**Aprendizado de Máquina 1**

**Trabalho 1**

Classificação do dataset “Personal Key Indicators of Heart Disease”

**Professor:** Prof. Dr. Murilo Naldi

**Grupo:**

Augusto dos Santos Gomes Vaz - 800268

Gabriel Lourenço de Paula Graton - 800432

Pietro Minghini Moralles - 792238

Vitor Matheus da Silva - 800260

São Carlos

29 de Julho de 2022

**Índice**

[**1. Resumo 1**](#_832br3m4isrq)

[**2. Introdução 2**](#_t0v0edy1v3cs)

[**3. Escolha do Conjunto de Dados 2**](#_87jrq0tq9adg)

[3.1. Justificativa da escolha do conjunto de dados 3](#_y93jti7yq61f)

[**4. Estudo do Conjunto de Dados 3**](#_oya1cp13rgiz)

[4.1. Entendendo o conjunto de dados 3](#_yb4egcma6bd9)

[4.2. Solução de problemas de atributos 4](#_dx36nz51uuaz)

[4.5. Métricas 8](#_4e18c84dg47n)

[4.6. Visualização dos dados e justificativa 8](#_4n67m3fpvq7n)

[**5. Classificação 18**](#_dtiyfxggt19m)

[5.1. Justificativa da escolha dos métodos de classificação 19](#_icad9zpwocub)

[5.2. Ajuste de parâmetros e hiperparâmetros 19](#_90q1s6w2fyly)

[5.3. Comparação dos resultados 20](#_edhl8kt8vmc4)

[**6. Conclusão 21**](#_2hh34asye060)

[**7. Bibliografia 23**](#_t7jbzabw7val)

# Resumo

A priori, o grupo verificou quais datasets tinham sido escolhidos para não haver nenhum tipo de conflito. Depois dessa decisão, optamos por procurar datasets que seriam desafiadores, mas que também nos agregasse em conhecimento.

Ademais, ao ter certeza sobre qual conjunto de dados escolheríamos, já verificamos o que deveria ser feito para facilitar o uso das ferramentas e possibilitar análises classificatórias mais resilientes, demonstrando quais abordagens optar e quais descartar.

Por fim, foram realizadas classificações com algoritmos diferentes visando obter as classificações das instâncias no conjunto de teste e verificar a comparação de desempenho entre os classificadores e definir qual foi o melhor e o que nos levou a tal resultado. Mas também, com abordagens diferentes, como iremos demonstrar.

# Introdução

Algoritmos de classificação são uma das grandes áreas em que o aprendizado de máquina se destaca, apresentando resultados surpreendentes para as mais diversas aplicações que precisam classificar objetos com base em seus atributos.

Para que esses resultados positivos apareçam, no entanto, é necessário o uso de datasets com dados sólidos e de um tratamento correto destes, além da escolha de algoritmos eficazes e eficientes.

Dessa forma, propomos neste trabalho um estudo acerca do dataset "Personal Key Indicator of Heart Disease", tratando certas inconsistências que apareceram, utilizando de abordagens diferentes para verificar as melhores estratégias. Principalmente, com base nas métricas fornecidas em sala de aula, como por exemplo acurácia, mas também matrizes de confusão, precisão e recall.

A acurácia é calculada pela divisão do número de previsões corretas pelo número total de previsões. Precisão refere-se a capacidade de um algoritmo de aprendizado de máquina supervisionado de classificar corretamente as instâncias positivas. Já o recall diz respeito à capacidade do modelo de identificar corretamente os exemplos positivos. Por fim, a matriz de confusão funciona como base para o cálculo das medidas de desempenho citadas acima, pois compara as classificações feitas pelo modelo com as classes reais dos exemplos.

# Escolha do Conjunto de Dados

Para a realização do trabalho, o grupo optou por escolher o conjunto de dados “**Personal Key Indicator of Heart Disease”** do repositório de dados Kaggle. A base de dados pode ser encontrada no link: [Dataset](https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease).

Além disso, todas as manipulações e aplicações realizadas sobre o banco de dados podem ser observadas no seguinte link do Colab: [Colab](https://colab.research.google.com/drive/1bp08j2Gxwl1KuNKSej1wXFWsWXL37GmD?usp=sharing).

O conjunto de dados foi criado pela CDC (Centers for Disease Control and Prevention), uma agência federal dos Estados Unidos responsável por proteger a saúde pública e promover a segurança através do controle e prevenção de doenças, lesões e deficiências. Os dados foram obtidos por meio de pesquisas telefônicas anuais para coletar dados sobre o estado da saúde dos Estados Unidos. O conjunto de dados contém mais de 300 mil objetos, compostos por 20 atributos que serão definidos mais adiante.

O nosso objetivo é predizer, a partir de uma série de informações sobre a saúde da pessoa, se ela possivelmente terá ou não terá algum tipo de doença cardíaca.

## Justificativa da escolha do conjunto de dados

Doenças cardíacas são um tema muito relevante, seja pela sua gravidade seja por sua altíssima ocorrência. No contexto brasileiro, entre 300 e 400 mil indivíduos sofrem de Infarto Agudo do Miocárdio (IAM) por ano, ocorrendo óbito em mais de 70% dos casos, segundo o [Ministério da Saúde.](https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/i/infarto) E [segundo este mesmo orgão](https://www.gov.br/saude/pt-br/composicao/saps/ecv), cerca de 30% de todas as mortes no território nacional advém de doenças cardiovasculares.

Em adição a isso, esse tipo de doença nem sempre apresenta sintomas claros, e ser capaz de diagnosticá-la antecipadamente é extremamente importante para iniciar tratamentos preventivos.

Por fim, quando nos deparamos com esse dataset, ficamos intrigados com ele, por suas dificuldades, mas também pelas peculiaridades. Desta forma, com ele podemos executar diversas funções das diversas bibliotecas do python, principalmente da sklearn. Abaixo, estão listados as motivações do grupo para lidar com esse dataset

* Presença de outliers e inválidos (BMI absurdos e SleepTime muito Grandes);
* Atributos Categóricos, Binários e Ordinais;
* Valores duplicados;
* Interesse pela área da medicina;
* Por ser um dataset de medicina, não há valores claros para obter a classe, é mais complexo;

# Estudo do Conjunto de Dados

## Entendendo o conjunto de dados

Como mencionado anteriormente, o nosso conjunto de dados contém cerca de 320 mil objetos, compostos por um total de 18 atributos, listados abaixo:

* **Heart Disease:** participantes da pesquisa que alegaram ter doença arterial coronariana ou infarto do miocárdio. Esse atributo é o que precisamos prever, com base nos demais atributos e é um atributo do tipo String binário, podendo ser apenas Yes, em caso afirmativo, ou No, em caso negativo.
* **BMI:** índice de massa corpórea. Representado por números de pontos flutuantes.
* **Smoking:** indica se a pessoa já fumou pelo menos 100 cigarros em toda sua vida. É um atributo do tipo String binário, podendo ser apenas Yes, em caso afirmativo, ou No, em caso negativo.
* **AlcohoolDrinking:** indica se a pessoa bebe constantemente (adultos do sexo masculino que tomam mais de 14 drinques em uma semana ou adultos do sexo feminino que tomam mais de 7 drinques em uma semana). É um atributo do tipo String binário, podendo ser apenas Yes, em caso afirmativo, ou No, em caso negativo.
* **Stroke:** indica se a pessoa já teve um acidente vascular cerebral (AVC). É um atributo do tipo String binário, podendo ser apenas Yes, em caso afirmativo, ou No, em caso negativo.
* **PhysicalHealth:** indica, no intervalo dos últimos 30 dias passados, durante quantos dias a saúde física da pessoa não estava boa. Consiste de um intervalo numérico que varia de 0 a 30.
* **MentalHealth:** indica, no intervalo dos últimos 30 dias passados, durante quantos dias a saúde mental da pessoa não estava boa. Consiste de um intervalo numérico que varia de 0 a 30.
* **DiffWalking:** indica se a pessoa tem dificuldades para andar ou subir escadas. É um atributo do tipo String binário, podendo ser apenas Yes, em caso afirmativo, ou No, em caso negativo.
* **Sex:** indica o sexo biológico da pessoa. É um atributo do tipo String binário, podendo ser apenas Male, em caso do sexo biológico masculino, ou Femalie, em caso do sexo biológico feminino.
* **AgeCategory:** indica 14 intervalo de idades, iniciando em “18-24”, aumentando de 5 em 5, até o intervalo “80 or older”.
* **Race:** indica a etnia racial da pessoal, sendo 6 às consideradas na pesquisa: “White”, “Black”, “Asian”, “American Indian/Alaskan Native”, “Hispanic” ou “Other”.
* **Diabetic:** indica se a pessoa já teve diabetes. É um atributo do tipo String binário, podendo ser apenas Yes, em caso afirmativo, ou No, em caso negativo.
* **PhysicalActivity:** indica se as pessoas realizaram atividades físicas nos últimos 30 dias, além do seu trabalho. É um atributo do tipo String binário, podendo ser apenas Yes, em caso afirmativo, ou No, em caso negativo.
* **GenHealth:** indica como a pessoa entrevistada classifica a sua saúde de maneira geral, podendo ser: “Poor”, “Fair”, “Good”, “Very good” e “Excellent”.
* **SleepTime:** indica, em média, quantas horas de sono a pessoa tem em um período de 24 horas. Consiste em valores numéricos que variam de 1 até 24.
* **Asthma:** indica se a pessoa tem/já teve asma. É um atributo do tipo String binário, podendo ser apenas Yes, em caso afirmativo, ou No, em caso negativo.
* **KidneyDisease:** indica se a pessoa já teve/tem alguma doença no rim, desconsiderando pedra no rim, incontinência urinária ou infecção urinária. É um atributo do tipo String binário, podendo ser apenas Yes, em caso afirmativo, ou No, em caso negativo.
* **SkinCancer:** indica se a pessoa já teve/tem câncer de pele. É um atributo do tipo String binário, podendo ser apenas Yes, em caso afirmativo, ou No, em caso negativo.

