UNIVERSITY CEU - SAN PABLO

POLYTECHNIC SCHOOL

BIOMEDICAL ENGINEERING DEGREE

BACHELOR THESIS

CRITICAL EVALUATION OF EXPLAINABLE MACHINE LEARNING

Author: María Palacios Pinedo
Supervisors: Constantino Antonio García Martínez

July 2021

UNIVERSIDAD SAN PABLO-CEU
ESCUELA POLITÉCNICA SUPERIOR
División de Ingeniería

Datos del alumno

NOMBRE:

Datos del Trabajo

TÍTULO DEL PROYECTO:

Tribunal calificador

PRESIDENTE:	FDO.:

SECRETARIO:	FDO.:

VOCAL:	FDO.:

Reunido este tribunal el _______/_______/_______, acuerda otorgar al Trabajo Fin de Grado presentado por Don___________________________________la calificación de ___________________.

ACKNOWLEDGMENTS

Finalizar esta tesis implica el final de un periodo muy especial e importante en mi vida. Han sido muchas las dificultades encontradas por el camino, pero todas ellas han sido superadas contando con el apoyo de las personas que me han ido acompañando a lo largo de estos años.

En primer lugar, quisiera agradecer el tiempo, esfuerzo y sobre todo apoyo de mis padres y hermano. Han sido ellos los primeros conocedores de mis inquietudes y dificultades y siendo quienes han estado siempre en los momentos más duros, mostrándome una confianza infinita, sin dejarme caer y apoyándome a seguir hacia delante siempre.

Me gustaría también agradecer a todos mis amigos, compañeros y profesores que me han ido acompañando durante esta etapa. En ellos también me he apoyado desde los inicios, he entablado grandes amistades y me han animado y alentado a continuar siempre trabajando sin rendirme nunca.

Por último, me gustaría agradecer en especial a mi tutor Constantino Antonio García Martínez por brindarme la oportunidad de trabajar en este proyecto junto a él. Gracias por dedicarme parte de su valioso tiempo y esfuerzo, ayudándome siempre con un constante aliento y agradable asesoramiento.

ABSTRACT

Nowadays, machine learning permits to build systems able to learn how to make very accurate predictions from loads of data. However, most of these models are very complex and hence, the reasons behind predictions cannot be fully understood. This prevents the adoption of such black-box machine learning models in sensible areas, such as healthcare. To tackle this issue, “explainable” machine learning models that try to explain black-box models have been created. However, there are critical voices concerned that explainable methods can be more harmful than helpful in facilitating machine learning adoption. This project aims to provide empirical evidence on actual cases where the use of explainable machine learning algorithms may result in harmful situations in the context of healthcare data. For finding these critical situations, LIME and SHAP algorithms were extensively tested on three different datasets, which were screened for explanations that inaccurately represent the underlying predictive model. The results for each of the experiments are presented supporting that explainable methods may 1) undermine trust on the underlying predictive model by providing explanations that are not consistent with medical knowledge or 2) boost confidence on wrong predictions by providing sound explanations.

RESUMEN

Hoy en día, el aprendizaje automático permite construir sistemas capaces de aprender a hacer predicciones muy precisas a partir de montones de datos. Sin embargo, la mayoría de estos modelos son muy complejos y, por lo tanto, no se pueden entender del todo las razones de las predicciones. Esto impide la adopción de estos modelos de aprendizaje automático de caja negra en ámbitos sensibles, como la sanidad. Para hacer frente a este problema, se han creado modelos de aprendizaje automático "explicables" que tratan de explicar los modelos de caja negra. Sin embargo, hay voces críticas que temen que los métodos explicables puedan ser más perjudiciales que útiles para facilitar la adopción del aprendizaje automático. Este proyecto pretende aportar pruebas empíricas sobre casos reales en los que el uso de algoritmos explicables de aprendizaje automático puede dar lugar a situaciones perjudiciales en el contexto de los datos sanitarios. Para encontrar estas situaciones críticas, los algoritmos LIME y SHAP se sometieron a pruebas exhaustivas en tres conjuntos de datos diferentes, que se examinaron para detectar explicaciones que representaran de forma inexacta el modelo predictivo subyacente. Los resultados de cada uno de los experimentos se presentan apoyando que los métodos explicables pueden 1) socavar la confianza en el modelo predictivo subyacente al proporcionar explicaciones que no son consistentes con el conocimiento médico o 2) aumentar la confianza en las predicciones erróneas al proporcionar explicaciones sólidas.

INDEX

FIGURE INDEX

TABLE INDEX

1 INTRODUCTION

Artificial intelligence (AI), as the name denotes, refers to the intelligence evidenced by machines and possesses a great potential to improve both private and public life. Constantly discovering patterns and structures in large amounts of data in an automated manner is a fundamental component of data science, and currently drives applications in diverse areas such as computational biology, law, and finance [1]. These systems are referred to as Classic Machine Learning or just Machine Learning (ML) plainly. Furthermore, for tasks such as prediction, simulation, and exploration, ML and AI have become indispensable tools in the sciences. They have great ability to handle, operate, and draw conclusions with immense amounts of data. In addition, its great precision in performing tasks such as image classification, exceeds human capabilities. Scientific understanding, inferring causal relationships from observational data, or even gaining new scientific insights are various of the main goals for using ML [2].

Nowadays, one of the most relevant techniques is Deep Learning, which is a growing trend in general data analysis. It refers to an improvement of artificial neural networks, consisting of more layers that permit higher levels of abstraction and improved predictions from data [3]. Improvements in this methodology, along with the availability of large databases, and computational gains obtained with powerful GPU cards, contributed to the immense successes of AI systems [4].

Machine learning models are opaque, non-intuitive, and difficult for people to understand. This problem leads to what they are considered as "black-boxes", or in other words, the inability to explain how a decision has been arrived or to show the procedures that have led to them.

Such lack of transparency may be not acceptable and hence the development of methods for visualizing, explaining, and interpreting deep learning models has recently attracted increasing attention [4]. This is particularly relevant in those environments where sensitive decisions need to be made, such as hospitals. This lack of transparency and inability to be understood and interpreted by humans is what limits the use of these models.

Therefore, understanding ML model behaviour beyond conventional performance metrics has become a necessary component of ML research, especially in healthcare. This has led to the development of “interpretable” or “explainable” machine learning models as a measure to overcome barriers of trust and adoption and this is the reason why Explainable Machine Learning (EML) models emerge. Among the EML models of interest for this project, the focus will be set on those that try to explain the predictions of a classic model without any assumption on its inner work.

In the field of medicine, the relevance of an interpretable and transparent decision is essential for both the patient and the clinician. Clinicians seek transparency and both reasoned and reasonable responses from machine learning models [5]. Moreover, patients have the right to the privacy of their data and a reliable, unprejudiced, and impartial response, and always certainly contrasted with a medical professional [6]. This demand for supervision by a clinical professional, shows how important the machine-human relationship is, where both parties benefit [7].

Moreover, when medical responses are made, lives may be at stake. To leave such important decisions to machines that could not provide accountabilities would be akin to shirking the responsibilities altogether. Apart from ethical issues, this is a serious loophole that could turn catastrophic when exploited with malicious intent [8].

On the other hand, throughout the exploratory interviews achieved in the study [5], it was clear that clinicians viewed explainability as a means of justifying their clinical decision-making (for instance, to patients and colleagues) in the context of the model’s prediction. To provide these explanations all clinicians expressed the need to understand the clinically relevant model features that align with current evidence-based medical practice.

Even though the application of AI and interpretable machine learning in the field of the medicine has its origins in the earliest days of area, it is only in the latest years when there has been an impetus towards the acknowledgement of the necessity to have healthcare solutions propelled by machine learning. Although much progress has been made, there are critical voices concerned that false explanations can be given. This project aims to look for biomedical problems where explainable machine learning fails. Figure 1 illustrates the interaction between the ML algorithm and the Explainable Machine Model (EML), which provides the intuition about when the EML model may fail.

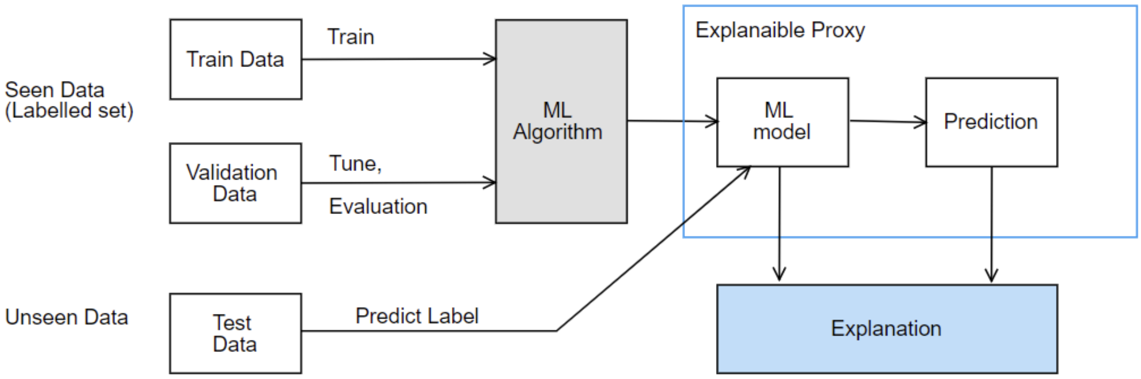

Figure 1. Explanatory scheme on which we will base the project

The ML model consists of the base classifier or predictor model, which will carry out the predictions. Then, on top of that, an explanatory model exists, which will try to explain the predictions that the first model provides. However, with this scheme a problem arises; what happens if the second model fails to capture what the first model is really doing?

In the field of the medicine, we are more interested in looking for two concrete situations. On one hand, we seek for scenarios in which we have correct predictions but bad explanations, referring to them as the explanations which are not plausible from a medical reasoning or that may deceive the physicians, being the real explanation a completely different one. On the other hand, we search for scenarios which consist of incorrect predictions but reasonable explanations. One simple example for this context is a patient suffering from heart attack or not. We encounter a prediction which stands that the patient is not suffering a heart attack. If the patient is actually suffering it, and the EML algorithm provides sound explanations able to convince the clinician, lives may be at stake. What it may be happening is that the patient does not have chest pain and seems fine. If the patient is a male this is correct but if the patient is a female, normally they do not suffer this symptom when having a heart attack.

Criticism to the EML scheme (see Figure 1) have already been considered [5],[9],[10].

In [5], clinicians repeatedly identified that knowing the subset of features deriving the model outcome, is crucial. This allows them to compare model decision to their clinical judgment, especially in case of a discrepancy. In time-constrained settings such as the Emergency Department, important features are perceived as a crucial metric to draw the attention of clinicians to specific patient characteristics to determine how to proceed. It is demonstrated how clinicians' views sometimes differ from existing notions of explainability in ML.

Another interesting example exposed in [10], was the one explaining the asthma case. The model (which was a classic one) was predicting that given asthma, a patient had a lower risk of in-hospital death when admitted for pneumonia. In fact, the opposite is true, since patients with asthma are at higher risk for serious complications and sequelae, including death, from an infectious pulmonary disease like pneumonia. The asthma patients were, in fact, provided more timely care of a higher acuity than their counterparts without asthma, thereby incurring a survival advantage. The EML system has not found the correct explanation and that prevents us from detecting that there is a problem in the ML system.

For finding these critical situations, the project was developed using three different experiments and three different datasets. With these three experiments, we will be testing the idea that if the EML fails to capture the way the predictor works, it can lead to very dangerous situations.

The remainder of this document is structured as follows. Section 2 of this work explains in more detail the materials and methods used for achieving the objectives. Section 3 describes the results obtained after the training and the critical situations encountered on them. The next section, Section 4, discusses those results. The final section, number 5, the conclusions of the project are presented, and some future lines are mentioned.

2 MATERIAL AND METHODS

In this section, the datasets used in this project are explained. Then, we introduce the experiments designed for testing the different algorithms. The classic black box ML algorithm of prediction that was considered suitable for this project was the Logistic regression model [11] and the Random Forest Classifier [11], while the EML algorithms chosen for the work were the LIME Algorithm [13],[14] and the SHAP Algorithm [13][15]. The programming language used for working on the whole project and for modelling the algorithms cited was Python. Code for this thesis can be found in [16].

2.1 Materials

2.1.1 Synthetic Data Prediction Datasets

The first datasets considered were two synthetic dataset that consists of several columns of which only two are relevant. Although synthetic, the datasets were intended to have a medical interpretation and therefore, the relevant features represent the age and the BMI of the patient.

The main feature of these datasets is that they can be easily visualized and interpreted. Hence, it was possible to visualize instances that could be problematic for the EML in the sense of providing unsensed explanations. With that in mind, datasets with non-linear boundaries between classes were built. The reason for using this is that EML algorithms (see Section 2.2.3) are based on linear explanations. Hence, these synthetic datasets are designed so that the EML algorithms should fail to capture the behaviour of the underlying classifier. More specifically, it is expected that around the green point shown in Figure 3, the EML algorithm would make good explanations, since near this point it is possible to approximate the boundary with a straight line. However, if explanations for the black point in Figure 3 are requested, worse performance is expected since there is no way to approximate such boundary with a straight line.

Both datasets were created with 10000 instances and 7 features. 5 of those features are useless for the prediction of the label. For both synthetic datasets, the range of the age feature was from 45 years to 90, while the values of BMI oscillated between

17 and 34. The outcome was a binary value (either 0 or 1) for interpreting the positive and negatives of the hypothetical diagnoses.

The idea for the first synthetic dataset is that only the older people which have a high value of BMI are the ones who have more risk to get the imaginary disease. Figure 2 represents the first synthetic dataset that was considered.

On the other hand, the idea for the second dataset is that the older you get, the greater the risk of having the imaginary disease. It is presented the dependency between this disease and the BMI which is U-shaped (See Figure 3). This makes sense because having either a very high or very low BMI leads to having a high risk of having this imaginary disease. When you move to the right you turn blue which means that the imaginary disease is present.

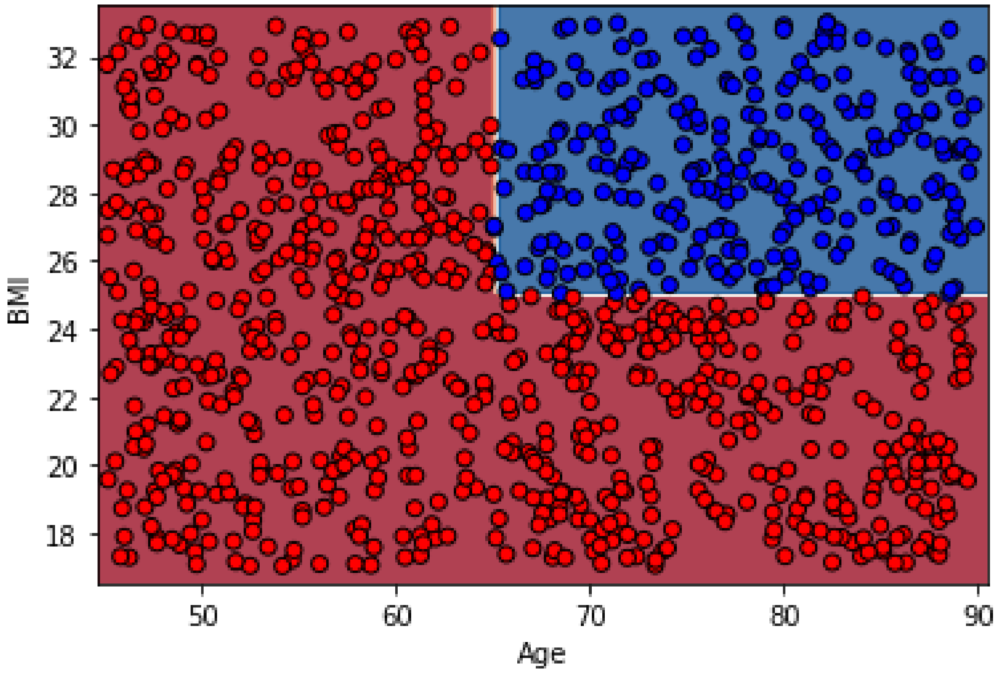

Figure 2. Plot of first synthetic dataset for visualizing dependencies.

