Identification of heart problems through predictive models based on supervised machine learning

Matheus Henrique Lopes Cunha

Paloma Melo Pereira

ABSTRACT

The predictive models of "Supervised machine learning" are becoming increasingly important in aiding decision-making for various areas of human knowledge and, consequently, will also be important for assisting in medical cases of greater complexity. The goal of the study is to develop a supervised machine learning algorithm that can have a high success rate in predicting whether a person has a heart problem or not. The article shows how the models were developed, the tests applied before the implementation of the models, the utilization rate of each model and an analysis of which is the most efficient model for a specific situation. The specifications of each supervised machine learning model and its impact on the development of the models that were used in the work, determined by testing and applications made in the Python programming language; The positive and negative results were considered to reach a final position on what was the best way to use the algorithms in this case. The article concludes that the application of supervised machine learning models in the diagnosis of heart problems can help many health institutions, both public and private, to streamline processes and increase the success rate when classifying a person as cardiac or non-cardiac, to have a high improvement in this process and consequently increase efficiency and profitability in the case of private institutions.

Keywords: XGBoost, Random Forest, KNeighbors Classifier, Health, Heart.

1 INTRODUCTION

A person's lifestyle habits are directly linked to their health and predisposition to the development of cardiovascular diseases. Cardiovascular disease is the leading cause of illness and death in Brazil, affecting both sexes. Atherosclerotic lesion is identified as the fundamental etiological factor of this condition. (DINIZ, 2011). Hypertension, high cholesterol and glucose levels, alcohol consumption, smoking, obesity and a sedentary lifestyle are all factors that contribute to the onset of cardiovascular diseases.

According to the Brazilian Society of Cardiac Arrhythmias, hypertension is related to the onset of cardiac arrhythmias, with an emphasis on atrial fibrillation, as well as heart failure with heart dilation (MAGALHÃES et al, 2016). This is justified because hypertension severely affects important arteries in the body, such as the coronary arteries, which irrigate the heart.

Cardiovascular diseases are closely related to age and biological sex, since certain diseases are linked to certain age groups. In children and young people, the occurrence of congenital anomalies and the first episodes of rheumatic disease are more common. From 20 to 50 years of age, Chagas disease and hypertension are more frequent. On the other hand, coronary artery disease, including angina pectoris and

acute myocardial infarction, is more common in people over fifty years of age (PORTO, 7th edition, 2014). It is also possible to affirm that biological sex has an impact on the types of cardiovascular diseases prevalent in men and women. Mitral lesions, such as mitral valve stenosis and prolapse, are more common in young women. On the other hand, coronary atherosclerosis is more prevalent in men up to 45 years of age. From this age group, the incidence of the disease becomes equal in both sexes (PORTO, 7th edition, 2014). Coronary atherosclerosis equals its incidence in men and women at the age of 45, because it is at this average age that women enter menopause and there is a reduction in the production of the hormone estrogen, which is naturally a protective substance of blood vessels. Estrogen promotes the release of vasodilatory substances in the endothelium, such as nitric oxide, and decreases the production of vasoconstrictor compounds. Therefore, before menopause, women have lower blood pressure values than men and are less likely to develop heart problems. (MACIEL et al, 2021).

Another factor related to cardiovascular diseases is smoking. It is associated with increased cholesterol levels, thrombosis, platelet aggregation, and coronary heart disease. The components present in cigarettes propel the activation and release of inflammatory cells that lead to an increase in inflammatory mediators in the bloodstream. In turn, these inflammatory regulators trigger the elevation of other compounds that are related to a higher risk of developing myocardial infarction and coronary heart disease. In addition, cigarettes lead to changes in the intimal layer of blood vessels, decreasing nitric oxide concentrations, i.e., smokers have lower amounts of this compound when compared to nonsmokers (NUNES et al, 2011).

A sedentary lifestyle is also considered an etiological agent of cardiovascular diseases. Physical inactivity causes there to be no appropriate venous return, because the muscles do not contract effectively, so the joint work of the muscles and venous valves is compromised, favoring the formation of thrombi. Thrombus are associated with major cardiovascular disorders, including ischemic heart disease. In addition, a sedentary lifestyle leads to an increase in heart rate and is a direct cause of obesity (GASPAR, 2004).

