UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

Daniel Ramos Correia dos Santos

Detecção de doença cardiovascular ou diabetes utilizando machine learning

Recife

2024

Daniel Ramos Correia dos Santos

Detecção de doença cardiovascular ou diabetes utilizando machine learning

Artigo apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Estatística e Informática

Curso de Bacharelado em Sistemas de Informação

Orientador: Gabriel Alves de Albuquerque Junior

Recife

2024

Daniel Ramos Correia dos Santos

Detecção de doença cardiovascular ou diabetes utilizando machine learning

Artigo apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Aprovada em: 07 de Março de 2024.

BANCA EXAMINADORA

Gabriel Alves de Albuquerque Junior (Orientador)
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

Cleviton Vinícius Fonsêca Monteiro
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

Agradecimentos

Agradeço primeiramente a Deus, que me concedeu força, sabedoria e perseverança para concluir este trabalho. Sua graça e orientação estiveram presentes em cada etapa desta jornada acadêmica, e por isso sou profundamente grato.

A minha mãe, Aderita, meus avós Miguel e Maria, minha irmã Danielly e toda minha família pelo o incentivo durante o curso.

A minha noiva, Eliene, por toda a companhia e força que me forneceu.

Aos meus amigos que fiz durante a graduação, em especial Wellington, pelo apoio concedido.

Aos meus irmãos em cristo pelo apoio e orações.

Ao meu orientador, Gabriel, por todo o apoio e todos os conselhos durante o desenvolvimento do trabalho.

Detecção de doença cardiovascular ou diabetes utilizando machine learning

Daniel R. C. dos Santos [1], **Gabriel A. de A. Junior** [1]

[1]Departamento de Estatística e Informática – Universidade Federal Rural de Pernambuco
Rua Dom Manuel de Medeiros, s/n, - CEP: 52171-900 – Recife – PE – Brasil

daniel.correia@ufrpe.br, gabriel.alves@ufrpe.br

Resumo. *Doenças cardiovasculares e diabetes representam desafios significativos para a saúde pública, demandando abordagens eficazes de diagnóstico e prevenção. Este trabalho propõe uma abordagem baseada em modelos de machine learning para oferecer suporte a esses processos. A partir de uma base de dados da pesquisa nacional de saúde do IBGE, o estudo investigou como diferentes variáveis afetam a detecção dessas doenças. Utilizando algoritmos como Random Forest, XGBoost e SVM, foram desenvolvidos modelos preditivos. Os resultados demonstraram uma acurácia de 71.96% para o algorimto Random Forest na classificação de pacientes com doenças cardiovasculares e 72.26% na classificação de pacientes com diabetes. Também foi realizada através do método SHAP, análise das variáveis mais influentes, que revelaram alguns insights sobre os dados.*

Abstract. *Cardiovascular diseases and diabetes represent significant challenges for public health, requiring effective diagnostic and prevention approaches. This work proposes an approach based on machine learning models to support these processes. Using a database from the IBGE national health survey, the study investigated how different variables affect the detection of these diseases. Using algorithms such as Random Forest, XGBoost and SVM, predictive models were developed. The results demonstrated an accuracy of 71.96% for the Random Forest algorithm in classifying patients with cardiovascular diseases and 72.26% in classifying patients with diabetes. Analysis of the most influential variables was also carried out using the SHAP method, which revealed some insights into the data.*

1. Introdução

As doenças cardiovasculares são a principal causa de morte não só no Brasil, mas em todo o mundo. Todos os anos, milhares de brasileiros vão a óbito em decorrência dessas doenças [Eyken and Moraes 2009]. Algumas enfermidades no coração podem ser descobertas ao longo da vida ou logo nos primeiros anos, como as cardiopatias congênitas. As doenças cardiovasculares podem afetar o coração e os vasos sanguíneos, como a doença arterial coronariana, que envolve dor no peito e infarto agudo do miocárdio. Os principais fatores de risco para o desenvolvimento de doenças cardiovasculares são o tabagismo, o colesterol em excesso, pois podem se acumular e levar à formação de placas de gordura, hipertensão, obesidade, estresse, depressão e diabetes [Barreto and Nogueira 2013].

De acordo com [Ramos 2021], o Brasil ocupa a quinta posição global em termos de incidência de diabetes, contando com 16,8 milhões de adultos afetados (idade entre 20

e 79 anos), ficando atrás apenas da China, Índia, Estados Unidos e Paquistão. Projeções do Atlas do Diabetes da Federação Internacional de Diabetes (IDF) indicam que a estimativa para a incidência da doença em 2030 é de 21,5 milhões de casos no país. O diabetes tornou-se uma séria questão de saúde pública em escala mundial, com as projeções sendo superadas a cada avaliação subsequente. Em 2000, a estimativa global de adultos vivendo com diabetes era de 151 milhões, aumentando para 285 milhões em 2009, um crescimento de 88%. A ascensão da prevalência do diabetes em todo o mundo é impulsionada por uma complexa interação de fatores socioeconômicos, demográficos, ambientais e genéticos. O aumento constante está vinculado principalmente ao crescimento do diabetes tipo 2 e aos fatores de risco associados, como o aumento da obesidade, adoção de dietas não saudáveis e a falta de atividade física. No entanto, é relevante notar que os níveis de diabetes tipo 1, com início na infância, também estão em ascensão.

Diante da problemática apresentada, surge uma oportunidade de desenvolver uma ferramenta por meio de algoritmos de machine learning. O intuito é utilizar as informações disponíveis em nossa base de dados, como idade, gênero, prática de atividade física, consumo de bebidas alcoólicas, tabagismo, hábitos alimentares, entre outras características. Com a implementação desses algoritmos, a proposta é classificar indivíduos como saudáveis ou portadores de doenças cardiovasculares ou diabetes. Essa abordagem visa fornecer suporte aos profissionais médicos durante o processo de diagnóstico, otimizando e aprimorando a análise clínica dos pacientes.

1.1. Objetivos

O presente trabalho tem como objetivo geral aprimorar a eficiência no diagnóstico de doenças cardiovasculares ou diabetes, aplicando modelos de *machine learning* em uma base de dados proveniente da pesquisa nacional de saúde disponibilizada pelo IBGE.

Como objetivos específicos:

- Analisar os dados aplicando *Explainable AI* para melhor compreensão;
- Analisar como as doenças crônicas impactam a precisão na detecção de doenças cardiovasculares ou diabetes;
- Examinar como a prática regular de atividade física contribuem para a redução do risco de doenças cardiovasculares ou diabetes;
- Examinar como ingestão de bebidas alcoólicas, tabagismo afetam a classificação de pessoas com problemas cardíacos;
- Examinar como hábitos alimentares contribuem para redução do risco do diabetes;

O restante deste artigo está organizado da seguinte forma: a Seção 2 aborda estudos correlatos sobre a detecção de doenças cardiovasculares ou diabetes por meio de machine learning; a descrição da teoria adotada encontra-se na Seção 3; na Seção 4, são detalhadas as ferramentas e métodos empregados; os resultados experimentais estão na Seção 5. Finalmente, apresentamos as conclusões e sugestões para trabalhos futuros.

2. Trabalhos relacionados

Nesta seção iremos abordar resumidamente alguns artigos encontrados que possuem relação com o tema proposto. O trabalho apresentado por [GUIMARÃRES 2019] descreve a implementação de um software baseado em machine learning para identificação

de doenças cardíacas por meio de eletrocardiogramas (ECG). Utilizando a base de dados "*The PTB Diagnostic ECG Database*", o autor enfrentou desafios devido ao baixo número de amostras para algumas classes de doenças, o que dificultou o treinamento confiável dos métodos. Foram empregadas técnicas como *K-Nearest-Neighbors Classifier, Support Vector Machine, Decision Tree Classifier e Random Forest Classifier*. O artigo destaca dados estatísticos que demonstram como essas técnicas auxiliaram na detecção de predisposição a problemas cardíacos com eficácia. Entre os resultados, destacamos com altas taxas de acerto os métodos utilizados, como *Support Vector Machine* com 100% e *Decision Tree Classifier* com 99,85%. O autor atribui esse sucesso à correlação no dataset e ao aumento na quantidade de dados utilizados para treinamento. Neste trabalho diferentemente do trabalho de [GUIMARÃRES 2019], foram aplicados alguns algoritmos em comum, como por exemplo, o *Random Forest Classifier* e o *Support Vector Machine* com o diferencial da aplicação do algoritmo *XGBoost*. Ambos algoritmos aplicados a uma base que não possui dados relacionados a eletrocardiogramas.

O trabalho apresentado por [Santos 2022] aborda a análise preditiva da probabilidade de ocorrência de doença cardiovascular (DCV) com base em dados de pacientes. Utilizando um conjunto de dados do Kaggle com 70.000 registros, o estudo destaca a pressão arterial, colesterol, idade e índice de massa corporal (IMC) como fatores com maior correlação ao risco de DCV. Diversos algoritmos de *machine learning*, incluindo *Decision Tree, Random Forest, Support Vector Machine, K-Nearest-Neighbors e Logistic Regression*, foram aplicados à análise. O modelo *Random Forest* destacou-se com uma acurácia de 80%, porém, o autor ressalta que este valor não é ideal para um modelo de classificação na área da saúde. Este resultado obtido na acurácia ocorreu devido a parametrização do algoritmo e também a grande quantidade de dados existentes na base de dados explorada. Embora eficaz em identificar quem não possui doença cardíaca (92% de precisão), de acordo com o autor o modelo não alcança uma discriminação satisfatória para identificar casos positivos de DCV. Assim, embora o estudo tenha identificado fatores de influência na DCV, o modelo proposto não é considerado aceitável para aplicação prática na área da saúde. Semelhantemente ao artigo [Santos 2022], neste trabalho a base de dados disponibilizada pelo IBGE possui algumas características em comum, como por exemplo, colesterol e idade. Além disso também foram aplicados os algoritmos *Random Forest* e o *Support Vector Machine* com o diferencial do algoritmo *XGBoost*.

O estudo de [Costalat and Tavares 2022] propõe uma abordagem moderna na detecção de doenças cardiovasculares, utilizando indicadores de saúde como idade, sexo, glicose e colesterol como entrada para sistemas de aprendizado de máquina. A pesquisa emprega a base de dados "*Heart Disease Data Set*"do *UCI Machine Learning Repository*, composta por 76 atributos, mas apenas 14 são utilizados nos algoritmos. Com 303 indivíduos (207 homens e 96 mulheres), a avaliação ocorre por meio de algoritmos de aprendizado supervisionado, incluindo *K-Nearest Neighbours, Decision Tree, Logistic Regression e Voting Classifier*. Os resultados indicam que os classificadores baseados em *Logistic Regression* e *K-Nearest Neighbours* alcançam uma acurácia de 80,26%, superando a *Decision Tree*, que atinge 73,68%. O *Voting Classifier* apresenta uma acurácia de 82,89%, sugerindo que a combinação de algoritmos de classificação pode aprimorar a capacidade preditiva. Este trabalho semelhantemente ao trabalho de [Costalat and Tavares 2022], possui uma base de dados envolvendo algumas características dos pacientes, porém como diferencial foram aplicados aplicados algoritmos distintos.