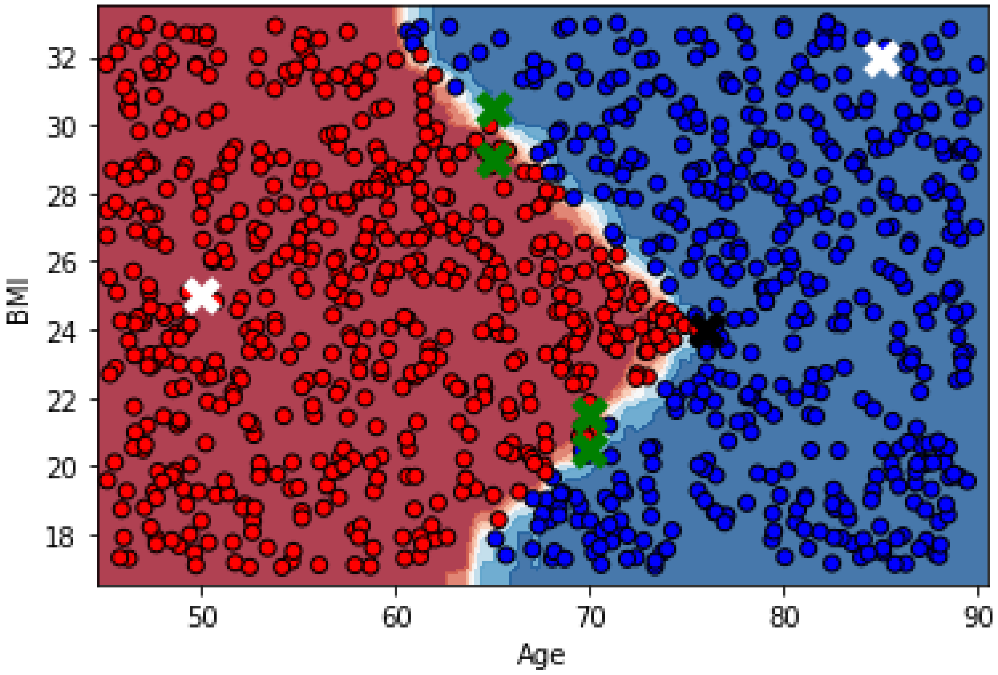

Figure 3. Plot of second synthetic dataset for visualizing dependencies.

2.1.2 Stroke Prediction Dataset

The first dataset considered is the Stroke prediction dataset. According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths [17]. This dataset is used to predict whether a patient is likely to get a stroke based on the input parameters. There are certain factors which influence the chances of getting a stroke. These factors are represented as features or attributes in the dataset.

Table 1 represents the 10 attributes (plus the label and unique identifier) that constitute this dataset. In the smoking_status feature, the possible value of “unknown” means that the information is unavailable for this patient. Finally, the stroke label divides its value in 249 positives (4.8%) and 4861 negatives (95.2%).

Table 1. Attributes that constitute the Stroke Prediction Dataset

Feature name	Description	Possible value
ID	Unique identifier that was removed before training	Integer number
stroke	Label, which indicates whether a patient has a stroke or not	Binary number (0 for negative and 1 for positive)
gender	Patient's gender	*"Male", "Female"* and *"Other".*
age	Patient's age	Integer number
hypertension	Describes whether the patient has hypertension or not	Binary number (0 if the patient does not suffer hypertension or 1 if it does)
heart_disease	Describes whether the patient has a heart disease or not	Binary number (0 if the patient does not suffer from any heart disease or 1 if it does)
ever_married	Answers the question if the patient has ever been married	Binary number (0 if the answer is no or 1 if the answer is yes)
work_type	Describes the type of work that the patient has	*"children", "govt_job", "never_worked", "private"* or *"self-employed"*
residence_type	Describes the type of residence that the patient has	*"rural"* or *"urban"*
avg_glucose_level	Represents the average glucose level in blood	Decimal number
BMI	Represents the body mass index	Decimal number
smoking_status	Describes the smoking status of the patient	*"formerly smoked"*, "*never smoked", "smokes"* or "*unknown"*

2.1.3 Heart Failure Prediction Dataset

The heart failure prediction dataset was also used in this work [18]. Every year, nearly 17 million individuals die from cardiovascular disorders, which primarily manifest as myocardial infarctions and heart failures. When the heart cannot pump enough blood to meet the body's needs, it is called heart failure [19].

Table 2. Attributes constituting the Heart Failure Prediction Dataset

Feature name	Description	Possible value
age	Patient's age	Integer number
creatinine phosphokinase (CPK)	Explains the level of the CPK enzyme in the blood (mcg/L)	Integer number
ejection fraction	Explains the percentage of blood leaving the heart at each contraction	Integer number
serum sodium	Explains the level of serum sodium in the blood (mEq/L)	Integer number
time	Corresponds to the follow-up period in days	Integer number
platelets	Represents the number of platelets in the blood (kiloplatelets/mL)	Decimal number
serum creatinine	Represent the level of serum creatinine in the blood (mg/dL)	Decimal number
anaemia	Corresponds to the decrease of red blood cells	Binary number (0 if negative or 1 if positive)
high blood pressure	Indicates whether the patient has hypertension or not	Binary number (0 if negative or 1 if positive)
diabetes	Indicates if the patient has diabetes or not	Binary number (0 if negative or 1 if positive)
smoking	Represents whether the patient smokes or not	Binary number (0 if negative or 1 if positive)
sex	Represents if the patient is a woman or a man	Binary number (0 if it is a woman or 1 if it is a man)
death_event	The target attribute, describing whether the patient deceased during the follow-up period	Binary number (0 if negative or 1 if positive)

This dataset contains 12 features (plus de label) that can be used to predict mortality by heart failure. Most cardiovascular diseases may be prevented by implementing population-wide programs to address some existing behavioural risk factors such as the use of cigarettes, unhealthy diet and obesity, physical inactivity, and damaging alcohol consumption. People with cardiovascular disease or who are at high cardiovascular risk, due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia, or previously established disease, require early detection and management, which can be greatly aided by a machine learning model. The attribute information is therefore fundamental for the detection part. The target attribute (see Table 2) divides its value in 96 positives (32.1%) and 203 negatives (67.9%).

2.1.4 Diabetes Prediction Dataset

The third dataset considered is the Diabetes Prediction dataset [20] with the objective of predicting whether a patient has diabetes or not based on diagnostic measurements. 382 million people worldwide are affected by diabetes mellitus, and the number of people with type 2 diabetes is growing in every country. If diabetes is untreated, it can cause many complications.

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The population selected was the Pima Indian population (subgroup of Native Americans) near Phoenix (Arizona), which has been under continuous study since 1965 due to its high incidence rate of diabetes. According to the World Health Organization Criteria, if the 2-hour post-load glucose was at least 200 mg/dl at any survey exam or if the Indian Health Service Hospital serving the community found a glucose concentration of at least 200 mg/dl during routine medical care, diabetes was diagnosed.

Several constraints were placed on the selection of these instances from a larger database. All patients here are females with at least 21 years old of Pima Indian heritage. The dataset consisted of 768 instances and 8 attributes, plus the label.

A particularly interesting attribute used in the study was the Diabetes Pedigree Function (see Table 3). It provided some data on diabetes mellitus history in relatives and the genetic relationship of those relatives to the patient. This measure of genetic influence provided an idea of the hereditary risk a patient might have with the onset of diabetes mellitus. Based on observations in the proceeding section, it is unclear how well this function predicts the onset of diabetes [21]. Examining the distribution of class values, it was concluded with 500 negative instances (65.1%) and 258 positive instances (34.9%).

Table 3. Attributes constituting the Diabetes Prediction Dataset

Feature name	Description	Possible value
pregnancies	Specifies the number of times the patient had been pregnant	Integer number
glucose	Indicates the plasma glucose concentration after 2 hours in an oral glucose tolerance test	Integer number
blood pressure	Describes the diastolic blood pressure in mm Hg	Integer number
skin thickness	Constitutes the triceps skin fold thickness in mm	Integer number
insulin	Demonstrates the 2-hour serum insulin in mu U/ml	Integer number
age	Patient’s age in years	Integer number
BMI	Represent the body mass index measured as weight in kg/ (height in m) ^2	Decimal number
diabetes pedrigree function	Represents the value of the diabetes pedigree function	Decimal number
outcome	Represents the label, describing if the patient has diabetes or not	Binary number (0 if tested negative for diabetes or 1 if tested positive for diabetes)

2.2 Methods

2.2.1 Preprocessing

Before starting with the classification algorithms, several modifications and filtering through several transformers had to be performed before fitting the logistic regression estimator. For the second and third dataset a standard scaler estimator and a transform method were used for the part of preprocessing the data. This method standardizes the features by removing the mean and scaling to unit variance. The standard score of a sample x can be calculated as: $z = \frac{(x-u)}{s}$, where u is the mean of the training samples and s is the standard deviation of the training sample. Later, with the transformation method which prevents repeated computations it is fitting to data and then transforming it, returning the final transformed array [22]. The estimator stores the mean and the standard deviation that it computed for the training set.

The One-Hot Encoding is also needed for feeding categorical data to many scikit-learn estimators, our linear regression model. Therefore, for the first and the second datasets, this encoder method which encodes categorical features as a one-hot numeric array, was used. The features are encoded using a one-hot (aka 'one-of-K' or 'dummy') encoding scheme. This creates a binary column for each category and returns it as a dense array. After this step, the transformation method was used once again to fit the encoder's output to the categorical values and then transform those categorical values.

Finally, another last step that was necessary to perform over the first dataset, was the part of imputing BMI's NAs (Not Available values). For this purpose, a simple imputer class, which provides basic strategies for imputing missing values, was used. Missing values can be imputed with a provided constant value, or using the statistics (mean, median or most frequent) of each column in which the missing values are located. In this case, the NAs were replaced by the mean, which was set by default. Once again, the transformation method was used as in previous steps to fulfill this purpose.

2.2.2 Classification algorithms

2.2.2.1 Logistic regression + l1 penalization

To facilitate the interpretation of the experiments, one of the classifiers that has been used is Lasso. The reason for choosing this classifier is because with this one it is possible to know which features in the datasets are used and which ones are ignored.

The Lasso model could be understood as a sort of linear regression in which only relevant variables contribute to the explanation and proceeds by removing non-relevant ones. In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function. Therefore, Lasso may be represented as logistic regression [11] + $\ell 1$ penalization regularized logistic regression, which solves the following optimization problem:

$$\min_{\boldsymbol{\omega},\boldsymbol{c}} \|\boldsymbol{\omega}\|_1 + \boldsymbol{C}\sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} \log(\boldsymbol{exp}\left(-\boldsymbol{y}_{\boldsymbol{i}}(\mathbf{X}_{\boldsymbol{i}}^{\mathbf{T}}\boldsymbol{\omega} + \boldsymbol{c})\right) + \mathbf{1}) \tag{1}$$

With this penalty what it is pretending is to encourage the qualifier to discard some of the variables. In the context of this thesis, this will prove useful to test if the explainer is using variables that Lasso does not actually use.

Two of the most used logistic regression model implementations in Python are: scikit-learn and statsmodels. Although both are highly optimized, Scikit-learn is mainly oriented to prediction, being this the reason of our choice [22].

For initializing the logistic regression estimator used for constituting the base classification model, some parameters were declared. The saga solver representing the algorithm to use in the optimization problem was one of the most relevant parameters to be indicated. This algorithm was selected due to l1 penalty, the multinomial loss, and the large size of the dataset, since it works faster. Whenever randomization is part of a Scikit-learn algorithm, a specific parameter may be provided to control the random number generator used. The passed value will influence the reproducibility, producing the same results across different calls. The maximum number of iterations taken for the solver to converge was 1000. Finally, a parameter which presents the corresponding weights associated with classes is needed. The “balanced” mode was the one selected since it uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data as $\frac{n\_samples}{n\_classes \times np.bincount(y)}$.

After fitting the Lasso, the relevant columns for each experiment were obtained (See Table 4). This is a fundamental part for comparing later with the EML model and looking for examples in which a non-relevant column appears as part of the explanation.

Table 4. Relevant columns extracted for each dataset.

Dataset	Relevant columns
Stroke Dataset	Age, BMI, Average Glucose Level
Heart Failure Dataset	Ejection Fraction, Serum Creatinine
Diabetes Dataset	Pregnancies, BMI, Glucose

2.2.2.2 Random Forest Classifier

For the synthetic dataset experiments, and due to their non-linear boundaries, the random forest classifier was selected [12]. In random forests, each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. A random forest is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset. Then, predictions are averaged to improve the predictive accuracy and control over-fitting.

The random forest classifier works in four steps:

1. Select random samples from a given dataset.
2. Create a decision tree for each sample and use it to generate a prediction result.
3. Perform a vote for each expected outcome.
4. As the final prediction, select the prediction result with the most votes.

2.2.3 EML algorithms

2.2.3.1 LIME algorithm

As it was introduced before, over the ML model we implement the EML model to obtain explanations for the predictions made by the ML model. The first algorithm that was used for this goal was the LIME (*Local Interpretable Model-Agnostic Explanations*) algorithm, whose aim is to explain the predictions of a classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. It identifies an interpretable model over the interpretable representation that is locally faithful to the classifier [24].

LIME can explain the prediction of any classifier, even without knowing its internals, by approximating it locally with an interpretable model. If the input domain of the surrogate function is human-interpretable, then LIME can even explain decisions of a model which uses non-interpretable features. LIME is one of the most effective and popular models today, although its high computational complexity.

LIME generates explanations proceeding as follows:

1. It selects the instance or sample of interest to calculate its local explanation.
2. For each selected *x* observation, sample *n* times to create a new dataset with data from the neighborhood of *x*.
3. Calculate the distance of all permutations to the original observation and convert that distance into a score of similarity with which to weight the importance of each instance of the sample.
4. Select the *m* variables that best describe the result of the complex model with the data to be included in the interpretable model.
5. Run an interpretable model with the new permuted dataset, explaining the result of the complex ML model with the *m* variables selected at the previous point and the observations weighted according to their similarity to the original observation *x*.
6. Extract the weight of the variables from the simple model and use it as a local interpretation of the behavior of the complex ML model for cases similar to *x* [25].

In formal terms, an explanation is defined as a model $g \in \mathrm{G}$, where G is a class of potentially interpretable models, such as linear models. The domain of g is $\{0,1\}^{d'}$, implying that g operates over absence or presence of the interpretable components. Because not every $g \in \mathrm{G}$ may be simple enough to be interpretable, $\Omega(g)$ is defined as a measure of complexity (as opposed to interpretability) of the explanation $g \in \mathrm{G}$. In this case, for linear models, $\Omega(g)$ may be the number of non-zero weights. The model being explained might be denoted f: $\mathbb{R}^d \rightarrow \mathbb{R}$. In classification, $f(x)$ is the probability that *x* belongs to a specific class. Additionally, it is defined $\pi_x(z)$ as a proximity measure between an instance *z* to *x*, to explain locality around *x* [24].

Finally, $\mathcal{L}(f, g, \pi_x)$ is defined as the fidelity, a measure of how unfaithful g is in approximating f in the locality explained by π_x, measuring how close the explanation is to the prediction of the original model f. To ensure both interpretability and local fidelity, it is necessary to minimize $\mathcal{L}(f, g, \pi_x)$, while having $\Omega(g)$ be low enough to be interpretable by humans. The explanation generated by LIME is achieved by the following:

$$\boldsymbol{\xi}(\boldsymbol{X}) = \underset{\boldsymbol{g} \in \mathbf{G}}{\mathbf{argmin}}\, \boldsymbol{\mathcal{L}}(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{\pi_x}) + \boldsymbol{\Omega}(\boldsymbol{g}) \tag{2}$$

This formulation can be used with different explanation families G, fidelity functions $\mathcal{L}$, and complexity measures Ω [26].

The function chosen for implementing LIME, was the Lime Tabular explainer [24]. This function explains classifiers (ending up with a prediction) that use tabular data. In our experiments, all non-categorical features were discretized into quartiles. For categorical features, LIME perturbs them by sampling according to the training distribution and making a binary feature that is 1 when the value is the same as the instance being explained.

Other important section is in relation to the kernel which is used to determine the way the sample weights are calculated and limits the locality in which perturbation can happen. It consists of a similarity kernel which takes euclidean distances and kernel width as input, which is set to $\sqrt{n^{\underline{o}}\ of\ columns} \times 0.75$, and outputs weights in (0,1). This is kept as a hyperparameter in the algorithm, and it defaults to an exponential kernel which is closely related to the Gaussian kernel, with only the square of the norm left out:

$$k(x, y) = \exp\left(-\frac{\|x - y\|}{2\sigma^2}\right) \tag{3}$$

2.2.3.2 SHAP algorithm

The second algorithm chosen for providing another type of explanations was the SHAP Algorithm [15]. Shapley additive explanations (SHAP) have been proposed to calculate the contribution of each feature (input) to the output (prediction). SHAP is an extension of a Shapley value in coalitional game theory as a method for calculating the contribution of each feature in machine learning [27].

SHAP values are calculated by determining the difference between the predicted values with and without the addition of each feature (i.e., the effect of adding each attribute) for all combinations and taking the average. Namely, it is the average marginal contribution of a feature value across all possible coalitions/permutation. Therefore, it becomes possible to understand which features have a significant influence on the output and whether the influence is negative or positive by calculating the SHAP values [28].