High cholesterol levels are linked to inflammatory events of the blood vessels that result in the formation of atheroma plaques that lead to ischemic heart problems. Some conditions such as diabetes, hypertension and smoking can damage the endothelium of the blood vessels, generating an inflammatory process that increases the permeability of the intima to lipoproteins and causes the subendothelial space to accumulate cholesterol and form the atheromatous plaque, which grows towards the lumen of the vessels and reduces blood flow. In addition, if the atherosclerotic plaque suffers an injury, coagulation occurs at the site and thrombus formation, which further impairs the passage of blood in this region. If the heart is irrigated by an artery that is obstructed by atheromatous plaques, the oxygen supply to the heart cells will be low, increasing the chances of ischemia and infarction (XAVIER et al, 2013). In addition, it is observed that in

populations with low plasma cholesterol levels, cases of heart attacks and atherosclerosis are infrequent (CASTRO et al, 2004).

Finally, people with high blood glucose levels are also more likely to develop heart problems. The intracellular increase in glucose and pro-inflammatory substances in people with hyperglycemia leads to tissue damage in the blood vessels capable of forming unstable atherosclerotic plaques that are more prone to rupture. In addition, elevated glucose levels decrease antiplatelet antiplatelet complexes while elevating platelet-activating compounds that together contribute to the formation of platelet aggregation and the generation of thrombi (Schaan; Portal, 2004).

Therefore, it is evident how several associated factors lead to the emergence of cardiovascular problems, and it is necessary to implement preventive measures and early diagnosis to avoid possible sequelae. According to the Ministry of Health, the early detection of diseases is based on the idea that some conditions have greater chances of cure, survival and/or quality of life when diagnosed as early as possible. There is significant value in identifying the disease in early and asymptomatic stages, because during this period treatment and cure can be achieved more easily (UMintimo da Saúde, 2010).

As it is a very serious problem that generates devastating sequelae, which can even lead to death. This work uses supervised machine learning models to predict whether the person has a heart problem or not, thus making the detection of heart problems faster and more accurate. The algorithms will take into account various characteristics and habits of the person who will have the heart analysis judged such as age, glycemic rate, cholesterol, whether the person is a smoker or not, whether the person consumes alcohol or not, whether the person is active in relation to physical activities, and the relationship that all the previous characteristics have with the evaluation of some people who have already had the diagnosis confirmed. It has been stated that machine learning can be understood as the solution of geometric problems in the field of artificial intelligence. According to his words, the models used in this context are programmed to analyze information from a data set, acquire knowledge, and improve themselves based on their experience within the parameters provided (Piovezan, 2022).

In view of the above, realize that this analysis is a technique that involves powerful technology, with the great purpose of speeding up the classification process and bringing more accurate results about the real cardiac diagnosis of a person. With this mentioned, it can be concluded that the study of the diagnosis of heart problems with "supervised machine learning" is extremely important and generates a lot of knowledge. Technology has the power to provide unprecedented opportunities, allowing visually impaired people to actively participate in social networks, interact, comment and have the same access as all other users. This inclusive capacity is considered wonderful, reflecting a vision of the future that seeks to incorporate more and more individuals. Children can now enjoy videos and have fun, and the expectation is that more and more people will be included in the web, resulting in a different world. Although some online environments

are still toxic, the author points out that there are people who want and support healthy environments (Junior, 2022).

This work aims to create a supervised machine learning algorithm that has a high accuracy rate when giving a diagnosis of a heart problem. The data presented were taken from the database on heart problems, which can be found on the Kaggle website. The data was analysed using the artificial intelligence of the Python programming language, through which statistical analysis and predictions with machine learning were provided.

2 MATERIAL AND METHODS

A heart problem, also known as heart disease, refers to any condition or disease that affects the normal functioning of the heart. When heart problems occur, it can lead to a number of serious and even life-threatening medical complications.

Preventing heart problems is crucial for maintaining heart health and reducing the risk of cardiovascular disease. Preventions can be: having a healthy diet, maintaining a normal weight, exercising frequently, limiting alcohol consumption, having blood pressure under control, controlling diabetes, reducing stress, getting enough sleep, having regular medical checkups, avoiding the use of illicit drugs, and following medical advice.

The research was designed with the objective of understanding the statistical information related to the analysis of those who have a heart problem or not. Through this analysis, we create supervised machine learning models to determine its final quality in a safe way, in order to understand how such a space was formed and developed.

The methodology used in this research is explanatory and descriptive in nature. The study was quantitative and qualitative, using data analysis and the use of supervised machine learning algorithms by the Python programming language to interpret the data. In view of this, the quantitative study is usually carried out by: In studies, data collection is commonly carried out through questionnaires and interviews, which cover different variables relevant to the research. This collected information is usually presented through tables and graphs during the analysis. (DALFOVO, LANA AND SILVEIRA, 2008). On the other hand, in the qualitative study Carspecken (2011, p.27), it is common for a qualitative social researcher to be interested in understanding the functioning of forms of power, especially in real interactions that he observes and in which he possibly participates (Carspecken, 2011).

