Explainability Analysis in Predictive Models Based on Machine Learning Techniques on the Risk of Hospital Readmissions

Juan Camilo Lopera Bedoya

GIDITIC, Universidad EAFIT
Medellín, Colombia
jcloperab@eafit.edu.co

Jose Lisandro Aguilar Castro

GIDITIC, Universidad EAFIT, Colombia
CEMISID, Universidad de Los Andes, Venezuela
IMDEA Networks Institute, Spain
aguilar@ula.ve; aguilarjos@gmail.com
Corresponding author

***Abstract*—**
Purpose: Analyzing the risk of re-hospitalization of patients with chronic diseases allows the healthcare institutions can deliver accurate preventive care to reduce hospital admissions, and the planning of the medical spaces and resources. Thus, the research question is: Is it possible to use artificial intelligence to study the risk of re-hospitalization of patients?
Methods: This article presents several models to predict when a patient can be hospitalized again, after its discharge. In addition, an explainability analysis is carried out with the predictive models to extract information to determine the degree of importance of the predictors/descriptors. Particularly, this article makes a comparative analysis of different explainability techniques in the study context.
Results: The best model is a classifier based on decision trees with an F1-Score of 83% followed by LGMB with an F1-Score of 67%. For these models, Shapley values were calculated as a method of explainability. Concerning the quality of the explainability of the predictive models, the stability metric was used. According to this metric, more variability is evidenced in the explanations of the decision trees, where only 4 attributes are very stable (21%) and 1 attribute is unstable. With respect to the LGBM-based model, there are 12 stable attributes (63%) and no unstable attributes. Thus, in terms of explainability, the LGBM-based model is better.
Conclusions: According to the results of the explanations generated by the best predictive models, LGBM-based predictive model presents more stable variables. Thus, it generates greater confidence in the explanations it provides.

Keywords*—*Explainability analysis, Prediction Models, Machine Learning, Hospital Readmission, Health decision-making Systems

I. INTRODUCTION

There are countless lessons that remain after facing the recent COVID-19 pandemic. The deterioration this crisis has caused in various sectors has forced the world today to rethink the idea of how to prepare for future situations. In particular, the health sector must be strengthened to guarantee a good quality service to the entire population in such situations. However, there are several challenges in the health sector, such as access and coverage for the entire population (including care in rural areas), and drug shortages, among others.

On the other hand, unplanned rehospitalizations correspond to hospital events that occur after a previous hospital discharge and during the following 30 days, where the discharge diagnosis of the preceding event is the same or similar to the admission diagnosis of the present event. These are more common than it seems, and it is estimated that about 20% of discharged patients re-enter hospitalization 30 days after discharge [1], which in turn translates into very high costs for the healthcare system. For example, it implied a cost of $17 billion dollars, in other words., 20% of the total hospital payment, of the Medicare program in the United States in 2008 [1] (social security coverage program administered by the U.S. government). According to the same study, it is estimated that 71% of this cost is potentially preventable [1].

However, the Medicare case is extrapolated to other healthcare entities. But identifying the problem is just the beginning. The real challenge lies in designing and implementing sustainable healthcare programs for high-risk patients who must be treated with precision [2]. To achieve this, Artificial Intelligence (AI) and Big Data emerge as real options in favor of improving the quality of patient care [3]. On the other hand, based on the review of the literature on the study of the problem of the risk of hospital readmission (See section II), it is observed that although different implementations have been put into practice, more comparative studies are needed to evaluate the real impact of these solutions, and in particular, it is required to incorporate explanatory analysis approaches in the domain of readmission risk prediction.

The general goal of this work is to predict hospital readmissions in the patient population using machine learning techniques, with an in-depth analysis of the explanatory characteristics of the models obtained. The specific objectives are. The main contributions of this paper are:

- The implementation of predictive models about hospital readmission.

- The definition of explainability methods for the predictive models, which are compared through different metrics.

This work is organized as follows: Section II presents the state of the art. Section III describes the work's theoretical framework, and then, the predictive models are defined in Section IV. Section V presents the experiments, and section VI analyzes the results. Finally, section VII describes the conclusions.

II. RELATED WORKS

Risk predictive models are used nowadays in different economic sectors, including the health sector [6]. Its applicability is of interest in certain healthcare niches, for example, in patients who have an associated disease that makes them more susceptible to consume healthcare services (e.g., chronic diseases).

The scientific interest to address this problem with analytical methods is not new. It was identified through a query made in the Web of Science database that since 2012 there are evidences of a growing number of articles related to patient re-admission. This growth had a jump in 2017, reaching its peak in 2021, as evidenced in Figure 1. Additionally, it is noted that for the first quarter of 2022, the number of publications has been important (just over 200), surpassing the total number of publications in years previous to 2014.

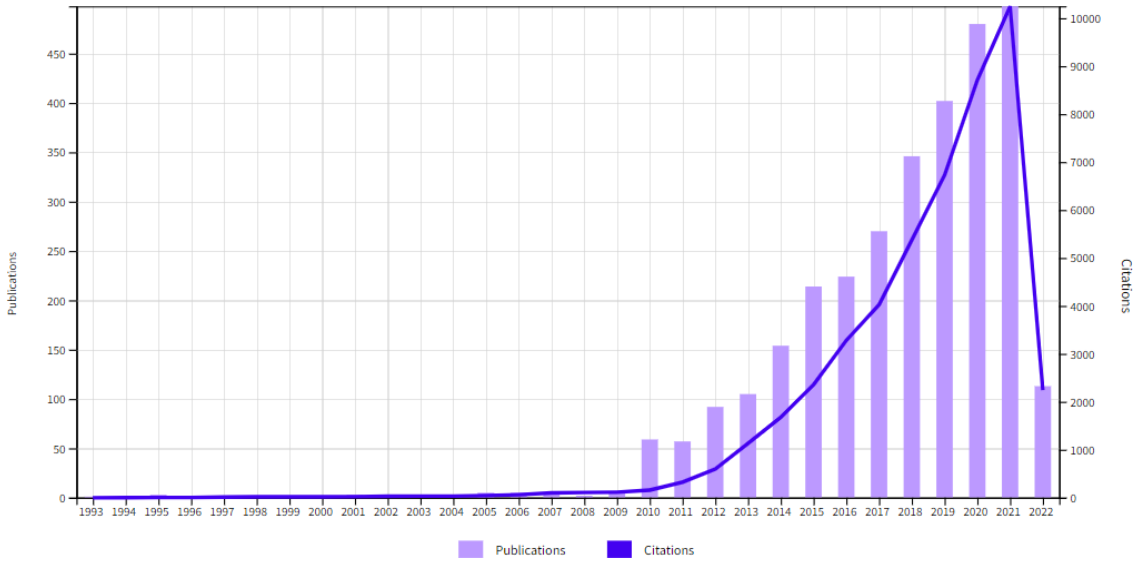

Figure 1. Number of publications per year

.