## Solução de problemas de atributos

Feita a leitura básica inicial dos dados, partimos para uma análise mais aprofundada dos mesmos. Em um primeiro passo, utilizamos a função describe, para ter uma noção mais aprofundada dos valores numéricos presentes na nossa base de dados. 

***Figura 1*** *: Tabela de distribuição dos dados antes do tratamento*

*A partir disso, foram realizadas diversas modificações nos atributos para que o conjunto de dados se tornasse mais verossímil com a realidade.*

1. Remoção de outliers e inválidos

Em um primeiro momento, os dados que mais chamaram a atenção foram os atributos **SleepTime** e **BMI.** No primeiro, percebemos que o valor máximo era de 24, o que quer dizer que em um intervalo de 24 horas, o indivíduo dorme 24 horas. Estes valores foram considerados absurdos pelo grupo, dado que estudos apontam que um ser humano saudável deveria dormir no máximo 13 horas. Por isso, optamos por remover aqueles que dormem mais de 20 horas por dia.

Quanto ao atributo **BMI** (Índice de Massa Corpórea), temos valores máximos de 94,85. Estes valores também foram considerados absurdos pelo grupo, uma vez que valores acima de 35 para o IMC são considerados obesos. Apenas para servir de exemplo, para que se tenha um IMC de 94,85, é necessário ter 1,80 de altura e 300kgs. Por conta disso, decidimos adotar um valor máximo para IMC de 70.

Após essas mudanças, rodamos o comando describe novamente, e ficamos com os valores de acordo com a imagem abaixo.

***Figura 2*** *: Tabela de distribuição dos dados após o tratamento*

*tabela feita através da função .describe() no dataset após a retirada*

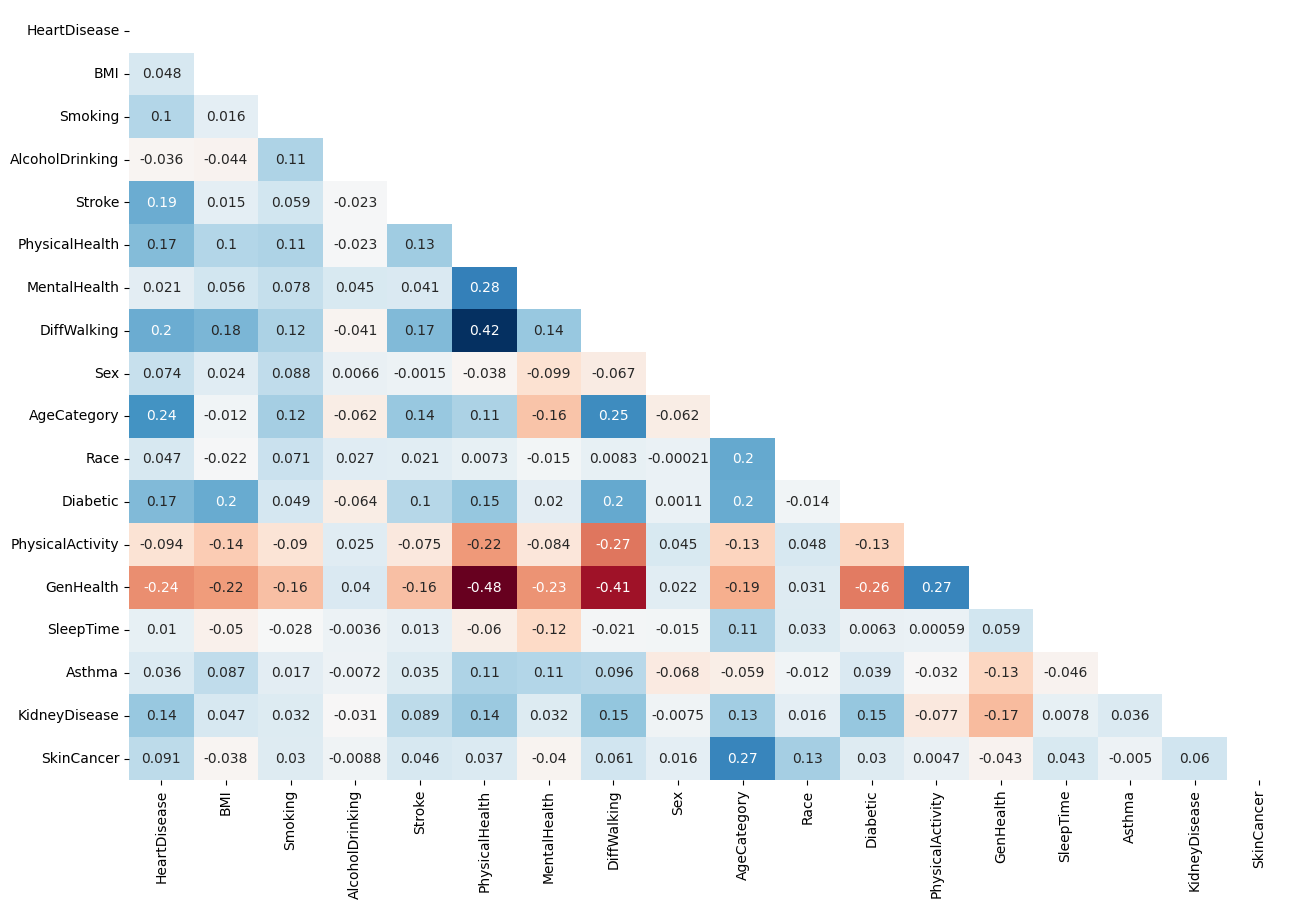
1. Dados duplicados

Por fim, resolvemos checar a existência de dados duplicados, pois os mesmos não acrescentariam positivamente para a aplicação do modelo. Dito isto, encontramos 18.078 objetos dos quais removemos todas as duplicatas, mantendo apenas os objetos originais.

* 1. **Correlação entre os atributos**

Feito isso, aplicamos a medida de correlação para descrever a relação entre duas variáveis. Essa medida estatística pode variar de -1 a 1, em que os extremos indicam uma correlação perfeita, enquanto valores próximos a zero indicam uma correlação fraca ou inexistente. Uma correlação de 1 significa uma correlação positiva perfeita, onde as variáveis têm uma relação linear crescente, ou seja, a medida que uma variável aumenta, a outra também aumenta na mesma proporção.

Já uma correlação de -1 significa que as variáveis têm uma relação linear decrescente, que indica que a medida que uma variável aumenta, a outra diminui na mesma proporção. Dito isto, segue abaixo as correlações obtidas entre nossos atributos e a classe de predição, Heart Disease.



***Figura 3*** *: Correlação entre os atributos*

De acordo com a imagem, podemos obter os atributos que mais influenciam na categorização da nossa classe, a HeartDisease. Em ordem decrescente, temos AgeCategory, com uma correlação de 0.24, seguida de DiffWalking, com uma correlação de 0.2, seguida de Stroke, com 0.19, logo após PhysicalHealth e Diabetic. É interessante ressaltar a covariância negativa com GenHealth, de -0.24, o que parece ser um pouco contra-intuitivo, mas que no nosso conjunto de dados faz sentido.

* 1. **Abordagens**

1. **Abordagem** 1

A priori, após o tratamento inicial, decidimos gerar abordagens diferentes, nesta primeira optamos por fazer algo simples, apenas trocar os atributos sem se importar com a verossimilhança dos dados. Dessa forma, foi realizado a troca dos atributos como “yes” e “no” por binários 1 e 0, além de - nas diabetes - onde há 2 tipos de “yes” e “no” foram todos convertidos para ter somente um tipo. Isso foi feito com a função .replace()

Mas também, intervalo de idades como 0 até 13. Ademais, ao abordar a qualidade de vida, foi colocado em ordem de qualidade, sendo a maior “excelente” como 4. Contudo, ao lidar com raça, colocamos a que possui maior porcentagem de dados com o maior float, ou seja, atribuindo uma ordem ao atributo categórico.Para isso, foi utilizado a função .map()

1. **Abordagem 2**

Na segunda abordagem, foi utilizado o codificador OneHotEncoder, optamos por ela devido a presença de atributos categóricos, essa estratégia é muito boa para esses tipos. Dessa forma, passamos eles apenas nesses tipos de dados, e nos outros usamos .replace() para “yes” e “no” assim como na primeira abordagem.

Como resultado, temos 38 colunas, o que irá aumentar e muito tempo computacional.

Ademais, a estratégia foi vista em aula, o que fez com que surgisse a vontade de aprender mais sobre ela.

