Article

Predicting Patient No-Show Using Machine Learning Techniques in the Healthcare Sector

Luiz Henrique A. Salazar [1], Valderi R. Q. Leithardt[2,3], Wemerson Delcio Parreira [1]* Anita M. da Rocha Fernandes [1]*, Jorge L. V. Barbosa [4] and Sérgio D. Correia [2,3]

[1] University of Vale do Itajai, 88302-901, Brazil. (L.H.A.S, W.D.P. and A.M.R.F); Email: luizhsalazar@edu.univali.br; {parreira, anita.fernandes}@univali.br
[2] VALORIZA, Research Center for Endogenous Resources Valorization, Instituto Politécnico de Portalegre, 7300-555, Portalegre, Portugal. (V.R.Q.L. and S.D.C.); E-mail: {valderi,scorreia}@ipportalegre.pt
[3] COPELABS, Universidade Lusófona de Humanidades e Tecnologias,1749-024 Lisboa, Portugal
[4] Apllied Computing Graduate Program, University of Vale do Rio dos Sinos, Av. Unisinos 950, Bairro Cristo Rei, São Leopoldo, RS, 93022-750, Brazil; Email: jbarbosa@unisinos.br (J.L.V.B)
* Correspondence: W.D.P. and A.M.R.F {parreira, anita.fernandes}@univali.br

Abstract: Today, across the most critical problems faced by hospitals and health centers are those caused by the existence of patients who do not attend their appointments. Among others, this practice generates waste of resources and increases the patients' waiting list. To handle these problems, hospitals are actively trying to implement methods to reduce the idle time caused by patient no-shows. Many scheduling systems developed require predicting whether a patient will show up for an appointment or not. Although, a challenging problem resides in obtaining these estimates precisely. The goal of this work is to analyze how objective factors influence a patient not to attending their appointment, to identify the main causes that contribute to a patient's decision, and to be able to predict whether or not the patient will attend the scheduled appointment. As a result, the obtained model is tested on a real dataset collected in a health center linked to the University of Vale do Itajaí (UNIVALI), which includes 25 features and about 5000 samples. The algorithm that produced the best results for the available dataset is the Random Forest classifier. It reveals the best recall rate (0.91), since it measures the ability of a classifier to find all the positive instances and achieves a receiver operating characteristic curve rate of 0.969.

Keywords: Artificial Intelligence, Data Science, HealthCare Applications, Machine Learning, Patient Attitudes

1. Introduction

The high rate of patients not showing up for examinations and medical appointments is a recurring problem in health care. "No-show" refers to a non-attending patient who neither uses nor cancels their medical appointments. This patients' behaviour is one of the main problems faced by health centres and has a significant impact on revenues, costs and the use of resources. Previous studies have looked at the economic consequences of the patient's absenteeism [1].

Each year, an average of about 30% of patients in the Brazilian state of Santa Catarina miss appointments, exams or scheduled surgeries. In 2018 alone, more than 52,000 patients did not attend scheduled procedures at health facilities in the state. This number represents 32.81% of the appointments given by regular centers. In absolute numbers, there were 52,710 patients who missed scheduled procedures in the first ten months of 2018, surpassing the figure for 2017, when 46,394 people failed to show up [2].

In other Brazilian states the situation is no different. In the city of Vitória, Espírito Santo, the number of absences from medical consultations in health centers reached 30% of the total performed in 2014-2015. According to the City of Vitória, the average cost of an appointment during these years was approximately 37 USD, which represented a loss of approximately 17.5 USD million for the government's coffers [3]. Given the impact that the non-use of health services causes in society, there is room for studies that present efficient solutions to this problem.

 doi:10.20944/preprints202109.0342.v1

There is a general consensus in the literature that patient nonattendance is not random, and several studies have recognized the need to statistically analyze the factors that influence the patients' no-show. Reduce the impact of missed appointments and improve health care operations are some benefits that this analysis may support. Some of the recent studies show that there is a relationship between the number of missed appointments and patients behavior [1,4]. Also, the work presented in [5] carried out a study in the field of hospital radiology and grouped the extracted data into three groups: patient, examination, and scheduling. The most informative aspects for predicting patients' non-attendance for the exam were those based on the type of exam and the scheduling attributes, such as the waiting time between the appointment and the performance of the exam.

Long waiting lines, lack of resources to meet the demand, and financial loss, consequences that the patient's absence from the scheduled appointment can cause. To reduce these adverse effects, health centers have implemented various strategies, including sanctions and reminders. Although, during the last decades, a sizable number of medical scheduling systems have been developed to achieve better appointment allocation based on predictive models. Therefore, machine learning (ML) algorithms can serve as efficient tools to help understand the patient's behavior concerning their presence in the medical appointment.

A better understanding of the patient's absenteeism phenomenon allows the development of solutions to mitigate the occurrence of no-shows and contribute to the management and planning of health services. According to the context in which ML algorithms are applied, different observations are made, and new solutions can be modeled to mitigate absenteeism in medical appointments and exams. Thus, data collection and the application of ML algorithms in different contexts in healthcare should be explored, providing relevant information to assist in the decision process regarding the scheduling of appointments and exams.

There is currently one publicly available dataset present for patients' no-show which most researchers have been using. The open database is available on the Kaggle[1] platform and refers to medical appointments scheduled in public hospitals in the city of Vitoria, in the state of Espírito Santo – Brazil. The downside of this dataset is the lack of information about how data was pre-processed and the huge class imbalance. Therefore, in our work, we have opted to collect the dataset on our own to grab the most valuable information about the problem and help us to build more accurate classifiers.

To help understand the problem of absenteeism in medical appointments, we analyze the reasons that leverage the decision of a patient not to attend their medical consultation. Starting from the initial dataset, we identify the main factors related to the patient's absence and propose a no-show classification model based on ML techniques. To compose the solution, algorithms that apply supervised ML techniques are used. As such, the present work brings to the current state-of-the-art a new contribution to elucidate the reasons for the no-show of healthcare patients. For that purpose, a new ML model is proposed and validated with real-life medical data, an essential tool for managing healthcare units.

The remainder of this paper is organized as follows. Section 1 presents an approach to the context of the research. Section 3 describes the methodology applied in this study, including data analysis and the algorithms. Next, Section 4 discusses the results. Finally, Section 5 presents the final remarks.

2. Materials and Methods

Machine learning, as a subset of artificial intelligence, is a field of computer science that aims to develop algorithms that can improve through experience and by the use of incremental data [6]. In the past decades, an increase in its research interest led to more and more areas of science finding application within ML algorithms. Application can be

[1] https://www.kaggle.com/joniarroba/noshowappointments

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0342.v1

found in agriculture [7], industry [8], sensor networks [9], fashion [10,11] or healthcare [12], just to mention a few examples.

When considering the healthcare environment, some studies apply ML algorithms to identify preeminent factors and characteristics of patients associated with the lack of attendance to the scheduled appointment [1,13,14]. Other studies use statistical predictive models capable of predicting whether a patient will be absent from the appointment based on their historical data [15]. Lee et al. [16] described the development process up to the implementation of the predictive model in a real clinical environment, as well as the insights acquired using the model.

Despite the relevance of patient absenteeism in medical appointments, only very few works use ML algorithms to identify and understand this problem. In this sense, ML was first applied in [17,18]. In these works, the publicly available dataset was used and guided the overall analysis. However, the lack of available attributes in the dataset did not allow the authors to build a solid predictive model. To the best of the authors' knowledge, a study of the new private dataset with different attributes to analyze the data distinctly and validate new ML models for the no-show problem is still missing in the literature.