Starting with the objectives of this work, there are two areas of review to address in this research. The first is associated with *predictive models applied to health*. In 2021, the work [7] presented a logistic regression model to predict the risk of hospital re-admission during the following 15 days to the patient's discharge. To achieve this, it was initially performed a descriptive analysis to identify those behavioral patterns subject to the risk of readmission during the following 15 days of hospital discharge. The predictive model ranked patients according to measured risk, with a model discrimination value of 0.81. Also, Dimitriadou and others [8] implemented Machine Learning techniques (ML) such as Support Vector Machine (SVM), balanced random forests and weighted random forests (Random Forest, RF), to predict the risk of readmission of a discharged patient. To this end, they collected 11,172 hospitalization records with 24 independent variables. The best technique was the balanced random forest (sensitivity of 0.70 and AUC of 0.78).

In the work of Zhang and others [9] studied patients discharged from acute care centers between 2015 and 2016 in Alberta, Canada. The objective was to predict hospital readmissions generated during the 30 days following to hospital discharge (independent of the cause of rehospitalization). Data from 428,669 patients were used, of which 5.83% were readmissions. It was identified that a patient was more likely to be re-hospitalized if they frequented hospital care more frequently, had more office visits, had more prescriptions, had a chronic condition, or if they were older than 65 years. They used the Learnable Adaptive Cosine Estimator (LACE) prediction model with an AUC of 0.66 and another ML model (the Gradient Boosting Machine) that improved the results with an AUC of 0.83. In a previous research [10], 77 articles related to hospital readmission were analyzed. It is evident how the problem has been addressed from different perspectives, and the utilization of multivariate statistics using survival and regression analysis, or ML techniques.

Hoyos et al. [13] carried out as Systematic Literature Review

(SLR) to analyze three modeling approaches of dengue: prediction, prescription and optimization. This SLR defines the state-of-the-art in dengue modeling, using ML, in the last years. In the paper [5] are studied several prediction models based on several ML techniques, considering different input dependences, which include temporal interdependence, dependence with context variables, and temporal intra-dependence. They analyse the quality of the approaches as predictive models for the SEIRD variables. Finally, Camargo et al [22] proposed an incremental learning algorithm to define predictive models of the SEIRD variables for COVID-19. This algorithm is a dynamic ensemble method based on a bagging scheme that allows the updating of incremental models or the addition of new models. The first component carries out an analysis of the interdependencies of the SEIRD variables and the second component is an incremental learning model that builds/updates the predictive models.

The other area of literature review to be addressed in this research is health *explanatory analysis*. As mentioned by Langs and others. in [11], challenges arise in medical decision support such as dealing with uncertainty, probabilistic, unknown, incomplete, unbalanced, heterogeneous, noisy, dirty, erroneous, inaccurate, or missing data, among other things. They conclude that since a great goal of future medicine is to model patient complexity to adapt medical decisions, healthcare practices, and therapies to the individual patient, AI must be able to explain the solutions it arrives at. The opacity of AI must be reduced to generate the necessary confidence.

In the same way, it is explained in [12] that the decision making behind AI black box models requires that they become more transparent, accountable and understandable to humans. It is where general (or global) explanation mechanisms emerge, which do not require predictions, just feature analysis often associated with training data (e.g., attribute weights extracted from linear regression models), or local (given a prediction model, analyze the information for a particular instance). Some of these techniques are Shapley values [21], which are created by a coalitional game theory method assuming that each feature value of the instance is a player in a game where the quality of the prediction is the payoff. The Shapley values distribute the payoff among the features.

Other recent works are the following. Baig et al. [31] developed a predictive model for the risk of 30-day hospital readmission using three ML techniques: XGBoost, Random Forests, and Adaboost. The predictive model was compared with two classic readmission models, patients at risk of hospital readmission and the LACE index, using data from two hospitals in the Auckland region, obtaining better quality metrics in all cases. Lo et al. [32] built four ML models (logistic regression, random forest, extreme gradient boosting, and categorical boosting) to predict 14-day unplanned readmissions. They conducted a study on 37,091 hospitalized adult patients with 55,933 discharges in an 1193-bed university hospital. They used 7 categories of variables extracted from the hospital's medical record dataset. The objective of the work [33] was to determine whether ML models for allocating readmission-mitigating interventions have different usefulness based on their overall utility and discriminative ability. They conducted a utility analysis using claims data acquired from the Optum Clinformatics Data Mart, with 513,495 commercially-insured inpatients. Utility analysis estimated the cost, in dollars, of allocating interventions for lowering readmission risk based on the reduction in the 90-day cost. The objective of the work of Zhao et al. [34] was to build an early prediction model of unplanned 30-day hospital readmission using a large and diverse sample. They also identified novel readmission risk factors and protective factors. They developed six candidate readmission models using different ML algorithms and the best performing model was XGBoost. On the other hand, the study of [35] aimed to select the most affecting features of COVID-19 readmission and compare the capability of ML algorithms to predict COVID-19 readmission based on the selected features. The LASSO feature selection algorithm was used to select the most important features related to COVID-19 readmission. HistGradientBoosting classifier, Bagging classifier, Multi-Layered Perceptron, Support Vector Machine, and XGBoost classifiers were used. XGBoost outperformed the other models. Also, Shang et al. [36] defined ML-based readmission prediction methods to predict the readmission risks of diabetic patients. They used ML classifiers such as random forest, Naive Bayes, and decision tree ensemble for classification, and the random forest algorithm had the highest AUC.

Finally, Huang et al. [37] carried out a review of the literature on ML methods and their performance for predicting hospital readmission in the US. They used the PRISMA methodology for the review and analyzed the next electronic databases: PUBMED, MEDLINE, and EMBASE. They concluded that the most common algorithms were tree-based methods (23, 53%), and neural network (NN) (14, 33%), and the range of variability of AUC reported by these studies was a median of 0.68 (IQR: 0.64–0.76). The purpose of the paper of Gatt et al. [38] was to identify and analyse the readmission risk prediction tools reported in the literature. They identified 34 readmission risk prediction models, with a predictive ability ranging from poor to good, and readmission rates ranged between 3.1 and 74.1%, depending on the risk category. They concluded that most prediction models were developed for specific populations, conditions or hospital settings, hence the generalisability and transferability of the predictions across wider or other contexts is difficult to achieve.

Taking into account the areas exposed, it is detected that the problem of measuring the risk of hospital readmission has been addressed in recent years using different methodologies, offering good results the application of ML models. However, nowadays it is still relevant to explain those results obtained through such models in such a way that they are understandable and facilitate the interpretability to the end user. Thus, an opportunity opens up to complement ML models with explainability approaches. This is even more important in ML approaches considered as "black boxes", in order to be able to interpret and explain their results.

Taking into account the review carried out, it is detected that the problem of measuring the risk of hospital readmission has been addressed in recent years using different methodologies, with the application of ML models offering good results. However, it can be detected that improvements are still required in the process, either by trying new prediction techniques [27], or by exploring with new paradigms 28, 29, 30]. Also, today it is still relevant to explain the results obtained through these models in such a way that they are understandable and facilitate interpretability by the end user. Therefore, an opportunity opens up to complement ML models with explainability approaches. This is even more important in ML approaches considered as "black boxes", to be able to interpret and explain their results.

