feelings raised contributed to a higher probability of bad sleep. The general contribution's behavior is less clear for leisure activities, but more extreme contribution values agree. One possible explanation for this scenario is that leisure activities themselves have less impact on the outcome than anxiety and depression symptoms. However, in cases where leisure activities have a significant impact, a higher frequency contributes to a lower probability of bad sleep while a lower frequency contributes to a higher probability of bad sleep. A similar case seems applicable to family relationships. In more significant scenarios, a bad relationship seems to contribute to a higher probability of bad sleep. However, good relationships do not contribute significantly to a lower probability of bad sleep.

At W2 (Figure 5.6), the subset of most important features decreased (m=10) in comparison to W1, and this can be an indicative of a less complex model. The features were related to anxiety (GAD) and depression (PHQ). The global SHAP plot describes a

Figure 5.5 – Analysis of SHAP values for W1.



more uniform relation over the frequency of anxiety and depression symptoms (features values) and how these values impacted the classifier's prediction (SHAP values). However, although this relation is more uniform than the W1 global SHAP plot, we can see a non-uniformity in instances having lower SHAP values.

Figure 5.6 – Analysis of SHAP values for W2.

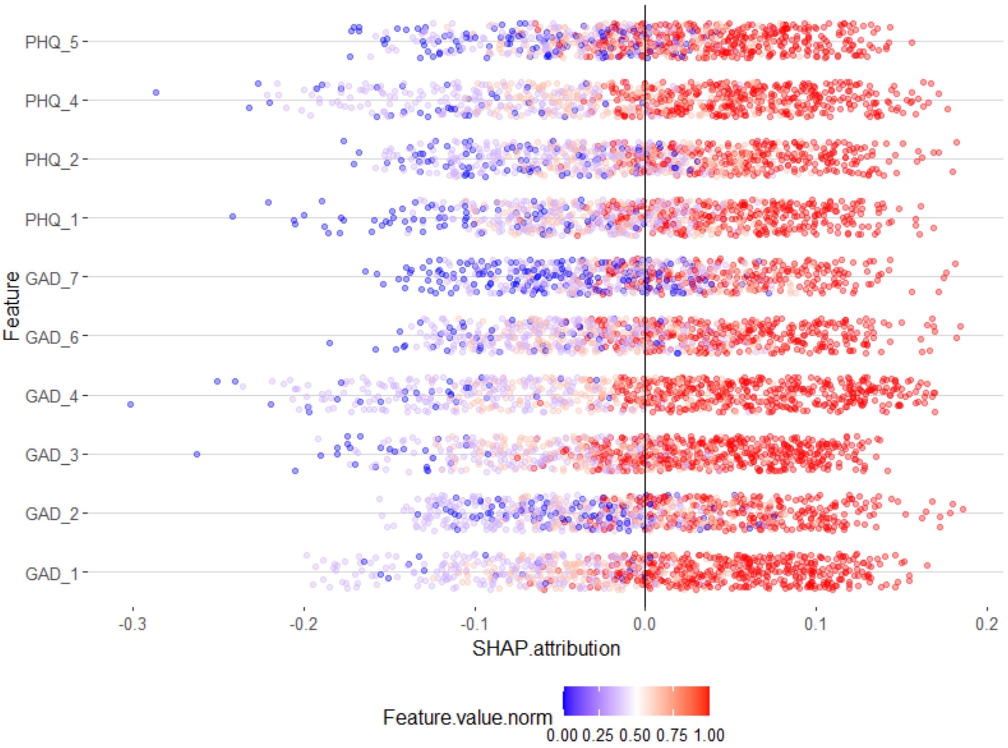
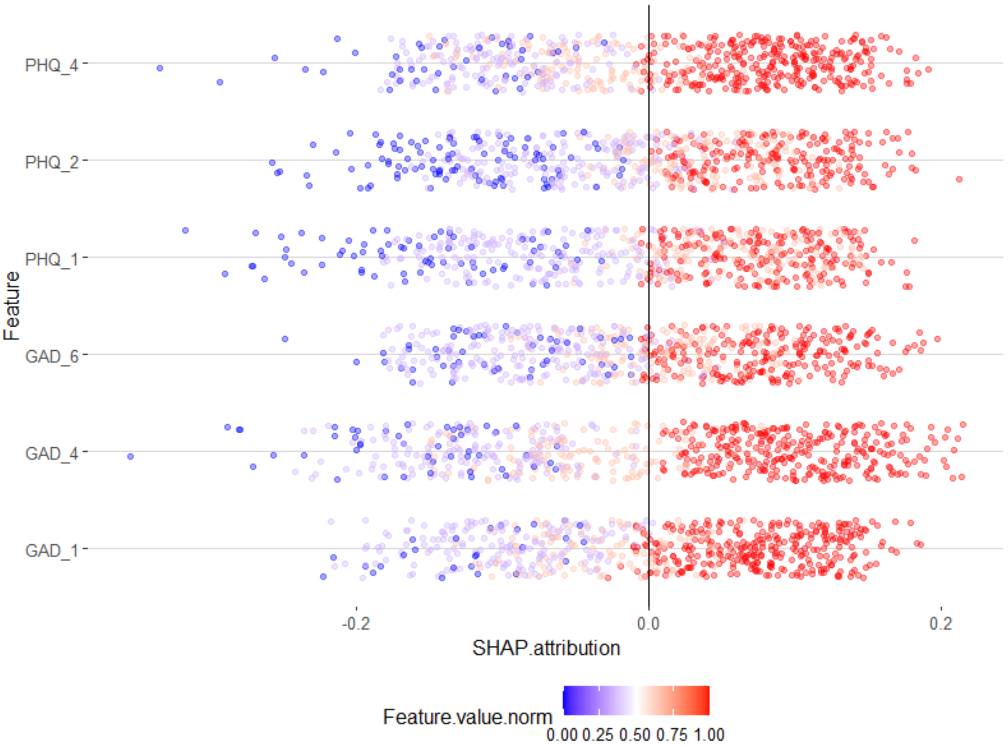


