**O uso da *Machine Learning* com o algoritmo de Regressão Logística para realizar análise de dados de Doenças Cardiovasculares de um Banco de Dados Público da CC By 4.0**

**The use of Machine Learning with the Logistic Regression algorithm to perform data analysis of Cardiovascular Diseases from a CC By 4.0 Public Database**

**¹SILVA, Ricardo, Santos; ²LIMA, Edilson Carlos, Silva; ³QUEIROS, Jonathan de Araújo.**

¹Engenharia de Computação - Universidade Ceuma (UniCEUMA) Caixa Postal 15.064 - 91.501-970 - São Luís - MA - Brasil

²Engenharia de Computação - Universidade Ceuma (UniCEUMA) Caixa Postal 15.064 - 91.501-970 - São Luís - MA – Brasil

3Engenharia de Computação - Universidade Ceuma (UniCEUMA) Caixa Postal 15.064 - 91.501-970 - São Luís - MA – Brasil

{SILVA, Ricardo, ricardosantoscomp@gmail.com; LIMA, Edilson Carlos Silva, edilsonlima3@gmail.com; QUEIROS, Jonathan de Araújo, queirozjth@gmail.com}

**Resumo.** A doença cardiovascular (DCV) é uma adversidade que afeta o coração e os vasos sanguíneos, e estima-se que 17,9 milhões de pessoas morreram em 2019 (WHO, 2021). A crescente importância da análise de dados em informática médica despertou o interesse na geração de modelos analíticos em aprendizado de máquina (ML). A previsão de DCV é um desafio complexo e a classificação usando ML desempenhará um papel importante na previsão e investigação de doenças cardíacas para mitigar os efeitos cardíacos e evitar a mortes prematuras. O objetivo deste trabalho é analisar dados treinando com *Machine Learning* e o algoritmo de Regressão Logística para avaliar modelos e aprendizado profundo para prever diagnósticos de doenças cardiovasculares.

**Palavras-chaves**: *Machine Learning*; Algoritmo de Regressão Logística; Análise de Dados.

***Abstract.*** *Cardiovascular disease (CVD) is an adversity that affects the heart and blood vessels, and an estimated 17.9 million people died in 2019 (WHO, 2021). The growing importance of data analysis in medical informatics has sparked interest in generating analytical models in machine learning (ML). CVD prediction is challenging and a classification using ML will play an important research complex in predicting heart disease to mitigate cardiac deaths and prevent premature death. The aim of this work is healthy and to evaluate machine and deep learning models to predict diagnoses of cardiovascular diseases.*

***Key-words:*** *machine learning; cardiovascular diseases; data analysis.*

**1. Introdução**

A doença cardiovascular (DCV) é uma adversidade que afeta o coração e os vasos sanguíneos. Também conhecida como doença cardíaca, a incidência dessas doenças se deve principalmente a bloqueios que estreitam as artérias, impedindo o fluxo sanguíneo para o coração ou cérebro. Isso inclui várias doenças, como doença cardíaca coronária, doenças cerebrovasculares, doenças arteriais periféricas, doenças reumáticas e outras doenças. Em 2019, estima-se que as DCV tenham sido a causa de óbito de 17,9 milhões de pessoas, ou seja, 32% das mortes daquele ano (WHO, 2021).

Os fatores de risco para doenças cardiovasculares incluem fatores comportamentais, como tabagismo, dieta não saudável, uso nocivo de álcool e inatividade física, e fatores fisiológicos (metabólicos), incluindo pressão alta e níveis elevados de colesterol e glicose. Devido à enorme influência da informação disponível em várias formas, o papel da análise de dados em informática médica cresceu rapidamente na última década. Isso também levou a um crescente interesse na construção de modelos analíticos baseados em dados de *Machine Learning* (ML, em português, Aprendizado de Máquina) para computação médica.

A previsão de doenças cardiovasculares é um dos desafios mais complexos na análise de dados clínicos. No entanto, a classificação usando ML desempenha um papel importante na previsão de doenças cardíacas e revisão de dados para reduzir o impacto cardíaco e prevenir possível morte prematura. A doença cardiovascular pode ser amenizada através do diagnóstico precoce, reduzindo assim a mortalidade. Portanto, o uso de ML para identificar fatores de risco é uma importante abordagem potencial para a melhora da doença. Vários estudos têm desenvolvido modelos preditivos com o objetivo de prever doenças cardiovasculares. Por exemplo, em Al-Absi et al. (2021), os autores tentaram desenvolver um modelo na base de dados do Qatar Biobank (QBB) para identificar indivíduos saudáveis ​​e pacientes com doença cardiovascular, que pudesse detectar e analisar os principais fatores de risco relacionados à doença no Qatar. Outro exemplo é o estudo desenvolvido por Smigiel et al. (2021), que realizou de diferentes classes diagnosticais de DCV usando monitoramento de eletrocardiogramas (ECG).

Dado o histórico, o principal objetivo do trabalho atual é treinar e avaliar modelos de aprendizado de máquina e aprendizado profundo para prever o prognóstico de doenças cardiovasculares. Como contribuição, este trabalho destaca o uso de Regressão Logística para avaliação de séries temporais e comparação com outros dados já apurados.

O trabalho está organizado da seguinte forma: a seção 2 apresenta alguns trabalhos relacionados a esta proposta de trabalho. A seção 3 apresenta a metodologia e seus passos específicos para orientar este trabalho. Os resultados e análises obtidos são apresentados e discutidos na Seção 4. A Seção 5 conclui com comentários sobre as considerações finais identificadas neste trabalho.

**2. Fundamentação Teórica**

O diagnóstico precoce é um passo importante para atingir o objetivo de reduzir o impacto e as consequências das doenças cardiovasculares. Ghosh et al. (2021), semelhante a muitos estudos, tentou classificar as DCV usando ML. Um modelo com 99,05% de precisão foi projetado com bons resultados utilizando Random Forest Packing (RFBM) para classificar doenças cardíacas.

