Politécnico
de Viseu

Escola Superior
de Tecnologia
e Gestão de Viseu

O uso de machine learning na prevenção de diabetes

Maria Alice Holanda Lopes

Trabalho de Projeto

Mestrado em Engenharia Informática - Sistemas de Informação

Trabalho efetuado sob a orientação de
Professora Doutora Cristina Wanzeller
Professora Doutora Joana Fialho

Março de 2025

Politécnico
de Viseu

Escola Superior
de Tecnologia
e Gestão de Viseu

O uso de machine learning na prevenção de diabetes

Maria Alice Holanda Lopes

Trabalho de Projeto

Mestrado em Engenharia Informática - Sistemas de Informação

Trabalho efetuado sob a orientação de

Professora Doutora Cristina Wanzeller
Professora Doutora Joana Fialho

Março de 2025

Agradecimentos

Gostaria, primeiramente, de expressar minha gratidão aos meus pais, que me proporcionaram não apenas o apoio necessário ao desenvolvimento desta tese, mas que sempre estiveram presentes em todas as etapas da minha vida, oferecendo incentivo constante e incondicional.

Agradeço também, às professoras Doutoras Ana Cristina Wanzeller Guedes de Lacerda e Joana Rita Silva Fialho pela orientação, paciência e conhecimento compartilhado, fundamentais para a realização deste trabalho.

Por fim, meu sincero obrigado aos amigos que estiveram ao meu lado durante este percurso académico.

Muito obrigado a todos.

Resumo

A *Diabetes Mellitus* é uma das doenças crónicas com crescimento mais acelerado no mundo, demandando soluções eficazes para diagnóstico e prevenção. Neste contexto, técnicas de *Machine Learning* (ML) apresentam potencial significativo na identificação de padrões relevantes ao controlo da doença. Este estudo utilizou a metodologia CRISP-DM para analisar dados do *Diabetes Health Indicators Dataset*, contendo informações sociodemográficas, clínicas e comportamentais.

Na fase de pré-processamento, aplicou-se o equilíbrio de classes por subamostragem (*NearMiss*) devido à baixa proporção de indivíduos diabéticos. Técnicas de seleção de características, como Eliminação Recursiva de Características (RFE) e Análise de Componentes Principais (PCA), foram utilizadas para avaliar a relevância das variáveis e reduzir a dimensionalidade. Avaliaram-se seis modelos: Floresta Aleatória, *Gradient Boosting*, KNN, Regressão Logística, Perceptron Multicamadas (MLP) e Redes Neuronais Recorrentes (RNN).

Os resultados mostraram que o equilíbrio das classes melhorou significativamente o desempenho, destacando-se a RNN, com acurácia acima de 86% e *F1-score* próximo a 0,87. A combinação da seleção RFE com MLP também apresentou resultados robustos. Conclui-se que ML e DL são promissores para priorizar acompanhamento clínico e apoiar políticas públicas, sendo necessário ampliar a representatividade dos dados, incorporar técnicas de *Explainable AI* para maior interpretabilidade, e ajustar limiares decisórios visando minimizar falsos negativos.

Palavras-Chave: Diabetes Mellitus, Machine Learning, Deep Learning, Redes Neuronais Recorrentes, Seleção de Características

Abstract

Diabetes Mellitus is one of the fastest growing chronic diseases in the world, requiring effective solutions for diagnosis and prevention. In this context, Machine Learning (ML) techniques have significant potential for identifying patterns relevant to disease control. This study used the CRISP-DM methodology to analyze data from the *Diabetes Health Indicators Dataset*, containing sociodemographic, clinical and behavioral information.

In the pre-processing phase, class balancing by undersampling (*NearMiss*) was applied due to the low proportion of diabetic individuals. Feature selection techniques, such as *Recursive Feature Elimination* (RFE) and *Principal Component Analysis* (PCA), were used to assess the relevance of the variables and reduce dimensionality. Six models were evaluated: Random Forest, Gradient Boosting, KNN, Logistic Regression, Multilayer Perceptron (MLP) and Recurrent Neural Networks (RNN).

The results showed that class balancing significantly improved performance, with RNN standing out with accuracy above 86% and an F1-score near 0.87. The combination of RFE feature selection with MLP also yielded robust results. It is concluded that ML and DL are promising for prioritizing clinical follow-up and supporting public policies. However, it is necessary to increase data representativeness, incorporate Explainable AI techniques for greater interpretability, and adjust decision-making thresholds aiming to minimize false negatives.

Keywords: Diabetes Mellitus, Machine Learning, Deep Learning, Recurrent Neural Networks, Feature Selection

Índice

Lista de Tabelas

Lista de Figuras

Lista de Códigos Fonte

Lista de Acrónimos

BRFSS	*Behavioral Risk Factor Surveillance System*
CRISP-DM	*Cross-Industry Standard Process for Data Mining*
DL	*Deep Learning*
DM1	*Diabetes Tipo 1*
DM2	*Diabetes Tipo 2*
GRU	*Gated Recurrent Unit*
IMC	*Índice de massa corporal*
KNN	*K-nearest neighbors*
LSTM	*Long Short-Term Memory*
ML	*Machine Learning*
MLP	Perceptron Multicamadas
PCA	Análise de Componentes Principais
PIB	Produto Interno Bruto
RFE	Eliminação Recursiva de Características
RNN	Redes Neuronais Recorrentes

Capítulo 1

Introdução

Segundo dados de 2021, os investimentos anuais em saúde representam 10,3% do Produto Interno Bruto (PIB) global (Sterlin, 2024). A inovação tecnológica como o uso da *Machine Learning* (ML) permitiu uma transformação na área da saúde. A análise de grande quantidade de dados na área da saúde pode mudar o panorama da prevenção, do diagnóstico e do tratamento de doenças.

Conforme a pesquisa, publicada no IDF diabetes atlas (Federation, 2019), pode-se afirmar que a diabetes é um dos problemas sanitários de mais rápido crescimento do século XXI. Com base nos resultados de 2019, estimava-se que 463 milhões de pessoas sofriam de diabetes (Federation, 2019), sendo esta uma doença metabólica complexa caracterizada pela regulação inadequada da glicose no sangue e suas ramificações significativas na saúde pública (Association, 2021).

A diabetes se divide em dois tipos principais. O tipo 1 resulta da destruição das células produtoras de insulina. O tipo 2 ocorre devido à resistência à insulina e à produção insuficiente dessa hormona (Federation, 2019).

A utilização da ML na prevenção da diabetes tem se destacado como uma ferramenta de impacto para apoiar como a saúde pública e os cuidados médicos analisam e enfrentam a crescente incidência desta doença. Por meio da análise criteriosa de dados clínicos, genéticos, comportamentais e ambientais, o uso da ML tem sido capaz de identificar padrões e prever, com maior precisão, quais indivíduos apresentam maior risco de desenvolver o diabetes. Essa previsão antecipada tem permitido a implementação de medidas preventivas, como mudanças no estilo de vida, orientações nutricionais personalizadas, incentivo à prática de atividades físicas e até

tratamentos direcionados para retardar ou evitar o surgimento da doença (Suryasa, Rodríguez-Gámez e Koldoris, 2021).

Além de promover benefícios diretos à população, como a melhoria da qualidade de vida e a prevenção de complicações associadas ao diabetes, essa abordagem traz impactos significativos para os sistemas de saúde, reduzindo a necessidade de intervenções emergenciais, internações hospitalares e tratamentos de longo prazo, que demandam investimento tanto financeiro quanto organizacional. Dessa forma, é possível fazer uma melhor gestão dos recursos disponíveis, direcionando-os para ações mais eficazes e sustentáveis (X. Zhou et al., 2020).

1.1 Motivação

Dado o impacto global da diabetes na qualidade de vida, diversos estudos (Firdous, Wagai e Sharma, 2022) têm explorado o uso de ML para aperfeiçoar os métodos preditivos da doença.

A diabetes é uma condição médica que está associada a vários fatores condicionantes. Portanto, por meio da agregação de elevados níveis de informações médicas, o uso de modelos de ML torna-se alternativa eficaz para a predição correta da condição. Esses modelos têm a capacidade de reconhecer padrões e relações relevantes entre as variáveis, permitindo uma compreensão mais abrangente e precisa dos fatores indicadores de risco para a diabetes. Nesse contexto, é essencial uma análise de alguns dos métodos existentes para entender as limitações e vantagens dos modelos preditivos atuais e como eles podem potencialmente ajudar a identificar outras características mais significativas correlacionadas ao risco de diabetes.

A utilização de modelos de ML para a previsão da diabetes permite identificar indivíduos em risco antes do desenvolvimento da doença, possibilitando intervenções precoces para evitá-la ou, pelo menos, retardar o seu início.

A diabetes pode ser gerida de forma mais eficaz quando os seus fatores de risco são identificados atempadamente, permitindo a adoção de medidas preventivas personalizadas, como alterações no estilo de vida, monitorização contínua e estratégias terapêuticas ajustadas. Estas ações ajudam a evitar a progressão desnecessária da doença e a reduzir complicações preveníveis.

Desta forma, a aplicação de ML na saúde reforça a eficácia dos cuidados médicos no combate à diabetes, otimizando recursos e melhorando os resultados clínicos.

1.2 Objetivos

O objetivo deste trabalho é contribuir para o desafio da previsão precoce da diabetes, por um estudo comparativo de algoritmos de ML utilizados na predição de diabetes, abrangendo tanto modelos supervisionados tradicionais quanto abordagens mais recentes baseadas em *deep learning* (Afsaneh et al., 2022). Além disso, será explorada a aplicação de técnicas de seleção de características *feature selection* para identificar as variáveis mais relevantes nos modelos preditivos, visando melhorar o desempenho da predição.

O trabalho também se propõe a examinar o papel da ML na previsão e gestão da diabetes, demonstrando os principais avanços e os desafios destacados na literatura científica. Essa análise busca demonstrar como os modelos mais precisos podem contribuir para intervenções antecipadas, otimizando o cuidado médico e auxiliando na prevenção do desenvolvimento da doença.

Por fim, o trabalho pretende fornecer direções relevantes para futuras investigações, promovendo a aplicação de ML como uma ferramenta prática e eficaz na antecipação de diagnósticos e mitigação dos impactos da diabetes. Serão identificadas metodologias que possam tornar essas tecnologias mais acessíveis e compreensíveis para profissionais da área médica, ampliando o seu uso e impacto.

1.3 Plano de Trabalhos

Para alcançar os objetivos e direcionar este trabalho aos resultados esperados, foram estabelecidas etapas específicas, representadas na Figura 1.1, que guiaram todo o desenvolvimento.

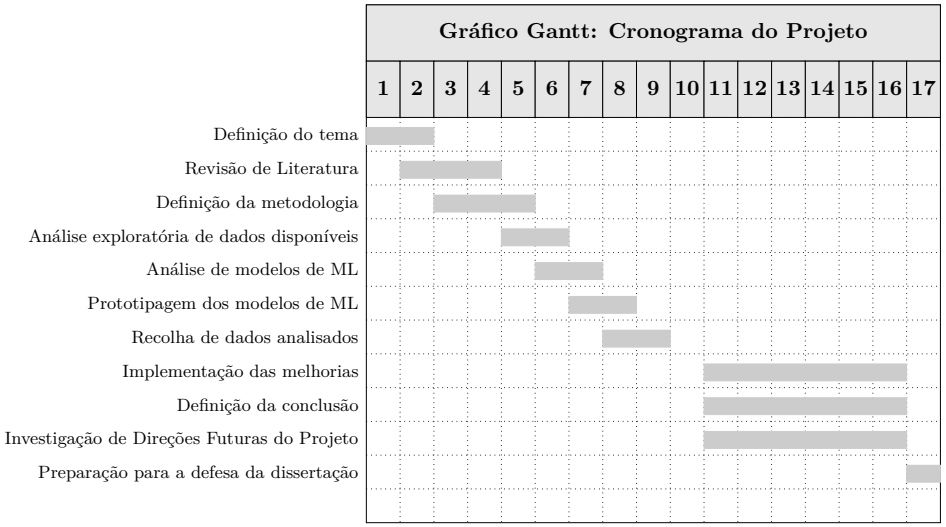

Figura 1.1: Cronograma de tarefas com início a Novembro de 2023 e término a Março de 2025.

O primeiro passo consistiu na definição do tema abordado, baseada em pesquisas sobre temas de grande impacto na área da saúde, amplamente explorados no contexto de ML. Após a escolha do tema, foi realizada uma revisão de literatura para compreender os conceitos, o histórico, os desafios e os trabalhos relacionados.

Para estruturar e organizar o desenvolvimento, foi definida a metodologia aplicada ao projeto, que incluem às etapas de modelagem e aplicação dos modelos de ML. Em seguida, foram identificados e selecionados conjuntos de dados confiáveis e validados para o contexto deste trabalho, os quais passaram por uma análise exploratória detalhada.

Na etapa de análise de modelos, foram investigados algoritmos relevantes para o contexto do trabalho, com base nos resultados da revisão de literatura. A prototipagem foi realizada utilizando os modelos previamente definidos e foram aplicadas as métricas de avaliação para comparar o desempenho de cada modelo. Com base nessa análise, o modelo de melhor desempenho foi selecionado, e seus resultados, bem como os resultados dos demais modelos, foram examinados de forma detalhada.

Após essa etapa, foram elaboradas as conclusões do trabalho. Por fim, foram exploradas direções futuras que este trabalho pode seguir. Todo o processo foi concluído com a organização e preparação da apresentação para a defesa da tese.

Na Figura 1.1 é apresentado um diagrama de Gantt com o Plano de trabalhos.

1.4 Estrutura do documento

A estrutura deste documento é a seguinte:

No capítulo de introdução, é realizado o enquadramento ao tema, destacando a sua relevância na área de estudo. Nesse capítulo, a motivação da pesquisa é definida, apresentam-se os objetivos do projeto e é discutido o plano de trabalho.

No capítulo de revisão de literatura são analisadas as fontes bibliográficas relacionadas ao tema do estudo. São investigados os conceitos, trabalhos relevantes e os principais desafios associados ao tema, fornecendo uma base teórica para o desenvolvimento do projeto.

No capítulo de desenvolvimento, são apresentadas a metodologia e as etapas práticas que conduziram à implementação dos modelos preditivos. Essa secção inclui a análise exploratória do conjunto de dados, a aplicação de técnicas de seleção de características, a avaliação comparativa dos resultados obtidos por diferentes algoritmos de ML e os ajustes realizados nos modelos para otimizar o seu desempenho.

Por fim, no capítulo de conclusões, são sintetizadas as principais conclusões do trabalho, destacando as contribuições para o campo de estudo. Além disso, são sugeridas possíveis direções para pesquisas futuras, apontando novas oportunidades e abordagens para aprofundar o tema investigado.

Capítulo 2

Revisão da Literatura

Neste capítulo apresentam-se os conceitos centrais do projeto, a começar pela caraterização da diabetes, a definição e uso de seleção de características no contexto da diabetes, a definição de *machine learning*, a definição de *deep Learning*, e apresentam-se os modelos usados e trabalhos relacionados ao presente tema.

2.1 Diabetes

A Diabetes Mellitus é a doença crónica que se caracteriza por níveis elevados de glicose no sangue devido a questões na produção ou atuação da insulina. Há duas formas mais comuns de diabetes: tipo 1 (DM1) e tipo 2 (DM2).

Diabetes Tipo 1 (DM1) é causada por uma reação autoimune na qual o sistema imunológico ataca as células beta do pâncreas responsáveis pela produção de insulina. Como resposta, o corpo acaba por produzir quantidades reduzidas ou até nulas de insulina. Pessoas com DM1 precisam administrar insulina diariamente para regular os níveis de glicose no sangue (Federation, 2019).

Já a *Diabetes Tipo 2* (DM2) é a forma mais comum de diabetes e ocorre devido à resistência à insulina e à produção insuficiente desta hormona no organismo. A resistência à insulina prejudica a capacidade das células responderem de forma adequada à insulina, o que resulta em níveis elevados de açúcar no sangue. A DM2 geralmente está relacionada a um estilo de vida sedentário, a uma alimentação pouco nutritiva,

à falta de exercícios físicos e a obesidade. Este tipo de diabetes tem elevada incidência em todo o mundo, sendo identificado um aumento generalizado (Federation, 2019).

Os dois tipos de diabetes estão relacionadas a complicações de saúde que afetam diversos sistemas do corpo humano. As complicações podem acontecer nos microvasos. Quando há danos aos pequenos vasos sanguíneos, podem ocorrer problemas de visão (retinopatia diabética), renais (nefropatia diabética) e neurológicos (neuropatia diabética). Também podem ocorrer complicações nos macrovasos que afetam os grandes vasos sanguíneos e aumentam o risco de doenças cardiovasculares, acidente vascular cerebral e doença arterial periférica (Mansour et al., 2023).

Além das categorias 1 e 2, existem outras formas menos frequentes de diabetes, como a diabetes gestacional – que pode acontecer em mulheres gestantes devido a alterações hormonais, entre outros tipos de diabetes relacionados a condições de saúde preexistentes.

A diabetes é conhecida há séculos. Registos históricos mostram que, por volta de 1500 a.C., o Papiro de Ebers, no Egito Antigo, já mencionava sintomas compatíveis com a doença. Tratamentos antigos incluíam dietas ricas em fibras, trigo e óxido de ferro (Bailey e Day, 1989). Ao longo da história, várias civilizações utilizaram ervas, especiarias e vegetais para controlar os sintomas (Bailey e Day, 1989). Apesar dos avanços científicos, algumas dessas abordagens ainda são utilizadas em regiões com poucos recursos (Bailey e Day, 1989).

O tratamento da diabetes passou por grandes mudanças ao longo dos anos. A mais significativa aconteceu em 1921, com a descoberta da insulina por Frederick Banting e Charles Best. Essa descoberta possibilitou avanços como medicamentos orais, insulinas análogas e tecnologias como bombas de insulina e monitores de glicose em tempo real para gestão dessa condição (Bliss, 1982).

Um dos fatores desafiantes da diabetes é o fator social no seu tratamento. Fatores como status socioeconómico, ambiente, insegurança alimentar, acesso aos cuidados de saúde e apoio social afetam significativamente os riscos e impactos da diabetes. Isto acontece porque indivíduos com menos recursos financeiros têm menor acesso a tratamentos adequados e enfrentam maiores complicações, além das condições de moradia instáveis e com baixa infraestrutura que dificultam a gestão da doença, incluindo o acesso a alimentos saudáveis e locais para atividades físicas (Hill-Briggs et al., 2021).

A identificação e o tratamento corretos da diabetes permitem evitar complicações mais graves no futuro. É essencial monitorizar regularmente os níveis de açúcar no sangue, alterar a alimentação, praticar atividades físicas e, em alguns casos, administrar medicamentos para o controlo da diabetes.

Uma compreensão mais profunda dessa condição e os fatores que mais impactam no avanço da diabetes permite identificar as possibilidades de amenizar a evolução

da doença ou melhorar a qualidade de vida dos indivíduos afetados.

2.2 Seleção de características

A seleção de características é um método que deteta e define um subconjunto mais relevante de variáveis de um conjunto maior, sendo usado como forma de minimizar o problema das caraterísticas excessivas e irrelevantes, reduzindo a dimensionalidade dos dados (Pudjihartono et al., 2022).

As variáveis redundantes não acrescentam valor preditivo e podem introduzir ruído, prejudicando a precisão do modelo. Além disso, aumentam a complexidade e tornam o treinamento mais lento. A seleção das características mais relevantes melhora o desempenho do modelo, reduzindo a sua complexidade e aumentando a precisão das previsões (Pathan et al., 2022).

