Identificação de problemas cardíacos através de modelos preditivos baseados em supervised machine learning

Matheus Henrique Lopes Cunha

Paloma Melo Pereira

RESUMO
Os modelos preditivos de "supervised machine learning", estão se tornando cada vez mais importantes no auxílio a tomada de decisões para diversas áreas do conhecimento humano e, consequentemente, também serão importantes para auxílio em casos médicos de maior complexidade. O objetivo do estudo é desenvolver um algoritmo de "supervised machine learning" que consiga ter uma alta taxa de aproveitamento em relação a previsão se determinada pessoa tem problema cardíaco ou não. O artigo mostra como foram desenvolvidos os modelos, os testes aplicados antes da implementação dos modelos, a taxa de aproveitamento de cada modelo e uma análise de qual é o modelo mais eficiente para situação em específico. As especificações de cada modelo de "supervised machine learning" e o seu impacto no desenvolvimento dos modelos que foram usados no trabalho, determinados por teste e aplicações feitas na linguagem de programação Python; foram considerado os resultados positivos e negativos para se chegar a uma posição final sobre qual era a melhor forma de se usar os algoritmos neste caso. O artigo conclui que a aplicação de modelos de "supervised machine learning" no diagnóstico de problemas cardíacos, pode ajudar muitas instituições de saúde, tanto públicas quanto privadas, a agilizar processos e a aumentar a taxa de acerto na hora de classificar uma pessoa como cardíaca ou não cardíaca, a terem uma elevada melhora nesse processo e consequentemente aumentar a eficiência e a lucratividade no caso das instituições privadas.

Palavras-chave: XGBoost, Random Forest, KNeighbors Classifier, Saúde, Coração.

1 INTRODUÇÃO

Os hábitos de vida de uma pessoa estão interligados, de forma direta, à sua saúde e à predisposição ao desenvolvimento de doenças cardiovasculares. A doença cardiovascular é a principal causa de doença e morte no Brasil, afetando ambos os sexos. A lesão aterosclerótica é identificada como o fator etiológico fundamental dessa condição. (DINIZ, 2011). Hipertensão, altos níveis de colesterol e de glicose, consumo de álcool, tabagismo, obesidade e sedentarismo são fatores que contribuem para o surgimento de doenças cardiovasculares.

De acordo com a Sociedade Brasileira de Arritmias Cardíacas, a hipertensão está relacionada ao surgimento de arritmias cardíacas, com ênfase na fibrilação atrial, como também está relacionada à insuficiência cardíaca com dilatação do coração (MAGALHÃES et al, 2016). Isso se justifica porque a hipertensão acomete gravemente artérias importantes do corpo, como as coronárias, que fazem a irrigação do coração.

As doenças cardiovasculares possuem íntima relação com a idade e o sexo biológico, uma vez que determinadas doenças estão ligadas a certas faixas etárias. Nas crianças e jovens, é mais comum a ocorrência

de anomalias congênitas e os primeiros episódios da moléstia reumática. Já dos 20 aos 50 anos de idade, a doença de Chagas e a hipertensão arterial são mais frequentes. Por outro lado, a doença arterial coronariana, incluindo a angina de peito e o infarto agudo do miocárdio, é mais comum em pessoas acima dos cinquenta anos de idade (PORTO, 7ª edição, 2014). Também é possível afirmar que o sexo biológico gera impacto nos tipos de doenças cardiovasculares predominantes em homens e mulheres. As lesões mitrais, como a estenose e o prolapso da valva mitral, são mais comuns em mulheres jovens. Por outro lado, a aterosclerose coronária é mais predominante em homens até os 45 anos de idade. A partir dessa faixa etária, a incidência da doença torna-se igual em ambos os sexos (PORTO, 7ª edição, 2014). A aterosclerose coronária iguala sua incidência em homens e mulheres no marco dos 45 anos, porque é nessa média de idade que as mulheres entram na menopausa e ocorre redução na produção do hormônio estrogênio que é naturalmente uma substância protetiva dos vasos sanguíneos. O estrogênio promove a liberação de substâncias vasodilatadoras no endotélio, como o óxido nítrico, e diminui a produção de compostos vasoconstritores. Sendo assim, antes da menopausa as mulheres possuem valores de pressão arterial mais baixo do que em homens e menor probabilidade de desenvolver problemas cardíacos. (MACIEL et al, 2021).

Outro fator relacionado às doenças cardiovasculares é o cigarro. Ele está associado ao aumento do nível de colesterol, trombose, agregação plaquetária e doenças coronarianas. Os componentes presentes no cigarro propulsionam a ativação e a liberação de células inflamatórias que levam ao aumento de mediadores inflamatórios na corrente sanguínea. Por sua vez, esses reguladores inflamatórios desencadeiam a elevação de outros compostos que estão relacionadas ao maior risco de desenvolver infarto do miocárdio e doença coronariana. Além disso, o cigarro leva a alterações na camada íntima dos vasos sanguíneos, diminuindo as concentrações de óxido nítrico, ou seja, pessoas fumantes possuem menores quantidades desse composto quando comparadas a não fumantes (NUNES et al, 2011).