Al-Absi et al. (2021) construíram um modelo de ML para distinguir indivíduos saudáveis ​​de pacientes com doenças cardiovasculares para revelar uma lista de potenciais fatores de risco relacionados à doença. O estudo examinou uma série de medidas biomédicas, incluindo medidas comportamentais, representando as várias medidas biomédicas CVD do Qatar Biobank (QBB). Portanto, os autores listaram o CatBoost como o melhor modelo com 93% de precisão. Além dos fatores de risco conhecidos, como doença renal, função hepática, aterosclerose, etc. Outros achados foram a identificação de um novo conjunto de fatores de risco.

Cui et al. (2020) indicaram que níveis elevados de lipídios no sangue são um dos fatores de risco mais importantes para doenças cardiovasculares. Assim, com a previsão precoce de valores anormais de lipídios, a intervenção precoce é possível e reduz o risco de desenvolver dislipidemia, ou seja, colesterol ou lipídios no sangue anormalmente elevados. Portanto, o objetivo do estudo foi prever o risco de desenvolver dislipidemia em trabalhadores de siderúrgicas. Naquele estudo, a rede neural LSTM foi o melhor modelo, com acurácia[[1]](#footnote-1) superior a 95%.

Strodthoff et al. (2020) Uma análise comparativa do conjunto de dados de ECG clínicos, o PTB-XL. Como contribuição, este trabalho implementa e aplica diversos modelos modernos de *Deep Learning* (DL), além de fornecer um banco de dados para mitigar o problema da análise automatizada de ECG. Dentre todos os modelos avaliados, uma rede neural convolucional com arquiteturas *ResNet* e *Inception* tem o melhor desempenho na tarefa proposta.

Este trabalho propõe um método para construção de modelos de redes neurais, usando algoritmos de ML. Para o modelo de treinamento e teste foi utilizado um classificador baseado em Regressão Logística, visto que a primeira proposta feita utilizando Random Forest teve uma acurácia menor que a atual. O modelo de Regressão Logística, é empregado para modelar a relação entre uma variável dependente categórica e um conjunto de variáveis explanatórias.

Os trabalhos apresentados nesta seção são consistentes com o presente trabalho na execução de tarefas de aprendizado de máquina para prever eventos de saúde, particularmente doenças cardiovasculares. Portanto, assim como o presente trabalho, ambos propõem uma tarefa de classificação. O trabalho a ser apresentado de maneira mais detalhada é uma nova contribuição para o estudo de diferentes modelos ML utilizando a base de classificação Creative Commons Atribuição 4.0 Internacional (CC BY 4.0).

*2.1. Metodologias*

A primeira fase do desenvolvimento da pesquisa consiste na busca e seleção de um bom banco de dados no nicho da saúde, mais especificamente sobre doença cardiovascular estruturado para estudos de aprendizagem supervisionada. O processo de pesquisa de ML requer atividades que incluem pré-processamento e preparação de dados após a aquisição do data set. O objetivo da etapa de pré-processamento é usar técnicas de visualização para entender os dados e fazer insights sobre tendências e qualidade dos dados, além de fazer hipóteses e suposições na análise antes de ir para parte prática. Durante a preparação dos dados, a limpeza dos dados é realizada para otimizar a qualidade dos dados, que visa reduzir as amostras com valores ausentes, atributos com valores nulos e dados discrepantes.

Uma etapa muito importante antes que o modelo preditivo receba os dados é a visualização e análise dos atributos. Em relação aos atributos, é possível editar atributos deletando, convertendo e buscando novos já fornecidos pelo conjunto de dados, isso se faz necessário quando existe atributos sem valor especificado ou com valor atribuído de forma erronia no data set. Além disso, é possível converter os dados brutos em formatos mais adequados ao processo, o que não foi necessário nesta pesquisa.

*2.1. Materiais e Métodos*

Para atingir os objetivos desejados e testar as hipóteses experimentais formuladas anteriormente, foi necessário decidir quais as ferramentas que seriam usadas para tal processo. Visto que este é um estudo sobre ML, a linguagem de programação é o ponto mais importante a ser decidido, existem diversas linguagens que tem o poder de fazer algoritmos de ML, as mais usadas são Python, Linguagem R, Scala, Java, Julia, C, etc. No presente trabalho, a linguagem de programão utilizada foi Python.

Python é uma linguagem ágil, simples e objetiva, principalmente se tratante da área de análise de dados e apesar da sua simplicidade, ela é uma linguagem muito robusta que exerce seu papel sem apresentar nenhum problema, conta também com uma sintaxe fácil de entender. Python desenvolveu uma comunidade grande e ativa de processamento científico e análise de dados. Nos últimos dez anos, Python passou de uma linguagem de computação científica inovadora, ou para ser usada “por sua própria conta e risco”, para uma das linguagens mais importantes em ciência de dados e aprendizado de máquina (MCKINNEY, 2018)

Temos também o Jupyter Notebook, que é uma excelente opção para quem deseja criar códigos em Python, ele é se utiliza de uma interface web que permite prototipagem rápida e compartilhamento de projetos relacionados a dados e foi utilizado para fazer este modelo.

O modelo presente utilizará algoritmos que se enquadrem como ML tradicional de modelo preditivo de regressão logística que faz parte da biblioteca Scikit-Learn. Esta biblioteca se transformou no principal kit de ferramentas para aprendizado de máquina. Além desta se fará o uso também do Pandas e Numpy.

**3. Estudo de caso**

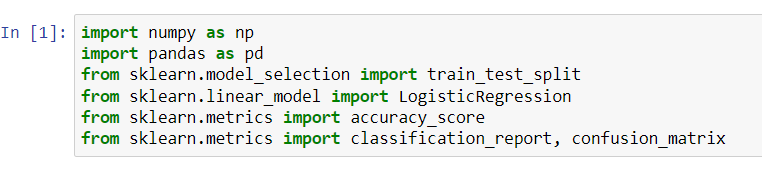
O trabalho presente utiliza dados do banco de dados CC BY 4.0 fornecido pelo UC *Irvine Machine Learning Repository* (UCI), um projeto com colaboração com Rexa na Universidade de Massachusetts Amherst e apoio financeiro da *National Science Foundation*, onde foi coletado dados de 4 banco de dados distintos (Hospital Universitário de Zurique - Suíça, Instituto Húngaro de Cardiologia – Hungria, Hospital Universitário de Basel - Suíça e VA Medical Center de Long Beach). Todos os dados coletados são reais e disponibilizados para pesquisas e estudos de casos.