Os métodos de seleção de características são geralmente classificados em três categorias principais que se explicam a seguir:

Os métodos baseados em filtros, que classificam as características ao calcular uma pontuação independente do modelo, selecionando apenas aquelas com as pontuações mais altas ou que ultrapassam um determinado limiar (Bommert et al., 2022). Esses métodos são eficientes do ponto de vista computacional, pois permitem o cálculo paralelo das pontuações, tornando-os ideais para conjuntos de dados de alta dimensionalidade. Eles podem ser divididos em dois grupos principais: os filtros univariados, que avaliam cada característica de forma independente, e os filtros multivariados, que consideram as interações entre diferentes características para a seleção final (Bommert et al., 2022).

Os métodos *wrapper* combinam duas abordagens: uma para escolher as características mais relevantes e outra para as classificar. Este processo ocorre de forma iterativa, onde diferentes subconjuntos de características são testados em um algoritmo de classificação até que um critério de paragem seja alcançado (Maldonado, Riff e Neveu, 2022). A principal vantagem desse método é que ele melhora a precisão do modelo, já que a seleção das características é ajustada especificamente para o classificador utilizado. Porém, essa abordagem pode ser computacionalmente cara, pois exige múltiplas reavaliações do modelo durante o processo (Maldonado, Riff e Neveu, 2022).

Os métodos *Embedded* tentam resolver os desafios entre os métodos Filtro e *Wrapper*, isto porque eles usam critérios estatísticos semelhantes aos métodos de filtros para pré-selecionar algumas características, mas também empregam um algoritmo de ML para escolher o subconjunto de características que proporciona o melhor desempenho na classificação. O processo de seleção ocorre durante o treinamento do modelo, reduzindo a complexidade computacional. Contudo, estes dependem fortemente do algoritmo usado (Effrosynidis e Arampatzis, 2021).

Apesar da seleção de características ser benéfica no contexto de ML esta enfrenta diversos desafios que impactam a sua eficiência e aplicação (Islam et al., 2022). A estabilidade da seleção é uma preocupação, já que pequenas variações nos dados podem gerar subconjuntos diferentes, dificultando a reprodução dos modelos (L. Jiang et al., 2022). Outro fator é conseguir um equilíbrio entre o custo computacional e desempenho, pois métodos como os filtros, apesar de eficientes, podem comprometer a qualidade da seleção, enquanto abordagens mais robustas, como os wrappers, são computacionalmente mais caras (Pudjihartono et al., 2022). Outro desafio é que a avaliação da qualidade das características ainda não possui um critério universalmente eficiente, dificultando a escolha do melhor subconjunto (Yang, Long Liu e Wen, 2024). Por fim, a configuração de hiperparâmetros depende de ajustes manuais que muitas vezes seguem um processo de tentativa e erro (Yang, Long Liu e Wen, 2024).

A escolha da técnica geralmente depende do tipo de desafio a ser resolvido e da quantidade de dados disponíveis. Neste projeto, serão aplicados dois métodos de seleção de características: a Análise de Componentes Principais (PCA), que é um processo que prioriza certas informações e descarta outras, permitindo reduzir a dimensionalidade de um conjunto de dados, transformando um grande número de variáveis em um número menor de variáveis não correlacionadas. (Rupapara et al., 2023), e o Eliminação Recursiva de Características (RFE), que é um método baseado em *wrappers* que elimina recursivamente as características menos importantes com base na pontuação de importância do modelo (Shantal, Alshareef e Ahmid, 2024).

2.2.1 Análise de Componentes Principais

A PCA é um dos métodos de seleção de caraterísticas muito utilizado para reduzir a dimensionalidade dos dados. Este projeta os dados em um espaço de menor dimensionalidade de forma ortogonal, preservando a máxima variância possível (Bajcsi, Andreica e Chira, 2021).

A PCA transforma um conjunto de observações em um novo espaço, onde cada amostra pertence a um espaço Euclidiano de dimensão reduzida. Para isso, calcula a matriz de covariância dos dados e extrai seus valores próprios e vetores próprios, selecionando aqueles com os maiores valores próprios para definir os novos eixos principais. Esse processo permite representar os dados de forma mais compacta para reduzir a perda de informação e facilitar a análise e aprendizado de padrões nos dados (Nguyen et al., 2021).

Com a redução das variáveis, o custo computacional dos algoritmos torna o treinamento mais eficiente e transforma os dados em um novo espaço onde as variáveis são ortogonais entre si, facilitando a interpretação e o desempenho dos modelos (Lenka et al., 2021).

Um dos desafios da PCA é a captura de relações lineares entre os dados, ou seja, se as estruturas relevantes nos dados forem não-lineares, a PCA pode não ser eficaz na redução de dimensionalidade (Ahmad e Nassif, 2022).

Reduzir a dimensionalidade pode afetar a precisão do modelo, por isso escolher o número ideal de componentes é um desafio crítico para evitar perda de informações relevantes. A PCA prioriza a preservação da variância dos dados, mas nem sempre as características com maior variância são as mais discriminativas para uma tarefa específica. (Mousavi et al., 2023)

2.2.2 Eliminação Recursiva de Características

O Método de RFE elimina recursivamente atributos do conjunto de dados e constrói um classificador com os atributos restantes. Este usa a precisão do classificador para identificar quais atributos contribuem significativamente para prever a variável alvo (S. A. Abdulkareem e Z. O. Abdulkareem, 2021).

Alguns dos principais desafios do RFE residem na possibilidade de descartar características importantes para a categorização, pois foca excessivamente no impacto inicial na acurácia da classificação e por vezes variáveis que não demonstram uma influência substancial na fase inicial de treinamento podem ser erradamente excluídas, mesmo que possuam relevância para a classificação num contexto mais abrangente (Priyatno e Widiyaningtyas, 2024).

Além disso, o seu uso exige treinamento e a avaliação de classificadores em cada iteração, sendo que quando o tamanho do conjunto de dados e a contagem de caraterísticas aumentam, a eficiência iterativa da RFE diminui, aumentando consequentemente a complexidade do processo iterativo (C. Chen et al., 2024).

Embora tenha seus desafios, o RFE consegue eliminar variáveis redundantes, contribuindo diretamente para a melhoria da precisão da classificação, concentrando o modelo nas informações mais relevantes (Priyatno e Widiyaningtyas, 2024). O RFE também consegue aumentar a eficiência computacional, diminuindo o tempo de treinamento e os recursos necessários para a execução do modelo, sendo principalmente benéfico em aplicações com limitações de hardware ou em contextos de tempo real (Priyatno e Widiyaningtyas, 2024).

2.3 Machine Learning

A *Machine Learning* tem como base o desenvolvimento de algoritmos e modelos que permitam aos sistemas aprender padrões a partir de dados sem a necessidade de programação explícita (Sharifani e Amini, 2023).

O termo *"machine learning"* popularizou-se na década de 1950, quando o cientista Arthur Samuel definiu-a como a capacidade de máquinas aprenderem sem programação explícita (Samuel, 1959).

Dentro do contexto do uso dos modelos de ML é importante citar alguns dos problemas comuns associados ao uso de modelos que podem ocorrer na fase de treinamento que são o sobreajuste e o subajuste.

O sobreajuste ocorre quando um modelo tem desempenho superior nos dados de treinamento não conseguindo generalizar para dados novos e, dessa forma, não consegue lidar com dados que não foram vistos durante o treinamento. Isso acontece porque o modelo memoriza detalhes específicos, incluindo ruídos presentes no conjunto de treinamento, e não aprende padrões importantes que permitam generalizar as características dos dados (Aburass e Rumman, 2024).

Existem algumas soluções que podem evitar o sobreajuste: uma delas é o *early stopping* que interrompe o treinamento antes que o modelo comece a aprender ruídos dos dados, de forma a garantir um equilíbrio entre aprendizado e generalização (Sabiri, Asri e Rhanoui, 2022). Outra abordagem é a redução da dimensionalidade, onde técnicas são usadas para remover variáveis irrelevantes ou menos significativas do modelo, reduzindo a sua complexidade (Maharana, Mondal e Nemade, 2022). Além disso, a expansão dos dados de treinamento ajuda a melhorar a capacidade de generalização (Maharana, Mondal e Nemade, 2022). Por fim, a regularização é utilizada para limitar a influência de características irrelevantes no modelo, eliminando variáveis menos importantes *L1 Regularization* ou reduzindo o peso das suas variáveis *L2 Regularization* (M. Li, 2023).

O problema de subajuste acontece quando o modelo não consegue captar a complexidade dos dados. Isso geralmente ocorre porque o modelo é muito simples para capturar os padrões presentes no conjunto de dados (S. Zhou et al., 2022).

Para solucionar o subajuste pode-se aumentar a complexidade do modelo ou da quantidade de dados, incrementar os parâmetros do modelo para aumentar a sua complexidade e fazer uma análise e seleção de modelos com base nos dados a serem trabalhados (Prajapati e Singh, 2023).

Na Figura 2.1 é possível analisar a curva de aprendizagem de um modelo com sobreajsute, com o subajuste, e considerado ótimo.

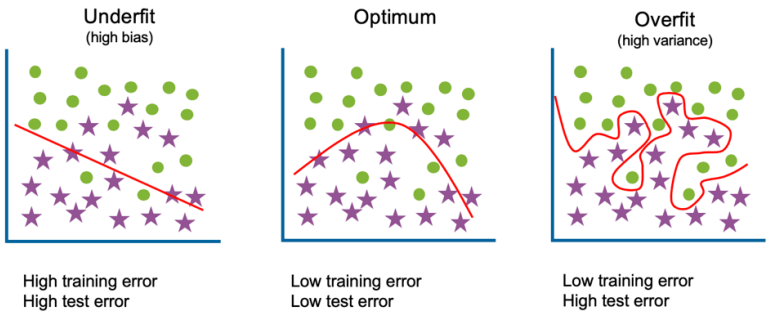

Figura 2.1: Curva de aprendizagem com sobreajuste vs subajuste (Concepts, 2025).

A ML pode ser dividida em três tipos principais: aprendizagem supervisionada (*supervised learning*), aprendizagem não supervisionada (*unsupervised learning*) e aprendizagem por reforço (*reinforcement learning*).

A aprendizagem supervisionada faz previsões com base num conjunto de exemplos, que incluem dados de entrada e de saída (Marsland, 2014).

Na aprendizagem supervisionada, existem duas categorias distintas: regressão e classificação.

Os algoritmos de regressão modelam as dependências e relações entre a variável alvo e as características de entrada para prever novos valores de dados não vistos (Mrabet, Makkaoui e Faize, 2021).

Na classificação, o algoritmo mapeia vetores de entrada para categorias predefinidas, de forma a prever rótulos para os novos dados com base nos padrões aprendidos a partir de exemplos que foram rotulados (T. Jiang, Gradus e Rosellini, 2020).

A aprendizagem supervisionada tem encontrado aplicações significativas em diversos domínios, como cuidados de saúde, finanças e educação. No sector dos cuidados de saúde, os modelos supervisionados têm sido utilizados para tarefas de previsão, incluindo o diagnóstico de doenças a partir de imagens médicas, como as fotografias do fundo da retina para detetar a retinopatia diabética (Ono e Goto, 2022). No domínio empresarial e financeiro, as técnicas de ML apoiam a avaliação do risco de crédito, na previsão do mercado, melhorando a tomada de decisões e a gestão do risco (Gao, Kou, Liang et al., 2024). Além disso, na educação, a ML está a impactar a pesquisa e o ensino, fornecendo conhecimento sobre o desempenho do aluno ou experiências de aprendizagem personalizadas (Ersozlu, Taheri e Koch, 2024).

Alguns dos desafios no uso da aprendizagem supervisionada podem incluir; a necessidade de grande quantidade de dados, problemas de sobreajuste impactando a capacidade de generalização e distribuição desequilibrada dos dados, podendo certas condições serem pouco representadas, tornando o modelo enviesado (Hong et al., 2023).

A aprendizagem não supervisionada identifica padrões em dados não rotulados ou estruturados. Neste caso não existem dados de entrada que se relacionem diretamente com o resultado. O objetivo é o modelo de forma autónoma encontrar padrões nos atributos de entrada para obter um dado de saída (Naeem et al., 2023).

Alguns dos seus modelos comuns incluem: *K-means* que se trata de um método de agrupamento onde os dados são divididos em k agrupamentos (*clusters*) com base nos seus centros de gravidade e o agrupamento hierárquico que agrega os dados de forma hierárquica, podendo ser divisivo ou aglomerativo (Rahmani et al., 2021).

Os desafios do método não supervisionado estão principalmente relacionados à ausência de referências diretas para validar os resultados, tornando difícil medir a

precisão dos modelos. A definição do número ideal de grupos em tarefas de agrupamento pode ser subjetiva e influenciar significativamente os resultados e as dificuldades na interpretação dos resultados sem um conjunto de dados rotulado que dificulta a extração de perceções relevantes, tornando essencial o uso de abordagens estratégicas para garantir a qualidade das inferências (Almuqati et al., 2024).

A aprendizagem por reforço é uma abordagem em que um agente interage com o ambiente que o rodeia mediante um método de tentativa e erro. É normalmente utilizada em situações em que um agente interage com um ambiente, aprendendo a melhorar o seu comportamento com base nas recompensas recebidas (Shakya, Pillai e Chakrabarty, 2023).

Existem diferentes tipos de algoritmos de reforço, alguns dos mais comuns incluem: os baseados em valores como o *Q-Learning* em que se armazena valores de qualidade (Q-values) para cada ação possível em cada estado e *Deep Q-Networks* que supera a limitação do Q-learning em problemas com espaços de estado muito grandes (Chadia e Mousannifa, 2023). Também existem métodos baseados em política que direcionam a otimização diretamente na política do agente, sem estimar diretamente valores de estado como *Proximal Policy Optimization* e *Trust Region Policy Optimization* (Byeon, 2023).

Alguns desafios no uso deste tipo de modelo é sua eficiência amostral, a atribuição de créditos as ações, funções de recompensa mal projetadas que pode levar a comportamentos indesejados e a sua explicabilidade (M. Li, 2023).

Neste trabalho, foram selecionados quatros modelos de ML supervisionada, cada um com características distintas que permitem explorar diferentes aspetos dos dados, aplicados ao contexto da predição de diabetes. São eles:

- **Regressão Logística:** por ser um modelo linear, se destaca pela simplicidade e fácil interpretação, facilitando a análise inicial.
- **Métodos baseados em árvores (Árvores de Decisão, Floresta Aleatória e Gradient Boosting):** eficazes para identificar padrões não lineares nos dados.
- **K-nearest neighbors (KNN):** método simples, que toma decisões com base na proximidade entre os dados, conseguindo capturar relações não lineares.

A combinação desses modelos proporciona uma abordagem mais ampla e comparativa, para a análise realizada neste projeto.

2.3.1 Árvores de Decisão

O modelo organiza os dados numa estrutura hierárquica, onde cada nó interno representa uma decisão baseada num atributo do conjunto de dados. O processo

continua até alcançar os nós folha, que representam os resultados da classificação ou regressão. (Bansal, Goyal e Choudhary, 2022).

O modelo enfrenta alguns desafios que incluem não tratar dados ausentes, dificuldade em lidar com grandes quantidades de atributos tornando a árvore complexa e difícil de interpretar. Dados ruidosos ou valores extremos podem afetar drasticamente a estrutura da árvore e dificuldade na interpretação quando se tornam muito complexas (Mienye e Jere, 2024).

O modelo consegue ser eficiente na gestão de dados mistos, ou seja, combinações de variáveis contínuas e categóricas. Ele possui diversas variações para superar as suas limitações, como métodos como Florestas Aleatórias e *Boosting Trees* que combinam múltiplas árvores para melhorar a precisão e robustez (Zhang, 2021).

No contexto da predição de diabetes, o uso de árvores de decisão pode ser interessante, pois são altamente interpretáveis, eficazes para identificar características relevantes e a sua facilidade na classificação de novos casos (Vakil et al., 2021).

2.3.2 Floresta Aleatória

O algoritmo Floresta Aleatória é usado para classificação e regressão, construindo múltiplas árvores de decisão e combinando os seus resultados para gerar uma previsão final, sendo que cada uma das árvores de decisão é treinada em subconjuntos aleatórios dos dados, reduzindo correlação entre árvores e problemas de sobreajuste (Salman, Kalakech e Steiti, 2024).

Na figura 2.2 é representada a estrutura do modelo, destacando as suas múltiplas árvores de decisão.

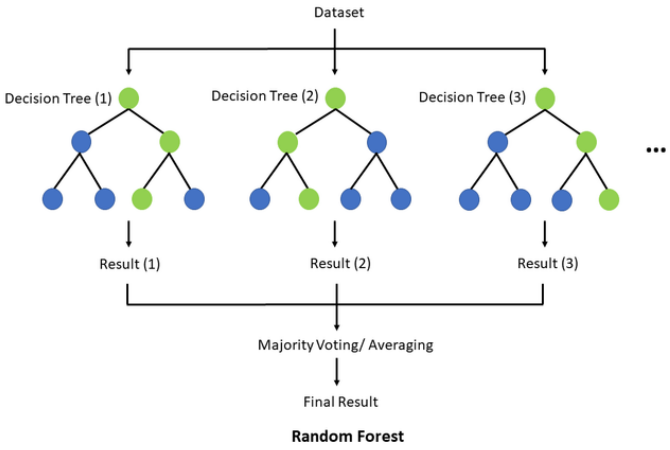

Figura 2.2: Ilustração do funcionamento do algoritmo Floresta Aleatória (Chaos, 2024).

Este modelo se destaca pela robustez contra ruídos, pois ao combinar múltiplas árvores, ele consegue lidar melhor com dados ruidosos e cenários mais complexos e

também por sua eficiência na seleção automática de características mais relevantes para a previsão, facilitando o processamento e melhorando a qualidade dos resultados (A. Wang et al., 2020).

Apesar de ser um modelo identificado como preciso, pode ser difícil de interpretá-lo devido ao grande número de árvores. Com florestas muito grandes aumenta-se a complexidade tornando o modelo computacionalmente caro, especialmente para grandes conjuntos de dados (Dorador, 2024).

O uso da floresta aleatória pode ser benéfico na prevenção da diabetes devido à sua capacidade de capturar relações complexas entre múltiplos fatores de risco (Ooka et al., 2021).

2.3.3 Regressão Logística

A regressão logística é geralmente utilizada em ML para abordar questões de classificação binária, onde a variável resposta é binária e assume valores de 0 e 1. Dessa forma, a variável resposta indica a probabilidade de ocorrência de um evento com base nas variáveis independentes (Silveira et al., 2021).

Ao contrário de modelos mais complexos, a regressão logística oferece algumas vantagens na sua aplicação (Panda et al., 2022): com a sua interpretação intuitiva, já que os coeficientes do modelo podem ser transformados em "razões de probabilidades" (*odd ratios*), permitindo uma análise do impacto de cada variável preditora. Além disso, a regressão logística não requer que as variáveis independentes sigam uma distribuição normal, tornando-se mais flexível para aplicação em diferentes cenários. Outra vantagem é sua robustez relativamente a variáveis categóricas, permitindo a inclusão de variáveis nominais e ordinais sem a necessidade de transformações complexas. O modelo também é relativamente simples de treinar e implementar, sendo compatível com diversas ferramentas estatísticas e linguagens de programação.

Alguns aspetos podem ter impacto no modelo, como correlações altas entre variáveis que podem tornar os coeficientes instáveis e modelos com muitas variáveis para poucos eventos, podendo gerar estimativas não confiáveis (Zabor et al., 2022).

2.3.4 Gradient Boosting

O modelo Gradient Boosting foi introduzido por Friedman, sendo amplamente reconhecido por sua capacidade de transformar preditores fracos num preditor mais robusto (Friedman, 2001).