O sedentarismo também é considerado um agente etiológico das doenças cardiovasculares. A inatividade física faz com que não haja um retorno venoso apropriado, porque os músculos não se contraem de forma eficaz, sendo assim, o trabalho conjunto dos músculos e das válvulas venosas fica comprometido, favorecendo a formação de trombos. Os trombos estão associados a transtornos cardiovasculares importantes, entre eles a doença cardíaca isquêmica. Além disso, o sedentarismo leva ao aumento da frequência cardíaca e é causa direta da obesidade (GASPAR, 2004).

Elevados níveis de colesterol estão ligados a eventos inflamatórios dos vasos sanguíneos que resultam em formação de placas de ateroma que desembocam em problemas cardíacos isquêmicos. Algumas condições como diabetes, hipertensão e tabagismo podem lesar o endotélio dos vasos sanguíneos gerando um processo inflamatório que aumenta a permeabilidade da íntima às lipoproteínas e faz com que o espaço subendotelial acumule colesterol e forme a placa de ateroma, que cresce em direção à luz dos vasos e reduz o fluxo sanguíneo. Além disso, se a placa aterosclerótica sofrer uma lesão, ocorre coagulação no local e

formação de trombos, o que prejudica ainda mais a passagem de sangue nessa região. Caso o coração seja irrigado por alguma artéria que possui obstrução por placas de ateroma, o suprimento de oxigênio para as células cardíacas será baixo, aumentando a chances de ocorrência de isquemia e infarto (XAVIER et al, 2013). Ainda, é observado que em populações com níveis baixos de colesterol plasmático, os casos de ataques cardíacos e aterosclerose são pouco frequentes (CASTRO et al, 2004).

Por fim, pessoas com altos níveis glicêmicos também possuem maior probabilidade de desenvolver problemas cardíacos. O aumento intracelular de glicose e de substâncias pró inflamatórias em pessoas com hiperglicemia levam a lesões teciduais nos vasos sanguíneos capazes de formar placas ateroscleróticas instáveis e mais propensas à ruptura. Além disso, elevadas taxas de glicose diminuem complexos anti-agregantes ao passo que eleva compostos ativadores de plaquetas que, juntos, contribuem para formar a agregação plaquetária e gerar trombos (Schaan; Portal, 2004).

Portanto, fica evidente como vários fatores associados proporcionam o surgimento de problemas cardiovasculares, sendo necessária a implantação de medidas preventivas e de diagnóstico precoce para se evitar possíveis sequelas. De acordo com o Ministério da Saúde, a detecção precoce de doenças tem como fundamento a ideia de que algumas condições possuem maiores chances de cura, sobrevida e/ou qualidade de vida quando diagnosticadas o mais cedo possível. Existe um valor significativo em identificar a doença em estágios iniciais e assintomáticos, pois nesse período o tratamento e a cura podem ser alcançados com maior facilidade (MINISTÉRIO DA SAÚDE, 2010).

Como é um problema muito grave e que gera sequelas devastadoras, podendo levar até a morte. Este trabalho utiliza de modelos de “supervised machine learning” para ter uma predição se a pessoa tem um problema cardíaco ou não, assim tornando mais rápido e certeira a detecção de problemas cardíacos. Os algoritmos vão levar em consideração várias características e hábitos da pessoa que terá análise do coração julgada tais como idade, taxa glicêmica, colesterol, se a pessoa é fumante ou não, se a pessoa consome álcool ou não, se a pessoa é ativa em relação a atividades físicas, e a relação que todas as características anteriores têm com a avaliação de algumas pessoas que já tiveram o diagnóstico confirmado. Foi afirmado que o aprendizado de máquina pode ser compreendido como a solução de problemas geométricos na área da inteligência artificial. Conforme suas palavras, os modelos utilizados nesse contexto são programados para analisar informações de um conjunto de dados, adquirir conhecimento e aprimorar-se com base em sua experiência dentro dos parâmetros fornecidos (Piovezan, 2022).