The programming codes used in this work and version of the software and packages used are on "Github", the link is available in the appendices section, this platform is an excellent alternative to store the codes, since the "scripts" are in the cloud, which makes it possible to avoid the loss of codes due to any problem that the computer presents.

The database that was used in this study was obtained through Kaggle, which is a famous machine learning competition community. And the database has eight columns with information from the variables that were analyzed, seventy thousand rows with information from the people who were analyzed, and it was submitted to the website in January 2023.

A previous study of each variable present in the database was also carried out, to validate the type that each variable presented, in addition to having a basic notion of how each information would be used. The description of each type of information can be seen in Table 1.

Table 1. Variables

Variable	Kind	Description of the data
Acts	*Entire*	Age of participants
Gender	*Text*	Gender of participants
Height	*Entire*	Height of participants
Weight	*Entire*	Weight of participants
Ap_hi	*Entire*	Participants' systolic pressure
Ap_lo	*Entire*	Diastolic pressure of participants
Cholesterol	*Entire*	Whether the patient smokes or not
Gluc	*Entire*	Whether the patient consumes alcohol or not
Smoke	*Boolean*	Whether the patient is active or not
Alco	*Boolean*	Whether the patient has a heart condition or not
Active	*Boolean*	
Cardio	*Boolean*	

Source: Original survey data

With the data already classified, and with the identification that it was not necessary to do a deeper cleaning in the database. The data were submitted to an exploratory analysis, in which it was understood which would be the best model options for the specific case.

The explanatory study is a method of scientific analysis that seeks to explain how it works and the performance achieved by the models of the segment studied. The work comes in the quantitative mold, as the intention is to bring an approach with numerical data, which will be presented in graphic, discursive and statistical format, in order to understand which is the "supervised machine learning" model that has greater applicability for the study in question. In general, quantitative field studies follow a research model similar to experimental research, in which the researcher uses well-structured conceptual frames of reference to

formulate hypotheses about the phenomena and situations he or she wishes to investigate (Dalfovo, Lana and Silveira, 2008). And the work also has qualitative traits, as there is an analysis of significant and transformative data, critical qualitative research is described as a truly stimulating, political and meaningful practice, capable of expanding the mind. Both fieldwork experiences and data analysis are mentioned as processes rich in meaning and with transformative potential (Carspecken, 2011).

Machine learning is defined as: the components of machine learning involve a set of variables called "features", which can be measured or pre-defined, along with a set of outputs, which can be known or unknown. The process of building the model is based on the use of a dataset composed of examples (Tome, 2017).

Machine learning is a complicated tool, but it can help in professional and personal development, the concepts and tools that exist in it, facilitate the ways to reach the goal. Incremental learning is characterized by the dynamic accumulation of information extracted from lived experiences. The adaptive approach to machine learning aims to integrate symbolic machine learning techniques with adaptive techniques in order to solve learning problems efficiently (Stange et al, 2011).

There are several ways to use machine learning, which are also very important factors, as it is with them that the best performance model that will be used in each situation is defined. During the training phase, the presence of irrelevant and redundant attributes makes it difficult to learn the classifier. One approach to dealing with this issue is to select the attributes that are most important for ranking, i.e., those that have the greatest ability to distinguish between positive and negative news. This selection aims to remove irrelevant attributes and improve the classifier's performance (Almeida, 2014).

The classifiers chosen for the efficiency tests of the supervised machine learning models that were used to achieve the results of this study use the methods of "Random Forest", "XGBoost" and "KNeighbors Classifier". The models were tested to see which would be the most efficient classifier.

In order not to generate "overfitting" and "underflighting" and to generate a more reliable model, the division of the data into test and training models was used, because when the data goes to the supervised machine learning algorithms directly, it tends to get biased predictions. In situations where the priority is to select the model with the best predictive capability, caution should be exercised with overfitting and underfitting. Overfitting occurs when the model overfits the training data, resulting in inappropriate predictions when applied to new data. Underfitting, on the other hand, refers to the situation in which the model does not fit properly even to the training set (Lopes, 2018).