III. THEORETICAL FRAMEWORK

A. Explainability Methods

Many may wonder why explainability in AI is important, and there are several answers to this. To begin with, the European Union, through Regulation 679 [18], gives the user the right to an explanation of a decision made, and more if it is made by AI algorithms. On the other hand, in contrast to traditional software, in many AI algorithms, it is not possible to show the users the logic behind the software by which it makes decisions. A misinterpreted and misapplied algorithm can lead to strange results. This lack of transparency can lead to rejection and distrust in the use of AI models.

There are interpretation methods that are independent of the ML technique used, that can be used in any of them, and that, in addition, are applied after the model has been trained (post hoc), so they are considered agnostic to the model. Such methods generally work by analyzing pairs of input and output features. By definition, these methods cannot have access to the internal components of the model, such as weights or structural information [19]. There are also model-specific interpretation methods, which depend on the internal structure of the model to get certain conclusions. These methods can interpret coefficient weights in generalized linear models (GLM), or weights and biases in the case of neural networks, among other things.

Separating the explanations of the ML model has some advantages [19]. The main one would be the flexibility to be able to apply in any model. Thus, when comparing the models used in terms of interpretability, it is easier to work with model-independent explanations, since the same method can be used for all models, which would not happen if the interpretation method were model-specific.

Inside the agnostic interpretability of the model, we find global and local methods. In accordance with [20], global methods describe the average behavior of an ML model, and are generally expressed as expected values based on the distribution of the data, and are useful when the modeler wishes to understand the general mechanisms in the data. The following are model–independent global interpretation techniques:

- *Partial Dependence Plot (PDP):* It is a diagram that shows the initial effect that one or two features have on the predicted result of an ML model. A partialdependence plot can show whether the relationship between the target and a feature is linear, monotone or more complex. For example, when it is applied to a linear regression model, the partial dependence graphs always show a linear relationship. The method considers all instances and gives a statement about the global relationship of a feature with the predicted result.
- *Accumulated Local Effects (ALE):* describe how the features influence in the prediction of an ML model on average. ALE plots average the changes in predictions and accumulate them on the grid. ALE calculates differences in the predictions, so that we replace the feature of interest with z-grid values. The difference in prediction is the effect the feature has for an individual instance at certain interval. The ALE value can be interpreted as the main effect of the feature at a certain value in comparison to the average prediction of the data. ALE's graphics are impartial, which means that they still work when the features are correlated.
- *Functional Decomposition:* A supervised ML model can be viewed as a function that takes a high-dimensional feature vector as input and produces a prediction of qualification or classification as output. Functional decomposition is an interpretation technique that deconstructs the high-dimensional function and expresses it as a sum of individual feature effects and interaction effects that can be visualized. Particularly, a prediction function takes p features as input $\hat{f}: R^p \rightarrow R$ and produces an output. It can be a regression function, but it can also be the classification probability of a determined class or the score of a determined group (unsupervised machine learning). Completely decomposed, it can represent the prediction function as the sum of the functional components. The number of possible sets to form is expressed as a combinatorial of the way $\sum_{i=0}^{p} \binom{p}{i} = 2^p$. For example, if a function uses 10 features, we can decompose the function into 1042 components.

The local methods are a set of techniques designed to answer questions such as Why did the model make this specific prediction? They are designed to explain how the label of a particular instance is predicted. To do so, they induce interpretable models in the neighborhood of the instance to explain. The following are agnostic methods of local explanation [20]:

- *Local Surrogate (LIME):* They are interpretable models that are used to explain the individual predictions of black-box ML models. Surrogate models are trained to approximate the predictions of the underlying black box model. Instead of training a global surrogate model, LIME focuses on training local surrogate models to explain individual predictions. LIME tests what happens to the predictions when it provides

variations of its data in the ML model. LIME generates a new data set, which consists of perturbed samples and the corresponding predictions from the black box model. On this new data set, LIME then trains an interpretable model, which is weighted according to the proximity of the sampled instances to the instance of interest.

- *Scoped Rules (Anchors):* The Anchors method explains the individual predictions of any black box classification model by finding a decision rule that "anchors" the prediction enough. Anchors uses reinforcement learning techniques in combination with an algorithm of graph search. Anchors' approach implements a perturbation-based strategy to generate local explanations for the predictions of black-box ML models. However, instead of surrogate models used by LIME, the resulting explanations are expressed as rules IF-THEN easy-to-understand called anchors. These rules are reusable since they have scope: anchors include the notion of coverage, indicating with accuracy to which other instances, possibly invisible, are applied. Finding anchors involves the problem of exploring the discipline of reinforcement learning. To this end, neighbors, or perturbations, are created and evaluated for each instance that is being explained. Doing so allows the approach to ignore the structure of the black box and its internal parameters, In order for these can remain unnoticed and unchanged. Therefore, the algorithm is independent of the model, which means that it can be applied to any kind of model.
- *SHAP (SHapley Additive exPlanations):* It is a method to explain individual predictions. SHAP is based on the Shapley values theoretically optimal from the game. The SHAP goal is to explain the prediction of an instance x by calculating the contribution of each feature to the prediction. The SHAP explanation method calculates Shapley values from coalition game theory. The feature values of a data instance play as players in a coalition. Shapley values tell us how to fairly distribute the "payoff" (= the prediction) among the features. A player can be an individual feature value. A player can also be a group of feature values. For example, to explain an image, pixels can be grouped into superpixels and the prediction distributed among them. One innovation that SHAP brings is that the Shapley value explanation is represented as an additive feature attribution method in a linear model. SHAP specifies the explanation as:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \qquad (1)$$

Where g is the explanation model, $z \in \{0,1\}^M$ is the coalition vector, M is the maximum coalition size and $\phi_j \in R$ is the Shapley value attribution for a feature j.

B. Stability metric (explainability)

The stability metric allows measuring the quality of an explanation, so it can be used to compare methods of explainability [21]. This metric has the main objective of measuring the confidence in the contribution values of each attribute in a given result, and thus determine if we can trust in an explanation. The idea behind local stability is the following: if the instances are very similar, we would expect the explanations to be similar as well. Therefore, locally stable explanations generate confidence around a method of explanation in particular Thus, for similar cases, are the explanations similar? Therefore, It is expected that:

- Instances must be close in feature space.
- The predictions of the predictive model considered must be close.

