

CENTRO ESTADUAL DE EDUCAÇÃO TECNOLÓGICA PAULA SOUZA

Faculdade de Tecnologia da Praia Grande

Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas

GUILHERME OZORES PIRES

GUSTAVO PEREIRA VIEIRA

**ALÉM DAS MÉTRICAS: ANÁLISE DE COMPORTAMENTO PREDITIVO NA
SELEÇÃO DE MODELOS DE MACHINE LEARNING PARA O RISCO DE
DIABETES**

Praia Grande

Junho/2025

GUILHERME OZORES PIRES

GUSTAVO PEREIRA VIEIRA

**ALÉM DAS MÉTRICAS: ANÁLISE DE COMPORTAMENTO PREDITIVO NA
SELEÇÃO DE MODELOS DE MACHINE LEARNING PARA O RISCO DE
DIABETES**

Trabalho de Conclusão de Curso
apresentado à Faculdade de
Tecnologia da Praia Grande para a
obtenção do título de Tecnólogo
em Análise e Desenvolvimento de
Sistemas.

Orientador: Prof. Leonardo Villani

Praia Grande

202

GUILHERME OZORES PIRES

GUSTAVO PEREIRA VIEIRA

**ALÉM DAS MÉTRICAS: ANÁLISE DE COMPORTAMENTO PREDITIVO NA
SELEÇÃO DE MODELOS DE MACHINE LEARNING PARA O RISCO DE
DIABETES**

Trabalho de Conclusão de Curso
apresentado à Faculdade de Tecnologia
da Praia Grande para a obtenção do título
de Tecnólogo em Análise e
Desenvolvimento de Sistemas.

Praia Grande, 24 de junho de 2025

Banca Examinadora

Leonardo Vilani

Faculdade de Tecnologia da Praia Grande
Presidente

XXXX

Faculdade de Tecnologia da Praia Grande

XXXX

Faculdade de Tecnologia da Praia Grande

Dedico esse trabalho a, Divo Ozores
e Wagner Gimenez Pires

Os dias prósperos não vem por acaso;
nascem de muita fadiga e persistência

(Henry Ford)

RESUMO

O Diabetes Mellitus tipo 2 é uma doença crônica com prevalência crescente, representando um significativo desafio para a saúde pública. A detecção precoce é um fator crucial para a prevenção de complicações graves e, neste contexto, o objetivo deste trabalho foi desenvolver e avaliar um modelo preditivo de aprendizado de máquina para identificar indivíduos de alto risco, utilizando o dataset *PIMA Indians Diabetes* como estudo de caso. A metodologia envolveu um rigoroso pipeline de pré-processamento, seguido da avaliação de nove algoritmos através de métricas quantitativas (AUC, Recall), otimização de limiar e, de forma decisiva, uma validação qualitativa com perfis clínicos sintéticos. Os resultados demonstraram que a análise de plausibilidade foi indispensável para desqualificar algoritmos com altas métricas que apresentaram comportamentos de risco, como o *Gradient Boosting*. Este processo consagrou o modelo *Random Forest* como a solução superior, por apresentar o melhor equilíbrio entre performance quantitativa e um comportamento preditivo seguro e coerente. Conclui-se que a validação de comportamento clínico é uma etapa indispensável no desenvolvimento de ferramentas de IA responsáveis para a saúde, superando a avaliação baseada unicamente em métricas estatísticas.

Palavras-chave: Aprendizado de Máquina, Diabetes Mellitus Tipo 2, Modelo Preditivo, Ciência de Dados, Pré-processamento de Dados.

ABSTRACT

Type 2 Diabetes Mellitus is a chronic disease with a growing prevalence, representing a significant public health challenge. Early detection is a crucial factor in preventing severe complications, and in this context, the objective of this work was to develop and evaluate a predictive machine learning model to identify individuals at high risk, using the PIMA Indians Diabetes dataset as a case study. The methodology involved a rigorous preprocessing pipeline, followed by the evaluation of nine algorithms through quantitative metrics (AUC, Recall), threshold optimization, and, decisively, a qualitative validation with synthetic clinical profiles. The results showed that, although several models had high metric performance, the plausibility analysis was indispensable for disqualifying algorithms with high-risk behaviors, such as Gradient Boosting. This process established the Random Forest model as the superior solution, presenting the best balance between quantitative performance and safe, coherent predictive behavior. It is concluded that the validation of clinical behavior is a fundamental step for the development of responsible AI tools for healthcare, surpassing evaluations based solely on statistical metrics.

Keywords: Machine Learning, Type 2 Diabetes Mellitus, Predictive Model, Data Science in Health, Data Preprocessing, Model Validation.

LISTA DE ILUSTRAÇÕES

Figura 1 – Ilustração da transformação de dados em informação.....	17
Figura 2 – Guarda-chuvas da ciência de dados.....	21
Figura 3 – Principais de aprendizado de Machine Learning.....	22
Figura 4 – Regressão linear.....	24
Figura 5 – Regressão logística.....	25
Figura 6 – Árvore de decisão.....	26
Figura 7 – Random Forest.....	27
Figura 8 – K-Means.....	29
Figura 9 – Aprendizado de reforço.....	31
Figura 10 – Um exemplo de rede neural profunda 2-(4,2,2)-3.....	32
Figura 11 - Principais fatores de risco.....	38
Figura 12 – Fluxograma para diagnóstico de diabetes mellitus tipo 2.....	42
Figura 13 – “Heatmap” das features para visualizar a correlação entre os itens.....	46
Figura 14 - Estratificação e fluxo ideal.....	49
Figura 15 – IQR.....	50
Figura 16 – Demonstração do funcionamento do SMOTE.....	51
Figura 17 – Normalização de dados.....	52
Figura 18 – Matriz de confusão.....	56
Figura 19 – Demonstração da validação cruzada.....	64
Figura 20 – Curva ROC.....	65
Figura 21 – Área sob a curva AUC.....	67
Figura 22 – Ciclo de vida da mineração de dados.....	72
Figura 23 - Histogramas de frequência e distribuição de classes para todas as variáveis do dataset.....	81
Figura 24 – Matriz de correlação entre todas as variáveis.....	82
Figura 25 - Análise de correlação/dispersão - glicose vs. outras variáveis.....	83
Figura 26 - Análise da Distribuição da glicose por desfecho/outcome.....	84
Figura 27 – Distribuição de <i>BMI/IMC</i> por status de diabetes.....	84
Figura 28 – Quadrantes de risco glicose vs IMC e taxa de diabetes associada.....	86
Figura 29 – Taxa de diabetes por número de fatores de risco.....	87
Figura 30 – Diferenças na distribuição de acordo com cada processamento.....	90
Figura 31 – Dispersão dos pontos inseridos de acordo com cada processamento.....	91
Figura 32 – <i>Boxplot</i> de acordo com cada processamento	92
Figura 33 – <i>Boxplot</i> de remoção de outliers com (1.5xIQR).....	97
Figura 34 - Percentual de <i>outliers</i> removidos por <i>feature</i>	97
Figura 35 - Histogramas mostrando a comparação geral "Antes e Depois" do balanceamento com SMOTE	99
Figura 36 – Dados de treino normalizados.....	100

Figura 37 – Análise completa da performance de todos os modelos	102
Figura 38 – Matriz de confusão dos modelos treinados	104
Figura 39 – Comparação validação vs teste, ranking de performance.....	105
Figura 40 – Comparação validação vs teste, todos os modelos, faixa de corte em 80%.....	106
Figura 41 – Distribuição da importância por <i>feature</i> de todos os modelos....	106
Figura 42 – Importância das features por cada modelos.	107
Figura 43 – Validação cruzada 5-Fold, média e desvio padrão.....	110
Figura 44 – Validação cruzada 5-Fold, performance a cada fold.	110
Figura 45 – Análise completa de <i>threshold</i> dos melhores modelos selecionados.	112
Figura 46 - Diferença da matriz de confusão com <i>threshold</i> 0,5 e 0,4.	115
Figura 47 – Análise dos modelos candidatos finais.....	117

LISTA DE TABELAS

Tabela 1 – Exemplo concreto da definição de Mitchell.....	19
Tabela 2 – Datas e marcos importantes do ML ao longo da história.....	20
Tabela 3 – Critérios laboratoriais para diagnóstico de diabetes.....	41
Tabela 4 – Comparação qualitativa de atributos diagnósticos: métodos tradicionais vs. baseados em ML.....	43
Tabela 5 – Componentes chave e relevância da análise exploratória de dados (AED).....	45
Tabela 6 - Vantagens e desvantagens do SMOTE.....	51
Tabela 7 – Como cada modelo funciona.....	53
Tabela 8 – Onde cada modelo funciona melhor.....	54
Tabela 9 - Estratégias de ajuste de hiperparâmetros.....	55
Tabela 10 – Síntese dos valores da matriz de confusão e siglas.....	58
Tabela 11 – Interpretações de AUC.....	66
Tabela 12 – Processo CRISP-DM adaptado.....	73
Tabela 13 – Modelos e parâmetros utilizados no treinamento.....	77
Tabela 14 – Descrição das variáveis do conjunto de dados/ <i>dataset</i>	79
Tabela 15 - Análise de ocorrência de valores zero por variável.....	80
Tabela 16 – Taxa de Diabetes por classificação de <i>BMI/IMC</i>	85
Tabela 17 - Perfis de risco sintéticos utilizados para análise de plausibilidade dos modelos.....	88
Tabela 18 – Estratificação dos dados.....	94
Tabela 19 - Comparativo Estatístico da Classe Minoritária (Diabético) Antes e Depois do SMOTE	98
Tabela 20 - Métricas de performance detalhadas dos modelos no conjunto de teste.	103
Tabela 21 – Valores da matriz de confusão.....	105
Tabela 22 – Importância das features por peso.....	109

SUMÁRIO

1 INTRODUÇÃO.....	13
1.1 OBJETIVO	15
1.2 JUSTIFICATIVA	15
2 REVISÃO DA LITERATURA.....	17
2.1 UMA BREVE INTRODUÇÃO A CIÊNCIA DE DADOS E <i>MACHINE LEARNING</i>	17
2.2 TIPOS DE <i>MACHINE LEARNING</i>	22
2.2.1 Aprendizado supervisionado	23
2.2.2 Aprendizado não supervisionado.....	29
2.2.3 Aprendizado por reforço	30
2.2.4 <i>Deep learning</i> (aprendizado profundo)	32
2.3 <i>MACHINE LEARNING</i> NA SAÚDE	33
2.4 DIABETES E SUA PREVISÃO	35
2.4.1 Definição abrangente.....	35
2.4.2 Classificação e tipos	35
2.4.3 Fatores de risco e importância da detecção precoce	38
2.4.4 Critérios técnicos para o diagnóstico laboratorial	41
2.4.5 Métodos tradicionais de diagnóstico e abordagens baseadas em <i>Machine Learning</i>	42
2.5 ESTEIRA (PIPELINE) DE <i>MACHINE LEARNING</i> : AED, PREPARAÇÃO, MODELAGEM E AVALIAÇÃO	44
2.5.1 AED – Análise exploratória de dados	44
2.5.2 Preparação dos dados.....	47
2.5.2.1 Técnicas de estratificação de “dataset”, conjunto de dados	47
2.5.2.2 Técnicas de remoção de outliers e balanceamento de “dataset” conjunto de dados	49
2.5.2.3 Técnicas de modelagem e treinamento de algoritmos	52
2.5.2.4 Avaliação de modelos de <i>Machine Learning</i>	55
2.6 INDIANS PIMA DATASET	68
3 MATERIAIS E MÉTODOS.....	70
3.1 FERRAMENTAS & TECNOLOGIAS	70
3.2 PROCESSO CRISP-DM ADAPTADO PARA ANÁLISE ACADÊMICA	72
3.3 ANÁLISE EXPLORATÓRIA DE DADOS (AED)	74
3.4 ETAPAS DO PRÉ-PROCESSAMENTO DE DADOS.....	74
3.4.1 Tratamento de valores zero em variáveis fisiológicas	74

3.4.2	Divisão e estratificação dos dados	75
3.4.3	Remoção de outliers.....	76
3.4.4	Balanceamento das classes	76
3.4.5	Normalização dos dados	76
3.5	MODELAGEM E AVALIAÇÃO	77
3.5.1	Treinamento dos modelos de classificação	77
3.5.2	Métricas de processo de avaliação.....	78
4	RESULTADOS E DISCUSSÕES.....	79
4.1	ANÁLISE EXPLORATÓRIA DE DADOS.....	79
4.2	PRÉ-PROCESSAMENTO DO CONJUNTO DE DADOS	88
4.2.1	Análise comparativa das técnicas de pré-tratamento de dados ...	88
4.2.2	Estratificação (Treino 80%, validação 20%, teste 20%)	93
4.2.3	Remoção de outliers (IQR) e	96
4.2.4	Balanceamento: SMOTE	98
4.2.5	Normalização dos dados	100
4.3	TREINAMENTO E MODELAGEM.....	101
4.3.1	Avaliação e métricas dos modelos	101
4.3.2	Otimização dos melhores modelos.....	111
4.3.3	Aplicação em perfis reais.....	116
5	CONCLUSÕES E RECOMENDAÇÕES	118
	REFERÊNCIAS.....	120

1 INTRODUÇÃO

O Diabetes Mellitus tipo 2 (DM2) consolidou-se como uma das doenças crônicas de maior impacto na saúde pública contemporânea, com uma prevalência crescente tanto em escala global quanto no cenário brasileiro. Caracterizada como uma desordem metabólica que leva à hiperglicemia persistente, a doença, quando não diagnosticada e tratada precocemente, acarreta um risco elevado de complicações severas e incapacitantes, como retinopatia, nefropatia e doenças cardiovasculares. Nesse contexto, a detecção precoce de indivíduos em risco emerge como uma estratégia fundamental, não apenas para a prevenção de desfechos clínicos adversos, mas também para a otimização de recursos e a redução de custos no sistema de saúde.

Diante deste desafio, a ciência de dados e o aprendizado de máquina (*machine learning*) apresentam-se como ferramentas poderosas, capazes de analisar grandes volumes de dados clínicos para identificar padrões complexos e predizer riscos de forma automatizada. Este trabalho aborda a predição de diabetes como um problema de classificação supervisionada. Essencialmente, o objetivo é treinar um algoritmo utilizando um conjunto de dados históricos que contém diversas variáveis preditoras (ou *features*), como níveis de glicose, índice de massa corporal (IMC), idade e histórico de gestações, associadas a um rótulo conhecido – neste caso, o diagnóstico de diabetes (positivo ou negativo). Uma vez treinado, o modelo deve ser capaz de estimar a probabilidade de um novo indivíduo, com base em suas características, pertencer à classe de risco para diabetes. Para este estudo de caso, foi utilizado o PIMA Indians Diabetes Dataset, um conjunto de dados canônico na literatura para esta finalidade.

