SEVEN
publicações acadêmicas

Comparação e seleção de algoritmos de aprendizado de máquina para predição de diabetes: Um estudo quantitativo exploratório baseado em análise de dados médicos

https://doi.org/10.56238/sevened2024.007-053

Vinicius de Souza Santos
Departamento de Engenharia de Computação, Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP) - Campus Birigui, Brasil.
E-mail: vinicius.santos@ifsp.edu.br

RESUMO
A prevalência global de diabetes está aumentando a uma taxa alarmante, tornando a detecção precoce e precisa uma área crítica de interesse. Este estudo emprega técnicas de Machine Learning para prever a incidência de diabetes em uma população de mulheres da herança Pima, conhecida por sua predisposição à doença. Usando um banco de dados de medidas diagnósticas, vários algoritmos foram aplicados, incluindo Support Vector Machines (SVM), Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), Decision Trees e Random Forest, para desenvolver modelos preditivos. A Análise de Componentes Principais (ACP) foi implementada para redução da dimensionalidade e realce das principais variáveis diagnósticas, otimizando o desempenho do algoritmo. Os resultados evidenciaram a superioridade da Floresta Aleatória, que apresentou maior acurácia e precisão, sugerindo sua viabilidade como ferramenta diagnóstica clínica. Este estudo contribui para o campo emergente das aplicações de inteligência artificial em saúde, fornecendo insights valiosos para a prevenção e tratamento precoce do diabetes.

Palavras-chave: Machine Learning, Diabetes, Análise de Componentes Principais, Floresta Aleatória.

1 INTRODUÇÃO

O diabetes tornou-se uma das maiores ameaças à saúde global, com sua prevalência aumentando de forma alarmante nas últimas décadas [1] [2]. Estima-se que, até 2045, mais de 700 milhões de pessoas serão afetadas pela doença, evidenciando a urgência do desenvolvimento de estratégias efetivas para sua prevenção e tratamento [3]. Apesar dos avanços da medicina, a detecção precoce do diabetes continua sendo um desafio significativo, sendo crucial para prevenir complicações graves e melhorar a qualidade de vida dos pacientes, evidenciando uma lacuna crítica no controle da doença [4].

Nesse contexto, a aplicação de algoritmos de aprendizado de máquina surge como uma abordagem promissora para melhorar a predição e o diagnóstico do diabetes [5]. Essas técnicas avançadas oferecem a capacidade sem precedentes de analisar grandes volumes de dados clínicos, detectando padrões ocultos que podem sinalizar o início da doença bem antes dos sintomas se manifestarem [5]. Em particular, este estudo avalia a eficácia de diferentes algoritmos de aprendizado de máquina, incluindo Redes Neurais Artificiais, Máquinas Vetoriais de Apoio, Vizinhos K-Mais Próximos, Árvores de Decisão e Floresta Aleatória. A seleção desses algoritmos baseou-se em sua comprovada eficácia em tarefas de classificação em domínios médicos de acordo com Paix˜ao et al. (2022)[7], bem como sua capacidade de lidar com dados complexos e de alta dimensão [7]. A metodologia comparativa adotada visa não apenas avaliar a acurácia desses modelos, mas também sua capacidade de generalização em diferentes contextos clínicos.

No entanto, é imperativo reconhecer que, apesar de seu imenso potencial, os modelos de aprendizado de máquina não são perfeitos e são suscetíveis a erros [6]. Cardozo, (2022)[5] abordou que a eficácia desses algoritmos depende intrinsecamente da qualidade dos dados, da precisão dos modelos e da adequação da escolha do algoritmo para a tarefa específica em questão. Consequentemente, este estudo não apenas aplica essas ferramentas, mas também propõe um arcabouço metodológico para melhorar continuamente sua acurácia e confiabilidade. A pesquisa em andamento e a colaboração multidisciplinar serão fundamentais para otimizar a aplicabilidade desses algoritmos na área da saúde, garantindo que contribuam positivamente para o diagnóstico precoce e o manejo efetivo dosdiabéticos[5].

Este estudo distingue-se por empregar uma abordagem de pesquisa quantitativa, com foco no método de comparação direta com múltiplos algoritmos de aprendizado de máquina para identificar o método mais eficaz. Os resultados deste estudo podem ter um impacto prático significativo, oferecendo insights valiosos para melhorar as práticas clínicas no diagnóstico e tratamento do diabetes, bem como informar o desenvolvimento de políticas de saúde pública mais eficazes. A aplicação do aprendizado de máquina para o diagnóstico do diabetes tem o potencial de revolucionar a forma como a doença é detectada e gerenciada.

1.1 DIABETES

O diabetes mellitus, comumente referido como diabetes, é uma doença metabólica crônica caracterizada por hiperglicemia persistente, ou seja, níveis elevados de glicose (açúcar) no sangue [49]. Essa condição surge devido a uma deficiência na produção ou ação da insulina produzida pelo pâncreas, hormônio essencial para o metabolismo da glicose [51]. Os critérios diagnósticos específicos incluem glicemia de jejum, tolerância à glicemia pós-prandial ou níveis aleatórios de açúcar no sangue. Os sintomas do diabetes podem incluir sede excessiva, micção frequente, fadiga e perda de peso [50].

No entanto, no contexto histórico e teórico do diabetes, observa-se que, na primeira metade do século 20, o diabetes mellitus se manifestou em crianças e adolescentes de diversas formas [52]. Uma parcela significativa dos pacientes apresentou sintomas agudos como poliúria, polidipsia, desidratação e cetose, com rápida deterioração do quadro clínico, necessitando da administração de insulina para reverter o quadro clínico condição [52]. Entretanto, também foram observados casos em que a doença se apresentava de forma mais insidiosa e muitas vezes sem associação de cetose [53]. Esses casos menos agudos, que constituíram minoria, não necessitaram de insulinoterapia para sobrevida nos estágios iniciais da doença.

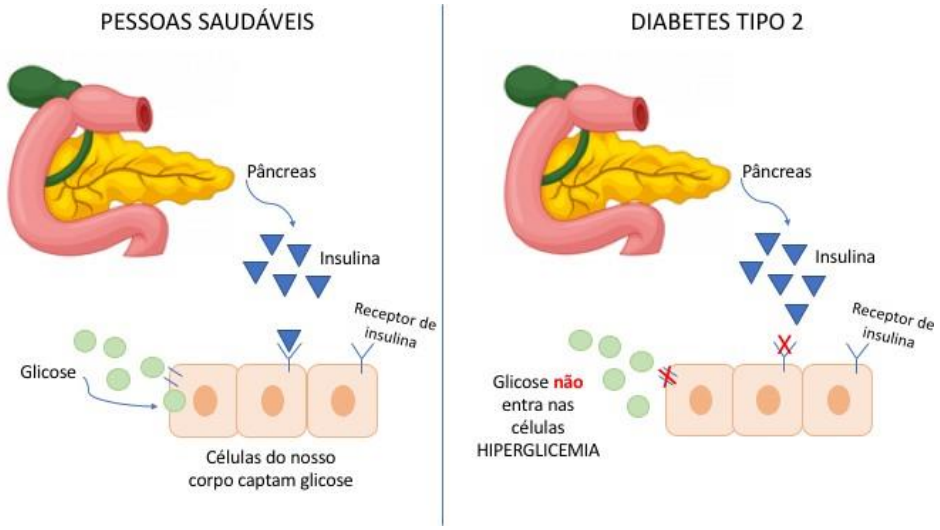
Existem vários tipos de diabetes, sendo os mais comuns na medicina o tipo 1, o tipo 2 e o diabetes gestacional [55]:

- **Diabetes tipo 1: Condição** autoimune em que o sistema imunológico ataca e destrói as células beta do pâncreas, responsáveis pela produção de insulina. Sem insulina suficiente, a glicose se acumula na corrente sanguínea em vez de ser usada como energia. Esse tipo geralmente se manifesta na infância e adolescência, mas também pode ser diagnosticado em adultos. O tratamento requer administração de insulina, planejamento dietético e atividades físicas [55].
- **Diabetes tipo 2:** No diabetes tipo 2, o corpo apresenta resistência à ação da insulina ou não produz insulina suficiente para manter um nível normal de glicose no sangue. É o tipo mais comum, que pode ser manejado, em muitos casos, com atividades físicas e planejamento alimentar. Em outros casos, pode ser necessário o uso de medicamentos ou insulina [55].
- **Diabetes Gestacional:** Ocorre durante a gravidez, quando há um aumento nos níveis de glicose no sangue, e o corpo não consegue produzir insulina suficiente para transportar toda a glicose para dentro das células, resultando em hiperglicemia. Pode causar complicações tanto para a mãe quanto para o bebê se não for manejada adequadamente [55].

No diabetes tipo 1, o sistema imunológico ataca e destrói as células beta produtoras de insulina no pâncreas. Em contraste, o diabetes tipo 2 envolve uma combinação de resistência à insulina e uma deficiência relativa em sua secreção [55]. A Figura 1 ilustra a diferença entre um indivíduo saudável e alguém com diabetes tipo 2. Em uma pessoa saudável, a insulina secretada pelo pâncreas após a ingestão ajuda a glicose a entrar nas células para ser usada como energia [54]. No entanto, no diabetes tipo 2, as

células do corpo não respondem adequadamente à insulina (resistência à insulina), e a glicose não consegue entrar efetivamente nas células, resultando em hiperglicemia [54]. A Figura 1 ilustra esse processo, mostrando os receptores de insulina não funcionando corretamente, impedindo a entrada de glicose nas células.

Fig. 1. Comparação entre a captação de glicose em indivíduos saudáveis e com diabetes tipo 2, destacando a função da insulina e o mecanismo de resistência à insulina.



Fonte: adaptado de [54].

A Figura 1 mostra claramente o funcionamento normal do pâncreas e a ação da insulina nas células de uma pessoa sem diabetes, em comparação com a disfunção observada no diabetes tipo 2, onde a insulina não é capaz de facilitar a entrada de glicose nas células, levando à hiperglicemia. Essa compreensão é crucial para o tratamento, manejo, prevenção e educação sobre o diabetes.