O trabalho apresentado por [Santos 2023] também aborda as Doenças Cardiovasculares (DCV) como a principal causa de morte global, sendo crescentemente diagnosticadas em jovens devido a hábitos não saudáveis. Destaca a importância da detecção precoce por meio de Eletrocardiograma (ECG) e menciona o desenvolvimento do Aprendizado de Máquina (AM) como ferramenta para interpretação do ECG. O trabalho propõe o uso de Aprendizado Profundo (AP), especificamente redes neurais *Transformer* e *Long Short-Term Memory* (LSTM), para classificação de diagnósticos de DCV. Os resultados experimentais indicam que a LSTM superou o *Transformer* em termos de precisão, revocação e *f1-score*, alcançando valores entre 70% e 84%, enquanto o *Transformer* obteve valores menores, até 48% em precisão, 62% em revocação e 36% em *f1-score*. Este trabalho diferentemente do trabalho de [Santos 2023], foram aplicados algoritmos distintos de *machine learning* e a base de dados explorada não possui dados relacionados a eletrocardiogramas.

O trabalho de [Cardozo 2022] aborda a questão do diagnóstico do Diabetes Mellitus (DM), indicando que, embora os exames tradicionais, como glicose em jejum (FPG) e hemoglobina glicada (HbA1c), sejam comuns, discrepâncias podem surgir devido às variações diárias no FPG. O estudo propõe a combinação de exames laboratoriais de rotina com técnicas de *Machine Learning* para prever o HbA1c e identificar falsos negativos no FPG. Uma metodologia, incluindo métodos como *KNN, SVM, Naïve Bayes, Random Forest e ANN*, foi aplicada a dados de 201.338 pacientes. O modelo de rede neural artificial destacou-se na predição de HbA1c, alcançando sensibilidade, precisão e F1-Score de 78,1%, 78,7% e 78,4%, respectivamente. Na detecção de falsos negativos de FPG, o modelo de regressão ANN apresentou aumento significativo na identificação de diabetes (16,6%) e pré-diabetes (35%). Os resultados sugerem que modelos de *Machine Learning* podem prever HbA1c a partir de exames de rotina, auxiliando no diagnóstico de diabetes. O ajuste dos valores de FPG pode melhorar a concordância com HbA1c, indicando a utilidade potencial desses modelos na rotina de exames laboratoriais e na sugestão de exames adicionais de confirmação. Este trabalho diferentemente do trabalho de [Cardozo 2022], não será focado apenas na questão do diabetes, como também nas doenças cardiovasculares, explorando uma base de dados composta por caracteristicas que diferem de glicose em jejum e hemoglobina glicada. Embora, utilizemos algoritmos em comum, como o *SVM* e o *Random Forest*.

O trabalho apresentado por [Araujo et al. 2022] aborda como o diabetes é considerada uma das principais crises de saúde do século 21, apresentando um aumento significativo nos últimos anos. Desta forma, o diagnóstico precoce é crucial, mas enfrenta desafios devido à sutileza dos sintomas iniciais. Este estudo propôs validar um método automatizado de análise de sintomas para detectar o risco de diabetes em estágios iniciais. Utilizando uma rede neural *Extreme Learning Machine (ELM)* com seleção de *features* por algoritmo genético, os resultados foram comparados com alguns algoritmos, destacando uma acurácia média de 98,64% para a ELM. Neste trabalho diferentemente do trabalho de [Araujo et al. 2022], foram aplicados algoritmos distintos e também engenharia de características como seleção de *features* para composição da base de dados.

De acordo com o trabalho apresentado por [Aquino et al. 2021], o aumento alarmante da diabetes tipo 2 globalmente, classificada como uma epidemia pela Organização Mundial de Saúde, impacta significativamente a vida social e econômica. O estudo utiliza dados da Vigitel (2006-2019) e a técnica *Random Forest* para identificar os 15 fatores

de risco mais relevantes entre 80 variáveis para o diabetes tipo 2. Modelos de regressão *Logit* e *Probit* foram usados para estimar os efeitos e assim conseguir identificar quais os 15 fatores de risco mais relevantes para a doença. As variáveis de idade, peso, índice de massa corporal, pressão alta, estado de saúde e tabagismo são destacados como fatores significativos que aumentam a probabilidade da doença. Neste trabalho diferentemente do trabalho de [Aquino et al. 2021], foi realizada a aplicação de algoritmos distintos, pórem a base de dados analisada contém semelhanças de características.

3. Referencial teórico

Nesta seção, apresentaremos uma introdução aos conceitos fundamentais, técnicas e temas discutidos ao longo deste artigo. Compreender esses elementos é essencial para alcançar uma compreensão plena da proposta deste trabalho.

3.1. Machine learning

Machine learning é um subconjunto da inteligência artificial (IA). Especificamente, se concentra no desenvolvimento de algoritmos e técnicas que permitem o aprendizado e o aperfeiçoamento de sistemas com base em dados. Esses algoritmos operam mediante a criação de modelos a partir de conjuntos de dados de entrada, extraídos de amostras individuais ou conjuntos diversos, com o propósito de realizar previsões ou tomar decisões orientadas pelos dados, em vez de simplesmente aderir a instruções programadas que sejam rígidas e estáticas. Técnicas baseadas em *machine learning* têm sido aplicadas com sucesso em diversos campos, desde reconhecimento de padrões, visão computacional, engenharia de naves espaciais, finanças, entretenimento e biologia computacional até aplicações biomédicas e médicas [El Naqa and Murphy 2015].

Existem alguns tipos de algoritmos de *machine learning* com abordagens distintas, cada uma com suas aplicações específicas, dentre eles estão o aprendizado não supervisionado, onde os algoritmos são adotados para analisar e agrupar conjuntos de dados sem rótulos. O objetivo é descobrir padrões intrínsecos nos dados, sem qualquer controle humano. Algoritmos de agrupamento, como *k-means*, são exemplos dessa abordagem [Hastie et al. 2009]. Outro tipo é o aprendizado por reforço, os agentes aprendem a tomar decisões através da interação com um ambiente. Eles recebem *feedback* na forma de recompensas ou penalidades, otimizando suas ações para maximizar recompensas ao longo do tempo. Este paradigma é frequentemente aplicado em jogos e robótica [Sutton and Barto 2018].

No aprendizado semi-supervisionado combina elementos da aprendizagem supervisionada e não supervisionada. O modelo é treinado em um conjunto de dados que contém tanto exemplos rotulados quanto não rotulados. Isso é útil quando rotular dados é caro ou demorado [Chapelle et al. 2006]. Também temos, a aprendizagem por transferência que utiliza conhecimento adquirido em uma tarefa para melhorar o desempenho em uma tarefa relacionada. Essa prática é valiosa quando há escassez de dados para a tarefa-alvo [Pan and Yang 2009].

Por ultimo aprendizado supervisionado, que será o tipo abordado neste estudo, onde os algoritmos são treinados em um conjunto de dados rotulado, onde as entradas estão associadas a rótulos ou saídas desejadas. Um exemplo é o uso de classificadores para prever categorias de dados com base em exemplos previamente etiquetados

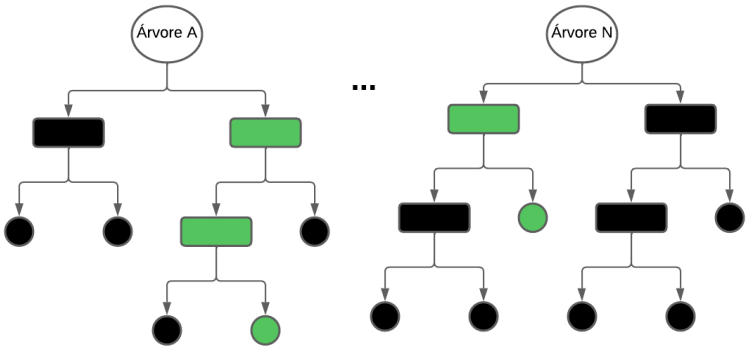
[Mitchell 1997]. Para ilustrar, consideremos o exemplo da base de dados selecionada para este estudo. A base de dados apresenta características específicas, uma coluna que indica se a pessoa foi diagnosticada ou não com um problema cardíaco e uma coluna que indica se a pessoa possui ou não diabetes. Com base nessas informações, o modelo é treinado usando os dados de teste, com o objetivo de realizar previsões que sejam mais precisas possível.

Quando falamos em modelos de aprendizagens de máquina, existem distintas abordagens, sendo as duas principais classificação e regressão. Os modelos de regressão têm como objetivo a previsão de valores dentro de um espectro contínuo, enquanto os modelos de classificação, foco deste estudo, referem-se aqueles em que os dados a serem utilizados pelo modelo estão devidamente rotulados, apresentando seus respectivos rótulos. Nesse tipo de cenário, cada rótulo corresponde a uma classe distinta, e o objetivo é que o algoritmo de classificação, ao realizar previsões, seja capaz de atribuir a previsão a uma das classes existentes. No contexto deste estudo, as previsões resultam em duas classes possíveis: determinar se uma pessoa pode ou não ter uma doença cardiovascular ou diabetes, com base nos dados da pesquisa respondida por cada individuo que compõem a base de dados.

3.1.1. *Random Forest*

O Random Forest [Breiman 2001], também conhecido como Floresta Aleatória, é um método de aprendizado de máquina empregado na resolução de desafios relacionados à classificação ou regressão. Este algoritmo fundamenta-se no conceito de árvore de decisão e opera ao criar uma floresta de forma aleatória. Essa floresta resultante é uma composição (*ensemble*) de múltiplas árvores de decisão, frequentemente treinadas utilizando o método de agregação *bootstrap (bagging)*.

Figura 1. Várias Árvores de Decisão



Fonte: O autor.

Os nós das árvores são criados a partir das características (*features*) do conjunto de dados. O nome *"Random Forest"*, do algoritmo, faz todo o sentido quando se pensa em seu funcionamento, pois *Random* significa "aleatório", e denota o comportamento do algoritmo ao selecionar subconjuntos de *features* e montar mini árvores de decisão. *Forest* significa"floresta"já que são geradas ao longo da resolução do problema mini árvores de

decisão.

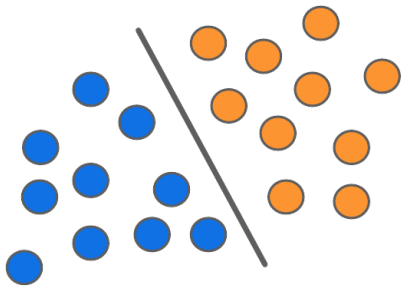
O algoritmo *Random Forest* se alinha de maneira ideal com os objetivos do projeto. A essência proposta é solucionar uma problemática de classificação, especificamente, estabelecer correlações nos dados fornecidos por meio dos atributos das tabelas contidas na base de dados.