1. Abordagem 3

Além disso, foi feita uma terceira abordagem com LabelEncoder com o intuito de aprender mais sobre esta. A estratégia foi usada em todo o dataset, para usá-la foi extremamente compacta, contudo a mesma acaba aumentando também o custo computacional.

## Métricas

A fim de medir a eficiência dos resultados serão usados certas médias, as quais serão:

1. **Acurácia**

Medida vista em aula para comentar sobre o desempenho do modelo em classificar corretamente os exemplos dos conjuntos de dados.

1. **Tempo**

Medida padrão apenas para verificar o quanto demora cada algoritmo.

1. **Sensibilidade** - [Definição](https://www.fleury.com.br/noticias/precisao-diagnostica)

Nesse quesito, a sensibilidade é a mesma coisa do que o Recall, iremos usá-la com esse nome uma vez que exploraremos o contexto de medicina do modelo, onde queremos saber se os classificadores conseguem acertar bem os verdadeiros positivos. Dessa forma, caso seja diagnosticado o exame de forma positivo, qual a chance de ser realmente um positivo.

1. **Especificidade** - [Definição](https://www.fleury.com.br/noticias/precisao-diagnostica)

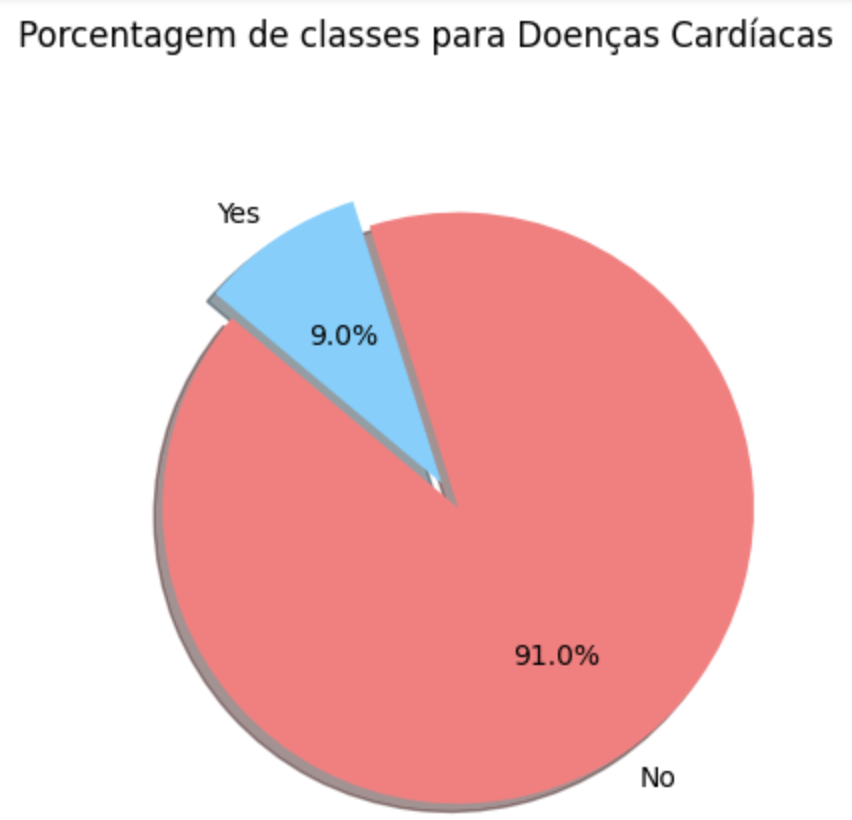
Assim como a Sensibilidade, este vai ser um termo medicinal para verificarmos a chance de acontecer um verdadeiro negativo. Ou seja, caso seja diagnosticado o exame de forma negativa, qual a chance de ser realmente um negativo.

1. **Matriz de Confusão**

Outra métrica vista em aula que vai ser muito importante para avaliarmos os modelos

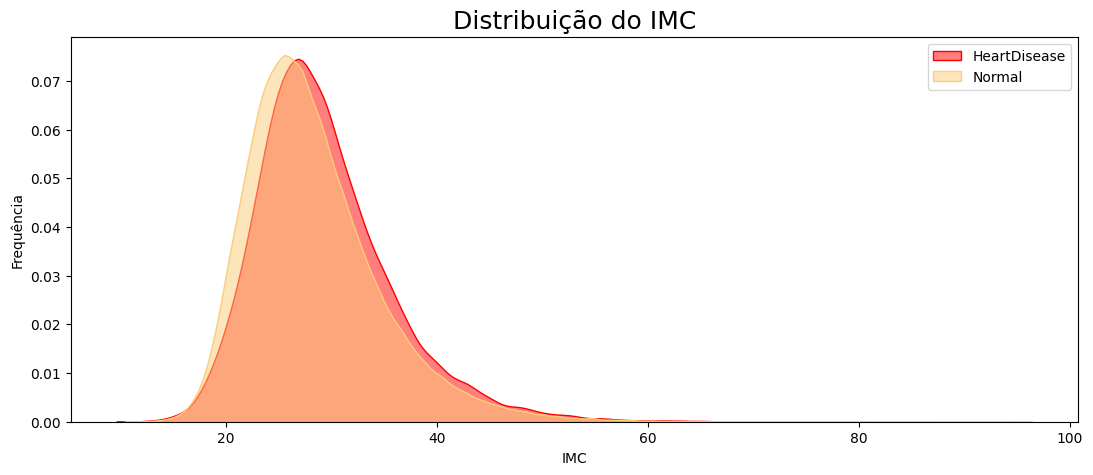
## Visualização dos dados e justificativa

Para compreender melhor o conjunto de dados, foram usados gráficos para representar a distribuição de cada atributo.



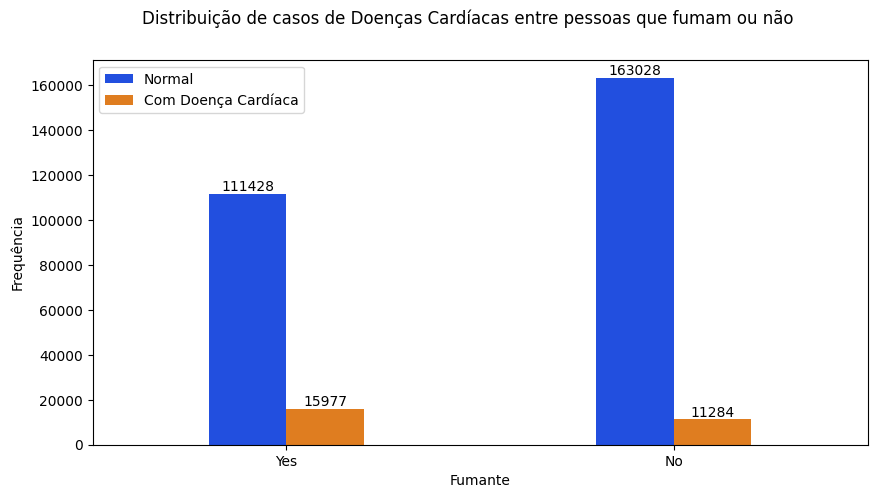
***Figura 4*** *: Distribuição de pessoas com doença cardíaco e sem doença cardíaco no dataset*

Neste primeiro gráfico, podemos observar que a quantidade de classes a serem previstas estão desbalanceadas, sendo que apenas 9% dos objetos são da classe que possui doenças cardíacas, enquanto 91% fazem parte do grupo que não possuem doenças cardíacas.



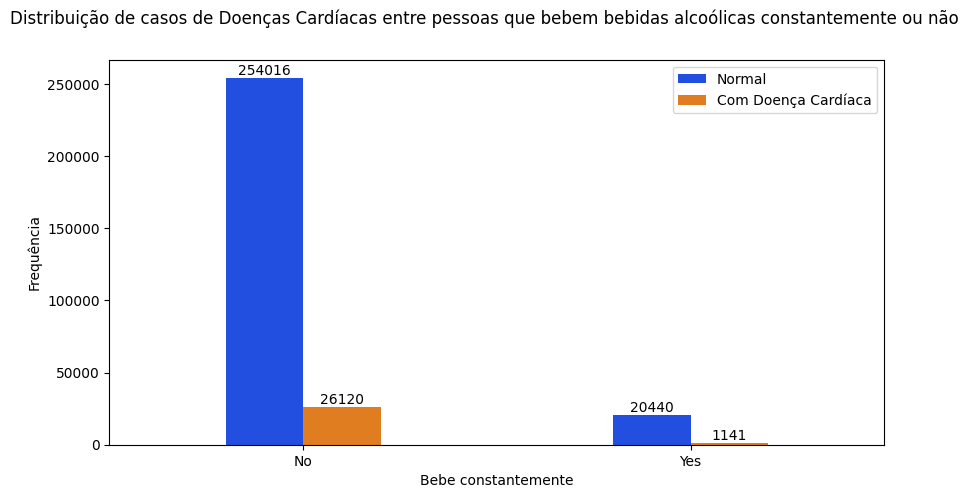
***Figura 5*** *: Distribuição dos casos de doença cardíaca com base no IMC*

Nesta figura, temos a distribuição entre casos com pessoas que possuem ou não doenças cardíacas relacionados com o índice de massa corpórea. É interessante notar que a maioria das pessoas que responderam a pesquisa se encontram em um IMC na faixa de 25 a 30, além de que, quanto maior for o IMC, maior é a proporção de pessoas que possuem doenças cardíacas.