*3.1. Dados e Ferramentas*

A primeira etapa do desenvolvimento da pesquisa foi analisar os atributos do banco de dados obtidos para o estudo, ao total são 303 linhas e 14 colunas compondo o Dataset, as linhas se referem a quantidade de pacientes analisados e as colunas ao total de atributos a serem analisados de cada paciente, a fim de gerar um modelo de treinamento e teste para a otimização de exames reais. Em cada paciente analisado, já vem o resultado se ele porta ou não doença cardiovascular e essa informação prévia será responsável para fazer o treinamento do modelo e calcular sua acurácia.

*3.2. Bibliotecas Utilizadas*

As bibliotecas Python são uma coleção de módulos e funções úteis que reduzem o uso de código de programação e sendo possível processar uma montanha de dados com elas de forma mais automatizadas. A figura 1, mostra quais bibliotecas foram utilizadas neste projeto.



**Figura 1:** Bibliotecas importadas

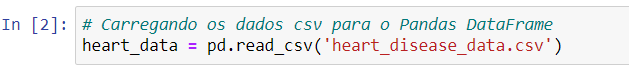
**Fonte:** Autoral, 2022.

O Numpy tem como função trabalhar operações numéricas, já o Pandas vai fazer a manipulação dos dados de forma geral, as bibliotecas baseadas em Sklearn vai servir para fazer a modelagem estatística, análise e mineração de dados, além de suporte ao aprendizado supervisionado e não supervisionado.

*3.3 Coleta e Processamento de dados*

Esses dados são devidamente preparados e passam por alguns processos de limpeza e ajuste de pré-processamento e seleção de variáveis ​​para que possam ser alimentados a um algoritmo de aprendizado de máquina que faz previsões e verifica novamente até que ponto o resultado está no valor correto. Ajustar os parâmetros usados ​​na previsão ajuda a obter um valor mais adequado.

A biblioteca Pandas terá um trabalho essencial na parte presente, onde será visualizado os dados em conjuntos de forma mais detalhadas. A figura 2, mostra o carregamento dos dados.

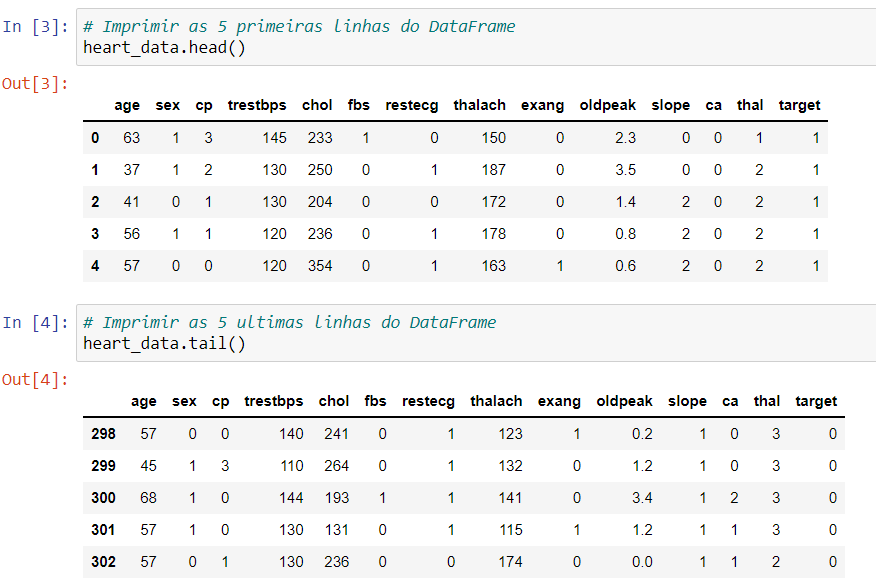


**Figura 2:** Carregamento dos dados

**Fonte:** Autoral, 2022.

Todos os dados estão armazenados em um arquivo de informações separadas por vírgulas (CSV) localizado na própria memoria do computador e precisa ser puxado através do código descrito na figura acima para, assim, fazer o tratamento e manipulação dos dados.

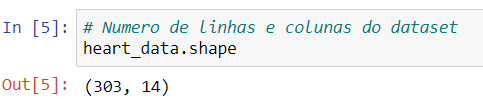
Agora vamos visualizar as 5 primeiras linhas e as 5 últimas do nosso Data Frame, conforme mostra a figura 3, para ter noção do modelo dos dados de cada atributo apresentado.



**Figura 3:** Visualização dos dados

**Fonte:** Autoral, 2022.

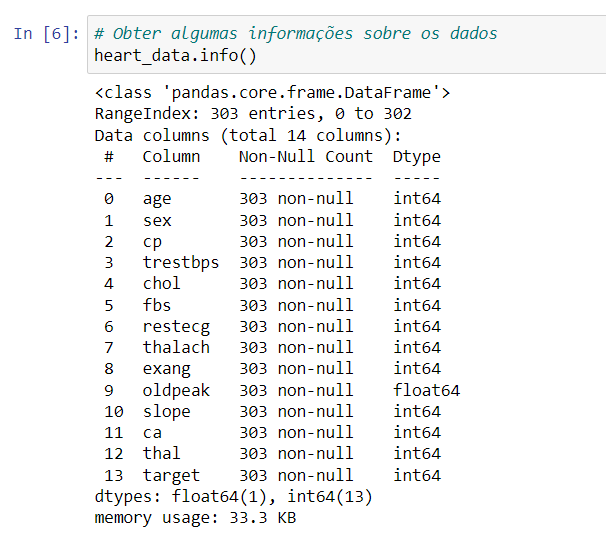
É notório que eles seguem um padrão, visto que todas essas informações foram coletadas de pessoas reais. Para uma melhor resolução da quantidade de dados presente no Data Frame foram utilizados o comando mostrado na figura 4.