To apply the supervised machine learning models with the already improved data, statistical tests were performed to see the importance of each attribute and consequently whether the attribute could be irrelevant to the model. Therefore, two columns were excluded from the database: height and weight. These two variables alone do not have an impact on cardiovascular problems, making it necessary to manipulate

these elements through the Body Mass Index. However, this index is not as effective because it does not take into account body composition as well as the distribution of fat in the individual's body (RECH et al, 2006). In addition, studies by Hans et al have shown that waist circumference is more closely related to cardiovascular disease risk factors (Hans et al, 1995). The boxplot was analyzed to detect outliers and correlation was analyzed to analyze possible relationships between the variables. In addition, the data were also checked for normality to validate whether the number of samples is sufficient for the model.

After all the static analyses, the "GridSearchCV" function was used, with cross-validation in 10 folds, to define the best parameters for each model. Then, the data were divided into 80% for training and 20% for testing, where the 70000 samples were divided into 56000 for training and 14000 for testing, randomly selected by the "train_test_split" function. The purpose of implementing the technical process standard is to minimize changes in process control parameters when introducing a new product. This aims to improve the efficiency of the machine setup, reduce productivity and quality losses, and eliminate the variability of specifications that arise during production (Campos and Miguel, 2013).

The models were evaluated for their accuracy, sensitivity and precision, calculated using their confounding matrix. Another important analysis for this study is the noise test. It was verified up to which percentage of noise inclusion in the models still maintain good performance. The data was found on Kaggle's website. The database has seventy thousand samples, and these samples have twelve variable samples analyzed.

Methods and tools that are not efficient need to be modified and improved. Thus, the use of several techniques was of great importance for the study, because in some very positive results were found and consequently they were maintained, and others were inefficient and discarded. It is noteworthy that the widely recognized technique as "machine learning" or "Machine Learning" has played a significant role in several areas. This approach consists of programming computers to learn from previous experiences, going beyond simply reproducing the data provided. The system develops its own cognitive capacity, allowing continuous learning based on successes and failures (De Figueiredo and Cabral, 2020).

In this context, the goal of the algorithm is to produce a classifier that can predict whether a certain person has a heart problem or not, even when the information is not very clear to statistical analyses.

Statistical equations used in the models. In "XGBoost" the equation eq was used. (1):

$$L_t = \sum_{i=1}^{n} l(y_i, \hat{y}_i^t) \quad (1)$$

In the "Random Forest" we used the formula for measuring gini inequality, which is shown in the equation eq. (2):

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i)), \tag{2}$$

The "KNeighbors Classifier" aims to calculate by proximity, so elements that are close to or have very similar characteristics are assigned as of the same class, and is represented equation eq. (3):

$$D_E(p,q) = \sqrt{(p_1 - q_1)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{3}$$

3 RESULTS AND DISCUSSION

The following work aims to show how supervised machine learning can be used to identify heart problems. Presenting what really works well and what needs to evolve in relation to machine learning tools within this segment. Machine learning is described as an intensified scientific method. It follows a similar process of generating, testing, and discarding or refining hypotheses. However, while it can take a scientist a lifetime to create and test a few hundred hypotheses, a machine learning system is capable of accomplishing the same in a fraction of a second. Machine learning automates the discovery process, which explains why it is revolutionizing both science and business (Domingos, 2017).

The elaboration of the model was based on the statistical analysis of some characteristics of people who have heart problems and people who do not have any heart disease, in view of the relationship that these characteristics have with the final evaluation of whether the person has heart problems or not. During the data analysis, the variables age, glycemic rate, cholesterol, whether the person is a smoker or not, whether the person consumes alcohol or not, and whether the person is active in relation to physical activities were characteristics taken into account. This evaluation brought satisfactory results to be used in the models.

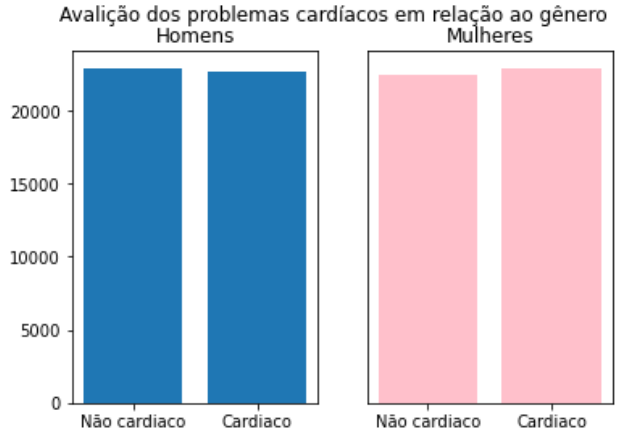
It is also important to highlight that normality tests were carried out on the variables that were introduced into the machine learning model. Normality distribution graphs were analyzed. And the tests found that all variables were positive for normality and consequently were used in this study.