Este utiliza a técnica de descida do gradiente para construir modelos preditivos de forma sequencial. Cada novo modelo é treinado para corrigir os erros dos modelos anteriores, aprimorando gradualmente a precisão das previsões. Esse processo

iterativo ajusta modelos fracos para minimizar a função de perda, resultando num modelo robusto e altamente eficaz (Miller, Panneerselvam e Lu Liu, 2022).

Além disso, o *Gradient Boosting* tem variantes que reforçam o seu uso. O *XGBoost* é uma variante que busca minimizar a função de perda e otimizar o desempenho do classificador (Gündoğdu, 2023). O *LightGBM* é uma variante alternativa que reduz a utilização de memória e acelera o processo de treinamento, procurando não comprometer o desempenho (Rufo et al., 2021). Por sua vez, o *CatBoost* inclui mecanismos internos que tratam de forma otimizada variáveis categóricas de forma a otimizar a parte de pré-processamento (Kumar et al., 2021).

2.3.5 K-nearest neighbors

O *K-nearest neighbors* (KNN) é um algoritmo usado para classificação e regressão. Este funciona identificando os vizinhos mais próximos de um novo dado, e com base em alguma métrica de distância, faz previsões com base nos valores desses vizinhos (Halder et al., 2024a).

No modelo KNN escolhe-se um valor para k, que representa o número de vizinhos mais próximos a serem considerados e selecionam-se as k instâncias do conjunto de treinamento que possuem as menores distâncias em relação à nova instância (S. Uddin et al., 2022). Com isso, ele busca a semelhança entre os dados e funciona identificando os vizinhos mais próximos de um determinado ponto de entrada para fazer as suas previsões.

O modelo pode ser benéfico devido à sua versatilidade para diferentes conjuntos e à sua facilidade de implementação (Hidayati e Hermawan, 2021).

O KNN pode enfrentar alguns desafios como: a escolha do valor de 'k', pois um valor muito baixo pode resultar em sobreajuste, enquanto um valor muito alto pode levar a subajuste; o custo computacional, pois exige o cálculo de distâncias entre cada ponto de consulta e todos os pontos do conjunto de dados e a dimensionalidade elevada, pois este modelo se deteriora em espaços de alta dimensão (Halder et al., 2024b).

2.4 Deep Learning

O Deep Learning pode ser caracterizado como um subcampo da ML que utiliza redes neuronais artificiais composta por várias camadas ocultas (Jeong, 2020). Antes de 2006, a maioria das pesquisas em ML focavam nos chamados métodos de aprendizado raso que são eficientes para problemas simples, mas não conseguem capturar padrões mais complexos, como reconhecimento de voz e visão computacional (Dong, P. Wang e Abbas, 2021).

Modelos de *Deep Learning* (DL) genericamente requerem muito mais dados do que modelos de ML, pois aprende automaticamente os padrões dos dados e padrões

mais complexos sem necessidade de intervenção humana direta, além de que as redes neuronais profundas podem ajustar automaticamente pesos e parâmetros. Isso melhora a capacidade de generalização e torna os modelos mais adaptáveis (Taye, 2023).

Apesar das suas vantagens, o DL apresenta desafios. O treinamento requer um grande volume de dados rotulados e demanda alto custo computacional. O processamento exige hardware especializado, como Unidades de Processamento Gráfico (GPUs) e Unidades de Processamento de Tensor (TPUs), o que pode dificultar a sua adoção em sistemas de baixo custo (Macas, Wu e Fuertes, 2022).

Outro problema é a falta de explicabilidade dos modelos. Muitos algoritmos são considerados caixas-pretas, o que dificulta a interpretação dos resultados. Além disso, os modelos são vulneráveis a ataques adversários. Pequenas alterações nos dados de entrada podem levar a previsões incorretas.

Por fim, a catástrofe do esquecimento é outro desafio. Quando os modelos são treinados com novos dados, podem perder informações aprendidas anteriormente. Esse problema compromete a retenção do conhecimento ao longo do tempo (Talaei Khoei, Ould Slimane e Kaabouch, 2023).

Neste projeto serão trabalhados dois modelos diferentes que são comummente usados em estudos de predição de diabetes e tem diferentes abordagens o que pode trazer diferentes perspetivas a este projeto. O Perceptron multicamadas é uma escolha eficiente se os dados forem estruturados e independentes, como o caso do conjunto de dados trabalhado e o modelo Rede Neuronal Recorrente que modela dependências temporais e pode ser útil para dados sequenciais.

2.4.1 Perceptron multicamadas

O Perceptron Multicamadas (MLP) é descrito como uma rede neuronal do tipo de "alimentação direta", que utiliza o algoritmo de retropropagação para o aprendizado, possuindo uma camada de entrada para receber os dados, uma ou mais camadas ocultas para processamento e uma camada de saída para a previsão (Desai e Shah, 2021).

Cada neurónio nas camadas ocultas aplica uma função de ativação para capturar padrões não lineares nos dados. O MLP utiliza o algoritmo de retropropagação para ajustar os pesos das conexões, minimizando o erro entre a saída prevista e a desejada (Naskath, Sivakamasundari e Begum, 2023).

O MLP enfrenta alguns desafios que pode afetar o seu desempenho (Bachmann, Anagnostidis e Hofmann, 2023). Um dos principais problemas é o sobreajuste, especialmente em conjuntos de dados pequenos. Além disso, o MLP não possui um viés indutivo forte, ou seja, ele não aproveita bem certas estruturas dos dados, o que o pode tornar menos eficiente para utilização em visão computacional. Outro desafio no treinamento de redes profundas envolve a atenuação do gradiente de ajuste, em

que as informações para ajustar os parâmetros tornam-se muito subtis nas camadas mais distantes, dificultando a atualização dos pesos e prejudicando o aprendizado – problema conhecido como *vanishing gradient*. Com isso, esse modelo pode não ser o mais indicado para cenários com menos recursos, pois seu desempenho tem a tendência a melhorar com um maior volume de dados e de poder computacional.

O MLP se sobressai em diversos campos, como a classificação de imagens de sensoriamento remoto, a deteção de invasões em redes, a identificação de ataques cibernéticos, a área médica, a análise de tráfego de veículos e a condução autónoma de robôs (Naskath, Sivakamasundari e Begum, 2023).

2.4.2 Redes neuronais recorrentes

As Redes Neuronais Recorrentes (RNN) são como um tipo de rede neuronal que processa dados sequenciais, utilizando um estado oculto que armazena informações sobre entradas anteriores para influenciar as saídas futuras. As RNN têm um sistema de equações que inclui pesos e funções de ativação que controlam como as informações de entradas e estados passados contribuem para o estado atual (Sherstinsky, 2020).

As RNN mantêm informações de entradas passadas por um longo período, o que a torna adequada para dados temporais e contextuais. Isso é útil em aplicações que dependem do contexto e da ordem dos dados, como no processamento de linguagem natural e na análise de séries temporais (Sherstinsky, 2020).

Na Figura 2.3 é possível ver a estrutura geral do RNN. A Camada de Entrada representada pelos neurónios à esquerda, que recebem os dados de entrada e os enviam para a primeira camada oculta. Nas Camadas Ocultas, que são pelo menos duas, os neurónios têm conexões recorrentes que retornam a eles próprios, permitindo o aprendizado de sequências e dependências temporais. Por fim, a Camada de Saída à direita, recebe os sinais da última camada oculta e gera a saída final da rede.

O modelo RNN, principalmente nas versões *Long Short-Term Memory* (LSTM) e *Gated Recurrent Unit* (GRU), tem ganho destaque em tarefas que envolvem dados espaço-temporais, devido à sua habilidade de compreender relações complexas no tempo e no espaço. Essas redes conseguem identificar padrões temporais contínuos e periódicos, além de capturar correlações espaciais que variam com o tempo (Fang, Y. Chen e Xue, 2021).

O RNN tem desafios no treinamento e modelagem de sequências longas, como problema de atenuação de gradientes, onde os pesos das camadas iniciais mudem muito pouco durante o treinamento, dificultando a aprendizagem de dependências de longo prazo. Além disso, pode ocorrer o problema de gradientes explosivos, que ocorre quando os valores dos gradientes aumentam exponencialmente, tornando as atualizações dos valores dos pesos muito grandes e fazendo com que a rede fique

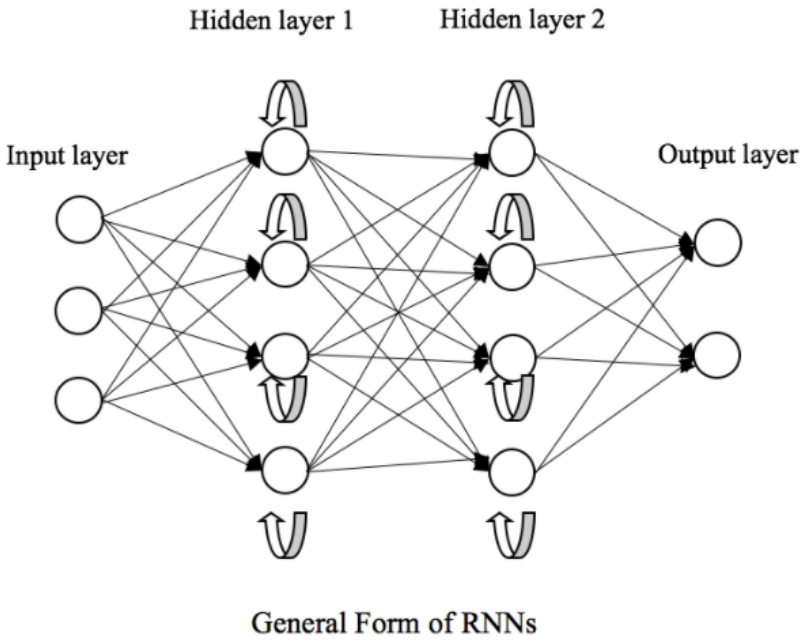

Figura 2.3: Representação simplificada de Rede Neuronal Recorrente (Science, 2024).

instável. Estas situações dificultam a capacidade da RNN de capturar relações distantes, pois a informação se perde ao longo da sequência (Ahmed et al., 2023).

2.5 Machine learning e Deep Learning aplicada à prevenção de diabetes

O uso da ML na prevenção de diabetes tem sido aplicado em diversos estudos, especialmente em alguns modelos supervisionados como o de Regressão logística, Árvores de decisão, Floresta Aleatória e *gradient boosting* e também modelos de DL que têm-se mostrado eficazes em tarefas como predição de risco, diagnóstico precoce e personalização de estratégias preventivas. Seguidamente destacam-se alguns destes estudos relevantes na análise e prevenção da diabetes, dando ênfase aos modelos já referidos.

As técnicas da ML têm sido utilizadas na predição e tratamento da diabetes em revisões sistemáticas como de (Tigga e Garg, 2020) que foi efetuada para prever a probabilidade de diabetes tipo 2 utilizando uma variedade de algoritmos de ML. O estudo utilizou 952 entrevistas realizadas por meio de formulários preenchidos por voluntários, incluindo 18 perguntas relacionadas com a saúde, o estilo de vida e os antecedentes familiares. O desempenho do classificador Floresta Aleatória foi considerado o mais exato.

Em outro estudo de (Jian et al., 2021) foram aplicados vários algoritmos de classificação supervisionada para construir diferentes modelos de previsão e classificação de oito complicações da diabetes, sendo elas, a síndrome metabólica, a neuropatia, a nefropatia, a hipertensão, a obesidade, a retinopatia, entre outros. Foram utilizados um conjunto de dados recolhidos pelo *Rashid Center for Diabetes and Research* (RCDR) composto por 884 registos com 79 características. Foi efetuada uma seleção de características para selecionar as cinco e dez principais características para cada complicação. Chegou-se a conclusão no fim deste estudo que as características mais dominantes que afetam as previsões e consequentemente mais úteis para prever as complicações, são o colesterol total, idade da diabetes, género, *Índice de massa corporal* (IMC) e pressão arterial.

Na análise de (Daghistani e Alshammari, 2020) foi feita a comparação dos algoritmos Floresta Aleatória e Regressão Logística na predição de diabetes usando 66.325 registos médicos da Arábia Saudita. Dados demográficos, índices como IMC e pressão arterial, e resultados laboratoriais foram usados como variáveis. Como resultado o Floresta Aleatória superou a Regressão Logística em todas as métricas avaliadas, mostrando ser mais adequada para lidar com múltiplos fatores de risco e dados complexos.

No artigo (Olisah, L. Smith e M. Smith, 2022) é apresentada uma abordagem para prever e diagnosticar diabetes explorando dois conjuntos de dados, um com informações de pacientes indígenas Pima e outro de um hospital no Iraque, para treinar modelos capazes de identificar a doença. Para melhorar os resultados, eles aplicaram métodos de seleção de características e preencheram dados ausentes usando regressão polinomial, garantindo que os modelos trabalhassem com informações mais relevantes e confiáveis. Foram testados diferentes algoritmos, incluindo Floresta Aleatória e Máquina de Vetores de Suporte, mas o destaque foi um modelo de rede neuronal profunda que superou os demais em termos de precisão. A pesquisa demonstrou que com a preparação dos dados e otimização dos modelos, é possível alcançar diagnósticos mais precisos.

O estudo conduzido por (Alzyoud et al., 2024) investigou como diferentes técnicas de ML podem ser aplicadas no diagnóstico precoce da diabetes. Os pesquisadores testaram quatro métodos de seleção de características e doze classificadores, representando seis estratégias distintas de aprendizado. Os resultados indicaram que o método *CorrelationAttributeEval* foi o mais eficiente para escolher os atributos mais relevantes. Já na etapa de classificação, o *MultiClassClassifier* apresentou o melhor desempenho, especialmente em conjuntos de dados compostos por variáveis contínuas.

Já a pesquisa de (Wee et al., 2024) analisou o uso de ML e DL na deteção da diabetes. Um dos pontos discutidos foi a dependência de exames laboratoriais invasivos em muitos conjuntos de dados existentes. Para contornar essa limitação, os autores

sugeriram que características antropométricas e informações não invasivas poderiam ser mais acessíveis para a criação de modelos preditivos. O estudo também avaliou diferentes estratégias de pré-processamento, além de comparar algoritmos como Floresta Aleatória, Máquina de Vetores de Suporte (SVM), Rede Neural Convolucional (CNN) e DNN. Os modelos baseados em DL alcançaram uma acurácia de até 98%. O estudo conclui que, embora os modelos de DL sejam mais precisos, eles exigem grandes quantidades de dados e poder computacional, enquanto os modelos de ML são mais interpretáveis e acessíveis.

O artigo de (Khan et al., 2024) apresenta um modelo para prever diabetes usando ML e DL. Os pesquisadores analisaram dois conjuntos de dados amplamente usados na área médica (PIMA Indian Diabetes e Early-Stage Diabetes Risk) e aplicaram técnicas avançadas de seleção de características para encontrar os fatores mais relevantes na previsão da doença. Eles testaram diferentes algoritmos, incluindo redes neuronais artificiais, Floresta Aleatória, *Gradient Boosting* e Máquina de Vetores de Suporte. Os resultados mostraram que a rede neuronal teve a melhor performance no conjunto de dados PIMA, com 99,35% de precisão, enquanto a Floresta Aleatória se destacou no conjunto de dados de risco inicial de diabetes, com 99,36% de precisão.

Numa investigação realizada por (Srinivasu et al., 2022) foi analisado como o modelo RNN e alguma das suas variantes pode ser utilizada para prever diabetes tipo 2 a partir de dados genómicos e tabulares. A partir de um modelo baseado em RNN foram analisados padrões genéticos associados ao diabetes testando sua eficácia usando dois conjuntos de dados, um extraído de sequências de ADN permitindo identificar padrões genéticos e outro que contém variáveis clínicas como glicose no sangue, pressão arterial e IMC. Os respetivos resultados mostraram bom desempenho ao usar a variante LSTM, principalmente quando combinando dados genómicos e tabulares.

O estudo de (El-Bashbishy e El-Bakry, 2024) propôs um método para previsão precoce do diabetes pediátrico utilizando um novo conjunto de dados clínicos, contendo 548 amostras de pacientes com idades entre 1 e 19 anos, com 18 atributos clínicos relevantes. Foi usado um modelo baseado em MLP com 10 camadas ocultas e utilizando diferentes funções de ativação, também incluindo técnicas de normalização de dados e balanceamento de classes para garantir um modelo mais confiável. O modelo conseguiu alcançar um índice de acurácia de 99,8%, superando outras abordagens destacadas no estado da arte.

Capítulo 3

Desenvolvimento

Neste capítulo, será descrito o processo de desenvolvimento do projeto. O primeiro passo é a definição da metodologia aplicada. Na secção seguinte, o conjunto de dados é apresentado. Em seguida, a análise a este conjunto de dados para entender melhor as informações disponíveis. Posteriormente, na preparação dos dados, serão explicadas as transformações aplicadas para tornar esse conjunto mais adequado para modelagem. A secção de Modelação apresentará o processo de construção dos modelos. Por fim, em Avaliação dos Resultados, será analisado o desempenho final dos modelos. Essa análise será baseada em métricas de avaliação.

3.1 Metodologia

Para a criação e aplicação dos modelos de ML que permitirão realizar análise preditiva de diabetes, foi utilizada a metodologia *Cross-Industry Standard Process for Data Mining* (CRISP-DM). Essa metodologia surgiu no final de 1996. Seu desenvolvimento contou com a participação de líderes da indústria e fornecedores de ferramentas e serviços de mineração de dados. O objetivo foi estabelecer um modelo que incentiva as melhores práticas com uma estrutura que busca garantir resultados eficazes na mineração de dados, conforme será explicado nesta secção (Shearer, 2000).

A metodologia é adaptativa e a sua abordagem permite a interação dinâmica entre as fases em que o processo possibilita o retorno às etapas anteriores para ajustes. Essas revisões melhoram a abordagem com base nos resultados obtidos. O

método também garante alinhamento com os objetivos do negócio (Purbasari et al., 2021).

A metodologia CRISP-DM possui seis fases, conforme ilustrado na Figura 3.1. A primeira é a compreensão do negócio. Em seguida, ocorre a compreensão dos dados disponíveis. A terceira fase trata da preparação dos dados. A modelação ocorre na quarta etapa. A avaliação vem na quinta fase. Por fim, a última etapa corresponde à implementação de uma solução em ambiente de produção (Schröer, Kruse e Gómez, 2021). No desenvolvimento deste projeto, a fase de implementação não foi realizada, uma vez que o foco do estudo era a análise e avaliação de diferentes modelos.

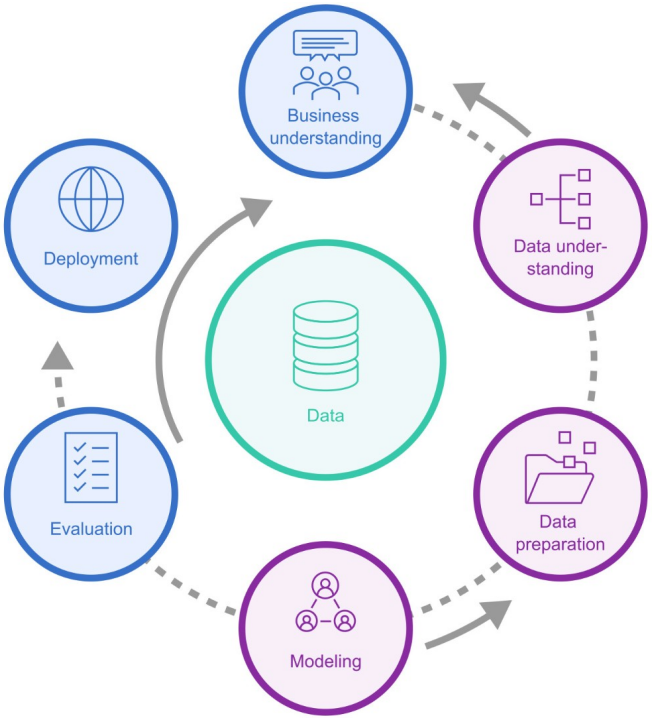

Figura 3.1: Cross-Industry Standard Process for Data Mining.[1]

Alguns dos principais desafios identificados (Schröer, Kruse e Gómez, 2021) na aplicação da CRISP-DM estão relacionados a inconsistências na fase de implantação, limitando a integração dos modelos em ambientes produtivos. Além disso, o método pode ter limitações em cobrir o ciclo de vida completo de projetos de ML. Por fim, há a necessidade de atualizar a metodologia para incorporar tecnologias modernas e automação no processo.