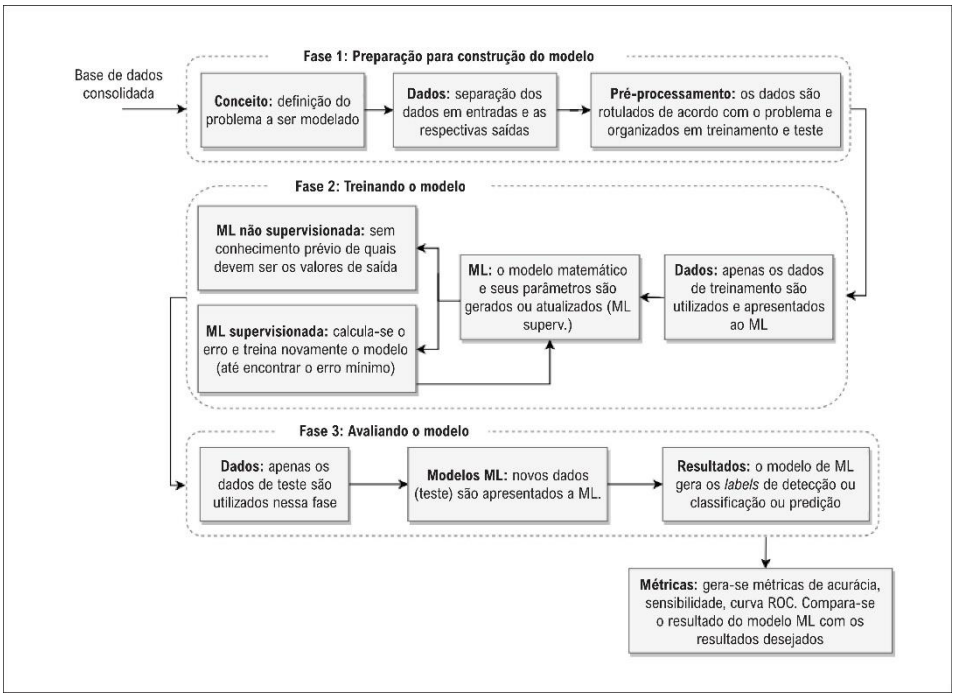
1.2 APRENDIZADO DE MÁQUINA

O aprendizado de máquina (ML), uma área crítica da ciência da computação, opera na confluência de técnicas matemáticas e estatísticas com algoritmos computacionais para identificar padrões e fazer previsões [8]. Na área médica, a ML avança além dos sistemas especialistas tradicionais baseados em regras, processando um volume substancial de variáveis em busca de novas combinações preditivas [8]. A era do big data, caracterizada pelo modelo "3 Vs" — grande volume, alta velocidade e uma grande variedade de informações — desafia as ferramentas tradicionais de gerenciamento de dados com seu enorme volume, alta velocidade e gama variada de informações, exigindo técnicas inovadoras de processamento [9].

O processo de criação de um algoritmo de ML, ilustrado na Figura 2, consiste em três fases: pré-processamento, treinamento e avaliação. Inicialmente, os dados são organizados, a pergunta de pesquisa é formulada e os dados são divididos em conjuntos de treinamento e teste [7]. Na fase de treinamento, o aprendizado pode ser supervisionado, com amostras corretamente classificadas, ou não supervisionado, onde o algoritmo aprende sem rótulos pré-definidos [7]. Na etapa final, o modelo é

testado e avaliado, estabelecendo-se um padrão de mapeamento para a classificação precisa e confiável de novos dados [7].

Fig. 2. Fases para o desenvolvimento de algoritmos de machine learning.



Fonte: [7]

É essencial que o desenvolvimento de algoritmos de ML seja conduzido em um banco de dados consolidado e validado, evitando assim a geração de resultados espúrios [7]. A aprendizagem de ML, supervisionada ou não, é um processo iterativo de observações repetidas. Na aprendizagem supervisionada, o algoritmo aprende com exemplos rotulados, enquanto na aprendizagem não supervisionada, o algoritmo identifica padrões nos dados sem rótulos pré-definidos. Esse processo permite que o algoritmo generalize informações e classifique com precisão novos conjuntos de dados [7].

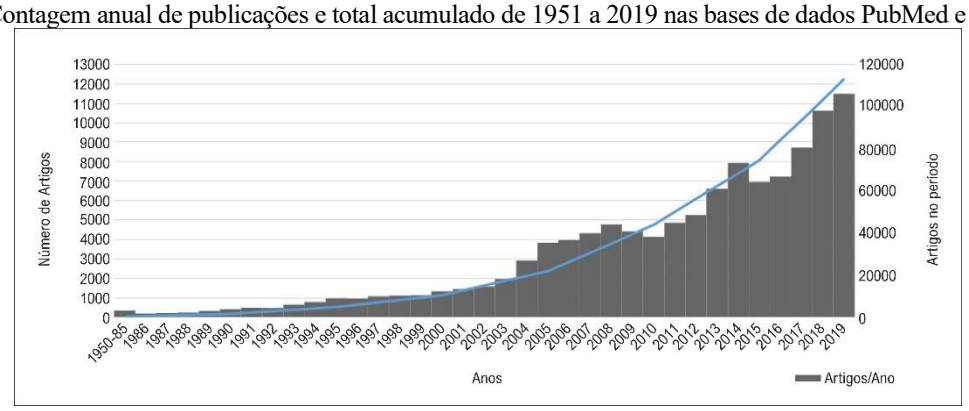
1.3 APLICAÇÕES DE MACHINE LEARNING EM MEDICINA

A medicina está passando por uma transformação impulsionada pelo rápido avanço das técnicas de aprendizagem ma- chine (ML). A aplicação dessas técnicas à prática médica tem mostrado potencial para revolucionar o diagnóstico, o tratamento e a prevenção de doenças [10]. À medida que a quantidade de dados gerados no setor de saúde continua a crescer exponencialmente, o ML oferece ferramentas para analisar eficientemente esses dados e extrair insights valiosos [11].

A Figura 3 ilustra a tendência crescente na produção da literatura científica relacionada à LM em medicina, demonstrando um aumento substancial no número de artigos publicados entre 1951 e 2019, conforme

indexado no PubMed e Medline. Esse crescimento reflete não apenas o interesse acadêmico, mas também o potencial prático da LM na medicina.

Fig. 3. Contagem anual de publicações e total acumulado de 1951 a 2019 nas bases de dados PubMed e Medline.



Fonte: [7]

As aplicações do ML na medicina são vastas e variadas. Vão desde sistemas de apoio à decisão clínica que auxiliam os profissionais de saúde na escolha de tratamentos baseados em padrões encontrados na história médica [12], até algoritmos de processamento de imagens que melhoram a acurácia diagnóstica em radiologia [13].

Os algoritmos de Machine Learning (ML) têm desempenhado um papel crucial na previsão de surtos de desanuviamento, na otimização de recursos hospitalares e no desenvolvimento de novos medicamentos. A efetividade das técnicas de ML aplicadas a séries temporais e a uma coleção de variáveis explicativas varia dependendo da variável resposta utilizada. Em estudos recentes, predic- ções de novos casos diários e mortes por Coronavírus em cidades brasileiras têm utilizado características como temperatura, qualidade do ar, umidade e buscas no Google relacionadas à Covid-19 como covariáveis, combinando-as com informações históricas para melhor prever tendências pandêmicas e direcionar intervenções apropriadas [15] [16].

No entanto, apesar dos avanços, a implementação da ML na medicina enfrenta desafios, incluindo a necessidade de grandes conjuntos de dados anotados, preocupações com a privacidade e segurança dos dados e a importância de resultados interpretáveis por profissionais de saúde [17].

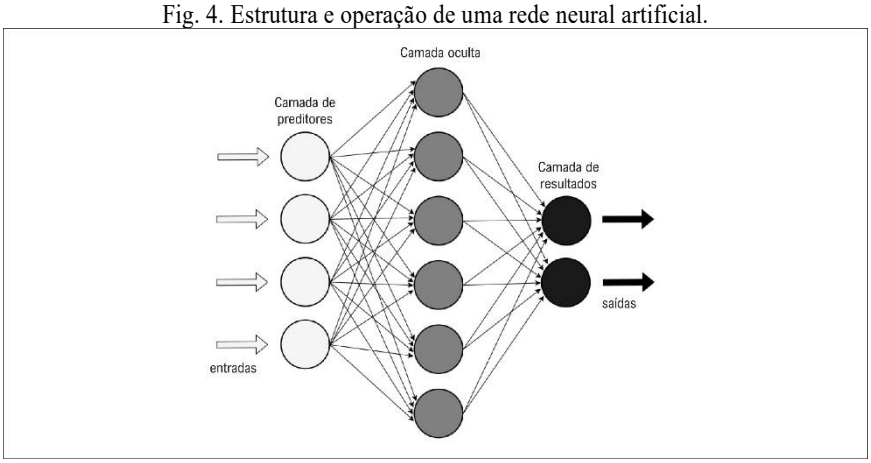
Portanto, o potencial do ML na medicina é evidente, mas sua aplicação efetiva requer uma abordagem multidisciplinar e colaborativa envolvendo médicos, cientistas de dados, engenheiros e formuladores de políticas de saúde [7]. À medida que avançamos, é essencial que as ferramentas de ML sejam validadas em ambientes clínicos e alinhadas com os melhores meios de comunicação. práticas para garantir que complementem - em vez de substituir - a experiência humana na prestação de cuidados de saúde.

1.4 TÉCNICAS DE MACHINE LEARNING

O aprendizado de máquina (ML) transformou várias áreas de pesquisa e aplicação prática, notadamente no campo da medicina, onde as técnicas de ML oferecem novas perspectivas para diagnósticos e tratamentos personalizados [11]. Este estudo concentra-se em métodos específicos de LM, cada um com sua figura representativa, para melhorar a predição e o diagnóstico do diabetes.

As Redes Neurais Artificiais (RNAs) são inspiradas no funcionamento biológico dos neurônios humanos e foram inicialmente propostas por McCulloch e Pitts em 1943. Esse modelo, representado na Figura 4, consiste em unidades de processamento conectadas por elos ponderados, cujos pesos são ajustados durante o treinamento. Uma RNA "aprende" ajustando esses pesos para minimizar o erro de previsão do modelo. Por exemplo, um estudo de Fonseca, Afonso Ueslei, et al (2023) aplicou RNA para identificar padrões em imagens médicas, facilitando o diagnóstico precoce da doença [18].

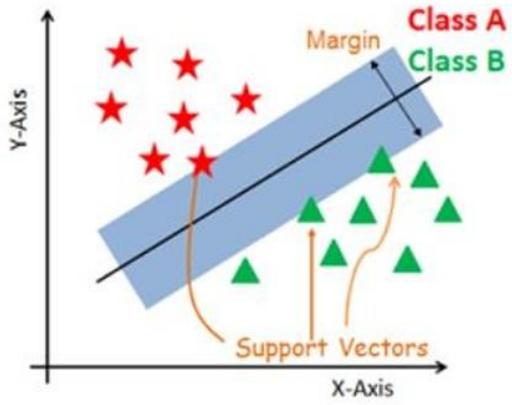
Fig. 4. Estrutura e operação de uma rede neural artificial.



Fonte: [7]

Support Vector Machines (SVM) são um modelo analítico robusto dentro do aprendizado de máquina, operando tanto na classificação quanto na regressão. Essa técnica, originalmente desenvolvida por Vapnik em 1995 [22], distingue-se pelo uso estratégico de hiperplanos que atuam como margens decisivas na separação de classes dentro de um conjunto de dados, como demonstrado na representação de seu hiperplano na Figura 5. A efetividade da SVM está na maximização dessas margens, pois intuitivamente entendida que quanto maior a distância entre hiperplanos paralelos, mais preciso será o modelo na previsão de novas instâncias [21].

Fig. 5. Representação de um hiperplano em um determinado conjunto de dados.