Regarding the age of the people analyzed, it was noticed that people over the age of eighteen tend to have more heart problems, with little or almost no number of people under this age who have some type of

heart problem. On the other hand, in relation to those who are over 18 years old. All age groups have a certain degree of people with heart problems.

Gender is another element that generated an interesting analysis, because it was an analysis in which it was possible to see that the number of men and women who have and do not have heart problems are very similar. The majority of the men analyzed were not having heart problems, but there was a small difference between the two classifications. Women, on the other hand, had the majority with heart problems, but also with a very small difference.

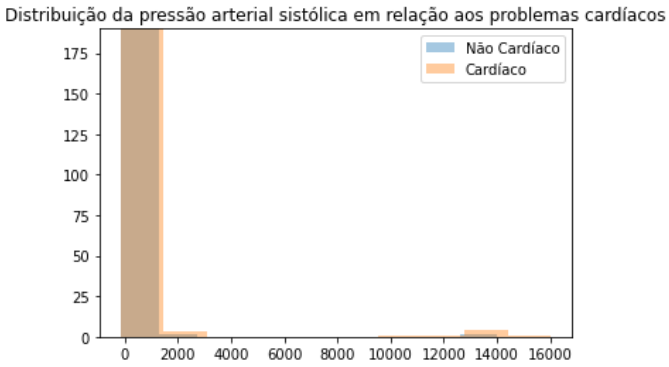
Figure 1. Assessment of heart problems in relation to gender



Source: Original survey results

Within systolic blood pressure, it is visible that a large part of cardiac and non-cardiac people have values concentrated on a certain side of the following graph, and a small part have values that are not concentrated in the standard places of the graph.

Figure 2. Distribution of systolic blood pressure in relation to heart problems

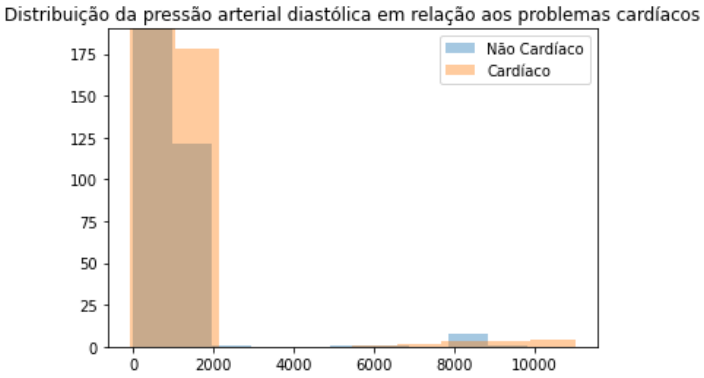


Source: Original survey results

In relation to diastolic blood pressure, a concentration of values is also perceived, but it is a slightly more distributed concentration, where it is possible to see more clearly the difference in the concentration

of people who have heart problems and those who do not have heart problems.

Figure 3. Diastolic Blood Pressure Distribution in Relation to Heart Problems

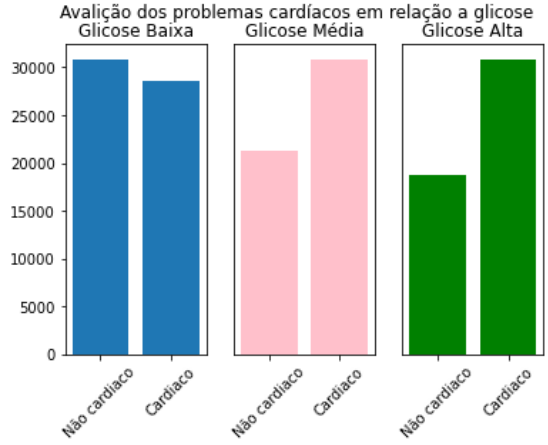


Source: Original survey results

Soon after, it was analyzed how much cholesterol influences heart problems, and cholesterol was divided into low, medium and high. Low cholesterol presented a sample containing more people without heart problems, but presented somewhat considerable compared to people who have heart problems as well. On the other hand, medium-level cholesterol, despite having a somewhat high number of people with heart problems, was the variable that had the most non-cardiac problems. The high cholesterol was mostly having heart problems.