The function used for creating the SHAP explainer was the Kernel SHAP (Linear LIME + Shapley Value) [13], which is the universal SHAP Explainer for any ML algorithm. It is an alternative, kernel-based estimation approach for Shapley values inspired by local surrogate models. This function estimates for an instance x the contributions of each feature value to the prediction. It uses a specially weighted linear regression as the local surrogate model by using your data, your predictions, and whatever function that predicts the predicted values, along with an appropriate weighting kernel, to compute the importance of each feature. The computed importance values are Shapley values from game theory and coefficients from a local linear regression.

In this project, the function performs a local regression by taking the prediction method and the data for which you want to compute the SHAP values. For this part, the Kernel Explainer requires a background dataset to generate the perturbed dataset required for training surrogate models. To determine the impact of a feature, that feature is set to "missing" and the change in the model output is observed. Since most models are not designed to handle arbitrary missing data at the test time, the "missing" part is simulated by replacing the feature with the values it takes on the background dataset. So, if the background dataset is a simple sample of zeros, then the feature being missing would be approximated by setting it to zero. For this project, the k-means function was used as the background dataset, to summarize the dataset.

The process consists of 5 steps [28]:

1. Sample coalitions $z'_k \in \{0,1\}^M,\ k \in \{1, \dots, K\}$ (1 = feature present in coalition, 0 = feature absent).
2. Get prediction for each z'_k by first converting z'_k to the original feature space and then applying model f: $f(h_x(z'_k))$.

3. Compute the weight for each z'_k with the SHAP kernel.
4. Fit weighted linear model.
5. Return Shapley values ϕ_k, the coefficients from the linear model.

This procedure is explained with the following example: in step 1, a vector of (1,0,1,0) would be created, representing that a coalition of the first and third features is existing. The K sampled coalitions become the dataset for the regression model. The target for the regression model is the prediction for a coalition. To get from coalitions of feature values to valid data instances, we need a function $h_x(z'_k)$ where h_x: $\{0,1\}^M, \mathbb{R}^p$. The function h_x maps 1's to the corresponding value from the instance x that it is pretended to explain. For tabular data, it maps 0's to the values of another instance that it is sampled from the data. This means that it is equated "feature value is absent" with "feature value is replaced by random feature value from data". h_x for tabular data treats x_c and x_s as independent and integrates over the marginal distribution:

$$f\big(h_x(z')\big) = E_{X_c}[f(x)]. \tag{4}$$

Sampling from the marginal distribution means ignoring the dependence structure between present and absent features. The base value (E_{X_c}) is just the average model prediction for the background dataset provided when initializing the explainer object. SHAP values are all relative to this base value, since the sum of SHAP contributions, plus this base value, equals to the model's output. Since we are explaining a logistic regression model the units of the SHAP values will be in the log-odds space [27]. To achieve Shapley compliant weighting, Lundberg et. al propose the SHAP kernel:

$$\pi_x(z') = \frac{(M-1)}{\binom{M}{|z'|}\,|z'|(M-\ |z'|)} \tag{5}$$

M represents the maximum coalition size and $|z'|$ the number of present features in instance z'. After this, with the corresponding data, target and weights the weighted linear regression model is created:

$$g(z') = \phi_0 + \sum_{\boldsymbol{j=1}}^{\boldsymbol{M}} \phi_j z'_j \tag{6}$$

The linear model g is trained by optimizing the following loss function $\boldsymbol{\mathcal{L}}$**,** where Z is the training data:

$$\boldsymbol{\mathcal{L}}(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{\pi_x}) = \sum_{z' \in Z} \left[f\left(h_x(z')\right) - g(z') \right]^2 \pi_x(z'). \tag{7}$$

Finally, the estimated coefficients of the model, the ϕ_j's are the Shapley values [13] are computed. $\boldsymbol{S}$ represents a feature subset in the set of total features $\boldsymbol{F}$**.** $f_{S \cup \{i\}}$ is a trained model with that feature present, and f_S is another model trained with the feature withheld, resulting in the current input $f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S\,(x_S)$, where x_S represents the values of the input features in the set S, the predictions from the two models are compared. The preceding differences are estimated for all potential subsets $S \subseteq F \setminus \{i\}$, due to the fact that the effect of withholding a feature is dependent on other characteristics in the model:

$$\boldsymbol{\phi_i} = \sum_{\boldsymbol{S \subseteq F \setminus \{i\}},} \frac{|\boldsymbol{S}|!\,(|\boldsymbol{F}| - |\boldsymbol{S}| - \boldsymbol{1})!}{|\boldsymbol{F}|!} \left[\boldsymbol{f}_{\boldsymbol{S} \cup \{i\}}\left(\boldsymbol{x}_{\boldsymbol{S} \cup \{i\}}\right) - \boldsymbol{f}_{\boldsymbol{S}}\,(\boldsymbol{x_S}) \right] \tag{8}$$

2.2.4 Experiments

2.2.4.1 Visualizing and understanding explainers

For both experiments in Section 2.1.1, the ML model used for obtaining the predictions was the random forest classifier. The EML model selected for the first synthetic dataset experiment was the LIME algorithm. After this, the same procedure was followed but selecting the SHAP algorithm for simulating the EML model for obtaining the explanations. As these experiments were based on artificial datasets, first all the instances were visualized in the corresponding plot (See Figure 2). Then, a few instances that could be problematic for the EML model were selected. For the second synthetic dataset, the same procedure was followed (See Figure 3).

2.2.4.2 Medical datasets

- For the medical dataset experiments it was the logistic regression with ℓ1 regularization the ML model chosen as the base classifier to create the corresponding predictions. First, the LIME algorithm was selected as the EML model for obtaining the explanations. After this, the SHAP algorithm was additionally selected as another EML model for retrieving the corresponding explanations. This same procedure was followed for the three medical datasets, consisting of the Stroke Prediction Dataset, Heart Failure Prediction Dataset and the Diabetes Prediction Dataset. For all the experiments, the following steps were followed: Quantifying explanations fidelity. The aim is to measure the percentage of correct and incorrect numbers of incorrect and correct explanations in order to see the accuracy for the models.
- Locating correct predictions with non-plausible explanations. These situations are considered interesting from a medical point of view since they would undermine the trust of the physician in the model.
- Locating incorrect predictions with plausible explanations. In this last part, these situations were also considered interesting since it is an explanation which a doctor could consider reasonable and making sense. This will cause the doctor to trust the classifier even though it is providing an incorrect prediction.

2.2.5 Medical criteria

For this thesis, it has been estimated that plausible and non-plausible explanations are considered with two different criteria. First, we consider an explanation correct if it is exclusively based on the features shown in Table 4 (it does not use features that are known to be irrelevant for the Lasso classifier). The second criteria is defined for examining the values corresponding to the features, both on Table 4 or not, describing if they are reasonable from a clinical point of view. We now explain the clinical criteria taken into account for considering an explanation as reasonable or not.

Regarding the relevant columns in Stroke Prediction Dataset, BMI values below 18.5 are considered low weight, between 18.5 and 24.9 it is considered a normal weight, between 25 and 29.9 it is considered overweight and values higher than 30 is considered as obesity [29]. The critical values for glucose levels in blood are the ones which are higher than 140 mg/dL. A range between 140 and 199 mg/dL is considered prediabetes and values higher than 200 are considered diabetes [30]. Regarding the age, it is seen that the older a person gets, the more risk it has of suffering from stroke [31].

Regarding the Heart Failure Prediction Dataset, a range of ejection fraction between 55% and 70% is considered normal, from 40% to 54% is considered slightly below normal, from 35% to 39% is considered moderately below normal and finally for values below 35% it is considered severely below normal [32]. It has also been demonstrated how women tend to develop heart failure later in life compared with men. On the other hand, in patients with heart failure, kidney failure has been a significant prognostic factor. To assess this renal function, the serum creatinine level is commonly used [33]. The serum creatinine levels that were considered normal were <1.3 mg/dL in men and <1.1 mg/dL in women [34]. Additionally, it has been shown how other factors such as smoking, physical activity or diet are some other important factors contributing [35].

Finally, for the Diabetes Prediction Dataset, the criteria used to select the reasonable glucose levels in blood and the BMI was the same as previously explained. Moreover, the number of times that the patient has been pregnant has been considered relevant in this case due to recent studies which found that the percentage of pregnant women with gestational diabetes (type of diabetes which develops during pregnancy) has increased 56% and the percentage of women with type 1 or type 2 diabetes before pregnancy increased 37% in the past few years [36]. Diabetes in pregnancy varies by race and ethnicity. Asian and Hispanic women tend to have higher rates of gestational diabetes while black and Hispanic women have higher rates of type 1 or type 2 diabetes during pregnancy.

3 RESULTS

In this section, results for each of the experiments carried out in this thesis will be presented and the corresponding percentages of the accuracies for the algorithms.

3.1 Synthetic datasets results

Careful examination of the explanations obtained with LIME lead to the conclusion that, since they are based on intervals, LIME returns different explanations by organizing the feature space in a set of squared bins, as illustrated in Figure 4.

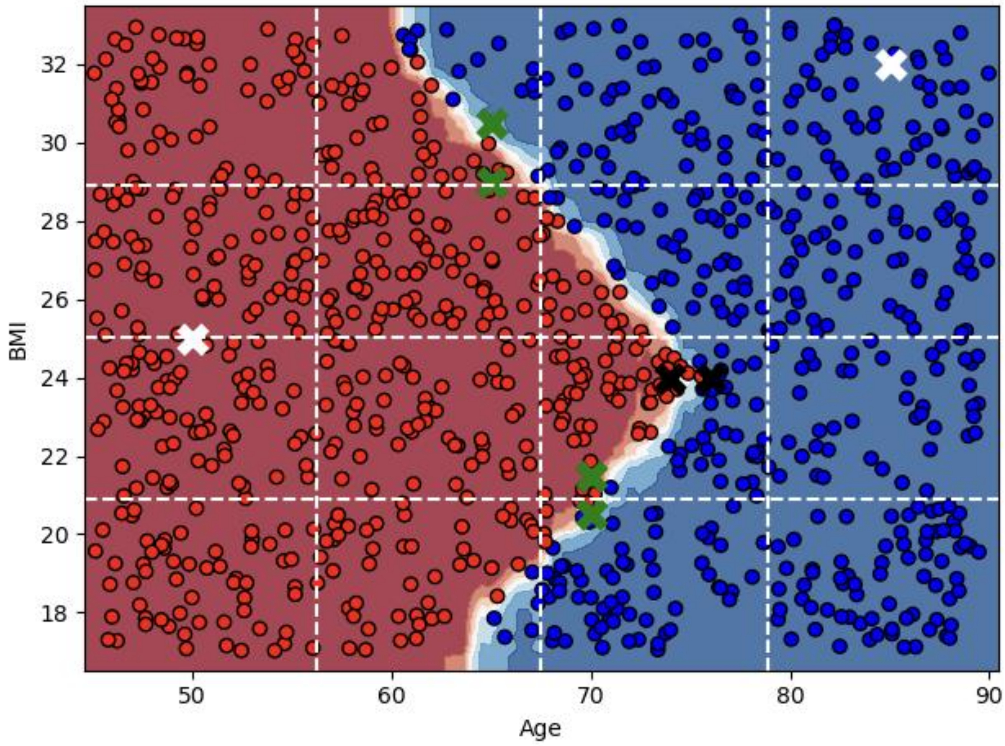

Figure 4. Intuition of grid of points for obtaining the explanations in second synthetic dataset

For this grid of squares, it has been observed a correct performance in the first dataset, since the boundaries in this case match shape of the squares, providing interpretable explanations (See Figure 5 and Figure 6). Figure 5 describes another test point which describes the situation of age of 65 and BMI of 26. The explanation describes how the age is between 56.27 and 67.75 and the BMI is between 25.02 and 28.89. The test point is lying on the red zone which indicates that the imaginary disease is not present.

On the same way, Figure 6 intends to represent the SHAP explanations for the test point describing the age of 65 and BMI of 25. Test point lies on the red zone representing that the patient had not the imaginary disease. Once again Feature 0 stands for the age and Feature 1 for the BMI.