This work presents a statistical analysis to enumerate the potential reasons why patients do not attend appointments. In addition, we apply ML techniques to create models that better fit the absenteeism problem. Thus, this work provides answers to some of the questions regarding patients' non-attendance, that is:

- What are the key indicators that signal that a patient will not attend a scheduled appointment?

- What is the probability of a patient not showing up for an appointment?

2.1. Research Approach

The present work brings a new contribution to elucidate the reasons for the no-show of patients and build a ML model according to the following steps (see Figure 1):

1. Collect the patient dataset consisting of data from both appointments and patients;
2. Apply data cleaning techniques to prepare the dataset;
3. Include peripheral databases to add more value to the initial dataset;
4. Analise potential correlations between attributes in the dataset;
5. Start a descriptive data analysis to detect key factors and trends that contribute to patients no-show;
6. Adapt the dataset for the training and testing phase, and try several classification algorithms to process the data;
7. Compare different performance metrics from the ML models, and select the model that provides the most accurate results for the problem at hand.

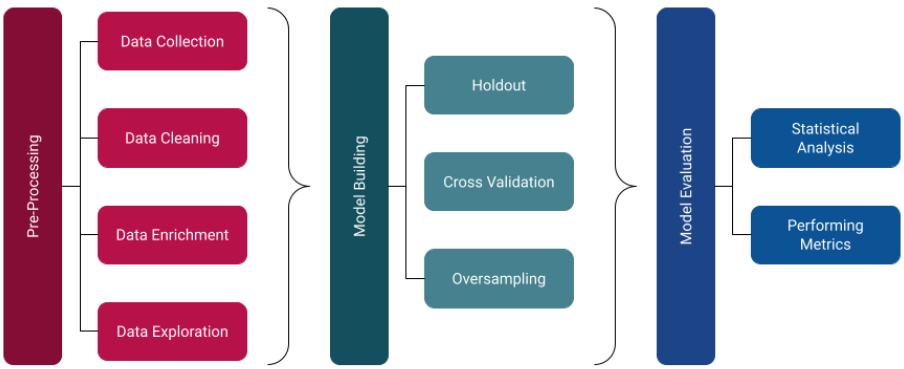

Figure 1. Steps in the analysis and modeling process.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0342.v1

Figure 1 summarizes the building steps of the ML model, aggregating preprocessing, model building, and model evaluation stages.

2.2. Data Preprocessing

Data preparation is one of the most relevant aspects of ML. This task is one of the most time-consuming and requires 60%, on average, of the time and energy spent on a data science project [19]. We include in this task the collection, cleaning, enrichment, and exploration of data, which are presented as follows.

Data Collection

The dataset used in this study consists of data obtained and extracted from the University of Vale do Itajaí Center of Specialization in Physical and Intellectual Rehabilitation (CER). The CER is an outpatient care service that performs diagnosis, assessment, guidance, early stimulation, and specialized care. It has acted in functional rehabilitation and psychosocial qualification to encourage the autonomy and independence of people with disabilities [20]. Firstly, we collected the relevant information – on the absenteeism problem – in loco at the rehabilitation center by transcribing 4812 medical records from an electronic spreadsheet of 2017 and 2019. In the initial dataset, each file is composed of the following attributes:

1. "Medical record number": unique identifier of the patient's record;
2. "Gender": male or female gender of the patient;
3. "Appointment Date": appointment date scheduled;
4. "Attended": given whether the patient attended the scheduled appointment or not;
5. "No-show Reason": description of the reason why the patient did not attend the scheduled appointment;
6. "Type of Disability": the patient's motor or intellectual disability;
7. "Date of Birth": the patient's date of birth;
8. "Date of Entry into the Service": date of the patient's first appointment at the CER;
9. "City": city where the patient resides;
10. "ICD": identifier of the patient's disease;
11. "UBS": basic health unit that sent the patient to be treated at the CER.

The dataset contains a target feature, identified by the variable "Attended" in which: "no" represents a patient that did not attend the medical appointment, and "yes" represents a patient that showed up. Unlike a system that performs a task by explicit programming, a ML system learns from data. It means that, over time, if the training process is repeated and conducted on relevant samples, the predictions will be more accurate.

Data Cleaning

This process converts (or maps) the data to another convenient format to carry out an analysis. In this work, the data manipulation process was performed in the virtual environment Google Colaboratory [21], through the Python programming language [22] with the help of libraries, such as pandas [23] and NumPy [24]. Firstly, we renamed the dataset columns. Secondly, we started the validation process. For the attribute "Attended", we founded that some values were in a different format than expected, such as "No", "no", and "Did not attend". To deal with this inconsistency, we adjusted all values to the value "No" to standardize this attribute value – in case the patient did not attend the scheduled appointment.

The expected value for the "Type of Disability" column was the letter "I" for intellectual disability and "F" for physical disability. However, we noticed that seven empty values, and three values outside the expected standard. In those cases, we amputated the empties values and corrected the others. Another validation of the data was related to the appointments date format. The initial data were not in the standard day/month/year. We provided the adjustment to this format. After this transformation, we considered only the data for 2019 and discarded the 90 medical records found for 2017.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0342.v1

Data Enrichment

In order to add more information to the collected data, some other databases were combined with the current database. Also, new columns were created based on the existing ones. The following items describe this process.

A. Disease Data: As the initial database only contained the patient's disease code, a new database with the names of related diseases was combined. With the inclusion of the disease names, data visualization and interpretation became more objective. The database with the International Classification of Diseases (ICD) was extracted from a file in PDF format on the government portal [25], and transcribed to a file in JSON format. Initially, we adjusted the ICD registered in the database to the disease codes – extracted from the government portal – for the data merging. After the code standardization and data merging, we identified 37 diseases with different ICDs, and 1662 medical records without the registered disease code.

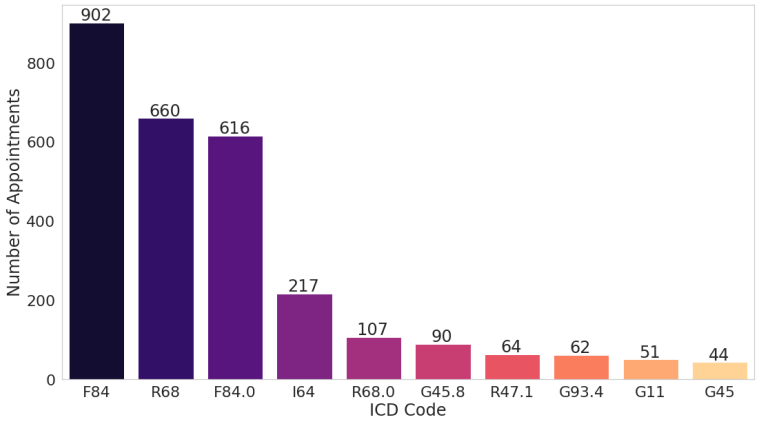

Figure 2. Higher Incidence Diseases.

According to the graph in Figure 2, we could check that some diseases stand out to the number of appointments. *Global Developmental Disorders*, *Other General Symptoms and Signs*, and *Child Autism* together, corresponded to 66.31% of consultations carried out at the CER in 2019.