Both aspects can occur in different ways depending on the ML model used. For the calculation of the metric, the following steps are carried out:

- The data are normalized.
- The top N of the nearest neighbors for each instance is chosen by the L1 standard, also known as the Manhattan distance.
- Those neighbors whose model output is very different from that of the instance (distances of less than 10% are considered) are rejected. For this, the maximum allowed difference is calculated from the following expression:

$$|Output_{instance} - Output_{neighbor}| < 10\% \qquad (2)$$

- Finally, a graph is generated where the neighborhood around each instance (neighborhood, in terms of features and model results) is analyzed. In this graph, the Y-axis shows the importance of each feature in the data set as a function of its contributions (using the SHAP value), such that the higher the value, the more important the feature. On the other hand, the X-axis shows the average variability of the feature in its instances neighborhood, and is calculated as:

$$Variability = \frac{\sum_1^N (x_i - \bar{X})^2}{N} \qquad (3)$$

Being x_i the prediction of each instance in the neighborhood, $\bar{X}$ the average of the predictions of the instances, and N the total number of instances in the neighborhood. The farther to the right, the more distant their instances are from each other, therefore more unstable the feature is.

- Then, it is intended to measure how important the feature is versus its variability. For the above reasons, stability can be expressed according to the following relationship:

$$Stability = \frac{Importance\ (Shap\ value)}{Variability} \quad (4)$$

The greater the importance and the lower the variability, the greater the stability of this characteristic in the explanation of the model.

IV. PREDICTIVE MODELS ON RE-ENTRY RISK HOSPITAL

The problem of developing predictive models to determine the risk of hospital readmission is addressed using the CRISP–DM (Cross-industry Standard Process for Data Mining) methodology [24]. In the following sections, the results obtained from its implementation will be presented.

A. *Business understanding*

For the development of this work, information is extracted concerning hospital discharges generated during July 2021 and September 2022 for a health institution from an anonymized database [26]. Of these, it is identified which corresponded to a re-hospitalization based on a mark in the system by the medical staff at the moment of hospitalization, and those that preceded them (which will be the focus of interest). This database consists of the following:

- Number of records (hospitalizations): 367,656
- Number of variables: 139
- Quantitative variables: 111
- Qualitative variables: 28
- Target variable: 1, which corresponds to the hospital readmission indicator (if the hospital event occurred during the first 15 days following discharge of the same patient).

Figure 2 shows the total number of hospital discharges generated in the period mentioned above.

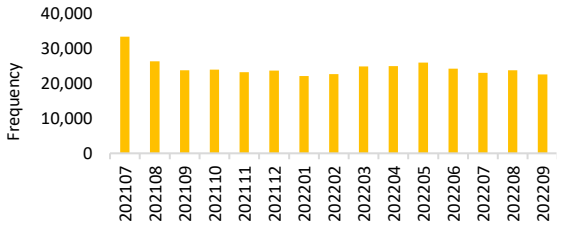

Figure 2. Behavior of hospital discharges from July 2021 to September 2022

Of these data is identified those events that correspond to hospital admissions occurring during the first 15 days after discharge. The behavior of these readmissions is shown in Figure 3.

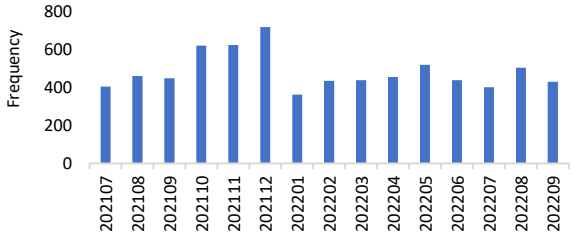

Figure 3. Behavior of hospital readmissions from July 2021 to September 2022

Given these frequencies, the participation of readmissions on the total hospital discharges, their cost is analyzed (see Figure 4), and the hospital readmission rate is calculated as follows:

$$Rehospitalization\ rate = \frac{Readmitted\ members}{Exposed\ members} * 100{,}000\ exposed \quad (5)$$

An exposed member is a member who is active and, given his/her conditions and level of risk, is susceptible to the materialization of an occurrence.

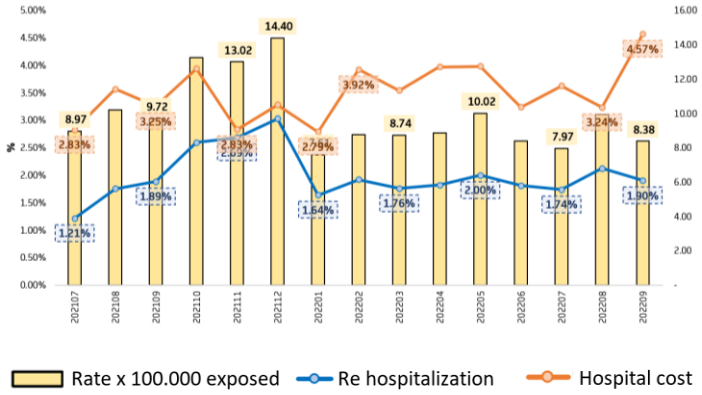

Figure 4. Rate, participation and cost measured in % of readmissions between July 2021 and September 2022

From January 2022, a stable behavior is observed in terms of readmissions (between 1.64% and 2.00% with respect to total hospital discharges), with a cost from 2.79% to 4.57% with a slightly increasing trend.

B. *Data preparation*

Initially, A filter of available information is made to consider the data between February and September 2022. This filter results in 191,644 records, of which 188,031 correspond to hospital discharges, which are not hospital readmissions and 3,613 records are associated with hospital readmissions.

Then, a process of atypical data identification (outliers) is undertaken to find those variables that are the furthest from the

center. The dataset correlation array is calculated. In this regard will be taken into account the next interpretation [14, 15, 17]:

- Values close to 0 indicate a low degree of correlation. Therefore, it can be said that the variables are independent of each other.
- Values close to 1 indicate a high degree of correlation, so it can be said that the variables are linearly dependent, so that the information provided by one of them corresponds to the same extent as that provided by the other. This will be important at the time of doing dimensionality reduction.

Thus, the Pearson correlation coefficient is calculated. Afterward, the covariance array is calculated, identifying the covariance among each pair of variables [25].

Next, a class balancing process is carried out taking into account the imbalance between the number of hospital readmissions for the period (February to September 2022) and the total number of hospital discharges; these readmissions represent 1.89% of the resulting information. This may cause the recall metric for class 1 results close to 0%, which increases the probability that the predictive model misclassifies the new records, favoring the presence of False Positives (this means, that the model assigns a discharge a high probability of hospital readmission during the next 15 days when it really is not). For this reason, balancing techniques are considered to obtain a better Recall. The techniques implemented were:

- NearMiss Subsampling of the majority class: Through this method, the aim is to balance the number of records for the classes by reducing the majority class.
- Random Oversampling of the minority class: In this case, synthetic samples of the minority class are created.

Of the ones already mentioned, Random Oversampling was used as the technique for balancing the minority class (hospital readmissions) since many records would be eliminated with Subsampling (98.11% of the information).

Finally, the resulting dataset is divided into 80% as training data and 20% for the evaluation of the predictive models obtained.

V. EXPERIMENTS

A. General Results

A series of experiments will be carried out implementing Grid Search and Cross Validation with k fold = 5, to optimize the parameters of each model to obtain the best results from each one [4, 23]. The Table shows the models used along with the different metrics that were calculated for each one.