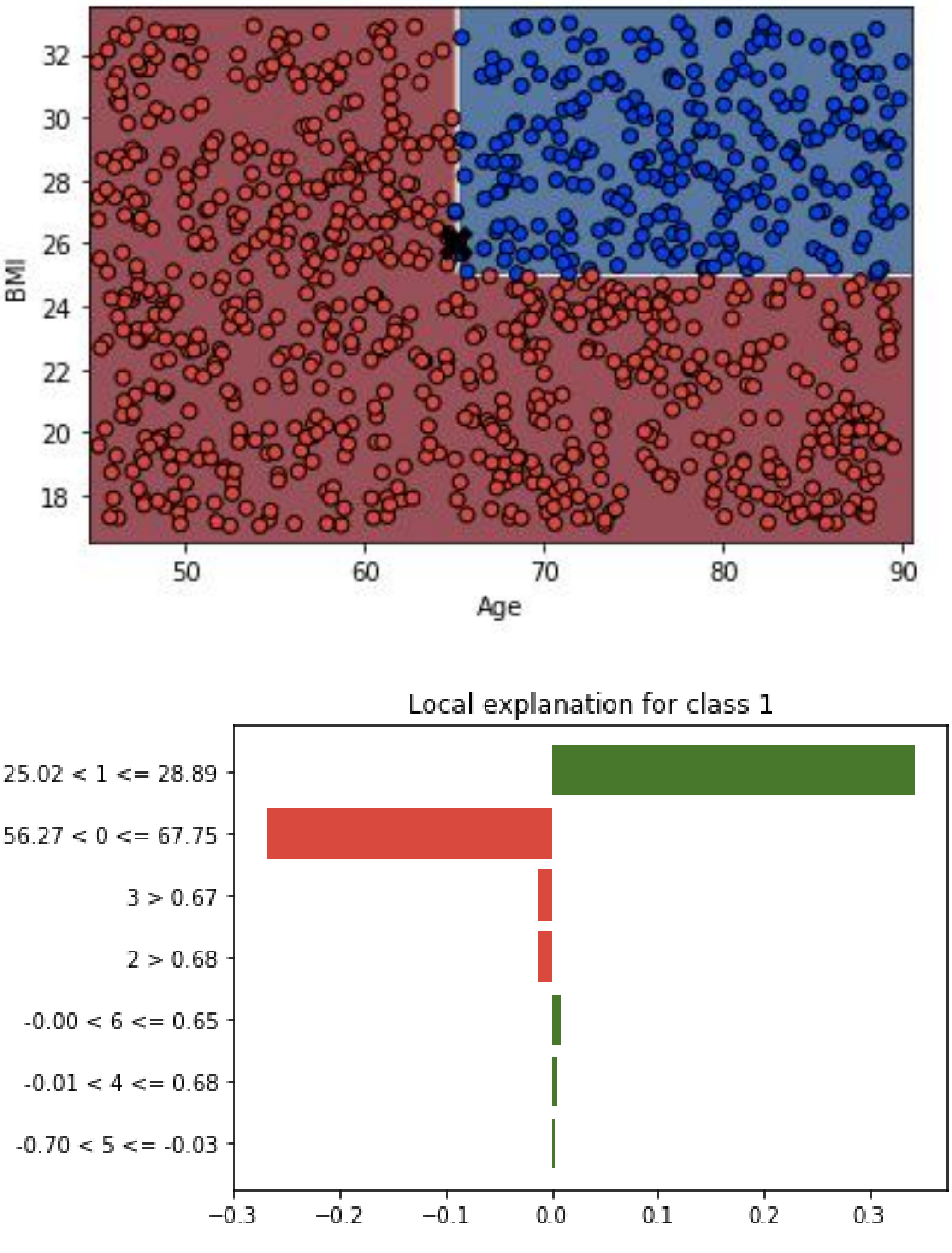

Figure 5. LIME explanation for test point in 1º synthetic dataset

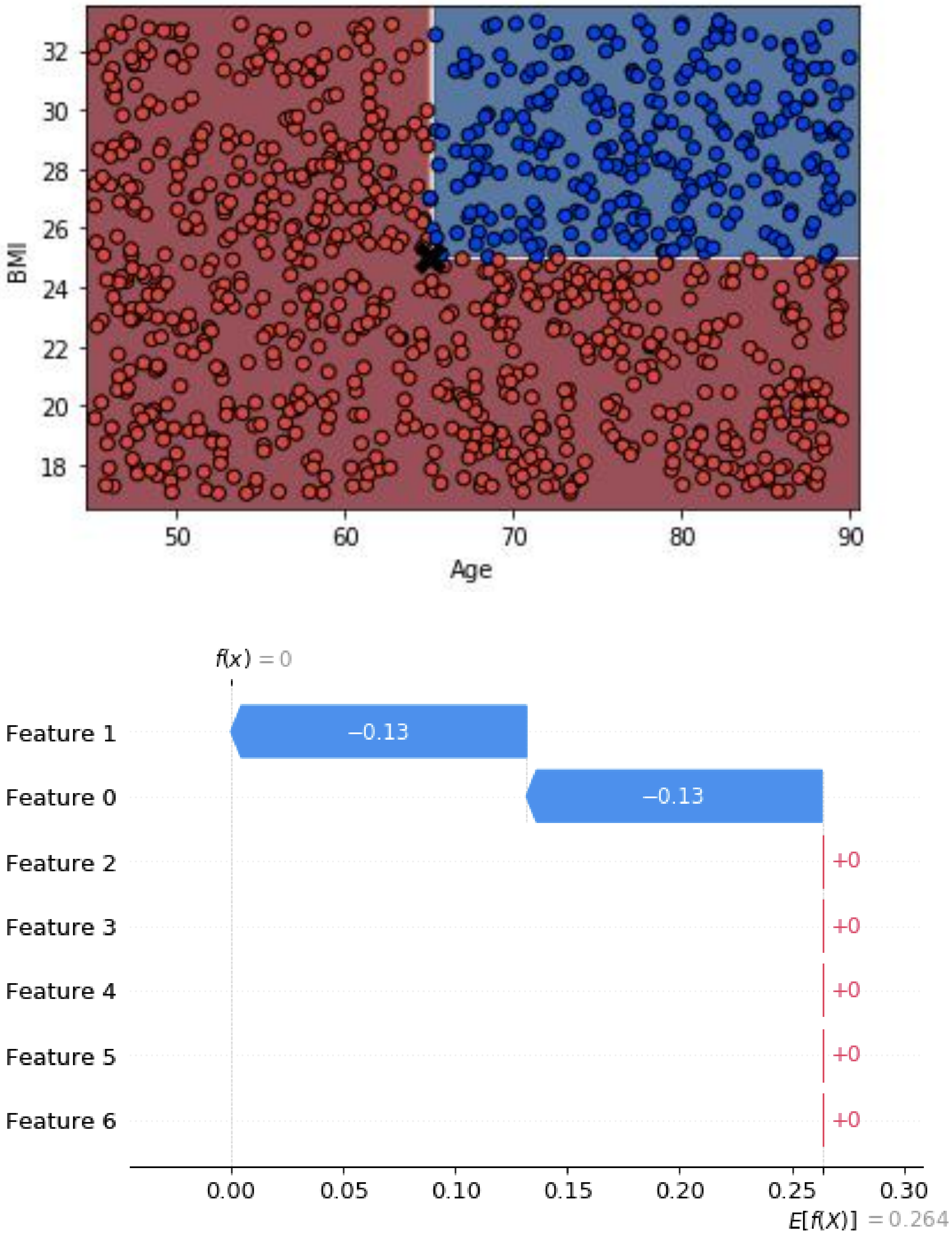

Figure 6. SHAP explanation for 1° synthetic dataset

However, in the second dataset, the fact that it is divided into squares does not help the explanation to be intuitive. For example, explaining the decisions of the classifier in the neighbourhood of the V-shaped region in terms of squared intervals is difficult to understand because it does not fit with the orientation of the variables (see Figure 7 and Figure 8).

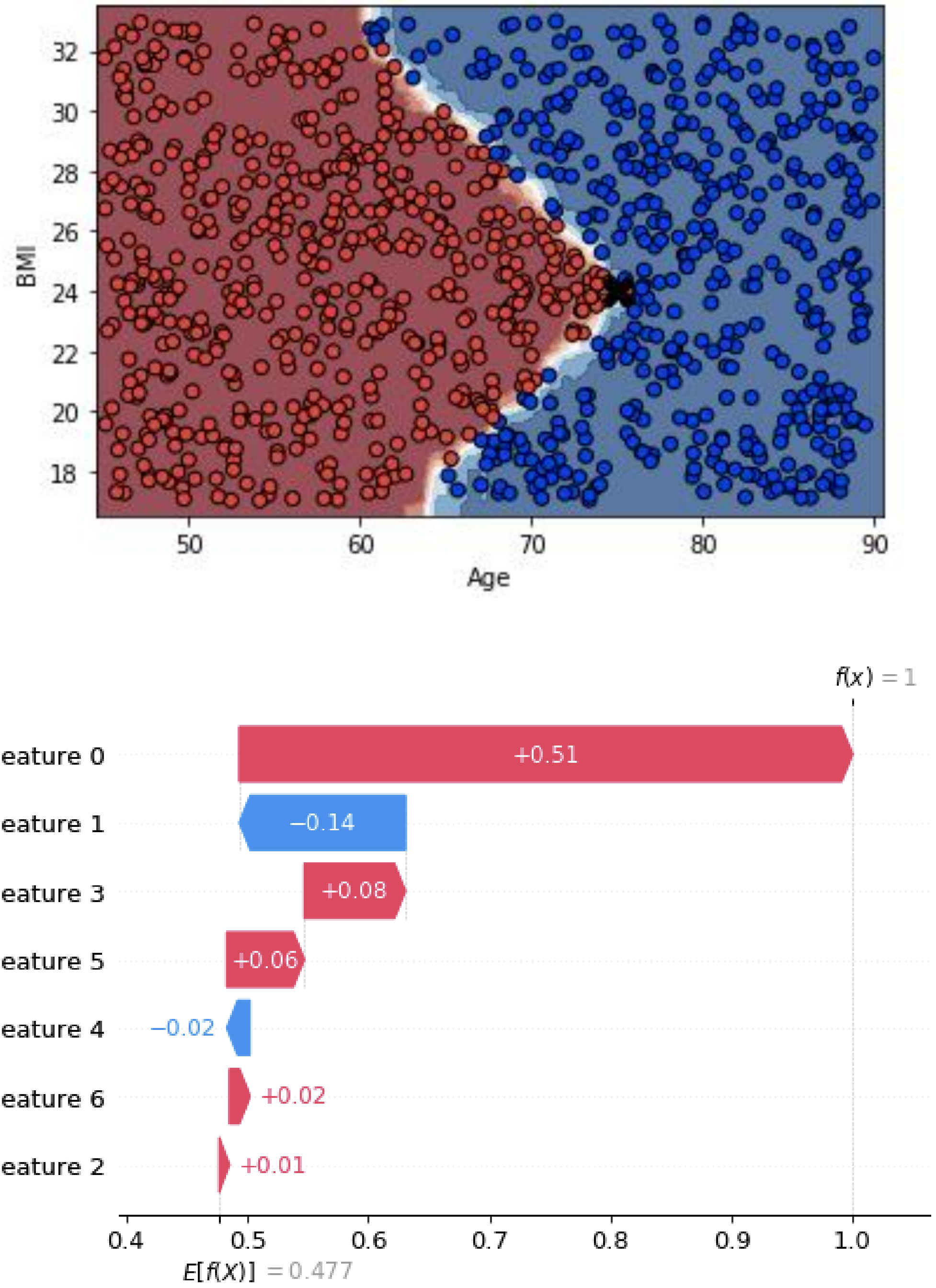

Figure 7. SHAP explanation for 2º synthetic dataset

Figure 7 represents the SHAP explanation in 2º synthetic dataset for a test point representing the age of 75 and a BMI rate of 24. Feature 0 in explanation represents the age while Feature 1 represents the BMI. Other features represent factors that should not be considered for the explanation. Finally, Figure 8 represents the LIME explanation in 2º synthetic dataset for a test point representing the age of 75 and a BMI rate of 24. Feature 0 in explanation represents the age while Feature 1 represents the BMI. Other features represent factors that should not be considered for the explanation in this case either.

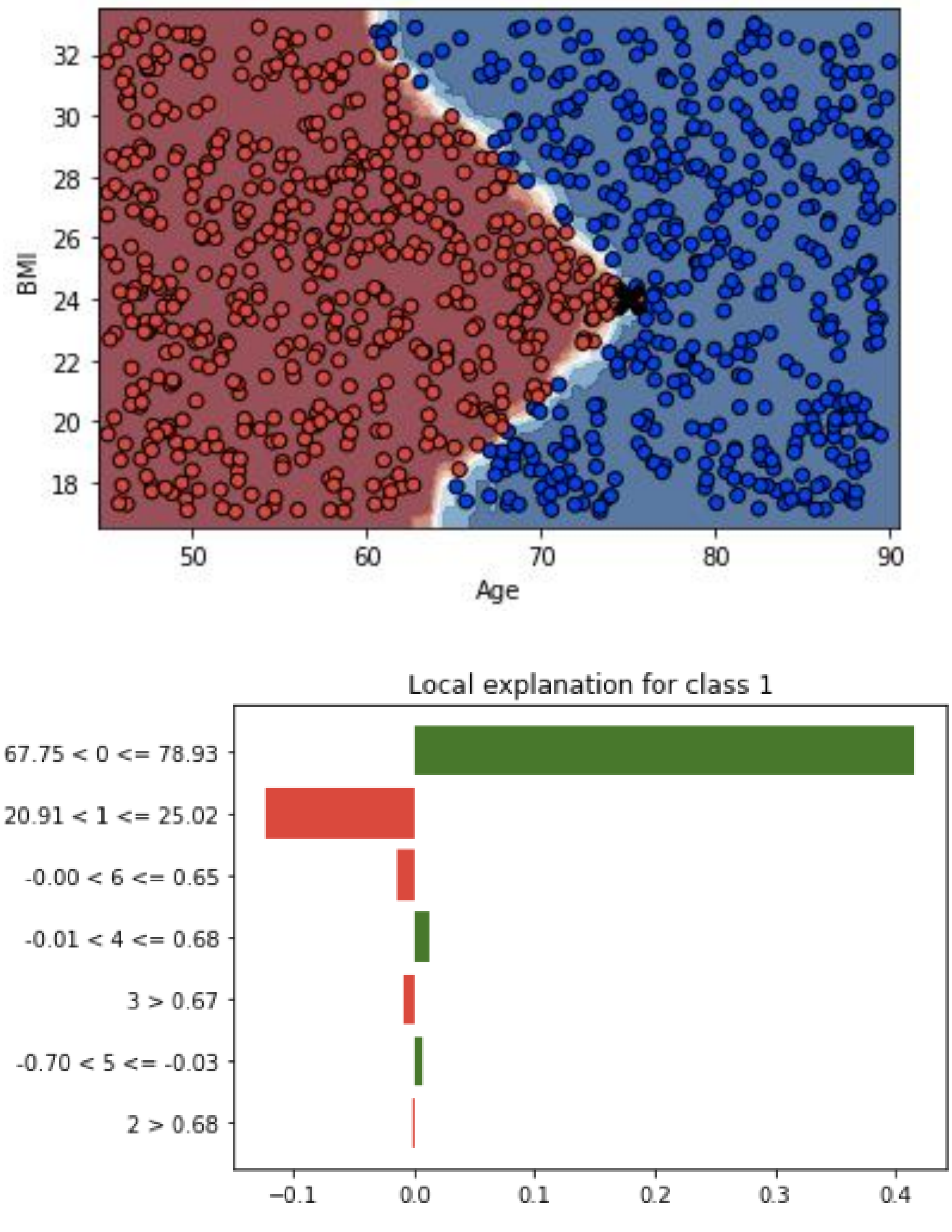

Figure 8. LIME explanation for 2º synthetic dataset

3.2 Quantifying explanations fidelity

The percentage of total correct explanations overall, correct explanations for correct predictions, and correct explanations for incorrect predictions are presented in Table 5 and Table 6 for the LIME and SHAP algorithms, respectively. Note that, the number of incorrect explanations is computed when the features appearing in the explanation do not correspond to the features in Table 4.

Table 5. Prediction/explanation percentages in LIME algorithm

Dataset	**Percentage of Correct Explanations**	**Percentage of Correct Explanations for Correct Predictions**	**Percentage of Correct Explanations for Incorrect Predictions**
Stroke Dataset	52.92%	53.8%	51.09%
Heart Failure Dataset	97.32%	96.90%	98.63%
Diabetes Dataset	70.44%	71.22%	68.18%

Table 6. Prediction/explanation percentages in SHAP algorithm

Dataset	**Percentage of Correct Explanations**	**Percentage of Correct Explanations for Correct Predictions**	**Percentage of Correct Explanations for Incorrect Predictions**
Stroke Dataset	65.04%	70.03%	54.94%
Heart Failure Dataset	98.99%	98.67%	100%
Diabetes Dataset	99.48%	99.47%	99.49%

3.3 Locating correct predictions with non-plausible explanations

3.3.1 LIME algorithm

3.3.1.1 Stroke Prediction Dataset

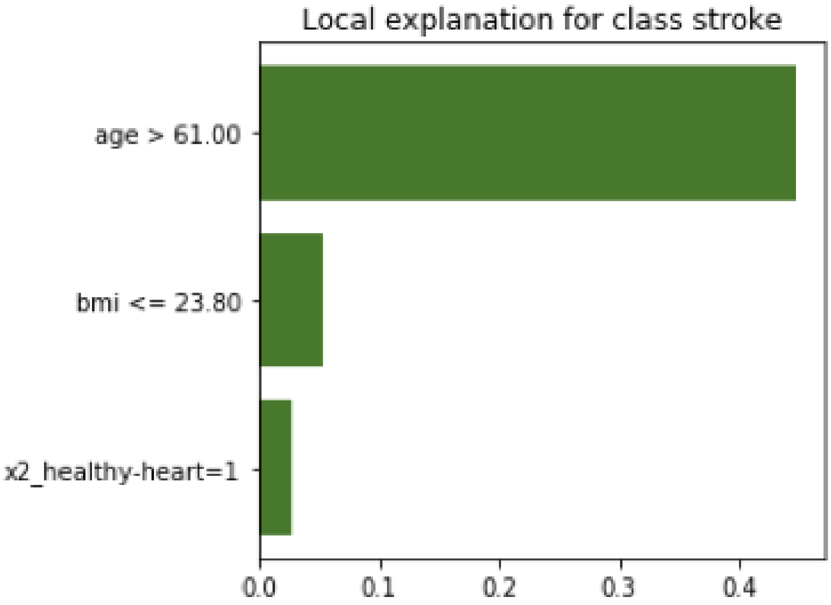

Figure 9. LIME explanation for stroke class in instance nº 7 in Stroke Prediction Dataset

In Figure 9, the explanation for a patient’s stroke condition is presented. The true class declared that the patient had a stroke. The probability of having a stroke deduced by the ML model was of 73.45%, thus positive in stroke and prediction is correct. In the explanation it is observed that features do not correspond to the features in Table 4, thus an incorrect explanation is given. Moreover, it declares that having a healthy heart contributes to have a stroke when it should be on the contrary, making the explanation unacceptable since it does not have any clinical sense. On the same way, Figure 10 represents the explanation for another patient’s stroke condition. The true class declared once again that the patient had a stroke. The probability of having a stroke deduced by the ML model this time was of 85.05%, thus positive in stroke and finally resulting on a correct prediction. This explanation, once more, exposes features which do not match with the ones in Table 4, and explains how being a non-smoker is contributing to suffer a stroke when it ought to be quite the opposite.