3.1.2. *Support Vector Machine* (SVM)

O SVM ou *Support Vector Machine*, é um algoritmo de aprendizado de máquina utilizado para classificação e regressão. Desenvolvido por Vapnik e Cortes [Cortes and Vapnik 1995], seu fundamento principal é a identificação de um hiperplano ótimo que separa eficientemente as diferentes classes no espaço de características. Este algoritmo destaca-se por sua capacidade de lidar com conjuntos de dados complexos e de alta dimensionalidade. Os SVMs demonstram eficácia em diversos campos, incluindo reconhecimento de padrões, bioinformática, finanças e diagnóstico médico. A abordagem baseada em maximização da margem entre as classes proporciona uma maior generalização do modelo, reduzindo o risco de *overfitting*.

Para este estudo será utilizado o SVM linear. Podemos analisar o uso do SVM com o seguinte exemplo, na Figura 2 é possível observar duas classes de objetos, azul ou laranja. Essa linha que os separa define o limite em que se encontram os pontos azuis e os pontos laranjas. Ao entrarem novos objetos na análise, estes serão classificados como laranjas se estiverem à direita e como azuis caso situam-se à esquerda. Neste caso, conseguimos separar, por meio de uma linha, o conjunto de objetos em seu respectivo grupo, o que caracteriza um classificador linear.

Figura 2. Classes de objetos do SVM



Fonte: O autor.

3.1.3. XGBoost

O algoritmo de aprendizado de máquina XGBoost (*Extreme Gradient Boosting*) pertence à classe de algoritmos *ensemble*, e destaca-se por sua eficácia em diversas tarefas, como classificação e regressão. Desenvolvido por Tianqi Chen [Chen and Guestrin 2016], o *XGBoost* se tornou amplamente utilizado devido à sua capacidade de oferecer resultados precisos e eficientes em conjuntos de dados complexos.

Este algoritmo utiliza uma abordagem de *boosting*, onde modelos mais fracos podem ser aprimorados para torna-se mais assertivo. Um modelo mais fraco é aquele cuja perfomance é no minimo, um pouco melhor que a probabilidade aleatória [Gregório 2018]. O processo de treinamento do XGBoost envolve a construção sequencial de árvores de decisão, onde cada nova árvore visa corrigir os erros residuais das árvores anteriores. O termo "*gradient boosting*" refere-se à otimização dos parâmetros do modelo através do gradiente da função de perda, o que contribui para uma convergência mais rápida e eficaz durante o treinamento [Friedman 2001]. Para este estudo será adotado o *XGBClassifier*, que é frequentemente utilizado para prever a classe de um determinado exemplo em um conjunto de dados.

3.1.4. Explainable AI

Explainable AI [Gunning et al. 2019] é um conjunto de processos e métodos que conseguem dispor informações sobre os dados de uma forma mais compreensível, referente a suas interpretações e processos de tomada de decisão. Uma técnica muito útil para isso é o SHAP (*SHapley Additive exPlanations*), que para explicar como o modelo de *machine learning* funciona utiliza uma abordagem baseada em teoria de jogos, que busca prever e entender como as decisões individuais afetam o resultado geral de um jogo, levando em conta as preferências e estratégias dos jogadores envolvidos. Na Seção 5, iremos aplicar esta técnica.

Este é um campo que tem ganhado destaque, impulsionado pela crescente complexidade dos modelos de *machine learning*, como as redes neurais profundas e os algoritmos de aprendizado de máquina mais avançados. Estes, por vezes, são considerados caixas-pretas devido à sua dificuldade de interpretação. Essa falta de transparência representa um desafio significativo em áreas onde a compreensão das decisões é fundamental, tais como medicina, sistema judiciário e finanças. Alguns membros da Organização das Nações Unidas (ONU) e União Europeia (UE) possuem políticas públicas internacionais implementadas ou em desenvolvimento, para regular a inteligência artificial. O documento exige que medidas sejam tomadas para garantir que a IA seja projetada, desenvolvida e aplicada de forma a proteger os direitos humanos, a democracia e o estado de direito conforme destaca [Melo et al. 2022].

Explainable AI é importante para análise dos dados e adoção de modelos de *machine learning* em áreas sensíveis, possibilitando que especialistas compreendam os dados fornecidos, as recomendações ou até mesmo decisões tomadas pelos sistemas de inteligência artificial.

3.2. KDD

O *Knowledge Discovery in Databases* (KDD), ou descoberta de conhecimento a partir de dados, é definida como o processo sequencial de extração de dados ou conhecimento de grandes conjuntos de dados [Fayyad et al. 1996]. Este processo é caracterizado por sua natureza iterativa e interativa, abrangendo várias etapas que podem ser repetidas conforme necessário. Ele conta com a participação de profissionais especializados, como analistas de dados, e dos usuários finais.

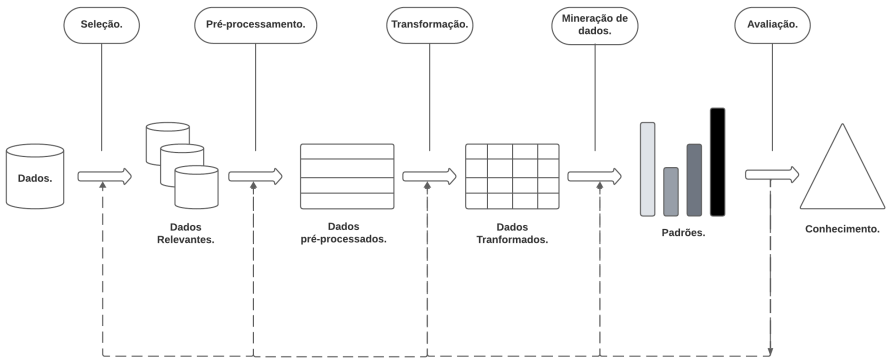
Conforme trabalho apresentado por [Fayyad et al. 1996], o processo de Desco-

berta de Conhecimento em Bases de Dados ou KDD, conforme demonstrado na Figura 3 compreende cinco fases: seleção dos dados, pré-processamento dos dados, transformação dos dados, mineração dos dados e interpretação. Na fase de seleção dos dados, os conjuntos a serem analisados são escolhidos, e perguntas como escopo, resultados esperados e consideração de dados qualitativos ou quantitativos são definidas. A segunda fase abrange o pré-processamento e aprimoramento dos dados, realizando operações para tratamento dos *outliers*, como eliminar ruídos, lidar com campos incompletos ou remove-los. A terceira fase, chamada de Transformação e Redução de Dados, padroniza o conjunto bruto, utilizando abordagens como discretização e redução de dimensionalidade.

A quarta fase, Mineração de Dados, envolve a geração de informações úteis a partir dos dados, utilizando técnicas como classificação, *clustering* e regressão. A escolha dos algoritmos nesta fase é crucial e depende do problema definido, por exemplo, a escolha de um algoritmo de classificação para prever doenças crônicas médicas. A etapa final, Interpretação dos Dados, é essencial para compreender e garantir a aplicabilidade prática dos resultados obtidos na fase de Mineração de Dados. Analisando minuciosamente os resultados em comparação com os objetivos iniciais, o conhecimento assimilado torna-se um recurso valioso para o processo de tomada de decisões em organizações.

Como é demonstrada na Figura 3 que ilustra o processo de descoberta de conhecimento, ou seja, as cinco etapas citadas anteriormente a respeito do KDD.

Figura 3. Processos de KDD



Fonte: O autor.

3.3. Técnicas computacionais na aréa da saúde

O estudo [Balbinot 2014] afirma que “mais pessoas morrem anualmente, pelo mundo, devido a doenças cardiovasculares do que por qualquer outra causa de morte.”. Atualmente, está em constante ascensão o movimento para integrar as inovações da computação ao campo da medicina, sendo o *machine learning* uma das abordagens proeminentes nesse cenário. Em vista da prevalência das doenças cardíacas, assistimos a uma crescente série de estudos voltados à aplicação do *machine learning* neste domínio, com o objetivo de desenvolver modelos capazes de antecipar a ocorrência de doenças cardíacas em pacientes. Essa capacidade preditiva possibilitaria intervenções preventivas e proativas no combate a esses problemas de saúde.

Quando falamos em diabetes, o diabetes mellitus figura entre as principais doenças crônicas não transmissíveis. Entre seus diversos tipos, o diabetes tipo 2 destaca-se como

a mais prevalente em adultos com mais de 40 anos, sendo estreitamente associada ao sedentarismo e à obesidade. Em muitos casos, o diagnóstico ocorre somente após o surgimento de complicações graves. Devido à complexidade da enfermidade, observa-se uma crescente aplicação de processos computacionais inteligentes para auxiliar os profissionais de saúde no diagnóstico precoce dessa condição, conforme destacado no artigo [Maciel 2020].

4. Abordagem proposta

Nesta seção iremos explorar sobre como foram realizadas as diferentes etapas do projeto utilizando os métodos do KDD, de forma que os conceitos, apresentados no referencial teórico, sejam detalhados e demonstrados de forma prática.

4.1. Seleção de dados

Inicialmente, com o propósito de conduzir uma pesquisa na área da saúde, optou-se por utilizar um conjunto de dados ou dataset proveniente da Pesquisa Nacional de Saúde (PNS), conduzida em colaboração com o IBGE e disponibilizada no site [de Geografia e Estatística 2022]. Esse conjunto de dados engloba informações obtidas a partir de entrevistas realizadas com participantes da pesquisa, abrangendo variáveis como idade, gênero, prática de atividade física, consumo de bebidas alcoólicas, tabagismo, presença de diabetes, níveis de colesterol hábitos alimentares e outras características. Entre as colunas fornecidas por esse conjunto de dados, encontra-se uma que representa a indicação de presença ou ausência de problemas cardíacos em cada participante da pesquisa e outra que indica se o participante tem diabetes ou não, o que possibilita a realização de diversas análises e a aplicação de algoritmos de aprendizado de máquina, visando antecipar a identificação de doenças cardíacas, mesmo antes que sintomas clínicos se manifestem.

Este conjunto de dados compreende cerca de 279 mil entrevistados e contempla 816 perguntas, abordando vários tópicos conforme demonstrados com os metadados nas Tabelas 1 e 2. Dentre esse tópicos, para o desenvolvimento deste trabalho iremos dar ênfase aos hábitos alimentares, dados clínicos entre outras variáveis conforme demonstrado na Tabela 4 em conjunto com o detalhamento destas perguntas selecionadas demosntrados nas Tabelas 5 e 6. Esse conjunto de dados são fundamentais para o nosso estudo visando a previsão de doenças cardíacas ou diabetes. Importante ressaltar que o conjunto de dados é de acesso público e está disponível para download no website do IBGE [de Geografia e Estatística 2022].