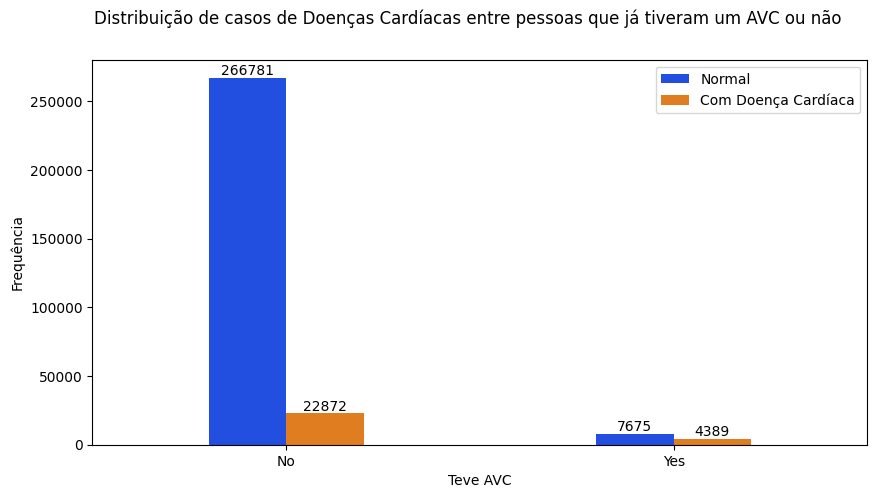
***Figura 6*** *: Distribuição dos casos de doença entre fumantes e não fumantes*

Neste gráfico temos a proporção de pessoas que fumam ou não e possuem ou não algum tipo de doença cardíaca. É interessante observar que, embora a quantidade de pessoas que fumam seja bem menor do que as que não fumam, a porcentagem de pessoas que fumam e têm doenças cardíacas é consideravelmente maior do que as pessoas que não fumam, sendo de 14,3% para a primeira e 6,91% para a segunda.



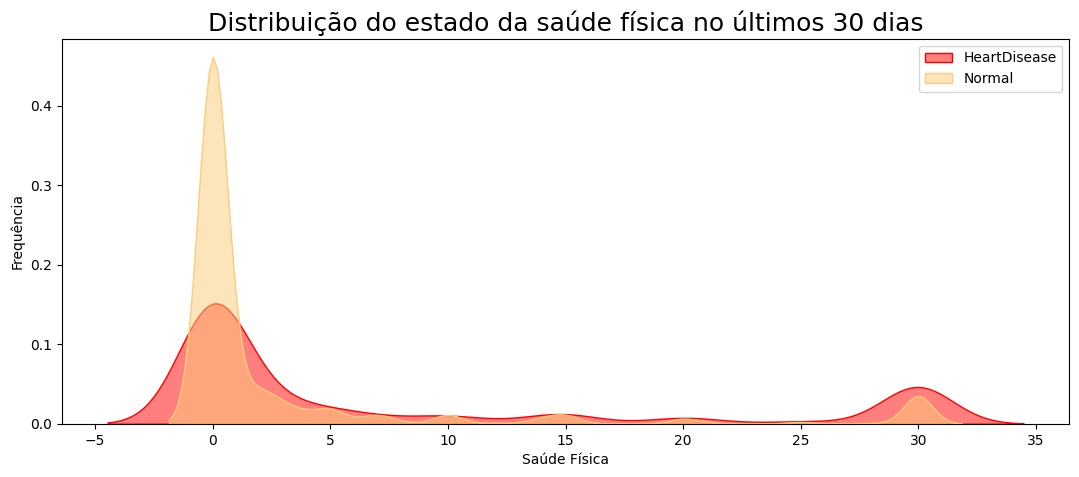
***Figura 7*** *: Distribuição dos casos de doença cardíaca entre pessoas que bebem constantemente e não bebem constantemente*

Já neste gráfico, temos a distribuição entre pessoas que bebem constantemente e a relação com doenças cardíacas. Esse atributo nos revela que a pessoa ser alcoólatra, não necessariamente influencia diretamente em ter doenças cardíacas, vide a pouca quantidade de pessoas que bebem e têm doenças cardíacas.



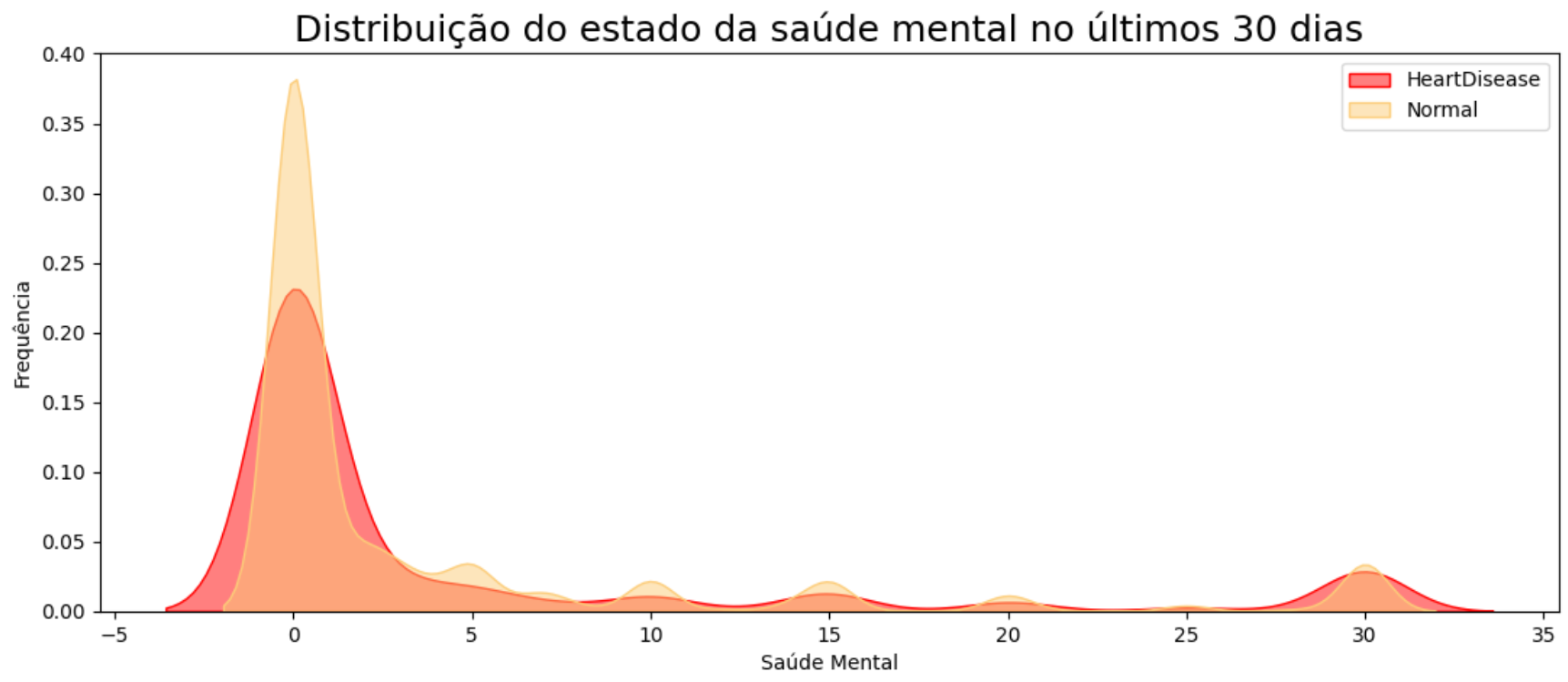
***Figura 8*** *: Distribuição dos casos de doença cardíaca entre pessoas que tiveram AVC constantemente e que não tiveram AVC*

No gráfico que relaciona acidente vascular cerebral com doenças cardíacas, é possível concluir que o AVC contribui de maneira significativa para ter doenças cardíacas. Isso ocorre pois mesmo que os números de pessoas que já tiveram AVC seja muito menor quando comparado com os que não tiveram, a porcentagem dessas pessoas que possuem doenças cardíacas é cerca de 57,1%.



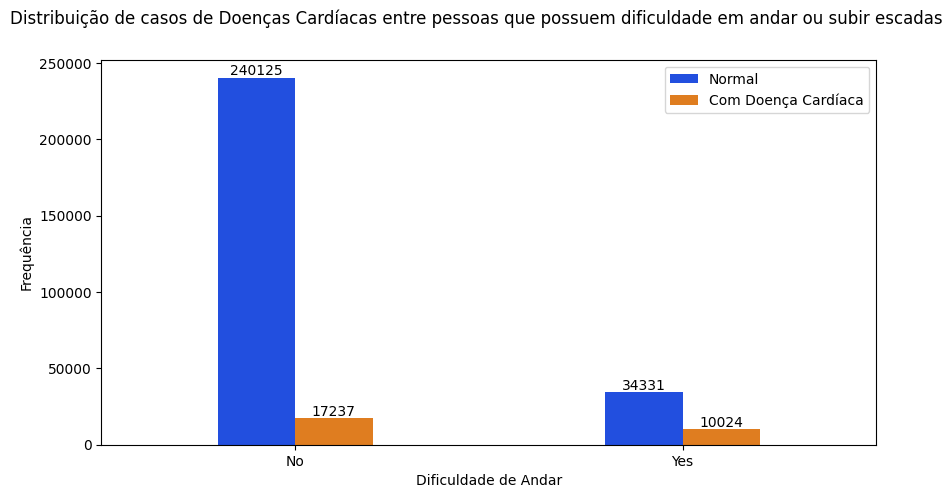
***Figura 9*** *: Distribuição dos casos de doença cardíaca com base em sua classificação de saúde física*

Neste gráfico, podemos observar que, em pessoas nas quais a saúde física foi pior no intervalo de 30 dias, a quantidade de doenças cardíacas tende a crescer, medida essa considerada intuitiva para categorização de doenças cardíacas.