**Figura 4:** Quantidade de dados

**Fonte:** Autoral, 2022.

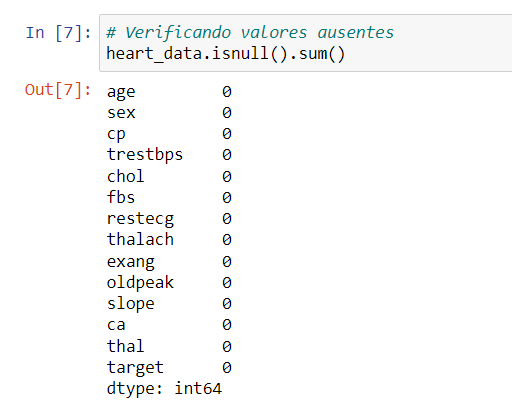
Vemos que, como descrito antes, temos dados de 303 pacientes, outras duas importantes funções apresentadas a seguir para a uma visualização mais profunda é saber quais os tipos de dados foram trabalhados, conforme mostra a figura 5.



**Figura 5:** tipo de dados

**Fonte:** Autoral, 2022.

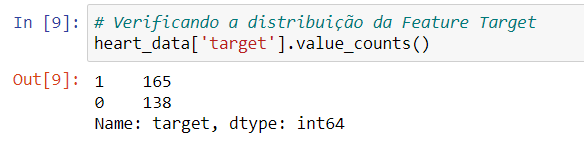
Portanto, é visível que entre os 14 atributos, 13 são do tipo inteiros (INT) e apenas 1 do tipo FLOAT, ou seja, números “quebrados”. É de suma importância também verificar se existe algum atributo com valores ausente, o comando para tal finalidade é mostrado na figura 6.



**Figura 6:** verificação dos atributos

**Fonte:** Autoral, 2022.

Dentre todos, nenhum atributo possui valor nulo, como visto na figura acima, isso é muito importante para o projeto, pois poderia acontecer algum erro na hora de treinar ou testar o modelo. Outro fator importante a ser apresentado, conforme é mostrado na figura 7, é a exibição de quantas pessoas têm ou não doenças cardiovasculares.



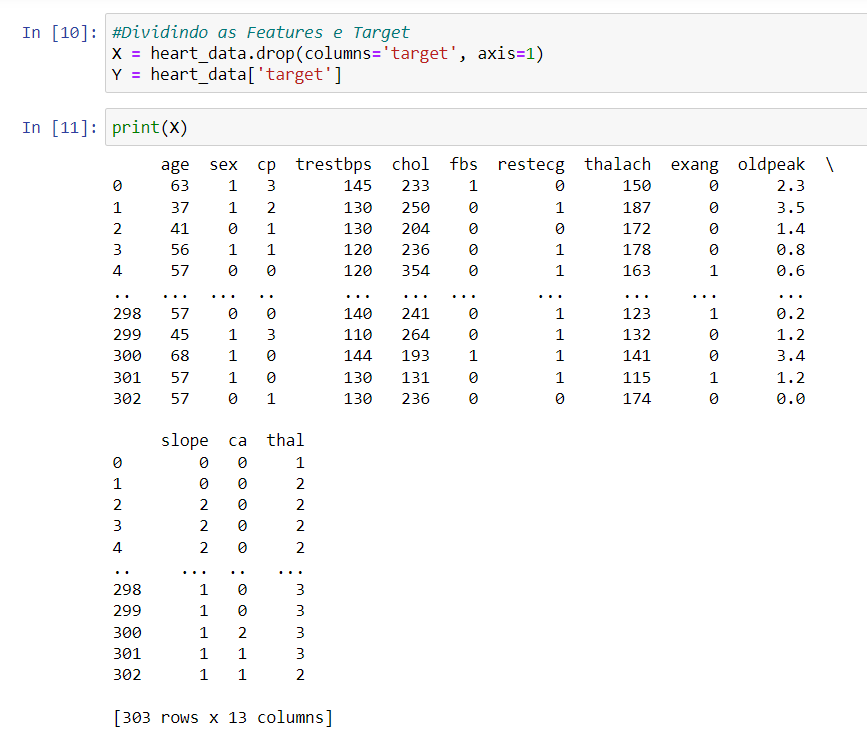
**Figura 7:** identificação dos pacientes

**Fonte:** Autoral, 2022.

Como mostrado na figura acima, o atributo chamado “target” é que mostra qual paciente tem o não doenças cardíacas, e no nosso Data Frame de estudo de caso, 165 pacientes têm e os outros 138 pacientes não têm, ou seja, o número 1 diz que o paciente porta doença e o número 0 que não porta.

*3.3. Dados de Treinamento e Dados de teste*

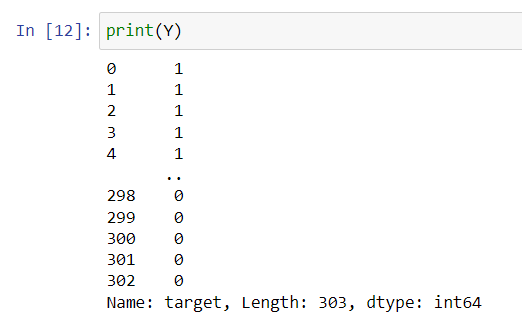
Uma das bases do aprendizado de máquina são os dados históricos, os algoritmos de aprendizado supervisionado de máquina precisam aprender e, para isso, quanto mais dados usar, melhor será o modelo, por isso acontece essa fase de treinar modelos antes de ir para os testes, conforme mostra a figura 8.