Diante do exposto, percebesse que essa análise é uma técnica que envolve poderosa tecnologia, tem como grande intuito agilizar o processo de classificação e trazer resultado mais precisos sobre o diagnóstico cardíaco real de uma pessoa. Com isso citado pode-se concluir que o estudo do diagnóstico de problemas cardíacos com “supervised machine learning”, é extremamente importe e gerador de muito conhecimento. A tecnologia tem o poder de proporcionar oportunidades sem precedentes, permitindo que pessoas com

deficiência visual participem ativamente das redes sociais, interajam, comentem e tenham os mesmos acessos que todos os demais usuários. Essa capacidade inclusiva é considerada maravilhosa, refletindo uma visão de futuro que busca incorporar cada vez mais indivíduos. Crianças agora podem desfrutar de vídeos e se divertir, e a expectativa é que cada vez mais pessoas sejam incluídas na web, resultando em um mundo diferente. Apesar de alguns ambientes online ainda serem tóxicos, o autor ressalta que existem pessoas que desejam e apoiam ambientes saudáveis (Junior, 2022).

Este trabalho tem como objetivo criar um algoritmo de "supervised machine learning" que tenha uma alta taxa de precisão ao dar um diagnóstico de um problema cardíaco. Os dados apresentados, foram retirados do banco de dados sobre problemas de coração, que se encontra no site Kaggle. Os dados foram analisados através da inteligência artificial da linguagem de programação Python, pela qual foram fornecidas análises estatísticas e previsões com "machine learning".

2 MATERIAL E MÉTODOS

Um problema cardíaco, também conhecido como doença cardíaca, refere-se a qualquer condição ou doença que afete o funcionamento normal do coração. Quando ocorrem problemas no coração, isso pode levar a uma série de complicações médicas sérias e até mesmo fatais.

A prevenção de problemas cardíacos é crucial para manter a saúde do coração e reduzir o risco de doenças cardiovasculares. As prevenções podem ser: ter uma dieta saudável, manter um peso dentro do padrão, fazer exercícios físicos com frequência, limitar o consumo de álcool, ter a pressão arterial sob controle, controlar o diabetes, reduzir o estresse, dormir o suficiente, fazer exames médicos com regularidade, evitar o uso de drogas ilícitas e seguir as orientações médicas.

A pesquisa foi elaborada com o objetivo de compreender as informações estatísticas referentes a análise de quem tem problema cardíaco ou não. Através dessa análise, criar modelos de "supervised machine learning" para determinar sua qualidade final de forma segura, de modo a entender como tal espaço foi formado e desenvolvido.

A metodologia usada nesta pesquisa é de caráter explicativo, e tem como sua natureza descritiva. O estudo foi do tipo quantitativo e qualitativo, utilizando análise de dados e utilização de algoritmos de "supervised machine learning" pela linguagem de programação Python, para interpretação dos dados. Diante disso, o estudo quantitativo, geralmente é realizado pela: nos estudos, a coleta de dados é comumente realizada por meio de questionários e entrevistas, que abrangem diferentes variáveis relevantes para a pesquisa. Essas informações coletadas são, em geral, apresentadas por meio de tabelas e gráficos durante a análise. (DALFOVO, LANA E SILVEIRA, 2008). Já no estudo qualitativo Carspecken (2011, p.27) é comum que um pesquisador social qualitativo tenha o interesse em compreender o funcionamento das

formas de poder, principalmente em interações reais que ele observa e nas quais possivelmente participa (Carspecken, 2011).

Os códigos de programação usados neste trabalho e versão do software e pacotes utilizados estão no "Github", o link está disponível na seção de apêndices, esta plataforma é uma excelente alternativa para se armazenar os códigos, uma vez que os "scripts" estão na nuvem, o que faz com que se evite a perca dos códigos por qualquer problema que o computador apresente.

O banco de dados que foi utilizado neste estudo foi obtido através do Kaggle, que é uma é uma famosa comunidade de competição de aprendizagem de máquina. E o banco de dados tem oito colunas com informações das variáveis que foram analisadas, setenta mil linhas com informações das pessoas que foram analisadas, e ele foi submetido ao site em janeiro de 2023.

Também foi feito um estudo prévio de cada variável presente no banco de dados, para validação do tipo que cada variável apresentava, além disso, ter uma noção básica de como seria usada cada informação. A descrição de cada tipo de informação pode ser vista na Tabela1.

Tabela1. Variáveis

Variável	Tipo	Descrição dos dados
Age	*Inteiro*	Idade dos participantes
Gender	*Texto*	Gênero dos participantes
Height	*Inteiro*	Altura dos participantes
Weight	*Inteiro*	Peso dos participantes
Ap_hi	*Inteiro*	Pressão sistólica dos participantes
Ap_lo	*Inteiro*	Pressão diastólica dos participantes
Cholesterol	*Inteiro*	Se o paciente fuma ou não
Gluc	*Inteiro*	Se o paciente consome álcool ou não
Smoke	*Booleano*	Se o paciente é ativo ou não
Alco	*Booleano*	Se o paciente é cardíaco ou não
Active	*Booleano*	
Cardio	*Booleano*	

Fonte: Dados originais da pesquisa

Com os dados já classificados, e com a identificação de que não era necessário fazer uma limpeza mais profunda no banco. Os dados foram submetidos a uma análise exploratória, na qual foi entendido quais seriam as melhores opções de modelo para o caso em específico.