Para tratar deste problema de forma robusta e metodologicamente rigorosa, o presente trabalho seguiu uma esteira estruturada de análise de dados. A jornada iniciou-se com uma Análise Exploratória de Dados (AED) aprofundada, visando compreender as particularidades, distribuições e correlações do dataset. Em seguida, foi executada uma fase crítica de pré-processamento, na qual foram investigadas e aplicadas técnicas para o tratamento de desafios como valores ausentes mascarados, remoção de *outliers* e o desbalanceamento entre as classes. Com os dados devidamente

preparados, procedeu-se ao treinamento e à avaliação comparativa de nove distintos algoritmos de aprendizado de máquina, com o intuito de identificar os de maior potencial preditivo.

Dessa forma, o que se buscou resolver com esta pesquisa foi além da simples construção de um classificador acurado. O objetivo central foi desenvolver e validar um pipeline completo que resultasse em um modelo preditivo não apenas performático sob a ótica das métricas quantitativas, mas também seguro, coerente e clinicamente plausível. Este trabalho, portanto, investiga como a combinação de análise estatística, técnicas de machine learning e uma validação qualitativa rigorosa pode gerar uma ferramenta de inteligência artificial responsável e verdadeiramente aplicável no apoio à tomada de decisão para a prevenção do diabetes

1.1 OBJETIVO

O objetivo geral deste trabalho é investigar a aplicação de técnicas de análise de dados e aprendizado de máquina na prevenção do diabetes tipo 2, por meio do desenvolvimento e avaliação de um modelo preditivo capaz de identificar, com base em dados clínicos, indivíduos com alto risco de desenvolver diabetes. Busca-se, com isso, demonstrar na prática como a ciência de dados pode auxiliar na detecção precoce e apoio à tomada de decisão em saúde, contribuindo para ações preventivas direcionadas, com os seguintes objetivos específicos:

1. Realizar uma revisão bibliográfica sobre fatores clínicos relacionados ao diabetes tipo 2 e técnicas de Machine Learning.
2. Realizar uma análise exploratória detalhada dos dados clínicos, identificando padrões, correlações e variáveis relevantes para o risco de diabetes tipo 2.
3. Executar o pré-processamento dos dados clínicos para tratamento, limpeza e preparação para análise.
4. Desenvolver e treinar modelos preditivos utilizando técnicas de Machine Learning para estimar o risco de diabetes tipo 2.
5. Avaliar e comparar o desempenho dos modelos, destacando os fatores clínicos com maior relevância para diagnóstico precoce.
6. Discutir os resultados obtidos quanto à aplicabilidade prática na prevenção do diabetes tipo 2 e no apoio à tomada de decisões em saúde.

1.2 JUSTIFICATIVA

O diabetes tipo 2 é uma doença crônica de prevalência crescente no Brasil, onde desafios socioeconômicos e de acesso à saúde dificultam a detecção precoce e a prevenção, tornando necessárias ferramentas mais precisas e acessíveis para identificar indivíduos em risco. Este trabalho propõe o estudo e análise de um modelo preditivo baseado em técnicas de análise de dados e aprendizado de máquina, utilizando dados clínicos para estimar o risco de diabetes, contribuindo tanto para a ciência de dados quanto para a prática em

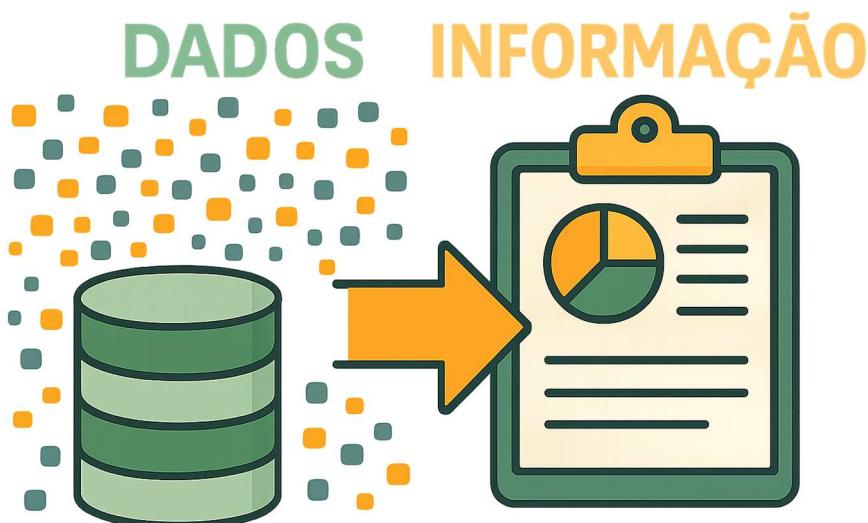
saúde. Por meio de revisão bibliográfica, pré-processamento de dados, treinamento e avaliação de modelos, além de uma análise exploratória detalhada, busca-se identificar fatores de risco relevantes e apoiar a tomada de decisão em saúde, possibilitando intervenções preventivas personalizadas que reduzam a incidência da doença e os custos associados, especialmente em um contexto de recursos limitados como o Brasil.

2 REVISÃO DA LITERATURA

2.1 UMA BREVE INTRODUÇÃO A CIÊNCIA DE DADOS E *MACHINE LEARNING*

A Ciência de Dados é uma área interdisciplinar que combina estatística, computação e conhecimento de domínio para extrair “*insights*” de grandes volumes de dados. A Ciência de Dados surgiu como resposta ao fenômeno do “*Big Data*”, caracterizado pelo alto volume, velocidade e variedade dos registros digitais que passam a ser gerados em praticamente todas as atividades humanas. Na prática, é transformar uma grande e complexa quantidade de dados em informação útil. Sua importância na era digital reside na possibilidade de transformar grandes volumes de dados em insights acionáveis, que podem auxiliar na tomada de decisões estratégicas em diversas áreas, incluindo a saúde e os estudos populacionais. A Figura 1 abaixo demonstra a transformação de dados em informação.

Figura 1 – Ilustração da transformação de dados em informação.



Fonte: elaborado pelo autor.

Ela abrange técnicas como análise estatística, visualização de dados e aprendizado de máquina, permitindo a transformação de grandes volumes de

dados em informações açãoáveis. Na era digital, caracterizada pela explosão de dados gerados por dispositivos, sistemas online e sensores, a Ciência de Dados é fundamental para impulsionar a inovação e a competitividade em setores como saúde, finanças, marketing e manufatura. Por exemplo, na saúde, ela permite identificar padrões em dados clínicos para prever doenças, enquanto em negócios, auxilia na criação de campanhas de marketing direcionadas e na otimização de processos conforme GeeksforGeeks (2025).

Dentro do vasto campo da Ciência de Dados, uma das ferramentas mais poderosas para extrair insights de grandes volumes de dados é o *Machine Learning (ML)*. Este subcampo da Inteligência Artificial (*IA*) concentra-se no desenvolvimento de algoritmos que permitem aos computadores aprenderem com dados e melhorar seu desempenho sem programação explícita. Como por exemplo: ele utiliza dados históricos para identificar padrões e fazer previsões ou classificações, como determinar a probabilidade de um paciente desenvolver diabetes com base em características como peso, glicemia e IMC, posteriormente discutido neste trabalho por este autor. Na Ciência de Dados, o *Machine Learning* é uma ferramenta essencial, pois fornece os métodos computacionais para construir modelos preditivos e analíticos que complementam outras técnicas, como estatística descritiva e visualização de dados. **Enquanto a Ciência de Dados abrange o processo completo de coleta, limpeza, análise e interpretação de dados, o ML é responsável por automatizar a descoberta de padrões e a geração de previsões**, sendo particularmente valioso em tarefas complexas que exigem análise de grandes conjuntos de dados. Essa integração torna o *Machine Learning* indispensável para aplicações preditivas, como as abordadas neste trabalho, que visa explorar a previsão de diabetes com base em dados clínicos.

Um dos conceitos clássicos em ML, apresentado por Mitchell (1997), define que:

Definition: A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P** if its performance at tasks in **T**, as measured by **P**, improves with experience **E**.

Traduzindo, um sistema de aprendizado melhora com experiência (E) em relação a um conjunto de tarefas (T) e uma medida de desempenho P, se sua performance nas tarefas (T), medida por (P), melhora com (E), a Tabela 1 abaixo pode ajudar a compreender melhor o entendimento de Mitchell (1997).

Tabela 1 – Exemplo concreto da definição de Mitchell.

Símbolo	Significado no exemplo de filtro de spam	Instância concreta
T(Task)	A tarefa que o sistema deve executar	Classificar cada e-mail que chega como “spam” ou “não spam”.
P(Performance)	A régua que mede quão bem o sistema cumpre T	Acurácia (ou precisão/recall): porcentagem de e-mails corretamente classificados.
E (Experience)	A fonte de “experiência” usada para aprender	Conjunto de milhares/milhões de e-mails rotulados anteriormente por humanos como spam ou legítimos.

Fonte: elaborado pelo autor com o conceito de Mitchell (1997).

De forma equivalente, pode-se dizer que a aprendizagem de máquina envolve algoritmos computacionais que aprimoram seu desempenho automaticamente, à medida que “aprendem” padrões presentes nos dados. Diferentemente de um programa tradicional, no qual um desenvolvedor codifica explicitamente as regras para resolver um problema, em *Machine Learning* o objetivo é que o próprio algoritmo descubra essas regras ou padrões internos, inferindo generalizações a partir de exemplos fornecidos.

Em termos formais, um algoritmo de ML busca otimizar seu comportamento em determinada tarefa conforme é exposto a mais dados (*experiência*), ajustando seus modelos internos de modo a produzir predições ou decisões cada vez mais acuradas.

Embora o termo *machine learning* tenha se popularizado apenas na última década, seu desenvolvimento acompanha a própria história da inteligência artificial. A Tabela 2 resume, em ordem cronológica, os marcos que impulsionaram essa trajetória — do *Teste de Turing* (1950) e do primeiro programa que “aprendeu” a jogar damas (1952) até a ascensão dos modelos fundacionais capazes de raciocinar em múltiplas etapas (2025). Esses marcos

mostram como cada salto teórico se apoia em avanços simultâneos de hardware (GPUs) e em uma disponibilidade crescente de dados, criando ciclos virtuosos de inovação. Também evidenciam a alternância entre períodos de grande entusiasmo, seguidos por fases de refinamento em que as limitações descobertas guiam pesquisas mais profundas. Dessa forma, observar a evolução do *machine learning* não é apenas recuperar datas históricas, mas entender como ciência, indústria e a necessidade da sociedade caminham juntas para transformar descobertas em tecnologias aplicáveis no cotidiano.

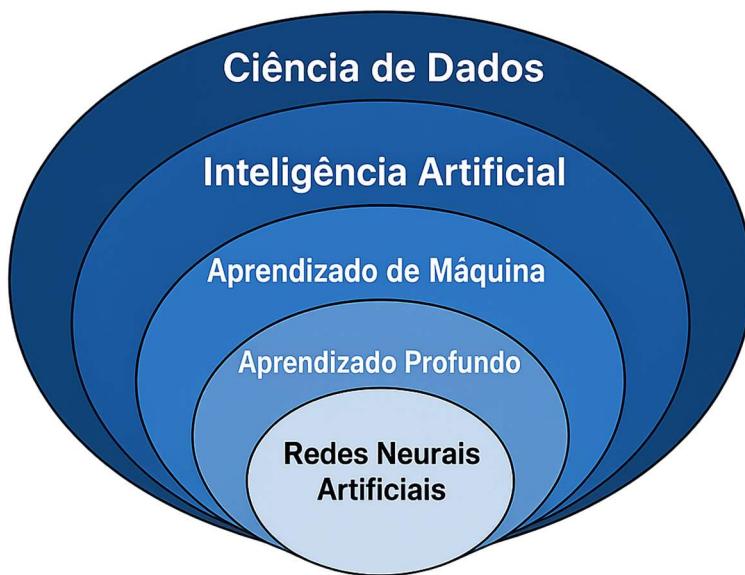
Tabela 2 – Datas e marcos importantes do ML ao longo da história.

Data	Acontecimento	Autor(es)
1950	Artigo Computing Machinery and Intelligence → proposta do Teste de Turing, marco conceitual da IA.	Alan M. Turing
1952	Primeiro programa de aprendizado de máquina (jogo de damas) → cunhado o termo “Machine Learning”.	Arthur L. Samuel
1956	Conferência de Dartmouth formaliza o campo da Inteligência Artificial.	John McCarthy, Marvin Minsky, Claude Shannon, Nathan Rochester
2012	Rede AlexNet vence a ImageNet, inaugurando a era do Deep Learning moderno.	Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton
2020	Lançamento do GPT-3, consolidando foundation models de linguagem.	OpenAI
2022-23	Popularização massiva da IA Generativa com ChatGPT (texto) e DALL·E (imagens).	OpenAI
2025	Claude 4 (Opus/Sonnet): novo modelo da Anthropic com raciocínio prolongado (capaz de manter foco por horas em tarefas de codificação e análises complexas).	Anthropic
2025	GPT o3: modelo OpenAI otimizado para “simulated reasoning” — o sistema executa múltiplos ciclos internos de reflexão antes de responder, melhorando cadeias longas de raciocínio.	OpenAI

Fonte: Adaptado pelo autor a partir de He et al. (2025); OpenAI (2018, 2020); Anthropic (2025).

A Figura 2 a seguir ilustra a relação hierárquica entre os principais conceitos envolvidos no campo do aprendizado de máquina. No centro estão as Redes Neurais Artificiais (ANN), base dos modelos de Aprendizado Profundo (*Deep Learning – DL*), que por sua vez compõem um subconjunto do Aprendizado de Máquina (Machine Learning – ML). O *Machine Learning* inclui diversos modelos e métodos, sendo parte integrante da Inteligência Artificial (IA), que por sua vez está inserida no domínio mais amplo da Ciência de Dados. Essa estrutura representa o "guarda-chuva" de técnicas seletivas da ciência de dados, onde a IA abrange desde a programação clássica até métodos baseados em aprendizado. A representação evidencia como cada camada contribui para a evolução de sistemas inteligentes orientados por dados.

Figura 2 – Guarda-chuvas da ciência de dados.