**Figura 8:** separação dos atributos

**Fonte:** Autoral, 2022.

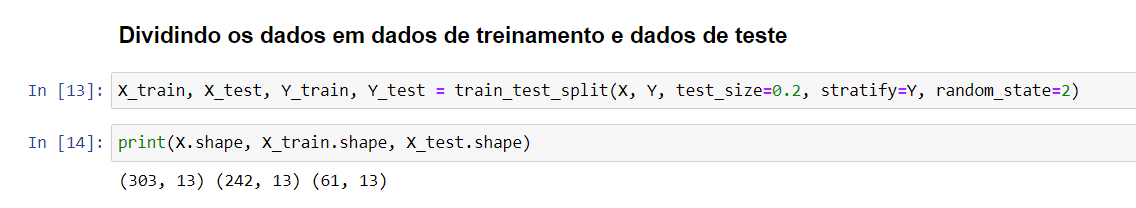
O primeiro passo antes de treinar o modelo é separar o atributo “target” de todos os outros, para assim, podermos saber se ele está avaliando corretamente, mas, então vemos a variável “X” recebendo todos os 13 atributos de treinamento e a variável “Y” recebe o atributo mencionado, segue a figura 9.



**Figura 9:** separação dos atributos 2

**Fonte:** Autoral, 2022.

Dados de treinamento são dados fornecidos a um algoritmo de aprendizado de máquina para construir um modelo. Esses dados geralmente representam cerca de 70% dos dados, por isso foram separados como mostra a figura 8 e figura 9. Segue a figura 10.



**Figura 10:** dados teste e dados treinamento

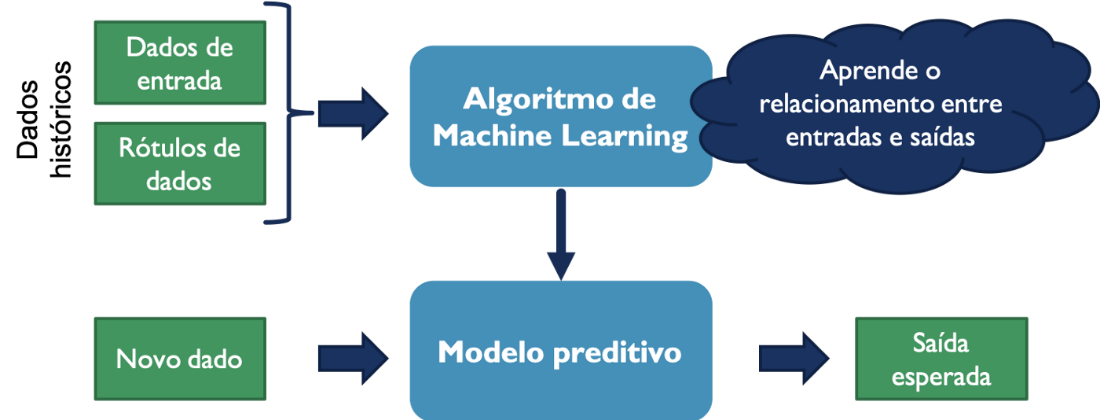
**Fonte:** Autoral, 2022.

Agora precisamos entender como funciona o a modelagem inicial do projeto, pois é aqui que o modelo será dividido para ser treinado, onde:

* X\_train - Isso inclui todas as suas variáveis ​​independentes, elas serão usadas para treinar o modelo, também como especificamos o test\_size = 0.2, esses 80% de seus dados completos será usado para treinar/ajustar o modelo e os 20% restantes serão usados para testar o modelo.
* X\_test - Esta é a parte restante dos 20% das variáveis ​​independentes dos dados que não serão usados ​​na fase de treinamento e serão usados ​​para fazer previsões para testar a precisão do modelo.
* Y\_train - Esta é sua variável dependente que precisa ser prevista pelo modelo, isso inclui rótulos de categoria contra suas variáveis ​​independentes, precisamos especificar nossa variável dependente enquanto treinamos/ajustamos o modelo.
* Y\_test - Esses dados têm rótulos de categoria para os dados de teste, esses rótulos serão usados ​​para testar a precisão entre as categorias reais e previstas.

*3.4. Algoritmo de Aprendizagem*

Os algoritmos de atividades preditivas podem ser classificados como sendo de regressão ou classificação, algoritmos de classificação tentam classificar objetos ou amostras com base nas propriedades observadas do operador, enquanto algoritmos de regressão tentam classificar a relação entre variáveis ​​usadas para prever valores. Segue o esquema (figura 11).



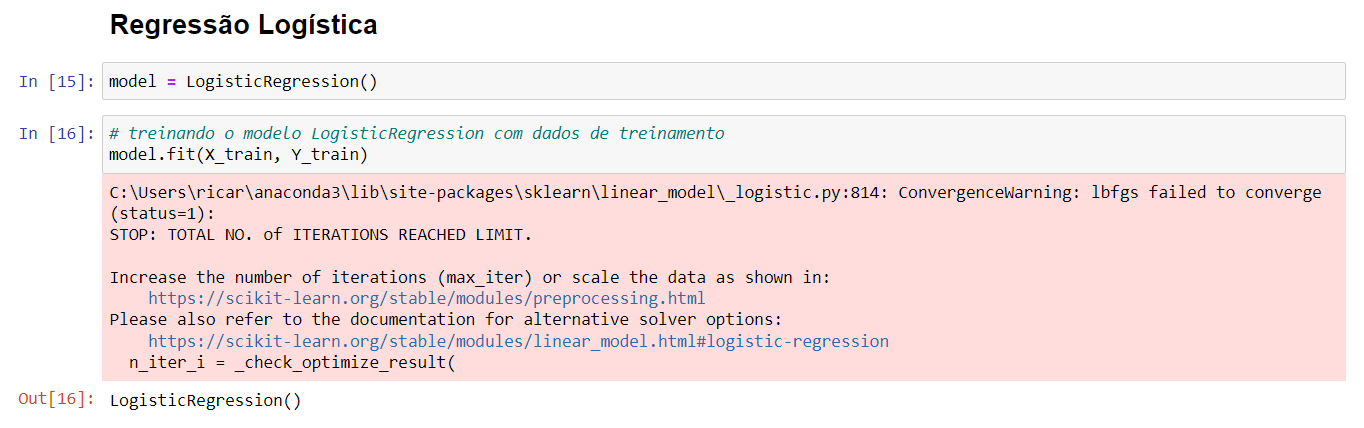
**Figura 11:** esquema do aprendizado

**Fonte:** Tatiana Escovedo, 2022.

Podemos concluir que este é um modelo supervisionado, pois o modelo de entrada é construído através de um conjunto de dados que foi recebido, e são apresentados como pares ordenados (entrada - saída). Dizemos que foram rotulados, porque conhecemos os detalhes de cada entrada e saída com antecedência.