O estudo de caráter explicativo é um método de análise científica que busca explicar como funciona e o desempenho alcançado pelos modelos do segmento estudado. O trabalho vem no molde quantitativo, pois a intenção é trazer uma abordagem com dados numéricos, os quais serão apresentados no formato gráfico, discursivo e estatísticos, com intuito de conseguir entender qual é o modelo de "supervised machine learning" que tem maior aplicabilidade para o estudo em questão. De modo geral, os estudos de campo quantitativos seguem um modelo de pesquisa semelhante à pesquisa experimental, em que o pesquisador utiliza quadros conceituais de referência bem estruturados para formular hipóteses sobre os fenômenos e situações que deseja investigar (Dalfovo, Lana e Silveira, 2008). E o trabalho também apresenta traços qualitativos, pois existe uma análise de dados significantes e transformadoras, a pesquisa qualitativa crítica é descrita como uma prática verdadeiramente estimulante, política e significativa, capaz de expandir a mente. Tanto as experiências de trabalho de campo quanto a análise de dados são mencionadas como processos ricos em significado e com potencial transformador (Carspecken, 2011).

O "machine learning" é definido como: os componentes do aprendizado de máquina envolvem um conjunto de variáveis denominadas "features", que podem ser medidas ou pré-definidas, juntamente com um conjunto de saídas, que podem ser conhecidas ou desconhecidas. O processo de construção do modelo se baseia na utilização de um conjunto de dados composto por exemplos (Tome, 2017).

O "machine learning" é uma ferramenta complicada, mas que pode ajudar no desenvolvimento profissional e no pessoal, os conceitos e ferramentas que nele existem, facilitam as formas de chegar no objetivo. A aprendizagem incremental é caracterizada pela acumulação dinâmica de informações extraídas das experiências vivenciadas. A abordagem adaptativa da aprendizagem de máquina visa integrar técnicas simbólicas de aprendizagem de máquina com técnicas adaptativas, a fim de resolver problemas de aprendizagem de forma eficiente (Stange et al, 2011).

Existem várias formas de se utilizar o "machine learning", que também são fatores muito importantes, pois é com elas que se define o modelo de melhor performance que vai ser usado em cada situação. Durante a fase de treinamento, a presença de atributos irrelevantes e redundantes dificulta o aprendizado do classificador. Uma abordagem para lidar com essa questão é selecionar os atributos mais importantes para a classificação, ou seja, aqueles que têm maior capacidade de distinguir entre notícias positivas e negativas. Essa seleção visa remover os atributos irrelevantes e melhorar o desempenho do classificador (Almeida, 2014).

Os classificadores escolhidos para os testes de eficiência dos modelos de "supervised machine learning" que foram usados para alcançar os resultados deste estudo, utilizam os métodos de "Random Forest", "XGBoost" e "KNeighbors Classifier". Os modelos foram testados para ver qual seria o classificador mais eficiente.

Para não gerar “overfitting” e o “underffiting” e gerar um modelo mais confiável, foi utilizado a divisão dos dados em modelos de teste e treino, pois quando os dados vão para os algoritmos de “supervised machine learning” de forma direta, ele tende a ficar com previsões viciadas. Em situações em que a prioridade é selecionar o modelo com a melhor capacidade de previsão, é necessário ter cautela com o “overfitting” e o “underfitting”. O “overfitting” ocorre quando o modelo se ajusta excessivamente aos dados de treinamento, resultando em previsões inadequadas quando aplicado a novos dados. Já o “underfitting” se refere à situação em que o modelo não se ajusta adequadamente nem mesmo ao conjunto de treinamento (Lopes, 2018).

Para aplicar os modelos “supervised machine learning” com os dados já melhorados, foram realizados testes estatísticos para ver a importância de cada atributo e consequentemente se o atributo poderia ser irrelevante para o modelo. Sendo assim, foram excluídas duas colunas do banco de dados: altura e peso. Essas duas variáveis sozinhas não são impactantes em problemas cardiovasculares, fazendo-se necessária a manipulação desses elementos por meio do Índice de Massa Corporal. Porém, esse índice não é tão eficaz, porque não leva em consideração a composição corporal, bem como a distribuição de gordura no corpo do indivíduo (RECH et al, 2006). Ainda, os estudos feitos por Hans et al mostraram que a circunferência da cintura está mais relacionada com os fatores de riscos das doenças cardiovasculares (HANS et al, 1995). Foi analisado o “boxplot” para detectar “outliers” e correlação para analisar possíveis relações entre as variáveis. Além disso, também os dados foram verificados quanto à normalidade para se validar se o número de amostras é suficiente para o modelo.