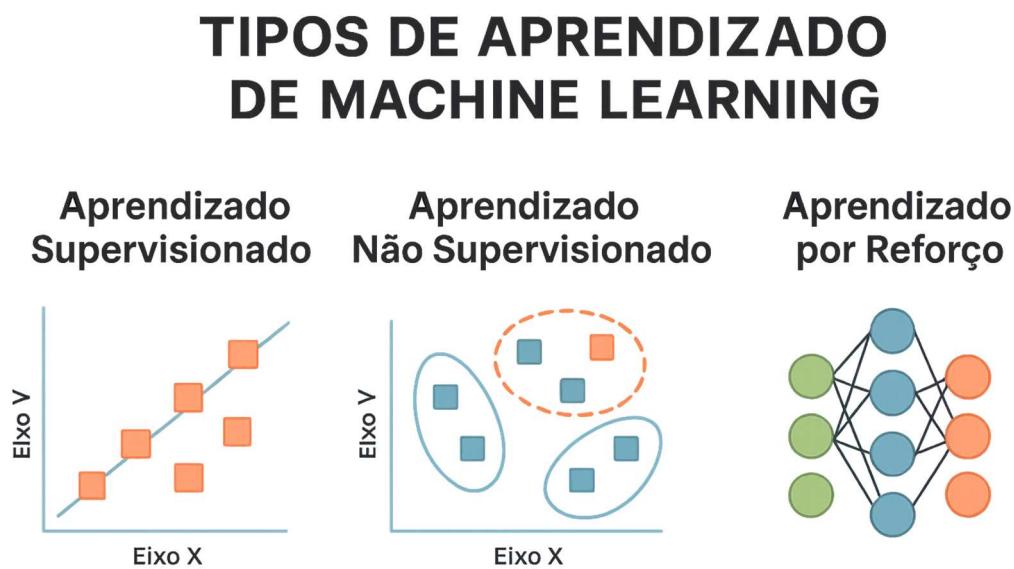


Fonte: adaptador pelo autor a partir de Choi et al. (2020).

2.2 TIPOS DE MACHINE LEARNING

Os principais tipos de Machine Learning discutidos nesta revisão bibliográfica são: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço. A Figura 3 abaixo ilustra brevemente a maneira na qual cada um trabalha.

Figura 3 – Principais de aprendizado de Machine Learning.



Fonte: Adaptado de Mitchell (1997)

Segundo Jordan e Mitchell (2015), o tipo que é vastamente utilizado é o supervisionado, como por ex: classificadores de spam, reconhecimento de imagem, entre outros. O Quadro 1 abaixo demonstra algumas utilidades de cada tipo de aprendizado.

Quadro 1 – Tipos de Machine Learning e suas utilidades.

Tipo	Aplicações
Aprendizado Supervisionado	Detecção de fraudes, previsão de preços, previsão de diabetes
Aprendizado Não Supervisionado	Segmentação de clientes, detecção de anomalias, compressão de imagens
Aprendizado por Reforço	Jogos, robótica, sistemas de controle

Fonte: Adaptado pelo autor a partir de Jordan e Mitchell (2015).

2.2.1 Aprendizado supervisionado

O aprendizado supervisionado utiliza conjuntos de dados rotulados, onde cada exemplo de treinamento contém uma entrada (características) e uma saída desejada (rótulo). O objetivo é aprender uma função que mapeie entradas para saídas, permitindo previsões em novos dados. Este processo de aprendizado envolve, tipicamente, a divisão do conjunto de dados original em subconjuntos de treinamento, validação e teste (CHOI et al., 2020). O modelo utiliza os dados de treinamento para inferir um algoritmo a partir dos pares de características e rótulos, sendo continuamente informado, através do rótulo, se suas previsões estão corretas. A performance do algoritmo é então ajustada e validada com o conjunto de validação e, finalmente, avaliada de forma rigorosa no conjunto de teste, que consiste em dados não vistos anteriormente pelo modelo, garantindo uma estimativa mais realista de sua capacidade de generalização.

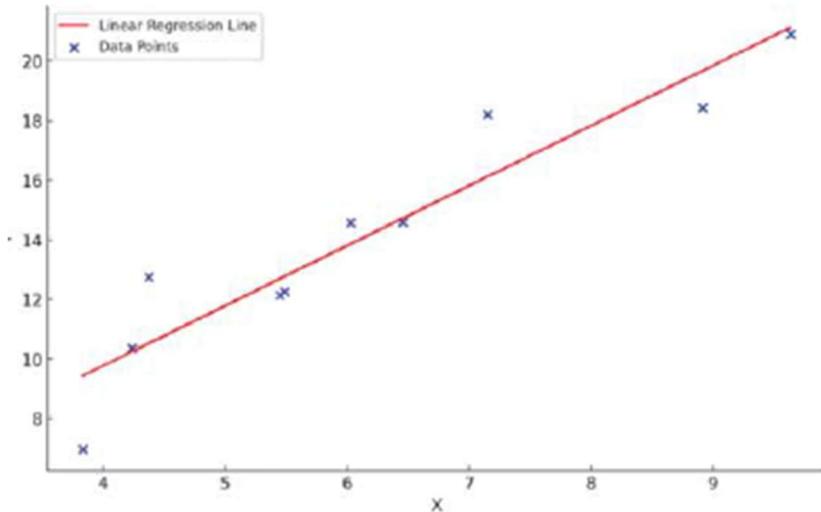
Este tipo de aprendizado é dividido em duas categorias principais: classificação, que prevê categorias discretas (ex.: diabético ou não diabético), e regressão, que prevê valores contínuos (ex.: nível de glicose no sangue). Algoritmos comuns incluem regressão linear, regressão logística, árvores de decisão, random forest, máquinas de vetores de suporte (SVM) e redes neurais artificiais. Na saúde, o aprendizado supervisionado é utilizado para tarefas preditivas, como a previsão de diabetes, onde modelos são treinados com dados clínicos rotulados (ex.: peso, glicemia, IMC) para classificar pacientes. Por exemplo, um modelo pode identificar padrões em dados de pacientes diagnosticados, permitindo prever o risco de diabetes em novos indivíduos com base em características semelhantes. Sua principal vantagem é a precisão quando há dados rotulados suficientes e de boa qualidade, mas requer grandes quantidades de dados anotados, o que pode ser um processo custoso e demorado, especialmente em aplicações médicas complexas.

I. Modelos lineares

Os modelos lineares são aqueles que procuram estabelecer uma relação de proporcionalidade direta ou linear entre as variáveis de entrada (features) e a variável de saída. A seguir, os dois principais modelos de acordo com a bibliografia

- **Regressão linear:** trata-se de um algoritmo basilar que visa modelar uma relação linear entre um conjunto de variáveis de entrada e uma variável de saída de natureza numérica contínua (CHOI et al., 2020). É frequentemente empregue na previsão de valores quantitativos, como preços de imóveis ou projeções de vendas. Dada a sua simplicidade, Cordella et al. (2024) observam que o seu desempenho pode ser limitado ao lidar com conjuntos de dados de maior complexidade, servindo muitas vezes como um ponto de partida ou linha de base para comparação com algoritmos mais sofisticados.
- Como por exemplo, a Figura 4 abaixo

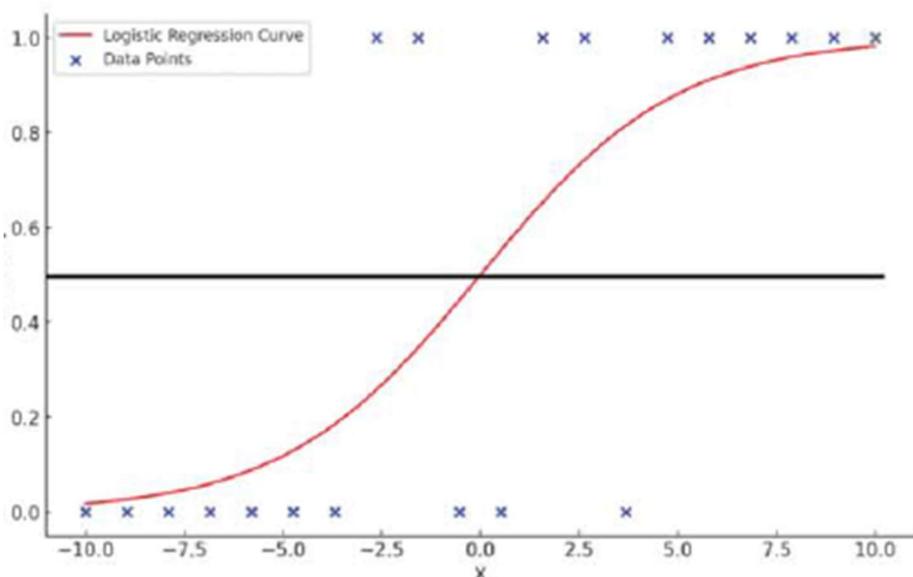
Figura 4 – Regressão linear.



Fonte: Cordella et al. (2024)

- **Regressão logística:** De forma similar à regressão linear em termos de simplicidade, a Regressão Logística é utilizada para modelar a relação entre variáveis de entrada e uma variável de saída categórica, usualmente binária (por exemplo, indicando presença/ausência, sim/não, ou classes como 0 e 1). É comumente aplicada em problemas como a avaliação de risco de crédito ou o auxílio em diagnósticos médicos. Apesar do termo "regressão" na sua designação, é fundamentalmente um algoritmo de classificação que estima a probabilidade de uma instância pertencer a uma determinada classe. Diferentemente da regressão linear, a regressão logística emprega uma curva sigmoidal para estimar a probabilidade da classe, em vez de uma linha reta. Esta curva é resultado da aplicação da função sigmoide, que converte as características de entrada (discretas ou contínuas) em um valor de probabilidade entre 0 e 1, garantindo que as previsões de probabilidade permaneçam dentro desses limites. Para a classificação binária, onde os rótulos são tipicamente 0 e 1, o algoritmo calcula uma pontuação de probabilidade para cada exemplo pertencer à classe 1. Usualmente, um limiar de 0,5 é aplicado: se a pontuação prevista for menor que 0,5, o exemplo é classificado como 0, caso contrário, como 1. Além disso, o método pode ser adaptado para problemas multinominais, lidando com três ou mais classes possíveis (CHOI et al., 2020; CORDELLA et al., 2024).

Figura 5 – Regressão logística.



Fonte: Cordella et al. (2024)

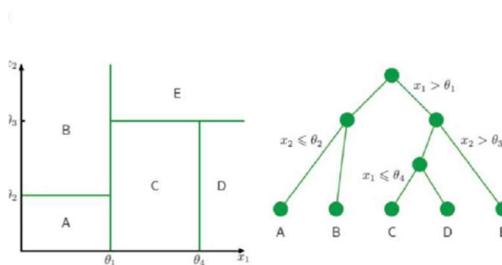
- **Máquina de Vetores de Suporte – SVM (kernel linear):** classifica os dados traçando o hiperplano de maior margem possível entre as classes. Quando se usa o kernel linear, a fronteira de decisão continua sendo um hiperplano — portanto, o modelo permanece na categoria dos lineares; kernels não lineares (RBF, polinomial) deslocam o algoritmo para a família “métodos de margem com kernel” (AKMEŞE, 2022).

II. Modelos baseados em árvores

Os modelos baseados em árvore utilizam uma estrutura hierárquica de decisões, semelhante a um fluxograma, para realizar previsões a partir dos dados de entrada. Estes modelos são capazes de aprender regras complexas e são utilizados tanto para tarefas de classificação quanto de regressão.

- **Árvore de Decisão (Decision Tree):** Uma Árvore de Decisão de acordo com Choi et al., (2020) e Cordella et al., (2024). é uma estrutura semelhante a um fluxograma onde cada nó interno representa um teste em uma *feature* (atributo), cada ramo representa o resultado do teste, e cada nó folha representa uma classe (em problemas de classificação) ou um valor numérico (em problemas de regressão). O modelo aprende a tomar decisões sequenciais sobre as *features* para produzir as previsões. Por exemplo, para cada instância de dados, o modelo percorre a árvore a partir do nó raiz, aplicando os testes das *features* até alcançar um nó folha que determina a previsão.

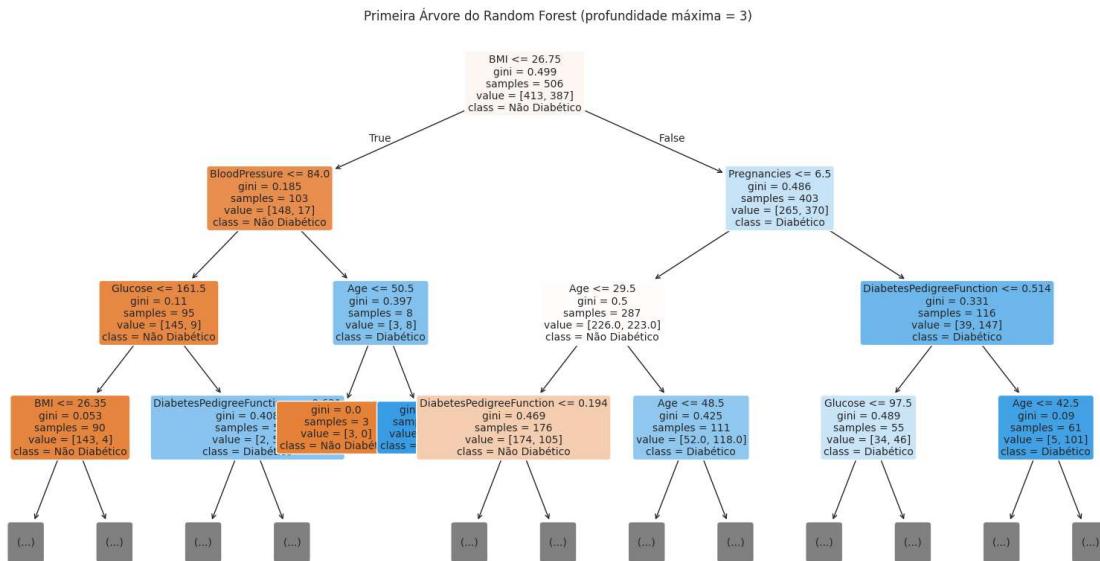
Figura 6 – Árvore de decisão.



Fonte: (Cordella et al., 2024)

- **Florestas Aleatórias (Random Forest):** As Florestas Aleatórias são um método de aprendizado de conjunto (*ensemble learning*) que consiste na construção de múltiplas Árvores de Decisão durante o treinamento. Em vez de depender de uma única árvore, a Random Forest agrupa as previsões de diversas árvores. Para tarefas de classificação, o resultado final é tipicamente a classe mais votada (moda) entre todas as árvores componentes. Esta abordagem de combinação de múltiplos aprendizes visa aprimorar a capacidade preditiva do modelo, segundo Choi et al., (2020) e Cordella et al., (2024).

Figura 7 – Random Forest.



Fonte: elaborado pelo autor.

- **Gradient Boosting (GBM):** cria sucessivamente pequenas árvores, cada uma focada em corrigir os erros da anterior; o resultado final é a soma ponderada dessas árvores fracas, produzindo um “comitê” de alto desempenho (AKMEŞE, 2022).
- **XGBoost (Extreme Gradient Boosting):** variação altamente otimizada do Gradient Boosting; explora regularização interna, paralelização e *shrinkage* para ganhar velocidade e reduzir *overfitting* (AKMEŞE, 2022).