4.2. Pré-processamento

Nesta seção, será demonstrada toda a etapa de pré-processamento que foi feita. Após a seleção das bases de dados, procedeu-se à elaboração de um documento no *Google Colaboratory*, onde foram importadas diversas bibliotecas destinadas a possibilitar a análise de dados e a geração de gráficos. Essas bibliotecas incluem pandas [McKinney 2010] utilizado na manipulação e analise dos dados, *seaborn* [Waskom et al. 2020] e matplotlib [Hunter 2007] utilizadas para criação de gráficos. Por fim, o *YData-profiling* que tem como seu principal objetivo fornecer uma análise exploratória de dados como uma solução consistente e rápida. Através da aplicação das bibliotecas mencionadas, gráficos foram criados e uma análise exploratória completa do conjunto de dados foi conduzida.

Tabela 1. Metadados - parte 1

Código	Tópicos
A	Informações do domicílio: • espécie do domicílio; • Para os domicílios particulares permanentes: – material das paredes, pisos e cobertura do prédio; – número de cômodos e de dormitórios; – forma de abastecimento de água e esgotamento sanitário; – destino do lixo; – existência de bens duráveis; – existência de animais de estimação.
B	Visitas domiciliares de Equipe de Saúde da Família e Agentes de Endemias
C	Características gerais dos moradores: • sexo; • idade; • cor ou raça; • condição no domicílio e na família.
D	Características de educação: • alfabetização; • escolarização; • série e grau frequentados pelos estudantes; • última série concluída, grau correspondente e conclusão do curso para pessoas que não são estudantes.
E	Características de trabalho e rendimento: • condição de atividade e de ocupação na semana de referência e no período de referência de 365 dias; • ocupação, atividade, posição na ocupação, categoria do emprego no trabalho principal da semana de referência, no trabalho principal do período de 365 dias ou no último trabalho do período de referência de cinco anos; • rendimento e horas trabalhadas no trabalho principal e em outros trabalhos da semana de referência; • Procura de trabalho; • Outras formas de trabalho: Cuidado de pessoas e afazeres domésticos.

Durante essa análise, foi possível identificar discrepâncias nos dados contidos na base escolhida. Esses pontos discrepantes serão destacados no decorrer desta seção. Além disso, insights valiosos foram obtidos, os quais foram devidamente discutidos na Seção 5 deste artigo, uma vez que um exame mais aprofundado do conjunto de dados foi realizado.

Tabela 2. Metadados - parte 2

Código	Tópicos
F	Rendimentos
G	Pessoas com deficiência
I	Cobertura de Plano de Saúde
J	Utilização de Serviços de Saúde
K	Saúde dos indivíduos com 60 anos ou mais
L	Crianças com menos de 2 anos de idade
M	Informações para futuros contatos, características do trabalho e apoio social do morador selecionado
N	Percepção do estado de saúde
O	Acidentes: acidentes de trânsito, acidentes de trabalho
P	Estilos de Vida: hábitos de alimentação, prática de atividade física, uso de bebidas alcoólicas e fumo
Q	Doenças crônicas
R	Saúde da Mulher
S	Atendimento Pré-natal
Z	Paternidade e pré-natal do parceiro (Homens)
V	Violência
U	Saúde Bucal
T	Doenças Transmissíveis
Y	Atividade Sexual
AA	Relações e condições de trabalho
H	Atendimento médico e de saúde

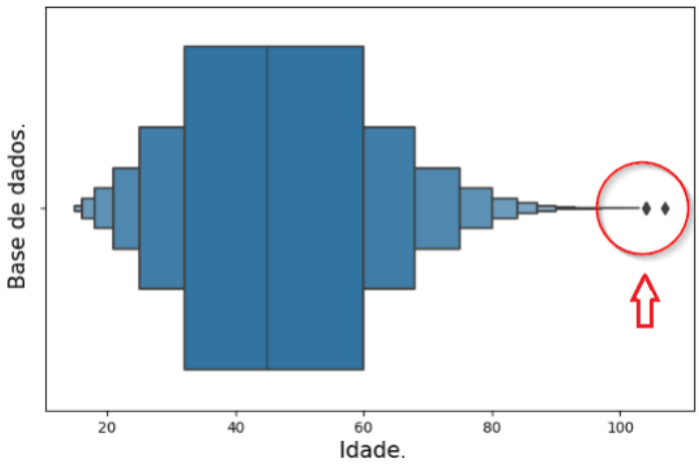
Na realização do pré-processamento dos dados, foi identificado a presença de outliers na coluna que representa a idade, onde alguns valores nas extremidades se destacavam em relação à maioria dos dados, como evidenciado na Figura 4. Na coluna que representa o peso, identificou-se valores discrepantes (outliers) nas faixas de 25 a 30 kg e 150 a 175 kg, como pode ser observado na Figura 5. Assim como identificamos valores discrepantes na coluna que representa a idade em relação ao conjunto principal, também identificamos ocorrências anômalas na coluna correspondente à altura. Foram registradas alturas abaixo de 120 centímetros e acima de 200 centímetros, conforme ilustrado na Figura 6.

Durante a análise exploratória dos dados, foi identificado a presença de valores discrepantes, conhecidos como outliers, como, por exemplo, casos em que a altura de indivíduos ultrapassa os 220 centímetros. Porém, estes dados foram mantidos para a análise, devido a sua pequena quantidade e pouca correlação com o tema proposto.

A engenheria de características também foi utilizada ajudando a destacar padrões nos dados e a melhorar a capacidade de generalização dos modelos para os novos conjuntos de dados, melhorarando o desempenho dos modelos preditivos e auxiliando para evitar que ocorra o vazamento de dados no modelo, como será demonstrado na etapa de mineração de dados.

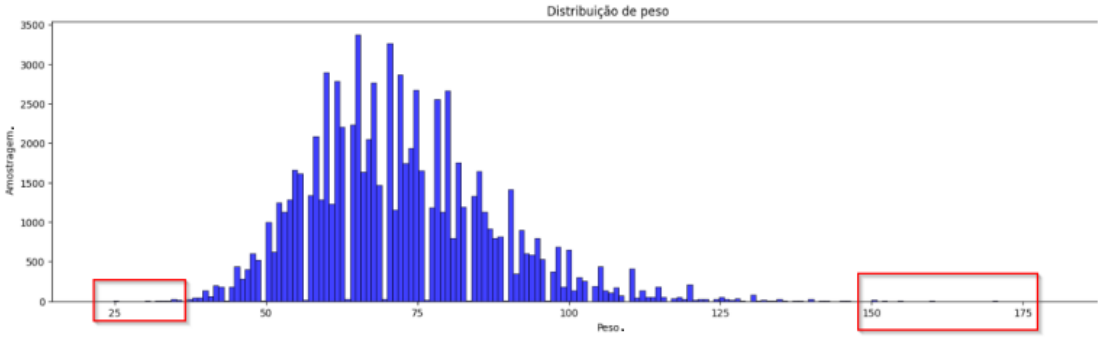
Além disso, foi observado um desbalanceamento no dataset, uma vez que o

Figura 4. Gráfico da idade



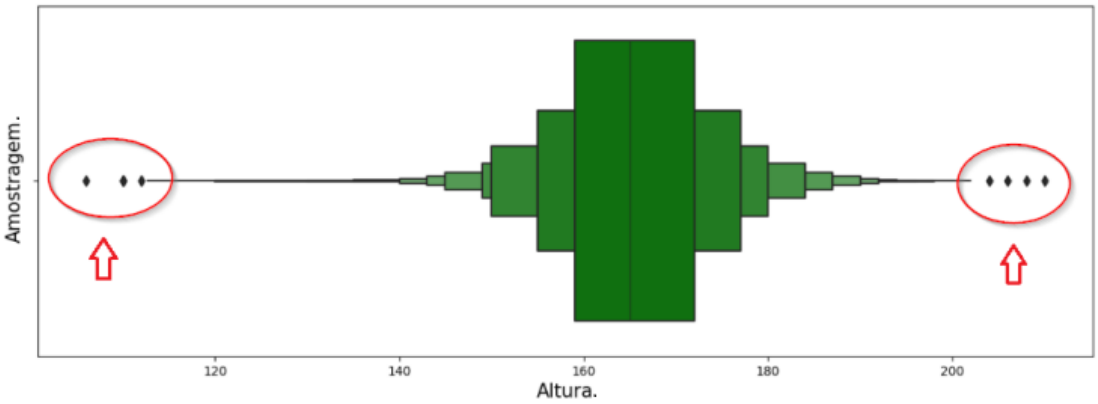
Fonte: O autor.

Figura 5. Gráfico do peso



Fonte: O autor.

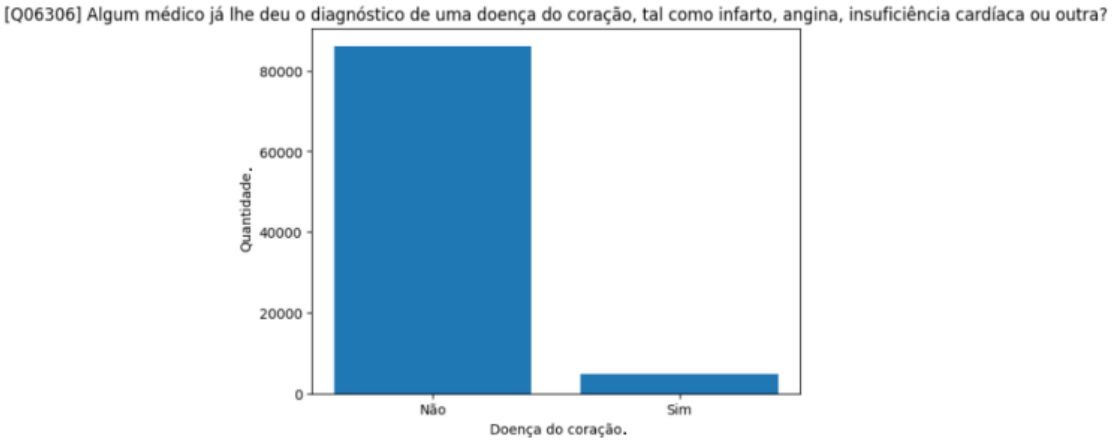
Figura 6. Gráfico da altura



Fonte: O autor.

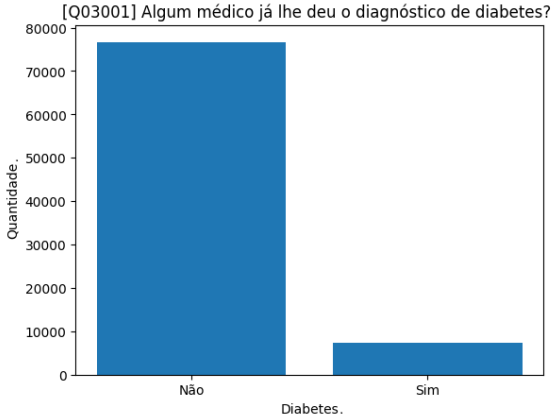
número de indivíduos sem doença cardiovascular é significativamente maior do que o número de pessoas com doença cardiovascular, como ilustrado na Figura 7. O mesmo também vale para a variável do diabetes, como podemos observar na Figura 8.