B. Weather Conditions Data: With the purpose of identifying whether weather conditions could influence patient absenteeism, we entered precipitation and temperature data into the dataset. We extracted the historical data with regard to the year 2019, from the National Institute of Meteorology (Inmet) [26]. INMET's database is feeding every 1 hour and, for each day, presents 24 measurements of temperature and precipitation. Another relevant factor regarding the obtained data, the measurements are carried out only in the city of the meteorological center. Since that Univali's CER serves the region with eleven cities of AMFRI (Association of Cities of Foz do Rio Itajaí) [27], only Itajaí has meteorological data from this data source. The measurement dates from the INMET dataset have converted to day/month/year format – they were in the US format, month/day/year – for data standardizing. Atmospheric pressure, speed and direction of the wind, and air humidity data were discarded from the original dataset, keeping only the temperature and precipitation data. After data validation and standardization, the mean and maximum temperature and precipitation for each day were calculated and merged with the dataset of the medical records. The highest average temperature found was during April and November, remaining around 25 degrees. The temperatures registered were highest in April and October, approaching 34 degrees. Finally, a qualitative value has been assigned to represent the temperature and precipitation range. For temperatures, five classifications had considered: very cold, cold, mild, warm, and very warm. These classes represent temperatures less than or equal to 15 degrees, greater than 15 degrees, greater than

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0342.v1

22 degrees, greater than 27 degrees, and greater than 32 degrees, respectively. As for precipitation, we entered the following values: no rain, weak, moderate, strong, and very strong. This classification refers to the maximum precipitation of the day that has been less than 1 millimeter (mm), greater than 1 mm, greater than 2.5 mm, greater than 10 mm, and greater than 50 mm, respectively.

C. Other Related Attributes: We have created new attributes for the dataset based on the existing ones. From the date of birth included in the data, we derived the patients' ages. In the same way, we extracted the month of the appointment from the date of the registration. Finally, we obtained the appointment shift based on the scheduled appointment time. The age attribute allowed us to analyze whether there is a relationship between the patient's age and the rate of abstention from appointments. The month of consultation helped us identify a correlation with the level of abstentions being severity in the months in which there is a drop in temperatures. After achieving the validation process, merging new databases, and inserting attributes in the original CER dataset, we kept 22 attributes to implement the initial data analysis.

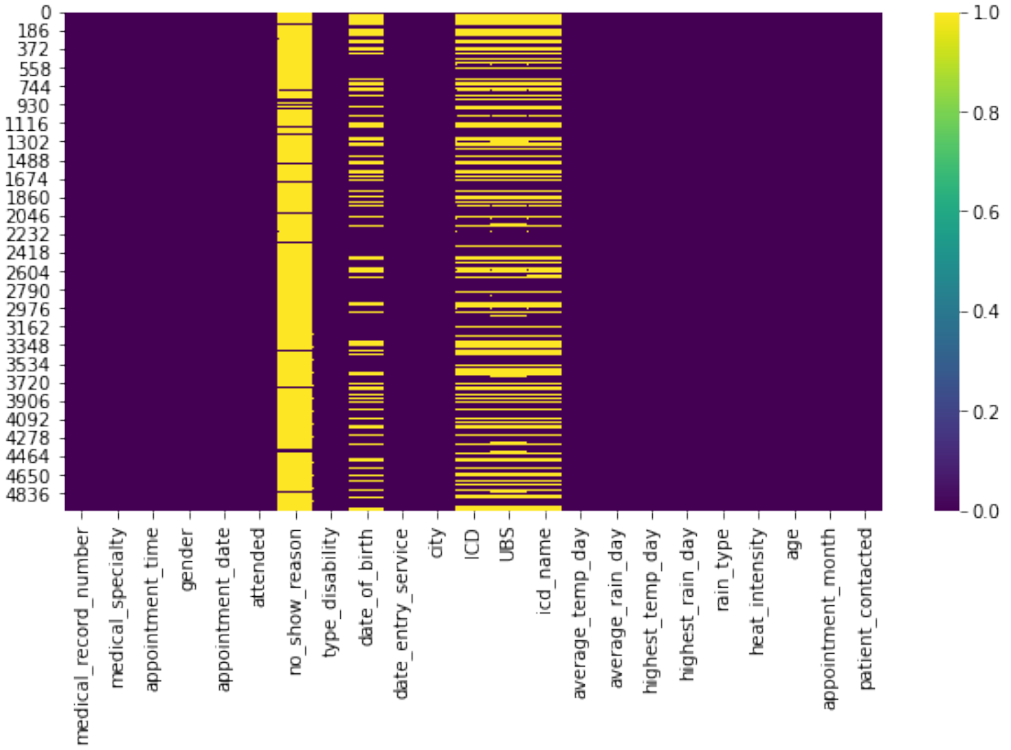

Figure 3. Heatmap for dataset attributes with null values.

Figure 3 presents the heat map of the attributes and their respective missing values: the dataset null value (in yellow) and the accurately filled value (in purple).

Data Exploration

Data exploration is one of the steps responsible for exploring and visualizing data so that it is possible to identify patterns contained in the data sample. In this way, we enable inference that can contribute to the understanding of the problem in question. One of the ways to summarize the data and get an overview of the attributes is through descriptive statistics. The use of descriptive statistics allows summarizing the main characteristics of the dataset numerical characteristic (continuous or discrete), such as top, frequency, mean, and standard deviation (std). However, in a scenario with a lot of categorical variables, other approaches may be more appropriate. An overview of categorical attributes is in Figure 4. Figure 5 presents the correlation matrix heat map. This figure illustrates the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0342.v1

medical_specialty	gender	attended	type_disability	city	ICD	icd_name	rain_type	heat_intensity	appointment_month	age_group	appointment_shift
Fisio	M	Yes	I	ITAJAÍ	NaN	NaN	rainless	warm	apr	Under Age	Afternoon
Fisio	M	Yes	I	ITAPEMA	NaN	NaN	rainless	warm	apr	Under Age	Afternoon
Fisio	M	Yes	F	ITAJAÍ	R68.0	Hipotermia Não Associada à Baixa Temperatura A...	rainless	warm	apr	Under Age	Afternoon
Fisio	F	Yes	F	CAMBORIU	I64	Acidente Vascular Cerebral, Não Especificado C...	rainless	warm	apr	Adult	Afternoon
Fisio	M	Yes	I	BALN. CAMBORIU	NaN	NaN	rainless	warm	apr	Under Age	Afternoon
Fisio	M	Yes	I	BALN. CAMBORIU	NaN	NaN	rainless	warm	apr	Under Age	Afternoon
Fisio	M	Yes	I	ITAJAÍ	NaN	NaN	rainless	warm	apr	Under Age	Afternoon
TO	M	Yes	I	BALN. CAMBORIU	F84.0	Autismo Infantil	rainless	warm	apr	Under Age	Afternoon
TO	M	Yes	I	ITAJAÍ	F84.0	Autismo Infantil	rainless	warm	apr	Under Age	Afternoon
TO	M	Yes	I	BALN. CAMBORIU	NaN	NaN	rainless	warm	apr	Under Age	Afternoon

Figure 4. Categorical values sample.

correlations between all variables: the gray fields do not represent any correlation, while the relative intensity of the yellow and blue colors represents an increase in correlation. In particular, it shows positive or direct correlation (in yellow) – in which the variation of one characteristic directly affects another – and negative or inverse correlation (in dark blue) – in which the fluctuation of one attribute inversely affects the other.

According to [28], the correlation refers to the linear relationship between variables. The correlation coefficient is a metric of the association between two numerical variables, usually denoted as x and y. Thus, Pearson's correlation is defined as a measure of linear association between quantitative variables and is frequently applied to explore the relationship between this type of variable in the data set. Nevertheless, we recommended employing another statistical method that supports the exploration of the relationship between variables of this type, because the available dataset has many categorical variables. Cramér's V (also known as Cramér's ϕ) is one of the statistical correlation techniques developed to measure the strength of the association between two nominal variables [29]. Unlike Pearson's correlation, Cramér's V assumes values in the interval $[0, 1]$. The value 0 corresponds to the absence of association between variables, values close to zero correspond to a weak association, and values closer to 1 correspond to a stronger association. Figure 5 presents the relationship of the numerical value using Pearson's coefficient and categorical variables using Cramér's V. By analyzing the correlations, in the heat map, we check for attributes with high correlations. For both numerical and categorical attributes, a direct correlation is measured between 0.7 and 1 and, from -1 to -0.7 for inverse correlation in numerical attributes. However, the attributes related to the patient's attendance at the appointment did not show a high correlation with any other attribute directly. The following features illustrate the correlations between the attributes of the dataset:

- The age of the patient ("age") and the ICD of the disease ("ICD"): patients in a certain age group are likely to have diseases caused by age;
- Patient attending the appointment ("attended") and the reason for not attending: the fact that the patient notified the reason for his/her non-attendance is directly related to the fact that he/she does not attend the scheduled appointment.