3.1.1 Compreensão do negócio

Na compreensão do negócio são definidos os objetivos e necessidades do projeto. Este processo inicia-se com um estudo sobre a área de negócio onde o modelo vai

[1] https://www.ist.fraunhofer.de/en/expertise/simulation-digital-services/data-acquisition-model-based-process-optimization/crisp-dm-surface-technology.html

atuar – neste caso, na área da saúde, e para impactar e suportar esta área na prevenção da diabetes.

A diabetes é uma doença de impacto mundial, e reduz a qualidade de vida da população, sendo que a identificação de indivíduos com alto risco de desenvolvimento da doença ou em estado inicial pode facilitar a implementação de medidas preventivas para garantir a não evolução dessa doença.

Assim sendo, o desenvolvimento deste projeto visa identificar características com maior correlação ao risco dessa doença, de forma que os profissionais da saúde possam atuar de forma mais focada em pessoas deste perfil.

3.1.2 Compreensão dos dados

Na etapa de compreensão dos dados, os possíveis conjuntos de dados existentes a serem usados no projeto foram analisados, para entender os dados disponíveis e se estes cumpriam as condições necessárias para aplicação neste trabalho. Para cada dataset analisado, as seguintes características foram elencadas: quais os descritores disponíveis, qual o tamanho do dataset, e quais os tipos de variáveis. Além disso, foi analisado se existiam valores discrepantes, valores ausentes ou desequilíbrios entre classes.

3.1.3 Preparação dos dados

A fase de preparação de dados dividiu-se em diversos passos, tendo iniciado com a limpeza de dados para remover inconsistências, com tratamento de valores ausentes, duplicados e discrepantes mediante a necessidade após análise. Em seguida, realiza-se a transformação dos dados para que estes fiquem no formato ideal para a modelação, utilizando técnicas de normalização e padronização. Também analisou-se a necessidade de seleção, criação ou conversão de atributos, contribuindo para a precisão do modelo – para este efeito, cogitou-se a criação de novas variáveis a partir das existentes no conjunto de dados. Por fim, a seleção de atributos foi conduzida para identificar as variáveis mais relevantes para a criação do modelo.

3.1.4 Modelação

Nesta etapa são aplicados diferentes modelos de ML e DL para predição de diabetes, tendo em conta os diversos dados de caráter comportamental, saúde, financeiro – entre outros – para uma população classificada como com e sem diabetes. Esta etapa decorre em conjunto com a preparação dos dados, pois a modelação contempla a escolha dos algoritmos mais adequados ao problema em questão. Mediante a escolha dos algoritmos e tendo definido os parâmetros relevantes, o próximo passo é estruturar o modelo aplicando tais algoritmos e também métodos de seleção de

características. Com a finalização da estruturação dos modelos, é conduzido o treinamento dos mesmos. Por fim, testamos o modelo com um conjunto de dados inédito para avaliar a sua precisão e capacidade de generalização.

3.1.5 Avaliação

Com a finalização da modelagem, segue-se a avaliação do desempenho do modelo, sendo necessário testar e analisar os resultados para perceber se os objetivos do projeto foram alcançados ou se é necessário revisar o processo. A avaliação do modelo acontece utilizando diferentes métricas que serão apresentadas mais à frente nesta secção. Caso algum objetivo não tenha sido alcançado, o modelo permite sempre voltar ao início deste processo, de forma a aplicar as melhorias para alcançar os objetivos pretendidos.

A avaliação dos modelos foi realizada por meio de métricas padrão (C.-Y. Chou, Hsu e C.-H. Chou, 2023) e também utilizando uma matriz de confusão, que é uma tabela que resume o desempenho de um modelo de classificação, apresentando o total de previsões corretas e incorretas (Sahid, Babar e M. P. Uddin, 2024).

A matriz de confusão contém quatro possíveis resultados:

- **Verdadeiros Positivos (TP)**: pacientes com diabetes que foram corretamente previstos como tendo diabetes.
- **Falsos Positivos (FP)**: pacientes sem diabetes que foram incorretamente previstos como tendo diabetes.
- **Falsos Negativos (FN)**: pacientes com diabetes que foram incorretamente previstos como não tendo diabetes.
- **Verdadeiros Negativos (TN)**: pacientes sem diabetes que foram corretamente previstos como não tendo diabetes.

As principais métricas de classificação derivadas da matriz de confusão utilizadas neste projeto são apresentadas a seguir.

A *Acurácia* refere-se à proporção entre o número de previsões corretas e o tamanho total da amostra (Ali et al., 2024). Esta é calculada utilizando a seguinte equação:

$$\text{acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

Numa situação de desequilíbrio entre classes, a acurácia pode apenas refletir a tendência do modelo para prever a classe maioritária (Sahid, Babar e M. P. Uddin, 2024). Por exemplo, o número de amostras de não diabéticos é muito superior ao número de amostras de diabéticos, e assim é fundamental considerar métricas adicionais para avaliar adequadamente o desempenho do modelo.

A *Precisão* é uma medida significativa para determinar a exatidão. Indica qual a percentagem de previsões positivas que estão efetivamente corretas (Rainio, Kukkonen e Pölönen, 2024). No contexto da previsão de diabetes, indica a frequência de quando o modelo está correto ao prever a diabetes.

Abaixo é representada a equação que define a precisão:

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{3.2}$$

O *Recall* indica a proporção de casos positivos reais que o modelo identifica corretamente (Rainio, Kukkonen e Pölönen, 2024). Quanto maior o *Recall* mais o modelo é capaz de detetar os casos de diabetes e minimizar a quantidade de falsos positivos.

É calculado pela divisão dos verdadeiros positivos por todos os positivos, como mostra a equação:

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

A precisão e o *recall* podem ser combinados em uma única métrica, conhecida como *F1-score*, que representa a média dessas duas medidas (Ali et al., 2024). O *F1-score* varia entre 0 e 1, sendo 1 o valor ideal, indicando um equilíbrio perfeito entre precisão e *recall*. Um *F1-score* elevado significa que o modelo tem um bom equilíbrio entre a precisão e *recall* indicado maior confiabilidade na deteção.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.4}$$

Como o foco do trabalho é um modelo preditivo para diagnósticos de diabetes, é preferível adotar uma postura mais cautelosa, onde um maior número de falsos positivos é aceitável em detrimento de falsos negativos. Isso porque suspeitar de uma condição, como diabetes, e tratá-la com a devida atenção pode evitar complicações graves decorrentes do não diagnóstico, mesmo que, em alguns casos, o paciente não seja realmente portador da doença. Essa abordagem assegura que casos verdadeiros não tendam a ser negligenciados, priorizando a segurança e a eficácia do acompanhamento clínico. Em outras palavras, é preferível um modelo com alto recall, mesmo que isso implique em mais falsos positivos, em vez de priorizar alta precisão.

3.2 Conjunto de dados

Para o desenvolvimento deste projeto foi utilizado o conjunto de dados para a previsão da diabetes, *"Diabetes Health Indicators Dataset"*, disponibilizado na plataforma Kaggle (Teboul, 2015). Esse conjunto de dados foi elaborado a partir da seleção de informações sobre saúde e comportamentos associados à diabetes,

extraídas do *Behavioral Risk Factor Surveillance System* (BRFSS), um sistema de vigilância epidemiológica dos Estados Unidos, com dados coletados em 2015.

O BRFSS é um inquérito telefónico sobre saúde realizado anualmente. Esse levantamento é conduzido pelo *Centers for Disease Control and Prevention (CDC)*. O inquérito reúne respostas de mais de 400.000 americanos. As perguntas abordam comportamentos de risco, doenças crónicas e uso de serviços preventivos.

Inicialmente, o conjunto de dados coletados pelo BRFSS possuía 330 variáveis, mas este foi reduzido numa seleção de 22 atributos pelo autor do repositório (Teboul, 2015), sendo essas as informações mais correlacionadas à diabetes. Este conjunto, após remoção de valores extremos e registos incompletos, representa 253.680 pessoas.

A escolha dessas variáveis no conjunto de dados Kaggle foram definidas com base no estudo de (Xie et al., 2019). Nesse estudo, os autores construíram modelos para predição de diabetes tipo 2 utilizando dados do BRFSS 2014 e revisaram a literatura para identificar fatores de risco relevantes da diabetes para selecionarem as variáveis usadas.

Grande parte das variáveis destacadas no estudo de (Xie et al., 2019) — e também presentes no conjunto de dados usado neste trabalho — têm suporte em estudos anteriores compilados pela literatura. Em particular, uma revisão sistemática de (Collins et al., 2011) que examinou 39 estudos nos quais foram desenvolvidos modelos de predição de diabetes tipo 2. Os autores dessa revisão levantaram os fatores de risco mais utilizados nos modelos ao longo dos estudos. Esse levantamento mostrou um consenso em torno de fatores como: história familiar de diabetes, IMC, tabagismo e sedentarismo.

Com base nessas evidências da literatura, o conjunto de dados utilizado neste trabalho foi elaborado utilizando a seleção dos 22 atributos elencados por (Teboul, 2015). A seguir são detalhados esses atributos, organizados em grupos para melhor compreensão da sua relevância para o modelo preditivo desenvolvido.

A Variável dependente (resposta da classificação) é a **Diabetes_binary**, uma variável binária que indica se o entrevistado já foi diagnosticado com diabetes (0 = Não, 1 = Sim).

Já as variáveis independentes são divididas em cinco grupos: Condições de Saúde Crónicas, Fatores de Estilo de Vida, Saúde Geral e Acesso a Cuidados Médicos, Saúde Física e Mental e Fatores Demográficos.

As Condições de Saúde Crónicas incluem:

- **HighBP** - Variável Categórica Binária (0 = Não, 1 = Sim) que indica se o entrevistado já recebeu diagnóstico de hipertensão arterial.
- **HighChol** - Variável Categórica Binária (0 = Não, 1 = Sim) que indica se o entrevistado já recebeu diagnóstico de colesterol alto.

- **CholCheck** - Variável Categórica Binária (0 = Não, 1 = Sim) que indica se realizou exame de colesterol nos últimos 5 anos (0 = Não, 1 = Sim).
- **Stroke** - Variável Categórica Binária (0 = Não, 1 = Sim) se já teve um acidente vascular cerebral.
- **HeartDiseaseorAttack** - Variável Categórica Binária (0 = Não, 1 = Sim) se já teve doença coronária ou infarto do miocárdio.

Já os fatores de Estilo de Vida são:

- **BMI** - Índice de Massa Corporal (Variável Numérica Contínua).
- **PhysActivity** -Variável Categórica Binária (0 = Não, 1 = Sim) se praticou atividade física ou exercício nos últimos 30 dias.
- **Fruits** - Variável Categórica Binária (0 = Não, 1 = Sim) se consome frutas pelo menos uma vez ao dia.
- **Veggies** - Variável Categórica Binária (0 = Não, 1 = Sim) se consome vegetais pelo menos uma vez ao dia.
- **HvyAlcoholConsump** - Variável Categórica Binária (0 = Não, 1 = Sim) para consumo excessivo de álcool (homens $>$ 14 doses/semana, mulheres $>$ 7 doses/semana).
- **Smoker** - Variável Categórica Binária (0 = Não, 1 = Sim) se já fumou pelo menos 100 cigarros ao longo da vida.

As variáveis relacionadas à Saúde Geral e Acesso a Cuidados Médicos são:

- **AnyHealthcare** - Variável Categórica Binária (0 = Não, 1 = Sim) se possui algum tipo de plano de saúde.
- **NoDocbcCost** - Variável Categórica Binária (0 = Não, 1 = Sim) se precisou de um médico nos últimos 12 meses, mas não pôde consultar devido ao custo.

Enquanto as variáveis relacionadas à Saúde Física e Mental são:

- **GenHlth** - Autoavaliação da saúde geral (Escala de 1 a 5, de “Excelente” a “Ruim”).
- **MentHlth** - Número de dias, nos últimos 30 dias, em que a saúde mental esteve ruim (0-30).
- **PhysHlth** - Número de dias, nos últimos 30 dias, em que a saúde física esteve ruim (0-30).

- **DiffWalk** - Variável Categórica Binária (0 = Não, 1 = Sim) se tem dificuldades para caminhar ou subir escadas.

As variáveis Demográficas são:

- **Sex** - Sexo do entrevistado (0 = Feminino, 1 = Masculino).
- **Age** - Faixa etária do entrevistado (14 categorias de idade).
- **Education** - Nível educacional mais alto atingido (Subdividida em 6 categorias).
- **Income** - Faixa de rendimento anual total do domicílio (Subdividida em 8 categorias).

3.3 Análise dos dados

Os dados foram analisados e preparados utilizando a linguagem Python. A biblioteca Pandas[2] foi usada para leitura, organização e manipulação das informações, permitindo a criação de um *dataframe* com as variáveis necessárias.

Os dados foram separados em classes, para facilitar a análise, sendo pessoas não diabéticas `Diabetes_binary == 0` e pessoas diabéticas `Diabetes_binary == 1`.

Para dar início à análise exploratória dos dados, foi realizada a remoção de linhas duplicadas que foram identificadas e removidas para garantir maior fiabilidade na análise, conforme excerto de código fonte 1.

```
print("\nQuantidade de linhas duplicadas antes da remoção:",
↪  df.duplicated().sum())
df.drop_duplicates(inplace=True)
print("Quantidade de linhas duplicadas após a remoção:",
↪  df.duplicated().sum())
```

Código fonte 1: Remoção de linhas duplicadas do *dataframe*.

Ao analisar a distribuição por género entre as classes, representada na figura 3.2, foi possível observar que, no caso de pessoas diabéticas, não há grande desequilíbrio entre a quantidade de amostras do género feminino ou masculino. Sendo que entre diabéticos, 52,09% são do género feminino e 47,91% do género masculino.

Também foi analisada a distribuição de pessoas diabéticas por idade, sendo removidos valores abaixo de 18 anos por não haver dados suficientes para análise nessa faixa etária – conforme evidenciado na Figura 3.3. Foi também possível perceber que neste conjunto de dados a diabetes tem maior predominância entre pessoas acima dos 45 anos com 93,75% da amostra. Existem poucos casos de diabetes em pacientes abaixo dos 40 anos, somando apenas 6,25%.

[2]https://pandas.pydata.org/

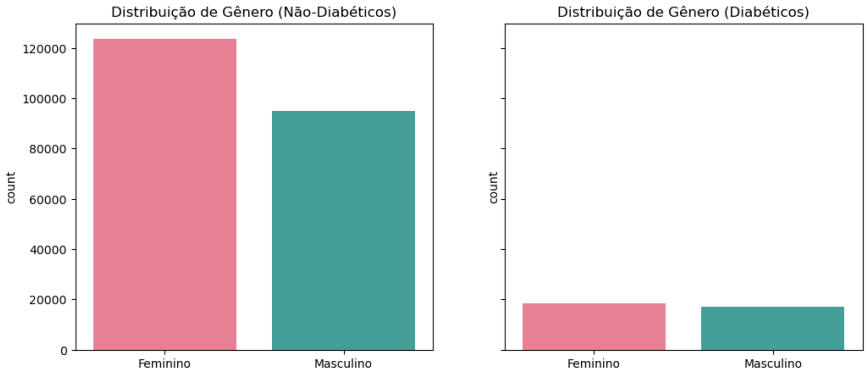

Figura 3.2: Distribuição por género.

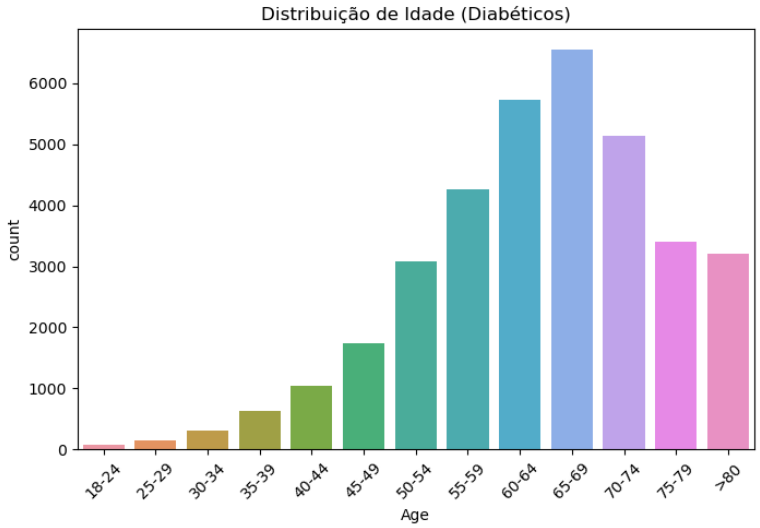

Figura 3.3: Distribuição por idade.

Em análise comparativa do IMC entre diabéticos e não diabéticos representada na Figura 3.4, um *boxplot* foi criado para comparar a distribuição. O *boxplot* apresenta a mediana (linha central da caixa), os quartis (limites da caixa), os valores extremos (whiskers) e outliers (pontos isolados) para cada grupo. Foi possível verificar na amostra que a mediana do IMC é visivelmente maior em pacientes diabéticos (aproximadamente 31) do que em não diabéticos (aproximadamente 27). Além disso, o grupo diabético apresenta maior variabilidade nos valores de IMC, com uma distribuição mais dispersa e maior concentração de valores acima de 30, indicando uma associação entre IMC elevado e diabetes.

Para variáveis categóricas como colesterol alto, hipertensão, tabagismo, consumo de álcool, atividade física e dificuldade para andar, foi criado um gráfico de contagem (*countplot*) para avaliar a distribuição dos dados deste conjunto para tais variáveis, conforme Figura A.1.

O colesterol alto foi identificado em mais pessoas diabéticas do que não diabéticos. A pressão alta também aparece mais entre diabéticos. A frequência de

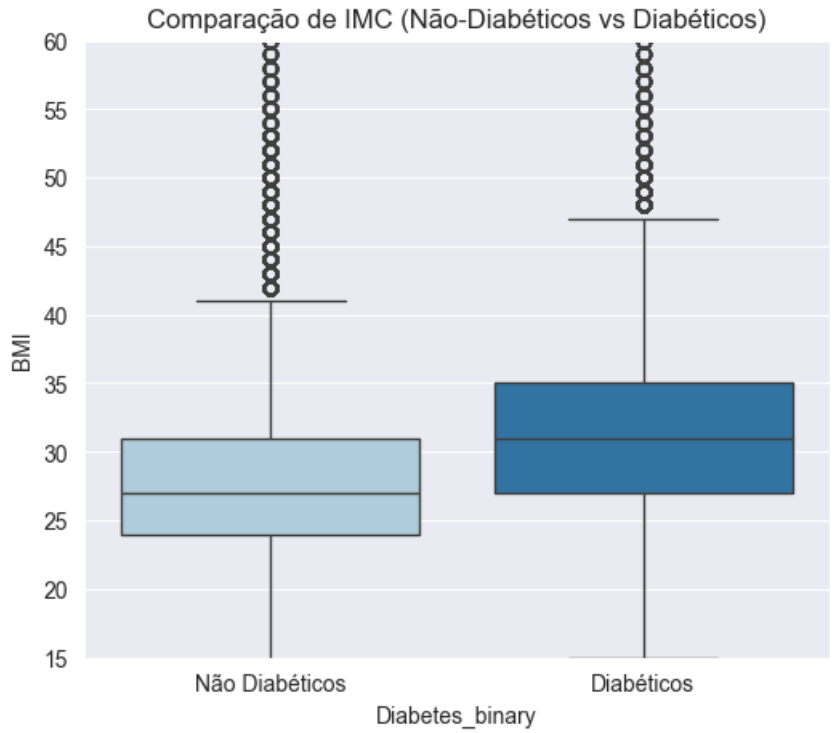

Figura 3.4: Comparação de IMC entre Diabéticos e Não Diabéticos.

tabagismo não apresentou uma diferença muito grande entre os grupos. O consumo excessivo de álcool também é baixo em ambos os grupos. A frequência de atividade física é muito superior em pessoas não diabéticas. No grupo diabético, há menos pacientes fisicamente ativos, também sendo mais frequente o relato de dificuldade para caminhar.