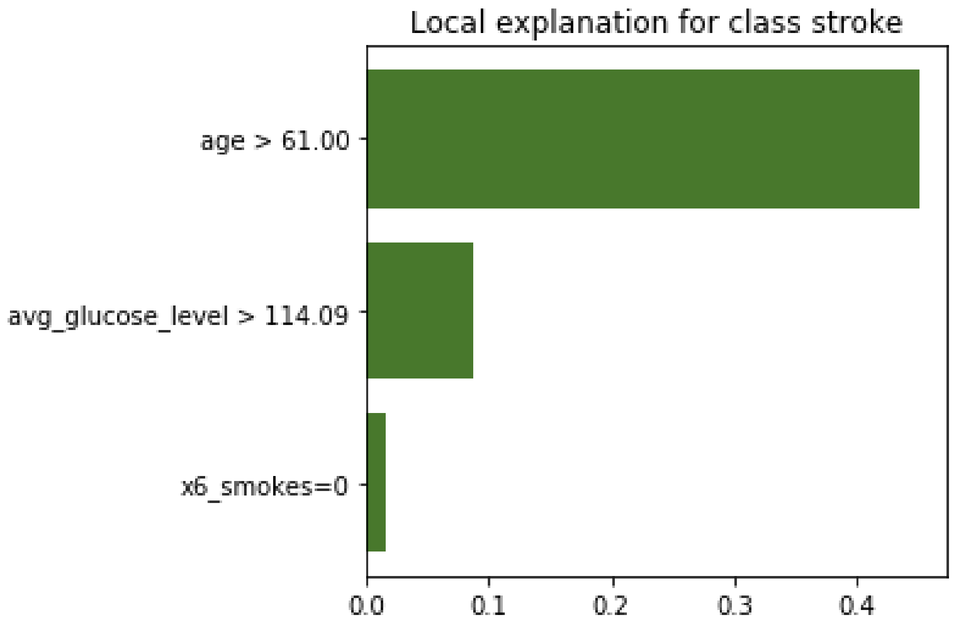

Figure 10. LIME explanation for class stroke in instance nº 139 in Stroke Prediction Dataset

3.3.1.2 Heart Failure Prediction Dataset

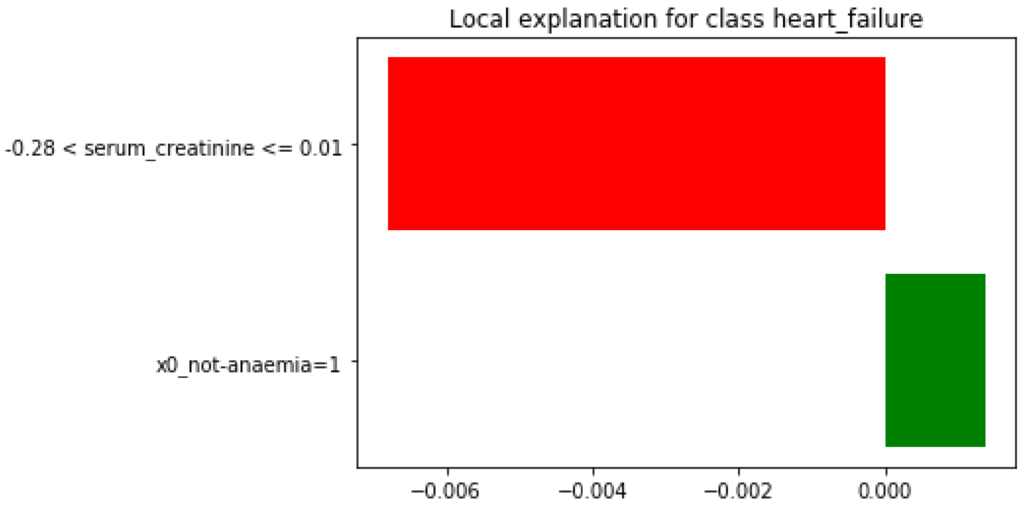

Figure 11. LIME explanation for class heart failure in patient nº 295 in Heart Failure Prediction Dataset

Figure 11 represents the explanation for the prediction of the patient nº 295. The serum creatinine corresponds to a range between 1.10 and 1.37 and does not suffer from anemia. The corresponding values for the serum creatinine were obtained after de-normalizing the values observed in Figure 11. The true class corresponding for this patient is healthy, while the probability deduced by the ML model for the heart failure prediction was of 49.33%, thus, negative for heart failure and therefore leading to a correct prediction. However, Figure 12 represents the case of patient nº 222, whose true class corresponds to healthy and the probability of the ML model for the heart failure prediction class is of 49.41% (negative in heart failure). This leads to a correct prediction, whose explanation corresponds to a range of creatinine between 0.89 and 1.10 (de-normalized values) and positive in suffering from anemia.

Both instances include the fact of having anemia which is not considered as a relevant feature and therefore it shall not appear in any of the explications. Furthermore, in Figure 11 it is exposed how not having anemia is contributing to have a heart failure which once again makes no clinical sense, while in Figure 12 it exposes the contrary explanation for a very similar healthy case.

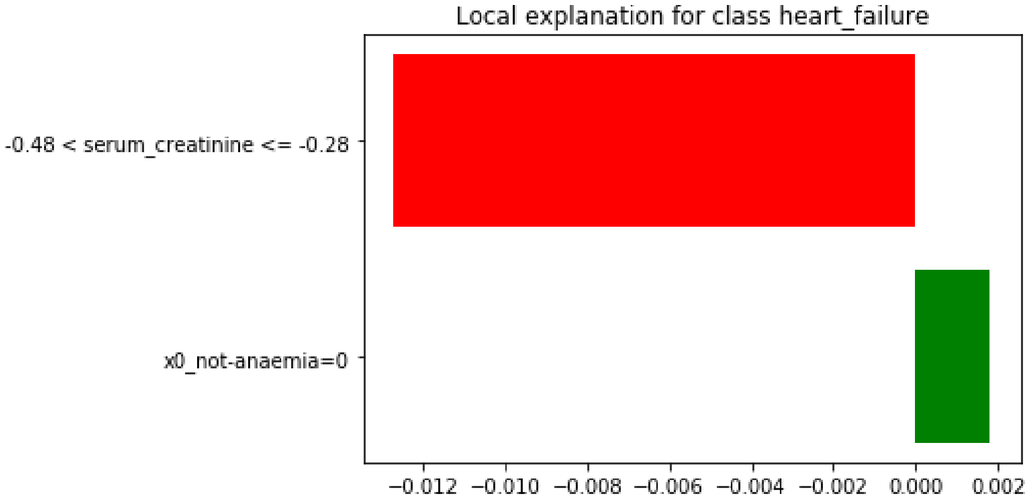

Figure 12. LIME explanation for the class heart failure in patient nº 222 in Heart Failure Prediction Dataset

3.3.1.3 Diabetes Prediction Dataset

The patient's nº 636 clinical case is presented in Figure 13. This patient's true class was of negative in diabetes and the probability deduced by the ML model for the diabetes prediction was of 38.32% and therefore negative for diabetes which leads to a correct prediction. The explanation presented for such prediction maintained how the glucose level in blood was in a range between 99 mg/dL and 117 mg/dL. The feature of age is indicating that patient is higher than 40 years old and the BMI is between a range of 27.26 and 32. Once again, these values were calculated de-normalizing the values observed in the Figure 13. It is shown how the feature of age appears when it should not as it has not been considered an important feature for contributing to the prediction in this dataset. Likewise, the BMI range in this explanation contributing to being healthy from diabetes makes no clinical sense once more.

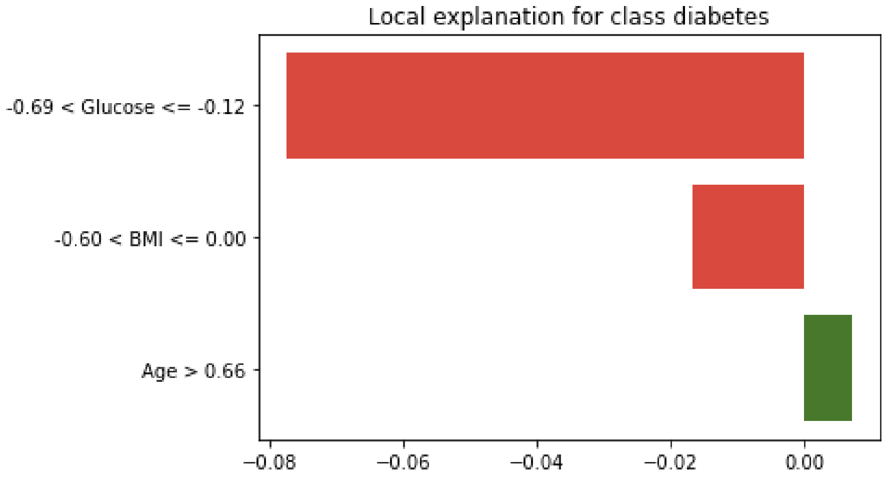

Figure 13. LIME explanation for the class diabetes in patient nº 636 in Diabetes Prediction Dataset

In the same way, the clinical case of patient nº 666 is presented in Figure 14. Its true class corresponds to positive in diabetes and the probability deduced by the ML model for the diabetes prediction was of 59.18%, resulting in a correct prediction. The explanation presented for this prediction consisted of the fact that the glucose level was higher than 140 mg/dL, that the range of BMI is between 32 and 36.5, and that the blood pressure was higher than 80 mmHg. This last feature is not considered relevant and therefore it should not be appearing in the explanation, and even less contributing to being healthy from diabetes, which once more makes no clinical sense. Diabetes or blood glucose levels do not immediately influence blood pressure, however, they do in the long term, because if glucose levels are high for a long time the kidneys are damaged, and with this the blood pressure rises.

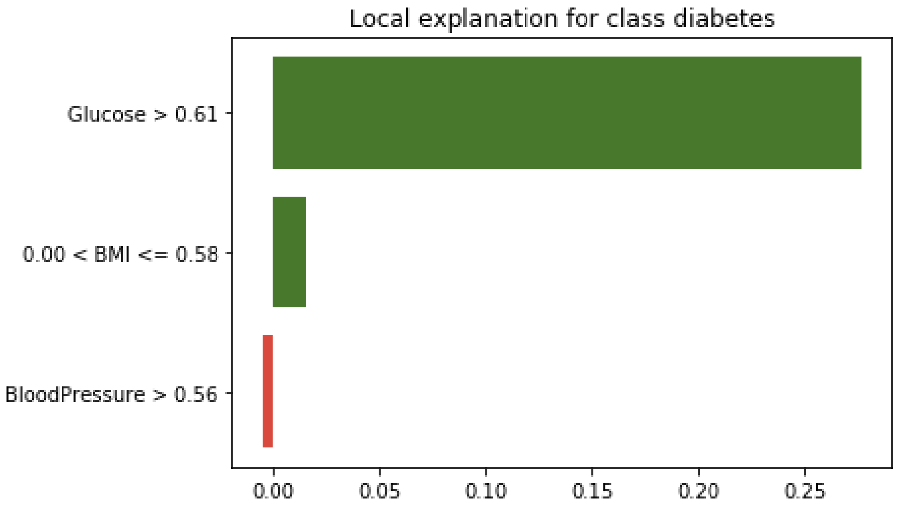

Figure 14. LIME explanation for the class diabetes in patient nº 666 in Diabetes Prediction Dataset

3.3.2 SHAP algorithm

3.3.2.1 Stroke Prediction Dataset

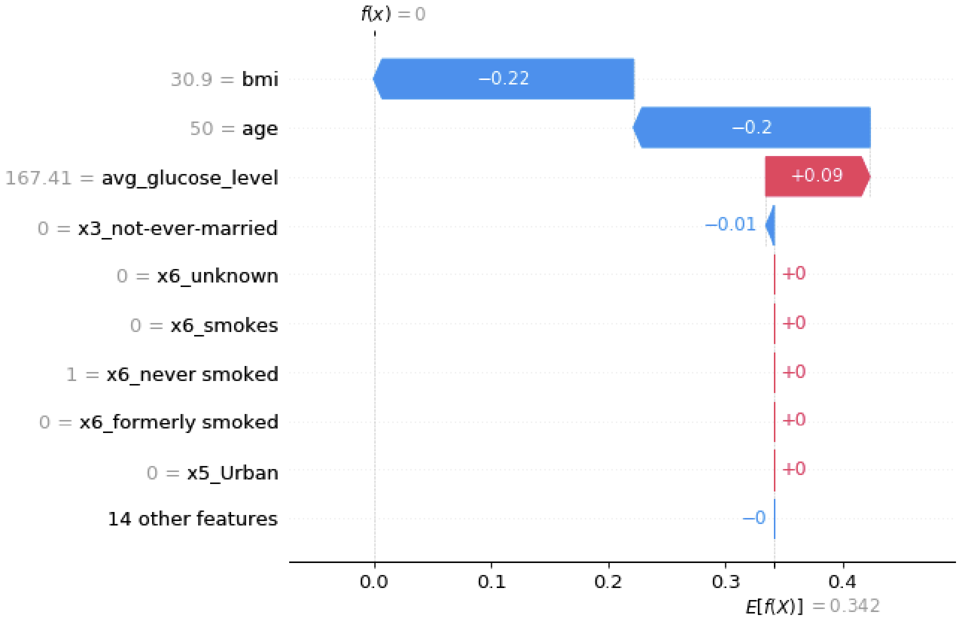

Figure 15. SHAP explanation for the class stroke in instance nº 15 in Stroke Prediction Dataset

In Figure 15 it is presented the case of a patient whose true class is positive in stroke and the deduced probability by the ML model is 51.53%. Therefore, the prediction is correct. This explanation shows more features contributing to the final prediction than it should, since ever being married is not considered a relevant column and should not appear on the explanation. Moreover, a BMI rate of 30.9 contributing to being healthy from stroke makes no clinical sense.

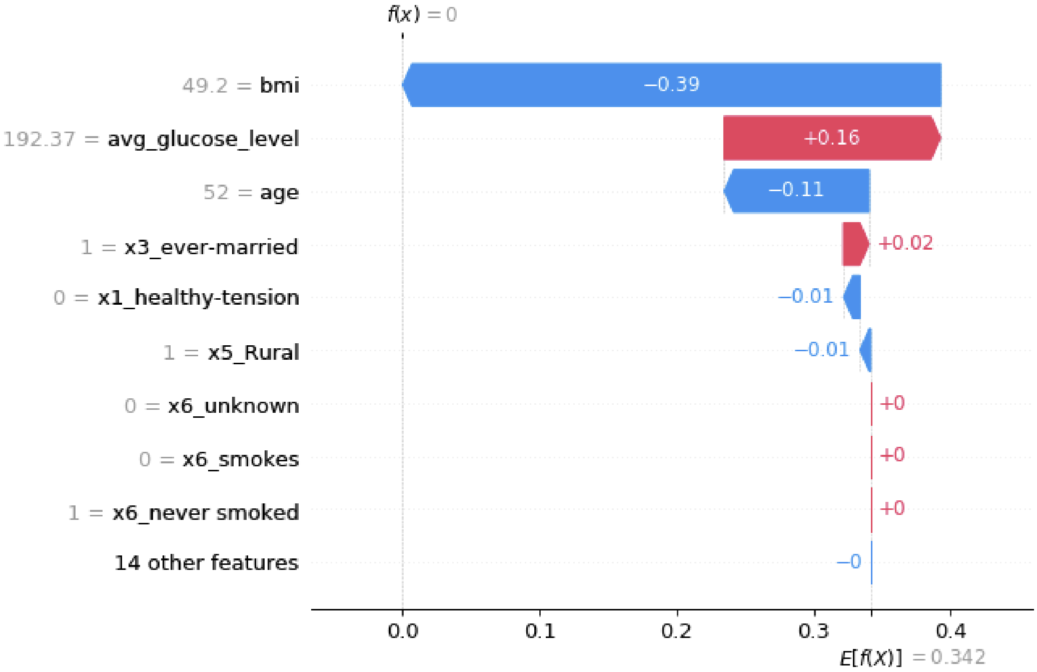

Figure 16. SHAP explanation for class stroke in patient nº 2284 in Stroke Prediction Dataset

Finally, in Figure 16 another example is presented with the case of a patient whose true class is negative in stroke, along with the deduced probability calculated by the ML model which consists of 47.36%. This represents a negative result in stroke and therefore the prediction is correct. Yet again, ever being married, hypertension and the type of residence are not referred to as relevant columns and should not appear on the explanation. Furthermore, a BMI rate of 49.2 and not having healthy tension contributing to being healthy from stroke makes no clinical sense. Both Figure 15 and Figure 16 represent the same condition of ever being married but contributing inversely on the final prediction, which makes the explanation unstable.

3.3.2.2 Heart Failure Prediction Dataset

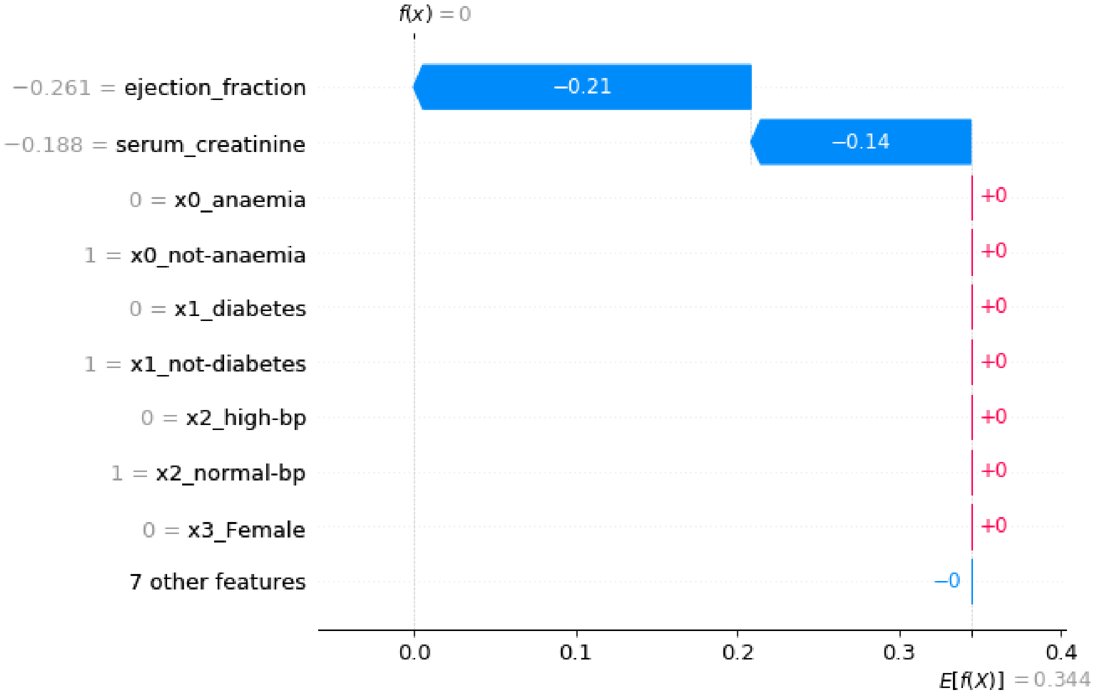

Figure 17. SHAP explanation for class heart failure in patient nº 111 in Heart Failure Prediction Dataset

In Figure 17, it is presented the clinical case for heart failure prediction in patient nº 111. This patient's original class indicated that the patient was healthy, along with the probability deduced by the ML model which was of 49.59%, indicating that the patient was healthy and making a correct prediction. The explanation for this prediction suggested that the patient's ejection fraction value was of 35% and the level of serum creatinine in blood was of 1.19 mg/dL. Both ejection fraction and serum creatinine values were de-normalized from Figure 17 before the interpretation. This explanation involves the only two relevant columns. However, having an ejection fraction percentage of 35% is not considered normal and should be contributing to having a heart failure, making the explanation a non-suitable one.