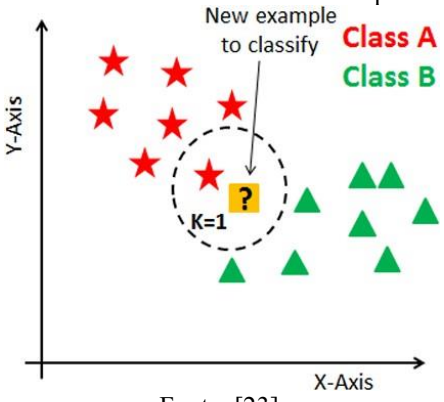


Fonte: [20]

Em termos de aplicação prática, estudo recente realizado por Costa e Gou- veia (2022) [19] exemplifica o potencial da SVM na área da saúde. Nesta pesquisa, o SVM foi aplicado na predição de doenças crônicas não transmissíveis (DCNT), alcançando uma notável acurácia de 97%. Esse resultado não apenas reforça a competência da SVM em lidar com dados complexos e de alta dimensão, mas também reforça a relevância da técnica em cenários onde decisões precisas e confiáveis são fundamentais para o diagnóstico e tratamento de condições de saúde.

O método K-Nearest Neighbors (KNN) é uma técnica de aprendizado de máquina poderosa e intuitiva para classificação e regressão. Proposto por Fix e Hodges em 1951 [24], esse método não-paramétrico atribui a classificação de um novo exemplo com base nas classes mais frequentes entre seus vizinhos mais próximos. No KNN, os dados k mais próximos do exemplo em questão são identificados, e a classificação é realizada por votação majoritária entre esses vizinhos, ou, no caso de k=1, o exemplo é simplesmente atribuído à classe de seu vizinho mais próximo, como ilustrado na Figura 6 abaixo.

Fig. 6. Exemplo de classificação com K-Vizinhos mais próximos (KNN), onde k=1 indica que o novo exemplo (marcado com um ponto de interrogação) é classificado de acordo com a classe de seu vizinho mais próximo.

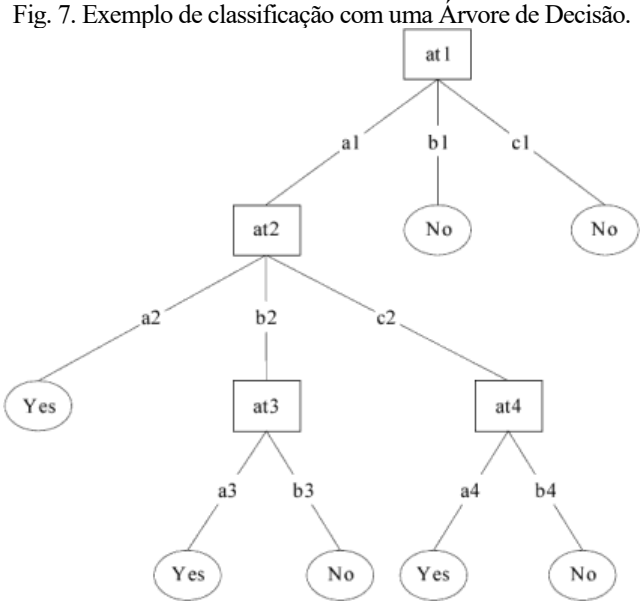


Fonte: [23].

Esse método é flexível em relação ao número de vizinhos (k), permitindo ajustes para melhorar a precisão da classificação. A KNN tem sido aplicada com sucesso em diversos contextos médicos, como demonstrado por Andr'e Oliveira (2016), que utilizou a KNN para classificar os tipos de diabetes com base em medidas clínicas [26]. Mais recentemente, KNN foi empregado com k=3 e a variante KNN ponderada, fornecendo uma capacidade refinada de discernir padrões complexos em dados de saúde, o que é crucial para a implementação de diagnósticos precisos e personalizados [23].

Decision Trees, criada por J. Ross Quinlan em 1983. Quinlan também é autor do livro "Machine Learning", publicado em 1983, que foi um dos primeiros livros a apresentar o conceito de machine learning [28]. Árvores de decisão são modelos preditivos que segmentam o espaço de dados em subconjuntos com base em decisões lógicas. Um exemplo prático é o trabalho de Carvalho e colaboradores (2015), que utilizaram árvores de decisão para criar sistemas de apoio à decisão clínica para o diagnóstico do diabetes tipo 2 [29].

Fig. 7. Exemplo de classificação com uma Árvore de Decisão.



Fonte: [27].

Random Forest, desenvolvido por Breiman em 2001, constitui um avanço significativo na análise preditiva, particularmente no contexto da classificação de dados complexos. Este método opera através da combinação de múltiplas árvores de decisão, cada uma construída a partir de uma amostra aleatória do conjunto de dados, com a seleção aleatória de variáveis em cada divisão do nó [60]. A essência do Random Forest reside em sua capacidade de reduzir o risco de overfitting – um problema comum em modelos complexos de aprendizado de máquina – enquanto mantém ou até aumenta a precisão preditiva [60].

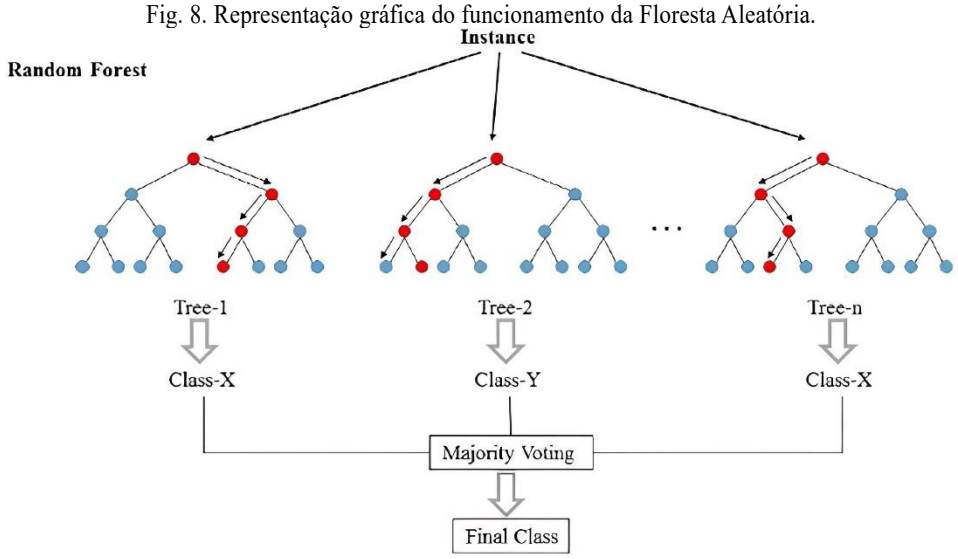
Um aspecto notável da Random Forest é sua adaptabilidade a diferentes tipos de dados e complexidades de problemas, tornando-a particularmente eficaz em contextos onde as relações entre variáveis são intrincadas e difíceis de modelar com abordagens lineares ou paramétricas simplificadas [61] [62]. A técnica baseia-se no princípio de que um grande número de modelos relativamente não

correlacionados (árvores) trabalhando juntos pode superar o desempenho de qualquer modelo individual, fornecendo assim uma abordagem poderosa para tarefas de classificação e regressão [62].

A implementação prática da Floresta Aleatória envolve o treinamento de inúmeras árvores de decisão em subconjuntos variados do conjunto de dados [60]. Cada árvore faz uma previsão independente, e a classificação final é determinada através de votação por maioria entre todas as previsões das árvores [63]. Esse processo de agregação, conhecido como "bagging", contribui para a capacidade da Random Forest de generalizar bem para novos dados, evitando o overfitting enquanto explora a diversidade das árvores constituintes [64].

Vários estudos têm demonstrado a eficácia da Floresta Aleatória em uma ampla gama de aplicações, desde a predição de doenças em áreas médicas [19] [66] até a modelagem de padrões de consumo de energia em ambientes urbanos, onde a complexidade e a interação entre múltiplas variáveis desafiam modelos mais simples [65]. A capacidade do Random Forest de lidar com grandes volumes de dados, sua tolerância a dados ausentes e a facilidade de interpretar resultados contribuem para sua popularidade e aplicabilidade em vários domínios de conhecimento.

Fig. 8. Representação gráfica do funcionamento da Floresta Aleatória.



Fonte: [61]

A Figura 8 ilustra a essência do algoritmo Random Forest, enfatizando sua estrutura colaborativa e descentralizada. Cada árvore individual na floresta realiza uma avaliação independente de uma instância, com base em uma amostra aleatória dos dados e um subconjunto aleatório de variáveis. O resultado de cada árvore é uma previsão que, quando combinada através do processo de votação por maioria, leva à classificação final fornecida pelo modelo. Esse mecanismo não apenas melhora a precisão da previsão por meio da diversidade e do número de árvores envolvidas, mas também mitiga o risco de sobreajuste, pois a probabilidade de todas as árvores cometerem os mesmos erros é reduzida. A representação visual capta esse

conceito, mostrando como árvores individuais contribuem para a decisão coletiva, exemplificando a abordagem de conjunto que é central para a Floresta Aleatória.

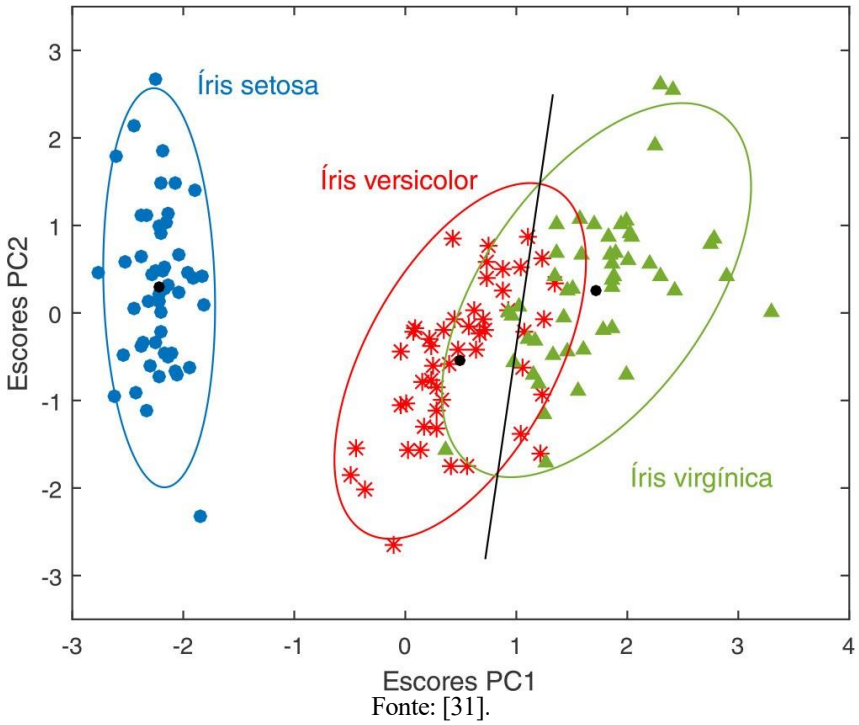
A Análise de Componentes Principais (ACP) é uma técnica de redução de dimensionalidade multivariada, crucial para o processamento e análise de conjuntos de dados de alta dimensão. Desenvolvida inicialmente por Karl Pearson em 1901, essa técnica transforma um conjunto de possíveis variáveis correlacionadas em um conjunto de valores de variáveis linearmente não correlacionadas chamados componentes principais [30]. A PCA baseia-se na ortogonalização do espaço de dados e na maximização da variância, que permite a compressão dos dados mantendo a maior parte da informação original – benefício explorado no trabalho de tese de Fernandes (2022) [69].