Figura 7. Gráfico de doença cardiovascular



Fonte: O autor.

Figura 8. Gráfico de diabetes



Fonte: O autor.

Dito isto, para aplicação dos algoritmos é fundamental efetuar o balanceamento dos dados a fim de promover um treinamento do modelo de forma mais equilibrada, o que, por sua vez, contribuirá para a obtenção de resultados aprimorados.

4.3. Transformação

A base de dados é composta por um conjunto de 279.382 entrevistados e apresenta 816 colunas que referem-se às perguntas feitas durante as entrevistas. No decorrer da análise, identificou-se que 90.846 indivíduos responderam afirmativamente à pergunta central do estudo, que era: "Algum médico já lhe deu o diagnóstico de uma doença do coração, tal como infarto, angina, insuficiência cardíaca ou outra?", como podemos observar na Figura 7, dos quais 86.114 responderam 'não' e 4.732 responderam 'sim'. Já quando

olhamos para os dados relacionados a identificação do diabetes na base de dados temos 84.067 pessoas que responderam a esta pergunta, dos quais 76.694 responderam 'não' e 7.373 responderam 'sim', como podemos observar na Figura 8.

No processo de preparação do conjunto de dados, as variáveis categóricas foram tratadas de acordo com as especificações detalhadas no dicionário de dados. Essa etapa é crucial para garantir que o modelo seja capaz de compreender e utilizar eficientemente informações qualitativas. As técnicas empregadas nesse contexto foram o *Ordinal Encoding* e o *One-Hot Encoding*.

O *Ordinal Encoding* foi utilizado nas variáveis categóricas ordinais, ou seja, aquelas com uma ordem específica. Nesse caso, cada categoria é mapeada para um número inteiro, preservando a relação de ordem entre elas. Por exemplo, a variável "[P01001] Em geral, o(a) Sr(a) costuma comer esse tipo de verdura ou legume:" com as categorias: "Uma vez por dia (no almoço ou no jantar)", "Duas vezes por dia (no almoço e no jantar)", "Três vezes ou mais por dia" e "Ignorado", cada categoria são mapeadas para um valor numérico correspondente, como "1", "2", "3" e "9", respectivamente. Em relação as duas variáveis alvo deste estudo, foi aplicada a codificação *One-Hot Encoding*, onde "0" reperesenta não ter a doença e "1" que possui a doença, seja diabetes ou doenças cardiovasculares. Esse tratamento permitiu que as variáveis categóricas fossem incorporadas de maneira adequada aos algoritmos de aprendizado de máquina, garantindo que a informação contida nessas características fosse devidamente interpretada. Os valores ausentes que existiam na base de dados foram substituidos por "9" ou "999", conforme definido no dicionário original de dados, onde esses números representam "Ignorado". Em particular, ao tratar os valores ausentes de peso e altura, foi aplicado o método de interpolação entre os dados disponíveis, seguido pelo arredondamento dos valores para uma casa decimal. A interpolação é um método matemático que consiste em estimar valores intermediários entre pontos conhecidos. No caso dos dados de peso e altura, utilizamos a interpolação linear, onde os valores ausentes são preenchidos com valores que são uma média ponderada dos pontos de dados vizinhos. Após a interpolação, os valores resultantes podem conter múltiplas casas decimais. Para simplificar e tornar os dados mais legíveis, foi realizado o arredondamento para uma casa decimal, o que significa que os valores foram ajustados para conter apenas uma casa decimal após o ponto. O conjunto de dados resultante, após a aplicação destes tratamentos, está agora pronto para ser utilizado na construção e treinamento de modelos preditivos.

4.4. Mineração de dados

Nesta seção, são inicializadas as técnicas de mineração de dados que são aplicadas quando deseja-se resolver problemas de classificação.

4.4.1. Escolha das features

Foi realizada a implementação do algoritmo de *Feature Importance* para a seleção das melhores *features* do *dataset*, ou seja, aquelas que mais contribuem para o resultado final do que deseja-se prever e estudar. Com base no resultado apresentado pelo algoritmo é possivel determinar algumas das *features* mais importantes e em conjunto com isso utilizando tecnica da engenharia de caracteristicas selecionar também *features* que não

possuem correlação alta para compor a base de dados, para evitar vazamento de dados e assim iniciar a preparação para executar os algoritmos de classificação.

Com a base de dados já composta por valores categóricos e como não possuía outliers visto que as repostas para as perguntas variam apenas dentro de uma escala já predeterminada de acordo com o dicionário de dados, foram selecionadas *features* para compor as análises, visto que a base no geral contava com 816 *features*. Foi utilizado também o conceito de *Feature Engineering* [Dong and Liu 2018] para a seleção das *features* que iriam compor o modelo, baseando-se na importância da feature para o modelo e também observando a sua correlação com a variavel alvo, para evitar features que possuiam alta correlação para não ocorrer vazamento de dados para o modelo. Apesar desta etapa ser um pouco exaustiva e manual, se torna necessária pois ajuda a melhorar o desempenho do modelo, interpretabilidade e redução da possibilidade de *overfitting*.

Dessa forma, a base de dados foi reduzida para 29 *features* referentes a alguns módulos do dicionário de dados, conforme demostrado na Tabela 3.

Tabela 3. Módulos do dicionário de dados

Identificação e Controle
Módulo C - Características gerais dos moradores
Módulo P - Estilos de vida
Módulo Q - Doenças crônicas

Na Tabela 4 é apresentado quais foram as *features* selecionadas para compor a base de dados. Já nas Tabelas 5 e 6 são demonstrados com mais detalhes as repostas para cada pergunta selecionada para base de dados.

A base de dados foi repartida em dois subconjuntos, o primeiro subconjunto refere-se à análise de doenças cardiovasculares no geral que possui como *target* a coluna "Algum médico já lhe deu o diagnóstico de uma doença do coração, tal como infarto, angina, insuficiência cardíaca ou outra?". Já o segundo subconjunto refere-se à análise especificamente sobre diabetes que possui como *target* a coluna "Algum médico já lhe deu o diagnóstico de diabetes?".

Como foi dito anteriormente, na Seção 5 serão demonstrados os resultados obtidos pelo algoritmo que selecionam as *features* mais importantes do *dataset*. O resultado retornado, corresponde ao índice de cada uma das colunas selecionadas como mais importante, no entanto, por meio de métodos da biblioteca "Pandas", se obtém o nome das *features* por meio de seus índices. Os resultados serão demonstrados, após a execução desta etapa.

4.4.2. Implementação e aplicação dos algoritmos de machine learning

Foram inicializadas pesquisas mais aprofundadas acerca do algoritmo citado no tópico anterior, o qual foi escolhido para realizar a classificação, e também iniciaram as implementações na linguagem de programação python, com o auxílio de algumas bibliotecas, as quais ajudam no uso do algoritmo.

Tabela 4. Colunas selecionadas para composição da base de dados

[V0001] Unidade da Federação
[C006] Sexo
[C008] Idade do morador na data de referência
[C009] Cor ou raça
[P00103] Peso - Informado (em kg) (3 inteiros e 1 casa decimal)
[P00403] Altura - Informada (em cm) (3 inteiros)
[P00901] Em quantos dias da semana, o(a) Sr(a) costuma comer pelo menos um tipo de verdura ou legume (sem contar batata, mandioca, cará ou inhame) como alface, tomate, couve, cenoura, chuchu, berinjela, abobrinha?
[P01001] Em geral, o(a) Sr(a) costuma comer esse tipo de verdura ou legume
[P01101] Em quantos dias da semana o(a) Sr(a) costuma comer carne vermelha (boi, porco, cabrito, bode, ovelha etc.
[P013] Em quantos dias da semana o(a) Sr(a) costuma comer frango/galinha?
[P015] Em quantos dias da semana o(a) Sr(a) costuma comer peixe?
[P02001] Em quantos dias da semana o(a) Sr(a) costuma tomar suco de caixinha/lata ou refresco em pó ?
[P01601] Em quantos dias da semana o(a) Sr(a) costuma tomar suco de fruta natural (incluída a polpa de fruta congelada)?
[P018] Em quantos dias da semana o(a) Sr(a) costuma comer frutas?
[P02002] Em quantos dias da semana o(a) Sr(a) costuma tomar refrigerante?
[P023] Em quantos dias da semana o(a) Sr(a) costuma tomar leite? (de origem animal: vaca, cabra, búfala etc.)
[P02501] Em quantos dias da semana o(a) Sr(a) costuma comer alimentos doces como biscoito/bolacha recheado, chocolate, gelatina, balas e outros?
[P02602] Em quantos dias da semana o(a) Sr(a) costuma substituir a refeição do almoço por lanches rápidos como sanduíches, salgados, pizza, cachorro quente, etc?
[P027] Com que frequência o(a) Sr(a) costuma consumir alguma bebida alcoólica?
[P034] Nos últimos três meses, o(a) Sr(a) praticou algum tipo de exercício físico ou esporte?
[P050] Atualmente, o(a) Sr(a) fuma algum produto do tabaco?
[Q03001] Algum médico já lhe deu o diagnóstico de diabetes?
[Q06306] Algum médico já lhe deu o diagnóstico de uma doença do coração, tal como infarto, angina, insuficiência cardíaca ou outra?
[Q068] Algum médico já lhe deu o diagnóstico de AVC (Acidente Vascular Cerebral) ou derrame?
[Q074] Algum médico já lhe deu o diagnóstico de asma (ou bronquite asmática)?
[Q079] Algum médico já lhe deu o diagnóstico de artrite ou reumatismo?
[Q092] Algum médico ou profissional de saúde mental (como psiquiatra ou psicólogo) já lhe deu o diagnóstico de depressão?
[Q120] Algum médico já lhe deu diagnóstico de câncer?
[Q124] Algum médico já lhe deu o diagnóstico de insuficiência renal crônica?

Antes dos algoritmos serem executados, na base de dados foram selecionadas as colunas que servirão como previsores e separadas da coluna que contém a classe. Isso foi realizado para que cada modelo possa aprender por meio dos previsores a como chegar em um dos resultados possíveis que somente é encontrado na coluna que contém a qual classe cada um dos registros pertence.