- **LightGBM:** também baseada em boosting de árvores, mas usa particionamento folha-a-folha e histograms para acelerar o treinamento e consumir menos memória, mantendo alta acurácia (AKMEŞE, 2022).
- **AdaBoost:** combina muitos *decision stumps* (árvores rasas) treinados em sequência, ajustando pesos das amostras a cada iteração para concentrar esforço nos exemplos mais difíceis; é considerado um ensemble de árvores quando o *base estimator* são stumps (AKMEŞE, 2022).

III. Outros métodos de classificação

Diferentemente dos modelos lineares ou baseados em árvores, estes algoritmos representam paradigmas alternativos de aprendizado supervisionado e costumam ser incluídos como benchmarks em experimentos de machine learning (AKMEŞE, 2022).

- **k-Nearest Neighbors (kNN)** – *instance-based / lazy learner, não paramétrico, discriminativo.*

Classifica cada amostra de teste pela votação majoritária dos k vizinhos mais próximos, calculados via distâncias Euclidiana, Manhattan, Minkowski ou Hamming. O “modelo” é, na prática, todo o conjunto de treino: não há fase de ajuste de parâmetros (lazy learning), e cada consulta exige a busca dos vizinhos no espaço de atributos.

- **Naive Bayes (Gaussian NB)** – *classificador probabilístico gerativo bayesiano, paramétrico, eager learner.*

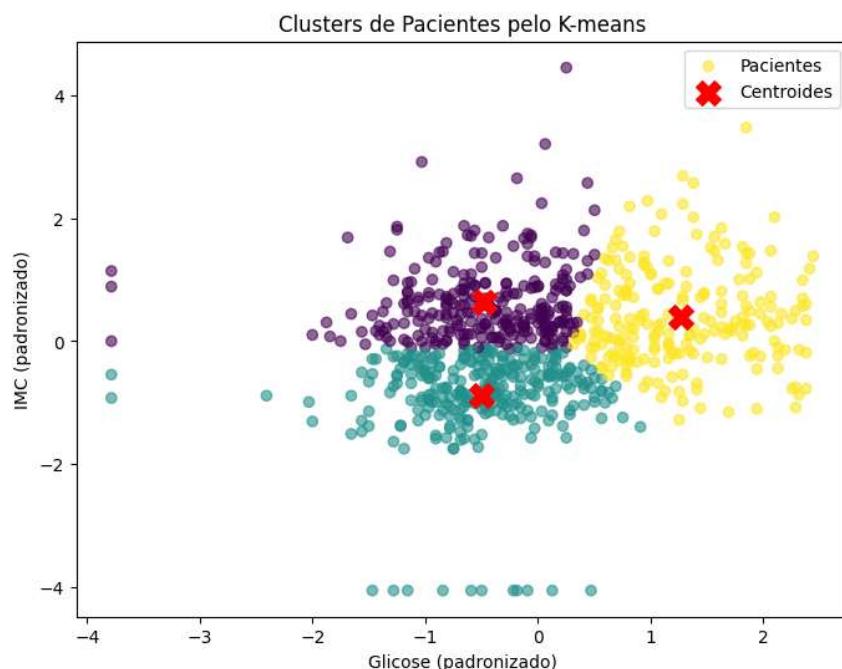
Baseia-se no Teorema de Bayes e assume independência condicional entre as variáveis de entrada dadas as classes (hipótese “naive”). Aprende, de forma paramétrica, as médias e variâncias gaussianas de cada feature por classe e, na predição, devolve a classe de maior probabilidade. Como ajusta esses parâmetros antes de qualquer consulta, é considerado um *eager learner*.

2.2.2 Aprendizado não supervisionado

Contrastando com o aprendizado supervisionado. A principal distinção reside nos dados de entrada: enquanto o aprendizado supervisionado trabalha com dados rotulados (pares de entrada e saída esperada), o aprendizado não supervisionado lida com dados não rotulados. Em outras palavras, os algoritmos de aprendizado não supervisionado são treinados em conjuntos de dados que contêm muitas *features* (características), mas não possuem um rótulo ou *target* associado para cada exemplo.

O objetivo do aprendizado não supervisionado não é prever um resultado *a priori* conhecido, mas sim descobrir propriedades estruturais ou encontrar padrões e similaridades nos dados. Isso é particularmente útil para analisar grandes conjuntos de dados complexos, ruidosos e volumosos onde a análise manual seria inviável e onde não se pressupõem hipóteses prévias específicas. A grande vantagem do aprendizado não supervisionado é que ele não requer a grande quantidade de trabalho humano e conhecimento especializado necessário para gerar dados rotulados, um dos principais modelo é o K-means (CHOI et al., 2020; CORDELLA et al., 2024).

Figura 8 – K-Means.



Fonte: elaborado pelo autor.

O principal modelo mais utilizado de acordo com Cordella et al. (2024) é o “*K-means*”. A clusterização “*K-means*” é um proeminente algoritmo de aprendizado não supervisionado para agrupamento, que partitiona dados em ‘*k*’ clusters distintos e não sobrepostos com base na similaridade. Seu mecanismo central envolve a predefinição do número ‘*k*’ de clusters, seguida pela busca iterativa por centroides (usualmente via distância euclidiana) e a atribuição de cada ponto de dado ao centroide mais próximo. Devido à sua simplicidade, interpretabilidade e capacidade de escalar para grandes volumes de dados, resultando em agrupamentos coesos, o “*K-means*” é amplamente aplicado em tarefas como segmentação de clientes e em sistemas de recomendação. Contudo, uma limitação crucial é a exigência da definição prévia do valor de ‘*k*’, além de apresentar dificuldades em lidar eficazmente com clusters que possuem tamanhos e densidades muito variados.

2.2.3 Aprendizado por reforço

De acordo com Jordan e Mitchell (2015) no aprendizado por reforço, as informações disponíveis nos dados de treinamento estão em um nível intermediário entre o aprendizado supervisionado e o não supervisionado. Em vez de exemplos de treinamento que indicam a saída correta para uma dada entrada, os dados de treinamento no aprendizado por reforço fornecem apenas uma indicação se uma ação está correta ou não. Se uma ação estiver incorreta, ainda há o problema de encontrar a ação correta.

Geralmente, no aprendizado por reforço, sinais de recompensa referem-se a sequências inteiras de entradas. A atribuição de crédito ou culpa a ações individuais na sequência não é fornecida diretamente.

O aprendizado por reforço tipicamente envolve um ambiente de controle onde a tarefa de aprendizado é aprender uma estratégia de controle (uma “política”) para um agente que atua em um ambiente dinâmico desconhecido. Essa estratégia aprendida é treinada para escolher ações para qualquer estado dado, com o objetivo de maximizar sua recompensa esperada ao longo do tempo.

Segundo Choi et al., (2020) o exemplo do jogo Super Mario Bros para ilustrar o aprendizado por reforço é uma forma fácil de compreender o seu funcionamento. O objetivo do jogo é mover o personagem Mario do lado esquerdo para o direito da tela para alcançar o mastro da bandeira no final de cada nível, evitando perigos como inimigos e poços. Não há uma sequência correta de comandos; existem sequências que levam a uma vitória e outras que não. Em aprendizado por reforço, um algoritmo "jogaria" sozinho, tentando diferentes entradas. Quando Mario se move para frente sem sofrer dano, o algoritmo é "recompensado" (ou seja, o comportamento é reforçado). Através desse processo, o algoritmo aprende qual comportamento é desejado (por exemplo, mover-se para frente é melhor do que para trás, pular sobre inimigos é melhor do que correr para eles). Eventualmente, o algoritmo aprende a se mover do início ao fim.

Figura 9 – Aprendizado de reforço.



Fonte: simplificado e elaborado pelo autor, adaptado de Choi et al., (2020).

2.2.4 Deep learning (aprendizado profundo)

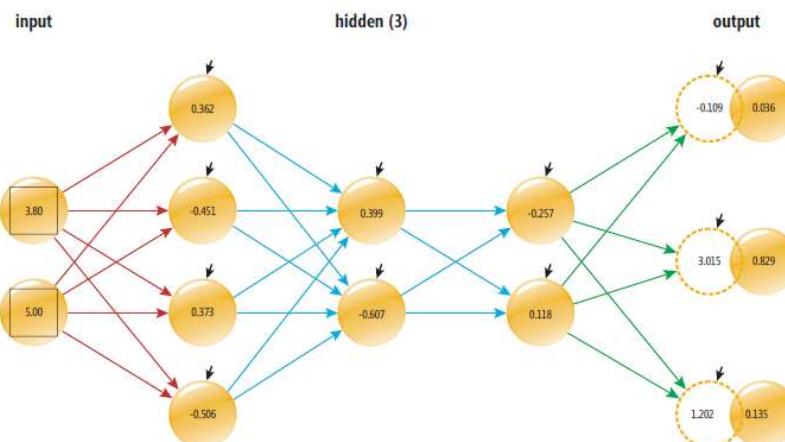
O aprendizado profundo (deep learning - DL) é um método específico para alcançar um objetivo da inteligência artificial, assim como o *Machine Learning* (ML). De fato, o DL está contido dentro do ML, que por sua vez está dentro da IA (CHOIT et al., 2020).

Os métodos de aprendizado profundo, juntamente com outros métodos de ML, se destacam em tarefas complicadas que exigem reconhecimento de padrões de alto nível, como reconhecimento de fala ou classificação de imagens, onde abordagens mais tradicionais de IA podem falhar.

As redes neurais profundas (DNNs), um tipo específico de modelo de aprendizado profundo mencionado é a rede neural profunda (Deep Neural Network - DNN) são difíceis de interpretar examinando o próprio modelo.

De acordo com Jordan e Mitchell (2015) as camadas internas das redes profundas podem ser vistas como fornecendo representações aprendidas dos dados de entrada. Embora muito do sucesso prático no aprendizado profundo venha de métodos de aprendizado supervisionado para descobrir essas representações, também existem esforços para desenvolver algoritmos de aprendizado profundo que descubram representações úteis sem a necessidade de dados de treinamento rotulados (aprendizado não supervisionado).

Figura 10 – Um exemplo de rede neural profunda 2-(4,2,2)-3.



Fonte: (STANEK, 2017)

2.3 MACHINE LEARNING NA SAÚDE

O *Machine Learning* (ML) tem se destacado como uma ferramenta poderosa na área da saúde, com aplicações que abrangem desde o diagnóstico até o tratamento personalizado de pacientes. Este subcapítulo apresenta uma visão geral das aplicações do ML na saúde, exemplos específicos de seu uso, os benefícios que ele traz e os desafios associados à sua implementação.

I. Visão geral das aplicações de *Machine Learning* na área da saúde

O ML é utilizado em diversas áreas da saúde para analisar grandes volumes de dados e extrair padrões que auxiliam na tomada de decisões clínicas. De acordo com Paixão et al. (2022), o ML é aplicado em cardiologia para predizer riscos cardíacos, na dermatologia para identificar câncer de pele, e na psiquiatria para diagnosticar transtornos mentais.

Além disso, Fonseca et al. (2024) destacam que “a inteligência artificial proporciona análises objetivas de exames de imagem, modificando fluxos de trabalho e aumentando a eficiência”, como em ressonâncias magnéticas e tomografias computadorizadas.

II. Exemplos de uso em diagnósticos, previsões e tratamentos personalizados

Um exemplo notável é o emprego do ML na dermatologia, onde algoritmos distinguem lesões cutâneas benignas de malignas com precisão comparável à de dermatologistas (PAIXÃO et al., (2022)).

Na psiquiatria, o ML reduziu os critérios diagnósticos para o transtorno do espectro autista, mantendo 100% de acurácia em um estudo com 612 pacientes (PAIXÃO et al., (2022)).

De acordo com Fonseca et al. (2024), no campo da imagem médica, o ML tem sido instrumental na detecção precoce de doenças, como o câncer, contribuindo para diagnósticos mais rápidos e precisos.

III. Benefícios do *Machine Learning* na saúde

Os benefícios do ML na saúde são significativos. Ele melhora a precisão dos diagnósticos ao identificar padrões imperceptíveis para profissionais de saúde. O ML também agiliza o processamento de informações, permitindo decisões mais rápidas. A capacidade de personalizar tratamentos com base nos dados de cada paciente promove uma medicina centrada no indivíduo, conforme Paixão et al., (2022).

IV. Desafios do *Machine Learning* na saúde

O autor Paixão et al., (2022) ressalta e conclui, apesar dos benefícios, o ML enfrenta desafios importantes. A necessidade de grandes conjuntos de dados para treinar algoritmos é um obstáculo em áreas com dados escassos. A interpretação dos modelos, frequentemente vistos como “caixas pretas”, dificulta a compreensão das decisões. Além disso, vieses nos dados podem levar a resultados injustos, comprometendo a equidade no atendimento.

Em conclusão, o Machine Learning representa uma revolução na saúde, oferecendo ferramentas para diagnóstico, previsão e tratamento personalizado. Contudo, desafios como a disponibilidade de dados e a interpretabilidade dos modelos exigem atenção para maximizar benefícios e minimizar riscos.

2.4 DIABETES E SUA PREVISÃO

2.4.1 Definição abrangente

O Diabetes Mellitus (DM) é reconhecido como uma doença metabólica crônica, cuja principal característica é a presença de níveis persistentemente elevados de glicose na corrente sanguínea, uma condição denominada hiperglicemia. Esta alteração metabólica fundamental decorre de deficiências na secreção e/ou na ação da insulina, um hormônio peptídico produzido pelas células beta das ilhotas pancreáticas, que desempenha um papel crucial na regulação do metabolismo da glicose e de outros nutrientes no organismo. A insulina facilita a captação de glicose pelas células para produção de energia e armazenamento, mantendo assim a homeostase glicêmica (SBD, 2024).

A manutenção prolongada de níveis glicêmicos elevados, ou hiperglicemia crônica, está intrinsecamente associada ao desenvolvimento de danos a longo prazo, disfunção progressiva e, eventualmente, falência de múltiplos órgãos e sistemas. Os órgãos mais vulneravelmente afetados incluem os olhos (retinopatia diabética), rins (nefropatia diabética), nervos periféricos e autonômicos (neuropatia diabética), coração e vasos sanguíneos (doença cardiovascular aterosclerótica e microvascular). A compreensão desta patologia complexa é vital, dada a sua crescente prevalência e o impacto significativo na saúde pública global (SBD, 2024).

2.4.2 Classificação e tipos

A classificação etiológica do DM é essencial para a compreensão da sua patogênese, para a estratificação de risco e para a definição de abordagens terapêuticas individualizadas. Embora existam diversas formas de diabetes, os tipos mais prevalentes e clinicamente significativos são o Diabetes Mellitus Tipo 1 (DM1), o Diabetes Mellitus Tipo 2 (DM2) e o Diabetes Mellitus Gestacional (DMG).