Table 1 shows that the decision tree model has good accuracy (83%), AUC (92%), Recall (84%) and good precision (82%). Using the F1 - Score as the reference metric for the selection of the best model, it is observed that the model based on decision trees presents the best result (83%). Additionally, its computation time is minimal, in comparison to the Light Gradient Boosting Machine and, even more, the Ada Boost Classifier.

Table 1. Métricas resultantes de los modelos predictivos

#	Model	Accuracy	AUC	Recall	Prec.	F1	Computing Time (Seg)
1	Decision Tree Classifier	**0.83**	**0.92**	**0.84**	**0.82**	**0.83**	**1.98**
2	Light Gradient Boosting Machine	0.68	0.75	0.66	0.68	0.67	4.16
3	Ada Boost Classifier	0.62	0.67	0.59	0.63	0.61	11.62
4	Logistic Regression	0.60	0.65	0.51	0.62	0.56	1.70

The confusion array for prediction is calculated with this model (see Figure 5):

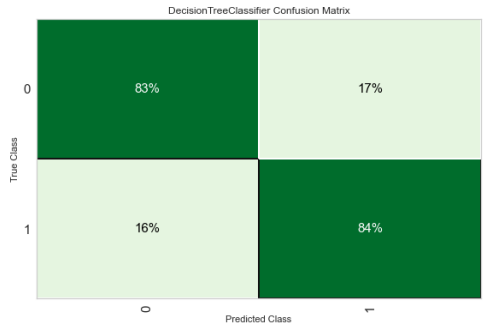

Figure 5. Confusion array (in %) with decision tree classifier

Class 0 represents the hospital discharge that is not preceded by hospital readmission, and 1 represents the hospital discharge that is preceded by hospital readmission. Generally, TP (True Positive) has 83%, TN (True Negative) has 17%, FP (False Positive) has 16% and FN (False Negative) has 84%.

Finally, the risk of hospital readmission to the 4,497 hospital discharges between 1 and 15 of January of 2023 is calculated based on the decision tree outcoming model. Zooming on this population finds the following:

- 70% of the population is between adulthood and old age.
- 79% of the population is sick.
- 61% had 1 or more admissions previous to hospitalization.
- 58% have more than 1 associated mark.
- 57% are polymedicated.
- 11% have hypertension and 9% have cancer.
- 8% were discharged with urinary tract infections.
- 95% were emergency hospitalizations.

B. *Results by chronic populations*

The above process is replicated for the 5 most costly comorbidities based on the source data obtained (see Table 2).

Table 2. Resulting metrics by pathology of chronic patients using decisión

Population	Accuracy	AUC	Recall	Prec.	F1	TT (Sec)
no label	77.76%	87.65%	77.88%	77.75%	77.80%	0.59
Hypertension	90.34%	95.65%	95.01%	86.78%	90.71%	0.07
Dyslipidemia	89.26%	95.26%	94.56%	85.50%	89.76%	0.05
Cancer	92.11%	94.92%	96.46%	88.85%	92.50%	0.04
Hypertension - Diabetes	92.74%	96.06%	97.51%	89.07%	93.10%	0.03
Hypertension - Dyslipidemia	90.40%	96.03%	95.58%	86.71%	90.93%	0.02

Pretty good results are obtained in the evaluation of the prediction for each type of chronic population.

C. *Explainability Analysis*

1. General Explainability Analysis

For this explanatory analysis, the best model obtained in the previous section is used. Initially, the ranking of the decision tree on the importance of the predictors on the result of the prediction is presented. Only the top 10 variables are shown by weight (see Figure 6).

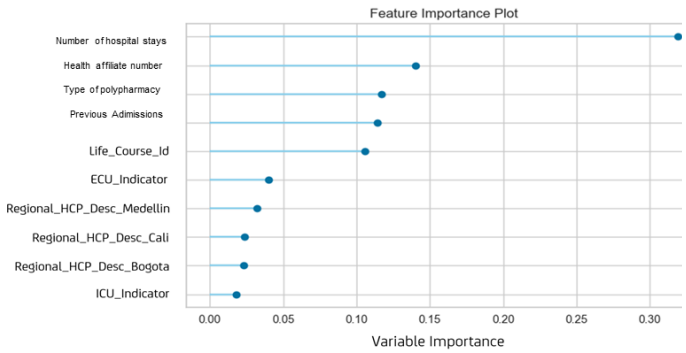

Figure 6. Importance of predictors in the decision tree model

It is identified that in the prediction of hospital readmission with the supplied data turn out that the variable with the greatest weight on the prediction is the "number of hospital stay days" (33%) followed by the "number affiliated health" (diseases) (14%), polymedicated level (12%), "previous admissions to hospitalization" (12%), "life course" (related to the age) (11%), and finally, other variables that together represent the remaining (18%) of the prediction.

Additionally, it is applied local explicability techniques on the outcoming model (decision trees) to understand how the individual predictions are explained. With this purpose, the method of values of Shapley based on traits is used. The calculations of the Shapley Values on the Decision Trees model are visualized in Figure 7.

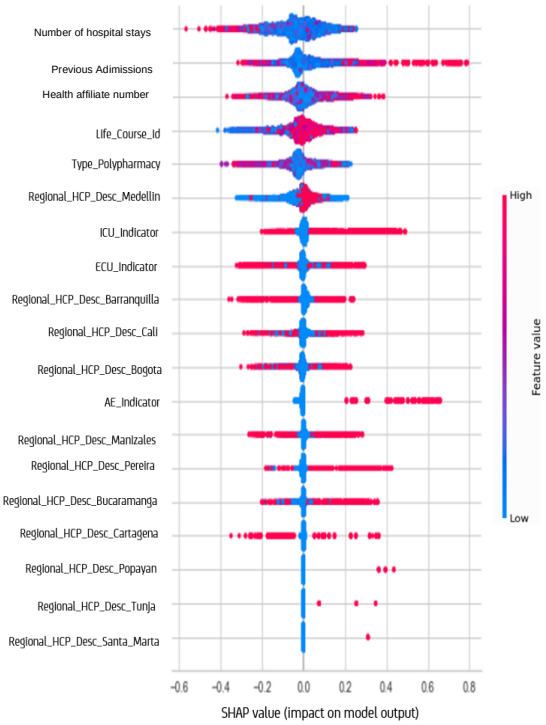

Figure 7. Shapley values for the model based on decision trees

To interpret the results in Figure 7, the following is taken into account:

- The order of the variables indicates which one predominates in terms of weight in the prediction of the model. The variables "number of hospital stays", "previous admissions", " health affiliate number", and "life course", are the most important variables.
- The color of the dots in each variable refers to the value of each instance. Thus, for the life course variable (which is an ordinal grouping of the affiliated member's age), the blue instances refer to low values, which means low ages (early childhood and infancy). On the other hand, the red instances refer to high values (adulthood and old age). The intermediate color corresponds to population groups in adolescence and young

adulthood age.