Normalmente, os dados de entrada (rotulados) são divididos em dois conjuntos: um conjunto de treinamento, que é usado para construir o modelo, e um conjunto de teste, que também é chamado de literatura de conjunto de validação.

Só podemos treinar uma máquina usando algoritmos de aprendizado de máquina. Em suma, os algoritmos de aprendizado de máquina são conjuntos organizados de regras, comandos e instruções. Ao combinar algoritmos com dados, eles obtêm um resultado sólido, como mostra a figura 12.



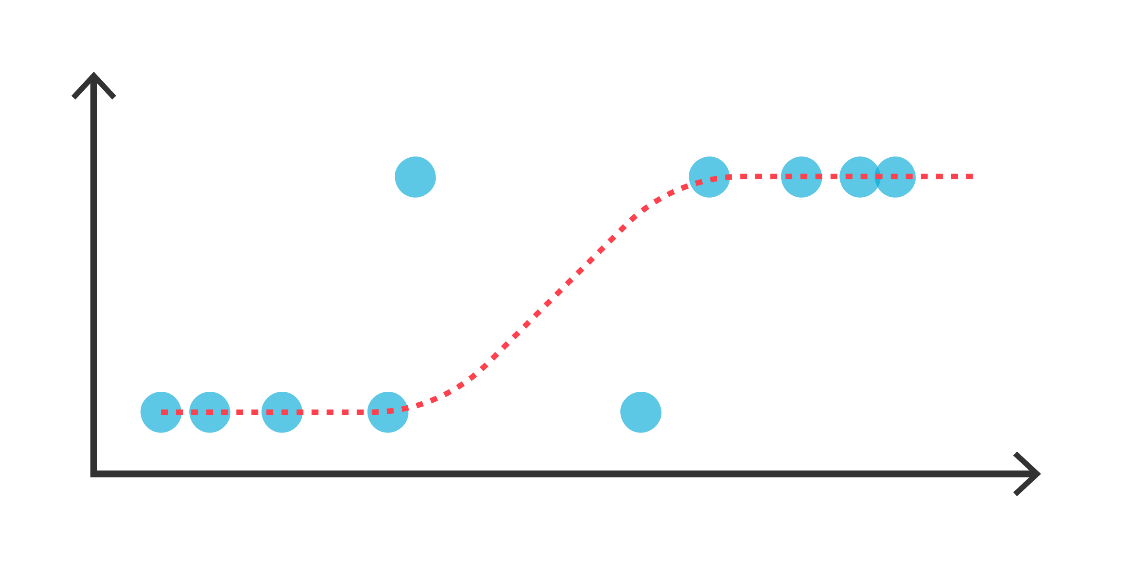
**Figura 12:** algoritmo usado

**Fonte:** Autoral, 2022.

Existem duas classificações principais de algoritmos de aprendizado de máquina:

* Aprendizado supervisionado: Os algoritmos de aprendizado de máquina supervisionados são aqueles que exigem entrada humana para manipular os dados. Nesse caso, já existe uma implementação do resultado correto (exemplos de rótulos), ou seja, o algoritmo já recebe determinados dados para realizar determinada operação.
* Aprendizado não supervisionado: Neste caso, as informações especificadas não são inseridas, portanto, o sistema não possui uma resposta correta. Assim, um algoritmo de aprendizado de máquina não supervisionado tem resultados variáveis ​​e imprevisíveis que geram novos padrões e filtros.

O algoritmo escolhido já citado para o presente projeto, que foi o Algoritmo de Regressão Logística, a figura 13, retrata como ele se apresenta.



**Figura 13:** curva da regressão logística

**Fonte:** TIBCO Software, 2022.

Este tipo de algoritmo de aprendizado de máquina trabalha com questões estatísticas e probabilísticas e lida com problemas de classificação. Para isso, a regressão logística calcula as variáveis e as classifica da melhor forma. Ao contrário da regressão linear, a regressão logística forma um padrão em S e apresenta valores entre 0 e 1. Na regressão logística a variável dependente tem apenas duas categorias. Em geral, a ocorrência do evento de interesse é codificada como “1” e a ausência como “0”. Lembrando que a codificação altera o sinal dos coeficientes e, portanto, sua interpretação substantiva.

A regressão logística difere de outras técnicas de mineração, principalmente pelo fato de sua variável dependente ser categórica, e mesmo quando ela não é dicotômica, é possível torná-la dicotômica, com a finalidade de aplicar esta técnica.