O primeiro componente principal é a direção no espaço de dados que maximiza a variância das projeções de dados, enquanto os componentes subsequentes são ortogonais aos anteriores e maximizam a variância restante. A técnica é particularmente útil na identificação de padrões, eliminação de redundâncias e interpretação de conjuntos de dados complexos [31].

Na área médica, a ACP tem sido aplicada para identificação de biomarcadores, visu- alização de doenças complexas e análise genômica. Por exemplo, Porreca e colaboradores (2021) usaram a PCA para investigar os principais fatores que influenciam os efeitos do uso de máscara facial no desempenho do exercício durante a pandemia de COVID-19. Essa aplicação destaca como a ACP pode desempenhar um papel na compreensão de fenômenos multifatoriais e no direcionamento de medidas de saúde pública [32].

A Figura 9 associada à ACP tipicamente mostra a dispersão dos dados nos dois primeiros componentes principais, oferecendo uma visualização clara da variabilidade dos dados e como diferentes grupos podem ser discriminados com base nessas projeções. As elipses de confiança ao redor dos agrupamentos fornecem uma compreensão visual do agrupamento e da confiança estatística de que uma amostra pertence a um grupo específico.

Fig. 9. Representação gráfica da ACP mostrando três espécies de flores de Iris. Os pontos pretos indicam os centroides de cada grupo. As elipses ao redor das amostras foram desenhadas com 95% de confiança, ilustrando a capacidade da ACP de discriminar diferentes categorias biológicas.



Fonte: [31].

Esse tipo de representação gráfica é uma ferramenta poderosa para a exploração inicial de dados, permitindo que os pesquisadores identifiquem agrupamentos naturais, outliers e tendências que podem não ser imediatamente aparentes em dados de alta dimensão.

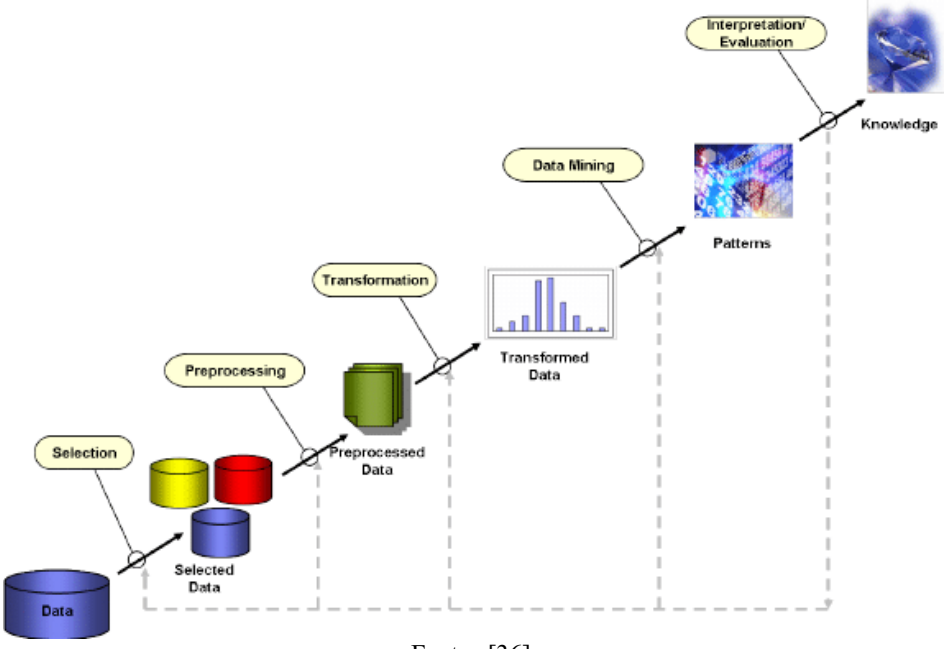
A mineração de dados, também conhecida como mineração de dados, constitui um campo interdisciplinar emergente, alimentado pelo crescimento exponencial da capacidade de armazenar e organizar dados massivos [33]. Essa evolução, decorrente dos avanços da tecnologia da informação, estimulou o desenvolvimento de métodos para extrair inteligência acionável de vastos repositórios de dados [34]. Assim, a mineração de dados configura-se como uma disciplina que congrega métodos estatísticos como demonstrado na Figura 10, princípios de aprendizado de máquina e técnicas de reconhecimento de padrões, para destilar conhecimento e descobertas significativas a partir de bancos de dados complexos e multidimensionais [34].

Fig. 10. Diversidade de Fontes de Mineração de Dados

Fonte: [33].

(1996) cunharam o termo Knowledge Discovery in Databases (KDD) para descrever o processo de ponta a ponta da descoberta do conhecimento, que começa com dados brutos e culmina no uso de insights derivados para a tomada de decisões estratégicas [35]. O processo KDD, que consiste em uma série de etapas iterativas e over-lapping - incluindo pré-processamento, limpeza, integração, seleção, transformação, a mineração em si, avaliação e, finalmente, apresentação - é descrito em vários modelos de processo [33]. Esse modelo estrutura a mineração de dados como uma sequência de etapas lógicas, garantindo rigor metodológico e replicabilidade.

Fig. 11. Processos de Descoberta de Conhecimento em Bancos de Dados (KDD)



Fonte : [36].

A limpeza de dados, etapa inicial crucial, lida com dados incompletos, incorretos ou inconsistentes, preparando o conjunto para posterior análise [37]. Técnicas como imputação, tratamento de outliers e normalização são empregadas para garantir a qualidade e confiabilidade dos dados [37].

A integração eficaz de dados busca consistência e coerência, reunindo informações de várias fontes, como arquivos de texto, bancos de dados, imagens e vídeos. Essa fase envolve análise detalhada dos dados para identificar redundâncias, dependências entre variáveis e conflitos de valor [37]. Após a integração, é realizada a seleção dos dados relevantes para as técnicas de mineração de dados, seguida do tratamento dos dados, que pode incluir a transformação ou consolidação dos dados no modelo mais adequado para o processo de mineração de dados [37]. Esse tratamento pode envolver a generalização de atributos detalhados e a normalização dos dados para se encaixarem em uma faixa específica, bem como a construção de novos atributos a partir dos já existentes, como o cálculo do IMC a partir de variáveis de peso e altura [40].

Na mineração de dados, a avaliação de algoritmos é crucial para garantir a confiabilidade dos resultados obtidos. Métricas de avaliação como acurácia, pontuação f1, precisão e matriz de confusão servem como indicadores-chave do desempenho dos modelos de classificação [41]. A acurácia é uma medida geral de desempenho que calcula a proporção de predições corretas em relação ao número total de casos, útil em conjuntos de dados balanceados [41]. O escore f1 é uma métrica que considera tanto a precisão (a proporção de cor-predições positivas em relação ao número total de predições positivas) e recall (a proporção de predições positivas corretas em relação ao número total de casos positivos reais), oferecendo um equilíbrio entre essas duas métricas, particularmente em situações de desequilíbrio de classe [42]. A matriz de confusão, por outro lado, proporciona uma visão detalhada do desempenho do modelo, representando as frequências de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos, permitindo uma análise mais granular do tipo de erros cometidos pelo modelo [43]. Essas métricas são fundamentais para o refinamento e seleção de modelos em aplicações de Data Mining, garantindo que as previsões não sejam apenas precisas, mas também aplicáveis e interpretáveis no contexto em que serão utilizadas [41].

1.5 APRENDIZAGEM SUPERVISIONADA E NÃO SUPERVISIONADA

O aprendizado supervisionado e não supervisionado são os pilares do campo de aprendizado de máquina, cada um servindo a propósitos distintos e fornecendo insights valiosos a partir de dados.

No contexto da Aprendizagem Supervisionada, a máquina é treinada com um conjunto de dados conhecido, onde tanto as entradas quanto as saídas desejadas são fornecidas, permitindo que o modelo estabeleça uma relação funcional entre elas [45]. Assim, o algoritmo aprende a mapear entradas para saídas, facilitando a previsão de resultados para dados novos e inéditos. Esse processo é descrito como um método de classificação ou regressão, dependendo da natureza da variável de saída – categórica para classificação e contínua para regressão [46].

A Aprendizagem Não Supervisionada, por outro lado, opera sem an- swers predefinidos, explorando a estrutura intrínseca dos dados de entrada. Aqui, o algoritmo procura descobrir padrões, agrupamentos ou associações subjacentes sem qualquer intervenção ou rótulos externos [47]. Esse tipo de aprendizagem é crucial quando a redundância nos dados de entrada permite a identificação de regularidades e, consequentemente, a formação de representações internas que categorizam os dados de forma autônoma [47] [45].

A Figura 12 ilustra a diversidade de aplicações e métodos dentro do Machine Learning, destacando a importância do aprendizado supervisionado e não supervisionado em campos avançados como visão computacional, marketing direcionado e desenvolvimento de sistemas de recomendação, todos essenciais na era do Big Data e da Inteligência Artificial.

Fig. 12. Aplicações de algoritmos de aprendizado de máquina

Fonte: [44].

É importante notar que ambas as abordagens têm suas vantagens e limitações, e a escolha entre supervisionado e não supervisionado muitas vezes depende da natureza do problema em questão e da disponibilidade e qualidade dos dados [44].

Para a implementação efetiva dessas técnicas, a compreensão da relação entre as características dos dados e os resultados desejados é fundamental, assim como a capacidade de traduzir essas relações em modelos preditivos precisos [48].