Figure 5.7 – Analysis of SHAP values for W3.



Finally, at W3 (Figure 5.7), the final model was trained with a smaller subset (m=6) as compared to W2, indicating an even less complex classifier for predicting sleep quality. The features were related to anxiety (GAD) and depression (PHQ). Compared to W1 and W2, the global SHAP plot shows a clear uniformity of anxiety and depression symptoms (features values) and how these values impacted the classifier's prediction (SHAP values). Different from W1 and W2, almost all high anxiety and high depression symptoms reported contributed to a bad sleep classification.

In summary, we observed that over time of social distancing, models trained to predict sleep quality become less complex, that is, fewer features are needed to develop a model with satisfactory performance, and the association of these features with the outcome of interest (i.e., sleep quality) becomes increasingly consistent.

# 6 CONCLUSION

This work proposed an interpretable ML approach for classification of sleep quality patterns during the COVID-19 pandemic, using the SHAP framework for interpreting predictions. We analyzed data from three temporal waves in an early stage of the COVID-19 pandemic in Brazil (from May 2020 to ..), collected through a web survey and encompassing several distinct factors that may contribute to sleep quality issues.

Previous works approaching sleep quality during the COVID-19 pandemic presented a non-uniformity regarding the leading causes of variations in sleep quality (KOCEVSKA et al., 2020), especially over different population subgroups. In addition to that, the applicability of ML approaches to this theme is still low explored since most of the works found utilizes classic statistical methods. Therefore, understanding the leading causes for this problem with more complex pattern recognition methods such as ML algorithms is still an open matter.

We selected three ML classifiers among different ML algorithms and balancing methods using cross-validation. The selected classifiers were trained with the downsampling balancing method, the Naïve Bayes algorithm, and the feature selection with RFE. These classifiers reached similar performance measures in the three temporal waves studied in this work (W1 - PR: 0.5892; Sensitivity: 0.7257; Specificity: 0.6595; W2 - PR: 0.5864; Sensitivity: 0.7715; - Specificity: 0.6282; W3 - PR: 0.5317; Sensitivity: 0.8364; Specificity: 0.6360).

By applying the SHAP framework to the classifiers and analyzing the global SHAP plot, we observed that, although the three classifiers were similar and that the most important features were generally related to anxiety and depression symptoms, the way each feature contributed to the final prediction seemed to vary significantly over the three temporal waves. The classifier for W1 was also more complex compared to W2 and W3 classifiers. In W1, we observed frequent symptoms of anxiety and depression, increasing the probability of both bad and normal sleep quality in different subgroups. At W2, we observed an increase in the uniformity of frequent anxiety and depression symptoms being related to a positive prediction for bad sleep quality, but a small sample of the studied population still diverged. At W3, we observed a clear correlation between frequent symptoms of anxiety and depression and a positive contribution to bad sleep quality prediction in all features.

This work allowed us to understand intrinsic classifier behaviors that standard ML metrics cannot point out. The three classifiers were trained using the same ML algorithm and balancing methods and reached similar performances, but the relationship between the values of the features and the outcome had significant differences that deserve further investigation.

# 7 FUTURE WORKS

Although we achieved the goal of proposing an interpretable ML approach for sleep quality during the COVID-19 pandemic in this work, the results allow future works to expand the interpretability approach in general ML problems. Applying the SHAP framework to specific subgroups of interest can allow a more interpretable analysis of how the values of the features can affect the classification. This analysis is even more critical in cases where the global SHAP plot presents a high level of overlap among the same SHAP value being affected by different features values, such as in W1.

For generalization purposes, where a subgroup of interest cannot or would not have to be specified, applying unsupervised ML algorithms using the SHAP values as features can support finding groups of instances that are not necessarily related to each other considering the values of the regular features, but are similar on how these values contribute to their classification. This would be an interesting strategy to apply to our dataset and analyze patterns at a higher granularity level, investigating regions of interest in the SHAP values plot individually and more carefully.