Uma análise da distribuição de valores representada na Figura 3.5 mostra que o conjunto de dados tem um desequilíbrio significativo: apenas 13,9% dos registos correspondem a pessoas diabéticas. Para corrigir esse problema, na secção a seguir será discutida a técnica usada nesse trabalho para garantir equilíbrio nos dados para melhorar a sua representatividade das classes.

A correlação entre atributos também foi analisada. O mapa de correlação, representado na Figura A.1, revelou padrões importantes: Saúde Geral (GenHlth), Saúde Física (PhysHlth) e Dificuldade para Caminhar (DiffWalk) apresentaram forte correlação positiva, mostrando que pode haver uma associação direta entre estas variáveis.

O rendimento (Income) teve correlação negativa com Saúde Geral e Dificuldade para Caminhar, o que pode indicar que pacientes com maiores rendimentos tendem a ter melhor saúde geral.

A presença de diabetes mostrou correlação com fatores como pressão alta, IMC, dificuldade de locomoção e doenças cardíacas.

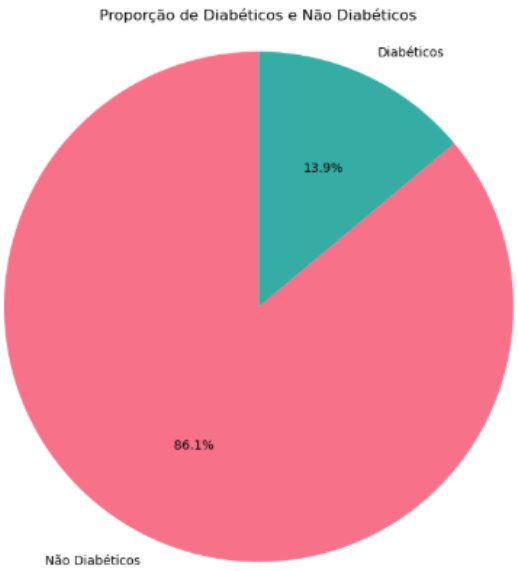

Figura 3.5: Proporção de Diabéticos.

3.4 Preparação dos dados

Para os seguintes passos foi feita a separação da variável alvo (*target*) e das restantes como y e X, respetivamente, conforme excerto de código fonte 2.

```
# Definir X (variáveis) e y (target)
X = df.drop('Diabetes_binary', axis=1)
y = df['Diabetes_binary']
```

Código fonte 2: Separação das variáveis a partir do *dataframe*.

Com já discutido na secção anterior, um dos desafios deste conjunto de dados é o desequilíbrio das classes, no qual pessoas com diabetes são menos representadas.

Para resolução deste problema, foi aplicada a técnica de subamostragem (*undersampling*) *NearMiss* para balancear as classes, conforme excerto de código fonte 3.

Esta técnica seleciona amostras da classe maioritária que estão mais próximas das amostras da classe minoritária para preservar a estrutura do problema. Esta definição de proximidade, em vez de apenas rotular os dados como "positivo" ou "negativo", usa um valor contínuo que indica o grau de proximidade com a classe positiva.

A versão 1 do algoritmo NearMiss, utilizada no desenvolvimento, realiza o balanceamento das classes por meio dos seguintes passos (Tanimoto et al., 2022):

1. Calcula a distância entre cada instância da classe maioritária e as instâncias da classe minoritária.

```
# Aplicar undersampling com NearMiss para balancear classes
nm = NearMiss(version=1, n_neighbors=10)
X_bal, y_bal = nm.fit_resample(X, y)
```

Código fonte 3: Aplicação de técnica NearMiss para balanceamento de classes.

2. Para cada instância da classe maioritária, computa-se a média das distâncias para as três instâncias minoritárias mais próximas.

3. Seleciona-se as instâncias da classe maioritária que apresentam a menor média de distância, enfatizando aquelas que estão na fronteira entre as classes.

4. Remove-se do conjunto original as demais instâncias da classe maioritária que não atendem a esse critério.

5. Retorna-se o conjunto de dados balanceado, favorecendo uma definição mais precisa da fronteira de decisão.

A Figura 3.6 ilustra um processo genérico de balanceamento de classes utilizando o método NearMiss.

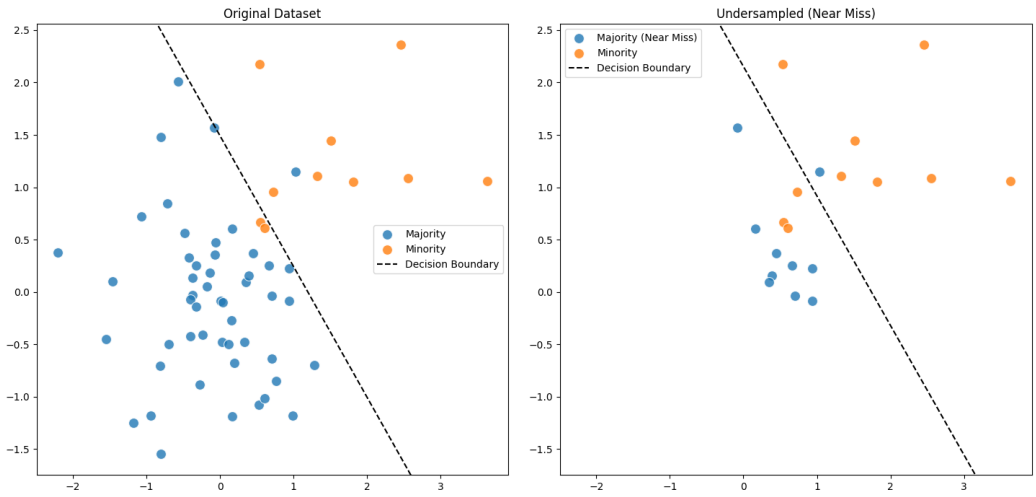

Figura 3.6: Processo de balanceamento de classes utilizando NearMiss versão 1.

Os dados foram divididos em dados de treinamento e teste, conforme mostrado no excerto de código fonte 4, sendo a divisão aleatória de 20% para teste e 80% para treinamento.

A padronização foi utilizada para ajustar algumas variáveis. Esse processo garantiu que os dados tivessem média zero ($\mu = 0$) e desvio padrão igual a um ($\sigma = 1$) (Sujon et al., 2024). A equação utilizada foi:

```
# Dividir em treinamento e teste
X_train, X_test, y_train, y_test = train_test_split(X_bal, y_bal,
↪  test_size=0.2,random_state=74)
```

Código fonte 4: Divisão dos dados em conjuntos de treinamento e teste.

$$X_{scaled} = \frac{X - \mu}{\sigma}$$

- X = valor original da variável
- μ = média da variável
- σ = desvio padrão da variável

Modelos que dependem de cálculos de distância ou gradiente foram beneficiados por essa normalização. O escalonamento evitou pesos desbalanceados em redes neuronais. A regressão logística também se beneficiou, reduzindo viés no aprendizado. Para a padronização foi usada a função *StandardScaler*, conforme mostrado no excerto de código fonte 5.

```
# Padronizar (importante para modelos sensíveis a escala)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Código fonte 5: Padronização dos dados usando *StandardScaler*.

Após o processo de padronização, foram adicionadas à análise duas das técnicas de seleção de características já mencionadas na secção de Revisão de Literatura.

A primeira técnica utilizada foi o RFE, utilizando diferentes quantidades de variáveis (10, 12 e 15). Como não foi observado diferenças relevantes no desempenho dos modelos com diferença de quantidade de variáveis, optou-se por trabalhar com apenas 10 variáveis.

Foram testados diferentes modelos que o RFE utiliza para avaliar a importância das variáveis, os modelos são usualmente chamados de estimadores. Foram utilizados os modelos Árvore de Decisão, Floresta Aleatória, Regressão Logística e *Gradient Boosting*.

Os seguintes passos foram seguidos para realizar a redução das características com o RFE (R.-C. Chen, Manongga e Dewi, 2022):

1. Seleciona o modelo (estimador) que irá avaliar a importância das variáveis.

2. Ajusta o modelo e classifica as caraterísticas com base nas pontuações de importância (por exemplo, coeficientes, importância das caraterísticas).
3. Elimina a variável com a pontuação de importância mais baixa.
4. Treina novamente o modelo utilizando o conjunto reduzido de caraterísticas.
5. Continua a remover a caraterística menos importante e a treinar novamente até restar apenas o número desejado de caraterísticas.

Na tabela 3.1 são apresentadas as dez variáveis selecionadas por cada um dos estimadores (modelos).

Tabela 3.1: Variáveis selecionadas pelos diferentes estimadores no RFE.

Estimador	**Variáveis Selecionadas**
Random Forest	HighBP, BMI, PhysActivity, GenHlth, MentHlth, PhysHlth, DiffWalk, Age, Education, Income
Decision Tree	HighBP, Smoker, PhysActivity, GenHlth, PhysHlth, Age, Education, Income
Gradient Boosting	BMI, HeartDiseaseorAttack, PhysActivity, Veggies, GenHlth, MentHlth, Age, Education, Income
Logistic Regression	BMI, Stroke, HeartDiseaseorAttack, PhysActivity, NoDocbcost, GenHlth, MentHlth, DiffWalk, Income

De forma a perceber quais eram as variáveis de maior consenso entre os estimadores foi criada uma série que contabilizou quantas vezes cada variável foi selecionada pelos diferentes estimadores. Na tabela 3.2 é possível ver a ocorrência de cada uma das variáveis listadas de forma decrescente.

As dez variáveis selecionadas com maior frequência usando os diferentes estimadores e que foram selecionadas para serem testadas são: Índice de Massa Corporal (BMI), Atividade Física (PhysActivity), Saúde Geral (GenHlth), Saúde Mental (MentHlth), Saúde Física (PhysHlth), Rendimento (Income), Dificuldade para Caminhar (DiffWalk), Escolaridade (Education), Histórico de Doença Cardíaca ou Ataque Cardíaco (HeartDiseaseorAttack) e Idade (Age).

Na seleção de características utilizando PCA foram testados diferentes níveis de variância explicada (92%, 95% e 97%). Como não houve grandes variações nos resultados entre essas configurações, optou-se por manter 95% da variância para a avaliação final deste trabalho.

Por meio dos passos a seguir foi feita a redução de características usando PCA (Kammoun, Ravier e Buttelli, 2024):

- Calcula a forma como as caraterísticas variam em conjunto, identificando padrões de correlação.

Tabela 3.2: Contagem de seleção das variáveis.

Variável	**Contagem de Seleção**
Income	4
BMI	4
PhysActivity	4
PhysHlth	4
GenHlth	4
MentHlth	4
Education	4
Age	3
DiffWalk	2
HeartDiseaseorAttack	1
HighBP	1
NoDocbcCost	1
HighChol	1
Veggies	1
Stroke	1
Smoker	1
AnyHealthcare	0
CholCheck	0
HvyAlcoholConsump	0
Sex	0

- Encontra os componentes principais resolvendo os valores próprios (importância) e os vetores próprios (direções no espaço de caraterísticas).
- Classifica os componentes por ordem decrescente de importância, com o primeiro a captar a maior variância.
- Escolhe o conjunto mais pequeno de componentes que explicam um limiar de variância de 95%.
- Projeta os dados originais nos componentes selecionados, reduzindo a dimensionalidade e preservando a informação chave.

A PCA identificou 19 componentes principais, que são novas variáveis linearmente combinadas a partir das originais. Estas componentes explicam pelo menos 95% da variância dos dados, conforme estabelecido.

3.5 Modelação

Em seguimento às etapas de análise e preparação dos dados, foi realizado o processo de modelação aplicando modelos de diferentes níveis de complexidade de forma a fazer uma comparação de seus resultados.

Os modelos foram aplicados em três diferentes cenários: usando as 21 variáveis disponíveis, utilizando as dez variáveis selecionadas pelo RFE e apenas usando as PCA.

Iniciamos a modelação pelo modelo KNN. Esse modelo realiza classificações de um novo ponto (nova entrada de dados) ao buscar os K pontos de treinamento mais próximos e usa a votação (ou média) dos vizinhos. Neste cenário, com um novo vetor de características do paciente atual, o modelo calcula a distância euclidiana entre este paciente e todos os pacientes do conjunto de treinamento. Assim, no processo são selecionados os K pacientes (vizinhos) mais próximos (com distâncias menores). Verifica quantos desses vizinhos têm diabetes e quantos não têm e faz uma votação. Se a maioria tem diabetes, o modelo prevê que o paciente atual provavelmente também tenha.

A Figura 3.7 representa como acontece a classificação no cenário descrito acima.

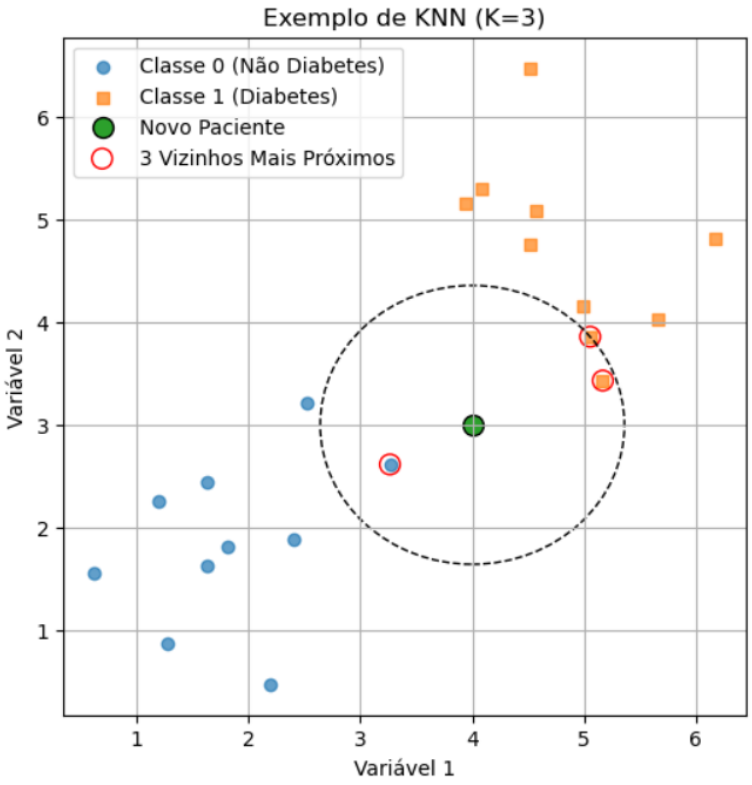

Figura 3.7: Gráfico ilustrativo do método KNN com duas variáveis arbitrárias.

O modelo linear de Regressão Logística analisa a relação direta entre as variáveis de entrada e a variável de saída. Este modelo ajusta uma função linear dos atributos do paciente e aplica a função sigmoide para transformar esse valor em uma probabilidade entre 0 e 1 de ter diabetes. Normalmente, se o resultado é maior que 0,5, classifica como “diabetes”; caso contrário, “não diabetes”.

Na figura 3.8 foi gerado um gráfico que demonstra como o cenário acima descrito acontece.

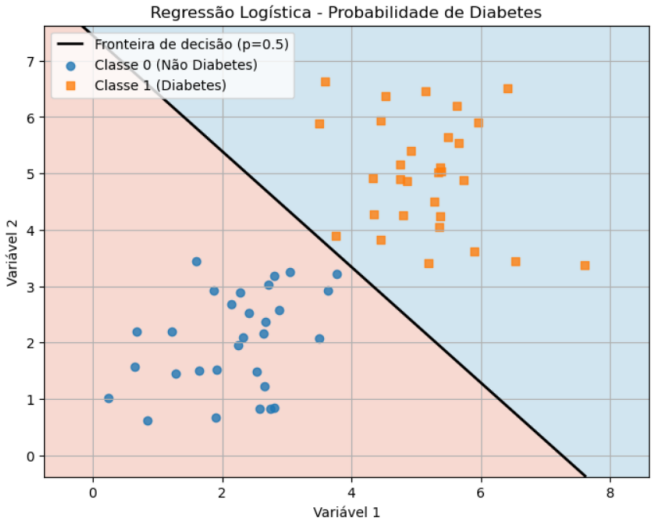

Figura 3.8: Cenário de Regressão Logística aplicada à previsão de diabetes.

O modelo de Árvore de Decisão cria divisões sucessivas nos dados com base em atributos que melhor separam as classes. O algoritmo constrói, a partir dos dados, uma sequência de questões (Exemplo: Age $>$ 5? - Se sim, BMI $>$ 30?). Cada pergunta subdivide mais os pacientes com e sem diabetes, para no fim chegar às "folhas" onde a maior parte dos pacientes tem ou não diabetes.

O modelo de Floresta Aleatória é representado por um conjunto de múltiplas árvores de decisão treinadas com amostras diferentes e subconjuntos de variáveis, e a predição final é feita por votação ou média dessas árvores. O modelo treina várias árvores de decisão em diferentes subconjunto de dados e subconjunto de variáveis. Cada árvore faz a sua predição se é diabetes ou não diabetes. A floresta agrega os resultados por votação e se a maioria for diabetes ou não diabetes é o que ganha a votação.

O modelo *Gradient Boosting* constrói sequencialmente árvores de decisão, onde cada nova árvore corrige os erros do conjunto anterior. O modelo começa com uma predição inicial da diabetes como uma árvore. Em seguida, cada nova árvore é treinada para corrigir os erros do conjunto anterior, de forma iterativa. No final, soma-se as árvores para chegar à predição final.

Já o modelo MLP funciona como redes neuronais compostas por camadas de neurónios totalmente conectadas. O modelo "alimenta" as variáveis (BMI, pressão, etc.) na camada de entrada. Essas informações passam por uma ou mais camadas ocultas, onde são combinadas linearmente e transformadas por funções de ativação, e a camada de saída fornece a probabilidade de diabetes.

Na Figura 3.9 foi gerado um gráfico que demonstra como o cenário acima descrito acontece.

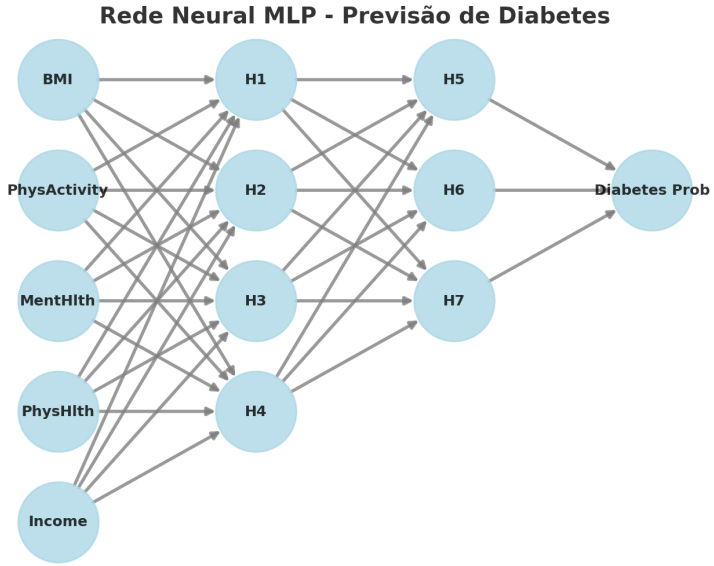

Figura 3.9: Cenário de MLP aplicado à previsão de diabetes.

No fragmento de código fonte 6 a seguir, os modelos anteriormente citados são definidos e é apresentado um ciclo que executa o treinamento e avaliação dos modelos.