3.3.2.3 Diabetes Prediction Dataset

Figure 18 represents the clinical case of patient nº 222, whose true class corresponds to being healthy from diabetes, and the probability deduced by the ML model for the diabetes class is of 44.36%. This leads to a correct prediction of negative in diabetes by ML model. The glucose level of this patient was of 119 mg/dL, the BMI corresponds to a value of 25.2 and the final fact of this explanation is that the patient suffered from 7 pregnancies, once more all de-normalized values. For this explanation, the three relevant columns are contributing to the final prediction. Regarding its content, once again the fact that the BMI with a value of 25.2 is contributing to the decision of being healthy is not clinically logical and therefore makes the explanation doubtful.

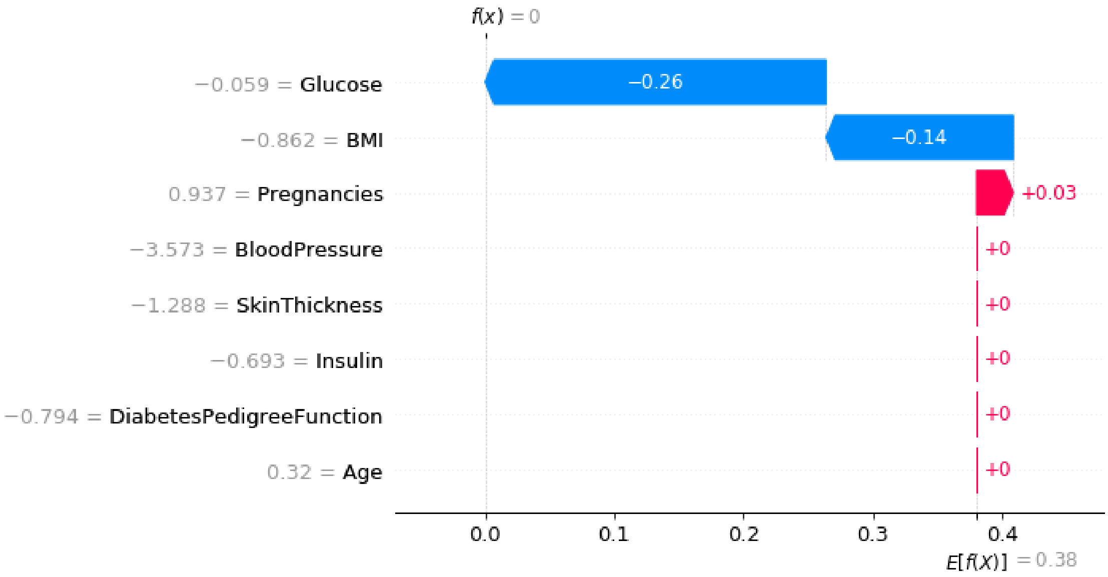

Figure 18. SHAP explanation for class diabetes in patient nº 222 in Diabetes Prediction Dataset

3.4 Locating incorrect predictions with plausible explanations

3.4.1 LIME algorithm

3.4.1.1 Stroke Prediction Dataset

In Figure 19, the explanation for a patient`s stroke condition is presented. The true class was of positive in stroke once again. The probability of having a stroke deduced by the ML model this time was of 40.45%, thus negative in stroke and prediction is incorrect. Although the explanation includes a feature which is not relevant, it is presenting facts which are clinically logical. It explains that the patient’s age should be between 45 and 61 years, which is not a critical age in stroke risking, an average glucose level in blood that is between 77.24 and 91.88 which is a normal healthy range, and it is a male who has more risk of stroke than a female. This is a potentially dangerous case since the true class is of stroke and the prediction stands the contrary, making even a reasonable explanation for it.

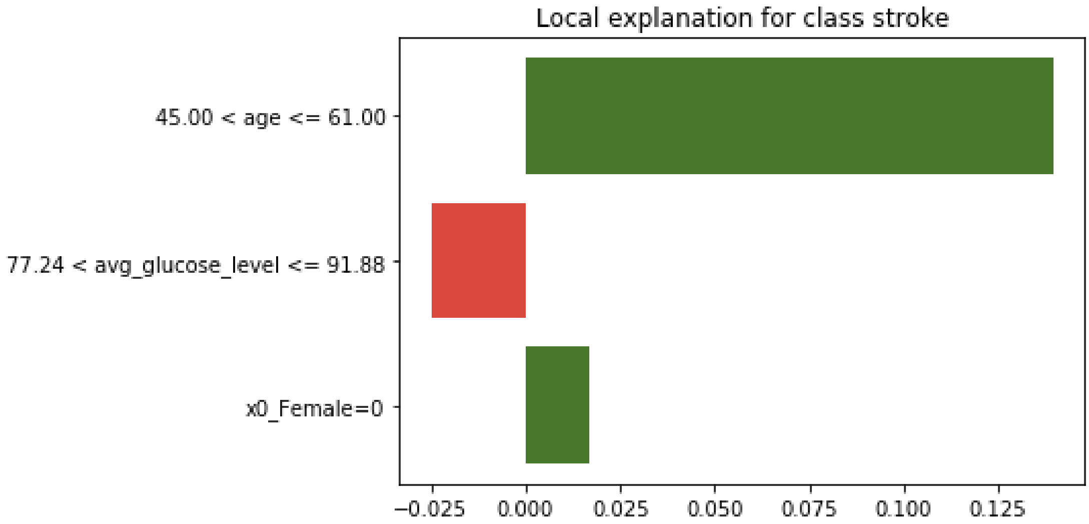

Figure 19. LIME explanation for the class stroke in instance nº 34 in Stroke Prediction Dataset

In the same way, Figure 20 represents the clinical case of patient nº 1000. The true class of this patient indicated that the patient was healthy, while the probability of having a stroke deduced by the ML model this time was of 70.07%, hence positive in stroke and leading to an incorrect prediction. The explanation stands that the age is above 61 and the average glucose level in blood is higher than 114.09 mg/dL and that is has been married, which is correlated to the age. This explanation is suitable for thinking that the prediction could be correct.

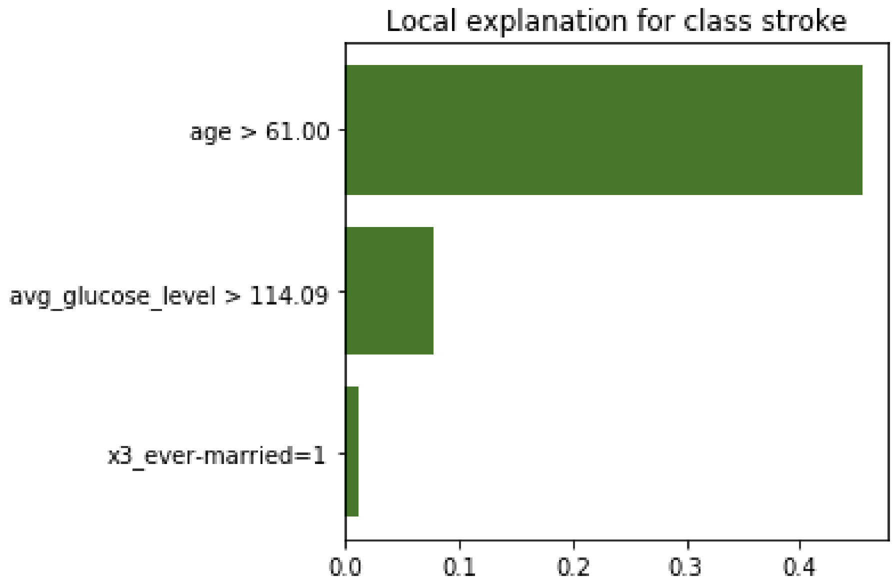

Figure 20. LIME explanation for the class stroke in instance nº 444 in Stroke Prediction Dataset

3.4.1.2 Heart Failure Prediction Dataset

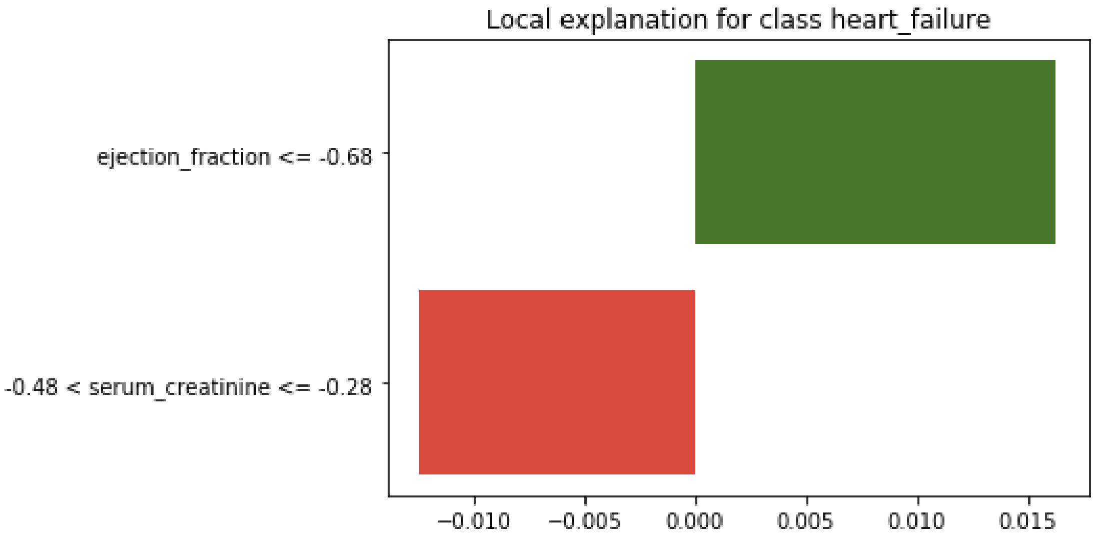

Figure 21. LIME explanation for class heart failure in patient nº 12 in Heart Failure Prediction Dataset

In Figure 21, it is shown the explanation for the patient’s diagnosis prediction. The true class corresponds to having a heart failure, while the probability deduced by the

ML model for the prediction of heart failure corresponds to a value of 49.85%, thus, negative in heart failure and therefore, leading to an incorrect prediction.

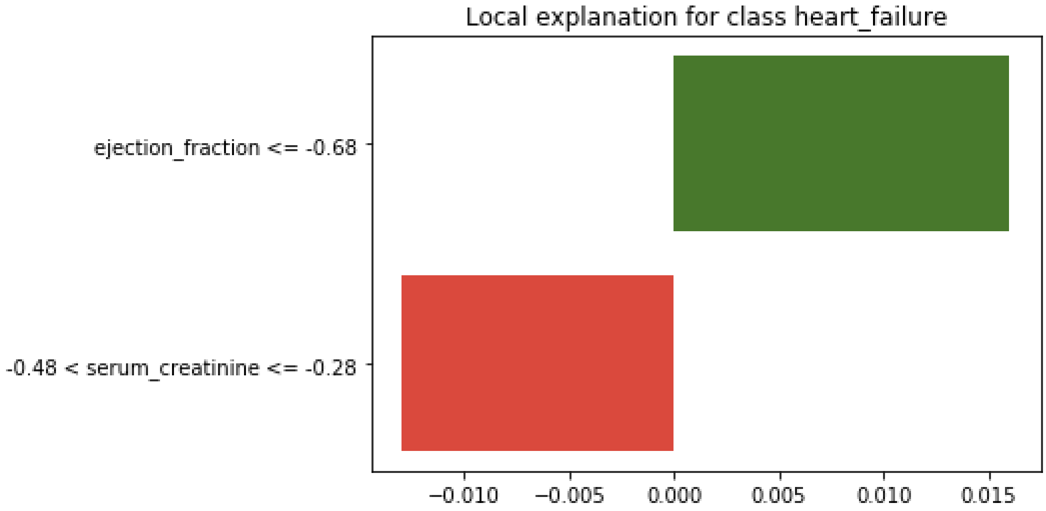

Figure 22. LIME explanation for class heart failure in patient nº 224 in Heart Failure Prediction Dataset

In Figure 22, it is presented the explanation of the prediction of patient nº 224. The true class of this patient was healthy, and the probability deduced for the prediction of heart failure by the ML model was of 50.10%, thus positive in heart failure and therefore, leading to an incorrect prediction.

Both Figure 21 and Figure 22 represent the same explanation in opposite classes and predictions, having a serum creatinine value between 0.89 and 1.10 which should be contributing to the prediction of being healthy as it does for both males and females and an ejection faction that corresponds to a value lower than 30% which should be contributing to having a heart failure as it does. Once again, the corresponding values were denormalized for the interpretation of these with the corresponding Figure 21 and Figure 22.

3.4.1.3 Diabetes Prediction Dataset

The patient`s nº 66 clinical case in presented in Figure 23. This patient’s true class was of positive in diabetes, but the probability deduced by the ML model for the diabetes prediction was of 40.54% and therefore negative for diabetes which leads to an incorrect prediction. The explanation presented for such prediction maintained how the glucose was in a range between 98 mg/dL and 117 mg/dL (de-normalized values), which is a normal healthy range that should be contributing to remaining healthy according to the medical criteria. The blood pressure indicates that is higher than 79.94 mmHg (de-normalized value), which could mean a value of 80 and still remaining within normal values. Finally, the number of pregnancies that this patient suffered was lower than 1 (de-normalized value) which should not be a risk for diabetes.

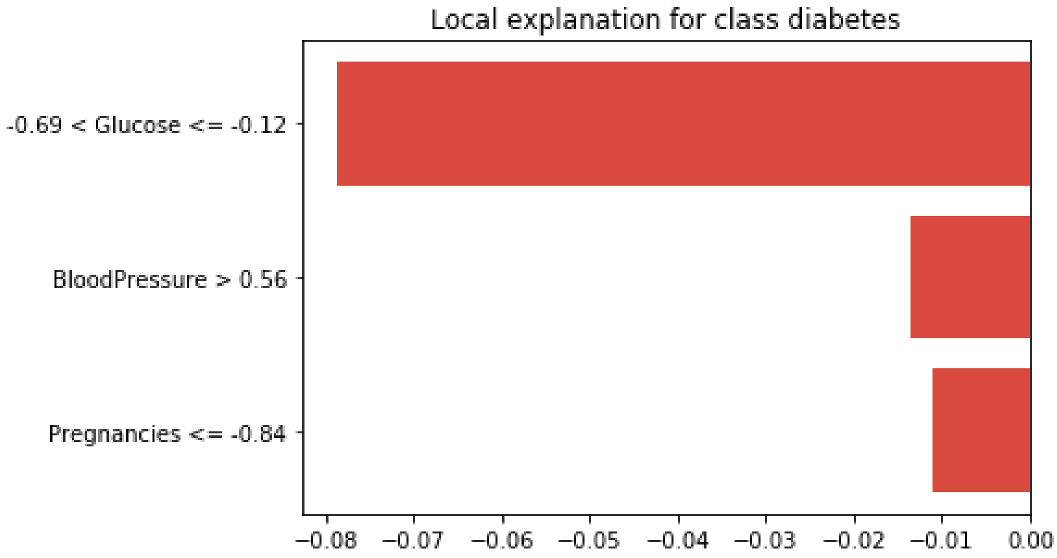

Figure 23. LIME explanation for diabetes class in patient nº 66 in Diabetes Prediction Dataset

In the same way, in Figure 24 it is presented the clinical case of patient nº 44 whose true class is of negative in diabetes. The deduced probability of the ML model for predicting diabetes was of 64.28%, standing for positive in diabetes, which leads to an incorrect prediction once again. The explanation shows as contributing facts having the glucose level above 141 mg/dL (de-normalized value), which is an indicator of prediabetes or even diabetes, and having more than 6 (de-normalized value) pregnancies (large number that increases the risk). On the other hand, the BMI range is between 27.26 and 32 (de-normalized values), which is considered as overweight/obesity.