2.3. Descriptive Analysis

The descriptive analysis process begins with the observation of the distribution of the target variable within the dataset. This step was conducted by relating each feature to the target variable "attended". In this section, we analyzed only the five most important characteristics.

Figure 6 shows the relationship between the patient's gender attribute and the fact that he (or she) attends the scheduled appointment. It should be noted that the proportion of consultations by women is much lower than that of men, 1333 and 3676 consultations for

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0342.v1

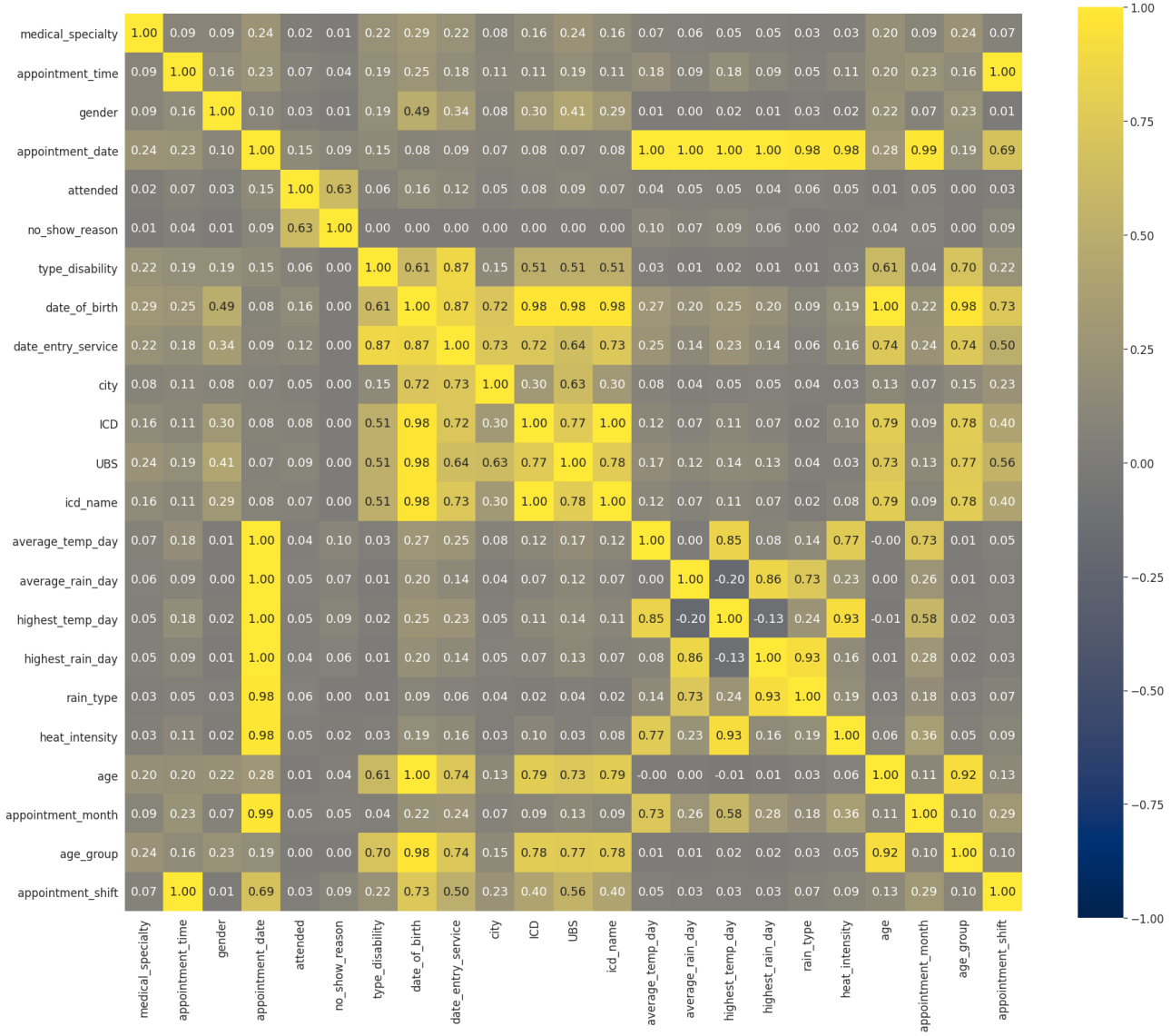

Figure 5. Correlation Heatmap.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0342.v1

women and men respectively. However, the amount of abstentions by women exceeds that of men, with female patients being responsible for 13.13% of abstentions against 10.45% for males.

Concerning the age of patients, we extracted two age categories for analysis: patients under 18 years old were labeled as "Minor Age" and the others as "Adult". It is observed in Figure 7 that there is an imbalance about these categories, with appointments for younger people being much more prevalent. Adults represent only 19% of consultations carried out at the specialized center.

We also note that most of the data collected are from patients under 18 years old. It may reveal some characteristics of the behavior of these patients that are not exclusively related to the patient who will receive care. In many cases, underage patients need someone close to accompany them in medical care. This fact implies a behavioral analysis of the patient and their companion. In this study, we did not collect data related to this issue.

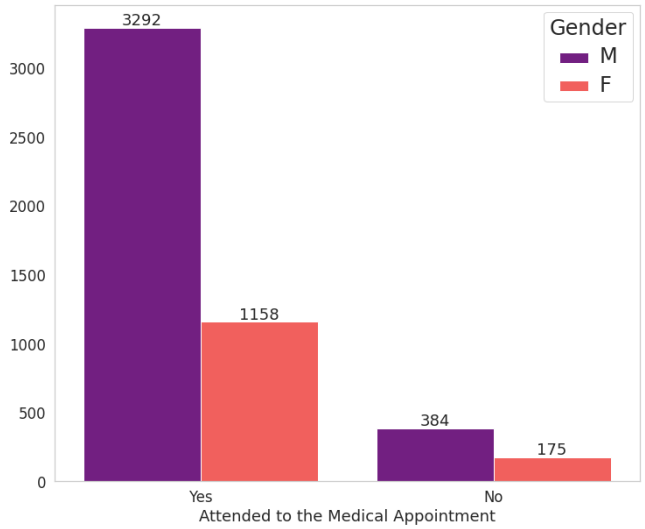

Figure 6. Analysis by gender.

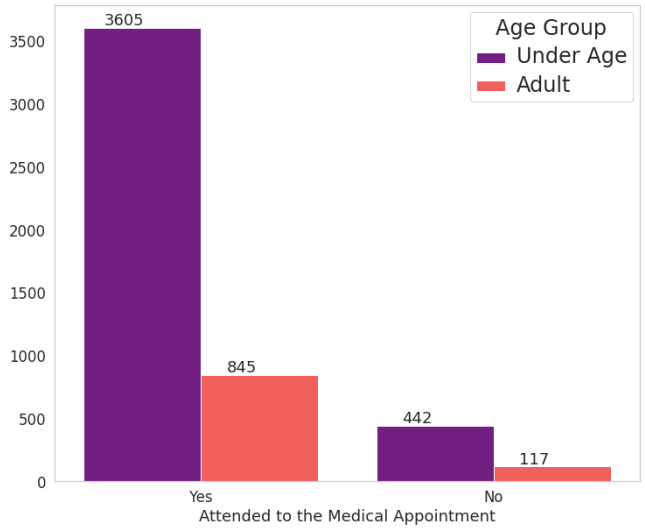

Figure 7. Analysis by age group.