I. Diabetes Mellitus Tipo 1 (DM1)

De acordo com a ADA (2024), a Diabetes Mellitus tipo 1 (DM1) é uma condição crônica caracterizada pela destruição autoimune das células beta do pâncreas, responsáveis pela produção de insulina, resultando em uma deficiência total desse hormônio. **Ele representa cerca de 5 a 10% dos casos de diabetes** e afeta aproximadamente 600 mil pessoas no Brasil (SBD, 2024). A doença tem causas multifatoriais, envolvendo uma predisposição genética ligada a certos alelos do sistema HLA (Antígeno Leucocitário Humano) e fatores ambientais que desencadeiam a resposta autoimune em indivíduos suscetíveis. Até o momento, não há formas conhecidas de prevenir o aparecimento do DM1.

Os sintomas do DM1 surgem de maneira repentina e incluem micção excessiva, sede intensa, fome aumentada, perda de peso sem explicação, alterações na visão e cansaço extremo, podendo evoluir rapidamente para uma complicaçāo grave chamada cetoacidose diabética se não forem tratados a tempo. O tratamento do DM1 é baseado no uso contínuo de insulina administrada externamente para substituir a produção ausente, aliado a um plano alimentar personalizado e à prática regular de exercícios físicos, que ajudam a controlar os níveis de glicose no sangue e a manter a saúde geral do paciente (ADA, 2024).

II. Diabetes Mellitus Tipo 2 (DM2)

O DM2 é a forma mais prevalente de diabetes, respondendo por 90-95% de todos os casos diagnosticados globalmente (ADA, 2024). Esta condição é caracterizada por um defeito duplo: resistência à ação da insulina nos tecidos periféricos (principalmente músculo, fígado e tecido adiposo) e uma deficiência relativa na secreção de insulina pelas células beta pancreáticas, que se torna progressivamente mais severa ao longo do tempo.

O desenvolvimento do DM2 está fortemente associado a uma constelação de fatores de risco, muitos dos quais são modificáveis. Estes incluem sobrepeso e obesidade (especialmente a obesidade central), um padrão dietético não saudável (rico em calorias, gorduras saturadas e açúcares refinados, e pobre em

fibras), sedentarismo e uma forte predisposição genética, frequentemente evidenciada por histórico familiar da doença (ADA, 2024).

Os sintomas do DM2 podem ser semelhantes aos do DM1, mas são frequentemente menos pronunciados e podem desenvolver-se de forma insidiosa ao longo de muitos anos. Esta natureza gradual pode levar a um diagnóstico tardio, muitas vezes ocorrendo quando as complicações crônicas já estão presentes. A apresentação clínica do DM2 é heterogênea, refletindo a variabilidade na predominância da resistência insulínica versus a deficiência secretória de insulina. Embora tradicionalmente considerado uma doença de adultos, a incidência de DM2 tem aumentado alarmantemente em crianças e adolescentes, um fenômeno diretamente correlacionado com o aumento da prevalência de obesidade infantil.

O manejo do DM2 é multifacetado, iniciando-se com modificações intensivas no estilo de vida, que incluem a adoção de um plano alimentar saudável, a prática regular de atividade física e a perda de peso, se indicada. Frequentemente, estas medidas necessitam ser complementadas por agentes farmacológicos orais e/ou injetáveis (não insulínicos). Com a progressão da doença e o declínio da função das células beta, muitos pacientes com DM2 eventualmente necessitarão de insulinoterapia para manter o controle glicêmico adequado.

III. Diabetes Mellitus Gestacional (DMG)

O Diabetes Mellitus Gestacional (DMG) é definido como qualquer grau de intolerância à glicose com início ou primeiro reconhecimento durante a gestação. Caracteriza-se por níveis de glicose sanguínea elevados acima dos valores considerados normais para a gravidez, mas que não atingem os critérios diagnósticos para diabetes mellitus franco preexistente. Segundo Bazon e Pereira (2023), cerca de 6% das mulheres em gestação possuem essa condição.

Mulheres que desenvolvem DMG apresentam um risco aumentado de complicações durante a gravidez e o parto, incluindo macrossomia fetal, distocia de ombro, pré-eclâmpsia e necessidade de cesariana. Além disso, tanto a mãe quanto o conceito têm um risco significativamente maior de desenvolver DM2.

2.4.3 Fatores de risco e importância da detecção precoce

O Diabetes Mellitus Tipo 2 (DM2) é uma condição multifatorial, resultante da interação entre suscetibilidade genética e uma miríade de fatores ambientais e comportamentais. Os fatores de risco para o desenvolvimento do Diabetes Mellitus Tipo 2 (DM2) podem ser agrupados em categorias que auxiliam na compreensão e abordagem preventiva da doença. Esses incluem fatores não modificáveis, que são intrínsecos ao indivíduo; fatores modificáveis, relacionados ao estilo de vida e passíveis de intervenção; e condições médicas preexistentes, que aumentam a suscetibilidade ao DM2. De acordo com a ADA (2024) e SBD (2024), os principais fatores de risco estão na Figura 11 abaixo, e descritos com mais detalhes em suas categorias na sequência.

Figura 11 - Principais fatores de risco.

PRINCIPAIS FATORES DE RISCO PARA DIABETES TIPO 2

NAO MODIFICAVEIS	MODIFICÁVEIS	CONDICOES MEDICAS ASSOCIADAS/PREEXISTENTES
<ul style="list-style-type: none"> ▸ Idade ≤ 35-45 anos ▸ Histórico familiar de DM (1º grau) ▸ Etnia (africana, hispânica, Indígena asiática) ▸ Histórico pessoal de DMG 	<ul style="list-style-type: none"> ▸ Sobre peso/Obesidade (IMC ≥ 25 kg/m²) ▸ Sedentarismo ▸ Dieta não saudável ▸ Tabagismo 	<ul style="list-style-type: none"> ▸ Pré diabetes (G/A ou TGD) ▸ Hipertensão arterial sistêmica ≥ 130/90 mmHg ▸ Dislipíderma (HDL baixo e/ou triglicerídeos altos) ▸ Síndrome dos ovários poliquísticos (SOP) ▸ Acanthosis nigricans ▸ Doença cardiorvascular estabelecida ▸ Apneia obstrutiva do sono ▸ Infecção por HIV/AIDS

Fonte: adaptador pelo autor a partir de SBD (2024).

Os fatores de risco para o desenvolvimento do Diabetes Mellitus Tipo 2 (DM2) podem ser agrupados em categorias que auxiliam na compreensão e abordagem preventiva da doença. Esses incluem fatores não modificáveis, que são intrínsecos ao indivíduo; fatores modificáveis, relacionados ao estilo de vida e passíveis de intervenção; e condições médicas preexistentes, que aumentam a suscetibilidade ao DM2.

- **Fatores de Risco Não Modificáveis:** Estes são aspectos sobre os quais não se pode intervir diretamente, mas que elevam a predisposição ao DM2. Incluem o avanço da idade (com risco aumentado a partir dos 35-45 anos), um histórico familiar de diabetes em parentes de primeiro grau (indicando componente genética), pertencimento a certas etnias/raças (como africana, hispânica, indígena e asiática, que apresentam maior prevalência) e um histórico pessoal de Diabetes Mellitus Gestacional (DMG) em mulheres, que consideravelmente aumenta o risco de desenvolvimento de DM2 posteriormente. Esses fatores são amplamente reconhecidos por diretrizes de saúde (ADA, 2024).
- **Fatores de Risco Modificáveis:** Diferentemente dos não modificáveis, estes fatores estão majoritariamente ligados a hábitos e estilo de vida, permitindo um manejo ativo para redução do risco de DM2. O sobrepeso e a obesidade, frequentemente avaliados por um Índice de Massa Corporal (IMC) igual ou superior a 25 kg/m², são dos principais fatores de risco. O sedentarismo, caracterizado pela ausência ou insuficiência de atividade física regular, e uma dieta não saudável (rica em alimentos processados, açúcares, gorduras saturadas, e pobre em fibras) também contribuem significativamente. O tabagismo é outro fator de risco comportamental importante que pode ser cessado. Intervenções como a adoção de uma alimentação equilibrada, a prática regular de exercícios físicos, a manutenção de um peso saudável e a cessação do tabagismo são cruciais para a prevenção do DM2 (ADA, 2024).

- **Condições Médicas Associadas/Preexistentes:** A presença de certas condições médicas aumenta substancialmente o risco de desenvolver DM2. O pré-diabetes (caracterizado por níveis de glicemia ou A1C acima do normal, mas abaixo do limiar para diagnóstico de diabetes), a hipertensão arterial sistêmica (pressão arterial $\geq 130/80$ mmHg ou em tratamento) e a dislipidemia (níveis baixos de HDL colesterol e/ou níveis elevados de triglicerídeos) são fortes preditores e componentes frequentes da síndrome metabólica. Outras condições incluem a Síndrome dos Ovários Policísticos (SOP), a presença de acantose nigricans (um marcador cutâneo de resistência à insulina), doença cardiovascular já estabelecida, apneia obstrutiva do sono e infecção por HIV/AIDS. Muitas dessas condições preexistentes podem ser manejadas ou melhoradas através de tratamento médico específico e modificações no estilo de vida – como o controle da pressão arterial, a melhoria do perfil lipídico e a reversão do pré-diabetes para normoglicemia através de dieta e exercício – o que pode ajudar a mitigar o risco de desenvolvimento do DM2 ou melhorar seu controle (ADA, 2024).

I. Importância da previsão e diagnóstico precoce do diabetes para a prevenção de complicações

O diagnóstico precoce do diabetes é crucial porque a hiperglycemia prolongada prova de forma silenciosa, danos vasculares, olhos, rins e nervos anos antes de aparecerem sintomas claros. Muitos adultos chegam ao consultório já com alguma perda de visão, comprometimento renal ou doença cardiovascular justamente porque passaram esse período sem saber que estavam doentes.

Quando a doença (ou mesmo o pré-diabetes) é detectada cedo, há tempo para medidas simples e de baixo custo — ajuste alimentar, atividade física, perda de peso e, se necessário, medicação — que podem evitar a progressão ou até reverter o quadro inicial, poupando o paciente de complicações irreversíveis e custos futuros elevados. Portanto, rastrear grupos de risco e diagnosticar o diabetes o quanto antes preserva qualidade de vida e reduz mortes e gastos em saúde (SBD, 2024).

2.4.4 Critérios técnicos para o diagnóstico laboratorial

O manejo adequado do diabetes mellitus reside no seu diagnóstico preciso e oportuno, que é fundamentalmente baseado na detecção laboratorial de hiperglicemia. Diversos testes são empregados para este fim, cada um com suas particularidades, indicações e limitações. As diretrizes de sociedades científicas renomadas, como a Sociedade Brasileira de Diabetes (SBD) e a American Diabetes Association (ADA). Os critérios para diagnóstico laboratorial podem ser definidos conforme a tabela abaixo.

Tabela 3 – Critérios laboratoriais para diagnóstico de diabetes.

Teste Laboratorial	Normalidade	Pré-Diabetes	Diabetes Mellitus
Glicemia Plasmática de Jejum (GPJ)	< 100 mg/dL (<5.6 mmol/L)	100 a 125 mg/dL (5.6 a 6.9 mmol/L) (Glicemia de Jejum Alterada)	$\geq 126 \text{ mg/dL} (\geq 7.0 \text{ mmol/L})$
TTGO - Glicemia 1h pós 75g glicose (SBD)	< 155 mg/dL (<8.6 mmol/L)	155 a 208 mg/dL (8.6 a 11.5 mmol/L)	$\geq 209 \text{ mg/dL} (\geq 11.6 \text{ mmol/L})$
TTGO - Glicemia 2h pós 75g glicose	< 140 mg/dL (<7.8 mmol/L)	140 a 199 mg/dL (7.8 a 11.0 mmol/L) (Tolerância à Glicose Diminuída)	$\geq 200 \text{ mg/dL} (\geq 11.1 \text{ mmol/L})$
Hemoglobina Glicada (A1C)	< 5,7% (<39 mmol/mol)	5,7% a 6,4% (39 a 47 mmol/mol)	$\geq 6,5\% (\geq 48 \text{ mmol/mol})$
Glicemia Plasmática Casual	N/A	N/A	$\geq 200 \text{ mg/dL} (\geq 11.1 \text{ mmol/L})$ (na presença de sintomas clássicos de hiperglicemia)

Nota: Os valores em mmol/L são conversões aproximadas e podem variar ligeiramente dependendo das fontes. A padronização da A1C em mmol/mol é o padrão SI, mas a porcentagem ainda é amplamente utilizada.

Fonte: (SBD, 2024)

A fidedignidade do diagnóstico de DM é crucial. Portanto, em indivíduos assintomáticos, um único resultado de teste laboratorial alterado (GPJ, TTGO ou A1C) não é suficiente para confirmar o diagnóstico. Nestes casos, a SBD (2024), recomenda seguir o fluxograma da Figura 12 a seguir.

Figura 12 – Fluxograma para diagnóstico de diabetes mellitus tipo 2.



Fonte: (SBD, 2024)

2.4.5 Métodos tradicionais de diagnóstico e abordagens baseadas em Machine Learning.

A análise comparativa entre os métodos tradicionais de diagnóstico do diabetes e as abordagens baseadas em Aprendizado de Máquina revela um panorama complexo e em rápida evolução, conforme Wang *et al.* (2025). Os métodos tradicionais, como, glicemia em jejum (FPG), HbA1c e TTGO (Teste oral de tolerância a glicose) (SBD, 2024), permanecem os pilares do diagnóstico, sustentados por critérios bem estabelecidos e uma longa história de utilização clínica. No entanto, as suas limitações intrínsecas – custo, acessibilidade e, em alguns casos, sensibilidade subótima para estados pré-diabéticos – criam uma necessidade premente de inovação. A Tabela 4 a seguir faz uma comparação entre os métodos.

Tabela 4 – Comparação qualitativa de atributos diagnósticos: métodos tradicionais vs. baseados em ML.