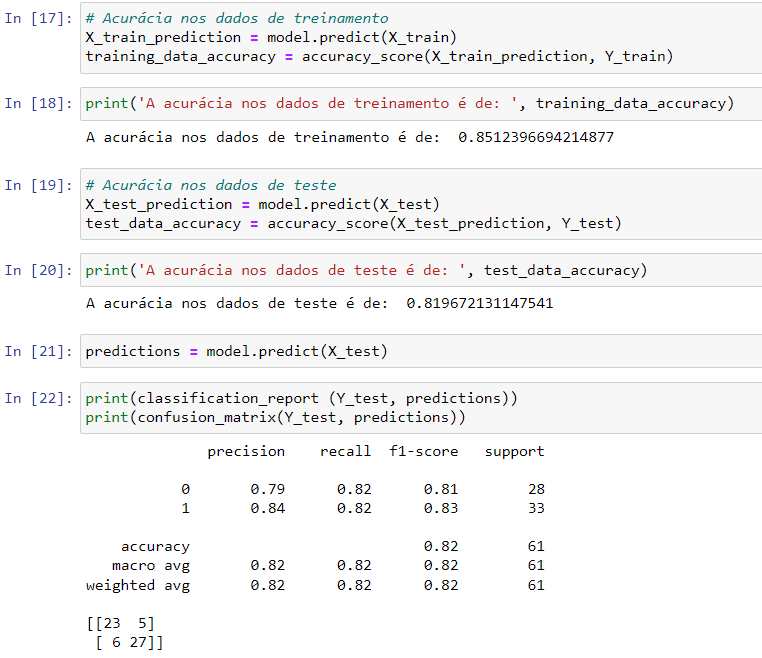
**4. Resultados e Discussões**

Como já foi citado anteriormente, o objetivo de um modelo de classificação de dados é fazer uma previsão com base em eventos passados. Para fazer isso, o modelo usa um conjunto de dados com entradas e atributos. Além disso, é necessário conhecer a saída esperada desse conjunto de dados.

Todas essas informações são usadas para treinar um modelo, que é usado para prever os resultados esperados de novos dados futuros. Ao treinar este modelo, foi usado um conjunto de dados (não aquele usado no treinamento) para testar o quão bem o modelo está correto. No entanto, não basta contar os sucessos do seu modelo, se foi bom ou não.

Nesta avaliação, diferentes métricas devem ser utilizadas dependendo do problema sob investigação. Mas antes de introduzir essas métricas, precisamos entender alguns conceitos de classificação binária: as classes que os dados previstos podem assumir.

Se falarmos especificamente sobre validação, então acurácia refere-se a quão próximos os resultados encontrados por meios automatizados ou soluções de inteligência artificial estão da realidade. Então, segue abaixo a figura 14, mostrando os resultados encontrados por meio do modelo criado.



**Figura 14:** acurácia do modelo

**Fonte:** Autoral, 2022.

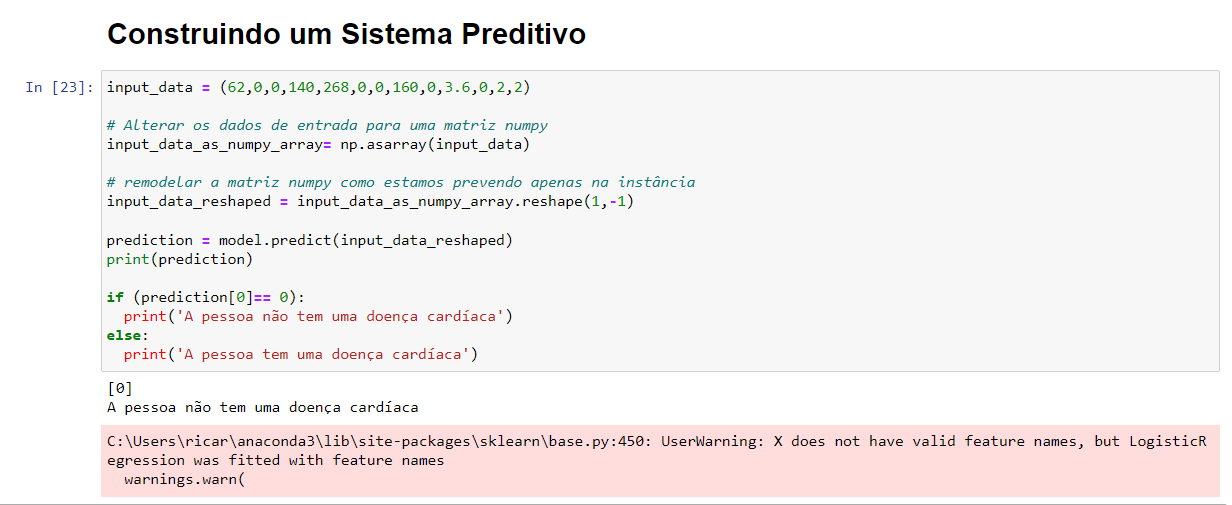
Vemos que a Acurácia do modelo de treinamento está equivalente a 85% e a acurácia do modelo de teste está a 82% arredondado, ou seja, a cada 10 indivíduos avaliados pelo modelo construído, cerca de 8 indivíduos foram diagnosticados correntemente, vale ressaltar que um projeto com acurácia acima de 70% já é considerado bom.

Nesta mesma figura, podemos observar algumas variáveis importantes e seus valores

* Acurácia indica o desempenho geral do modelo, quantas de todas as classificações o modelo classificou corretamente.
* Precisão: indica as classificações de classe positivas feitas pelo modelo estão corretas.
* Recall indica de todas as situações da categoria positiva, como valor esperado, quantas estão corretas
* F1 score é simplesmente uma maneira de rastrear apenas uma métrica em vez de duas (precisão e recall) em determinadas situações. É a média harmônica dos dois, que está muito mais próxima dos valores mais baixos do que a média aritmética simples. Ou seja, se a pontuação da F1 for baixa, indica que a precisão ou a recuperação são baixas.

Temos também a Matriz de Confusão mostrada no final da Figura 13, uma maneira simples de apresentar os resultados de um método de classificação de dados é com a chamada matriz de confusão, ela mostra o número de ocorrências que o programa teve de cada uma das quatro categorias.

Para finalizar o projeto apresentado até aqui, foi construído um sistema preditivo, como mostra a figura 15, com objetivo de receber os valores dos atributos de pessoas para serem avaliadas individualmente, como visto anteriormente, a acurácia está a cerca de 80%.