We extract the "appointment month" attribute from the existing column labeled "appointment date". Figure 8 shows that the month with the highest number of scheduled appointments is the month of September, followed by the months of October and August. In addition, the month with the fewest medical appointments is November. However, May,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0342.v1

July, and April are responsible for the highest abstentions from scheduled appointments. In May alone, the number of patients who scheduled medical appointments at the specialized center and did not show up totaled more than 16% of the number of medical appointments.

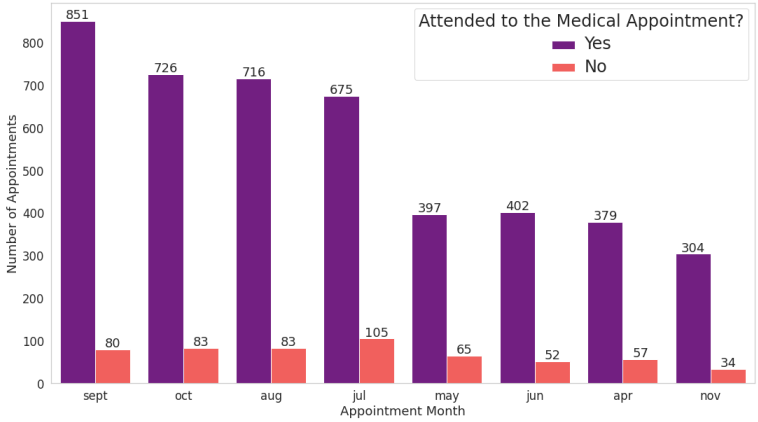

Figure 8. Analysis by appointment month.

As presented in Section 2.2, we included climate data in the original dataset. Figure 9 shows a way to visualize the relationship between weather conditions and the probability of the patient not attending the scheduled medical care. From Figure 9, we observe that in autumn and winter, temperatures tend to be lower in the city of Itajaí, region of the medical center.

In July, the number of patients who did not attend the scheduled appointment reached more than 15%. This fact may be related to the average temperature this month being the lowest in the year. However, based on Figure 10, it is not possible to establish a direct relationship between the patient's behavior and his abstention from rain strength on the appointment day.

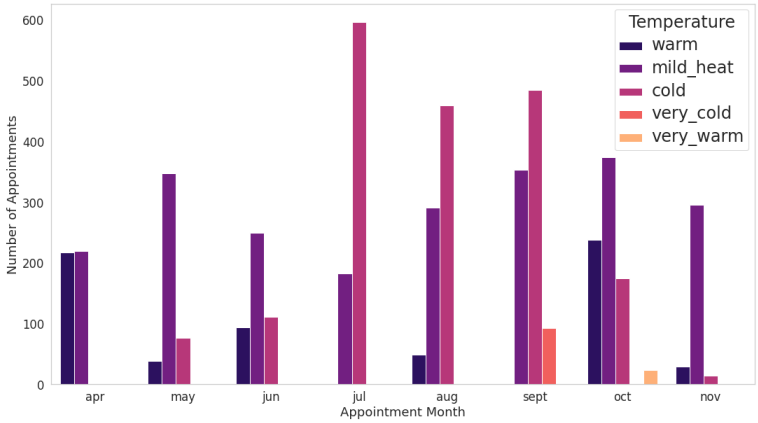

Figure 9. Analysis by day temperature.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0342.v1

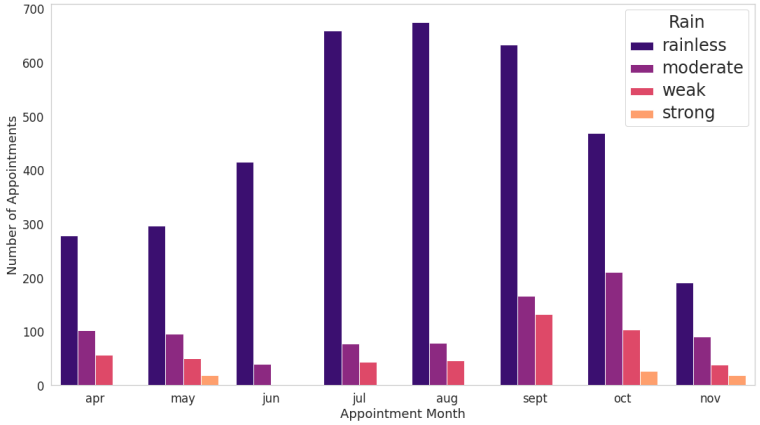

Figure 10. Analysis by rainfall.

We extracted the appointment shifts based on the appointment time and divided them into "morning" and "afternoon" shifts. The morning shift refers to the period until midnight and the afternoon shift from this time on. Scheduled appointment times at the health center range from 7 am to 6:20 pm. In the analyzed data, there are 35 different scheduled times, which the time of 8:40 am with the highest number of medical appointments.

The no-show proportion is similar for both shifts, in which 12.24% for the morning shift and 10.34% for the afternoon shift. Figure 11 shows that the absolute number of abstentions based on appointment times has low variability. When the highest frequency of scheduled medical appointments, the number of no-shows drops to around 9%, about 3% lower than the total number of abstentions for the morning.

In this way, although the hours most frequently present fewer abstentions, we can infer that the patients' profile has not impacted the appointments during business hours. It is relevant to clarify that the focus diseases at the study's medical center mainly concern patients with motor disabilities. For this reason, possibly, they already have a routine with different hours.

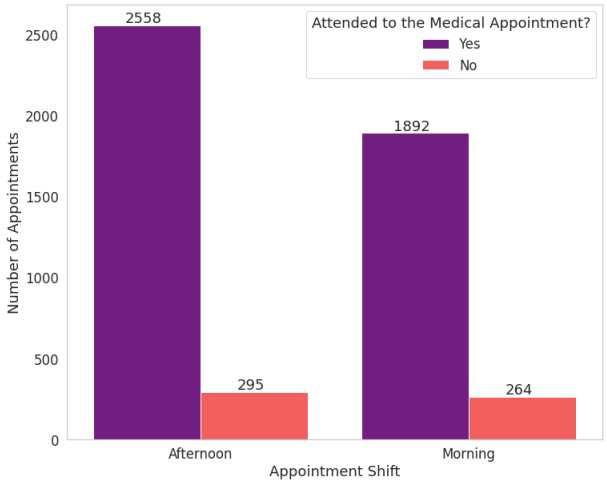

Figure 11. Analysis by appointment shift.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0342.v1

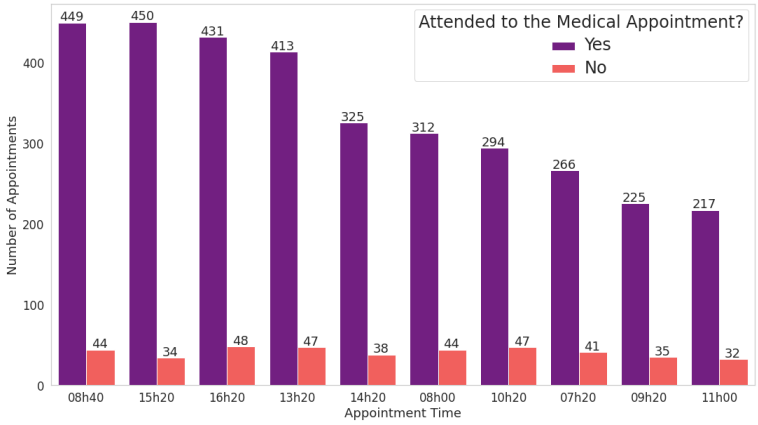

Figure 12. Analysis by appointment time.