Atributo	Métodos Tradicionais (FPG, HbA1c, TTGO)	Abordagens Baseadas em ML
Base Primária	Medição bioquímica direta	Reconhecimento de padrões em dados complexos
Invasividade	Geralmente levemente invasivo (colheita de sangue)	Frequentemente não invasivo ou utiliza dados existentes
Custo (Rastreio)	Moderado a Alto	Potencialmente mais baixo (especialmente não invasivos)
Velocidade do Resultado	Pode ser mais lento (processamento laboratorial)	Potencialmente mais rápido (análise computacional)
Requisito de Dados	Biomarcador(es) único(s) ou poucos	Múltiplas/diversas características (clínicas, estilo vida)
Interpretabilidade (Típica)	Alta (base fisiológica clara)	Frequentemente Baixa a Moderada (melhorando com XAI)
Potencial de Detecção Precoce (Pré-diabetes)	Moderado (variável com o teste)	Alto (foco em predição de risco)
Escalabilidade para Rastreio em Massa	Mais Baixa (devido a custo/invasividade)	Potencialmente mais Alta
Maturidade/Adoção Clínica	Alta (padrão-ouro estabelecido)	Emergente/Baixa (apesar da pesquisa intensiva)
Força Primária	Padrão-ouro estabelecido, medição direta	Detecção de padrões, predição, não invasividade
Desafio Primário	Custo, invasividade, acessibilidade, sensibilidade	Generalização, interpretabilidade, validação clínica

Fonte: adaptado pelo autor a partir de Wang et al. (2025), SBD (2024), ADA (2024)

Conforme os autores citam ADA (2024) e Wang et al. (2025), o *Machine Learning* é uma técnica excelente de “screening”, ou seja, rastreamento. A confirmação somente é realizada a partir de estudo clínico laboratorial (conforme Figura 12 anteriormente apresentada) e quadro do paciente.

2.5 ESTEIRA (PIPELINE) DE MACHINE LEARNING: AED, PREPARAÇÃO, MODELAGEM E AVALIAÇÃO

2.5.1 AED – Análise exploratória de dados

Análise Exploratória de Dados (AED) constitui uma etapa primordial e investigativa no ciclo de trabalho com dados, antecedendo fases mais complexas como a modelagem estatística ou o treinamento de algoritmos de Machine Learning. Trata-se do momento inicial em que o analista ou cientista de dados estabelece um contato direto e aprofundado com o conjunto de dados disponível.

A Análise Exploratória de Dados (AED) é fundamentalmente a fase em que se busca "conhecer" os dados. O seu propósito central é permitir a compreensão das características primárias e essenciais de um conjunto de dados. Esta etapa não se configura como uma observação passiva, mas sim como um processo ativo de questionamento e investigação inicial. A intenção é ir além da simples visualização superficial, procurando ativamente "descobrir o que os dados podem revelar além da tarefa de modelagem ou teste de hipóteses" (AKMEŞE, 2022).

A profundidade e a qualidade com que a AED é conduzida impactam diretamente o nível de conhecimento adquirido sobre os dados. Este entendimento inicial é crucial, pois influencia a robustez e a validade de todo o processo subsequente de análise de dados, incluindo o desenvolvimento e a avaliação de modelos de Machine Learning. Uma AED superficial ou negligenciada pode levar a interpretações equivocadas, à omissão de aspectos críticos dos dados e, consequentemente, a conclusões ou modelos falhos. Ter métodos, objetivos e realizar a esteira de maneira completa podem mudar os resultados significativamente.

A tabela abaixo resume os componentes chave e a relevância da Análise Exploratória de Dados para o pré-processamento de dados, etapa posterior a AED.

Tabela 5 – Componentes chave e relevância da análise exploratória de dados (AED).

Componente da AED	Descrição Concisa
Definição	Fase de "conhecimento" dos dados, buscando entender suas principais características.
Objetivos Principais	Identificar padrões, anomalias (outliers), valores ausentes; visualizar distribuições das variáveis e as relações entre elas; formular hipóteses.
Métodos Comuns	Visualização de dados (gráficos, heatmaps), estatísticas descritivas, análise de atributos e correlações.
Importância no Ciclo de Dados/Machine Learning	Informa decisões na preparação de dados, melhora a qualidade dos dados, aumenta a confiabilidade dos modelos de ML e permite a descoberta de insights.

Fonte: elaborado pelo autor a partir de Kuhn e Johnson (2019) e Akmeşe (2022)

Para atingir seus objetivos, a AED emprega uma variedade de métodos e técnicas, com destaque para a visualização de dados. A EDA utiliza "métodos visuais para resumir suas características principais", e a "visualização de dados é a representação visual de dados quantitativos para comunicação e análise". Ferramentas como "gráficos e tabelas" são comumente utilizadas. Um exemplo específico de técnica visual poderosa é o "*heatmap*" (mapa de calor), que pode ser usado para "examinar dados multivariados, mostrar a variância entre variáveis, indicar se alguma variável é semelhante a outra e detectar se há correlação entre as variáveis".

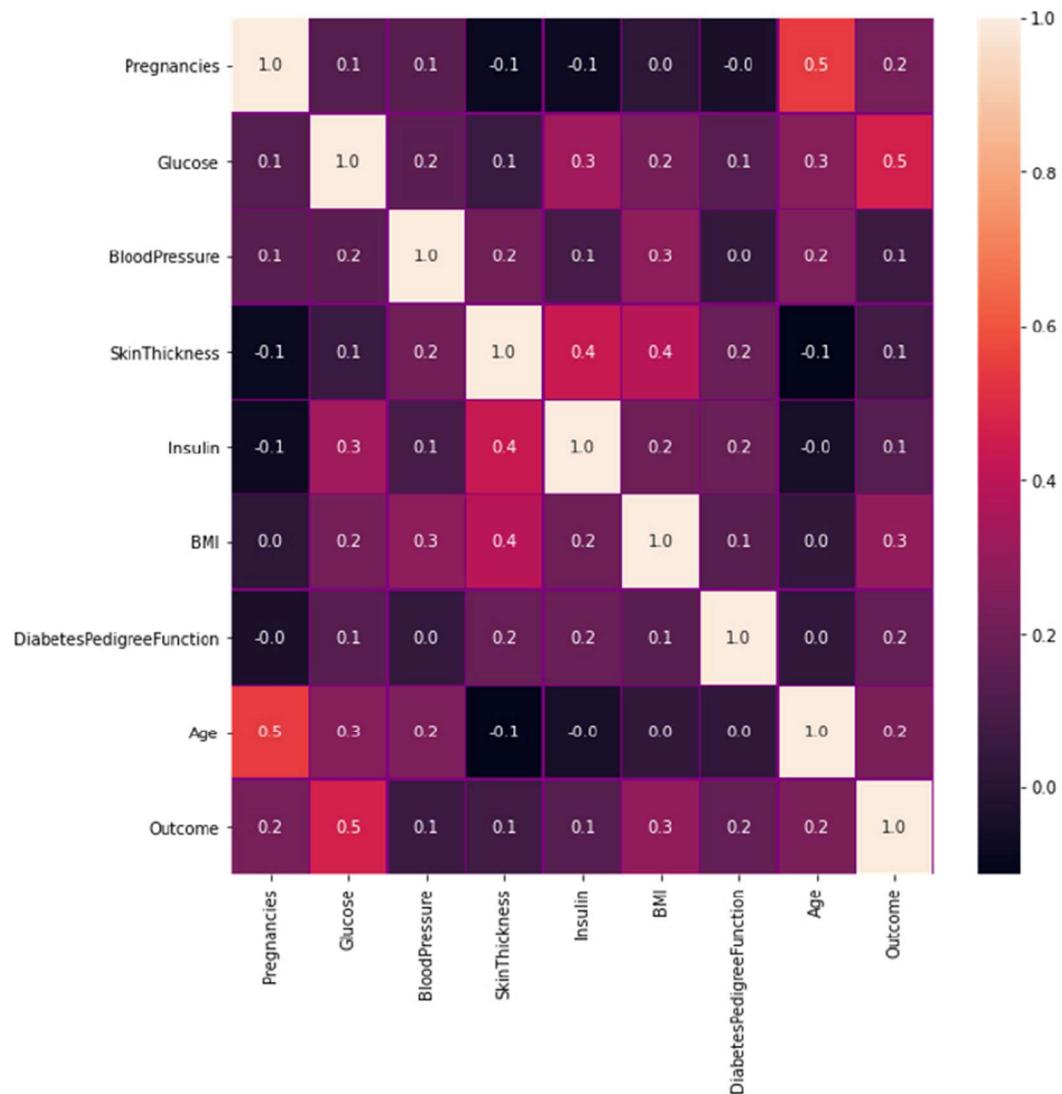
Paralelamente às técnicas visuais, a utilização de estatísticas descritivas é fundamental para resumir os dados. Embora o termo não seja explicitamente conectado à AED em todas as fontes fornecidas, sua prática é inerente ao processo de "conhecer os dados". Um estudo que, embora não utilize o termo AED, descreve uma metodologia análoga, menciona o uso de estatística descritiva, a partir de frequência absoluta (n) e relativa (%) das variáveis qualitativas e medidas de tendência central e dispersão dos dados para as variáveis quantitativas (BAZON; PEREIRA, 2023). Essas medidas (média, mediana, desvio padrão, etc.) são cruciais para entender as características centrais e a variabilidade de cada atributo.

A análise de atributos e suas correlações também é uma prática comum. A "Análise de Atributos" pode identificar variáveis preditoras importantes; por

exemplo, em um contexto de diagnóstico de diabetes, a Glicose foi identificada como o melhor indicador do resultado. O mesmo estudo aponta que "Características fortemente correlacionadas podem prever a classe alvo mais facilmente e gerar resultados mais significativos".

A escolha dos métodos de AED deve ser criteriosa, guiada pela natureza dos dados (quantitativos, qualitativos) e pelas questões de pesquisa. A Figura 13 abaixo exemplifica algumas práticas de AED.

Figura 13 – “Heatmap” das features para visualizar a correlação entre os itens.



Fonte: (AKMEŞE, 2022).

2.5.2 Preparação dos dados

O pré-processamento e a seleção de atributos são etapas cruciais antes de treinar um modelo de Machine Learning. Eles preparam seus dados para que o modelo possa aprender da melhor forma possível.

Importância do pré-processamento de dados na qualidade dos modelos: O pré-processamento é fundamental para garantir que os modelos de Machine Learning aprendam corretamente e façam previsões precisas. Dados "sujos", com inconsistências, valores faltantes ou escalas muito diferentes, podem levar a modelos com baixo desempenho ou que não generalizam bem para novos dados (WANG et al, 2025). De forma genérica e simples, é como preparar os ingredientes antes de cozinhar: bons ingredientes bem-preparados resultam em um prato melhor.

2.5.2.1 Técnicas de estratificação de “dataset”, conjunto de dados

De acordo com Hastie *et al.* (2009), a separação de dados em conjuntos de treino, validação e teste é uma prática fundamental em *Machine Learning* para desenvolver modelos robustos e generalizáveis. O conjunto de treino é utilizado para que o modelo aprenda os padrões presentes nos dados. O conjunto de validação serve para ajustar os hiperparâmetros do modelo e evitar o superajustamento “*overfitting*” aos dados de treino. Por fim, o conjunto de teste é usado para avaliar o desempenho final do modelo em dados não vistos anteriormente, fornecendo uma estimativa imparcial de sua performance em produção.

Proporções comuns para essa divisão variam, mas abordagens como 70% para treino, 15% para validação e 15% para teste, ou 80% para treino e 20% para teste (com validação cruzada no conjunto de treino) são frequentemente adotadas. A importância dessa separação reside na capacidade de avaliar se o modelo consegue generalizar o aprendizado para novos dados, em vez de simplesmente memorizar os dados de treino.

Complementarmente à divisão de dados, quando a validação cruzada (CV) é empregada — frequentemente sobre o conjunto de treino para a seleção de hiperparâmetros ou como alternativa a um conjunto de validação estático —

Hastie et al. (2009) ressaltam um mecanismo crucial para sua eficácia na estimativa do erro de generalização, o que é particularmente relevante em cenários de alta dimensionalidade onde ajustes espúrios aos dados são mais prováveis. Os autores elucidam que a confiabilidade da CV advém da exigência de que todo o processo de modelagem, incluindo a seleção de preditores e a otimização de seus respectivos parâmetros (como os pontos de corte em árvores de decisão, por exemplo), seja integralmente refeito para cada 'fold' ou subconjunto de dados da validação cruzada, utilizando-se exclusivamente os dados de treino daquele fold específico. Esta completa re-estimação a cada iteração assegura que a avaliação do modelo em cada fold seja realizada sobre dados efetivamente não utilizados no treinamento daquele modelo particular, prevenindo uma subestimação do erro de predição e permitindo que a CV forneça uma estimativa mais fidedigna do desempenho do modelo em dados genuinamente novos, conforme demonstrado pelos autores através de simulações.

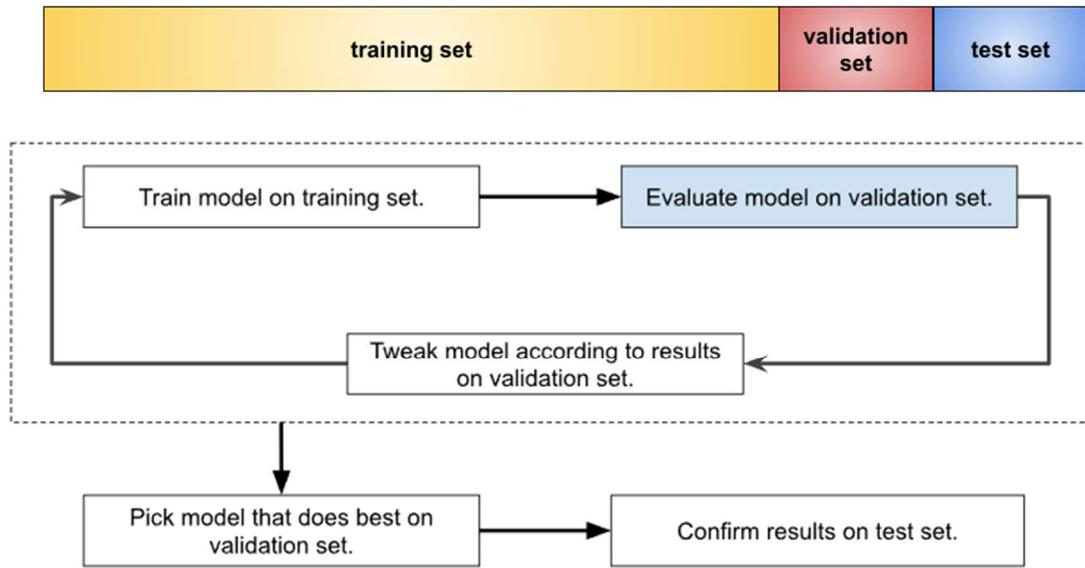
De acordo com Kuhn e Johnson (2019), um dos problemas mais comuns e que podem arruinar o treinamento de um modelo é o *data leakage*, ocorre quando informações de fora do conjunto de treinamento são usadas para criar o modelo, levando a uma avaliação de desempenho excessivamente otimista e irrealista. Essencialmente, o modelo "aprende" com dados que não estariam disponíveis em um cenário de previsão real no momento da predição.