The glucose of the people analyzed was also analyzed in three categories, low, medium and high. The low glucose classification had the majority having non-cardiac patients, but the difference was small. On the other hand, the average glucose was mostly for people with heart problems. And those who have high glucose have mostly heart problems.

Figure 4. Evaluation of heart problems in relation to glucose

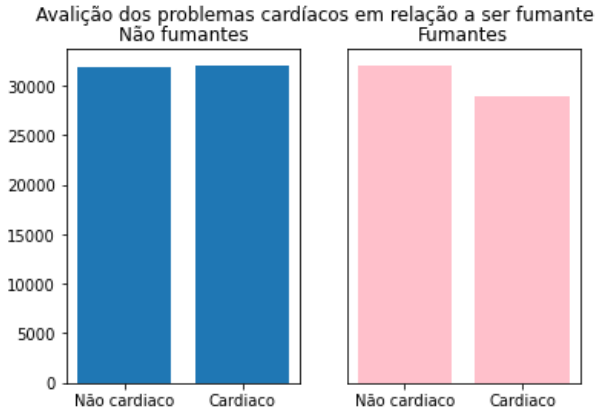


Source: Original survey results

The variable smoker was divided into smokers and nonsmokers. In relation to non-smokers, the ratio

between cardiac and non-cardiac patients was very similar, with cardiac patients being slightly higher. And the smokers had mostly as non-cardiac patients, but there were many people with heart problems as well.

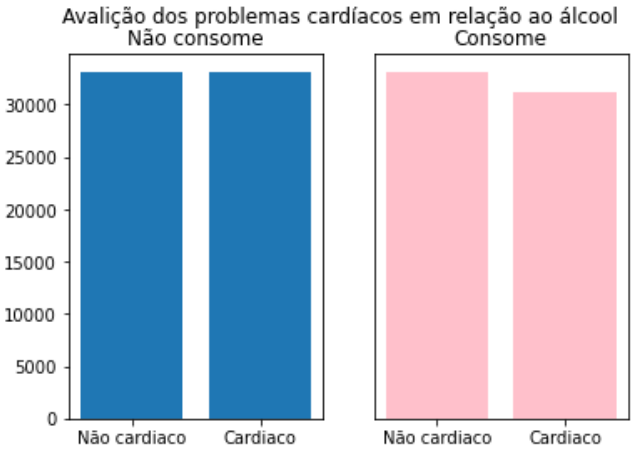
Figure 5. Assessment of heart problems in relation to being a smoker



Source: Original survey results

Alcohol consumption was also taken into account. Those who did not consume alcohol had almost identical results for having or not having heart problems, with non-cardiac ones slightly higher. On the other hand, those who consume alcohol had a non-cardiac condition.

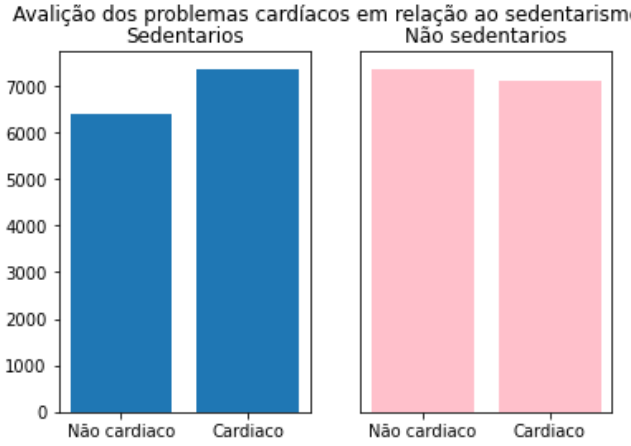
Figure 6. Evaluation of heart problems in relation to alcohol



Source: Original survey results

A sedentary lifestyle is also an important factor in identifying heart problems. The sedentary people were mostly people with heart problems, but with a high rate of not having heart problems. On the other hand, the non-sedentary had a higher rate of non-cardiac patients.

Figure 7. Evaluation of heart problems in relation to a sedentary lifestyle



Source: Original survey results

After the analyses were done, the "supervised machine learning" models were deployed, the model in which the first tests were carried out, was the "Random Forest", first the best parameters were found with "GridSearchCV". Then the data were divided into test and training, then the accuracy was 66%, the average sensitivity was 66.66%, the average accuracy was 66.60% and finally the confusion matrix had 4675 errors out of 14000 possible variables.