- In terms of the horizontal axis, the SHAP value measures the impact of the input on the output (prediction) of the model. To the left it indicates a negative impact with respect to the class, which means there is a low probability of hospital readmission; and to the right, it indicates a positive impact, which means there is a high probability of hospital readmission.

Figure 7 provides the following conclusions in terms of explainability:

- The "number of hospital stays" is the variable with the highest weight in the prediction. For low values, the probability of hospital readmission tends to be slightly higher than when the patient stays hospitalized for several days.
- Regarding previous admissions to hospitalization, the more previous admissions the patient has had, the higher the probability of readmission is and vice versa.
- In terms of the "health affiliate number" (diseases) associated with the patient, the ones that don't have any or a few number of marks, don't have an important role in the prediction; meanwhile, those that have associated marks can impact positively or negatively in the prediction.
- In terms of "life course", the older the patient is, the highest the predisposition to readmission to hospitalization.
- Regarding the indicator of whether the person was admitted to the ICU, those who were admitted have a higher risk of readmission.
- The same for those events that were marked as adverse events, if the event was marked in this way the risk of hospital readmission is higher.

By comparing the results obtained in the measurement of the importance of the features using the ranking of the predictive model (Figure 6) and the Shapley Values (Figure 7), it can be observed that both methods conclude that the features "Number of hospital stay" is the most important one to explain the risk of hospital readmission from the data considered. Besides, the top 5 most important variables according to each method contain the same attributes. However, the importance order changes for the following attributes. For example, while the " health affiliate number" is more important than "Previous admissions" according to the decision tree ranking, the opposite happens for the Shapley Values. The same situation occurs when comparing the variables "polymedicated Type" with "Life course", being more important in this order in the decision tree ranking, and vice versa in the Shapley Values. In conclusion, both methods give the same variables as relevant, with different orders of relevance.

2. Quantitative comparison of Explainability Analysis

To quantitatively compare the explainability analysis of different ML methods, a sample of 3,000 instances is taken and a stability metric is used to assess the quality of the explanations (see Figure 8 for the decision tree method). The step-by-step explanation of the representation of each instance of the stability graph is explained in III.B section.

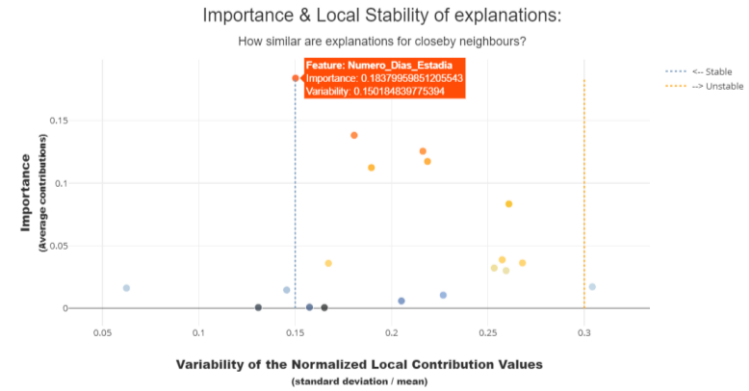

Figure 8. Stability metrics on the model based on decision trees

Based on the aforementioned, we can identify that for example, the variable "Number of hospital stays" is indeed the most important variable in the prediction, besides of being stable (it does not overpass the stability limit). Of the 19 variables:

- 4 fulfill the stability standard (21%).
- 14 are relatively stable in average (74%).
- 1 don't fulfill the standard (5%).

We compare the results obtained with the LGBM classifier, which was the second best model according to the metrics, with an F1-Score (67%). The importance of the predictors on the prediction (based on permutations) is calculated for this new model, see Figure 9.

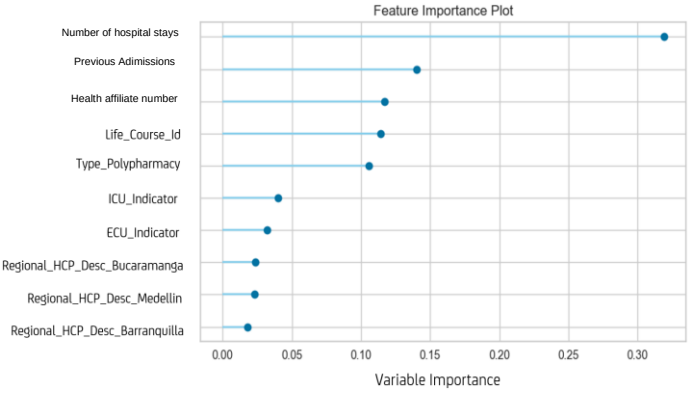

Figure 9. Importance of predictors in the model based in LGBM

It can be seen that the "Number of hospital stays" continues being a predominant variable in the explanation of the model,

followed by the "Previous admissions". Similarly, the calculation of the Shapley Values on LGBM is presented (see Figure 10).

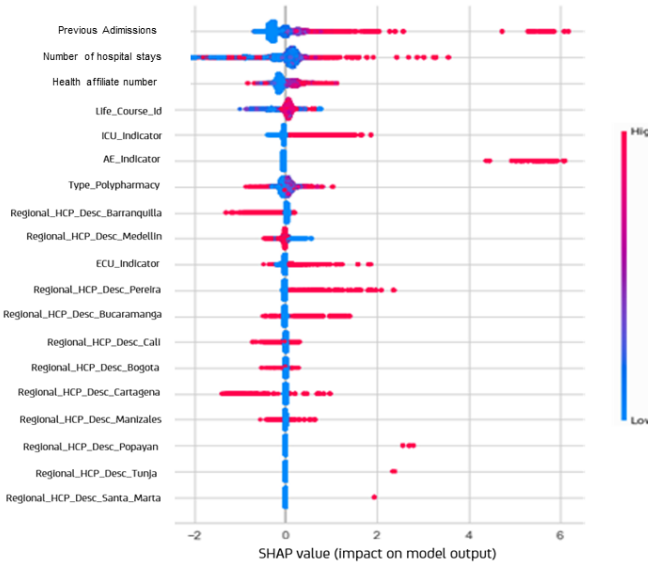

Figure 10. Shapley Values in the LGBM-based Model

Table 3 shows the differences between the two models:

Table 3. Comparison between the important variables according to the Shapley values for each model

Order of importance	Decision Tree	LGBM
1	Number of hospital stays	Number of hospital stays
2	Previous admissions	Previous admissions
3	Health affiliate number	Health affiliate number
4	Life course	Life course
5	Type of polypharmacy	ICU indicator

It is observed that for both the "Number of hospital stays" is the more important in the prediction. Similarly, polymedicated type is an important variable in the top 5 for the decision tree, but is not for LGBM, as is the ICU Indicator. As for the impact (positive or negative) on prediction, we can also observe differences such as:

- The positive impact on the risk of hospital readmission of the instances with high values of the "number of stay days" in LGBM is more evident than in the decision tree.
- Similarly, the positive impact of the "health affiliate number" associated with a patient on the risk of hospital readmission in LGBM is more noticeable.
- The ICU indicator also shows a more expressive behavior in terms of impact: negative on the prediction for minimum values (meaning, 0 when the patient was not admitted to the ICU) and positive for maximum values (1, when the patient was admitted) in the LGBM.