**Figura 15:** sistema preditivo para avaliação

**Fonte:** Autoral, 2022.

O primeiro passo é colocar todos os dados de cada atributo, precisa serem preenchidos todos na seguinte ordem:

1- Age (idade)

2- Sex (sexo)

0: feminino

1: masculino

3- Cp (tipo de dor torácica)

1: angina típica

2: angina atípica

3: dor não anginosa

4: assintomático

4- Trestbps (pressão arterial de repouso)

5- Chol (colesterol sérico em mg/dl)

6- Fbs (glicemia em jejum > 120 mg/dl)

0: falso

1: verdadeiro

7- Restecg (resultados eletrocardiográficos em repouso)

0: normal

1: com anormalidade da onda ST-T (inversões da onda T e/ou elevação ou depressão do segmento ST > 0,05 mV)

2: mostrando provável ou definitiva hipertrofia ventricular esquerda pelos critérios de Romhilt-Estes

8- Thalach (frequência cardíaca máxima alcançada)

9- Exang (angina induzida por exercício)

0: não

1: sim

10- Oldpeak (depressão ST induzida pelo exercício em relação ao repouso)

11- Slope (a inclinação do pico do exercício segmento ST)

1: ascendente

2: plano

3: downsloping

12- Ca (número de vasos principais (0-3) coloridos por fluorosopia)

13- Thal (3 = normal; 6 = defeito corrigido; 7 = defeito reversível)

O indivíduo usado como teste para o sistema preditivo foi diagnosticado sem doença cardiovascular. Lembrando que o algoritmo avalia de acordo com os dados apresentados a ele, os dados que foram submetidos anteriormente, tiveram finalidade apenas de criar o modelo te treinamento e fazer testes.

**5. Conclusão**

As doenças cardiovasculares são as principais causas de mortes no mundo. Neste trabalho, o modelo de aprendizado de máquina para classificação multe classes de diagnósticos de doenças cardiovasculares foi proposto para apoiar a tomada de decisão, promover a pesquisa de materiais e fornecer direcionamento para diagnósticos de reconhecimento e, assim, ajudar na indicação de tratamento.

Foi escolhido um modelo de algoritmo de *Machine Learning* para construir o modelo de aprendizado de máquina de classificação, que é: Regressão Logística; um método usado para prever variáveis ​​de resultado categóricas a partir de variáveis ​​preditoras categóricas ou contínuas. As variáveis ​​de resultado são muitas vezes referidas como dependentes e os preditores como independentes. A previsão de uma categoria específica pode ser chamada de classificação.

Em trabalhos futuros, pretende-se avaliar diferentes estratégias de balanceamento, explorar a otimização de mais hiper parâmetros para os modelos e realizar um pré-processamento que defina melhor as classes de amostra com base na probabilidade de diagnósticos implícitos nos dados.

Também pretende-se criar outros modelos de predição e classificação baseado em novos tipos de algoritmos de machine Learning, tais como Regressão Linear, Arvores de Classificação e regressão, Naive Bayes, KNN, Random Forest, etc. Visando buscar a melhor acurácia possível.

**Referências**

A.M. Aguilera, M. Escabias, M.J. Valderrama, Using principal components for estimating logistic regression with high-dimensional multicollinear data, Computational Statistics & Data Analysis, 55 (2006), 1905-1924.

Al-Absi, H. R. H., Refaee, M. A., Rehman, A. U., Islam, M. T., Belhaouari, S. B., andAlam, T. (2021). Risk factors and comorbidities associated to cardiovascular diseasein qatar: A machine learning based case-control study. IEEE Access, 9:29929–29941

MCKINNEY, Wes. Python for Data Analysis: tratamento de dados com pandas, numpy e ipython. O'Reilly, 2018. ISBN: 9781491957660.

Chollet, F. (2017). Deep Learning with Python. Manning.

Cui, S., Li, C., Chen, Z., Wang, J., and Yuan, J. (2020). Research on risk predictionof dyslipidemia in steel workers based on recurrent neural network and lstm neuralnetwork.IEEE Access, 8:34153–34161

LIAN, Suyun; GUAN, Lixin; PENG, Zhongzheng; ZENG, Gui; LI, Mengshan; XU, Yin. Retrieval of leaf chlorophyll content in Gannan navel orange based on fusing hyperspectral vegetation indices using machine learning algorithms. Rural Science, [S.L.], v. 53, n. 3, p. 6-10, out. 2022. FapUNIFESP (SciELO)

Mohan, S., Thirumalai, C., and Srivastava, G. (2019). Effective heart disease predictionusing hybrid machine learning techniques.IEEE Access, 7:81542–81554.

Raschka, S. and Mirjalili, V. (2019). Python Machine Learning. Packt Publishing, Bir-mingham, UK, 3 edition

Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. M. J. M., Ignatious, E., Shul-tana, S., Beeravolu, A. R., and De Boer, F. (2021). Efficient prediction of cardiovas-cular disease using machine learning algorithms with relief and lasso feature selectiontechniques.IEEE Access, 9:19304–19326.

Ravi, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., and Yang,G.-Z. (2017). Deep learning for health informatics. IEEE Journal of Biomedical andHealth Informatics, 21(1):4–21.

Smigiel, S., Pałczy ́nski, K., and Ledzi ́nski, D. (2021). Ecg signal classification using deeplearning techniques based on the ptb-xl dataset.Entropy, 23(9).

Strodthoff, N., Wagner, P., Schaeffter, T., and Samek, W. (2020). Deep learning for ECGanalysis: Benchmarks and insights from PTB-XL.CoRR, abs/2004.13701.

Wagner, P., Strodthoff, N., Bousseljot, R., Kreiseler, D., Lunze, F. I., Samek, W., andSchaeffter, T. (2020). Ptb-xl, a large publicly available electrocardiography dataset.Scientific Data, 7.

WHO, W. H. O. (2021).Cardiovascular diseases (cvds).Disponivel em: <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)>. Acesso em: 29 jun. 2022

World Health Organization. (‎2018)‎. Technical package for cardiovascular disease management in primary health care: healthy-lifestyle counselling. World Health Organization.

1. Acurácia é uma espécie de soma entre exatidão e precisão [↑](#footnote-ref-1)