```
modelos_baseline = {
    'LogisticRegression': LogisticRegression(random_state=74,
    ↪  max_iter=200),
    'KNN': KNeighborsClassifier(n_neighbors=5),
    'DecisionTree': DecisionTreeClassifier(random_state=74, max_depth=12),
    'RandomForest': RandomForestClassifier(n_estimators=100, max_depth=12,
    ↪  random_state=74),
    'GradientBoosting': GradientBoostingClassifier(random_state=74),
    'MLP': MLPClassifier(hidden_layer_sizes=(100,50), max_iter=300,
    ↪  random_state=74)
}

for nome, modelo in modelos_baseline.items():
    modelo.fit(X_train_scaled, y_train)
    y_pred = modelo.predict(X_test_scaled)
    acc = accuracy_score(y_test, y_pred)
    print(f"Acurácia: {acc:.4f}")
    print(classification_report(y_test, y_pred, digits=3))
```

Código fonte 6: Definição, treinamento e avaliação dos modelos de aprendizagem de máquina.

Abaixo na tabela 3.3 são apresentados os parâmetros utilizados na configuração desses modelos.

Tabela 3.3: Parâmetros utilizados nos modelos de aprendizagem de máquina.

Parâmetro	**Descrição**	**Modelo(s) Utilizado(s)**
`random_state=74`	Define o parâmetro aleatório (*seed*) para reprodutibilidade.	Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, MLP
`max_iter=200`	Número máximo de iterações para convergência.	Logistic Regression
`n_neighbors=5`	Número de vizinhos para considerar na classificação.	KNN
`max_depth=12`	Profundidade máxima da árvore de decisão, limitando complexidade.	Decision Tree, Random Forest
`n_estimators=100`	Número de árvores na floresta aleatória.	Random Forest
`hidden_layer_sizes=(100,50)`	Número de neurónios em cada camada oculta de uma rede neural.	MLP
`max_iter=300`	Número máximo de iterações para o treinamento da rede neuronal.	MLP

O modelo de RNN processa os dados em sequência, mantendo as informações anteriores através das suas conexões internas recorrentes. Com isso, a previsão feita pelo RNN em um determinado passo é influenciada pelos dados do passo atual e também pelo histórico das previsões e entradas anteriores. No caso da predição de diabetes, RNN pode capturar relações temporais, considerando como as mudanças anteriores influenciam o estado atual do paciente.

Também foi realizado o treinamento inicial do modelo baseado em RNN. Esse modelo requer que os dados estejam estruturados como sequências temporais, mesmo que não haja dependência temporal explícita entre as variáveis deste estudo. Por isso, conforme mostrado no excerto de código fonte 7, foi necessário reorganizar os dados em um formato composto por amostras, passos temporais e características. Como o conjunto não possui informações temporais, cada registo foi tratado como uma

sequência com apenas um passo temporal, resultando em uma estrutura compatível com o modelo RNN.

```
X_train_rnn = X_train_scaled.reshape(X_train_scaled.shape[0], 1,
↪  X_train_scaled.shape[1])
X_test_rnn = X_test_scaled.reshape(X_test_scaled.shape[0], 1,
↪  X_test_scaled.shape[1])
```

Código fonte 7: Reformatação dos dados para uso em RNN.

Em seguida, foi realizado o treinamento do modelo. Os hiperparâmetros selecionados para esse modelo são apresentados na tabela 3.4, juntamente com uma breve explicação sobre o papel que cada um desempenha no processo de treinamento.

No excerto de código fonte 8 é demonstrada a implementação e o treinamento do modelo RNN utilizando a linguagem Python.

```
rnn_baseline = Sequential([
    SimpleRNN(50, activation='tanh', return_sequences=False,
    ↪  input_shape=(1, X_train_scaled.shape[1])),
    Dense(25, activation='relu'),
    Dense(1, activation='sigmoid')
])

# Treinamento
rnn_baseline.compile(optimizer=Adam(learning_rate=0.001),
                     loss='binary_crossentropy',
                     metrics=['accuracy'])

rnn_baseline.fit(X_train_rnn, y_train, epochs=10, batch_size=32, verbose=1)
y_pred_rnn_baseline = (rnn_baseline.predict(X_test_rnn) >
↪  0.5).astype(int).flatten()