Table 2. Evaluating the performance of the Random Forest machine learning model

Metric	Percentage
Accuracy	66%
Confusion Matrix	4675
Medium sensitivity	66,66%
Medium accuracy	66,60%

Source: Original survey results

Soon after, the "XGBoost" model was tested, in which the tests followed the same pattern as the first test, with the definition of the parameters made by "GridSearchCV" and with the division into test and training data. It was the model that presented the best results, and for this, the calculation criterion used in the model was the mean squared error, which is a widely used metric to evaluate the quality of a regression model, in which the predictions made by the model are compared with the actual values of the data. Also used was "Log Loss", also known as "Logarithmic Loss" or "Cross-Entropy Loss", is a metric mainly used in binary classification problems or with many classes. With a maximum density of one, and the maximum control of the moment of the divisions of the model characteristic was set to automatic.

The accuracy was 72.97%, the mean sensitivity was 72.92%, the mean accuracy was 73.13%, and the confounding matrix showed 3784 errors also within the same 14000 test data.

Table 3. Evaluating the performance of the XGBoost machine learning model

Metric	Percentage
Accuracy	72,97%
Confusion Matrix	3784 errors
Medium sensitivity	72,92%
Medium accuracy	73,13%

Source: Original survey results

Finally, the "KNeighbors classifier" model was used, which is also known as knn, in this model the standard procedure of the other models was also maintained, with the definition of the best parameters and training and testing the data, with this an accuracy of 71.12%, an average sensitivity of 71.06%, the precision was 71.55% and the confusion matrix showed 4043 errors within the same 14000 test variables.

Table 4. Evaluating the performance of the KNeighbors classifier machine learning model

Metric	Percentage of success
Accuracy	71,12%
Confusion Matrix	4043 errors
Medium sensitivity	71,06%
Medium accuracy	71,55%

Source: Original survey results

After the whole process was carried out, noise was added to the data to prove the good performance of the models. The model that had the best performance in this regard was the "XGBoost" which managed to have a good performance with up to 70% noise included in the seat, where it maintained 60.42% accuracy. The "KNeighbors classifier" model also maintained a good performance, having a good performance with up to 60% noise included, with an accuracy of 60.36%. "Random Forest", on the other hand, had a good result until the inclusion of 50% of noise, obtaining an accuracy of 61.45%.

Table 5. Performance of all machine learning models used at work when subjected to noise.

Model	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**
Random Forest	68.37%	67,84%	64,67%	63,44%	61,45%	60,36%	58,67%
KNeighbors classifier	69,32%	67,40%	65,62%	63,51%	62,01%	60,36%	58,67%
XGBoost	71,82%	70,35%	68,18%	65,18%	64,20%	61,63%	60,42%

Source: Original survey results

4 FINAL THOUGHTS

The model that presented the best performance was the "XGBoost", even though the other models presented interesting results, the algorithm managed to present superior results. It was the one that obtained the best performance in relation to the noise test, thus being the model with the most efficient result and being considered the best model. It presented 72.97% accuracy compared to the original data and lower performance when faced with noise. In future studies, more robust models should be sought.

The work has limitations regarding the information found and consequently has a limited result due to the small amount of information, and another problem found was that the work was done on a computer that did not have adequate processing potential, which slowed down all processes. With more information that can be gathered over time, and with the use of hardware with greater potential, the results would be better.

It is concluded that this work provides the beginning of the study of the creation of a tool for the cardiac diagnosis of a given person. It is believed that the use of classifiers will allow the further development and study of the "supervised machine learning" tool, a parameter that will define whether the person analyzed is cardiac or non-cardiac in the future. This will be useful for entities looking for more efficient ways to identify heart problems.

REFERENCES

ALMEIDA, Filipe Guedes de Oliveira. Classificadores de polaridade de notícias utilizando ferramentas de machine learning: o caso da Vale SA. 2014.

Brasil. Ministério da Saúde. Secretaria de Atenção à Saúde. Departamento de Atenção Básica. Rastreamento / Ministério da Saúde, Secretaria de Atenção à Saúde, Departamento de Atenção Básica. – Brasília: Ministério da Saúde, 2010.

DALFOVO, Michael Samir; LANA, Rogério Adilson; SILVEIRA, Amélia. Métodos quantitativos e qualitativos: um resgate teórico. Revista interdisciplinar científica aplicada, 2008.

DE FIGUEIREDO, Carla Regina Bortolaz; CABRAL, Flávio Garcia. Inteligência artificial: machine learning na Administração Pública: Artificial intelligence: machine learning in public administration. International Journal of Digital Law, v. 1, n. 1, p. 79-96, 2020.