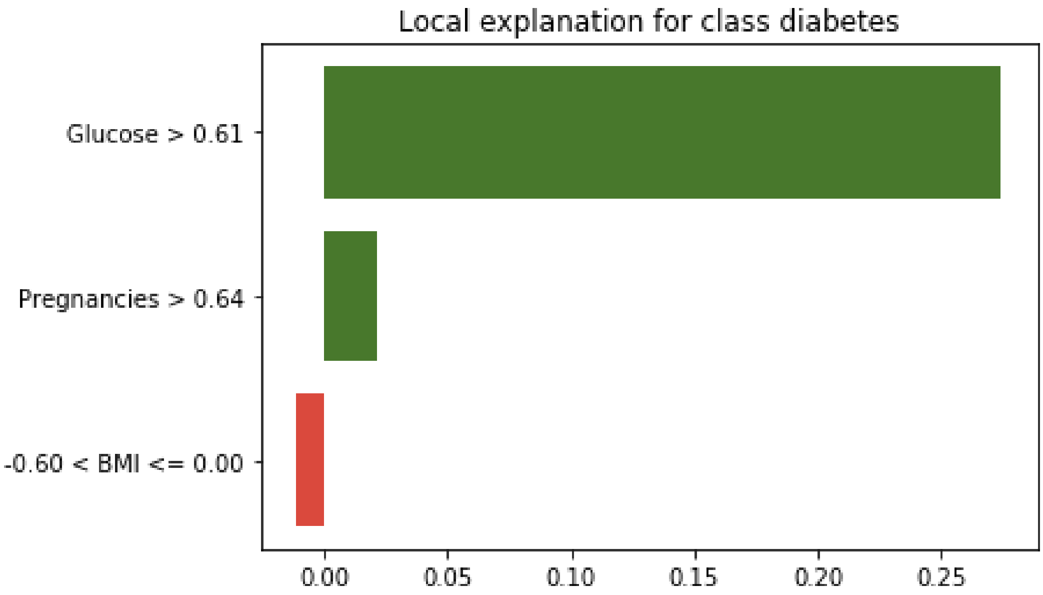

Figure 24. LIME explanation for diabetes class in patient nº 44 in Diabetes Prediction Dataset

3.4.2 SHAP algorithm

3.4.2.1 Stroke Prediction Dataset

Another example is presented in Figure 25, where the patient's true class is negative for stroke, while the probability deduced for the class by the ML model is of 51.03%, this is, positive for stroke and therefore returning an incorrect prediction. The SHAP explanation stands that both BMI with a value of 20.5, the average glucose level in blood with a value of 97.06 and an age of 54 years are contributing correctly from a clinical point of view to the healthy class, which according to the medical criteria is clinically logical. On the other hand, in Figure 26 it is presented the case of patient nº 34, whose true class indicated that the patient suffered from stroke, while the the probability deduced for the class by the ML model is of 40.45%, therefore negative in stroke and hence, leading to an incorrect prediction. The arguments for this explanation are that the age with a value of 48 years and the average glucose with a value of 84.2 mg/dL, are both contributing to the healthy class, while the BMI with a value of 29.7 is contributing to the stroke class. All these arguments are considered as relevant columns in Table 4 and make sense in the clinical way, so therefore the explanation is considered plausible. This case is a potentially dangerous case, since the patient is suffering from a stroke in the reality while the prediction is standing the contrary and it is supporting it with an explanation which is completely suitable from a medical point of view.

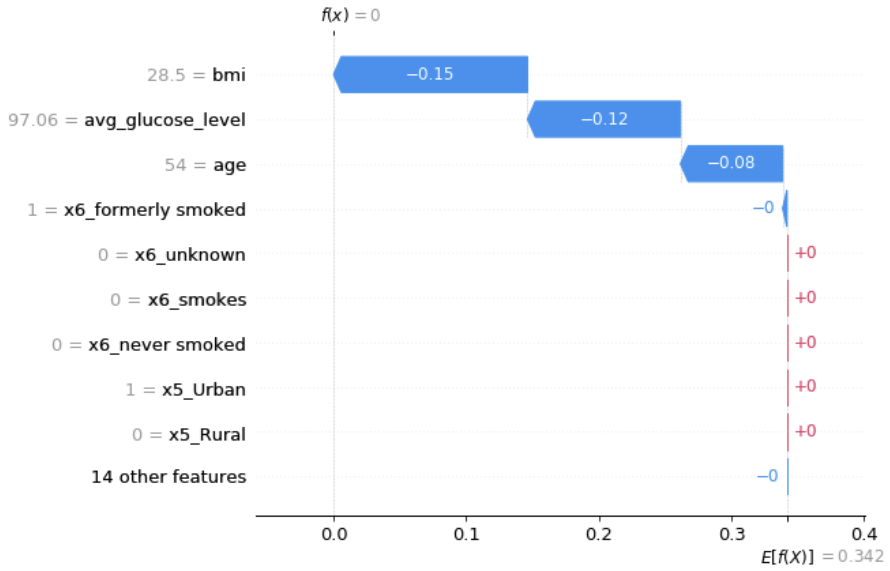

Figure 25. SHAP explanation for the class stroke in patient nº 759 in Stroke Prediction Dataset

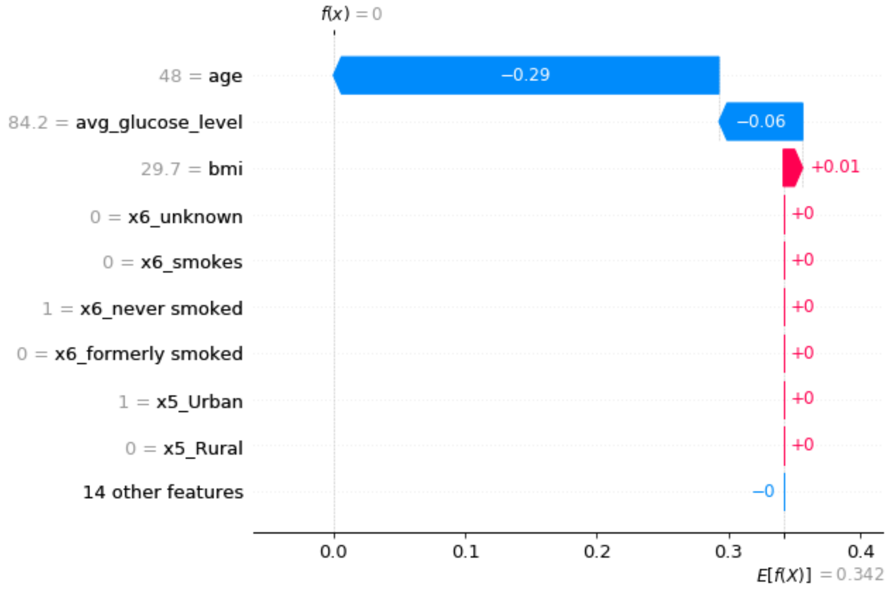

Figure 26. SHAP explanation for the class stroke in patient nº 34 in Stroke Prediction Dataset

3.4.2.2 Heart Failure Prediction Dataset

Figure 27 represents the clinical case of the patient nº 12, which is a potentially dangerous case with the original patient's class of having a heart failure and the ML model's prediction stands for remaining healthy with a probability of 49.85%. This prediction is therefore incorrect, and its explanation involves the fact of having a level of serum creatinine in blood of 1.096 (de-normalized value), which is clinically correct for a male patient; and having the ejection fraction on 30% (de-normalized value), contributing in a small way to the heart failure class, which makes sense according to the medical criteria.

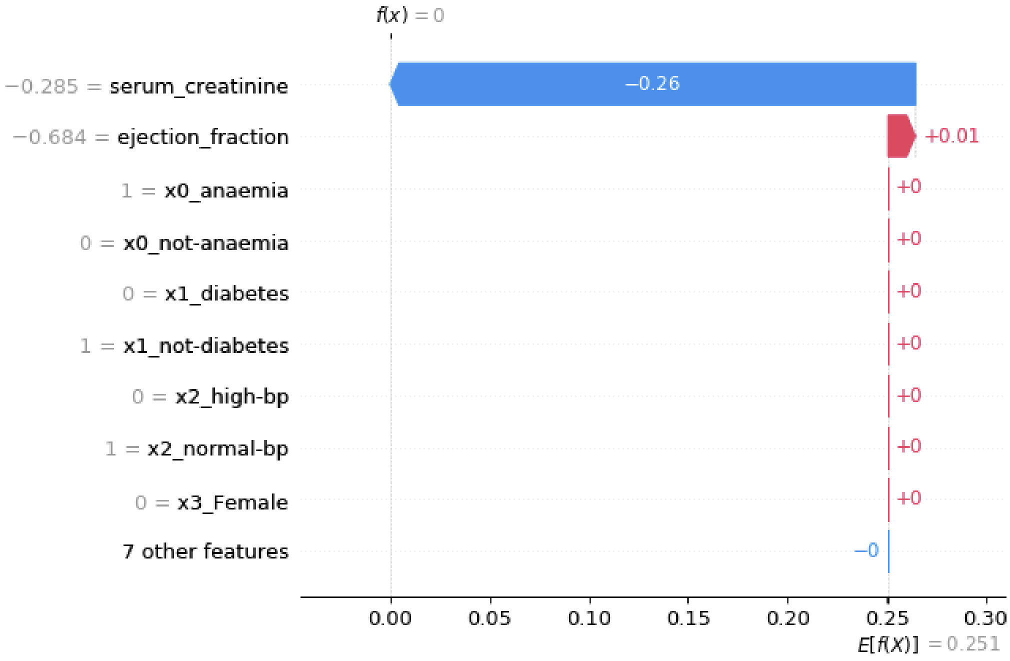

Figure 27. SHAP explanation for heart failure class in patient nº 12 in Heart Failure Prediction Dataset

3.4.2.3 Diabetes Prediction Dataset

In Figure 28 it is explained the clinical case of patient nº 44 for predicting diabetes. The true class of this patient indicates that the patient did not suffer from diabetes, even though the probability deduced by the ML model for the diabetes class is of 64.28% and therefore positive in diabetes and finally resulting on an incorrect prediction. The explanations for supporting this prediction maintained that the patient had a glucose level in blood of 159.01 mg/dL (de-normalized value), which is considered as prediabetes; the BMI was of 27.39 (de-normalized value), which is considered overweight; and the number of pregnancies that the patient experimented was of 7 (de-normalized value), which is a quite large number. All these arguments lead to a plausible explanation for matching the clinical criteria and the fact that all features are considered relevant (see Table 4).

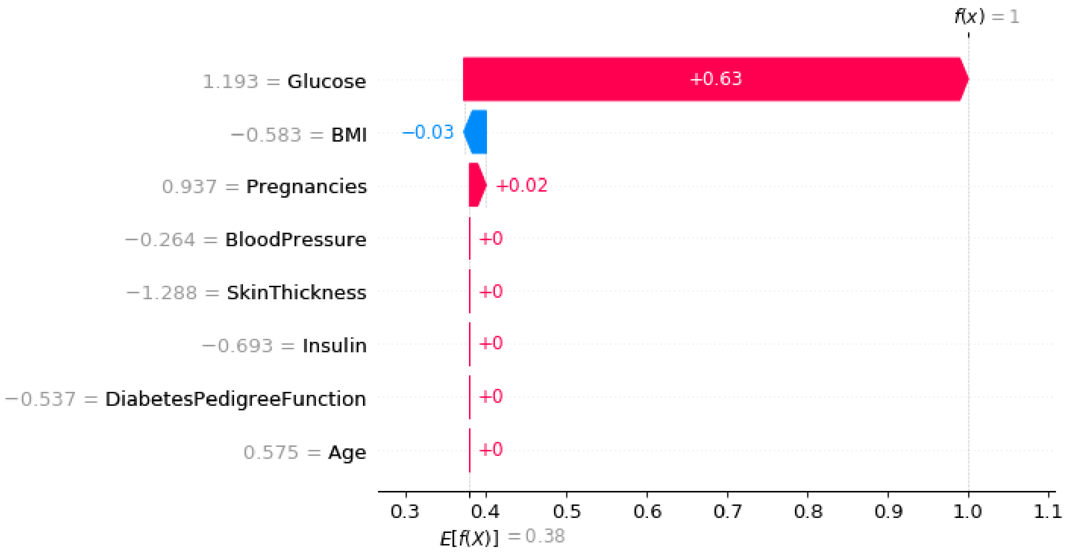

Figure 28. SHAP explanation for diabetes class in patient nº 44 in Diabetes Prediction Dataset

On the same manner, Figure 29 represents the clinical case of patient nº 469, whose true class maintained that the patient was negative for the diabetes test. The probability deduced by the ML model for the diabetes class is of 69.79%, hence positive in diabetes and finally deducing an incorrect prediction. The explanation for this instance this time indicated that the patient's glucose level in blood at the time was of 154 mg/dL (de-normalized value) which is considered as prediabetes; the BMI has the extremely large value of 46.10 (de-normalized value), which indicated obesity; and the total number of pregnancies that the patient suffered at the time was of 6 (de-normalized value). All of the items in the explanation increase the chance of having diabetes, which is clinically logical and leading therefore to a reasonable explanation.

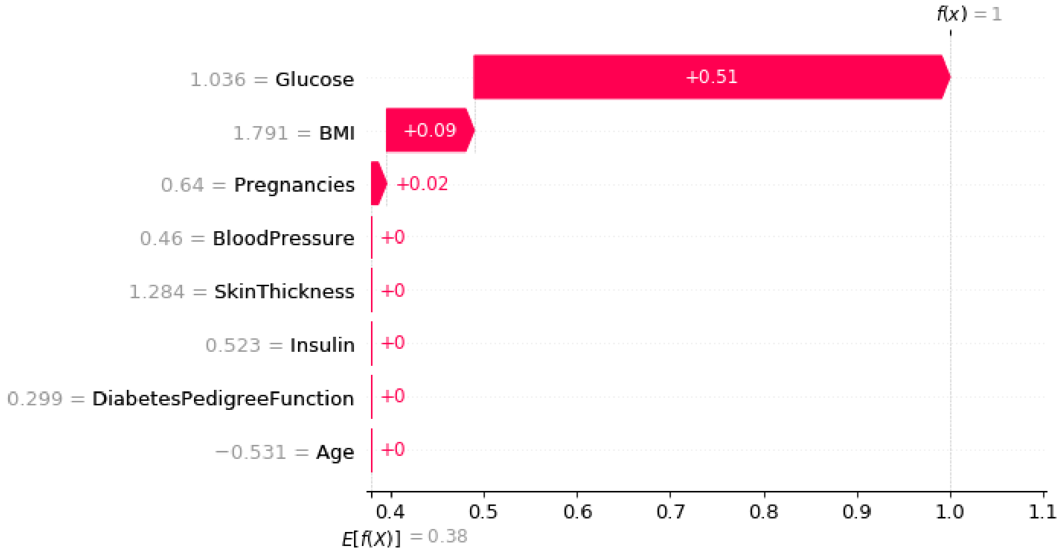

Figure 29. SHAP explanation for diabetes class in patient nº 469 in Diabetes Prediction Dataset

3.5 Contradictions in explanations

Figure 30 represents the clinical case of patient nº 50, whose true class corresponds to having a heart failure, and the probability deduced by the ML model for the heart failure class is of 50.10%. This leads to a correct prediction of positive in heart failure by ML model. The ejection fraction of this patient is of 25% and the level of serum creatinine in blood corresponds to a value of 0.99. On the same way, Figure 31 represents the clinal case of patient nº 224, whose true class corresponds to healthy, and the probability deduced by the ML model for the heart failure class is of 50.10%. This leads to an incorrect prediction of positive in heart failure by ML model. The ejection fraction of this patient is of 25% and the level of serum creatinine in blood corresponds to a value of 0.99. Once more, all the values were de-normalized for the better interpretation.

As it can be seen, the same explanation is given for contrary situations in which the probability deduced by the ML model is the same, leading to the same prediction, but the original true classes for the patients are the opposite.