1.6 DESAFIOS PRÁTICOS NA IMPLEMENTAÇÃO DE MACHINE LEARNING NA MEDICINA

A integração do Machine Learning (ML) na área médica, embora promissora, enfrenta desafios práticos significativos que podem impactar sua eficácia e adoção. Esses desafios podem ser agrupados em várias categorias principais:

- **Aquisição e Qualidade de Dados:** A eficiência dos algoritmos de ML é diretamente proporcional à qualidade e quantidade de dados disponíveis [33]. A obtenção de grandes conjuntos de dados médicos anotados e confiáveis é uma tarefa complexa devido à sensibilidade dos dados e à necessidade de proteger a privacidade do paciente.
- **Privacidade e Segurança de Dados:** Regulamentações rígidas sobre dados de saúde, como HIPAA nos EUA, GDPR na Europa e LGPD no Brasil, representam desafios significativos no uso de dados para treinamento de ML sem comprometer a privacidade do paciente [56].
- **Interpretabilidade do modelo:** A natureza de "caixa preta" dos algoritmos de ML pode ser um obstáculo na prática médica, onde a compreensão do "porquê" e do "como" das previsões é crucial para a confiança e aceitação pelos profissionais de saúde [57].
- **Integração no fluxo de trabalho clínico:** A integração de ferramentas de ML no ambiente clínico requer uma adaptação do fluxo de trabalho existente, que pode enfrentar resistência dos profissionais de saúde devido à curva de aprendizado ou desconfiança em relação às novas tecnologias [58].
- **Variações entre pacientes e condições:** A diversidade genética, comportamental e ambiental dos pacientes significa que os algoritmos de ML precisam ser extremamente robustos e capazes de generalizar bem em diferentes subpopulações.
- **Colaboração multidisciplinar:** A eficácia do ML na medicina depende da colaboração entre médicos, cientistas de dados, engenheiros de software e outros profissionais, o que pode ser desafiador devido às diferentes linguagens e abordagens de cada disciplina.
- **Atualização e manutenção contínuas: os** modelos de ML precisam ser continuamente atualizados com novos dados para manter sua precisão, o que requer um compromisso contínuo de recursos e conhecimento especializado.

A superação desses desafios requer uma abordagem multidisciplinar e colaborativa, bem como um compromisso com a educação continuada e a adaptação das práticas clínicas para incorporar de forma responsável e ética os avanços tecnológicos.

2 MATERIAIS E MÉTODOS

2.1 BASE DE DADOS

O banco de dados utilizado neste estudo foi adquirido do repositório Kaggle, (2024), originário do National Institute of Diabetes and Digestive and Kidney Dis- eases [59]. O objetivo do conjunto de

dados é prever diagnóstica se um paciente tem diabetes, com base em medidas diagnósticas específicas incluídas no conjunto de dados. Todos os pacientes são do sexo feminino, com idade mínima de 21 anos e de ascendência indígena Pima. A base de dados está publicamente disponível sob a licença CC0: Public Domain [59].

2.2 PROCESSAMENTO E ANÁLISE DE DADOS

O processamento e a análise de dados foram essenciais para preparar o conjunto de dados para algoritmos de aprendizado de máquina. A metodologia adotada para o tratamento dos dados seguiu uma série de etapas estruturadas, garantindo a qualidade e confiabilidade dos dados para posterior modelagem preditiva.

Inicialmente, o banco de dados, contendo medidas médicas relevantes para o diagnóstico de diabetes, foi carregado e lido. As colunas do banco de dados incluem o número de preg- nancies, níveis de glicose, pressão arterial, espessura da pele, insulina, Índice de Massa Corporal (IMC), função do pedigree do diabetes, idade e o desfecho binário indicando a presença ou ausência de diabetes.

Durante o carregamento dos dados, identificou-se a presença de valores faltantes, representados pelo caractere '?'. Estes foram tratados substituindo-os pelos valores médios de suas respectivas colunas, método estatístico padrão que mantém a distribuição original dos dados sem introduzir viés significativo, conforme estudado por Cardoso (2022) [67].

Após a correção dos valores faltantes, uma técnica de normalização foi aplicada para padronizar a escala de dados. Duas abordagens de normalização foram utilizadas: normalização do escore Z e normalização Mín-Máx. A normalização do escore Z transforma os dados para ter uma média zero e um desvio padrão, enquanto a normalização Min-Max redimensiona os dados para um intervalo [0, 1], onde os valores mínimo e máximo da coluna se tornam 0 e 1, respectivamente. Cada abordagem de normalização tem seu conjunto de vantagens e é selecionada com base nos requisitos específicos do algoritmo de aprendizado de máquina e na natureza dos dados, conforme discutido no estudo de Maniezzo (2022) [68].

Além disso, a redução da dimensionalidade foi realizada por meio da Análise Principal de Compósitos (ACP). A ACP é uma técnica estatística que converte um conjunto de possíveis variáveis correlacionadas em um conjunto de valores de variáveis linearmente não correlacionadas denominadas componentes principais. Essa etapa é crucial, pois reduz a complexidade do modelo sem perder informações significativas, o que pode melhorar a eficiência computacional e evitar o problema de overfitting ao treinar modelos de aprendizado de máquina.

A visualização da ACP, por meio de gráficos, proporcionou uma compreensão intuitiva da distribuição e separação dos dados, permitindo uma análise preliminar de como os dados poderiam ser agrupados ou classificados.

2.3 ALGORITMOS DE APRENDIZADO DE MÁQUINA USADOS

Os algoritmos de aprendizado de máquina usados para a classificação de dados médicos incluem:

- Máquinas de vetor de suporte (SVM)
- Redes Neurais Artificiais (RNA)
- K-Vizinhos mais próximos (KNN)
- Árvores de Decisão
- Floresta aleatória

Antes da aplicação desses algoritmos, o banco de dados passou por um processo de pré-processamento para garantir a qualidade e uniformidade dos dados. Esse processo incluiu a normalização dos dados e a divisão dos dados em um conjunto de treinamento e um conjunto de testes, usando uma proporção de 70:30, respectivamente. Essa abordagem garante que o modelo seja treinado em uma parte significativa dos dados, mantendo uma parte separada para testar a eficácia do modelo em dados inéditos.

Após o pré-processamento, cada algoritmo foi ajustado e validado para obter o melhor desempenho possível. A seleção do modelo baseou-se na precisão da classificação e no relevância clínica dos desfechos. Para garantir uma avaliação abrangente e justa de cada modelo, foram utilizadas as seguintes métricas:

- **Precisão:** A proporção de predições corretas em relação ao número total de casos.
- **Precisão:** A proporção de previsões positivas corretas em relação ao número total de previsões positivas.
- **F1-Score:** A média harmônica de precisão e recordação, proporcionando um equilíbrio entre essas duas métricas.
- **Matriz de Confusão:** Uma tabela que permite a visualização da performance do algoritmo, incluindo verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

O objetivo dessa avaliação é identificar o modelo que não apenas apresenta a maior acurácia, mas também equilibra efetivamente a precisão e a capacidade de recordação, cruciais para a aplicação prática na área médica.

Support Vector Machines (SVM): Um classificador SVM com um kernel linear foi instanciado, o que é adequado para dados que são linearmente separáveis. O parâmetro gama foi ajustado como 'auto', o que significa que o valor de gama é calculado automaticamente como 1/n características, e o C, que é o parâmetro de regularização, foi ajustado como 3,0. Um C maior pode levar a um modelo com uma margem menor, mas pode se ajustar melhor aos dados de treinamento. O parâmetro de estado aleatório foi configurado para garantir a reprodutibilidade dos resultados.

O modelo SVM foi treinado usando o conjunto de dados de treinamento X train para atributos e y train para os rótulos correspondentes.

Após o treinamento, o modelo foi utilizado para fazer predições no conjunto teste X. O desempenho do modelo foi avaliado por meio de várias métricas. A precisão (svm accuracy) fornece a fração de previsões corretas, enquanto o F1 Score (svm f1) é a média ponderada de precisão e sensibilidade e fornece uma medida de precisão e recordação. A precisão (svm precision) mede a proporção de identifieações positivas que foram realmente corretas, e a matriz de confusão (confusion matrix) oferece uma visão detalhada do desempenho do modelo, mostrando o frequências de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

Esta tradução mantém o detalhe técnico e a clareza do texto original, adequado para um público acadêmico e científico de língua inglesa. Se houver mais seções que você precisa de ajuda para traduzir ou ajustes específicos necessários, por favor me avise.

Redes Neurais Artificiais (RNA): A aplicação do algoritmo de Redes Neurais Artificiais (RNA) para o modelo de classificação no contexto desta pesquisa seguiu uma abordagem sistemática. A RNA foi configurada com duas ocultas camadas, cada uma contendo 10 neurônios, e o treinamento iterado por um máximo de 1000 épocas. A escolha da arquitetura da RNA, incluindo o número de camadas ocultas e neurônios, é influenciada por considerações empíricas e pela complexidade do problema em análise. A arquitetura selecionada visa capturar a complexidade dos dados sem incorrer em overfitting, alinhada com as diretrizes de Goodfellow et al (2016) sobre a profundidade das redes neurais [70].

A RNA foi treinada por meio do conjunto de treinamento, composto pelas variáveis independentes (X trem) e pela variável dependente (y trem). A função de ajuste da classe MLPClassifier do scikit-learn foi utilizada para ajustar o modelo aos dados, onde o estado aleatório foi ajustado para garantir a reprodutibilidade dos resultados.

Após o treinamento, a RNA foi utilizada para fazer predições no conjunto teste (teste X), resultando em uma série de classificações que foram comparadas com os valores verdadeiros (teste y). A avaliação do modelo foi realizada por meio de várias métricas de desempenho, incluindo a precisão, que mede a proporção de predições corretas; o escore F1 (escore F1), que é a média ponderada de precisão e recordação; precisão (preci- sion score), que avalia a acurácia de previsões positivas; e a matriz de confusão (matriz de confusão), que fornece uma visão detalhada da performance do modelo na categorização correta ou incorreta das observações em suas respectivas classes.

A precisão da RNA (precisão do RNA) reflete a capacidade geral do modelo de classificar corretamente as instâncias. O escore F1 (rna f1) é particularmente útil em situações com classes desbalanceadas, pois leva em conta tanto a precisão quanto a recordação. A precisão ponderada (precisão de rna) é calculada levando em conta o equilíbrio de classes e é útil para entender como o modelo se comporta em cada classe individualmente. A matriz de confusão (matriz de confusão de rna) oferece insights sobre os tipos de erros cometidos pelo modelo, como falsos positivos e falsos negativos.

K-Nearest Neighbors (KNN): Para a implementação do algoritmo KNN, foi utilizada a biblioteca Python Scikit-learn, que oferece ferramentas eficientes para análise de dados e modelagem preditiva. O KNeighborsClassifier foi instanciado com o número de vizinhos k definido como 1. A métrica de distância escolhida foi a de Minkowski com p=2, correspondente à distância euclidiana, apropriada para o nosso espaço característico.

O modelo foi treinado utilizando o conjunto de treinamento X trem com as classes y trem correspondentes. O ajuste do modelo foi realizado pelo método de ajuste, que é o processo de treinamento do algoritmo com os dados fornecidos.

Após o treinamento, o modelo foi utilizado para fazer predições no conjunto teste X. As predições foram armazenadas na variável predições knn. A eficácia do modelo foi então avaliada comparando-se as predições com os valores reais y teste do conjunto de testes. As métricas utilizadas para avaliação incluíram:

Precisão (knn accuracy**):** A proporção de previsões corretas a partir do total de previsões feitas.

- **F1-Score (**knn f1**):** Uma medida que combina precisão e recordação. É o meio harmônico de precisão e recordação, onde um F1-Score atinge seu melhor valor em 1 (precisão e recall perfeitos) e pior em 0.
- **Precisão (**knn precision**):** A proporção de predições positivas corretas do total de previsões positivas feitas.
- **Matriz de Confusão (**knn confusion matrix**):** Uma tabela que é frequentemente usada para descrever o desempenho de um modelo de classificação.