acc_rnn_baseline = accuracy_score(y_test, y_pred_rnn_baseline)
print(f"\n[RNN Baseline] Acurácia: {acc_rnn_baseline:.4f}")
print(classification_report(y_test, y_pred_rnn_baseline, digits=3))
```

Código fonte 8: Definição, compilação e treinamento do modelo RNN.

3.6 Avaliação dos resultados

Para a análise dos resultados, foram avaliados seis cenários diferentes, incluindo modelos aplicados com dados desequilibrados e equilibrados e também a comparação do desempenho dos modelos sem seleção de características, com seleção de características utilizando RFE, e com redução de dimensionalidade utilizando PCA.

A seguir, na tabela 3.5, são apresentados os resultados obtidos no cenário em que não foi realizado o equilíbrio dos dados nem a seleção de características.

Tabela 3.4: Descrição detalhada dos hiperparâmetros utilizados no treinamento da RNN.

Hiperparâmetro	**Descrição e configuração utilizada**
Tipo da camada recorrente	**SimpleRNN**: tipo de rede recorrente utilizado, sendo a estrutura mais simples, adequada para dados sem dependências temporais complexas.
Neurónios da camada recorrente	Número de unidades neuronais na camada recorrente. Foram utilizados 50 neurónios, permitindo capturar relações relevantes nos dados.
Camada densa intermediária	Uma camada totalmente conectada após a camada recorrente, contendo 25 neurónios com ativação ReLU, para aumentar a capacidade do modelo de aprender representações não-lineares.
Taxa de aprendizado (*learning rate*)	Define a velocidade com que o modelo atualiza seus pesos durante o treinamento. Foi utilizado o valor padrão de 0,001 para permitir uma convergência estável.
Função de ativação da camada intermediária	Função de ativação ReLU (*Rectified Linear Unit*), escolhida por sua capacidade de reduzir problemas de gradiente durante o treinamento.
Função de perda	**Binary Crossentropy**, função adequada para problemas de classificação binária, como este caso (diabetes/não diabetes).
Épocas (*epochs*)	Foram utilizadas 10 épocas, ou seja, o número de vezes que o conjunto completo de dados é passado integralmente pelo modelo durante o treinamento.
Tamanho do lote (*batch size*)	O modelo atualizou os pesos a cada 32 exemplos observados.
Taxa de aprendizado (*learning rate*)	Utilizou-se 0,001, uma taxa padrão que permite um treinamento gradual e estável.

Tabela 3.5: Resultados dos Modelos Sem Seleção de Características - Dados Desequilibrados.

Modelo	**Acurácia**	**Precisão**	**Recall**	**F1-score**
KNN	0.8339	0.641	0.580	0.595
Regressão Logística	0.8511	0.699	0.562	0.574
Árvores de Decisão	0.8435	0.666	0.577	0.592
Floresta Aleatória	**0.8541**	**0.737**	0.547	0.550
Gradient Boosting	0.8540	0.717	0.570	0.586
MLP	0.8432	0.666	0.579	0.595
RNN	0.8538	0.712	**0.580**	**0.599**

O modelo que obteve maior acurácia foi a Floresta Aleatória (0.8541), seguido pelo RNN (0.8538). A Floresta Aleatória obteve também a maior precisão (0.737), sugerindo que houve menor quantidade de falsos positivos. O RNN obteve melhor recall (0.580) que a Floresta Aleatória, ou seja, identificou corretamente uma quantidade maior de cenários positivos. O RNN também obteve maior valor de *F1-score* (0.599). Neste contexto, Floresta Aleatória seria a escolha preferível para minimizar falsos positivos. Entretanto, se o objetivo for um desempenho geral mais equilibrado e alinhado aos objetivos do modelo, o RNN seria uma melhor escolha.

Na tabela 3.6 são apresentados os resultados obtidos no cenário em que não foi realizado o equilíbrio dos dados e foi aplicada a seleção de características com RFE.

Tabela 3.6: Resultados dos Modelos com Seleção de Características (RFE) - Dados Desequilibrados.

Modelo + RFE	**Acurácia**	**Precisão**	**Recall**	**F1-score**
KNN + RFE	0.8343	0.643	0.583	0.598
Regressão Logística + RFE	0.8506	0.696	0.559	0.570
Árvores de Decisão + RFE	0.8483	0.684	0.574	0.591
Floresta Aleatória + RFE	0.8541	**0.725**	0.557	0.567
Gradient Boosting + RFE	0.8541	0.720	0.565	0.579
MLP + RFE	0.8491	0.689	**0.578**	**0.596**
RNN + RFE	**0.8547**	0.723	0.567	0.582

O modelo com melhor desempenho foi o RNN + RFE (0.8547). A Floresta Aleatória + RFE possui a maior precisão (0.725), entretanto teve uma queda em comparação ao cenário do qual não usava RFE o que pode sugerir um leve aumento de falsos positivos. O MLP + RFE tem o melhor recall (0.578) e (F1-score) de (0.596) significando uma melhor relação entre falsos positivos e verdadeiros positivos.

Na tabela 3.7 são apresentados os resultados obtidos no cenário em que não foi realizado o equilíbrio dos dados e foi aplicada a seleção de características com PCA.

O modelo com melhor acurácia foi o RNN + PCA (0.8536), embora não haja

Tabela 3.7: Resultados dos Modelos com Seleção de Características (PCA) - Dados Desequilibrados.

Modelo + PCA	**Acurácia**	**Precisão**	**Recall**	**F1-score**
KNN + PCA	0.8346	0.641	0.578	0.592
Regressão Logística + PCA	0.8510	0.700	0.557	0.566
Árvores de Decisão + PCA	0.8404	0.653	0.570	0.584
Floresta Aleatória + PCA	0.8519	**0.730**	0.531	0.522
Gradient Boosting + PCA	0.8531	0.725	0.546	0.549
MLP + PCA	0.8441	0.670	0.582	0.600
RNN + PCA	**0.8536**	0.710	**0.583**	**0.602**

diferença significativa em relação aos demais modelos anteriores (sem PCA). A maior precisão foi obtida pelo modelo Regressão Logística + PCA (0.700) e outros modelos como Gradient Boosting + PCA (0.725) e Floresta Aleatória + PCA (0.730) também apresentaram precisões superiores. O modelo RNN + PCA obteve maior *Recall* e *F1-score*. Usando PCA houve uma leve diminuição em precisão, indicado aumento de falsos positivos.

Em seguida, os modelos foram testados com o conjunto de dados já equilibrado e, a seguir, serão comparados os seus resultados.

Na tabela 3.8 são apresentados os resultados obtidos no cenário em que os dados estão equilibrados e que não foi aplicada a seleção de características.

Tabela 3.8: Resultados dos Modelos Sem Seleção de Características - Dados Equilibrados.

Modelo	**Acurácia**	**Precisão**	**Recall**	**F1-score**
KNN	0.7964	0.814	0.797	0.794
Regressão Logística	0.8504	0.859	0.850	0.850
Árvores de Decisão	0.8445	0.866	0.845	0.842
Floresta Aleatória	0.8629	**0.878**	0.863	0.862
Gradient Boosting	0.8615	0.870	0.862	0.861
MLP (Neural Network)	0.8570	0.863	0.856	0.856
RNN (LSTM)	**0.8684**	0.877	**0.868**	**0.868**

O modelo de RNN se destacou em todas as métricas analisadas. A Floresta Aleatória também obteve ótimos resultados e próximos ao RNN, inclusive ligeiramente com maior precisão. Caso a complexidade do RNN não seja desejável, a Floresta Aleatória é uma boa alternativa.

Com dados equilibrados, houve uma melhoria geral. O recall (0.868) indicando uma boa taxa de positivos verdadeiros, e maior precisão (0.877) indicando menos falsos positivos.

Em seguida foi analisado o desempenho dos modelos após a seleção de características com RFE, e os resultados demonstraram que não houve um impacto significativo na redução das características.

Na tabela 3.9 são apresentados os resultados obtidos no cenário em que os dados estão equilibrados e em que foi aplicado o RFE.

Tabela 3.9: Resultados dos Modelos com RFE (10 variáveis) - Dados Equilibrados.

Modelo	**Acurácia**	**Precisão**	**Recall**	**F1-score**
KNN + RFE	0.8238	0.835	0.824	0.822
Regressão Logística + RFE	0.8424	0.852	0.842	0.841
Árvores de Decisão + RFE	0.8471	0.864	0.847	0.845
Floresta Aleatória + RFE	0.8587	0.873	0.859	0.857
Gradient Boosting + RFE	0.8590	0.868	0.859	0.858
MLP + RFE	**0.8634**	**0.875**	**0.863**	**0.862**
RNN + RFE	0.8600	0.868	0.860	0.859

O modelo MLP + RFE obteve o melhor desempenho em todas as métricas avaliadas. O modelo RNN + RFE também permaneceu consistente, tendo apresentado desempenho muito próximo.

O modelo manteve um desempenho alto utilizando RFE apesar de não haver grandes melhorias. A utilização do RFE pode facilitar a adoção do modelo, reduzindo a complexidade por meio da redução da quantidade de dados necessários.

Na tabela 3.10 são apresentados os resultados obtidos no cenário em que os dados estavam equilibrados e em que cenário foi aplicado a PCA.

Tabela 3.10: Resultados dos Modelos com PCA - Dados Equilibrados.

Modelo	**Acurácia**	**Precisão**	**Recall**	**F1-score**
KNN + PCA	0.7983	0.813	0.798	0.796
Regressão Logística + PCA	0.8412	0.849	0.841	0.840
Árvores de Decisão + PCA	0.8109	0.821	0.811	0.809
Floresta Aleatória + PCA	0.8304	0.841	0.830	0.829
Gradient Boosting + PCA	0.8347	0.843	0.835	0.834
MLP + PCA	0.8514	0.856	0.851	0.851
RNN + PCA	**0.8572**	**0.863**	**0.857**	**0.857**

O modelo RNN + PCA obteve o melhor resultado em todas as métricas. O modelo MLP + PCA também apresentou desempenho próximo, com precisão de (0.856) e F1-score (0.857). O modelo Gradient Boosting + PCA apresentou bom desempenho em seu recall de (0.835).

O uso de PCA manteve bom desempenho, mas com uma leve redução comparado ao cenário sem seleção de características. A PCA apresentou menor eficiência do

que RFE, entretanto, ainda pode viável em termos médicos mantendo níveis bons de verdadeiros positivos e falsos positivos moderados.

Em resumo, os melhores cenários envolvem o balanceamento dos dados, confirmando que essa técnica melhora significativamente o desempenho dos modelos. As técnicas de seleção ou redução de dimensionalidade tiveram um leve impacto positivo.

O modelo RNN destacou-se como o melhor modelo na maioria dos cenários com dados equilibrados, especialmente em termos de F1-score.

A técnica de seleção com RFE teve um desempenho positivo, especialmente para o MLP, mas não superou significativamente o RNN original sem seleção de características.

A redução de dimensionalidade com PCA também obteve bons resultados, mas inferiores à utilização direta dos dados equilibrados e sem seleção de características.

Capítulo 4

Conclusões

O objetivo do presente trabalho centrou-se em desenvolver e analisar modelos preditivos capazes de identificar indivíduos com maior risco de desenvolver diabetes. Os resultados apresentados anteriormente mostram que os modelos têm a capacidade de realizar boas previsões, e que uma série de características físicas e comportamentais são correlacionadas à propensão ao desenvolvimento da diabetes.

Com o trabalho desenvolvido, pode-se afirmar que os objetivos definidos foram todos cumpridos com sucesso.

Os resultados demonstraram que o equilíbrio dos dados permitiu melhoria significativa no desempenho geral dos modelos, destacando a importância dessa técnica na obtenção de resultados mais robustos.

O modelo de RNN se destacou consistentemente como uma das melhores opções em todos os cenários analisados, especialmente nos dados equilibrados, alcançando o melhor desempenho geral com acurácia de 0,8684 e *F1-score* de 0,868 no cenário sem aplicação de seleção de características.

Além disso, o modelo Floresta Aleatória também apresentou resultados próximos ao RNN. Este modelo pode ser alternativa para um cenário onde as necessidades são de menor complexidade, sendo especialmente adequada para contextos onde é importante minimizar falsos positivos.

A técnica de RFE obteve bons resultados, principalmente com o modelo MLP. Este modelo apresentou melhor desempenho no cenário de seleção de características e dados equilibrados. Utilizar esta técnica de seleção de características pode ser

relevante quando se pretende reduzir a complexidade do conjunto de dados sem perder valores significativos na performance.

A PCA manteve desempenho bom, porém inferior em relação aos dados equilibrados sem seleção ou com uso de RFE, indicando que a PCA pode não ser a opção mais interessante para manutenção da interpretabilidade e precisão.

Estes resultados reforçam a importância da seleção cuidadosa de modelos e métodos de pré-processamento, demonstrando que nem sempre o uso de técnicas de seleção de características garante melhores resultados. A escolha de métodos de pré-processamento e de modelos depende de cada aplicação – é preciso considerar o balanço entre interpretabilidade, simplicidade do modelo e recursos computacionais disponíveis.

Com isso, a relevância deste estudo está na análise de estratégias para deteção dos grupos mais atingidos pela diabetes, proporcionando reflexões acerca desses fatores para profissionais da saúde e contribuindo para o aprimoramento de sistemas de suporte à decisão médica.

4.1 Limites e Desafios

A limitação primária deste trabalho está relacionada à dificuldade de encontrar um conjunto de dados relevantes para o estudo. No conjunto que foi selecionado, os dados coletados referem-se a um único período e a uma única região geográfica, o que pode limitar a generalização dos modelos para outras populações. Além disso, o conjunto de dados utilizado possui um desequilíbrio entre as classes, o que tende a influenciar na capacidade dos modelos de prever corretamente casos de diabetes.

Outro desafio está relacionado ao uso de técnicas de DL, como as Redes Neuronais Recorrentes, que exigem maior capacidade computacional e tempo de treinamento. Esses fatores podem dificultar a implementação em alguns cenários de recursos computacionais mais limitados.

A seleção de atributos também é um fator crítico, que exigiu extensa revisão da literatura para garantir a coerência da solução desenvolvida. A aplicação do RFE e da PCA demonstrou que, em alguns casos, a remoção de variáveis pode comprometer a acurácia dos modelos. Isso evidencia a necessidade de um estudo mais aprofundado sobre a relevância de cada atributo na previsão do diabetes.

Por fim, a interpretação dos modelos e dos resultados obtidos representa um desafio adicional. Modelos como *Gradient Boosting* e Redes Neuronais, embora eficazes, possuem menor interpretabilidade em comparação a modelos mais simples, como a Regressão Logística. Isso pode impactar a compreensão e aceitação das previsões por profissionais de saúde e outras partes interessadas.

4.2 Trabalho futuro

Apesar do desempenho positivo dos modelos de ML e DL, o seu grande desafio na área da saúde e mais especificamente na diabetes é a necessidade de maior transparência no processo de tomada de decisão, de forma que os profissionais de saúde possam compreender e confiar nas análises preditivas geradas por esses modelos (Hendawi, J. Li e Roy, 2023).

Uma direção promissora para trabalhos futuros é o ajuste fino dos limiares de decisão dos modelos. Em particular, a aplicação de técnicas de ajuste do limiar (*thresholding*) na probabilidade de classificação pode ser realizada com o intuito de que, quando ocorrerem erros, estes se manifestem predominantemente como falsos positivos e não como falsos negativos. Essa estratégia é fundamental na área da saúde, pois a minimização dos falsos negativos previne a omissão de casos positivos, mesmo que isso conduza a um aumento na taxa de falsos positivos. Assim, a calibração dos limiares, orientada para uma abordagem mais conservadora, pode aprimorar a segurança e a eficácia dos diagnósticos, contribuindo para uma maior confiabilidade das análises preditivas por parte dos profissionais de saúde.

Além disso, com o objetivo de tornar a aplicação do projeto mais relevante e confiável para a área médica, conceitos de Inteligência artificial explicável (*Explainable AI*) podem ser aplicados a análises futuras envolvendo modelos mais complexos.

A inteligência artificial explicável é um conjunto de técnicas e métodos que visam tornar os modelos mais compreensíveis, ajudando médicos e especialistas a entender como as decisões são tomadas de forma mais compreensível (Chaddad et al., 2023).

Estes métodos poderiam agregar ao estado da arte em diagnósticos de diabetes, fornecendo ferramentas para realizar a seleção e interpretação de variáveis relevantes, além de ajudar médicos a entender quais fatores deveriam ter maior atenção, e para identificar possíveis vieses nos modelos preditivos.

Referências

Abdulkareem, Sulyman Age e Zainab Olorunbukademi Abdulkareem (2021). «An Evaluation of the Wisconsin Breast Cancer Dataset using Ensemble Classifiers and RFE Feature Selection Technique». Em: *International Journal of Sciences: Basic and Applied Research (IJSBAR)* 55.2, pp. 67–80. ISSN: 2307-4531. URL: http://gssrr.org/index.php?journal=JournalOfBasicAndApplied.

Aburass, Sanad e Maha Abu Rumman (2024). «Quantifying Overfitting: Introducing the Overfitting Index». Em: *IV International Conference on Electrical, Computer and Energy Technologies (ICECET)*. IEEE, pp. 1–6. DOI: 10.1109/ICECET61485.2024.10698575. URL: https://ieeexplore.ieee.org/document/10698575.

Afsaneh, Elaheh et al. (2022). «Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review». Em: *Diabetology & Metabolic Syndrome* 14.196, pp. 1–39. DOI: 10.1186/s13098-022-00969-9. URL: https://doi.org/10.1186/s13098-022-00969-9.

Ahmad, Noor e Ali Bou Nassif (2022). «Dimensionality Reduction: Challenges and Solutions». Em: *ITM Web of Conferences* 43, p. 01017. DOI: 10.1051/itmconf/2022430101717.

Ahmed, Shams Forruque et al. (2023). «Deep learning modelling techniques: current progress, applications, advantages, and challenges». Em: *Artificial Intelligence Review* 56, pp. 13521–13617. DOI: 10.1007/s10462-023-10466-8. URL: https://doi.org/10.1007/s10462-023-10466-8.

Ali, Haider et al. (2024). «Comparison of Machine Learning Models for Predicting Interstitial Glucose Using Smart Watch and Food Log». Em: *Electronics* 13.16. ISSN: 2079-9292. DOI: 10.3390/electronics13163192. URL: https://www.mdpi.com/2079-9292/13/16/3192.

Almuqati, Mohammed Tuays et al. (2024). «Challenges in Supervised and Unsupervised Learning: A Comprehensive Overview». Em: *International Journal on Advanced Science, Engineering and Information Technology (IJASEIT)* 14.4. ISSN: 2088-5334.

Alzyoud, Mazen et al. (2024). «Diagnosing diabetes mellitus using machine learning techniques». Em: *International Journal of Data and Network Science* 8,

pp. 179–188. DOI: 10.5267/j.ijdns.2023.10.006. URL: https://www.GrowingScience.com/ijds.

Association, American Diabetes (2021). «Standards of Medical Care in Diabetes». Em.

Bachmann, Gregor, Sotiris Anagnostidis e Thomas Hofmann (2023). *Scaling MLPs: A Tale of Inductive Bias*. arXiv: 2306.13575 [cs.LG]. URL: https://arxiv.org/abs/2306.13575.

Bailey, Clifford J. e Caroline Day (1989). «Traditional Plant Medicines as Treatments for Diabetes». Em: *Diabetes Care* 12.8, pp. 553–564. DOI: 10.2337/diacare.12.8.553. URL: https://doi.org/10.2337/diacare.12.8.553.

Bajcsi, Adél, Anca Andreica e Camelia Chira (2021). «Towards feature selection for digital mammogram classification». Em: *Procedia Computer Science* 192, pp. 632–641. DOI: 10.1016/j.procs.2021.08.065. URL: https://www.sciencedirect.com/science/article/pii/S1877050921015520.

Bansal, Malti, Apoorva Goyal e Apoorva Choudhary (2022). «A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning». Em: *Decision Analytics Journal* 3, p. 100071. ISSN: 2772-6622. DOI: https://doi.org/10.1016/j.dajour.2022.100071. URL: https://www.sciencedirect.com/science/article/pii/S2772662222000261.

El-Bashbishy, Abeer El-Sayyid e Hazem M. El-Bakry (2024). «Pediatric diabetes prediction using deep learning». Em: *Scientific Reports* 14, p. 4206. DOI: 10.1038/s41598-024-51438-4. URL: https://doi.org/10.1038/s41598-024-51438-4.

Bliss, Michael (1982). *The Discovery of Insulin*. Chicago: University of Chicago Press.

Bommert, Andrea et al. (2022). «Benchmark of filter methods for feature selection in high-dimensional gene expression survival data». Em: *Briefings in Bioinformatics* 23.1, pp. 1–13. DOI: 10.1093/bib/bbab354. URL: https://doi.org/10.1093/bib/bbab354.

Byeon, Haewon (2023). «Advances in Value-based, Policy-based, and Deep Learning-based Reinforcement Learning». Em: *International Journal of Advanced Computer Science and Applications* 14.8, pp. 348–354. DOI: 10.14569/IJACSA.2023.0140838. URL: https://www.researchgate.net/publication/373550702.

Chaddad, Ahmad et al. (jan. de 2023). «Survey of Explainable AI Techniques in Healthcare». Em: *Sensors* 23.2, p. 634. DOI: 10.3390/s23020634. URL: https://www.mdpi.com/1424-8220/23/2/634.

Chadia, Mohamed-Amine e Hajar Mousannifa (2023). «Understanding Reinforcement Learning Algorithms: The Progress from Basic Q-learning to Proximal

Policy Optimization». Em: *arXiv preprint arXiv:2304.00026*. URL: https://arxiv.org/abs/2304.00026.

Chaos, Interactive (2024). *Random Forest | Interactive Chaos*. URL: https://interactivechaos.com/en/wiki/random-forest (acedido em 12/02/2025).

Chen, Chao et al. (2024). «An automatically recursive feature elimination method based on threshold decision in random forest classification». Em: *Geo-spatial Information Science*. DOI: 10.1080/10095020.2024.2387457. URL: https://doi.org/10.1080/10095020.2024.2387457.

Chen, Rung-Ching, William Eric Manongga e Christine Dewi (2022). «Recursive Feature Elimination for Improving Learning Points on Hand-Sign Recognition». Em: *Future Internet* 14.12. ISSN: 1999-5903. DOI: 10.3390/fi14120352. URL: https://www.mdpi.com/1999-5903/14/12/352.

Chou, Chun-Yang, Ding-Yang Hsu e Chun-Hung Chou (2023). «Predicting the Onset of Diabetes with Machine Learning Methods». Em: *Journal of Personalized Medicine* X.Y, Z. DOI: 10.XXXX/jpmXXXX. URL: https://www.mdpi.com/journal/jpm.