3. Model Building

Predictive analysis uses statistical and ML techniques to understand the data structure and predict future outcomes based on historical data. The modeling process selects models, which are based on ML algorithms and used in the experimentation. In the model development and testing stage, the subset of ML algorithms required to solve patient attendance prediction is classification, since the records presented in this dataset are labeled as "yes" (or "no") to the target variable "attended".

In ML, classification problems are a form of supervised learning. These types of algorithms have multiple predictors and also a variable of interest responsible for guiding the analysis [30]. On the other hand, unsupervised learning has only predictors (covariates) available and focuses on identifying patterns in data sets containing data points that are neither classified nor labeled. The classification algorithms taken into consideration in this work are:

- Logistic Regression classifier
- Decision tree classifier,
- Random forest classifier

This paper made use of the open-source scikit-learn (SKLearn)[2] library to develop supervised learning models. SKLearn library implements various ML algorithms and model performance analysis functions using the Python programming language. In the scope of this paper, we define a classification model as an algorithm implemented by a pre-defined function from the SKLearn library, which takes a distinct possible set of parameters.

3.1. Class Imbalance

Upon inspecting the percentage distribution of the records between the "yes" (or "no") attribute in the target variable "attended", we find a considerable imbalance between both classes. An imbalanced classification problem is well-known, in which the distribution is biased. Figure 13 shows 90% of the dataset's records are labeled as "yes" and 10% are labeled as "no".

Imbalanced classifications are a challenge for predictive modeling as most of the ML algorithms used for classification were designed based on the assumption of an equal number of examples for each class. This model generally results in poor predictive performance, specifically for the minority class. Thus, the problem is more sensitive to classification errors for the minority class than the majority class [31].

To solve the problem of class imbalance, and after testing several different undersampling and oversampling algorithms, the Synthetic Minority Oversampling Technique

[2] https://scikit-learn.org/stable/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0342.v1

(SMOTE) algorithm provided by the Imbalanced-learn (imblearn)[3] library yielded the most considerable performance improvement. Not only did SMOTE solved the issue of class imbalance, but it also improved the classification performance metrics, mainly in precision and recall measures. Figure 14 shows the target class after applying the oversampling algorithm.

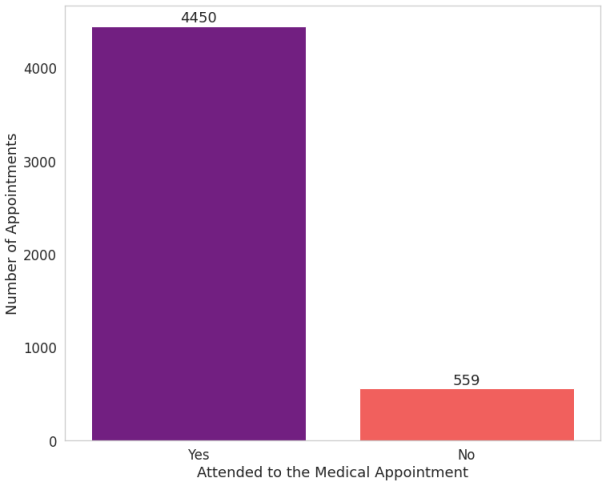

Figure 13. Class imbalanced target variable.

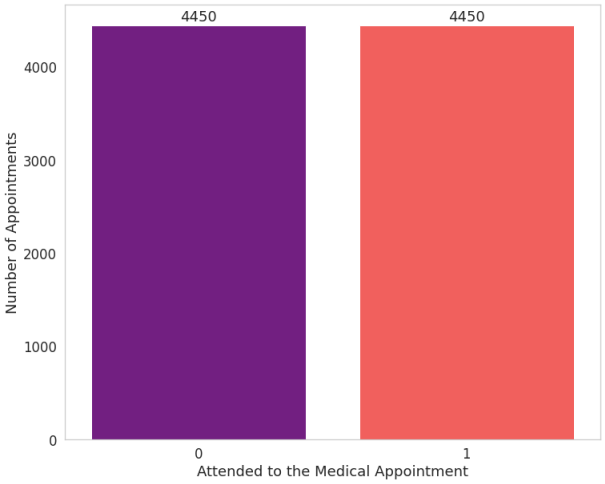

Figure 14. Class balanced after oversampling process.

3.2. Holdout and Cross Validation

The model must be trained on a consistent number of observations to refine its prediction ability to train the model to classify new patterns. If possible, two distinct datasets are the best choice: one for training and a second to be used as a test. In this case, as two dedicated datasets were not available, the original dataset was split in one part for training (70%) and another used for testing (30%), called the holdout method [32].

The `train_test_split` function, available in the SKLearn library, splits the data in both training and testing subsets. In the dataset split step, we need to keep the same distribution of target variables within both the training and test datasets. It is necessary to avoid that a random subdivision can change the proportion of the classes present in the training and test datasets from that in the original. Thus, even after the process of oversampling described in Section 3.1, we apply the parameter "stratify" to the `train_test_split` function to preserve the proportion of classes.

[3] https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0342.v1

We applied the cross-validation technique to prevent overfitting problems and to estimate the performance of the model. According to [33], in this technique, a dataset is randomly divided into k disjoint folds of approximately equal size, and each fold is in turn used to test the model derived by a classification algorithm from the other $k1$ fold. Then, the performance of the classification algorithm is evaluated using the average of the k accuracies resulting from the k-fold cross-validation. There is no defined rule for choosing k, although splitting the data into 5 or 10 parts is more common. Figure 15 shows an overview of the cross-validation method.

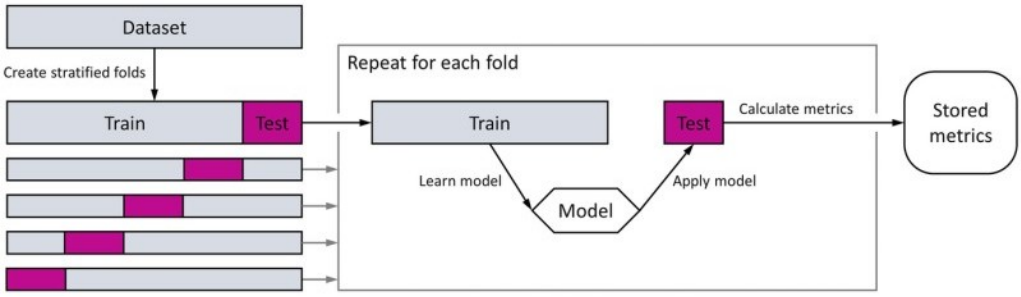

Figure 15. Schematic overview of k-fold cross-validation [34].

4. Results and Discussion

This section presents the results obtained after performing the data pre-processing, adjusting the class imbalance using SMOTE oversampling, performing feature selection, and tuning algorithms hyperparameters per classifier. The results illustrate the performance obtained when testing the models mentioned in Section 3 of this paper via multiple, defined metrics. To report on and evaluate model performance, we apply the following performance metrics [35]:

1. Accuracy: a simple ratio of total correct predictions over total wrong predictions.
2. Precision (Positive predictive value): the ratio of correctly predicted instances per class to all predictions made for that same class.
3. Recall (Sensitivity): the ratio of the correctly predicted instances per class to the total amount of actual instances labeled to that class.
4. F1-score: a harmonic mean of recall and precision.