Lastly, the dataset used in this work is part of a longitudinal study. In this work, we focused on analyzing the three temporal waves separately, investigating and comparing patterns among the three waves. Another interesting study that could be conducted is using the temporal relation of instances from W1 and W2 to improve the classifier's training phase to predict sleep quality on W3. Although the heterogeneity among the patterns found for different waves may be a challenge for this approach, being able to predict which individuals will continue suffering from sleep quality issues in advance may be useful to develop preventive measures.

# REFERENCES

ANTONELLI-SALGADO, T. et al. Loneliness, but not social distancing, is associated with the incidence of suicidal ideation during the covid-19 outbreak: a longitudinal study. **Journal of Affective Disorders**, Elsevier B.V., v. 290, p. 52–60, 7 2021. ISSN 15732517.

ARTUR, M. Review the performance of the bernoulli naïve bayes classifier in intrusion detection systems using recursive feature elimination with cross-validated selection of the best number of features. In: . [S.l.]: Elsevier B.V., 2021. v. 190, p. 564–570. ISSN 18770509.

BLUME, C.; SCHMIDT, M. H.; CAJOCHEN, C. **Effects of the COVID-19 lockdown on human sleep and rest-activity rhythms**. [S.l.]: Cell Press, 2020. R795-R797 p.

BRASIL, M. da S. **COVID-19 Painel Coronavírus.** 2021. (Accessed October 30, 2021). Available from Internet: <https://covid.saude.gov.br/>.

CANDIDO, D. S. et al. Evolution and epidemic spread of SARS-CoV-2 in brazil. **Science**, American Association for the Advancement of Science, v. 369, n. 6508, p. 1255–1260, 2020.

CHAWLA, N. V. et al. **SMOTE: Synthetic Minority Over-sampling Technique**. 2002. 321-357 p.

COBB, R. J. et al. Covid-19 hardships and self-reported sleep quality among american adults in march and april 2020: Results from a nationally representative panel study. **Sleep Health**, 4 2022. ISSN 23527218. Available from Internet: <https://linkinghub.elsevier.com/retrieve/pii/S2352721822000079>.

CUCINOTTA, D.; VANELLI, M. Who declares COVID-19 a pandemic. **Acta Bio Medica: Atenei Parmensis**, Mattioli 1885, v. 91, n. 1, p. 157, 2020.

HUANG, Y.; ZHAO, N. Generalized anxiety disorder, depressive symptoms and sleep quality during covid-19 outbreak in china: a web-based cross-sectional survey. **Psychiatry Research**, Elsevier Ireland Ltd, v. 288, 6 2020. ISSN 18727123.

KOCEVSKA, D. et al. Sleep quality during the covid-19 pandemic: not one size fits all. **Sleep Medicine**, Elsevier B.V., v. 76, p. 86–88, 12 2020. ISSN 18785506.

LALMUANAWMA, S.; HUSSAIN, J.; CHHAKCHHUAK, L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. **Chaos, Solitons & Fractals**, Elsevier, v. 139, p. 110059, 2020.

LUNDBERG, S.; LEE, S.-I. A unified approach to interpreting model predictions. 5 2017. Available from Internet: <http://arxiv.org/abs/1705.07874>.

MAKSYMIUK ALICJA GOSIEWSKA, P. B. S. **shapper: Wrapper of Python Library 'shap'**. 2020. (Accessed November 30, 2021). Available from Internet: <https://CRAN.R-project.org/package=shapper>.

MORENO, C. et al. **How mental health care should change as a consequence of the COVID-19 pandemic**. [S.l.]: Elsevier Ltd, 2020. 813-824 p.

PETCH, J.; DI, S.; NELSON, W. Opening the black box: the promise and limitations of explainable machine learning in cardiology. **Canadian Journal of Cardiology**, Elsevier BV, 9 2021. ISSN 0828282X.

QUEL Óscar Martínez-de et al. Physical activity, dietary habits and sleep quality before and during covid-19 lockdown: A longitudinal study. **Appetite**, Academic Press, v. 158, 3 2021. ISSN 10958304.

SCHAAR, M. van der et al. How artificial intelligence and machine learning can help healthcare systems respond to COVID-19. **Machine Learning**, Springer, v. 110, n. 1, p. 1–14, 2021.

SCOTT, A. J. et al. Improving sleep quality leads to better mental health: A meta-analysis of randomised controlled trials. **Sleep Medicine Reviews**, Elsevier BV, v. 60, p. 101556, 12 2021. ISSN 10870792.

TANG, N. K. et al. Changes in sleep duration, quality, and medication use are prospectively associated with health and well-being: Analysis of the uk household longitudinal study. **Sleep**, Associated Professional Sleep Societies,LLC, v. 40, 3 2017. ISSN 15509109.

WHO. **WHO Coronavirus (COVID-19) Dashboard**. 2021. (Accessed October 30, 2021). Available from Internet: <https://covid19.who.int/>.