The stability metric is calculated for this model (see Figure 11).

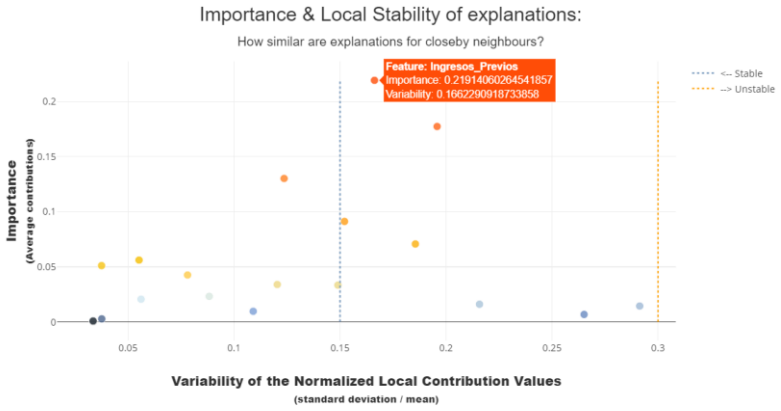

Figure 11. Stability metrics based on the LGBM model

Of the 19 variables:

- 12 fulfill stability standard (63%),
- 7 are relatively stable on average (27%),

Unlike the decision tree, in LGBM all variables fulfill the stability standard, and a large proportion are very stable. Thus, we can conclude that LGBM presents more stable variables, so we can have more confidence in using it to give explanations.

VI. ANALYSIS OF THE RESULTS

According to the obtained results, it is evident that almost 2.06% of hospital discharges emanate from readmissions. As well, 3.49% cost of hospital discharges corresponds to readmissions. On the other hand, among the predictive models, the one based on decision trees stands out with 83% reliability. Its relevant variables to explain the risk of readmission were: hospital stays (33%), previous admissions (14%), health affiliate number (14%), polymedicated type (12%), and (12%) and life course (11%).

According to the best model, 5.63% of the total number of patients discharged within 15 days have a high risk (between 80% and 100%) of hospital readmission. On the other hand, in the selected populations of patients with chronic diseases, the predictive model of the risk of hospital readmission obtains excellent results (on average, the F1 - Score is 80%).

With respect to the Shapley values to explain the weight of the variables in the prediction, those of higher importance in the prognosis of readmissions were identified, as well as their impact

(positive or negative) on the risk of readmission. For example, according to the results in Figure 7, again, the “number of hospital stays” is the variable with the highest weight in the prediction. Also, for low values of this variable, the probability of hospital readmission tends to be slightly higher than when the patient stays hospitalized for several days. It is also compared the explainability analysis of this model with the one constructed with LGBM (67% F1-Score). A close behavior of the Shapley values was obtained with respect to the model based on decision trees. For example, in Figure 9, the positive impact of the variables "Previous admissions" and "Number of hospital stay days" on the prediction.

For the stability metric, in terms of explainability, its value is better for the LGBM model, given that 100% of its predictors turn out to be stable vs. 95% of the decision tree model. This is interpreted as there is a low variability in the prediction instances with the LGBM model, which favors the reliability in the results. Therefore, and based on the proposed objective of good quality and explainability, this model is the best since each variable is stable with respect to its explanation.

VII. CONCLUSIONS

On the basis of the proposed objectives, an ML model was built to predict predictable hospital readmissions in the population of patients. The best model corresponds to a classifier based on decision trees with an F1-Score of 83% followed by a model based in LGMB with an F1-Score of 67%. For these two models, Shapley values were calculated as a method of explainability from which some differences were found.

Overall, the impact explanation on the prediction of the risk of hospital readmission of the LGBM model is more in line with the results obtained according to the Shapley values. For example, for the "Number of hospital stay days", high values (corresponding to longer in-hospital days) correspond to a positive impact on the risk of readmission, unlike the explanation resulting from the Shapley values based on the model based on binary trees, where the more days, the lower the risk of readmission. This same behavior is evidenced for the attribute " health affiliate number", where for the model based on LGBM, the more comorbidities the patient has, the higher the risk of readmission. This is different from the explanation of the model based on the decision tree, which was not conclusive. The same is true for other attributes such as ICU indicator and previous admissions.

Concerning the metric used to evaluate the quality of the explainability of the predictive model, the stability metric, which is based on the importance of the variable and its variability in its neighborhood, was used to assess its explanations. According to this metric, more variability is evidenced in the explanations by the model based on decision trees, where only 4 attributes are very stable (21%) and 1 attribute is unstable. With respect to the LGBM-based model, there are 12 stable attributes (63%) and no unstable attributes. Thus, according to the stability metric on the explanations generated by the two best predictive models, it is concluded that in terms of explainability, the LGBM-based model is better. This metric indicates that the LGBM-based predictive model presents more stable variables, thus generating greater confidence in the explanations it provides.

As future works, it is considered the implementation of techniques based on Deep learning for the hospital readmission risk calculation, and the utilization of other kinds of explanatory methods (global or local), to cross-check and determine in which context each one has to be used.

DECLARATIONS

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Authors' contributions

Concept and design: All authors; Acquisition, analysis, or interpretation of data: Lopera; Drafting of the manuscript: All authors; Results analysis: All authors; Obtained funding: Aguilar.

Funding

Jose Aguilar was partially supported by grant 22-STIC-06 (HAMADI 4.0 project) funded by the STIC-AmSud regional program.

Ethics statement

The study was conducted in accordance with relevant guidelines and regulations, and approved by the EAFIT University ethics committee.

Consent to participate

The Sura health center has signed an anonymized data use agreement with the EAFIT University.

Data Availability Statement

The datasets used in the current study are available from the corresponding author on reasonable request.

VIII. REFERENCES

[1] S. Jencks, M. Williams y E. Coleman, “Rehospitalizations among patients in the Medicare fee-for-service”, N. Engl. J. Med., 360, 1418-1428, 2009.

[2] Kansagara D, “Risk prediction models for hospital readmission, a systematic review”, JAMA 306 (15), 1688–1698, 2011.

[3] D. Insight, “56% of Hospitals Lack Big Data Governance”, Analytics Plans, Health It analytics. 2017. [Online]. Available: https://healthitanalytics.com/news/56-of-hospitals-lack-big-data-governance- analytics-plans.

[4] W. Hoyos, J., Aguilar, M. Toro, “A clinical decision-support system for dengue based on fuzzy cognitive maps”. Health Care Manag Sci, 25, 666–681, 2022.