These results are acquired by drawing a confusion matrix for each classifier and then taking the average of the results. The confusion matrix enables visualization of the metrics, where the diagonal elements represent the number of points for which the predicted label is equal to the true label, while off-diagonal elements are those that are mislabeled by the classifier. After training all the three algorithms and fitting the Random Forest model with the test data, we obtained the confusion matrix shown in Figure 16.

In the considered case study, we are interested in predicting the high number of patients who could not attend medical appointments by minimizing the incidence of false negatives. Thus, we selected the Random Forest classifier as the best classification algorithm able to achieve the objective of the analysis. In Table 1, we summarize the classification report of the Random Forest model, which is related to precision, recall, and F1-score statistical metrics.

Table 1. Random Forest Classifier Metrics. Note, 0 = not attend and 1 = attend.

	Precision	Recall	F1-Score	Support
0	0.91	0.93	0.92	1469
1	0.93	0.91	0.92	1468
accuracy			0.92	2937
macro avg	0.92	0.92	0.92	2937
weighted avg	0.92	0.92	0.92	2937

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0342.v1

After the holdout process in the oversampled data described in Section 3.2, we trained and tested the algorithm with 5963 and 2937 instances, respectively. The Random Forest algorithm correctly classified 2703 out of the total amount of test instances. Thus, the classifier obtained:

- The highest true positive rate of approximately 92%, correctly predicting 1363 out of 1469 of the patients who do not attend the scheduled appointment;
- The lowest false positive rate of approximately 0.07%. It only failed to detect 106 patients who had attended the appointment, getting the best recall score of 0.91;

The recall was identified as the most relevant performance metric to ensure the minimum number of false negatives (patients who may potentially attend the appointment but are not classified as such by the system) to a lack of precision resulted in considerable numbers of false positives. The Area Under the Curve (AUC) of Receiver Characteristic Operator (ROC) curve is an often-used performance metric for classification problems. It is one way to assess the rate of observed positives predicted as positives (sensitivity) and the proportion of observed negatives predicted as negatives (specificity). As closest to 1 the AUC is, better the model is at distinguishing between patients who will attend and not attend the appointment.

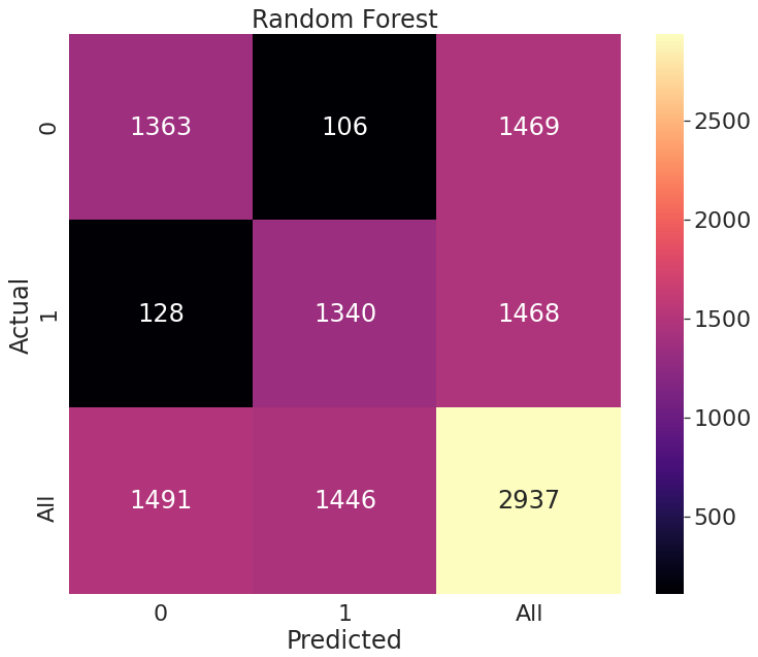

Figure 16. Confusion Matrix. Note, 0 = not attend and 1 = attend.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0342.v1

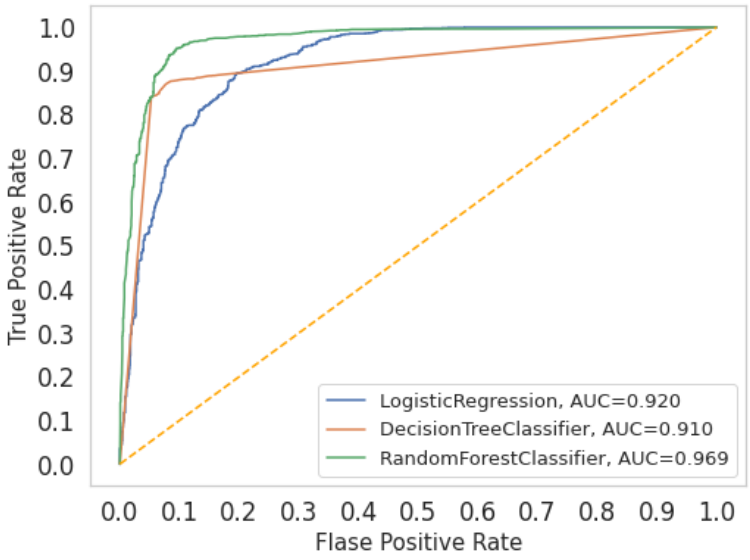

Figure 17. AUC-ROC curve performance.

5. Conclusions

This study aimed to build a no-show classification model for patients and investigate the meaningful features that signal that a patient will not attend the scheduled medical appointment. A better understanding of the absenteeism phenomenon in patients and the development of solutions to prevent the no-show behaviour; can improve healthcare quality worldwide.

To achieve the results, we applied some ML techniques to identify the factors that may contribute to the absenteeism of the patients and, above all, to predict the likelihood of individual patients not attending the scheduled appointments. Firstly, we extracted and preprocessed the data collected from the University of Vale do Itajaí CER. Secondly, we assess the data statistically and then analyze the most relevant features. Since the initial dataset only contained features that correlated directly to the no-show problem, we boosted the information to the dataset through a data enrichment step.

Followed by the initial preprocessing step, we handled the class imbalance problem by applying oversampling techniques, and then tested the ML algorithms to the data. In the first step, we split data into training and test data sets, to guarantee equivalence of the target variable and avoid the bias influence on the algorithms. We selected many classification algorithms. For each of them, we carried out the training and validation phases. The predicted results were collected and fed into the respective confusion matrices to evaluate the algorithm's performance. Based on the confusion matrix report, it was possible to calculate the metrics necessary for an overall evaluation (precision, recall, accuracy, F1-score, and AUC-ROC curve) and to identify the most suitable classifier to predict whether a patient was likely to not attend to the scheduled appointment.

As a final step, we analysed the performance metrics for each model, and the Random Forest seemed to be a good choice to be the final model for the available dataset. It revealed the best recall rate (0.91), a performance metric that shows the ability of a classifier to find all the positive instances. It achieved an overall false-negative rate equal to 0.08% of the total observations. The proposed predictor and analysis results demonstrate that the attributes that can influence the patient's attendance in a medical consultation are the lower weather temperature and the appointment time. However, many other factors may implicitly influence the patient's no-show that could not be inferred from the analyzed dataset.

The results obtained from the data analysis represent a starting point in the development of efficient patient no-show classifiers. The availability of additional information on patients and medical appointments may help to improve this knowledge. In the future,

we intend to study an enlarger number of records and attributes to increase the size of the dataset acquired and improved the ability of the model to learn new behaviours.