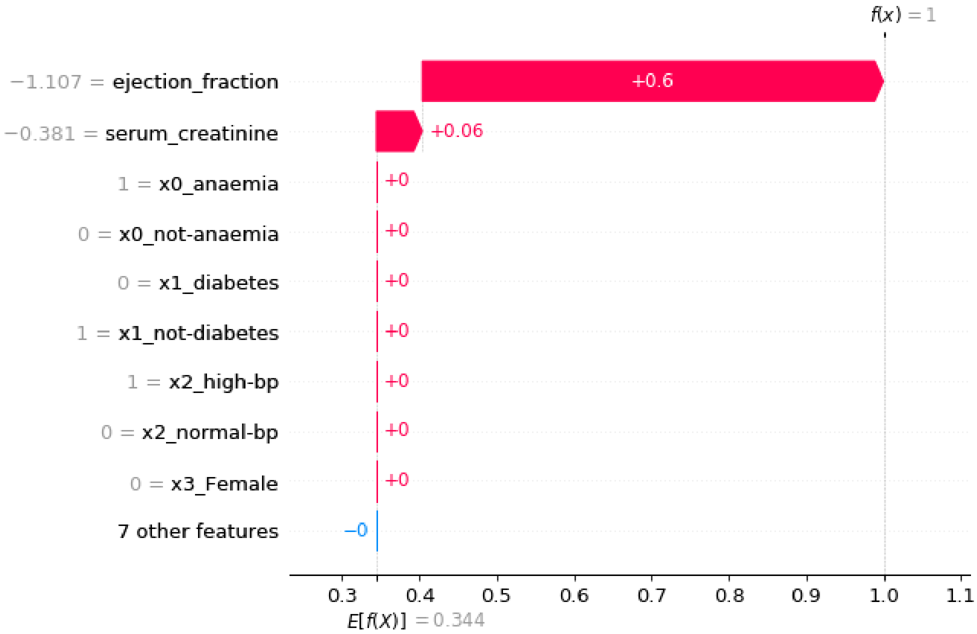

Figure 30. SHAP explanation for class heart failure in patient nº 50 in Heart Failure Prediction Dataset

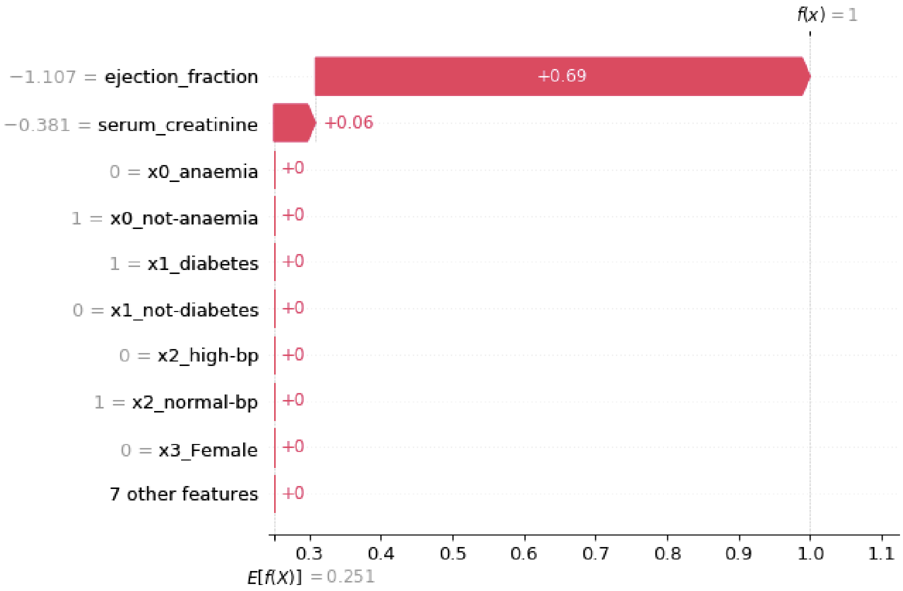

Figure 31. SHAP explanation for heart failure class in patient nº 224 in Heart Failure Prediction Dataset

3.6 Variability in explanations

Along the experiments in this thesis, it has been detected some variabilities in LIME explanations. Figure 32 and Figure 33 represent how for a single instance, LIME generates two different explanations. They both represent the explanation for patient nº 139 in Stroke Prediction Dataset. The true class declares that the patient had a stroke, while the probability deduced by the ML model for the stroke class was of 85.05%, standing for a correct prediction of positive in stroke. The female attribute referring to the gender feature which is not even considered as a relevant column appears to be contributing to the stroke class in Figure 33 while in the Figure 32 it is contributing in the opposite way to the healthy class.

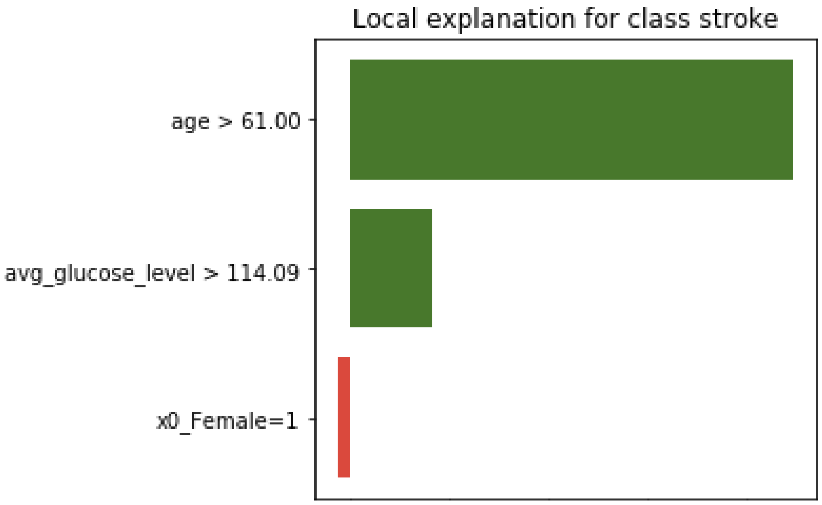

Figure 32. LIME explanation for class stroke in instance nº 139 in Stroke Prediction Dataset representing variability

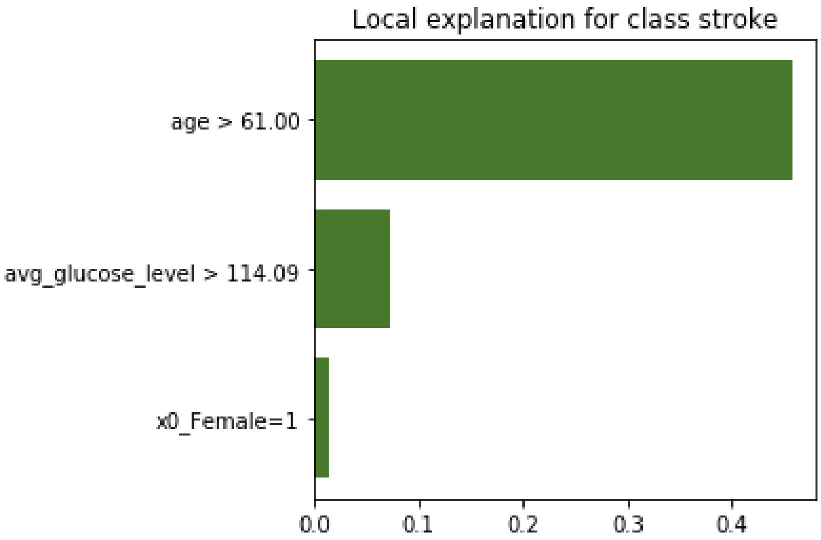

Figure 33. LIME explanation for class stroke in instance nº 139 in Stroke Prediction Dataset representing variability

4 DISCUSSION

This section will focus on demonstrating how explanation algorithms might be "dangerous" in biomedical engineering. For this, different experiments which were carried out will be discussed individually.

Firstly, the synthetic datasets were created based on the intuition of why explanatory algorithms might be dangerous. If the explanatory algorithm does not capture the functioning of the predictor algorithm, the explanations will be useless. In the experiments, it has been created a dataset where it is known that it is going to fail because the explanations are based on linear models and the underlying model is non-linear. It has been demonstrated how the explainers used in this thesis can fail when confronted with non-linear regions. The fact that EML algorithms fail is also supported by Tables 5 and 6. Referring to Table 5 and Table 6, it is shown that not always all the columns that are relevant appear in the explanations or that sometimes even some that are not relevant appear. Again, this shows that indeed the explanations do not always capture the true behavior of the classifier. Moreover, basing on this quantification it is seen how the performance of SHAP seems to be better than LIME.

These experiments demonstrate that LIME is not able to provide good explanations in the neighborhood of non-linear boundaries, as expected. This result highlights that EML explanations where interactions between variables are expected should not be trusted, since they cannot capture the behavior of the underlying algorithm. Furthermore, the experiments also highlight that if boundaries are not parallel to the features-axes, they can be hard to interpret, even if those boundaries are linear. This is due to LIME using intervals for making up its explanations.

Regarding the non-plausible explanations for correct predictions in LIME and SHAP algorithms, Figure 9 to Figure 18 demonstrate this situation clearly, since they expose columns which are not consider relevant or clinical arguments which do not make any clinical sense. Even if a classifier had a 100% accuracy, if the physician observes that the explanations are bizarre, he will not pay attention to it. This prevents very promising artificial intelligence methods from reaching clinical practice. This is not a dangerous situation, but it is not desirable either.

Regarding the plausible explanations for incorrect predictions, Figure 19 to Figure 29 demonstrate this situation in a clear manner, since the explanations show reasonable and convincing arguments from a medical point of view. All the explained cases are dangerous situations where incorrect predictions involving human lives are being taken, which could lead to fatal consequences. Consider, for example, the case of incorrectly predicting that an individual has not cancer and supporting this prediction with a convincing explanation. These explanations are hazardous because they could mislead a physician and cause a disease to go undiagnosed.

Examples of contradictions in explanations are presented through Figure 30 and Figure 31, and also Figure 21 and Figure 22. This situation once more makes no sense in the clinical practice, because two different clinical cases give opposite diagnosis cannot have the same explanation. Again, this type of behavior could undermine the confidence of the clinician in the predictive model, even in the case that it was 100% accurate.

Finally, the variability in LIME explanations is reproduced through the Figure 27 and Figure 28, with different explanations corresponding to the same instance in the Stroke Prediction Dataset. Although some variability due to the stochastic nature of LIME is expected, this can result in distrust on the part of clinicians if they are not correctly instructed in how the algorithm work. On the positive side, it should be noted that weights associated with varying explanations are usually close to 0.

CONCLUSIONS AND FUTURE WORK

In this thesis, we have presented a critical evaluation on EML in the context of healthcare data. To that end, LIME and SHAP algorithms were tested using three different datasets, which were then screened for explanations that may result harmful in the clinical practice.

First, we demonstrated that EML algorithms do not always capture the behavior of the classifier they try to explain. Specifically, we have shown that the problem may be particularly important if the underlying classifier has non-linear boundaries. These conclusions are also supported by Table 5 and Table 6, which show that features that are not really used by the classifier can appear in the explanations. Furthermore, Table 5 and Table 6, show that SHAP algorithm usually performs better than LIME, although this was not the main concern of the thesis. We have also demonstrated the existence of two potential critical situations in the clinical context. On one hand, of the existence of plausible explanations for incorrect predictions, which may boost confidence on wrong predictions by providing sound explanations. This situation is particularly dangerous since it can bias life or death decisions. On the other hand, we have also shown examples of non-plausible explanations for correct predictions. Although this situation is not as dangerous as the first one, it is undesirable, since it can undermine trust on the underlying predictive model (even if it is 100% accurate) by providing explanations that are not consistent with medical knowledge.

Under these circumstances and due to the fact that there is not any clear evidence of the algorithms working properly, using these tools in for clinical problems should be done under a lot of caution, especially when a final diagnose is wanted. As a suitable alternative for using ML in the clinical practice, we suggest the use of inherently interpretable models such as, for example, the Lasso or decision trees. A model is inherently interpretable if humans can understand the decisions it makes. By using interpretable models, we can avoid the use of EML algorithms, since they provide their own explanations.

We do not claim, however, that EML algorithms should be always avoided. Indeed, they can be very helpful if they are not used as a proxy between the clinician and the classifier. For example, they can help researchers in detecting problems with datasets or with the convergence of a learning algorithm. In this context, strange explanations can be a sign of some of the previous issues. In line with the previous conclusions, we believe that the only way that EML methods can be used in clinical practice is by designing EML algorithms with guarantees that the explanations capture the behavior of the underlying model. If such an algorithm exists should be subject of future research.

5 REFERENCES

[1] Belle, V. and Papantonis, I., 2020. Principles and practice of explainable machine learning. *arXiv preprint arXiv:2009.11698*.

[2] Roscher, R., Bohn, B., Duarte, M.F. and Garcke, J., 2020. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, *8*, pp.42200-42216. (ISSN: 2169-3536)

[3] Greenspan, H., Van Ginneken, B. and Summers, R.M., 2016. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, *35*(5), pp.1153-1159. (ISSN: 0278-0062)

[4] Samek, W. and Müller, K.R., 2019. Towards explainable artificial intelligence. In *Explainable AI: interpreting, explaining and visualizing deep learning* (pp. 5-22). Springer, Cham. (ISBN: 978-3-030-28953-9)

[5] Tonekaboni, S., Joshi, S., McCradden, M.D. and Goldenberg, A., 2019, October. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine Learning for Healthcare Conference* (pp. 359-380). PMLR.

[6] Ploug, T. and Holm, S., 2020. The four dimensions of contestable AI diagnostics-A patient-centric approach to explainable AI. *Artificial Intelligence in Medicine*, *107*, p.101901. (ISSN:0933-3657)

[7] Holzinger, A., 2018, August. From machine learning to explainable AI. In *2018 world symposium on digital intelligence for systems and machines (DISA)* (pp. 55-66). IEEE. (ISBN: 978-1-5386-5103-2)

[8] Tjoa, E. and Guan, C., 2020. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*. (ISSN: 2162-237X)

[9] Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), pp.206-215.

[10] Ahmad, M.A., Eckert, C. and Teredesai, A., 2018, August. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics* (pp. 559-560). (ISBN: 978-1-450-35794-4)

[11] Yu, H.F., Huang, F.L. and Lin, C.J., 2011. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, *85*(1-2), pp.41-75.

[12] Web page of Scikit-learn, Machine Learning in python. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (Accessed: June 2021)

[13] Lundberg, S. and Lee, S.I., 2017. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

[14] Web page of Christoph Molnar of Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book/lime.html (Accessed: June 2021)

[15] Web page of Shap Documentation. https://shap.readthedocs.io/en/latest/index.html# (Accessed: June 2021)

[16] Web page of Github. https://github.com/palaciosmaria/TFG-EML (Accessed: July 2021)

[17] Web page of Stroke Prediction Dataset | Kaggle. https://www.kaggle.com/fedesoriano/stroke-prediction-dataset (Accessed: June 2021)

[18] Web page of the UCI Machine Learning Repository, Center for the Machine Learning and Intelligent Systems. https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records# (Accessed: June 2021)

[19] Chicco, D. and Jurman, G., 2020. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC medical informatics and decision making*, *20*(1), p.16. ISSN: 1472-6947

[20] Web page of Pima Indians Diabetes Database | Kaggle. https://www.kaggle.com/uciml/pima-indians-diabetes-database (Accessed: June 2021)

[21] Web page of Machine Learning Mastery. https://machinelearningmastery.com/case-study-predicting-the-onset-of-diabetes-within-five-years-part-1-of-3/ (Accessed: June 2021)

[22] Web page of Scikit-Learn, Machine Learning in python. https://scikit-learn.org/stable/user_guide.html (Accessed: July 2021)

[23] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J. and Layton, R., 2013. API design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*.

[24] Web page of LIME. https://lime-ml.readthedocs.io/en/latest/lime.html#module-lime.lime_tabular (Accessed: July 2021)

[25] Web page of Quanam | Camila Palomeque, Consultora unidad Data & Analytics. https://quanam.com/interpretabilidad-de-los-modelos-de-machine-learning-segunda-parte/ (Accessed: June 2021)

[26] Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). (ISBN: 978-1-4503-4232-2)

[27] Web page of Christoph Molnar of Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book/shapley.html#shapley (Accessed: July 2021)

[28] Web page of Christoph Molnar of Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book/shap.html#kernelshap (Accessed: July 2021)

[29] Web page of CDC, Centers for Disease Control and Prevention. https://www.cdc.gov/obesity/adult/defining.html (Accessed: June 2021)

[30] Web. Page of Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/diabetes/diagnosis-treatment/drc-20371451 (Accessed: June 2021)

[31] Yousufuddin, M. and Young, N., 2019. Aging and ischemic stroke. *Aging (Albany NY)*, *11*(9), p.2542-2544.

[32] Web page of Cleveland Clinic. https://my.clevelandclinic.org/health/diseases/17069-heart-failure-understanding-heart-failure (Accessed: June 2021)

[33] Metra, M., Cotter, G., Gheorghiade, M., Dei Cas, L. and Voors, A.A., 2012. The role of the kidney in heart failure. *European heart journal*, *33*(17), pp.2135-2142. (ISSN: 0195-668X)

[34] Zamora, E., Lupón, J., Urrutia, A., González, B., Mas, D., Díez, C., Altimir, S. and Valle, V., 2007. Prognostic significance of creatinine clearance in patients with heart failure and normal serum creatinine. *Revista Española de Cardiología*, *60*(12), pp.1315-1318. (ISSN: 1311-3938)

[35] Kelly-Hayes, M., 2010. Influence of age and health behaviors on stroke risk: lessons from longitudinal studies. *Journal of the American Geriatrics Society*, *58*, pp. S325-S328. (ISSN: 0002-8614)

[36] Web page of CDC, Centers for Disease Control and Prevention. https://www.cdc.gov/reproductivehealth/maternalinfanthealth/diabetes-during-pregnancy.htm (Accessed: June 2021)