Depois de todas as análises estáticas, foi usado a função “GridSearchCV”, com validação cruzada em 10 “folds”, para definir os melhores parâmetros para cada modelo. Depois, os dados foram divididos em 80% para treino e 20% para teste, onde as 70000 amostras foram divididas em 56000 para treino e 14000 para teste, aleatoriamente selecionados pela função “train_test_split”. O objetivo da implementação do padrão técnico de processo é minimizar as alterações nos parâmetros de controle do processo ao introduzir um novo produto. Isso visa melhorar a eficiência do setup da máquina, reduzir as perdas de produtividade e qualidade, além de eliminar a variabilidade das especificações que surgem durante a produção (Campos e Miguel, 2013).

Os modelos foram avaliados pela sua acurácia, sensibilidade e precisão, calculados por meio de sua matriz de confusão. Outra análise importante para este estudo é o teste de ruídos. Foi verificado até qual percentual de inclusão de ruídos nos modelos ainda mantêm boa performance. Os dados foram achados no site do Kaggle. O banco de dados tem setenta mil amostras e essas amostras têm doze amostras variáveis analisadas.

Os métodos e ferramentas que não são eficientes precisam ser modificados e melhorados. Dessa forma foi de grande importância para o estudo a utilização de várias técnicas, pois em algumas foram

achados resultados muito positivos e consequentemente foram mantidos e outros foram ineficientes e descartados. Vale ressaltar que a técnica amplamente reconhecida como "machine learning" ou "Aprendizado de Máquina" tem desempenhado um papel significativo em diversas áreas. Essa abordagem consiste em programar computadores para aprender com experiências anteriores, indo além da simples reprodução dos dados fornecidos. O sistema desenvolve uma capacidade cognitiva própria, permitindo um aprendizado contínuo com base em acertos e falhas (De Figueiredo e Cabral, 2020).

Neste contexto, o objetivo do algoritmo é produzir um classificador que consiga prever se determinada pessoa tem problema cardíaco ou não, mesmo quando as informações não são muito claras às análises estatísticas.

Equações estatísticas usadas nos modelos. No “XGBoost” foi usada a equação eq. (1):

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

Já no “Random Forest” foi usada a fórmula de medida de desigualdade gini, que está demonstrado na equação eq. (2):

:

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

O “KNeighbors Classifier” que tem como objetivo calcular por proximidade, assim os elementos que estão perto ou tem características muito parecidas são atribuídos como da mesma classe, e está representado equação eq. (3):

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

3 RESULTADOS E DISCUSSÃO

O seguinte trabalho tem como objetivo central mostrar como o “supervised machine learning” pode ser utilizado para a identificação de problemas cardíacos. Apresentando o que realmente tem um bom funcionamento e o que precisa evoluir em relação as ferramentas de aprendizado de máquina dentro deste seguimento. O “machine learning” é descrito como um método científico intensificado. Ele segue um processo semelhante de geração, teste e descarte ou refinamento de hipóteses. No entanto, enquanto um cientista pode levar uma vida inteira para criar e testar algumas centenas de hipóteses, um sistema de “machine learning” é capaz de realizar o mesmo em uma fração de segundo. O “machine learning” automatiza o processo de descoberta, o que explica o motivo de estar revolucionando tanto a ciência quanto os negócios (Domingos, 2017).

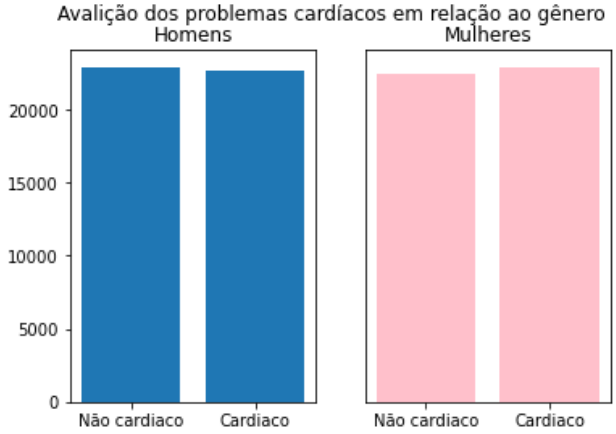
A elaboração do modelo se deu a partir da análise estatística de algumas características de pessoas que tem problemas cardíacos e de pessoas que não tem nenhum mal no coração, tendo em vista a relação que essas características têm com a avaliação final a relação se a pessoa tem problemas cardíacos ou não. Durante a análise dos dados as variáveis idade, taxa glicêmica, colesterol, se a pessoa é fumante ou não, se a pessoa consome álcool ou não e se a pessoa é ativa em relação a atividades físicas foram características levadas em consideração. Essa avaliação trouxe resultados satisfatórios para serem usados nos modelos.