Uma das fontes mais comuns de data leakage, é algo que o autor enfatiza, é realizar etapas de pré-processamento (como normalização, escalonamento, imputação de dados faltantes, e especialmente seleção de características supervisionada) usando informações de *todo* o conjunto de dados *antes* de dividi-lo em treino e teste, ou antes de realizar a reamostragem (como validação cruzada).

As principais consequências são: estimativas de desempenho inflacionadas, o modelo parece performar muito bem nos dados de teste ou validação, mas falha miseravelmente em dados verdadeiramente novos. A seleção de modelos/características incorreta: pode levar à escolha de modelos ou características que na verdade não generalizam bem. E os resultados enganosos, pode levar a conclusões errôneas sobre a predição de um problema.

A Figura 14 a seguir elucida a melhor forma de se realizar a estratificação de um conjunto de dados/*dataset*.

Figura 14 - Estratificação e fluxo ideal.



Fonte: (GOOGLE, s. d.)

2.5.2.2 Técnicas de remoção de outliers e balanceamento de “*dataset*” conjunto de dados

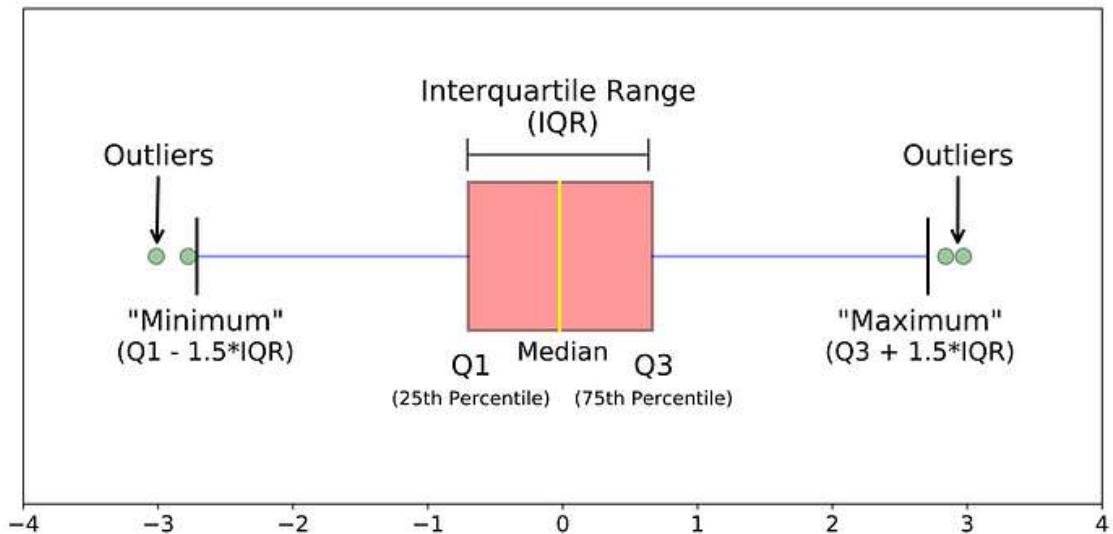
I. IQR -

O Intervalo Interquartil (IQR) é uma técnica estatística utilizada para a detecção de outliers, particularmente popular em domínios da engenharia através da aplicação em *boxplots*. O método define o IQR como a diferença entre o terceiro quartil (Q_3) e o primeiro quartil (Q_1) dos dados ($IQR = Q_3 - Q_1$). Observações que se encontram fora dos limites definidos por Limite Inferior = $Q_1 - 1,5 \cdot IQR$ e Limite Superior = $Q_3 + 1,5 \cdot IQR$ são tipicamente consideradas potenciais outliers.

Apesar de sua fácil implementação, abordagens estatísticas como o IQR para detecção de outliers, de forma geral, apresentam algumas limitações. Elas podem ser sensíveis aos próprios outliers, pois métricas como quartis podem ser influenciadas por valores extremos. Além disso, muitas dessas técnicas pressupõem uma distribuição normal dos dados, o que nem sempre reflete a realidade de conjuntos de dados complexos. Em contextos multivariados e com

alta dimensionalidade, a eficácia de métodos univariados como o IQR para identificar todos os tipos de outliers pode ser limitada (LARTEY *et al.*, 2024). A Figura 14 abaixo demonstra o seu funcionamento em um *boxplot*.

Figura 15 – IQR.



Fonte: (IGUENFER, 2020)

II. SMOTE e Oversampling

Chawla *et al.* (2002) propõem o SMOTE como uma técnica de oversampling sintético eficaz para lidar com desbalanceamento de dados.

O desbalanceamento de classes, onde uma classe (majoritária) possui um número significativamente maior de instâncias do que outra (minoritária), é um problema comum que pode levar modelos de aprendizado a serem enviesados em favor da classe majoritária. O SMOTE surge como uma solução sofisticada para o oversampling (sobresampling) da classe minoritária.

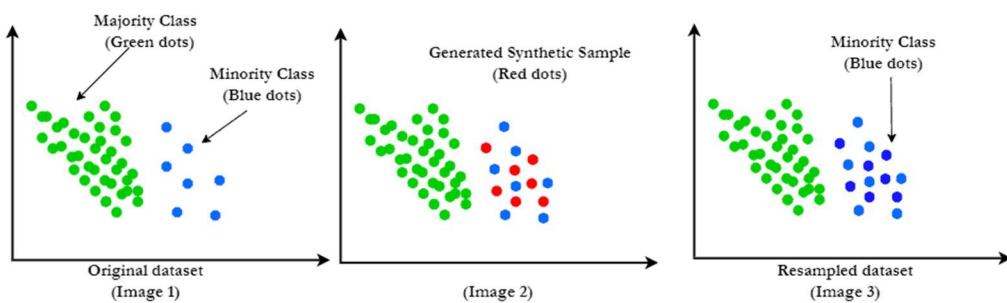
A técnica SMOTE, criada e introduzida por Chawla *et al.* (2002), visa balancear a distribuição das classes através da criação de instâncias sintéticas da classe minoritária. O processo fundamental envolve:

1. Para cada instância da classe minoritária, são identificados seus k vizinhos mais próximos (geralmente $k=5$) pertencentes à mesma classe.
2. Aleatoriamente, um ou mais desses vizinhos são selecionados.
3. Novas instâncias sintéticas são geradas por interpolação linear entre a instância original e os vizinhos selecionados.

Essencialmente, uma nova amostra é criada em algum ponto ao longo do segmento de linha que une a amostra minoritária original e um de seus vizinhos escolhidos.

Vantagens do SMOTE sobre o *Oversampling* Tradicional (*Random Oversampling*): O *oversampling* tradicional, que consiste em simplesmente duplicar aleatoriamente as instâncias da classe minoritária (*upsampling*), pode facilmente levar ao *overfitting*. Isso ocorre porque o modelo pode aprender regiões de decisão excessivamente específicas para as instâncias minoritárias replicadas, comprometendo sua capacidade de generalização para dados não vistos. A Figura 15 abaixo ilustra brevemente o seu funcionamento e na sequência a Tabela 6 demonstra as diferenças entre o SMOTE e o *random oversampling*.

Figura 16 – Demonstração do funcionamento do SMOTE.



Fonte: (PARASCHIV, 2024)

Tabela 6 - Vantagens e desvantagens do SMOTE.

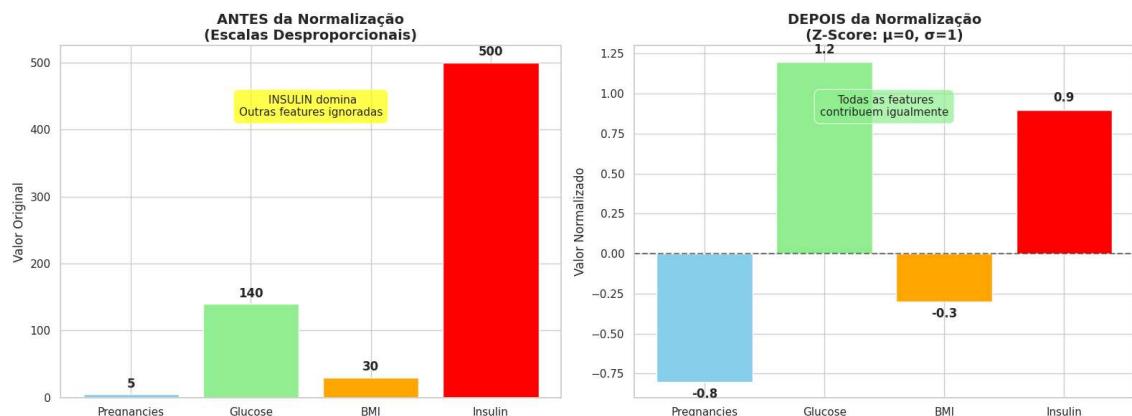
Vantagens	Desvantagens
Melhora a performance de modelos em dados desbalanceados	Pode gerar ruído e aumentar a sobreposição entre classes
Reduz o risco de <i>overfitting</i> comparado ao <i>random oversampling</i>	Menos eficaz em espaços de características de alta dimensionalidade
Flexível e compatível com diversos algoritmos de ML	Pode não capturar padrões complexos na distribuição da minoria
Fácil implementação em bibliotecas populares	Aumenta a complexidade computacional do treinamento
Ajuda a criar modelos mais generalizáveis	Pode perder sutilezas e a variabilidade fina dos dados originais

Fonte: adaptado pelo autor a partir de Chawla et al. (2002).

III. Normalização dos dados

A normalização das variáveis é uma etapa fundamental no pré-processamento de dados, uma vez que algoritmos de aprendizado de máquina são sensíveis à escala das features. Conforme destacado por Hastie et al. (2009), a padronização dos dados de entrada para média zero e variância um é essencial, pois a escala das variáveis afeta significativamente a performance de algoritmos como k-nearest neighbors e SVM e outros. Já modelos como árvore de decisão, *Random Forest*, *Gradient Boosting* são poucos sensíveis a normalização devido a sua característica, a Figura 16 a seguir demonstra o seu funcionamento na prática.

Figura 17 – Normalização de dados.



Fonte: elaborado pelo autor.

2.5.2.3 Técnicas de modelagem e treinamento de algoritmos

Os estudos de Wang et al. (2025) e Akmese (2022) empregam um conjunto diversificado de algoritmos de aprendizado de máquina supervisionado para tarefas de classificação no domínio médico. Especificamente, o trabalho de Wang et al. (2025) investigou a predição de tolerância à glicose diminuída isolada (I-IGT) em homens chineses da etnia Han, utilizando algoritmos como Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbors (KNN), Naive Bayes (NB), Adaptive Boosting (AdaBoost) e Gradient Boosting Machines (GBM). Por sua vez, Akmese (2022) focou no diagnóstico de diabetes, aplicando Random Forest, Gradient

Boosting, XGBoost (XGB), LightGBM (LGBM), Decision Tree, AdaBoost, Support Vector Machine, Logistic Regression, KNN e Naive Bayes.

A escolha destes algoritmos é recorrente em pesquisas na área médica, refletindo a sua capacidade de modelar relações complexas frequentemente encontradas em dados clínicos e laboratoriais. Tal fato pode indicar um grau de maturidade ou consenso nas abordagens metodológicas dentro desta área de pesquisa, onde estas técnicas são frequentemente escolhidas pela sua eficácia comprovada e familiaridade entre a comunidade de pesquisa médica e bioinformática para tarefas de classificação. A Tabela 7 abaixo demonstra o funcionamento e a Tabela 8 na sequência onde cada um funciona melhor.

Tabela 7 – Como cada modelo funciona.

Grupo de modelo	Como aprende / faz previsões (versão condensada)
Modelos baseados em árvores (Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM)	Criam regras hierárquicas que segmentam os dados em nós cada vez mais puros. Random Forest treina muitas árvores em paralelo sobre amostras bootstrap; métodos de boosting treinam árvores pequenas em sequência, cada uma corrigindo o erro da anterior. A saída final combina todas as árvores (votação ou soma ponderada) .
Support Vector Machine (SVM)	Encontra o hiperplano que maximiza a margem entre classes. Se os dados não forem linearmente separáveis, aplica kernels para projetá-los em espaços de dimensão maior onde a separação é possível.
Logistic Regression (LR)	Ajusta uma função sigmoide cujos coeficientes transformam as variáveis de entrada em probabilidades de pertencimento a cada classe.
K-Nearest Neighbors (KNN)	“Treina” apenas armazenando os dados. Na predição, mede a distância da nova instância a todos os pontos, busca os K mais próximos e atribui a classe majoritária.
Naive Bayes (NB)	Calcula probabilidades a priori das classes e probabilidades condicionais de cada atributo assumindo independência entre eles; classifica pela maior probabilidade posterior (teorema de Bayes).

Fonte: adaptado pelo autor a partir de Wang et al. (2025) e Akmese (2022).

Tabela 8 – Onde cada modelo funciona melhor.

Grupo / Algoritmo	Cenários em que se destaca	Motivos práticos
Árvores e ensembles	Dados tabulares heterogêneos, presença de não linearidades complexas, necessidade de ranking de importância das variáveis	Robustos a escalas distintas, toleram valores faltantes, fornecem interpretabilidade (árvores) e alta acurácia (ensembles); LightGBM e XGBoost são padrão-ouro para bases com milhões de linhas
SVM	Conjuntos de tamanho médio (até ≈ 100 k linhas) com alta dimensionalidade, detecção de anomalias quando margem larga é desejada	Usa kernels para padrões não lineares, mantém bom desempenho com poucos exemplos positivos
Logistic Regression	Modelos de linha de base rápidos, requisitos de interpretabilidade (ciências da saúde, marketing, áreas reguladas)	Treinamento ágil, coeficientes fáceis de explicar e probabilidades calibradas
KNN	Datasets pequenos e de baixa dimensionalidade, prototipagem ou recomendação ad-hoc	Não requer treinamento, funcionamento intuitivo; custo de predição cresce com o volume de dados
Naive Bayes	Classificação de texto (spam, sentimentos), dados com independência aproximada entre atributos	Extremamente rápido, exige poucos dados, lida bem com alta dimensionalidade

Fonte: adaptado pelo autor a partir de Wang et al. (2025) e Akmese (2022).

A otimização de hiperparâmetros pode ser realizada a partir de busca em referências bibliográficas existentes, conforme encontrado em Akmese (2022), ou utilizar metodologias de busca de hiperparâmetros existentes, a Tabela 9 a seguir resume três estratégias amplamente empregadas para ajuste de hiperparâmetros em modelos de machine learning. Cada método, do varrimento exaustivo ao uso de abordagens probabilísticas baseadas em histórico, apresenta vantagens e limitações que dependem do tamanho do espaço de busca, do orçamento computacional disponível e da necessidade de automação.

A “Visão rápida” destaca o princípio geral de cada técnica, enquanto a coluna “Como funciona” descreve, em poucas linhas, o fluxo operacional necessário para aplicá-la na prática.