Com o objetivo de realizar as classificações necessárias para abordar as indagações propostas no início deste artigo, os algoritmos de *machine learning* escolhidos foram *Random Forest*, *Support Vector Machine* (SVM) e *XGBoost*. Esses algoritmos foram escolhidos por serem amplamente utilizados e estudados em uma variedade de problemas de classificação, incluindo diagnósticos médicos, e têm demonstrado consistentemente boas performances como observamos na Seção 2. Para aplicação dos algoritmos foram utilizados os parâmetros padrões fornecidos na importação da biblioteca, devido ao baixo tempo para execução do projeto, inviabilizando a configuração e testes dos parâmetros para refinamento do modelo. Na execução dos algoritmos nos subconjuntos para classificação de doença cardiovascular e do diabetes, foram utilizados a divisão em treino-teste. Ao dividir o conjunto de dados em duas partes, 70% dos dados foi utilizado para treinar o modelo, ou seja, ajustar os parâmetros do modelo com base nesses dados. Os outros 30% para testar o modelo, ou seja, avaliar o desempenho do modelo em dados não vistos durante o treinamento. Os critérios que serão utilizados para selecionar o melhor modelo será a acurácia em conjunto com os resultados da classificação. Após isso, entre os algotrimos escolhidos, será aplicado a análise SHAP apenas no modelo que atingir o melhor resultado.

5. Resultados

Nesta seção, será apresentado uma análise da aplicação dos algoritmos de *machine learning* na base de dados e as respostas para as perguntas que nortearam este estudo. Para a aplicação dos algoritmos a base de dados foi balanceada utilizando a técnica do *Undersampling*, que consiste em reduzir o número de instâncias da classe majoritária para equilibrar com a classe minoritária. Isso é feito removendo aleatoriamente exemplos da classe majoritária até que haja um equilíbrio entre as classes. No conjunto de dados, quando a variável alvo refere-se a doenças cardiovasculares reduziu-se a classe majoritária para igualarmos a classe minoritária restando 4.731 pessoas sem doenças cardiovasculares da classe “0” e 4.731 pessoas com doenças cardiovasculares da classe “1”, resultando em um total de 9.462. Em contrapartida, quando a variável alvo passa a ser o diabetes reduziu-se a classe majoritária para igualarmos a classe minoritária restando 7.373 pessoas sem diabetes representada pela classe “0” e 7.373 pessoas com diabetes representadas pela classe “1”, resultando em um total de 14.746. Após isso, para aplicação dos algoritmos, foi considerada a separação de 70% do conjunto para treino e 30% para teste. Também foram analisadas com o auxilio do método SHAP a contribuição e comportamento das variáveis apenas para o modelo que obteve maior acurácia.

5.1. Predição de doenças cardiovasculares

Aplicando o algoritmo *Random Forest*, configurado com 100 árvores e com foco e doenças cardiovasculares, foi obtido uma acurácia de 71.96%. Na matriz de confusão, a diagonal principal que são os casos em que o modelo previu corretamente a classe positiva

Tabela 5. Detalhamento das perguntas selecionadas - parte 1

Código Pergunta	Respostas
[V0001]	• 11 - Rondônia • 12 - Acre • 13 - Amazonas • 14 - Roraima • 15 - Pará • 16 - Amapá • 17 - Tocantins • 21 - Maranhão • 22 - Piauí • 23 - Ceará • 24 - Rio Grande do Norte • 25 - Paraíba • 26 - Pernambuco • 27 - Alagoas • 28 - Sergipe • 29 - Bahia • 31 - Minas Gerais • 32 - Espírito Santo • 33 - Rio de Janeiro • 35 - São Paulo • 41 - Paraná • 42 - Santa Catarina • 43 - Rio Grande do Sul • 50 - Mato Grosso do Sul • 51 - Mato Grosso • 52 - Goiás • 53 - Distrito Federal
[C006]	• 1 - Homem • 2 - Mulher
[C008]	• 000 a 130 - Idade (em anos) • 999 - Ignorado
[C009]	• 1 - Branca • 2 - Preta • 3 - Amarela • 4 - Parda • 5 - Indígena • 9 - Ignorado

Tabela 6. Detalhamento das perguntas selecionadas - parte 2

Código Pergunta	Respostas
[P00103]	• 1 a 599 - Quilogramas
[P00403]	• 1 a 299 - Centímetros • 999 - Ignorado
[P00901], [P01101], [P013], [P015], [P02001], [P01601], [P018], [P02002], [P023], [P02501], [P02602]	• 1 a 7 - Dias • 0 - Nunca ou menos de uma vez por semana • 9 - Ignorado
[P01001]	• 1 - Uma vez por dia (no almoço ou no jantar). • 2 - Duas vezes por dia (no almoço e no jantar). • 3 - Três vezes ou mais por dia. • 9 - Ignorado
[P027]	• 1 - Não bebo nunca • 2 - Menos de uma vez por mês • 3 - Uma vez ou mais por mês • 9 - Ignorado
[P034], [Q03001], [Q06306], [Q068], [Q074], [Q079], [Q092], [Q120], [Q124]	• 0 - Não • 1 - Sim • 9 - Ignorado
[P050]	• 0 - Não fumo atualmente • 1 - Sim, diariamente • 2 - Sim, menos que diariamente • 9 - Ignorado

e negativa, obtendo 995 acertos referente às pessoas que não tem doenças cardiovasculares e 1.048 acertos referente às pessoas que possuem doenças cardiovasculares. Quando olhamos para diagonal secundária, que são os casos em que o modelo previu incorretamente a classe negativa como positiva e positiva como negativa, observamos que o modelo obteve 424 erros classificando a classe negativa como positiva e 372 erros classificando a classe positiva como negativa.

Já os resultados da classificação, que fornece uma visão detalhada das métricas de desempenho para cada classe do conjunto de dados, podemos observar que a variável *f1-score* que traz uma média harmônica entre as variáveis *precision* e *recall* obteve 71% para classificação de pessoas sem a doença cardiovascular e 72% para classificação de pessoas com a doença cardiovascular, conforme exibido na Figura 9 que traz algumas informações do modelo.

Também, diante das análises realizadas, podemos observar as *Features Importan-*

Figura 9. Classification Report Doença Cardiovascular - Random Forest

	precision	recall	f1-score	support
0	0.73	0.70	0.71	1419
1	0.71	0.74	0.72	1420
accuracy			0.72	2839
macro avg	0.72	0.72	0.72	2839
weighted avg	0.72	0.72	0.72	2839

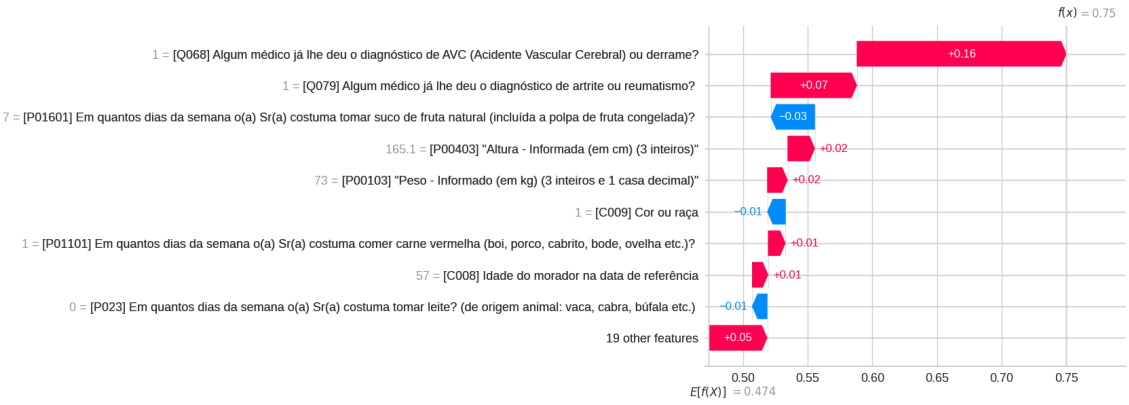
Fonte: O autor.

ces, que nos mostram a contribuição relativa de cada característica (ou *feature*) no nosso modelo. Aplicando esta análise, obtemos que "idade", "peso" e "altura" lideram o ranking, respectivamente, seguidos por "unidade federativa", "hábitos alimentares/estilo de vida" e "doenças crônicas". Dentre estas variáveis, a variável "idade" possui um destaque atingindo 19% de contribuição para o nosso modelo.

Foi analisado também o método SHAP (*SHapley Additive exPlanations*) que consiste em uma técnica de interpretação de modelos de *machine learning*, útil para entender como as características ou variáveis de entrada contribuem para as predições feitas pelo modelo.

Observando o valor SHAP absoluto, que nos mostra o quanto um único recurso afetou a previsão apenas na variável alvo, temos que o "Acidente Vascular Cerebral (AVC)/derrame" contribuiu mais, seguido por "artrite/reumatismo" em segundo lugar, as outras *features* tiveram contribuição, porém com valores bem baixos como podemos observar na Figura 10.

Figura 10. SHAP Absoluto - Doença Cardiovascular

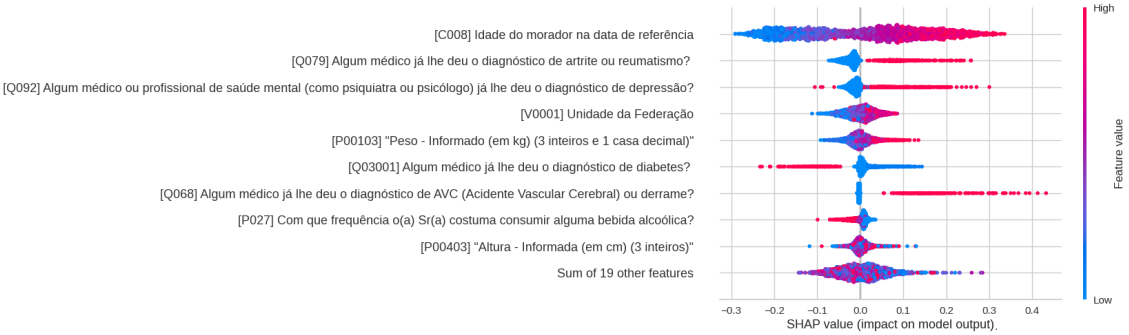


Fonte: O autor.

Para entender a importância ou contribuição das características para todo o conjunto de dados, podemos utilizar o gráfico do enxame de abelhas que é demonstrado na Figura 11.