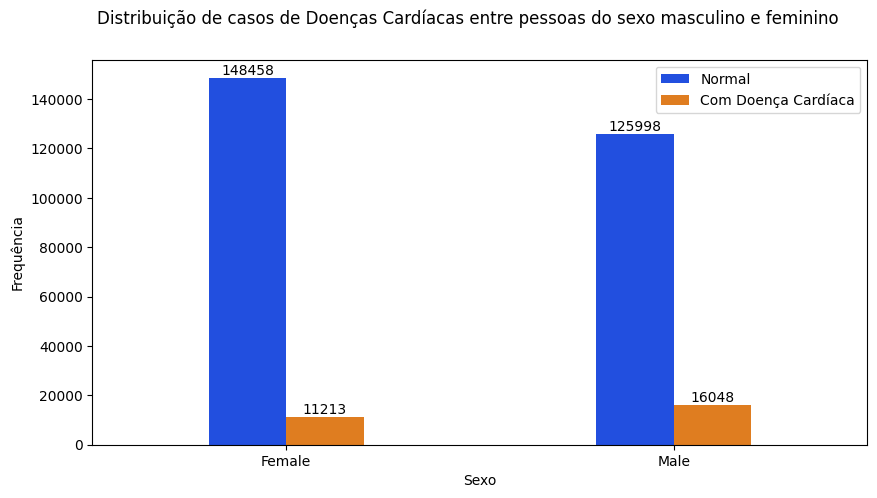
***Figura 10****: Distribuição dos casos de doença cardíaca de acordo com a classificação de saúde mental*

Em contrapartida, quando se trata de saúde mental, podemos perceber que independente do estado da saúde mental da pessoa nos últimos 30 dias, esse atributo não possui uma influência expressiva para a categorização de doenças cardíacas.



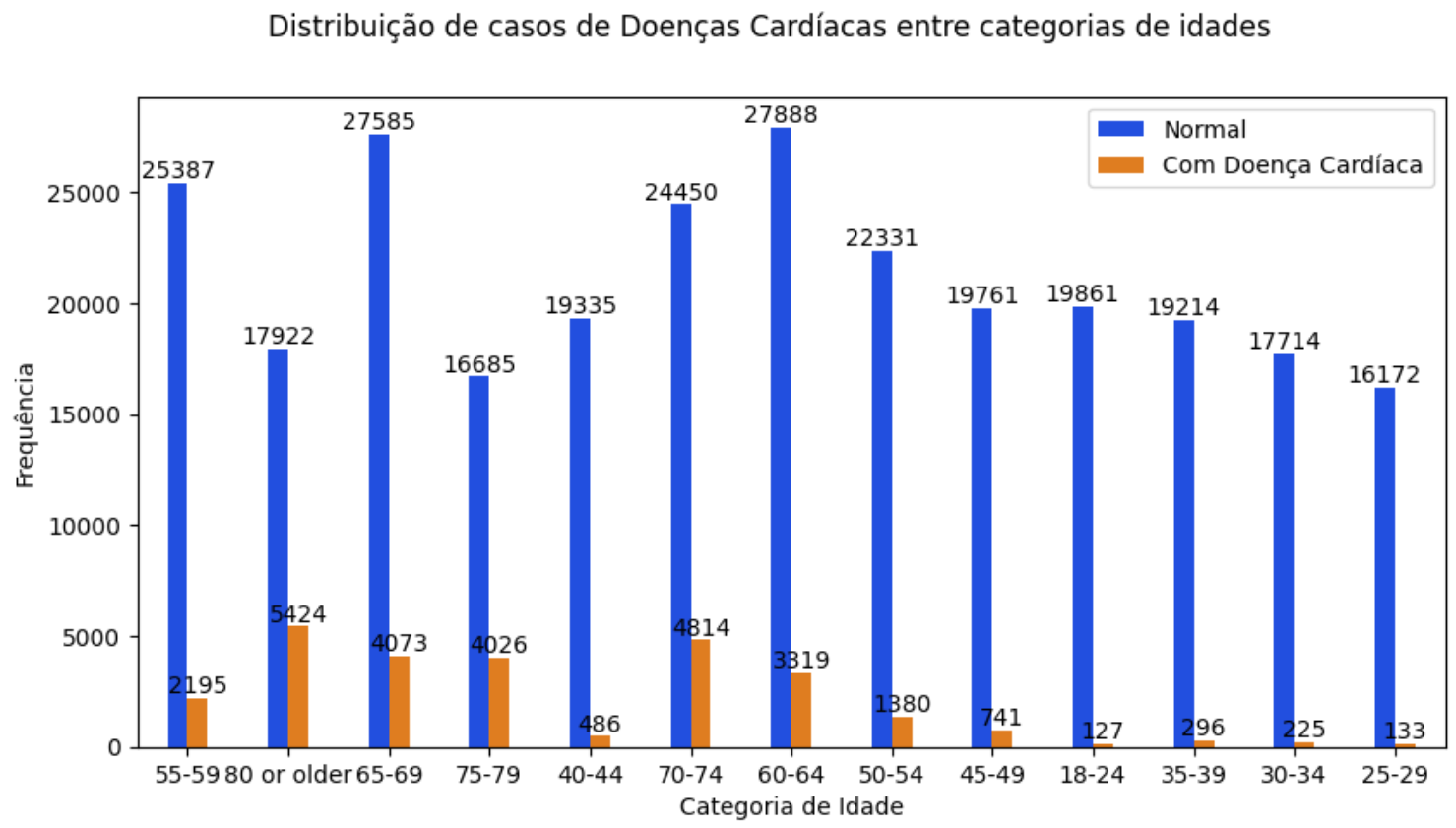
***Figura 11****: Distribuição dos casos dos casos de doença entre pessoas com dificuldade ou não de locomoção*

Neste gráfico, temos a relação entre dificuldade para andar com doenças cardíacas. Novamente, a quantidade de pessoas que não possuem dificuldade para andar é muito maior do que a quantidade das que possuem dificuldade. Porém, a porcentagem de pessoas que possuem dificuldade para andar e têm doença cardíaca é muito maior quando comparada com aquelas que não tem dificuldade para andar e possuem doenças cardíacas, sendo 29,21% para o primeiro e 7,17% para o segundo.



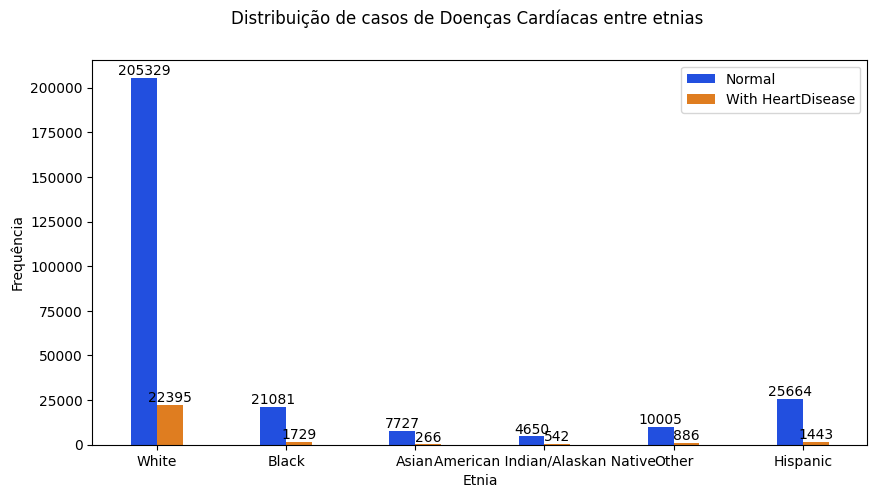
***Figura 12****: Distribuição dos casos de doença cardíaca de acordo com o sexo*

Aqui temos a diferenciação entre sexo biólogico. Neste caso, temos mais mulheres respondentes da pesquisa do que homens, porém, a probabilidade de um homem possuir doenças cardíacas quando comparado com uma mulher é cerca de 5 vezes maior.