A matriz de confusão fornece insights valiosos sobre a natureza dos erros cometidos pelo modelo, permitindo identificar se o modelo está confundindo uma classe com outra.

Árvores de Decisão: A Árvore de Decisão foi implementada usando o algoritmo DecisionTreeClas- sifier da biblioteca sklearn.tree, configurado com um critério 'gini' para medir a qualidade das divisões, uma divisão mínima de amostras de 2 para o número mínimo de amostras necessárias para dividir um nó interno e uma profundidade máxima de 11, que limita a profundidade máxima da árvore. O conjunto de dados foi dividido em subconjuntos de treinamento e teste, onde o modelo foi treinado com o subconjunto de treinamento usando o método de ajuste e as previsões foram feitas no subconjunto de testes.

O desempenho do modelo da Árvore de Decisão foi avaliado por meio de métricas como acurácia, F1-Score e precisão, obtidas com as funções accuracy score, f1 score e precision score da biblioteca sklearn.metrics. Uma matriz de confusão foi gerada com a função de matriz de confusão para visualizar a performance do classificador em termos de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. A matriz de confusão fornece insights valiosos para a interpretação do modelo, especialmente em relação ao equilíbrio entre sensibilidade e especificidade.

Floresta Aleatória: A implementação e previsão usando o modelo de Floresta Aleatória foram realizadas seguindo etapas cuidadosamente definidas para a classificação do conjunto de dados. O modelo foi estabelecido com base no RandomForestClassifier da biblioteca scikit-learn, uma escolha popular devido à sua eficácia no manuseio de conjuntos de dados para classificação e regressão.

Inicialmente, o modelo de Floresta Aleatória foi configurado com 100 árvores de decisão (n estimadores = 100), utilizando a entropia como critério para medir a qualidade de uma divisão. O estado aleatório (estado aleatório) foi ajustado como 0 para garantir a reprodutibilidade do modelo.

O treinamento modelo foi realizado com o conjunto de treinamento (X train, y train), onde o modelo aprendeu a identificar padrões e relações nos dados indicativos do desfecho diagnóstico de diabetes.

Após o treinamento, o modelo foi utilizado para fazer predições no conjunto teste (teste X), resultando em um vetor de predições (predições rf).

Para avaliar o desempenho do modelo Floresta Aleatória, várias métricas estatísticas foram calculadas para fornecer uma avaliação do modelo Floresta Aleatória:

- A acurácia (precisão rf) mediu a proporção de predições corretas relativas a todas as previsões feitas, fornecendo uma visão geral da eficácia do modelo.
- O escore F1 (rf f1) forneceu uma medida de teste de precisão, combinando precision e recall em uma única métrica, o que é particularmente útil quando as classes estão desequilibradas.
- A precisão (rf precision) avaliou a acurácia das predições positivas feitas pelo modelo.
- A matriz de confusão (matriz de confusão rf) ofereceu uma visão detalhada do desempenho do modelo, indicando onde o modelo está confundindo as classes.

3 CONSIDERAÇÕES ÉTICAS

Embora os dados estejam disponíveis publicamente e não contenham informações pessoalmente identificáveis, todas as práticas recomendadas para a ética em pesquisa foram seguidas. Isso inclui a anonimização de qualquer informação potencialmente identificável e a confir- mação de que o uso dos dados está em conformidade com os termos de uso estabelecidos pelo repositório Kaggle e regulamentos de dados relevantes.

3.1 LIMITAÇÕES

As limitações do estudo incluem a especificidade da população do conjunto de dados (mulheres de origem Pima com 21 anos ou mais), que pode não ser generalizável para outras populações. Além disso, a qualidade dos dados e a representatividade das variáveis podem influenciar os resultados dos algoritmos de aprendizado de máquina. Outras limitações devem ser consideradas na interpretação dos

resultados deste estudo. O banco de dados utilizado não faz distinção entre os diferentes tipos de diabetes (tipo 1, tipo 2 e gestacional) nos casos positivos, rotulando-os genericamente como positivos para a doença. Essa ausência de diferenciação impede uma análise mais aprofundada que poderia levar a insights específicos para cada tipo de diabetes e suas nuances fisiológicas e epidemiológicas.

Outra limitação significativa é o número de instâncias no banco de dados, que compreende 768 casos. Esse tamanho amostral, embora suficiente para realizar uma análise preliminar e desenvolver modelos preditivos, pode não ser grande o suficiente para capturar toda a heterogeneidade e complexidade associadas à condição diabética. O volume limitado de dados pode afetar a capacidade dos algoritmos de aprendizado de máquina de generalizar suas previsões para uma população mais ampla, potencialmente reduzindo a aplicabilidade prática e a precisão das conclusões tiradas deste estudo.

Essas limitações ressaltam a necessidade de cautela na generalização dos resultados obtidos e sugerem a importância de estudos futuros que incluam conjuntos de dados mais compreensivos e detalhados. Tais estudos devem permitir a distinção entre diferentes tipos de diabetes e considerar uma amostra mais representativa da população em geral.

4 RESULTADOS E DISCUSSÃO

Esta seção discute os resultados obtidos a partir da aplicação de algoritmos de aprendizado de máquina na previsão de diabetes e explora o impacto de valores faltantes e a contribuição da Análise de Componentes Principais (ACP).

4.1 IMPACTO DA IMPUTAÇÃO DE VALOR FALTANTE NA ANÁLISE PREDITIVA

A imputação de valores faltantes é uma etapa crítica na preparação de dados para análise predicativa. No presente estudo, variáveis importantes como IMC, Glicemia e Pressão Arterial continham valores faltantes representados por '?', conforme indicado na Tabela 1. Esses dados foram imputados com a média da variável correspondente, uma abordagem tradicional que visa minimizar o impacto na distribuição global dos dados.

A decisão de utilizar a média para imputação baseou-se na premissa de que os dados faltantes são MCAR (Missing Completely At Random). No entanto, essa presunção nem sempre se sustenta, e sua aplicação deve ser encarada com cautela. Embora essa técnica seja eficiente e de fácil implementação, ela pode levar a uma subestimação da variabilidade e potenciais vieses na estimação do modelo, especialmente se o mecanismo de dados faltantes estiver relacionado à própria variável faltante.

A avaliação do modelo considerou a influência potencial da imputação na acurácia predictiva, com análises adicionais realizadas para validar a imputação. A análise dos resultados indicou que, apesar da imputação, os modelos mantiveram desempenho adequado, sugerindo que a estratégia de

imputação empregada não introduziu um viés significativo que afetou negativamente a capacidade preditiva dos modelos nesse contexto específico.

Tabela 1. Resumo dos valores em falta

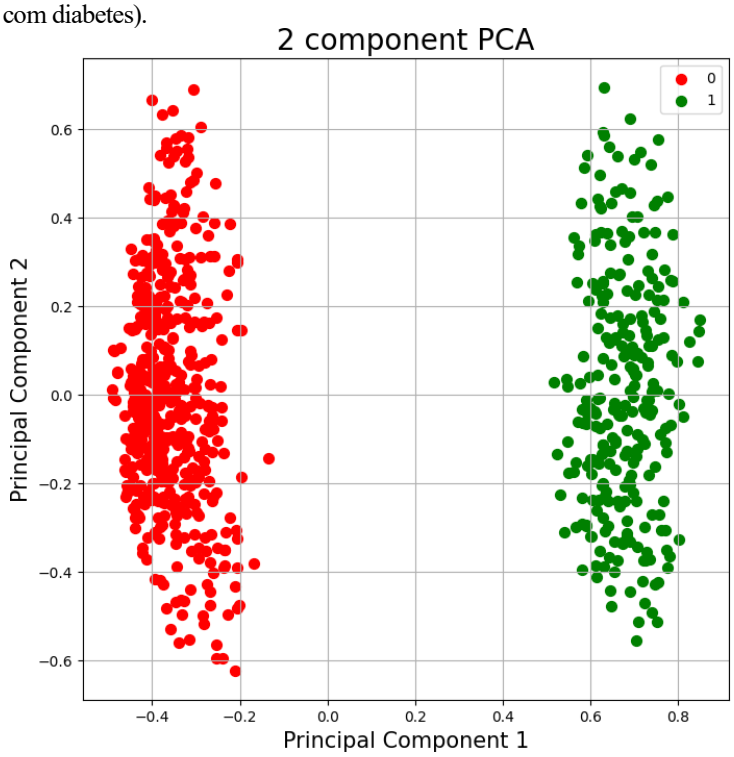
Variável	Valores ausentes
Número de gestações	0
Glicose	5
Tensão arterial	35
Espessura da pele	0
Insulina	0
IMC	11
Diabetes Função Pedigree	0
Idade	0
Resultado	0

4.2 INTERPRETAÇÃO DA ACP

A Análise de Componentes Principais (ACP) foi aplicada para reduzir a dimensionalidade do conjunto de dados e identificar as variáveis mais significativas que contribuem para a variação dos dados de pacientes com e sem diabetes. A Figura 13 ilustra a projecção dos dados em dois componentes principais. Observa-se que os pacientes são agrupados distintamente ao longo do Primeiro Componente Principal, o que sugere que esse eixo capta uma variação significativa relacionada ao estado de diabetes.

A mínima sobreposição entre os grupos na Figura 13 indica que o modelo de ACP conseguiu extrair características relevantes diferenciando pacientes diabéticos de não diabéticos. Esse resultado justifica o uso da ACP como etapa preliminar da análise preditiva, pois proporciona uma simplificação do espaço de feição, ao mesmo tempo em que retém as informações mais relevantes para a classificação.

Fig. 13. Distribuição dos dados nos dois principais componentes da ACP, demonstrando a separação entre pacientes com e sem diabetes (0: sem diabetes, 1: com diabetes).



A interpretação dos componentes principais em relação às variáveis originais é uma etapa subsequente essencial. Enquanto o Primeiro Componente Principal pode estar associado a fatores como níveis glicêmicos e IMC, o Segundo Principal O componente pode representar outras variáveis clínicas. Análises futuras poderiam se concentrar na carga de cada variável nos componentes principais para entender melhor como cada característica contribui para a condição de diabetes.

4.3 COMPARAÇÃO DO DESEMPENHO DO CLASSIFICADOR

A avaliação comparativa dos classificadores revelou variações notáveis em sua perfor- mância, como demonstrado na Tabela 2. O Random Forest emergiu como o modelo mais preciso, alcançando as pontuações mais altas em todas as métricas consideradas. Especificamente, uma acurácia de 0,86, F1-Score de 0,86 e precisão de 0,87 sugerem maior confiabilidade desse algoritmo na classificação correta dos pacientes. As Redes Neurais Artificiais também apresentaram desempenho robusto, com escores consistentes de 0,81 em acurácia, F1-Score e precisão, indicando sua capacidade de modelar as complexidades do conjunto de dados.