Collins, Gary S. et al. (2011). «Developing Risk Prediction Models for Type 2 Diabetes: A Systematic Review of Methodology and Reporting». Em: *BMC Medicine* 9, p. 103. DOI: 10.1186/1741-7015-9-103.

Concepts, ML (2025). *(12) Overfitting and Underfitting in Machine Learning | LinkedIn*. Visitado em 4 de Janeiro de 2025. URL: https://www.linkedin.com/pulse/overfitting-underfitting-machine-learning-ml-concepts-com/ (acedido em 11/02/2025).

Daghistani, Tahani e Riyad Alshammari (2020). «Comparison of Statistical Logistic Regression and RandomForest Machine Learning Techniques in Predicting Diabetes». Em: *Journal of Advances in Information Technology* 11.2, pp. 78–83. DOI: 10.12720/jait.11.2.78-83. URL: https://doi.org/10.12720/jait.11.2.78-83.

Desai, Meha e Manan Shah (2021). «An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN)». Em: *Clinical eHealth* 4, pp. 1–11. DOI: 10.1016/j.ceh.2020.11.002. URL: https://doi.org/10.1016/j.ceh.2020.11.002.

Dong, Shi, Ping Wang e Khushnood Abbas (2021). «A survey on deep learning and its applications». Em: *Computer Science Review* 40, p. 100379. DOI: 10.1016/j.cosrev.2021.100379.

Dorador, Albert (2024). «Theoretical and Empirical Advances in Forest Pruning». Em: *arXiv preprint* 2401.05535v3. URL: https://arxiv.org/abs/2401.05535.

Effrosynidis, Dimitrios e Avi Arampatzis (2021). «An evaluation of feature selection methods for environmental data». Em: *Ecological Informatics* 61, p. 101224. DOI:

10.1016/j.ecoinf.2021.101224. URL: https://doi.org/10.1016/j.ecoinf.2021.101224.

Ersozlu, Z., S. Taheri e I. Koch (2024). «A review of machine learning methods used for educational data». Em: *Education and Information Technologies* 29, pp. 22125–22145. DOI: 10.1007/s10639-024-12704-0. URL: https://doi.org/10.1007/s10639-024-12704-0.

Fang, Wei, Yupeng Chen e Qiongying Xue (2021). «Survey on Research of RNN-Based Spatio-Temporal Sequence Prediction Algorithms». Em: *Journal on Big Data* 3.3, pp. 98–110. DOI: 10.32604/jbd.2021.016993.

Federation, International Diabetes (2019). «IDF Diabetes Atlas». Em.

Firdous, Shimoo, Gowher A. Wagai e Kalpana Sharma (nov. de 2022). «A survey on diabetes risk prediction using machine learning approaches». Em: *Journal of Family Medicine and Primary Care* 11.11, pp. 6929–6934. DOI: 10.4103/jfmpc.jfmpc_502_22. URL: https://doi.org/10.4103/jfmpc.jfmpc_502_22.

Friedman, Jerome H. (2001). «Greedy function approximation: A gradient boosting machine». Em: *The Annals of Statistics* 29.5, pp. 1189–1232.

Gao, H., G. Kou, H. Liang et al. (2024). «Machine learning in business and finance: a literature review and research opportunities». Em: *Financial Innovation* 10, p. 86. DOI: 10.1186/s40854-024-00629-z. URL: https://doi.org/10.1186/s40854-024-00629-z.

Gündoğdu, S. (2023). «Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique». Em: *Multimedia Tools and Applications* 82, pp. 34163–34181. DOI: 10.1007/s11042-023-15165-8.

Halder, Rajib Kumar et al. (ago. de 2024a). «Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications». Em: *Journal of Big Data* 11.1, p. 113. ISSN: 2196-1115. DOI: 10.1186/s40537-024-00973-y. URL: https://doi.org/10.1186/s40537-024-00973-y.

— (2024b). «Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications». Em: *Journal of Big Data* 11, p. 113. DOI: 10.1186/s40537-024-00973-y. URL: https://doi.org/10.1186/s40537-024-00973-y.

Hendawi, Rasha, Juan Li e Souradip Roy (2023). «A Mobile App That Addresses Interpretability Challenges in Machine Learning–Based Diabetes Predictions: Survey-Based User Study». Em: *JMIR Formative Research* 7, e50328. DOI: 10.2196/50328. URL: https://formative.jmir.org/2023/1/e50328.

Hidayati, Nur e Arief Hermawan (2021). «K-Nearest Neighbor (K-NN) algorithm with Euclidean and Manhattan in classification of student graduation». Em: *Journal of Engineering and Applied Technology* 2.2, pp. 86–91. DOI: 10.21831/jeatech.v2i2.42777. URL: https://journal.uny.ac.id/index.php/jeatech.

Hill-Briggs, Felicia et al. (2021). «Social Determinants of Health and Diabetes: A Scientific Review». Em: *Diabetes Care* 44, pp. 258–279. DOI: 10.2337/dci20-0053. URL: https://doi.org/10.2337/dci20-0053.

Hong, Gil-Sun et al. (2023). «Overcoming the Challenges in the Development and Implementation of Artificial Intelligence in Radiology: A Comprehensive Review of Solutions Beyond Supervised Learning». Em: *Korean Journal of Radiology* 24.11, pp. 1061–1080. DOI: 10.3348/kjr.2023.0393. URL: https://doi.org/10.3348/kjr.2023.0393.

Islam, MD Rashedul et al. (2022). «A Comprehensive Survey on the Process, Methods, Evaluation, and Challenges of Feature Selection». Em: *IEEE Access* 10, pp. 99595–99625. DOI: 10.1109/ACCESS.2022.3205618.

Jeong, Gwang Hoon (mar. de 2020). «Artificial intelligence, machine learning, and deep learning in women's health nursing». Em: *Korean Journal of Women Health Nursing* 26.1, pp. 5–9. DOI: 10.4069/kjwhn.2020.03.11. eprint: 2020Mar17. URL: https://doi.org/10.4069/kjwhn.2020.03.11.

Jian, Yazan et al. (2021). «A Machine Learning Approach to Predicting Diabetes Complications». Em: *Healthcare* 9.12. DOI: 10.3390/healthcare9121712. URL: https://www.mdpi.com/2227-9032/9/12/1712.

Jiang, Lingjing et al. (2022). «Utilizing stability criteria in choosing feature selection methods yields reproducible results in microbiome data». Em: *Biometrics* 78.3, pp. 1155–1167. DOI: 10.1111/biom.13481. URL: https://doi.org/10.1111/biom.13481.

Jiang, Tammy, Jaimie L. Gradus e Anthony J. Rosellini (2020). «Supervised Machine Learning: A Brief Primer». Em: *Behavior Therapy* 51.5, pp. 675–687. DOI: 10.1016/j.beth.2020.05.002. URL: https://doi.org/10.1016/j.beth.2020.05.002.

Kammoun, Amal, Philippe Ravier e Olivier Buttelli (2024). «Impact of PCA Pre-Normalization Methods on Ground Reaction Force Estimation Accuracy». Em: *Sensors* 24.4. ISSN: 1424-8220. DOI: 10.3390/s24041137. URL: https://www.mdpi.com/1424-8220/24/4/1137.

Khan, Qazi Waqas et al. (2024). «An intelligent diabetes classification and perception framework based on ensemble and deep learning method». Em: *PeerJ Computer Science* 10, e1914. DOI: 10.7717/peerj-cs.1914. URL: https://doi.org/10.7717/peerj-cs.1914.

Kumar, P. Suresh et al. (2021). «CatBoost Ensemble Approach for Diabetes Risk Prediction at Early Stages». Em: *2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology(ODICON)*, pp. 1–6. DOI: 10.1109/ODICON50556.2021.9428943.

Lenka, Sudhansu R. et al. (2021). «An Effective Credit Scoring Model Implementation by Optimal Feature Selection Scheme». Em: *2021 International Conference*

on Emerging Smart Computing and Informatics (ESCI). IEEE, pp. 106–109. DOI: 10.1109/ESCI50559.2021.9396911.

Li, Muyuan (2023). «A Practical Significant Technic in Solving Overfitting: Regularization». Em: *Proceedings of the 2nd International Conference on Computing Innovation and Applied Physics (CONF-CIAP 2023)*. DOI: 10.54254/2753-8818/5/20230433. URL: https://doi.org/10.54254/2753-8818/5/20230433.

Macas, Mayra, Chunming Wu e Walter Fuertes (2022). «A survey on deep learning for cybersecurity: Progress, challenges, and opportunities». Em: *Computer Networks* 212, p. 109032. ISSN: 1389-1286. DOI: https://doi.org/10.1016/j.comnet.2022.109032. URL: https://www.sciencedirect.com/science/article/pii/S1389128622001864.

Maharana, Kiran, Surajit Mondal e Bhushankumar Nemade (2022). «A review: Data pre-processing and data augmentation techniques». Em: *Global Transitions Proceedings* 3, pp. 91–99. DOI: 10.1016/j.gltp.2022.04.020. URL: https://doi.org/10.1016/j.gltp.2022.04.020.

Maldonado, Javier, María Cristina Riff e Bertrand Neveu (2022). «A review of recent approaches on wrapper feature selection for intrusion detection». Em: *Expert Systems With Applications* 198, p. 116822. DOI: 10.1016/j.eswa.2022.116822. URL: https://doi.org/10.1016/j.eswa.2022.116822.

Mansour, Afnan et al. (2023). «Microvascular and macrovascular complications of type 2 diabetes mellitus: Exome wide association analyses». Em: *Frontiers in Endocrinology* 14, p. 1143067. DOI: 10.3389/fendo.2023.1143067. URL: https://www.frontiersin.org/articles/10.3389/fendo.2023.1143067/full.

Marsland, Stephen (2014). *Machine Learning: An Algorithmic Perspective*. Chapman e Hall/CRC.

Mienye, Ibomoiye Domor e Nobert Jere (2024). «A Survey of Decision Trees: Concepts, Algorithms, and Applications». Em: *IEEE Access*. This work is licensed under a Creative Commons Attribution 4.0 License. DOI: 10.1109/ACCESS.2024.3416838.

Miller, Anthony, John Panneerselvam e Lu Liu (2022). «A review of regression and classification techniques for analysis of common and rare variants and gene-environmental factors». Em: *Neurocomputing* 489, pp. 466–485. DOI: 10.1016/j.neucom.2021.08.150.

Mousavi, Azita et al. (2023). «Comparison of Feature Extraction with PCA and LTP Methods and Investigating the Effect of Dimensionality Reduction in the Bat Algorithm for Face Recognition». Em: *International Journal of Robotics and Control Systems* 3.3, pp. 500–509. DOI: 10.31763/ijrcs.v3i3.1057.

Mrabet, Mohammed Amine El, Khalid El Makkaoui e Ahmed Faize (2021). «Supervised Machine Learning: A Survey». Em: *4th International Conference on*

Advanced Communication Technologies and Networking (CommNet), pp. 1–10. DOI: 10.1109/CommNet52204.2021.9641998.

Naeem, Samreen et al. (2023). «An Unsupervised Machine Learning Algorithms: Comprehensive Review». Em: *International Journal of Computing and Digital Systems* 13.1, pp. 911–921. DOI: 10.12785/ijcds/130172. URL: https://www.researchgate.net/publication/368983958_An_Unsupervised_Machine_Learning_Algorithms_Comprehensive_Review.

Naskath, J., G. Sivakamasundari e A. Alif Siddiqua Begum (2023). «A Study on Different Deep Learning Algorithms Used in Deep Neural Nets: MLP, SOM, and DBN». Em: *Wireless Personal Communications* 128, pp. 2913–2936. DOI: 10.1007/s11277-022-10079-4. URL: https://doi.org/10.1007/s11277-022-10079-4.

Nguyen, Van Quan et al. (2021). «A Robust PCA Feature Selection to Assist Deep Clustering Autoencoder-Based Network Anomaly Detection». Em: *2021 8th NAFOSTED Conference on Information and Computer Science (NICS)*. IEEE, pp. 335–341. DOI: 10.1109/NICS54270.2021.9701456.

Olisah, Chollette C., Lyndon Smith e Melvyn Smith (2022). «Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective». Em: *Computer Methods and Programs in Biomedicine* 220, p. 106773. DOI: 10.1016/j.cmpb.2022.106773. URL: https://doi.org/10.1016/j.cmpb.2022.106773.

Ono, S e T Goto (jul. de 2022). «Introduction to supervised machine learning in clinical epidemiology». Em: *Annals of Clinical Epidemiology* 4.3, pp. 63–71. DOI: 10.37737/ace.22009.

Ooka, Tadao et al. (2021). «Random Forest Approach for Determining Risk Prediction and Predictive Factors of Type 2 Diabetes: Large-Scale Health Check-up Data in Japan». Em: *BMJ Nutrition, Prevention & Health* 4.1, e000200. DOI: 10.1136/bmjnph-2020-000200. URL: https://bmjnph.bmj.com/content/4/1/e000200.

Panda, Nihar Ranjan et al. (2022). «A Review on Logistic Regression in Medical Research». Em: *National Journal of Community Medicine* 13.4, pp. 265–270. DOI: 10.55489/njcm.134202222. URL: https://www.njcmindia.org.

Pathan, Muhammad Salman et al. (2022). «Analyzing the impact of feature selection on the accuracy of heart disease prediction». Em: *Healthcare Analytics* 2, p. 100060. DOI: 10.1016/j.health.2022.100060. URL: https://doi.org/10.1016/j.health.2022.100060.

Prajapati, Anil Kumar e Umesh Kumar Singh (set. de 2023). «An Optimal Solution to the Overfitting and Underfitting Problem of Healthcare Machine Learning Models». Em: *Journal of Systems Engineering and Information Technology*

(JOSEIT) 02.02, pp. 77–84. DOI: 10.29207/joseit.v2i2.5460. URL: https://doi.org/10.29207/joseit.v2i2.5460.

Priyatno, Arif Mudi e Triyanna Widiyaningtyas (fev. de 2024). «A Systematic Literature Review: Recursive Feature Elimination Algorithms». Em: *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)* 9.2, pp. 196–207. DOI: 10.33480/jitk.v9i2.5015. URL: https://www.researchgate.net/publication/377888030.

Pudjihartono, Nicholas et al. (2022). «A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction». Em: *Frontiers in Bioinformatics* 2, p. 927312. DOI: 10.3389/fbinf.2022.927312. URL: https://www.frontiersin.org/articles/10.3389/fbinf.2022.927312/full.

Purbasari, Ayi et al. (2021). «CRISP-DM for Data Quality Improvement to Support Machine Learning of Stunting Prediction in Infants and Toddlers». Em: *2021 International Conference on Advanced Informatics, Concepts, Theory and Applications (ICAICTA)*. IEEE, pp. 1–8. DOI: 10.1109/ICAICTA53211.2021.9640294.

Rahmani, Amir Masoud et al. (2021). «Machine Learning (ML) in Medicine: Review, Applications, and Challenges». Em: *Mathematics* 9.22, p. 2970. DOI: 10.3390/math9222970. URL: https://www.mdpi.com/2227-7390/9/22/2970.

Rainio, Antti, Mika Kukkonen e Ilkka Pölönen (2024). «Evaluation metrics and statistical tests for machine learning». Em: *Scientific Reports* 14.1, p. 6086. DOI: 10.1038/s41598-024-56706-x. URL: https://doi.org/10.1038/s41598-024-56706-x.

Rufo, D.D. et al. (2021). «Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM)». Em: *Diagnostics* 11.9, p. 1714. DOI: 10.3390/diagnostics11091714.

Rupapara, Vaibhav et al. (2023). «Chi-Square and PCA Based Feature Selection for Diabetes Detection with Ensemble Classifier». Em: *Intelligent Automation & Soft Computing* 36.2, pp. 1932–1949. DOI: 10.32604/iasc.2023.028257. URL: https://doi.org/10.32604/iasc.2023.028257.

Sabiri, Bihi, Bouchra El Asri e Maryem Rhanoui (2022). «Mechanism of Overfitting Avoidance Techniques for Training Deep Neural Networks». Em: *Proceedings of the 24th International Conference on Enterprise Information Systems (ICEIS 2022)*. SCITEPRESS – Science e Technology Publications, Lda., pp. 418–427. DOI: 10.5220/0011114900003179. URL: https://www.researchgate.net/publication/360503583.

Sahid, M. A., M. U. H. Babar e M. P. Uddin (2024). «Predictive modeling of multiclass diabetes mellitus using machine learning and filtering iraqi diabetes data dynamics». Em: *PloS one* 19.5, e0300785. DOI: 10.1371/journal.pone.0300785.

Salman, Hasan Ahmed, Ali Kalakech e Amani Steiti (2024). «Random Forest Algorithm Overview». Em: *Babylonian Journal of Machine Learning* 2024, pp. 69–

79. DOI: 10.58496/BJML/2024/007. URL: https://mesopotamian.press/journals/index.php/BJML.

Samuel, A. L. (1959). «Some studies in machine learning using the game of checkers». Em: *IBM Journal of Research and Development* 3.3, pp. 210–229.

Schröer, Christoph, Felix Kruse e Jorge Marx Gómez (2021). «A Systematic Literature Review on Applying CRISP-DM Process Model». Em: *Procedia Computer Science* 181, pp. 526–534. DOI: 10.1016/j.procs.2021.01.199.

Science, Open Data (2024). *Understanding the Mechanism and Types of Recurring Neuronal Networks*. Accessed: 20 November 2024. URL: https://opendatascience.com/understanding-the-mechanism-and-types-of-recurring-neuronal-networks/.

Shakya, Ashish Kumar, Gopinatha Pillai e Sohom Chakrabarty (2023). «Reinforcement learning algorithms: A brief survey». Em: *Expert Systems With Applications* 231, p. 120495. DOI: 10.1016/j.eswa.2023.120495. URL: https://www.sciencedirect.com/science/article/pii/S0957417423009971.

Shantal, Mohammed, Almahdi Alshareef e Omar Ahmid (2024). «The Impact Of Feature Selection On Diabetes Prediction». Em: *African Journal of Advanced Pure and Applied Sciences (AJAPAS)* 3.3, pp. 373–377. URL: https://aaasjournals.com/index.php/ajapas/index.

Sharifani, Koosha e Mahyar Amini (2023). «Machine Learning and Deep Learning: A Review of Methods and Applications». Em: *World Information Technology and Engineering Journal* 10.07.

Shearer, Colin (2000). «The CRISP-DM Model: The New Blueprint for Data Mining». Em: *Journal of Data Warehousing* 5.4, pp. 13–22.

Sherstinsky, Alex (2020). «Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network». Em: *Physica D: Nonlinear Phenomena* 404, p. 132306. DOI: 10.1016/j.physd.2019.132306. URL: https://www.sciencedirect.com/science/article/abs/pii/S0167278919305974.

Silveira, Maria Beatriz Galdino da et al. (2021). «Aplicação da regressão logística na análise dos dados dos fatores de risco associados à hipertensão arterial». Em: *Research, Society and Development* 10.16, e20101622964. ISSN: 2525-3409. DOI: 10.33448/rsd-v10i16.22964. URL: http://dx.doi.org/10.33448/rsd-v10i16.22964.

Srinivasu, Parvathaneni Naga et al. (2022). «Using Recurrent Neural Networks for Predicting Type-2 Diabetes from Genomic and Tabular Data». Em: *Diagnostics* 12.12, p. 3067. DOI: 10.3390/diagnostics12123067. URL: https://doi.org/10.3390/diagnostics12123067.

Sterlin, Elena (2024). *Health spending takes up 10% of the global economy: How can tech help reduce costs and improve lives?* Acessado em: 21 jan. 2025. URL:

https://www.weforum.org/stories/2024/08/healthcare-costs-digital-tech/?utm_source=chatgpt.com.

Sujon, Khaled Mahmud et al. (set. de 2024). «When to Use Standardization and Normalization: Empirical Evidence From Machine Learning Models and XAI». Em: *IEEE Access* 12, pp. 135300–135314. DOI: 10.1109/ACCESS.2024.3462434. URL: https://doi.org/10.1109/ACCESS.2024.3462434.

Suryasa, I Wayan, María Rodríguez-Gámez e Tihnov Koldoris (2021). «Health and Treatment of Diabetes Mellitus». Em: *International Journal of Health Sciences* 5.1, pp. i–v. DOI: 10.53730/ijhs.v5n1.2864. URL: https://doi.org/10.53730/ijhs.v5n1.2864.

Talaei Khoei, T., H. Ould Slimane e N. Kaabouch (2023). «Deep learning: systematic review, models, challenges, and research directions». Em: *Neural Computing and Applications* 35, pp. 23103–23124. DOI: 10.1007/s00521-023-08957-4. URL: https://doi.org/10.1007/s00521-023-08957-4.

Tanimoto, Akira et al. (2022). «Improving imbalanced classification using near-miss instances». Em: *Expert Systems with Applications* 201, p. 117130. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2022.117130. URL: https://www.sciencedirect.com/science/article/pii/S0957417422005280.

Taye, Mohammad Mustafa (2023). «Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions». Em: *Computers* 12.5, p. 91. DOI: 10.3390/computers12050091. URL: https://doi.org/10.3390/computers12050091.

Teboul, Alex (2015). *Diabetes Health Indicators Dataset*. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset. Accessed: 2024-12-17. URL: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset.

Tigga, Neha Prerna e Shruti Garg (2020). «Prediction of Type 2 Diabetes using Machine Learning Classification Methods». Em: *Procedia Computer Science* 167. International Conference on Computational Intelligence and Data Science, pp. 706–716. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2020.03.336. URL: https://www.sciencedirect.com/science/article/pii/S1877050920308024.

Uddin, Shahadat et al. (2022). «Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction». Em: *Scientific Reports* 12, p. 6256. DOI: 10.1038/s41598-022-10358-x. URL: https://doi.org/10.1038/s41598-022-10358-x.

Vakil, Vishal et al. (2021). «Explainable predictions of different machine learning algorithms used to predict Early Stage Diabetes». Em: *arXiv:2111.09939*. URL: https://arxiv.org/abs/2111.09939.

Wang, Aiguo et al. (2020). «Evaluation of Random Forest for Complex Human Activity Recognition Using Wearable Sensors». Em: *2020 International Conference on Networking and Network Applications (NaNA)*. IEEE, pp. 310–316. DOI: 10.1109/NaNA51271.2020.00060. URL: https://ieeexplore.ieee.org/document/9290356.

Wee, Boon Feng et al. (2024). «Diabetes detection based on machine learning and deep learning approaches». Em: *Multimedia Tools and Applications* 83, pp. 24153–24185. DOI: 10.1007/s11042-023-16407-5. URL: https://doi.org/10.1007/s11042-023-16407-5.

Xie, Zidian et al. (2019). «Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques». Em: *Preventing Chronic Disease* 16, E130. DOI: 10.5888/pcd16.190109.

Yang, Kaixin, Long Liu e Yalu Wen (2024). «The impact of Bayesian optimization on feature selection». Em: *Scientific Reports* 14.1, p. 3948. DOI: 10.1038/s41598-024-54515-w. URL: https://doi.org/10.1038/s41598-024-54515-w.

Zabor, Emily C. et al. (2022). «Logistic Regression in Clinical Studies». Em: *International Journal of Radiation Oncology, Biology, Physics* 112.2, pp. 271–277. DOI: 10.1016/j.ijrobp.2021.08.007. URL: https://www.redjournal.org/article/S0360-3016(21)02646-8/fulltext.

Zhang, Jinxiong (2021). «Dive into Decision Trees and Forests: A Theoretical Demonstration». Em: *arXiv preprint* cs.LG.2101.08656v1. URL: https://arxiv.org/abs/2101.08656.

Zhou, Sheng et al. (2022). «Why is the prediction wrong? Towards underfitting case explanation via meta-classification». Em: *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, pp. 1–10. DOI: 10.1109/DSAA54385.2022.10032332.

Zhou, Xilin et al. (2020). «Cost-effectiveness of Diabetes Prevention Interventions Targeting High-risk Individuals and Whole Populations: A Systematic Review». Em: *Diabetes Care* 43.7, pp. 1593–1616. DOI: 10.2337/dci20-0018. URL: https://doi.org/10.2337/dci20-0018.

Anexo A

Imagens referenciadas nos capítulos da dissertação são reproduzidas, em tamanho completo, nas seguintes páginas dos anexos.

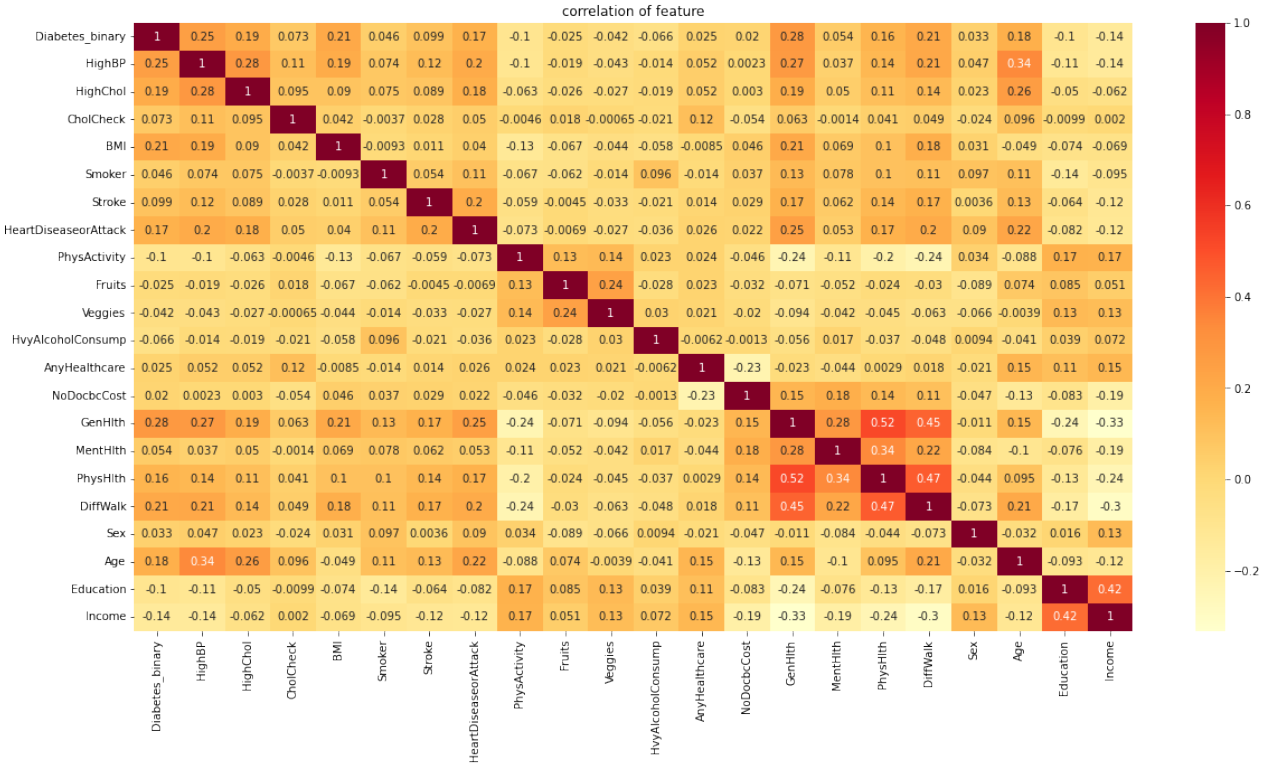

Figura A.1: Mapa de Correlação

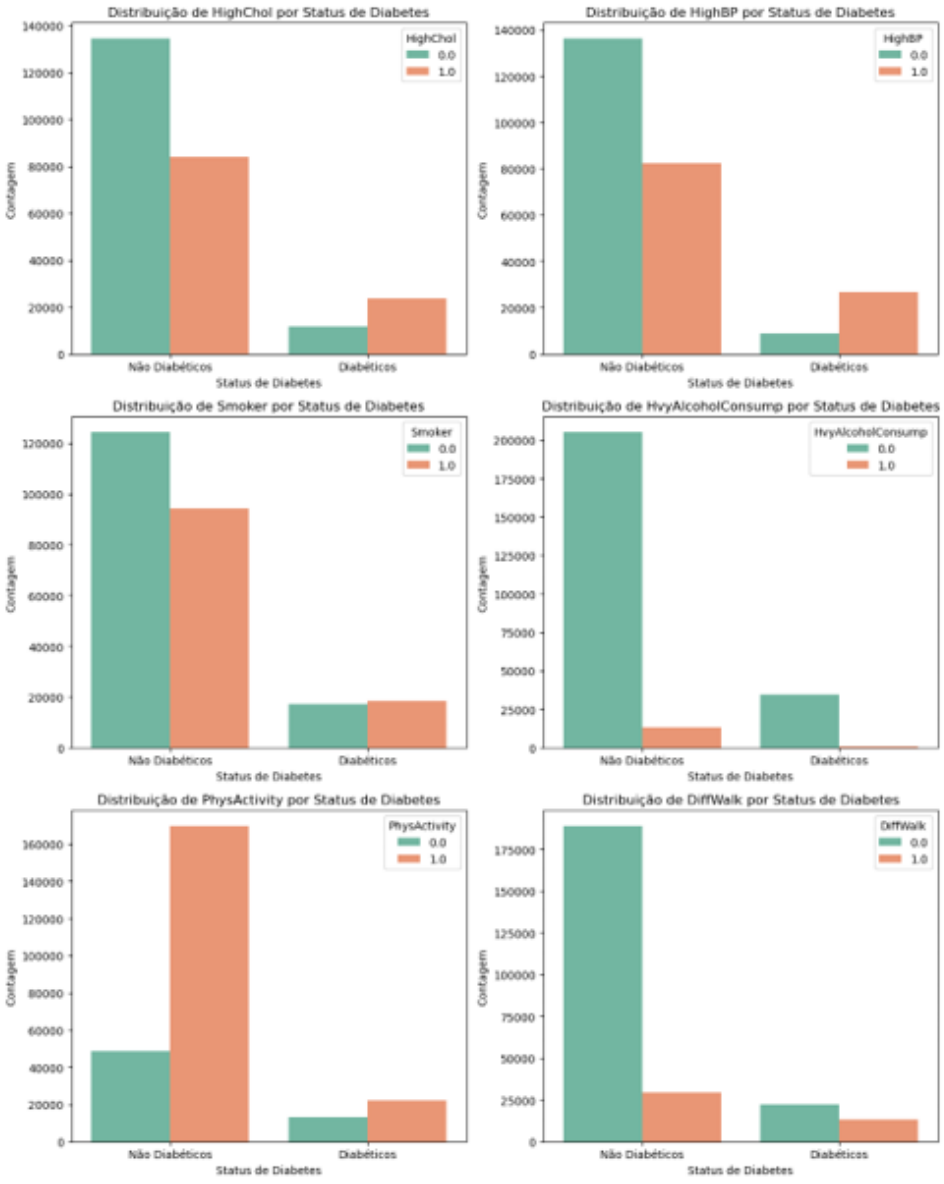

Figura A.2: Fatores de Risco e Diabetes