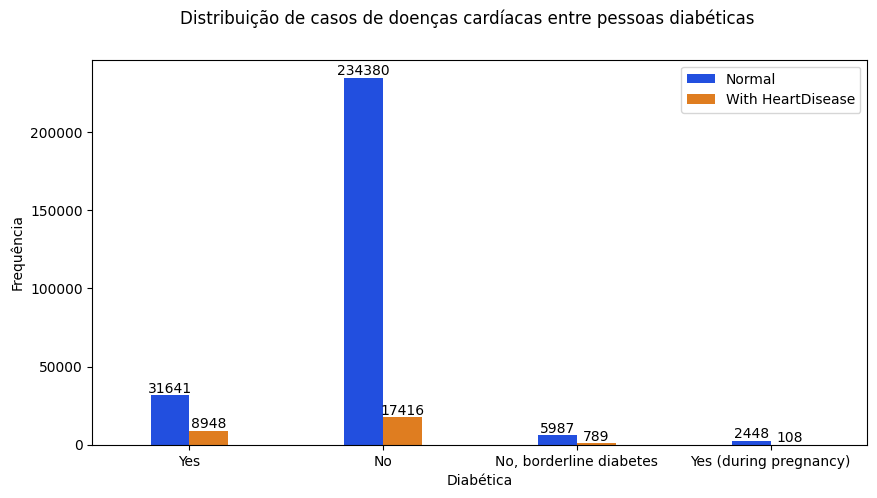
***Figura 13****: Distribuição dos casos de doença cardíaca de acordo com o intervalo de idade*

Neste gráfico temos a distribuição de pessoas que possuem doenças cardíacas entre os intervalos de idade identificados na pesquisa, sendo o primeiro intervalo de 18-24 anos e o último intervalo de 80 or older. Estes intervalos variam de 5 em 5 anos e são interessantes para informar que quanto mais elevada a idade da pessoa, maior o número de casos de doenças cardíacas.



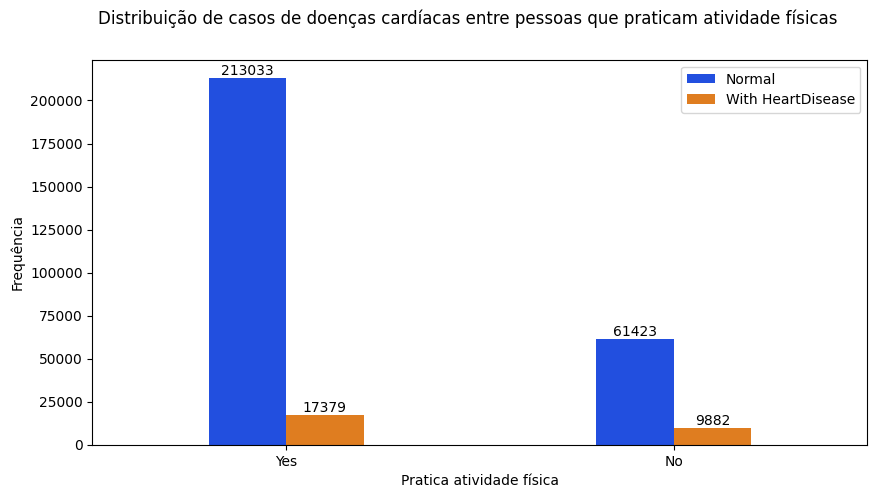
***Figura 14****: Distribuição dos casos de doença cardíaca de acordo com a etnia*

Quanto às etnias, pouco é possível concluir, pois a quantidade de pessoas que estão distribuídas entre as etnias são desproporcionais, uma vez que a quantidade de pessoas da etnia branca que responderam a pesquisa compõe mais de ⅔ do banco de dados.



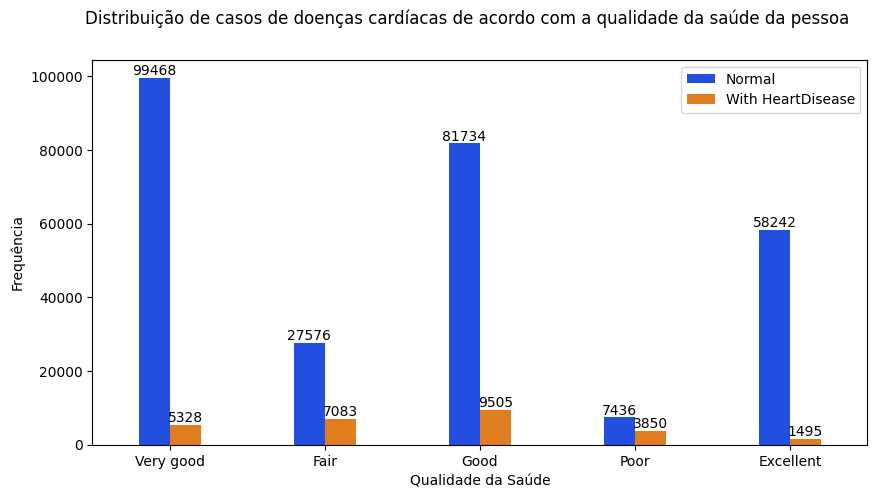
***Figura 15****: Distribuição dos casos de doença cardíaca de acordo com o grau de diabetes*

Neste gráfico, temos a distribuição da porcentagem de doenças cardíacas entre pessoas que possuem diabetes ou não. Neste caso, optamos por transformar as resposta de “No, borderline diabetes” e “Yes (during pregnancy)” em “No” ou “Yes”, pois elas acabam representando o mesmo tipo de resposta.



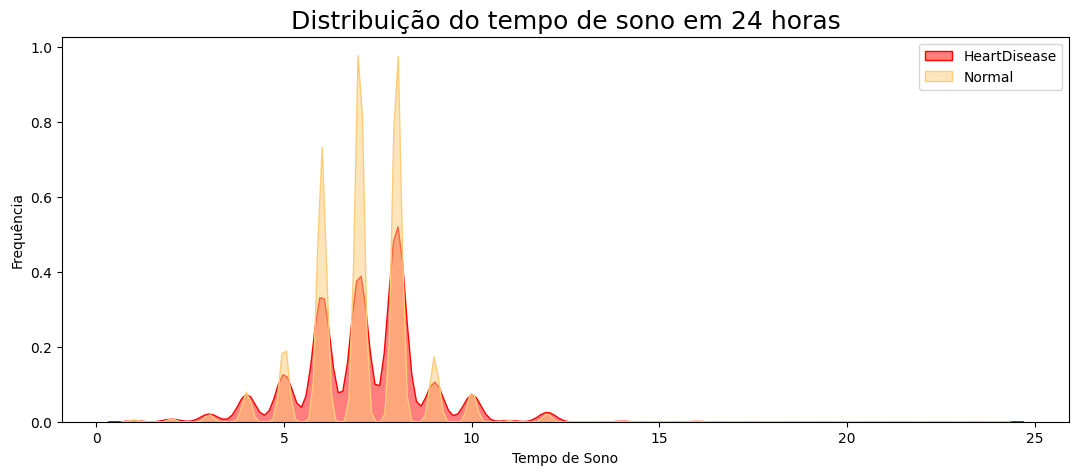
***Figura 16****: Distribuição dos casos de doença cardíaca de acordo com a prática de atividade física*

Neste gráfico, temos a distribuição doenças cardíacas entre as pessoas que praticam atividades físicas ou não. Neste caso, a porcentagem de pessoas que possuem doenças cardíacas e não praticam atividades físicas é o dobro das que praticam e possuem doenças cardíacas, sendo uma medida bem intuitiva quando comparada com a realidade.



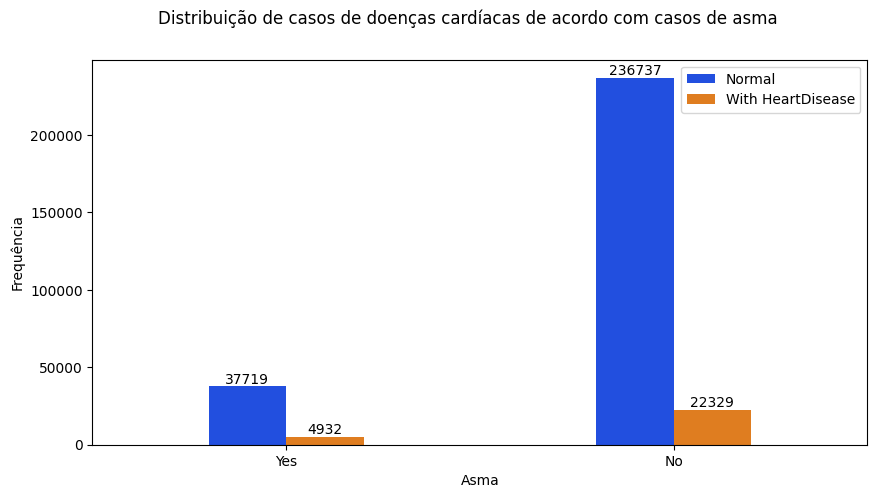
***Figura 17*** *: Distribuição dos casos de doença cardíaca de acordo com as classes de qualidade de saúde*

Na figura acima, é interessante observar que quanto pior a qualidade de vida, maior é o número de pessoas com doenças cardíacas, como no caso “Poor” e “Fair”, que possuem, respectivamente, cerca de 51,8% e 25,71% das pessoas com doenças cardíacas.