Nela podemos observar os valores elevados (cor vermelha) e os valores mais bai-

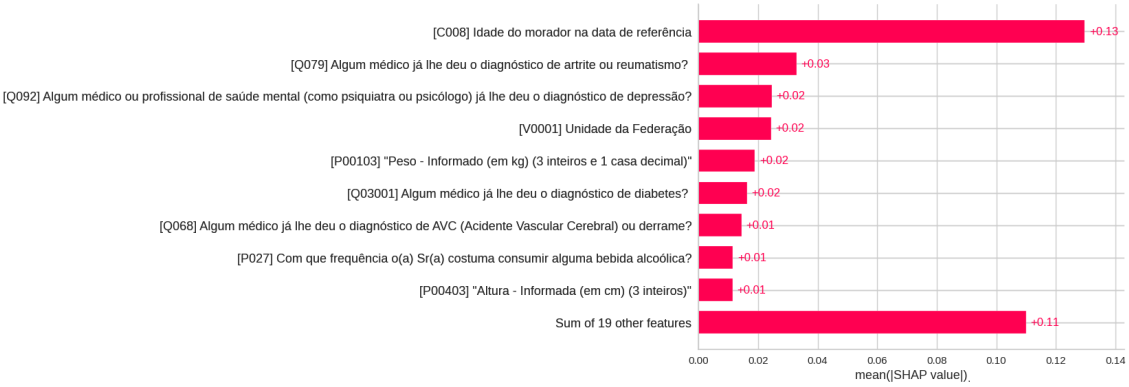
Figura 11. SHAP contribuição das características - Doença Cardiovascular



Fonte: O autor.

xos (cor azul) de cada *feature*. Também podemos observar a forma de contribuição de cada *feature* sendo negativa se estiver do lado esquerdo e positiva se estiver do lado direito. Quando olhamos para esta análise na nossa base de dados, vale a pena destacar que, de acordo com o gráfico "idades", "peso", "artrite/reumatismo", "depressão" e "AVC" têm contribuições positivas quando seus valores são altos, ou seja, neste conjunto de dados estas variáveis são um forte indicador de que pessoas com estas caracteristicas sejam acometidas por um doença cardiovascular. Em contrapartida as variáveis de "diabetes" e "consumo de bebida alcoólica" têm contribuições positivas quando seus valores são baixos, ou seja, o individuo que possui diabetes ou consome bebida alcoólica, neste conjunto de dados tem poucas chances de desenvolvimento de doença cardiovascular. Quando queremos verificar a importância das *features* levando em consideração o valor absoluto do SHAP, não importando se o recurso afeta a previsão de forma positiva ou negativa, temos as *features* ordenadas do maior para o menor efeito na previsão, conforme Figura 12.

Figura 12. SHAP Feature Importance - Doença Cardiovascular



Fonte: O autor.

Nela podemos observar como a "idade" dos entrevistados, "artrite" ou "reumatismo" e a "depressão" lideram o ranking e interferem, principalmente a "idade", na classificação de doenças cardiovasculares, seguido pelas demais *features*, porém com valores mais baixos e próximos.

Aplicando o algoritmo SVM do tipo linear e com foco em doenças cardiovasculares, foi obtido uma acurácia de 71.71%. Na matriz de confusão, a diagonal principal que são os casos em que o modelo previu corretamente a classe positiva e negativa, obteve 1.031 acertos referente às pessoas que não tem doenças cardiovasculares e 1.005 acertos referente às pessoas que possuem doenças cardiovasculares. Quando olhamos para diagonal secundária, que são os casos em que o modelo previu incorretamente a classe negativa como positiva e positiva como negativa, observamos que o modelo obteve 388 erros classificando a classe negativa como positiva e 415 erros classificando a classe positiva como negativa.

Nos resultados da classificação, conseguimos observar que a variável *f1-score* que traz uma média harmônica entre as variáveis *precision* e *recall* obteve 72% para classificação de pessoas sem a doença cardiovascular e 71% para classificação de pessoas com a doença cardiovascular, conforme exibido na Figura 13, que traz algumas informações do modelo.

Figura 13. Classification Report Doença Cardiovascular - SVM

	precision	recall	f1-score	support
0	0.71	0.73	0.72	1419
1	0.72	0.71	0.71	1420
accuracy			0.72	2839
macro avg	0.72	0.72	0.72	2839
weighted avg	0.72	0.72	0.72	2839

Fonte: O autor.

Aplicando o algoritmo XGBClassifier e com foco em doenças cardiovasculares, foi obtido uma acurácia de 68.68%. Na matriz de confusão, a diagonal principal que são os casos em que o modelo previu corretamente a classe positiva e negativa, obteve 992 acertos referente às pessoas que não tem doenças cardiovasculares e 958 acertos referente às pessoas que possuem doenças cardiovasculares. Quando olhamos para diagonal secundária, que são os casos em que o modelo previu incorretamente a classe negativa como positiva e positiva como negativa, observamos que o modelo obteve 427 erros classificando a classe negativa como positiva e 462 erros classificando a classe positiva como negativa.

Nos resultados da classificação, observamos que a variável *f1-score* obteve 69% para classificação de pessoas sem a doença cardiovascular e 68% para classificação de pessoas com a doença cardiovascular, conforme exibido na Figura 14.

5.2. Predição de diabetes

Aplicando o *random forest* no conjunto de dados para classificação do diabetes, também contendo 100 árvores na sua configuração, foi obtido uma acurácia de 72.26%. Na matriz de confusão, a diagonal principal que são os casos em que o modelo previu corretamente a classe positiva e negativa, obtendo 1.506 acertos referente às pessoas que não tem diabetes e 1.691 acertos referente às pessoas que possuem diabetes. Quando olhamos para

Figura 14. Classification Report Doença Cardiovascular - XGBoost

	precision	recall	f1-score	support
0	0.68	0.70	0.69	1419
1	0.69	0.67	0.68	1420
accuracy			0.69	2839
macro avg	0.69	0.69	0.69	2839
weighted avg	0.69	0.69	0.69	2839

Fonte: O autor.

diagonal secundária, que são os casos em que o modelo previu incorretamente a classe negativa como positiva e positiva como negativa, observamos que o modelo obteve 692 erros classificando a classe negativa como positiva e 535 erros classificando a classe positiva como negativa.

Já nos resultados da classificação, podemos observar que a variável *f1-score* que traz uma média harmônica entre as variáveis *precision* e *recall* obteve 71% para classificação de pessoas sem diabetes e 73% para classificação de pessoas com diabetes, conforme exibido na Figura 15, que traz algumas informações do modelo.

Figura 15. Classification Report Diabetes - Random Forest

	precision	recall	f1-score	support
0	0.74	0.69	0.71	2198
1	0.71	0.76	0.73	2226
accuracy			0.72	4424
macro avg	0.72	0.72	0.72	4424
weighted avg	0.72	0.72	0.72	4424

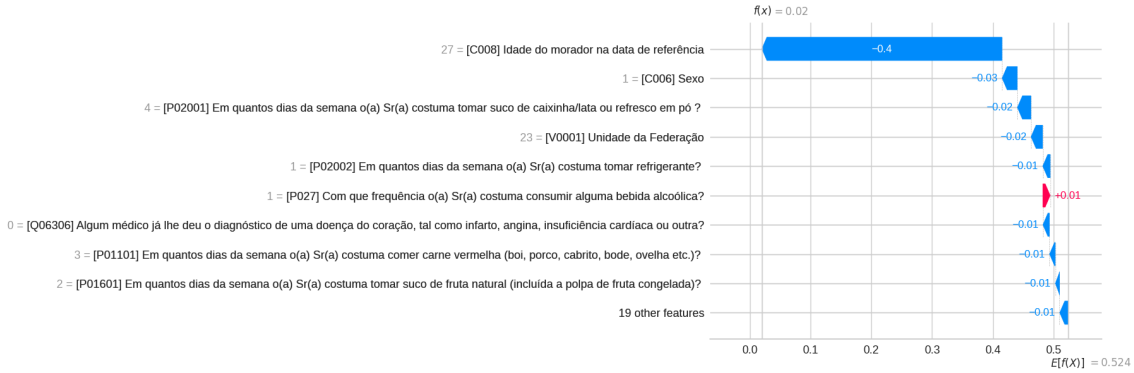
Fonte: O autor.

Também, diante das análises realizadas, podemos observar as *Features Importances*, que nos mostram a contribuição relativa de cada característica (ou *feature*) no nosso modelo. Aplicando esta análise, obtemos que "idade", "peso" e "altura" lideram o ranking, respectivamente, seguidos por "unidade federativa", "hábitos alimentares/estilo de vida" e "doenças crônicas". Dentre estas variáveis, a variável "idade" possui um destaque atingindo 25% de contribuição para o nosso modelo.

Quando analisamos o método SHAP, observando especificamente o valor SHAP absoluto, que nos mostra o quanto um único recurso afetou a previsão apenas na variável alvo, temos que a "idade" contribuiu mais, seguido pelo "sexo" em segundo lugar, em terceiro se o entrevistado costuma "tomar suco de caixa/lata ou refresco em pó", seguido pelas outras *features* que também tiveram contribuição, porém com valores mais baixos como podemos observar na Figura 16.

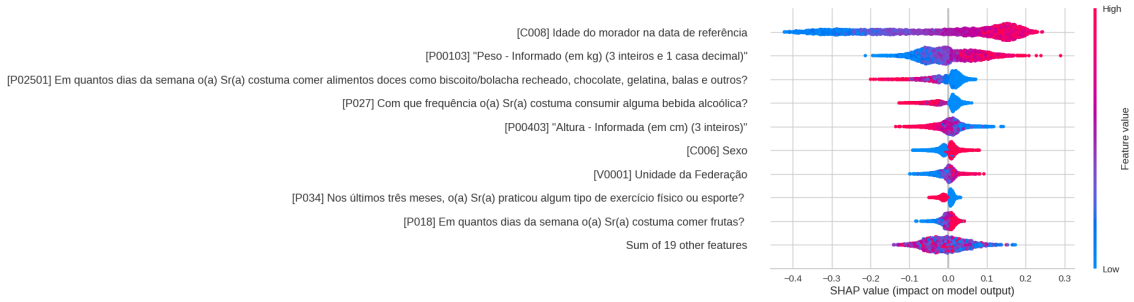
Para entender a importância ou contribuição das características para todo o con-

Figura 16. SHAP Absoluto - Diabetes



Fonte: O autor.

Figura 17. SHAP contribuição das características - Diabetes



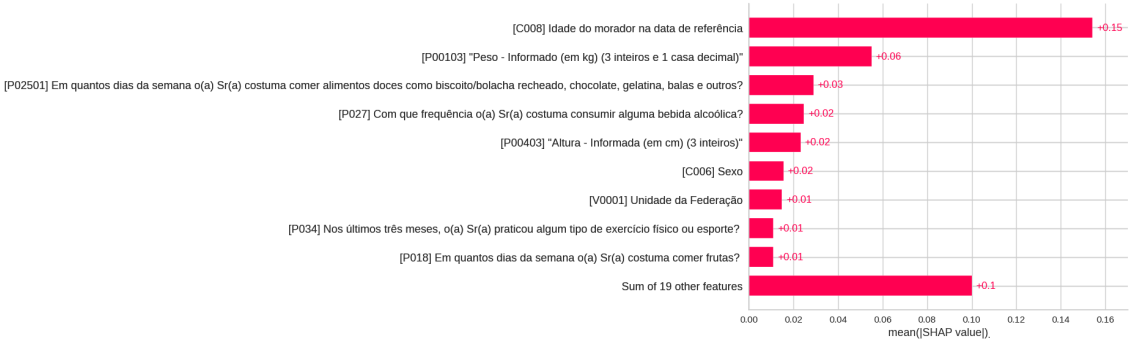
Fonte: O autor.

junto de dados, podemos utilizar o gráfico conforme Figura 17. Nela podemos observar os valores elevados (cor vermelha) e os valores mais baixos (cor azul) de cada *feature*. Também podemos observar a forma de contribuição de cada *feature* sendo negativa se estiver do lado esquerdo e positiva se estiver do lado direito. Vale a pena destacar que, de acordo com o gráfico "idade" e "peso" elevados têm contribuições positivas quando seus valores são altos, ou seja, neste conjunto de dados estas variáveis são um forte indicador de que pessoas com estas caracteristicas sejam acometidas por uma diabetes. Em contrapartida, as variáveis referente a comer "alimentos doces", "consumo de bebida alcoólica" têm contribuições positivas quando seus valores são baixos e ao analisar estas duas variavéis percebemos que isto ocorreu devido aos individuos presentes na base de dados já possuirem diabetes, então consequentemente não irão comer alimentos doces e nem consumir bebida alcoólica.