DINIZ, C. A. P. M. et al. Os efeitos do tabagismo como fator de risco para doenças cardiovasculares. Revista Eletrônica Saúde em Foco, 2011.

DOMINGOS, Pedro. O algoritmo mestre: como a busca pelo algoritmo de machine learning definitivo recriará nosso mundo. Novatec Editora, 2017.

CAMPOS, Roni CP; MIGUEL, Paulo A. Cauchick. Melhoria do processo de produção por meio da aplicação do Desdobramento da Função Qualidade. Sistemas & Gestão, v. 8, n. 2, p. 200-209, 2013.

CARSPECKEN, Phil Francis. Pesquisa qualitativa crítica: conceitos básicos. Educação & Realidade, v. 36, n. 2, p. 395-424, 2011.

CASTRO, L. C. V. et al. Nutrição e doenças cardiovasculares: os marcadores de risco em adultos. Revista de Nutrição, v. 17, n. Ver. Nutr., 2004 17 (3), p. 369- 377, jul. 2004.

GASPAR, João. Efeitos do sedentarismo a nível cardiovascular: a importância da actividade física na manutenção da saúde. 2004. Tese (Mestrado em Comunicação e Educação em Ciência)- Universidade de Aveiro, Aveiro, 2004.

HANS T, S.; VAN LEER E. M.; SEIDELL, J. C.; LEAN, M. E. Wiast circunference in the identification of cardiovascular risk factors: prevalence atudy in a random sample. BMJ. p. 311-1401, 1995.

JUNIOR, Bendev. Transformando códigos em sonhos: conselhos que gostaria de receber ao entrar na área da tecnologia. SEVEN publicações acadêmicas, 2022.

KAGGLE.2023.DATABASE. Risk Factors for Cardiovascular Heart Disease. Dsiponivél em:< https://www.kaggle.com/datasets/thedevastator/exploring-risk-factors-for-cardiovascular-diseas>. Acesso em: 20 mar. 2023.

LOPES, Lucas Pereira. Predição do preço do café Naturais Brasileiro por meio de modelos de statistical machine learning. Sigmae, v. 7, n. 1, p. 1-16, 2018.

MACIEL E. L. S. da R. et al. Efeito do estrogênio no risco cardiovascular: uma revisão integrativa. Revista Eletrônica Acervo Médico, v. 1, n. 1, p. e8527, 31 ago. 2021.

MAGALHÃES, Luiz Pereira de et al. Diretriz de Arritmias Cardíacas em Crianças e Cardiopatias Congênitas SOBRAC e DCC-CP. Arquivos Brasileiros de Cardiologia, v. 107, p. 1-58, 2016.

NUNES, S. O. B., CASTRO, M. R. P., CASTRO, M. S. A. Tabagismo, comorbidades e danos à saúde. In NUNES, SOV., and CASTRO, MRP., orgs. Tabagismo: Abordagem, prevenção e tratamento [online]. Londrina: EDUEL, 2011. pp. 17-38.

PIOVEZAN, Raphael Paulo Beal et al. Método de aprendizagem de máquina visando prever a direção de retornos de exchange traded funds (ETFs) com utilização de modelos de classificação e regressão. 2022.

PORTO, Celmo. Semiologia médica. 7ª edição. Rio de Janeiro: Guanabara Koogan, 2014.

RECH, C. R.; PETROSKI, E. L.; SILVA, R. C. R; SILVA, J.C.N;. Indicadores antropométricos de excesso de gordura corporal em mulheres. Rev. Bras. Med. Esporte, v.12, n.3, p.119-124, jun 2006.

SCHAAN, B. D., PORTAL, V. L. Fisiopatologia da Doença Cardiovascular no Diabetes. Revista da Sociedade de Cardiologia do Rio Grande do Sul, Ano XIII nº 03, 2004

STANGE, R. L.; GIANNINI, T.C.;SANTANA, F. S.; JOSE, J.; MAURO SARAIVA, A.: Evaluation of Adaptive Genetic Algorithm to Environmental Modeling of Peponapis and Curcubita. Revista IEEE América Latina, v. 9, p. 171-177, 2011

TOMÉ, Vívian Tostes. Utilização de machine learning para categorização dos gastos de bitcoin no Brasil. 2017. Tese de Doutorado.

XAVIER, H. T. et al.. V Diretriz Brasileira de Dislipidemias e Prevenção da Aterosclerose. Arquivos Brasileiros de Cardiologia, v. 101, n. Arq. Bras. Cardiol., 2013 101(4) suppl 1, p. 1–20, out. 2013