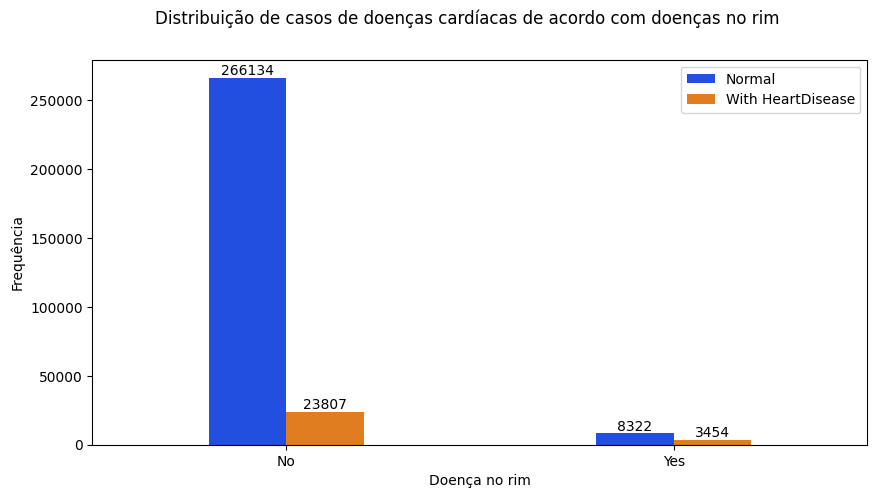


***Figura 18*** *: Distribuição dos casos de doença cardíaca de acordo com o tempo de sono*

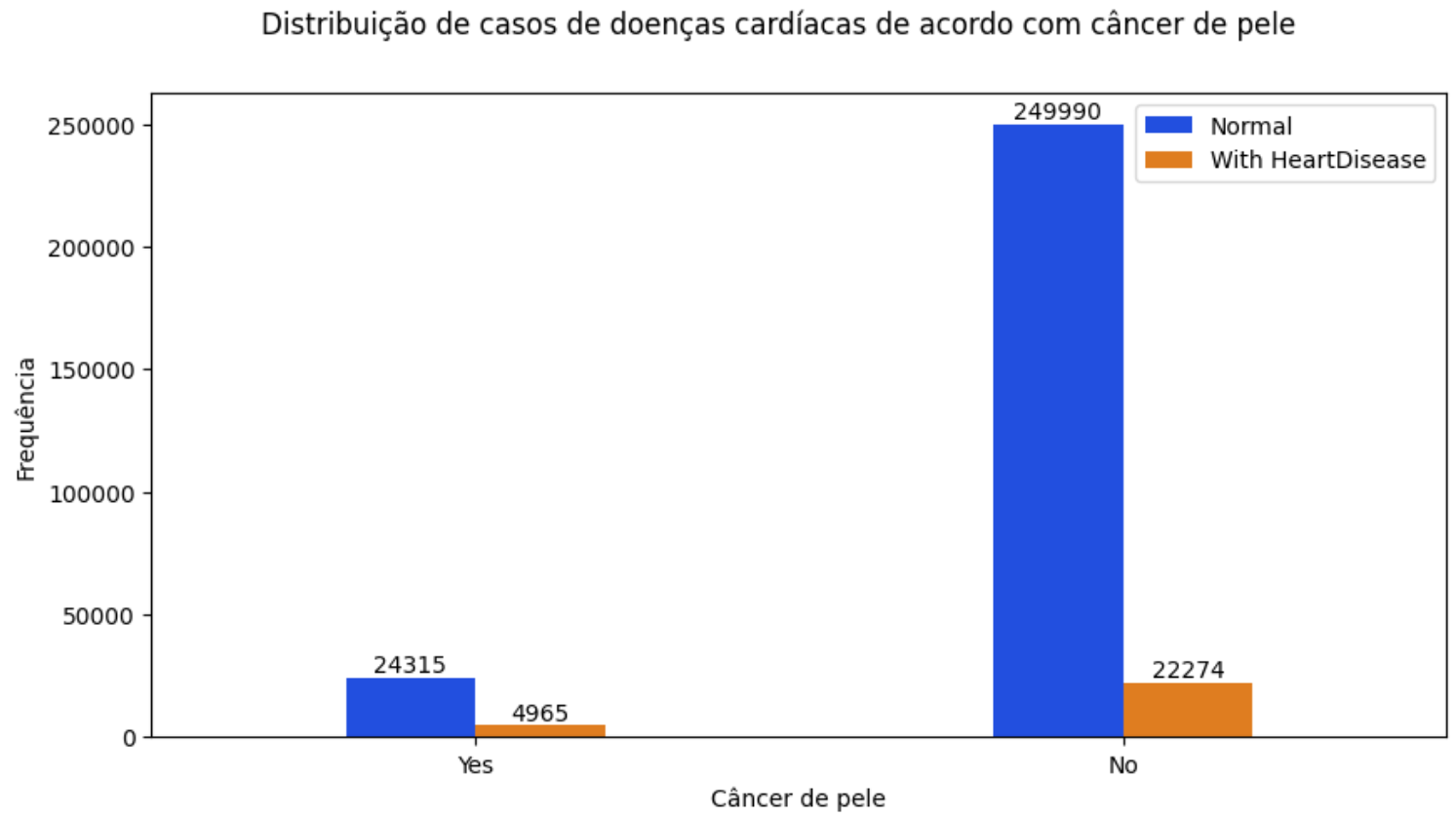
Neste gráfico temos a última variável de valor numérico presente na base de dados antes das transformações. É interessante observar que a maior distribuição de pessoas com doenças cardíacas está presente nas pessoas que possuem uma quantidade de sono dentro da média, entre 6 a 8 horas.



***Figura 19*** *: Distribuição dos casos de doença cardíaca de acordo com a presença de asma*



***Figura 20*** *: Distribuição dos casos de doença cardíaca de acordo com a presença de doenças renais*



***Figura 21*** *: Distribuição dos casos de doença cardíaca de acordo com a presença de câncer de pele*

Os últimos 3 gráficos, das figuras 19, 20 e 21, informam sobre casos isolados de pessoas que possuem doenças cardíacas ou não e também possuem alguma outra doença, como asma, doenças no rim ou câncer de pele. Deste dados, é interessante ressaltar a correlação entre doenças no rim com a possibilidade de ter doenças cardíacas ou não, uma vez que seu valor é muito maior quando comparada a outras duas doenças.

# Classificação

Com os dados tratados, foi possível aplicar os algoritmos de classificação para cada abordagem feita, a fim de obter melhores resultados. Para tal, foram criadas duas variáveis, que são divididas em: conjunto de dados **X**, que possui todos os atributos sem a classe a ser identificada, e o **Y,** que possui a classe. Além disso, foi dividido o conjunto de treinamento e teste na proporção de 80% e 20%, respectivamente, realizada com *holdout validation - random\_state = 44*

## Justificativa da escolha dos métodos de classificação

O grupo optou por realizar a classificação do conjunto de dados utilizando três dos diversos algoritmos vistos em aula, em todas as abordagens, de modo a comparar de maneira justa o desempenho. Os algoritmos são:

1. **Árvore de decisão**

Optamos pela árvore de decisão por ser um dos algoritmos mais simples e bem utilizados para realizar a classificação, além de ser relativamente rápida. Por se tratar de um algoritmo rápido e simples, foi um bom ponto de partida para a classificação do nosso conjunto de dados.

1. **Cross-Validation com KNN**

Quando utilizamos o KNN sem Cross-Validation, encontramos um problema em relação ao tempo do treinamento do algoritmo, devido a grande quantidade de dados presentes na nossa base de dados. Por conta disso, aplicamos o Cross-Validation, além de que pode ser utilizado como solução para o nosso problema de classes desbalanceadas.

1. **Cross-Validation com RandomFlorest**

Por fim, com o intuito de melhorar os resultados obtidos pela Cross-Validation com KNN, optamos por utilizar o algoritmo de Random Forest. Além disso, consiste de uma técnica que não temos muito conhecimento e queríamos estudar mais sobre.

## Ajuste de parâmetros e hiperparâmetros

Para cada um dos algoritmos classificadores utilizados, temos os seguintes casos:

**a. Árvore de decisão**

Ao usarmos esse classificador, utilizamos seus parâmetros padrões, como por exemplo: definir o nível da árvore máximo como “None”, random\_state como 0 e a estratégia de gini.

**b. Cross Validation com o KNN**

Na utilização do KNN, estipulamos apenas um hiperparâmetro, sendo ele a quantidade total de vizinhos. No caso, esse valor foi de 3, enquanto a quantidade de folds do Cross-Validation para este algoritmo foi de 100.

**c. CrossValidation com RandomForest**

Por fim, no RandomForest, utilizamos apenas os hiperparâmetros padrões. Já para o Cross-Validation nesse caso, optamos por utilizar apenas 10 folds, devido ao custo computacional para treinamento do modelo.