Quando queremos verificar a importância das *features* levando em consideração o valor absoluto do SHAP, não importando se o recurso afeta a previsão de forma positiva ou negativa, temos as *features* ordenadas do maior para o menor efeito na previsão, conforme Figura 18.

Nela podemos observar como a idade, peso e o consumo de "alimentos doces" dos entrevistados interferem bastante na classificação do diabetes, seguido pelas demais *features*, porém com valores mais baixos e próximos.

Figura 18. SHAP Feature Importance - Diabetes



Fonte: O autor.

Aplicando o algoritmo SVM do tipo linear e com foco em diabetes, foi obtido uma acurácia de 71.56%. Na matriz de confusão, a diagonal principal que são os casos em que o modelo previu corretamente a classe positiva e negativa, obteve 1.464 acertos referente às pessoas que não tem diabetes e 1.702 acertos referente às pessoas que possuem diabetes. Quando olhamos para diagonal secundária, que são os casos em que o modelo previu incorretamente a classe negativa como positiva e positiva como negativa, observamos que o modelo obteve 734 erros classificando a classe negativa como positiva e 524 erros classificando a classe positiva como negativa.

Nos resultados da classificação, vemos que a variável *f1-score* que traz uma média harmônica entre as variáveis *precision* e *recall* obteve 70% para classificação de pessoas sem diabetes e 73% para classificação de pessoas com diabetes, conforme exibido na Figura 19, que traz algumas informações do modelo.

Figura 19. Classification Report Diabetes - SVM

	precision	recall	f1-score	support
0	0.74	0.67	0.70	2198
1	0.70	0.76	0.73	2226
accuracy			0.72	4424
macro avg	0.72	0.72	0.71	4424
weighted avg	0.72	0.72	0.71	4424

Fonte: O autor.

Aplicando o XGBClassifier com foco em diabetes, foi obtido uma acurácia de 70.27%. Na matriz de confusão, a diagonal principal que são os casos em que o modelo previu corretamente a classe positiva e negativa, obteve 1.465 acertos referente às pessoas que não tem diabetes e 1.644 acertos referente às pessoas que possuem diabetes. Quando olhamos para diagonal secundária, que são os casos em que o modelo previu incorretamente a classe negativa como positiva e positiva como negativa, observamos que o modelo obteve 733 erros classificando a classe negativa como positiva e 582 erros classificando a

classe positiva como negativa.

Nos resultados da classificação, podemos observar que a variável *f1-score* obteve 69% para classificação de pessoas sem diabetes e 71% para classificação de pessoas com diabetes, conforme exibido na Figura 20.

Figura 20. Classification Report Diabetes - XGBoost

	precision	recall	f1-score	support
0	0.72	0.67	0.69	2198
1	0.69	0.74	0.71	2226
accuracy			0.70	4424
macro avg	0.70	0.70	0.70	4424
weighted avg	0.70	0.70	0.70	4424

Fonte: O autor.

De forma geral, observando as métricas dos algoritmos abordados neste estudo e focando e suas respectivas acurácia, temos que o algoritmo *Random Forest* obteve melhor desempenho na detecção de doenças cardiovasculares (DCV) ou Diabetes comparado ao XGBoost e o SVM conforme demonstrado na Tabela 7.

Tabela 7. Resultados dos algoritmos

Algoritmos	Acurácia
Random Forest - DCV	71.96%
SVM - DCV	71.71%
XGBoost - DCV	68.68%
Random Forest - Diabetes	72.26%
SVM - Diabetes	71.56%
XGBoost - Diabetes	70.27%

6. Conclusões

Ao revisitar os objetivos definidos para este trabalho, podemos reiterar o propósito central que era aprimorar a eficiência no diagnóstico de doenças cardivasculares ou diabetes, através da aplicação de modelos de machine learning em uma base de dados proveniente da pesquisa nacional de saúde do IBGE. Essa meta foi cumprida através da análise das variáveis presentes nos dados e da aplicação de algoritmos de classificação, com o intuito de identificar padrões e construir modelos preditivos.

Ao longo deste estudo houve a aplicação da *Explainable AI* com um conjunto de processos e métodos a fim de facilitar o entendimento a respeito das variáveis presente na base de dados, em conjunto com a aplicação de alguns algoritmos, conforme visto na Seção 5. Dentre os modelos de machine learning, o *Random Forest*, demonstrou uma acurácia maior na classificação de pacientes com e sem doenças cardiovasculares, ficando com 71.96%, como também na classificação de pacientes com e sem diabetes, obtendo

72.26%. A análise detalhada das métricas de desempenho, como *precision*, *recall* e *f1-score*, revelou a capacidade desse modelo, comparado aos outros analisados, porém vale ressaltar que para apoio na aréa de saúde o resultado obtido não alcançou o satisfatório.

A aplicação de modelos de machine learning é uma abordagem promissora para auxiliar na detecção precoce e no diagnóstico de doenças cardiovasculares ou diabetes. A qualidade dos dados dispostos na base de dados também deve ser levada em consideração, pois pode influenciar negativamente no modelo preditivo. Também foi realizada uma análise das variáveis mais influentes nos modelos, realizada através do método SHAP, que proporcionou alguns insights, como por exemplo, na classificação de doenças cardiovasculares as variáveis "AVC/Derrame", "Artrite/Reumatismo" e ingestão de "suco de fruta" natural ou polpa influenciaram no modelo. Assim como na classificação do diabetes as variáveis "idade", "sexo" e ingestão de "suco de caixinha/lata ou refresco em pó" influenciaram no modelo. De forma geral, sobre os fatores que contribuem para o desenvolvimento dessas condições nesta base de dados, destacam-se a importância de intervenções focadas em hábitos de vida saudáveis e na gestão de doenças crônicas.

No que diz respeito aos trabalhos futuros, sugerimos a continuação da pesquisa neste campo, explorando outros algoritmos, considerando um conjunto mais amplo de variáveis, pois na fase de filtragem utilizando o *feature engineering* existem variáveis que podem ser escolhidas arbitrariamente para ajudar o modelo nos resultados, causando vazamento de dados, como também podem ser selecionadas variáveis que prejudiquem o modelo e distõem do assunto abordado. Otimização de hiperparâmetros para o refinamento do modelo. Além disso, seria interessante investigar o impacto de intervenções específicas, como programas de promoção da saúde e mudanças nos hábitos alimentares e de atividade física, na prevenção e no controle dessas doenças.

Referências

Aquino, L. S. d. et al. (2021). Fatores de risco para diabetes tipo 2 no brasil: uma análise machine learning.

Araujo, L. V., da Silva Miranda, M. H., de Souza Fontenele, M. H., Neto, O. F. D., Batista, J. G., de Lima, A. F., and de Souza, D. A. (2022). Detecção do risco de diabetes em estágio inicial utilizando redes elm e seleção de features baseada em algoritmo genético: Early stage diabetes risk prediction using elm and ga-based feature selection. *Brazilian Journal of Development*, 8(7):54179–54190.

Balbinot, R. A. A. (2014). Diabetes, doenças cardiovasculares e obesidade: análise da legislação na argentina, no brasil e na colômbia. *Revista de Direito Sanitário*, 15(2):91–107.

Barreto, B. F. and Nogueira, M. F. (2013). Fatores de risco para doenças cardiovasculares: identificando a exposição de idosos assistidos na estratégia saúde da família.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Cardozo, G. (2022). Um modelo computacional utilizando técnicas de machine learning e exames laboratoriais de rotina na triagem e apoio ao diagnóstico de diabetes mellitus.

Chapelle, O., Scholkopf, B., and Zien, A. (2006). Semi-supervised learning. 2006. *Cambridge, Massachusettes: The MIT Press View Article*, 2.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20:273–297.

Costalat, T. R. M. and Tavares, G. F. (2022). Machine learning techniques comparison for risk assessment of cardiovascular disease development by health indicators. *Brazilian Journal of Development*, 8(1):6851–6862.

de Geografia e Estatística, I. B. (2022). Pns - pesquisa nacional de saúde.

Dong, G. and Liu, H. (2018). *Feature engineering for machine learning and data analytics*. CRC press.

El Naqa, I. and Murphy, M. J. (2015). *What is machine learning?* Springer.

Eyken, E. B. B. D. V. and Moraes, C. L. (2009). Prevalência de fatores de risco para doenças cardiovasculares entre homens de uma população urbana do sudeste do brasil. *Cadernos de Saúde Pública*, 25:111–123.

Fayyad, U. M., Haussler, D., and Stolorz, P. E. (1996). Kdd for science data analysis: Issues and examples. In *KDD*, pages 50–56.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Gregório, R. L. (2018). Modelo híbrido de avaliação de risco de crédito para corporações brasileiras com base em algoritmos de aprendizado de máquina.

GUIMARÃRES, T. J. R. (2019). Identificação de doenças cardíacas a partir de eletrocardiogramas utilizando machine learning.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., and Yang, G.-Z. (2019). Xai—explainable artificial intelligence. *Science robotics*, 4(37):eaay7120.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(03):90–95.

Maciel, M. (2020). Utilização de técnicas computacionais de mineração de dados para auxiliar no diagnóstico de diabetes mellitus tipo 2.

McKinney, W. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, 445(450):51–56.

Melo, A. K. A., Souza, G. C., Vasco, A. C., and Reis, B. S. (2022). Regulação da inteligência artificial: benchmarking de países selecionados.

Mitchell, T. M. (1997). Machine learning.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Ramos, T. (2021). Uso de *Machine Learning* para predição de pacientes com diabetes mellitus. *Medium*.

Santos, B. B. S. d. (2022). Uma análise exploratória de dados e o uso de aprendizado de máquina para classificação de doenças cardiovasculares. B.S. thesis, Universidade Federal do Rio Grande do Norte.

Santos, S. R. d. (2023). O uso de aprendizado profundo para predição de diagnósticos de doenças cardiovasculares.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.

Waskom, M., Gelbart, M., Botvinnik, O., Ostblom, J., Hobson, P., Lukauskas, S., Gemperline, D. C., Augspurger, T., Halchenko, Y., and Warmenhoven, J. (2020). mwaskom/seaborn: v0. 11.2 (august 2021). *Zenodo*.