[5] Y. Quintero, D. Ardila, E. Camargo, F. Rivas, J Aguilar, “Machine learning models for the prediction of the SEIRD variables for the COVID-19 pandemic based on a deep dependence analysis of variables, Computers in Biology and Medicine, 134, 2021.

[6] J. Jaana, “The Diabetes Risk Score: A practical tool to predict type 2 diabetes risk,” Expert Systems with Applications, 26 (3), 725- 731, 2003.

[7] M. Ortiz, Z. Altamar, C. Martínez, A. Petrillo, G. Jiménez, A. García y A. Medina, “Predicting 15-day unplanned readmissions in hospitalization departments: an application of logistic regression”. Ingeniare. Revista chilena de ingeniería, 29 (2), 378-398, 2021.

[8] P. Michailidis, A. Dimitriadou, T. Papadimitriou y P. Gogas. “Forecasting Hospital Readmissions with Machine Learning”. Healthcare 10, 981. 2022.

[9] D. Zhang y J. Lee, “Effective hospital readmission prediction models using machine-learned features”. BMC Health Serv Res 22, 1415. 2022.

[10] G. Arkaitz, “Predictive models for hospital readmission risk: A systematic review of methods”, Computer Methods and Programs in Biomedicine. 164, 49-64, 2018.

[11] H. Langs, G. Denk y K. Müller, “Causability and explainability of artificial intelligence in medicine”, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 9 (4), 2019.

[12] N. Burkart y M. Huber, “A survey on the explainability of supervised machine learning”. J. Artif. Intell. Res. 70: 245–317. 2021.

[13] W. Hoyos, J. Aguilar, M. Toro, “Dengue models based on machine learning techniques: A systematic literature review, Artificial Intelligence in Medicine, 119, 2021.

[14] A. Breiman, “Classification and Regression Trees”, New York, 1984.

[15] L. Breiman, “Statistical modeling: The two cultures (with comments and a rejoinder by the author)”, Statistical Science 16 (3): 199-231, 2001.

[16] F. Junliang, M. Xin, W. Lifeng, Z. Fucang, Y. Xiang y Z. Wenzhi, “Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data”, Agricultural Water Management 225, 2019.

[17] Y. Freund, R. Schapire, “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting” Journal of Computer and System Sciences 55(1):119-139, 1997.

[18] Unión Europea. “Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo”, Madrid. 2016. [Online]. Available: https://www.boe.es/doue/2016/119/L00001-00088.pdf

[19] C. Molnar, “Interpretable machine learning. A Guide for Making Black Box Models Explainable”. Leanpub. 2019.

[20] M. Ribeiro, S. Singh y C. Guestrin, “Model-agnostic interpretability of machine learning”. Chapter 6, In Interpretable Machine Learning: A Guide For Making Black Box Models Explainable (C. Molnar ed.), Independently published, 2022.

[21] R. Marco, S. Sameer y G. Carlos. "Why Should I Trust You? Explaining the Predictions of Any Classifier”. International Conference on Knowledge Discovery and Data Mining. 2016.

[22] Camargo, E., Aguilar, J., Quintero, Y. F. Rivas & D. Ardila An incremental learning approach to prediction models of SEIRD variables in the context of the COVID-19 pandemic. Health Technol. 12, 867–877, 2022.

[23] J. Vizcarrondo, J. Aguilar, E. Exposito and A. Subias, "ARMISCOM: Autonomic reflective middleware for management service composition," *Global Information Infrastructure and Networking Symposium (GIIS)*, 2012.

[24] C. Shearer. “The CRISP-DM Model: The New Blueprint for Data Mining”, Journal of Data Warehousing 5, 13-22. 2000.

[25] O, Ledoit, M. Wolf. “Honey, I Shrunk the Sample Covariance Matrix”. Journal of Portfolio Management 30, 110-119. 2004.

[26] Anonymized database, https://www.epssura.com/

[27] M. Araujo, J. Aguilar, H. Aponte, "Fault detection system in gas lift well based on artificial immune system," Proc. International Joint Conference on Neural Networks, pp. 1673-1677 vol. 3, 2003.

[28] J. Aguilar, M. Jerez, E. Exposito, T. Villemur, "CARMiCLOC: Context Awareness Middleware in Cloud Computing," Proc. *Latin American Computing Conference (CLEI)*, 2015.

[29] L. Morales, C. Ouedraogo, J. Aguilar, C. Chassot, S. Medjiah, K. Drira. Experimental comparison of the diagnostic capabilities of classification and clustering algorithms for the QoS management in an autonomic IoT platform. Service Oriented Computing and Applications, 13, 199–219, 2019.

[30] M. Sánchez, J. Aguilar, J. Cordero, P. Valdiviezo-Díaz, L. Barba-Guamán, L. Chamba-Eras, “Cloud Computing in Smart Educational Environments: Application in Learning Analytics as Service”. In: Rocha, Á., Correia, A., Adeli, H., Reis, L., Mendonça Teixeira, M. (eds) New Advances in Information Systems and Technologies. Advances in Intelligent Systems and Computing, 444, 2016.

[31] M Baig, N. Hua, E. Zhang, R. Reece, A. Spyker, D. Armstrong, R. Whittaker, T. Robinson, E. Ullah. “A machine learning model for predicting risk of hospital readmission within 30 days of discharge: validated with LACE index and patient at risk of hospital readmission (PARR) model”. Med Biol Eng Comput, 58, 1459–1466, 2020.

[32] Y. Lo, J. Liao, M. Chen, C. Chang. C. Li “Predictive modeling for 14-day unplanned hospital readmission risk by using machine learning algorithms”. BMC Med Inform Decis Mak, 21, 2021.

[33] M. Ko, E. Chen, A. Agrawal, P. Rajpurkar, A. Avati, A. Ng, S. Basu, N. Shah, “Improving hospital readmission prediction using individualized utility analysis”, Journal of Biomedical Informatics, 119, 2021.

[34] P. Zhao I. Yoo S. Naqvi, “Early Prediction of Unplanned 30-Day Hospital Readmission: Model Development and Retrospective Data Analysis”, JMIR Med Inform, 9(3), 2021.

[35] M. Afrash, H. Kazemi-Arpanahi, M. Shanbehzadeh, R. Nopour, E. Mirbagheri, “Predicting hospital readmission risk in patients with COVID-19: A machine learning approach”, Informatics in Medicine Unlocked, 30, 2022.

[36] Y. Shang, K. Jiang, L. Wang, L. Z. Zhang, S. Zhou, Y. Liu, J. Dong, H. Wu “The 30-days hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers”. BMC Med Inform Decis Mak. 21, 2021.

[37] Y. Huang, A. Talwar, S. Chatterjee, R. Rajender, R. Aparasu. “Application of machine learning in predicting hospital readmissions: a scoping review of the literature”. BMC Med Res Methodol 21, 2021.

[38] M. Gatt, M. Cassar, S. Buttigieg, "A review of literature on risk prediction tools for hospital readmissions in older adults", Journal of Health Organization and Management, 36 (4), 521-557. 2022