## Comparação dos resultados

* + 1. **Conclusão das Abordagens**

Após os resultados obtidos, foi confecionado uma tabela, presente na **Figura 20**, com a relação abordagens x classificadores, favorecendo a confecção dessas conclusões:

**Abordagem** 1

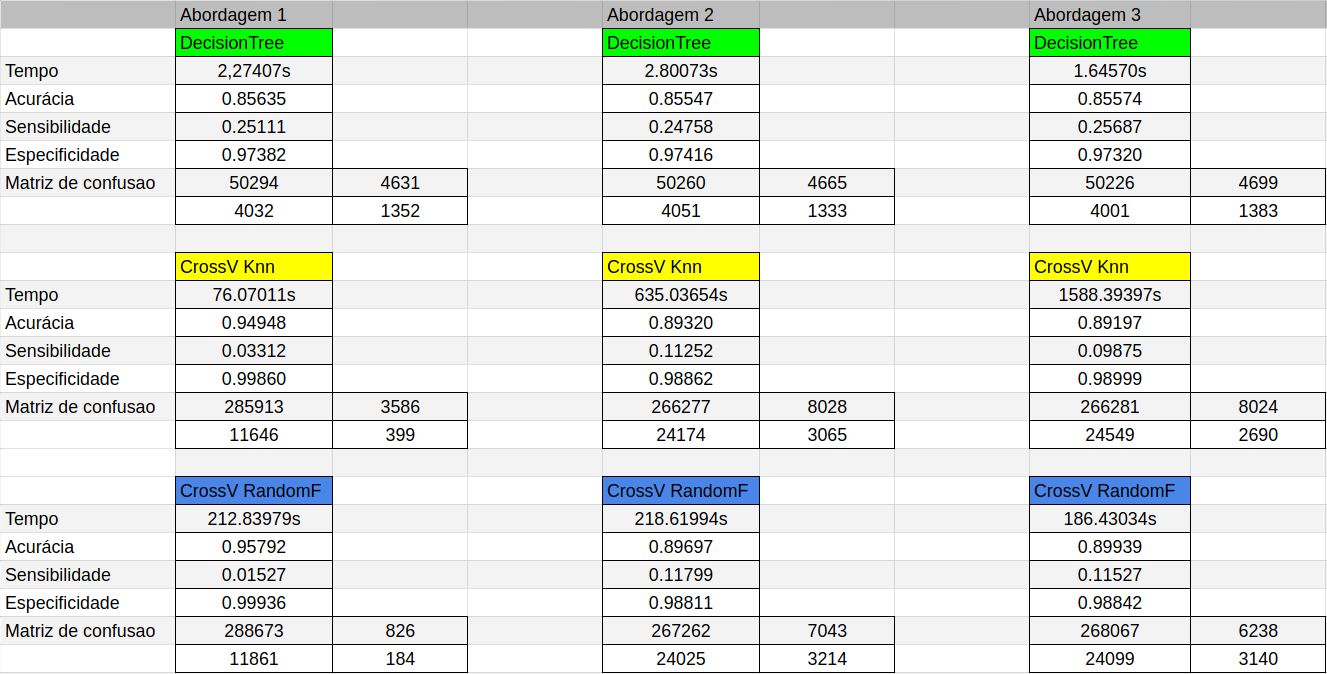
* + A Decision Tree teve um resultado melhor do que os outros algoritmos na métrica de Sensibilidade (25%), o que quer dizer que possui uma capacidade melhor de classificar verdadeiros positivos. Sua acurácia foi de 85% e sua especificidade de 97,3%. Um dos destaques da Decision Tree é o seu custo computacional, levando apenas 2,27 segundos para ser treinado.
  + O Cross-Validation com KNN teve uma piora na taxa de sensibilidade, obtendo um valor de 3%, mas tanto sua acurácia quanto sua especificidade aumentaram, tendo valores de 95% e 99%, respectivamente. Entretanto, o tempo computacional teve um aumento significativo para 76 segundos.
  + Por fim, o Cross-Validation com RandomForest teve a menor das taxas de sensibilidade, com apenas 1%, mas teve uma especificidade de 99,9% e acurácia de 96%, com tempo computacional de 217 segundos.

**Abordagem 2**

* + A Decision Tree, novamente, teve um resultado melhor do que os outros algoritmos no quesito sensibilidade, com valor de 24,7%. Os valores de acurácia e especificidade foram muito parecidos com a abordagem 1, com valores de 85,5% e 97,4% respectivamente. Novamente, o Decision Tree se destacou consideravelmente em seu tempo computacional, levando apenas 2,8 segundos.
  + Cross Validation com o KNN - Acabou sendo muito pior que na primeira abordagem. No geral, melhorou na Sensibilidade com 11%, mas no resto acabou piorando. A especificidade diminui para 98%, A Acurácia ficou em 89% e o tempo computacional ficou em torno de 635 segundos. Apesar de ter sido melhor que o decision tree, perde muito em tempo.
  + O Cross Validation com KNN, nesta segunda abordagem, teve uma queda em sua acurácia para 89,3%, porém, sua sensibilidade elevou para 11,2% e sua especificidade reduziu para 98,8%. O tempo, dessa vez, foi muito maior, levando 635 segundos (cerca de 10 minutos).
  + Assim como o Cross Validation com KNN, teve redução em sua precisão para 89,6% e em sua especificidade para 98,8%. Em contrapartida, sua sensibilidade aumentou para 11,7% e seu tempo computacional aumentou apenas 6 segundos, tendo um total de 218 segundos.

**Abordagem 3**

* + Por fim, na abordagem 3, a Decision Tree teve valores muito semelhantes a sua aplicação nas abordagens 1 e 2. Sua acurácia foi de 85,5%, sensibilidade de 25,6% e especificidade de 97,3%. O seu tempo computacional foi reduzido para 1.6 segundos.
  + O Cross Validation com KNN também manteve valores muito semelhantes a abordagem 2, com valores de acurácia de 89,1%, sensibilidade de 9,8% e especificidade de 98,9%. A dificuldade do KNN é em relação a seu tempo computacional, levando cerca de 25 minutos para a abordagem 3.
  + Por fim, o Cross Validation com RandomForest também ficou muito semelhante com a abordagem 2, tendo acurácia de 89,9%, sensibilidade de 11,5% e especificidade de 98,8%. Seu tempo computacional foi de 186 segundos.



***Figura 22:*** *Métricas de performance dos algoritmos em cada abordagem*

*Resultados com base na abordagem e estratégia*

# Conclusão

Com base nas informações fornecidas, podemos concluir que, à medida que o estudo do conjunto foi concluído e as manipulações foram aplicadas para executar as técnicas de classificação, os algoritmos utilizados foram se complementando de forma incremental. Embora a Abordagem 1 tenha apresentado resultados superiores quanto a acurácia dos algoritmos de CrossValidation com RandomForest e KNN, quando houve a mudança na estrutura dos dados, perdemos a estruturação básica dos mesmos, o que pode comprometer a precisão dos resultados obtidos. Portanto, é aconselhável concentrarmo-nos apenas nos resultados das duas últimas abordagens.

No geral, ao olharmos o contexto todo, é de suma importância comentar que os algoritmos trabalham bem a Especificidade - verdadeiros negativos - variando de 97,3 % até 99,9%. Isso é muito bom para detectar se a pessoa não tem doenças. Em contraposição, a Árvore de Decisão é a melhor em detectar a sensibilidade e em tempo computacional, ou seja, caso o algoritmo de um valor positivo, é muito capaz da pessoa realmente estar com a doença - ou seja, trabalha bem com verdadeiros positivos. Embora isso seja um ponto positivo, o valor da sua sensibilidade ainda é muito pequeno, com cerca de apenas 25%.

Podemos concluir então que, para definir qual algoritmo é o mais adequado ao nosso conjunto de dados, não basta apenas observar a sua acurácia, mas sim também outras métricas, como a sensibilidade e especificidade. Do ponto de vista medicinal, a Decision Tree pode ser bem relevante, mesmo possuindo uma acurácia menor quando comparado com os outros métodos, pois prevê mais corretamente classes verdadeiras positivas.

Além disso, ao colocarmos em prática os conhecimentos adquiridos em sala de aula, obtivemos uma compreensão mais profunda da linguagem de programação Python, bem como da realização de análises de dados e da obtenção de conclusões a partir delas. Assim, todas as atividades realizadas serviram como oportunidades de aprendizado, permitindo que os membros do grupo consolidaram o conhecimento adquirido e também buscassem informações adicionais sobre o mesmo tema, como é o caso do algoritmo RandomForest. Essa experiência contribuiu significativamente para o nosso desenvolvimento e enriquecimento no campo da análise de dados.

# Bibliografia

[Fleury](https://www.fleury.com.br/noticias/precisao-diagnostica)

[AM - Apresentação](https://docs.google.com/presentation/d/1yJEVHckY72Dh66GmSV0WAZiJKjaa5td_UuVAUMN6GBs/edit#slide=id.g1e441c91229_0_86)

[Pasta AM](https://drive.google.com/drive/u/0/folders/1BFMpsXYWUiVmPFopBm7QAZLqHFUvXTJF)

[Dataset](https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease)

[Planilha Comparações](https://docs.google.com/spreadsheets/d/1T6UEvD7kRoBa8FojJCFChHCxBbrDjQlseqyv2Zj2qGU/edit#gid=0)

[Matérias Saúde](https://bvsms.saude.gov.br/use-o-coracao-para-vencer-as-doencas-cardiovasculares-29-9-dia-mundial-do-coracao/#:~:text=No%20Brasil%2C%20as%20doen%C3%A7as%20cardiovasculares,250%25%20desses%20eventos%20no%20pa%C3%ADs.)

[IMC](https://endocrinologiacuritiba.com.br/blog/perguntas-frequentes/qual-o-meu-peso-ideal/attachment/attachment-tabela-imc/)