Também é importante destacar que foram realizados testes de normalidade em cima das variáveis que foram introduzidas no modelo de “machine learning”. Foram feitas análises com os gráficos de distribuição de normalidade. E os testes constataram que todas as variáveis atestaram positivo para normalidade e consequentemente foram aproveitadas neste trabalho.

Em relação a idade das pessoas analisadas, percebesse que pessoas com idade acima dos dezoito anos tendem a ter mais problemas cardíacos, sendo pouca ou quase nenhuma o número de pessoas abaixo dessa idade que apresenta algum tipo de problema cardíaco. Já em relação aos que estão acima dos 18 anos. Todas as faixas etárias têm um certo grau de pessoas com problemas cardíacos.

O gênero é outro elemento que gerou uma análise interessante, pois foi uma análise na qual deu para perceber que o número de homens e mulheres que tem e não tem problemas cardíacos, são bem parecidos. Os homens analisados apresentaram ser em maioria não tendo problemas cardíacos, mas foi uma diferença pequena entre as duas classificações. Já as mulheres, tiveram a maioria com problemas cardíacos, mas também com uma diferença bem pequena.

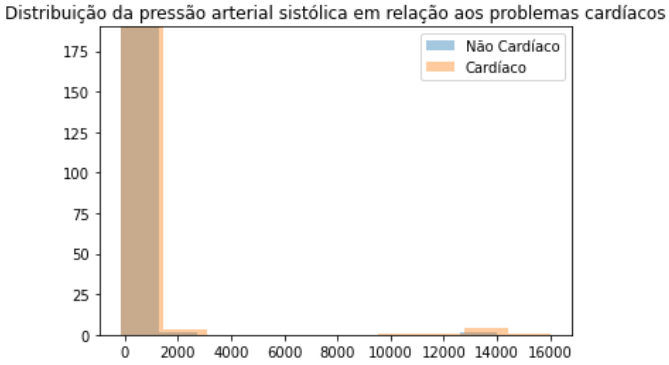
Figura 1. Avaliação dos problemas cardíacos em relação ao gênero



Fonte: Resultados originais da pesquisa

Dentro da pressão arterial sistólica, é visível que grande parte das pessoas cardíacas e não cardíacas apresentam valores concentrados em um determinado lado do gráfico a seguir, sendo que uma pequena parte tem valores que não estão concentrados nos lugares padrões do gráfico.

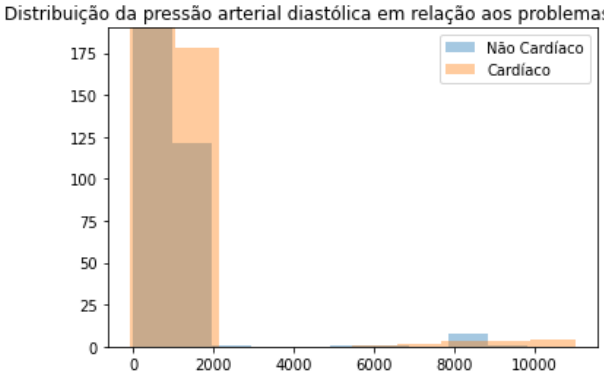
Figura 2. Distribuição da pressão arterial sistólica em relação aos problemas cardíacos



Fonte: Resultados originais da pesquisa

Em relação a pressão arterial diastólica, também se percebe uma concentração dos valores, porém é uma concentração um pouco mais distribuída, onde dá para ver com mais clareza a diferença da concentração das pessoas que tem problemas cardíacos e as que não tem problemas cardíacos.

Figura 3. Distribuição da pressão arterial diastólica em relação aos problemas cardíacos

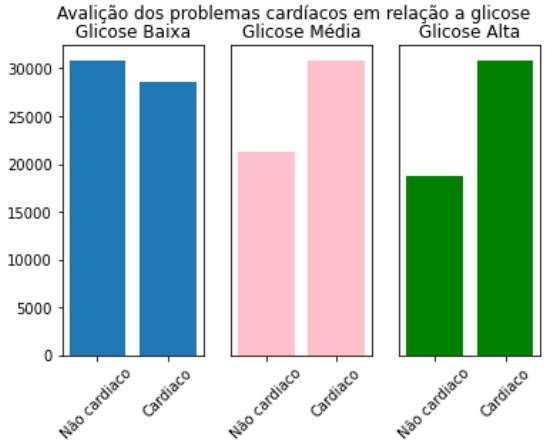


Fonte: Resultados originais da pesquisa

Logo após foi analisado o quanto o colesterol influência nos problemas cardíacos, sendo que o colesterol foi dividido em baixo, médio e alto. O colesterol baixo, apresentou uma amostra que contêm mais pessoas sem problemas cardíacos, porém apresentou um tanto considerável em relação as pessoas que têm problemas cardíacos também. Já o colesterol de nível médio, apesar de ter um tanto elevado de pessoas com problemas cardíacos, foi a variável que mais teve não cardíacos. O colesterol alto foi em grande maioria tendo problemas cardíacos.