Author Contributions: Conceptualization, L.H.A.S, W.D.P, and A.M.R.F; methodology, L.H.A.S, W.D.P, and A.M.R.F; analysis, L.H.A.S, W.D.P, S.D.C. and A.M.R.F; writing–original draft preparation, L.H.A.S, V.H.Q.L, W.D.P, S.D.C. and A.M.R.F; writing–review and editing, L.H.A.S, V.R.Q.L, S.D.C., W.D.P, and A.M.R.F

Funding: This research was funded in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil (CAPES) – Finance Code 001 and Fundação de Amparo à Pesquisa do Estado de Santa Catarina (FAPESC) – EDITAL DE CHAMADA PÚBLICA FAPESC No. 06/2017. We would like to thank Seed Funding ILIND—Instituto Lusófono de Investigação e Desenvolvimento, under project COFAC/ILIND/COPELABS/1/2020 and COFAC/ILIND/COPELABS/3/2020.

Conflicts of Interest: Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kheirkhah, P.; Feng, Q.; Travis, L.M.; Tavakoli-Tabasi, S.; Sharafkhaneh, A. Prevalence, predictors and economic consequences of no-shows. BMC health services research **2016**, 16, 13. doi: 10.1186/s12913-015-1243-z.
2. More than 50,000 patients do not show up for exams, appointments and scheduled surgeries - Mais de 50 Mil Pacientes Não Comparecem para Exames, Consultas e Cirurgias Agendadas.
3. Furtado, L.P.; Fernandes, P.C.; dos Santos, J.H. Redução de faltas em consultas médicas e otimização dos recursos da saúde pública em Vitória-ES por meio de Mineração de Dados e Big Data. Congresso Brasileiro de Informática em Saúde (CBIS), 2017, Vol. 15, pp. 23–25.
4. Menendez, M.E.; Ring, D. Factors Associated with Non-Attendance at a Hand Surgery Appointment. Hand **2015**, 10, 221–226. doi:10.1007/s11552-014-9685-z.
5. Mieloszyk, R.J.; Rosenbaum, J.I.; Bhargava, P.; Hall, C.S. Predictive modeling to identify scheduled radiology appointments resulting in non-attendance in a hospital setting. 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2017, pp. 2618–2621. doi:10.1109/EMBC.2017.8037394.
6. Kubat, M. An Introduction to Machine Learning; Springer, 2021.
7. Liakos, K.G.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D. Machine learning in agriculture: A review. Sensors **2018**, 18, 2674. doi:10.3390/s18082674.
8. Ge, Z.; Song, Z.; Ding, S.X.; Huang, B. Data Mining and Analytics in the Process Industry: The Role of Machine Learning. IEEE Access **2017**, 5, 20590–20616. doi:10.1109/ACCESS.2017.2756872.
9. Correia, S.D.; Tomic, S.; Beko, M. A Feed-Forward Neural Network Approach for Energy-Based Acoustic Source Localization. Journal of Sensor and Actuator Networks **2021**, 10, 29. doi: 10.3390/jsan10020029.
10. Henrique, A.S.; Fernandes, A.M.D.R.; Lyra, R.; Leithardt, V.R.Q.; Correia, S.D.; Crocker, P.; Dazzi, R.L.S. Classifying Garments from Fashion-MNIST Dataset Through CNNs. Advances in Science, Technology and Engineering Systems Journal **2021**, 6, 989–994. doi: 10.25046/aj0601109.
11. Gaussmann, R.; Coelho, D.; Fernandes, A.M.R.; Crocker, P.; Leithardt, V.R.Q. Using Machine Learning for Road Maintenance Cost Estimates in Brazil: a case study in the Federal District. 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), 2020, pp. 1–7. doi:10.23919/CISTI49556.2020.9141148.
12. Bhardwaj, R.; Nambiar, A.R.; Dutta, D. A Study of Machine Learning in Healthcare. 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), 2017, Vol. 2, pp. 236–241. doi:10.1109/COMPSAC.2017.164.
13. Alaeddini, A.; Yang, K.; Reeves, P.; Reddy, C.K. A hybrid prediction model for no-shows and cancellations of outpatient appointments. IIE Transactions on Healthcare Systems Engineering **2015**, 5, 14–32. doi:10.1080/19488300.2014.993006.
14. Huang, Y.; Hanauer, D.A. Patient No-Show Predictive Model Development using Multiple Data Sources for an Effective Overbooking Approach. Applied Clinical Informatics **2014**, 5, 836–860. doi:10.4338/ACI-2014-04-RA-0026.
15. Alaeddini, A.; Yang, K.; Reddy, C.; Yu, S. A probabilistic model for predicting the probability of no-show in hospital appointments. Health Care Management Science **2011**, 14, 146–157. doi: 10.1007/s10729-011-9148-9.

16. Lee, G.; Wang, S.; Dipuro, F.; Hou, J.; Grover, P.; Low, L.; Liu, N.; Loke, C. Leveraging on Predictive Analytics to Manage Clinic No Show and Improve Accessibility of Care. IEEE International Conference on Data Science and Advanced Analytics (DSAA) **2017**, pp. 429–438. doi:10.1109/DSAA.2017.25.
17. Salazar, L.H.; Fernandes, A.; Dazzi, R.L.S.; Garcia, N.; Leithardt, V.R.Q. Using Different Models of Machine Learning to Predict Attendance at Medical Appointments. Journal of Information Systems Engineering and Management **2020**, 5. doi:10.29333/jisem/8430.
18. Salazar, L.H.; Fernandes, A.M.R.; Dazzi, R.; Raduenz, J.; Garcia, N.M.; Leithardt, V.R.Q. Prediction of Attendance at Medical Appointments Based on Machine Learning. 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), 2020, pp. 1–6. doi: 10.23919/CISTI49556.2020.9140973.
19. Flower, C. Data Science Report. (Accessed on September 2021) **2016**. Available at https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf.
20. Dias, A.M.; Marcelino, A.P.; Viana, S.B.P.; Pagnossin, D.F.; Fialho, I.M.; Schillo, R.; Portes, J.R.M. CER II - Centro Especializado Em Reabilitação Física E Intelectual: Cartilha Informativa, 2016. Available at https://www.univali.br/noticias/Documents/Cartilha%20CER%20II.pdf.
21. Google Colaboratory. Available at https://colab.research.google.com.
22. Python. Available at https://www.python.org/.
23. Pandas Library. Available at https://pandas.pydata.org/.
24. NumPy Library. Available at https://numpy.org/.
25. Portal, F.G. Tabelas - CID-10. Available at https://www.gov.br/previdencia/pt-br/images/arquivos/compressed/1a_130129-160538-109.zip.
26. INMET. Annual Historical Data. Available at https://portal.inmet.gov.br/dadoshistoricos.
27. Association of Cities of Foz do Rio Itajaí – AMFRI. Available at https://www.amfri.org.br/.
28. Jupp, V. The Sage Dictionary of Social Research Methods; SAGE Publications, 2006. doi: 10.4135/9780857020116.
29. Frey, B. The SAGE encyclopedia of educational research, measurement, and evaluation; Vol. 1-4, SAGE Publications, 2018. doi:10.4135/9781506326139.
30. Abbott, D. Applied Predictive Analytics: principles, and techniques for the professional data analyst; John Wiley & Sons, 2014.
31. Abd Elrahman, S.M.; Abraham, A. A review of class imbalance problem. Journal of Network and Innovative Computing **2013**, 1, 332–340.
32. Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society. Series B (Methodological) **1974**, 36, 111–147.
33. Wong, T.T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognition **2015**, 48, 2839–2846. doi:10.1016/j.patcog.2015.03.009.
34. Dankers, F.J.W.M.; Traverso, A.; Wee, L.; van Kuijk, S.M.J., Prediction Modeling Methodology. In Fundamentals of Clinical Data Science; Springer, 2018; chapter 8, pp. 101 – 120. doi: 10.1007/978-3-319-99713-1_8.
35. Hossin, M.; Sulaiman, M.N. A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process **2015**, 5, 1. doi: 10.5121/ijdkp.2015.5201.