Tabela 2. Comparação do desempenho de diferentes classificadores

Classificador	**Precisão F1-Score Precisão**	**Classificador**	**Precisão F1-Score Precisão**
KNN	0.76	0.78	0.78
SVM	0.79	0.81	0.79
ANN	0.81	0.81	0.81
Árvore de Decisão	0.80	0.78	0.78
Floresta aleatória	0.86	0.86	0.87

Esses resultados indicam que métodos mais complexos capazes de captar interações não lineares entre variáveis, como a Floresta Aleatória, podem ser mais adequados para esse tipo de análise de dados médicos. No entanto, é importante ressaltar que a escolha do classificador não deve ser baseada apenas em métricas de desempenho, mas também deve considerar a interpretabilidade do modelo e o contexto clínico em que será aplicado.

4.4 INTERPRETAÇÃO DE MATRIZES DE CONFUSÃO

As matrizes de confusão para cada classificador foram analisadas para avaliar sua capacidade de identificar corretamente os casos de diabetes. Conforme ilustrado nas tabelas (3,4,5,6 e 7) abaixo, o classificador Floresta Aleatória apresentou menor incidência de falsos negativos, destacando sua eficiência no reconhecimento dos casos positivos da doença. Esse é um resultado significativo, pois, na prática médica, minimizar os falsos negativos é fundamental para garantir que os pacientes recebam o tratamento necessário.

Tabela 3. Matriz de Confusão - KNN

	Não-diabéticos (0)	**Diabético (1)**
Não diabético previsto (0)	111	45
Diabético previsto (1)	30	114

Tabela 4. Matriz de Confusão - SVM

	Não-diabéticos (0)	**Diabético (1)**
Não diabético previsto (0)	127	29
Diabético previsto (1)	36	108

Tabela 5. Matriz de Confusão - ANN

	Não-diabéticos (0)	**Diabético (1)**
Não diabético previsto (0)	125	31
Diabético previsto (1)	25	119

Tabela 6. Matriz de Confusão - Árvore de Decisão

	Não-diabéticos (0)	**Diabético (1)**
Não diabético previsto (0)	114	42
Diabético previsto (1)	19	125

Tabela 7. Matriz de Confusão - Floresta Aleatória

	Não-diabéticos (0)	**Diabético (1)**
Não diabético previsto (0)	123	33
Diabético previsto (1)	10	134

O classificador KNN exibiu um balanço relativamente bom entre verdadeiros positivos e verdadeiros negativos, enquanto o SVM e a Árvore de Decisão tenderam a classificar mais casos como negativos, como indicado pelo maior número de falsos negativos. Em contraste, a RNA demonstrou um compromisso efetivo entre sensibilidade e especificidade, evidenciado pela proporção de verdadeiros positivos e verdadeiros negativos.

A análise detalhada das matrizes de confusão sugere que a Floresta Aleatória pode ser mais adequada para o diagnóstico de diabetes no conjunto de dados estudado, fornecendo uma base para a seleção de algoritmos em futuras implementações clínicas.

5 CONSIDERAÇÕES GERAIS

A escolha de um classificador adequado para o diagnóstico médico deve equilibrar a acirracia e a sensibilidade. Os modelos devem minimizar tanto os falsos positivos, que podem levar a procedimentos médicos desnecessários, quanto os falsos negativos, que podem resultar em atrasos no tratamento. Neste estudo, a Floresta Aleatória destacou-se, sugerindo sua viabilidade para a detecção do diabetes. Além do desempenho estatístico, a clareza na interpretação dos resultados é essencial, reforçando o valor dos algoritmos explicáveis na prática médica, onde as decisões baseadas em dados devem ser transparentes e justificáveis.

6 CONCLUSÃO

Esta investigação revelou que, embora a seleção de um classificador para predição de diabetes deva ser informada por métricas de desempenho, a aplicabilidade clínica e a interpretabilidade dos resultados são igualmente cruciais. O Random Forest, destacando-se nos critérios estatísticos, é sugerido como uma opção robusta devido à sua capacidade de minimizar falsos negativos, o que é vital para garantir a identificação adequada dos pacientes que necessitam de intervenção. A inclusão da ACP como parte do processo de modelagem preditiva foi validada, contribuindo para uma compreensão mais profunda de características influentes e apoiando a seleção de características relevantes para futuras iterações de modelos de previsão. Este estudo ressalta a importância de uma abordagem holística na análise preditiva de saúde, priorizando não apenas a acurácia, mas também a usabilidade clínica e a transparência na tomada de decisões médicas.

REFERÊNCIAS

de Oliveira Santos, Givanildo, et al. ”Exerc´ıcios f´ısicos e diabetes mellitus: Revis˜ao.” Brazilian Journal of Development, vol. 7, no. 1, 2021, pp. 8837-8847.

da Silva, Maria Eduarda, et al. ”Promo¸c˜ao da homeostase glicˆemica em indiv´ıduos diab´eticos atrav´es do exerc´ıcio f´ısico: Uma revis˜ao narrativa.” Brazilian Journal of Development, vol. 6, no. 7, 2020, pp. 44576-44585.

WHITING, David R. et al. ”IDF diabetes atlas: global estimates of the prevalence of diabetes for 2011 and 2030.” Diabetes research and clinical practice, vol. 94, no. 3, 2011, pp. 311-321.

Mendes, Alana Caroline Alves, et al. ”Promo¸c˜ao em sau´de para condutas de h´abitos saud´aveis para redu¸cao de diabetes tipo II e hipertens˜ao na aten¸c˜ao prim´aria.” Revista JRG de Estudos Acadˆemicos, vol. 6, no. 13, 2023, pp. 1773-1792.

Cardozo, Glauco. ”Um modelo computacional utilizando t´ecnicas de machine learning e exames laboratoriais de rotina na triagem e apoio ao diagn´ostico de diabetes mellitus.” 2022.

Monteiro, Rosangela, et al. ”INTELIGEˆNCIA ARTIFICIAL, DEEP LEARNING, MACHINE LEARNING, REDES NEURAIS NA MEDICINA E BIOMARCADORES VOCAIS: CONCEITOS, ONDE ESTAMOS E PARA ONDE VAMOS.” Rev Soc Cardiol Estado de S˜ao Paulo, vol. 32, no. 1, 2022, pp. 11-17.

Paixao, Gabriela Miana de Mattos, et al. ”Machine Learning na Medicina: Revis˜ao e Aplicabilidade.” Arquivos Brasileiros de Cardiologia, vol. 118, 2022, pp. 95-102.

Surden, Harry, Tourinho Leal, Saul, e Silva Neto, Wilson Seraine da. ”Machine learning e o direito.” Suprema-Revista de Estudos Constitucionais, vol. 3, no. 1, 2023, pp. 353-389.

Ramos, Ma´ıra Catharina, et al. ”Big Data e Intelig ̂encia Artificial para pesquisa translacional na Covid-19: revis˜ao r´apida.” Sau´de em Debate, vol. 46, 2023, pp. 1202-1214.

Dias, Caio Eduardo, Saqui, Diego, e Moreira, Heber Rocha. ”Desenvolvimento de um aplica- tivo para classificaca˜o de doen¸cas e pragas em folhas de caf´e utilizando deep learning.” 15º JORNADA CIENT´IFICA E TECNOLO´GICA E 12º SIMPO´SIO DE PO´S-GRADUAC¸ A˜O DO IFSULDEMINAS, vol. 15, no. 3, 2023.

Souza, Ewerton Pacheco de, et al. ”Aplica¸c˜oes do Deep Learning para diagn´ostico de doen¸cas e identifica¸cao de insetos vetores.” Sau´de em Debate, vol. 43, 2020, pp. 147-154.

Vitoria, Sergio Ricardo Pacheco da. ”Machine learning e an´alise preditiva em sau´de: um estudo de caso sobre detec¸cao de anomalias em contas m´edicas do Ex´ercito.” 2022.

Souza, Amanda Cristina Eleut´erio, and Saqui, Diego. ”ME´TODO DE PRE´-DIAGNO´STICO DA COVID-19 E PNEUMONIA UTILIZANDO IMAGENS DE RADIOGRAFIA DO TO´RAX E CNN.” 15º JORNADA CIENT´IFICA E TECNOLO´GICA E 12º SIMPO´SIO DE PO´S- GRADUAC¸ A˜O DO IFSULDEMINAS, vol. 15, no. 3, 2023.

Santos, Jefferson Pacheco dos. ”Proposta de um sistema para avalia¸c˜ao de riscos de infec¸c˜ao do s´ıtio ciru´rgico utilizando t´ecnicas de intelig ̂encia artificial.” 2021.

Medeiros et al. ”Short-term COVID-19 forecast for latecomers.” arXiv preprint arXiv:2004.07977, 2020.

Vaishya, R., Javaid, M., Khan, I. H., e Haleem, A. ”Artificial intelligence (AI) applications for COVID-19 pandemic.” Diabetes and Metabolic Syndrome: Clinical Research and Reviews, 2020.

Stakoviak, Flavio Henrique Moura. ”Inova¸c˜ao em acesso a` informa¸c˜ao em biotecnologia: desenvolvimento de um mecanismo de busca inteligente baseado em processamento de linguagem natural.” 2023.

Fonseca, Afonso Ueslei, et al. ”Diagnosticando Tuberculose com Redes Neurais Artificiais e Recursos BPPC.” Journal of Health Informatics, vol. 15, Edi¸c˜ao Especial, 2023.

Costa, Oberdan, and Gouveia, L. ”Uma proposta para um Sistema Inteligente de Previs˜ao do Risco de Doen¸cas Cronicas.” In GASPAR, J. et al., eds. Sistemas Inteligentes para a Sau´de: desafios da ´etica e governan¸ca. Anais do CBIS, 2022, pp. 243-248.

DataCamp. ”Support Vector Machines with Scikit-learn.” Acessado em 02 de mar¸co de 2024. https://www.datacamp.com/community/tutorials/.

Veiga, D. M., & Ferreira, D. ”Ser´a Poss´ıvel Melhorar O Diagn´ostico Da Icter´ıcia Neonatal? Aplica¸cao De T´ecnicas De Data Mining.” 2011.

Boswell, Dustin. ”Introduction to support vector machines.” Departamento de Ciˆencia da Computa¸cao e Engenharia, Universidade da Calif´ornia, San Diego, vol. 11, 2002.

Teixeira, Pedro Miguel Bento. ”Classifica¸c˜ao autom´atica de termogramas do p´e diab´etico usando t´ecnicas de Machine Learning.” 2021.

Fix, E. and Hodges, J.L. ”Discriminatory Analysis, Nonparametric Discrimination: Consis- tency Properties.” Technical Report 4, USAF School of Aviation Medicine, Randolph Field, 1951.

Urso, A., Fiannaca, A., La Rosa, M., Rav`ı, V., Rizzo, R. ”Data mining: Prediction meth- ods.” In: Encyclopaedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, Elsevier, 2018, pp. 413–430. https://doi.org/10.1016/B978-0-12-809633-8.20462-7.