A glicose das pessoas analisadas, também foi analisada em três categorias, baixa média e alta. A classificação de baixa glicose teve como maioria tendo não cardíacos, mas a diferença foi pequena. Já a glicose média teve como maioria pessoas com problema cardíaco. E quem tem a glicose alta tem em grande maioria problemas cardíacos.

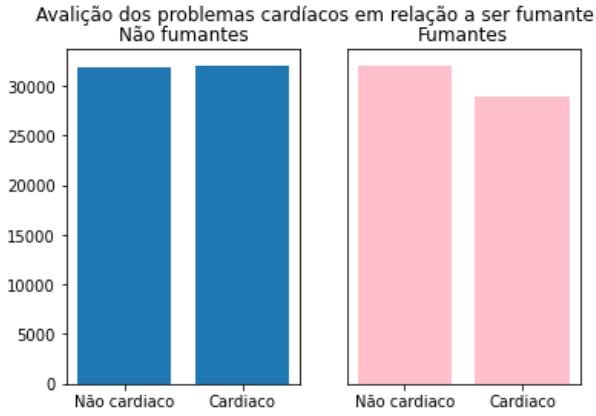
Figura 4. Avaliação dos problemas cardíacos em relação a glicose



Fonte: Resultados originais da pesquisa

A variável fumante foi dividida em fumante e não fumante. Em relação aos não fumantes, a relação entre cardíacos e não cardíacos foi bem parecida, com os cardíacos sendo levemente superior. E os fumantes tiveram a sua maioria como não cardíacos, mas tiveram muitas pessoas com problema de coração também.

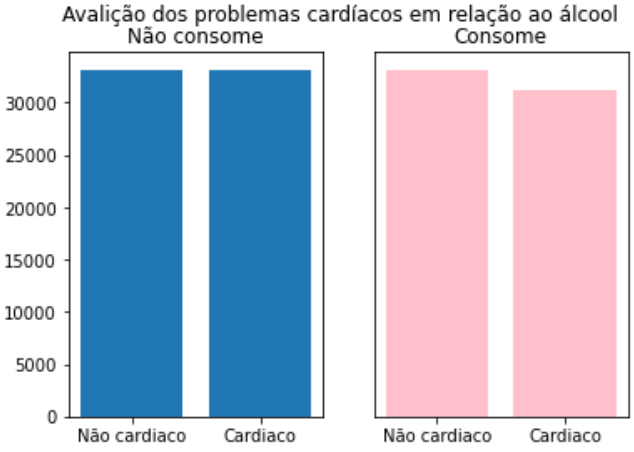
Figura 5. Avaliação dos problemas cardíacos em relação a ser fumante



Fonte: Resultados originais da pesquisa

O consomo de álcool também foi levado em consideração. Quem não consome álcool teve resultados quase que idênticos para ter ou não ter problemas cardíacos, com os não cardíacos levemente superior. Já os que consomem álcool tiveram em sua maioria como não cardíacos.

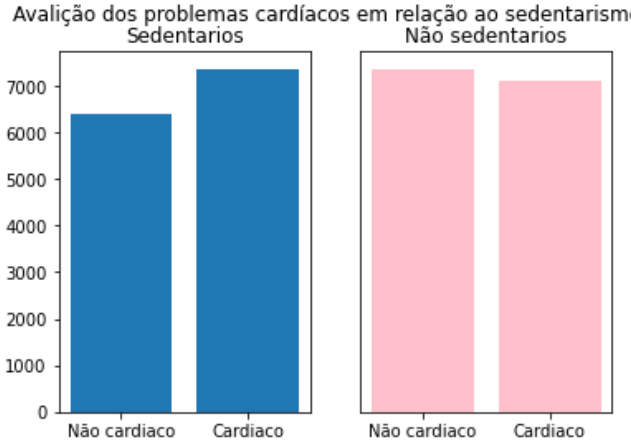
Figura 6. Avaliação dos problemas cardíacos em relação ao álcool



Fonte: Resultados originais da pesquisa

O sedentarismo também é um fator importante para identificar problemas de coração. As pessoas sedentárias foram em maioria pessoas com problemas cardíacos, mas com uma grande taxa de não ter problemas de coração. Já os não sedentários tiveram uma taxa maior de não cardíacos.

Figura 7. Avaliação dos problemas cardíacos em relação ao sedentarismo



Fonte: Resultados originais da pesquisa

Depois que as análises foram feitas, foram implantados os modelos de "supervised machine learning", o modelo no qual foram realizados os primeiros testes, foi o "Random Forest", primeiro foram encontrados os melhores parâmetros com "GridSearchCV". Depois os dados foram divididos em teste e treino, após foi medida a acurácia que foi de 66%, sensibilidade média que foi de 66.66%, a precisão média que é 66.60% e por último a matriz de confusão que teve 4675 erros dentro de 14000 variáveis possíveis.