Olivera, Andr´e Rodrigues. ”Compara¸c˜ao de algoritmos de aprendizagem de m´aquina para construca˜o de modelos preditivos de diabetes n˜ao diagnosticado.” 2016.

Maglogiannis, Ilias G., ed. ”Emerging artificial intelligence applications in computer engi- neering: real world AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies.” Vol. 160. IOS Press, 2007.

QUILAN, J. Ross. ”C4.5: Programs for Machine Learning.” San Mateo, CA: Morgan Kauf- mann Publishers, 1993. 312 p.

Carvalho, Deborah Ribeiro, Dallagassa, Marcelo Rosano, e da Silva, Sandra Honorato. ”Uso de t´ecnicas de minera¸cao de dados para a identifica¸c˜ao autom´atica de benefici´arios propensos ao diabetes mellitus tipo 2.” Informa¸c˜ao & Informa¸c˜ao, vol. 20, no. 3, 2015, pp. 274-296.

Pearson, Karl. ”LIII. On lines and planes of closest fit to systems of points in space.” The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 2, no. 11, 1901, pp. 559-572.

Ferreira, M´arcia Miguel Castro. ”Quimiometria III-Revisitando a an´alise explorat´oria dos dados multivariados.” Qu´ımica Nova, vol. 45, 2023, pp. 1251-1264.

Porreca, Paloma Priscila, et al. ”Efeitos do uso da m´ascara facial no exerc´ıcio durante a pandemia da covid-19: uma an´alise de componente principal (PCA).” Biol´ogicas And Sau´de, vol. 11, no. 38, 2021, pp. 13-14.

dos Santos Nunes, Florbela. ”Previs˜ao de nu´mero de dias de internamento em doentes diab´eticos-Uma abordagem de Machine Learning.” 2020.

Galvao, Noemi Dreyer, and Marin, Heimar de F´atima. ”T´ecnica de minera¸c˜ao de dados: uma revis˜ao da literatura.” Acta Paulista de Enfermagem, vol. 22, 2009, pp. 686-690.

Fayyad, Usama; Piatetski-Shapiro, Gregory; Smyth, Padhraic. ”The KDD Process for Extract- ing Useful Knowledge from Volumes of Data.” In: Communications of the ACM, Nov. 1996, pp. 27-34.

Guerra-Hernandez, Alejandro, Mondrag´on-Becerra, Rosbelda, and Cruz-Ram´ırez, Nicandro. ”Explorations of the BDI Multi-agent support for the Knowledge Discovery in Databases Pro- cess.” Research in Computing Science, vol. 39, 2008, pp. 221-238.

Oliveira, Paulo, Rodrigues, Fatima, and Henriques, P. ”Limpeza de dados: Uma vis˜ao geral.” Data Gadgets, 2004, pp. 39-51.

da Costa Cortes, S´ergio, Porcaro, Rosa Maria, and Lifschitz, S´ergio. ”Minera¸c˜ao de dados-funcionalidades, t´ecnicas e abordagens.” PUC, 2002.

Rigo, Sandro Jos´e, et al. ”Aplica¸coes de Minera¸c˜ao de Dados Educacionais e Learning Analytics com foco na evasao escolar: oportunidades e desafios.” Revista Brasileira de Inform´atica na Educa¸cao, vol. 22, no. 01, 2014, p. 132.

Mello, Joao Alexandre Bonin de. ”Uma proposta de modelo de dados para suporte ao proces- samento transacional e de data warehouse simultaneamente.” 2002.

Junior, Guanis B. Vilela, et al. ”M´etricas utilizadas para avaliar a efici^encia de classificadores em algoritmos inteligentes.” Revista CPAQV–Centro de Pesquisas Avan¸cadas em Qualidade de Vida, vol. 14.2, 2022, p. 2.

Milani, Anna, et al. ”A Deep Learning Application for Psoriasis Detection.” Anais do XX Encontro Nacional de Intelig^encia Artificial e Computacional, SBC, 2023.

Souza, Emanuel G. ”Entendendo o que ´e Matriz de Confus˜ao com Python - Data Hackers - Medium.” Medium, 2019. Dispon´ıvel em: https://medium.com/data-hackers/ entendendo-o-que-%C3%A9-matriz-de-confus%C3%A3o-com-python-114e683ec509. Acesso em: 2 mar. 2024.

Clavera, Walter. ”Aprendizado de M´aquina (Machine Learning).” SEJAT- ECH e FALETECH, 2019. Dispon´ıvel em: https://www.redesdesaude.com.br/ aprendizado-de-maquina-machine-learning/. Acesso em: 2 mar. 2024.

Dijkinga, Fernando Jean. ”Utiliza¸c˜ao de aprendizagem supervisionada de m´aquina para predi¸cao de valores gen´eticos com base em duas gera¸c˜oes de ascendentes.” Research, Society and Development, vol. 12, no. 6, 2023, e2812641904-e2812641904.

Fernandes, Fernando Timoteo, and Chiavegatto Filho, Alexandre Dias Porto. ”Perspectivas do uso de minera¸cao de dados e aprendizado de m´aquina em sau´de e seguran¸ca no trabalho.” Revista Brasileira de Sau´de Ocupacional, vol. 44, 2019, e13.

Fernandes, Fernando Timoteo. ”Machine learning em sau´de e seguran¸ca do trabalhador: perspectivas, desafios e aplica¸coes.” Disserta¸c˜ao de Mestrado, Universidade de S˜ao Paulo, 2021.

Oliveira, David Fernandes Neves. ”Interpretabilidade de modelos de aprendizado profundo aplicados ao diagn´ostico e progn´ostico n˜ao supervisionado de falhas.” Disserta¸c˜ao de Mestrado, Universidade de S˜ao Paulo, 2022.

SANCHES, Jonas Rodrigues. ”An´alise comparativa dos efeitos morfofuncionais e ultraestru- turais dos diabetes mellitus tipo 1 e diabetes mellitus tipo 2 no f´ıgado de ratos.” 2023.

Giroldo, Julio Cesar, and Gabriel, Anderson Luiz. ”Diabetes mellitus tipo 2: a interven¸c˜ao da atividade f´ısica como forma de aux´ılio e qualidade de vida.” Revista Carioca de Educa¸c˜ao F´ısica, vol. 15, no. 1, 2020, pp. 28-39.

Silva, Tha´ıza Morais da. ”Revisao bibliogr´afica sobre o diagn´ostico e o tratamento do diabetes mellitus.” 2019.

Costa, Fabrycianne Gon¸calves. ”Bem-estar subjetivo, resilicircencia e representa¸c˜oes sociais no contexto do diabetes mellitus.” 2017.

Waldman, Beatriz Ferreira. ”Envelhecimento bem-sucedido: uma metodologia de cuidado a pessoas com diabetes mellitus.” 2006.

OCRC, CEPID. ”A rela¸cao entre Diabetes tipo 2 e a maior gravidade da COVID-19: o que sabemos?” Sobre Peso — Dicas e segredos para man- ter o peso sob controle! 2020. Dispon´ıvel em: https://www.sobrepeso.com.br/ a-relacao-entre-diabetes-tipo-2-e-a-maior-gravidade-da-covid-19-o-que-sabemos/. Acesso em: 2 mar. 2024.

SBD. ”Tipos de Diabetes - Sociedade Brasileira de Diabetes.” Sociedade Brasileira de Diabetes, 2024. Dispon´ıvel em: https://diabetes.org.br/tipos-de-diabetes/#:~:text=Cerca%20de%2090%25%20das%20pessoas,atividade%20f%C3%ADsica%20e%20planejamento%20alimentar.. Acesso em: 2 mar. 2024.

Dadamos, Fernando Magalh˜aes. ”Fatores cr´ıticos de sucesso para ado¸c˜ao da LGPD nas em- presas brasileiras: um estudo Delphi com especialistas.” Disserta¸c˜ao de Mestrado, 2022.

Vidal, Maciel Calebe, and Machado, Arthur Cisotto. ”Intelig circencia Artificial Explic´avel (XAI) na a´rea m´edica.” 2023.

SaudeBusiness. ”Como vencer a resist circencia tecnol´ogica em m´edicos e pacientes.” saudebusiness.com, 2021. Dispon´ıvel em: https://www.saudebusiness.com/ti-e-inova%C3%A7%C3%A3o/como-vencer-resist%C3%AAncia-tecnol%C3%B3gica-em-m%C3%A9dicos-e-pacientes. Acesso em: 2 mar. 2024.

Chauhan, Aman. ”Predict Diabetes.” Kaggle.com, 2023. Dispon´ıvel em: https://www.kaggle.com/datasets/whenamancodes/predict-diabities/data. Acesso em: 2 mar. 2024.

Breiman, Leo. ”Random forests.” Machine Learning, vol. 45, 2001, pp. 5–32.

Junior, ICMC. ”Random Forest — Entenda Agora Esse Algoritmo Poderoso.” 2021. Dispon´ıvel em: https://icmcjunior.com.br/random-forest. Acessado em 16 de dezembro de 2023.

de Alvarenga Ju´nior, Wagner Jos´e. ”M´etodos de otimiza¸c˜ao hiperparam´etrica: um estudo comparativo utilizando a´rvores de decis˜ao e florestas aleat´orias na classifica¸c˜ao bin´aria.” 2018.

Ribeiro, Eduardo, et al. ”Modelo preditivo para a prolifera¸c˜ao do Aedes aegypti em Itaja´ı (Santa Catarina): Uma abordagem integrando fatores clim´aticos locais e globais.” Estrab˜ao, vol. 5, 2024, pp. 81-91.

Santos, Andraws Steve. ”Modelos de machine learning para identifica¸c˜ao de clientes com risco acrescido de branqueamento de capitais e financiamento do terrorismo.” 2023.

Santos, J. A. A. and Chaucoski, Y. ”Previs˜ao do consumo de energia el´etrica na regi˜ao sudeste: um estudo de caso usando sarima e lstm.” Revista Cereus, vol. 12, no. 4, 2020, pp. 93–104.

Rodrigues, Gustavo, and Kreutz, Diego. ”Modelo preditivo para classifica¸c˜ao de risco de o´bito de pacientes com COVID-19 utilizando dados abertos.” Anais do XXII Simp´osio Brasileiro de Computa¸cao Aplicada a` Sau´de, SBC, 2022.

Cardoso, Isaac. ”T´ecnicas de otimiza¸c˜ao e m´etricas de avalia¸c˜ao aplicadas a machine learning.” 2022.

Maniezzo, Giovanni Vallim, Oliveira, Jorge Luiz Arantes de, e Pereira, Maria Eduarda Riskalla. ”Processamento e classificaca˜o de sinais de EEG por meio de machine learning para identifica¸c˜ao de emo¸coes b´asicas. 2022.”

Fernandes, Lucinara Kecia Silva. ”Aprendizagem de m´aquina para previs˜ao de predisposi¸c˜ao ao medo do crime.” 2022.

Goodfellow, I., Bengio, Y., e Courville, A. ”Deep Learning.” MIT Press, 2016. Dispon´ıvel em: http://www.deeplearningbook.org. Acesso em: 3 mar. 2024.