Tabela2. Avaliação do desempenho do modelo de machine learning Random Forest

Métricas	Porcentagem
Acurácia	66%
Matriz de confusão	4675
Sensibilidade média	66,66%
Precisão média	66,60%

Fonte: Resultados originais da pesquisa

Logo após foi testado o modelo "XGBoost", no qual os testes seguiram o mesmo molde do primeiro teste, com a definição dos parâmetros feitos pelo "GridSearchCV" e com a divisão em dados de teste e treino. Ele foi o modelo que apresentou os melhores resultados, e para isso, o critério de cálculo usado no modelo foi o erro médio quadrático que é uma métrica amplamente utilizada para avaliar a qualidade de um modelo de regressão, em que se compara as previsões feitas pelo modelo com os valores reais dos dados. Também foi usado o "Log Loss", também conhecida como "Logarithmic Loss" ou "Cross-Entropy Loss", é uma métrica usada principalmente em problemas de classificação binária ou com muitas classes. Com densidade máxima de um, e o controle máximo do momento das divisões da característica do modelo ficou em automático.

A acurácia foi de 72.97%, a sensibilidade média foi de 72.92%, a precisão média foi de 73.13% e a matriz de confusão apresentou 3784 erros também dentro dos mesmos 14000 dados de teste.

Tabela3. Avaliação do desempenho do modelo de machine learning XGBoost

Métricas	Porcentagem
Acurácia	72,97%
Matriz de confusão	3784 erros
Sensibilidade média	72,92%
Precisão média	73,13%

Fonte: Resultados originais da pesquisa

Por último foi usado o modelo “KNeighbors classifier”, que também é conhecido como knn, neste modelo também foi mantido o procedimento padrão dos outros modelos, com a definição dos melhores parâmetros e treinar e testar os dados, com isso se obteve uma acurácia de 71,12%, uma sensibilidade média de 71,06%, a precisão foi de 71,55% e a matriz de confusão mostrou 4043 erros dentro das mesmas 14000 variáveis de teste.

Tabela4. Avaliação do desempenho do modelo de machine learning KNeighbors classifier

Métricas	Porcentagem de aproveitamento
Acurácia	71,12%
Matriz de confusão	4043 erros
Sensibilidade média	71,06%
Precisão média	71,55%

Fonte: Resultados originais da pesquisa

Depois que todo o processo foi realizado, foram adicionados ruídos dentro dos dados, para comprovar o bom desempenho dos modelos. O modelo que teve o melhor desempenho nesse quesito foi o “XGBoost” que conseguiu ter uma boa performance com até 70% de ruído incluso no banco, onde ele manteve 60,42% de acurácia. O modelo “KNeighbors classifier”, também manteve um bom desempenho, tendo uma boa performance com até 60% de ruído incluso, com a acurácia de 60,36%. Já o “Random Forest”, teve um bom resultado até a inclusão de 50% de ruído, obtendo uma acurácia de 61,45%.

Tabela5. Desempenho de todos os modelos de machine learning usados no trabalho quando submetidos a ruídos.

Modelo	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%

Fonte: Resultados originais da pesquisa

4 CONSIDERAÇÕES FINAIS

O modelo que apresentou melhor performance, foi o “XGBoost”, mesmo os outros modelos tendo apresentado resultados interessantes, o algoritmo conseguiu apresentar resultados superiores. Ele foi o que obteve melhor desempenho em relação ao teste de ruído, sendo assim o modelo com resultado mais eficiente e sendo considerado o melhor modelo. Apresentou 72,97% de acurácia diante dos dados originais e desempenho inferior ao se deparar com ruídos. Em trabalhos futuros devem ser buscados modelos mais robustos.

O trabalho apresenta limitações quanto as informações encontradas e consequentemente têm um resultado limitado pela pouca quantidade de informação, e outro problema encontrado foi o trabalho ter sido feito em computador que não tinha potencial de processamento adequada, o que deixou todos os processos mais lentos. Com mais informações que podem ser coletadas com o tempo, e com uso de hardware com maior potencial, os resultados seriam melhores.

Conclui-se que este trabalho proporciona o início do estudo da criação de uma ferramenta para o diagnóstico cardíaco de determinada pessoa. Acredita-se que o uso de classificadores permitirá alcançar no futuro com maior desenvolvimento e estudo da ferramenta de “supervised machine learning”, um parâmetro que definira se a pessoa analisada é cardíaca ou não cardíaca. Isso será útil para entidades que busquem formas mais eficientes de identificar problemas cardíacos.

REFERÊNCIAS

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