Tabela 9 - Estratégias de ajuste de hiperparâmetros.

Método	Visão rápida	Como funciona
Grid Search	Busca exaustiva e sistemática	Define-se uma grade discreta de valores para cada hiperparâmetro e avaliam-se todas as combinações com validação cruzada; retorna a configuração com melhor métrica média.
Random Search	Amostragem aleatória do espaço	Em vez de varrer toda a grade, sorteia-se um número fixo de combinações; é mais eficiente quando apenas alguns hiperparâmetros influenciam fortemente o desempenho.
Meta-aprendizado / AutoML	Aprendiz que aprende a buscar	Usa histórico de tarefas parecidas, modelos probabilísticos e heurísticas (p. ex. SMBO, ensembles) para iniciar e guiar a busca em regiões promissoras, automatizando seleção de algoritmo, pré-processamento e ajuste de hiperparâmetros.

Fonte: Adaptado de Hastie et al. (2009); Bergstra; Bengio (2012); Hutter et al. (2019).

2.5.2.4 Avaliação de modelos de *Machine Learning*

A avaliação de modelos de Machine Learning é um passo crucial no desenvolvimento de soluções inteligentes, assegurando sua capacidade de generalizar para dados não vistos e realizar previsões precisas. A escolha das métricas deve ser criteriosa, alinhada aos objetivos da tarefa e aos custos dos erros de predição. Para uma avaliação fidedigna, é fundamental utilizar um conjunto de dados de teste independente do treino, prática que previne o superajuste (overfitting) – onde o modelo memoriza os dados de treino mas falha com novos – e garante uma estimativa realista do seu verdadeiro desempenho, de acordo com Kuhn e Johnson (2019).

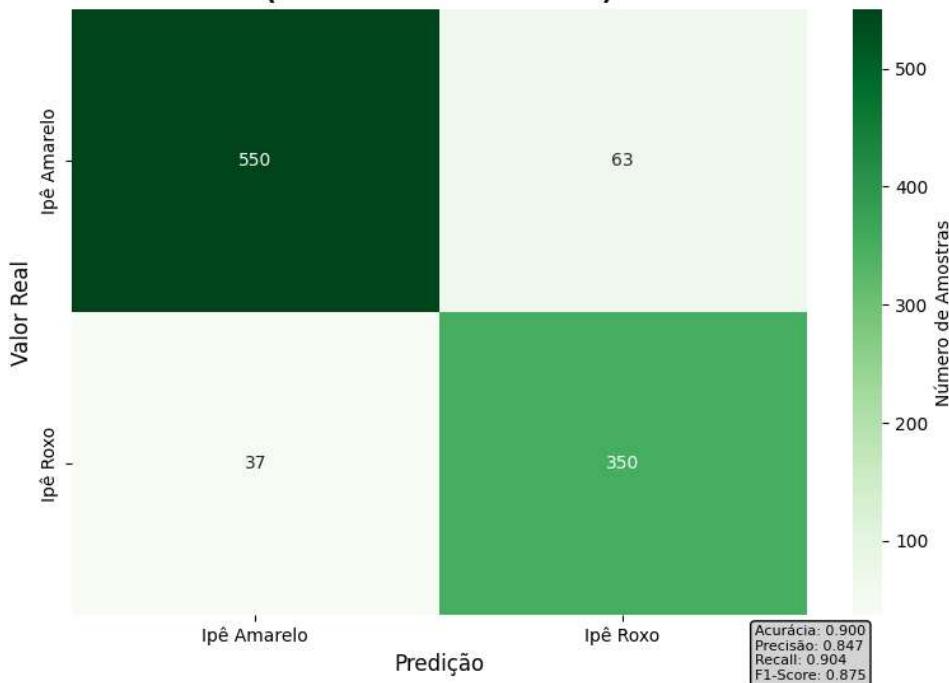
Com base nisso, passaremos a discutir as principais técnicas, métodos e métricas de avaliação, como a matriz de confusão, acurácia, F1-score, recall e AUC ROC.

I. Matriz de confusão

A Matriz de Confusão é uma ferramenta fundamental e essencial na avaliação do desempenho de algoritmos de classificação. Ela oferece uma visualização clara e concisa de como as previsões do modelo se comparam com os valores reais, permitindo uma análise mais profunda do que uma simples taxa de acerto geral. Ao invés de apenas indicar se uma previsão está certa ou errada, a matriz de confusão detalha os *tipos* de acertos e erros cometidos pelo modelo, revelando suas forças e fraquezas. A Figura X a seguir demonstra o seu funcionamento.

Figura 18 – Matriz de confusão.

**Matriz de Confusão - Classificação de Espécies de Ipê
(Dataset Demonstrativo)**



Fonte: elaborado pelo autor.

De acordo com os resultados apresentados na Figura 18, analisa-se o entendimento detalhado do que a matriz de confusão reportou. Para tal, é imprescindível analisar seus quatro componentes fundamentais, que quantificam os acertos e erros do modelo, a interpretação e a importância de cada um desses elementos, conforme elucidado por autores como Wang et al. (2025) e Akmese (2022), serão exploradas a seguir, permitindo uma compreensão completa do desempenho do classificador.

1. Verdadeiros Positivos (VP/True Positives - TP): 350

- O modelo previu "Ipê Roxo" e a árvore era *realmente* um Ipê Roxo.
- Na imagem: o quadrante inferior direito mostra 350 acertos para o Ipê Roxo.

2. Verdadeiros Negativos (VN / True Negatives - TN): 550

- O modelo previu "Ipê Amarelo" (ou seja, *não* Ipê Roxo) e a árvore era realmente um Ipê Amarelo.
- Na imagem: o quadrante superior esquerdo mostra 550 acertos para o Ipê Amarelo (que são os negativos corretos, se o positivo é Roxo).

Obs: O termo "Verdadeiro Negativo", embora possa confundir e sugerir um erro, na verdade indica um acerto: o modelo previu corretamente que uma instância pertence à "classe negativa". Em classificação, é comum definir uma "classe positiva" (o foco da detecção, como "Ipê Roxo") e uma "classe negativa" (seu oposto, como "Ipê Amarelo").

3. Falsos Positivos (FP / False Positives - FP): 63 (Erro Tipo I)

- O modelo previu "Ipê Roxo", mas a árvore era *na verdade* um Ipê Amarelo.
- Na imagem: o quadrante superior direito mostra que 63 Ipês Amarelos foram classificados erroneamente como Roxos. O modelo deu um "alarme falso" para Ipê Roxo.

4. Falsos Negativos (FN / False Negatives - FN): 37 (Erro Tipo II)

- O modelo previu "Ipê Amarelo", mas a árvore era *na verdade* um Ipê Roxo.
- Na imagem: O quadrante inferior esquerdo mostra que 37 Ipês Roxos não foram identificados pelo modelo e foram classificados erroneamente como Amarelos. O modelo "deixou passar" esses Ipês Roxos.

II. Métricas de avaliação

Após analisar os componentes da Matriz de Confusão – os Verdadeiros Positivos (VP/TP), Verdadeiros Negativos (VN/TN), Falsos Positivos (FP/FP) e Falsos Negativos (FN/FN) – métricas quantitativas que resumem o desempenho do modelo podem ser calculadas, conforme Hastie et al., (2009). Essas métricas são essenciais porque:

- Quantificam a performance do modelo de maneira objetiva.
- Permitem comparar diferentes modelos ou diferentes versões do mesmo modelo.
- Ajudam a entender as forças e fraquezas específicas do modelo, indicando onde ele pode ser melhorado.
- A escolha da métrica mais relevante depende do objetivo específico do problema e dos custos associados a cada tipo de erro.

A Tabela 10 abaixo sintetiza os valores de cada variável utilizada pelo autor para melhor entendimento e utilização nas fórmulas na sequência.

Tabela 10 – Síntese dos valores da matriz de confusão e siglas.

Sigla (português/inglês)	Valor	Descrição no Contexto do Problema
VP ou TP	350	Modelo classificou corretamente como "Ipê Roxo" (era realmente Ipê Roxo).
VN ou TN	550	Modelo classificou corretamente como "Ipê Amarelo" (era realmente Ipê Amarelo).
FP ou FN	63	Modelo classificou incorretamente como "Ipê Roxo" (mas era Ipê Amarelo).
FN ou FP	37	Modelo classificou incorretamente como "Ipê Amarelo" (mas era Ipê Roxo).
Total	1000	Número total de instâncias de Ipê classificadas pelo modelo.

Fonte: elaborado pelo autor.

1) Acurácia (Accuracy)

- **O que mede:** a proporção de todas as classificações que o modelo acertou (tanto Ipês Roxos quanto Amarelos). É uma visão geral do desempenho.

Fórmula:

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN}$$

Cálculo com os dados:

$$\text{Acurácia} = \frac{350 + 550}{350 + 550 + 63 + 37} = \frac{900}{1000} = 0.900$$

- **Interpretação:** O modelo acertou **90%** de todas as classificações de espécies de Ipê no conjunto de dados demonstrativo. Isso significa que, de cada 100 árvores, o modelo classificou 90 corretamente.

2) Precisão (*Precision*) - para a classe "Ipê Roxo"

- **O que mede:** Das instâncias que o modelo classificou como "Ipê Roxo", quantas eram *realmente* "Ipê Roxo"? Ajuda a entender a confiabilidade das previsões positivas. Uma baixa precisão significa muitos Falsos Positivos (FP).

Fórmula:

$$\text{Precisão} = \frac{VP}{VP + FP}$$

Cálculo com os dados:

$$\text{Precisão} = \frac{350}{350 + 63} = \frac{350}{413} \approx 0.847$$

- **Interpretação:** Quando o modelo previu que uma árvore era "Ipê Roxo", ele estava correto em aproximadamente **84.7%** das vezes. Os outros 15.3% foram classificações incorretas de Ipê Amarelo como Roxo (Falsos Positivos, FP).

3) Sensibilidade (*recall*) - para a classe "Ipê Roxo"

- **O que mede:** De todos os "Ipês Roxos" que realmente existiam no dataset, quantos o modelo conseguiu identificar corretamente? Ajuda a entender se o modelo está "deixando passar" muitos exemplos da classe positiva. Um baixo recall significa muitos Falsos Negativos (FN).

Fórmula:

$$\text{Sensibilidade} = \frac{VP}{VP + FN}$$

Cálculo com os dados:

$$\text{Recall} = \frac{350}{350 + 37} = \frac{350}{387} \approx 0.904$$

- **Interpretação:** O modelo conseguiu identificar corretamente aproximadamente **90.4%** de todos os Ipês Roxos verdadeiros presentes no conjunto de dados. Isso significa que cerca de 9.6% dos Ipês Roxos não foram detectados pelo modelo (Falsos Negativos, FN).

4) Especificidade (*Specificity*) ou Taxa de Verdadeiros Negativos, VPN (True Negative Rate - TNR)

- **O que mede:** Dentre todas as instâncias que são realmente da classe negativa (no nosso exemplo, "Ipê Amarelo"), qual é a proporção que o modelo conseguiu classificar corretamente como negativa? Ela foca na capacidade do modelo de evitar alarmes falsos para a classe negativa.

Fórmula:

$$\text{Especificidade} = \frac{VN}{VN + FP}$$

Cálculo com os dados:

$$\text{Especificidade} = \frac{550}{550 + 63} = \frac{550}{613} \approx 0.897$$

- **Interpretação:** O modelo conseguiu identificar corretamente aproximadamente **89.7%** de todos os Ipês Amarelos verdadeiros como sendo Ipês Amarelos.

5) F1-Score – para a classe “Ipê Roxo”

- **O que mede:** É a média harmônica da Precisão e do *Recall*. Busca um equilíbrio entre as duas métricas. É particularmente útil quando as classes são desbalanceadas ou quando o custo dos Falsos Positivos (FP) e Falsos Negativos (FN) é similar. Um F1-Score alto indica que o modelo tem boa precisão e bom recall.

Fórmula:

$$F1 - Score = 2 \cdot \frac{Precisão \cdot Recall}{Precisão + Recall}$$

Cálculo com os dados:

$$F1 - Score = 2 \cdot \frac{0.847 \cdot 0.904}{0.847 + 0.904} = 2 \cdot \frac{0.765688}{1.751} \approx 0.875$$

- **Interpretação:** O F1-Score de aproximadamente **0.875** para a classe "Ipê Roxo" indica um bom equilíbrio entre a precisão (quão corretas são as previsões de "Ipê Roxo") e o recall (quantos "Ipês Roxos" reais foram encontrados).

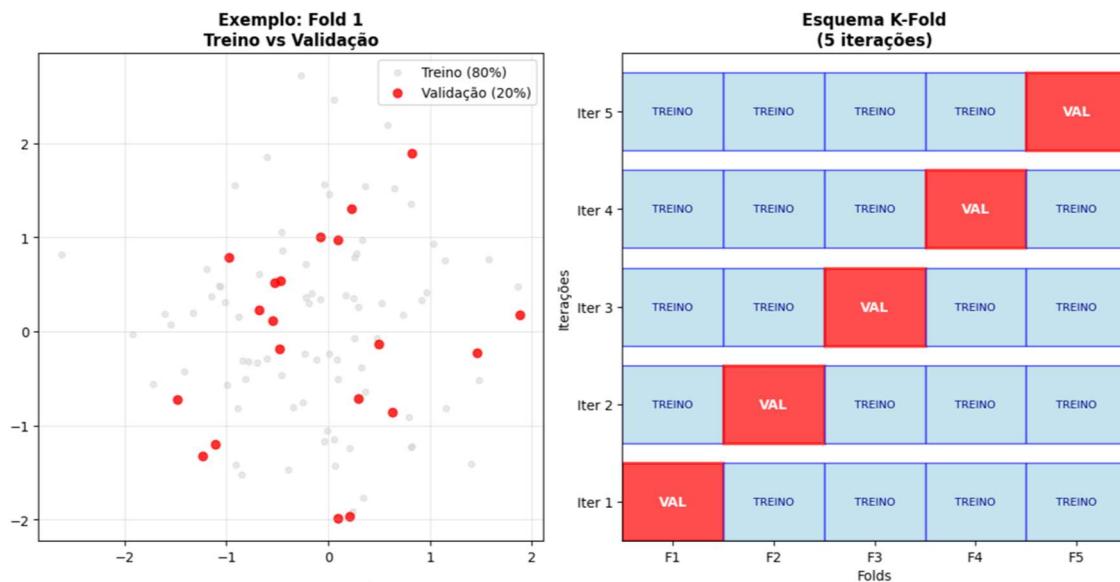
Esses cálculos, passo a passo, mostram como cada métrica é derivada da Matriz de Confusão e oferece uma perspectiva diferente sobre o desempenho do seu modelo. Ao analisar essas métricas em conjunto, você obtém uma compreensão mais completa e robusta da capacidade do seu modelo na tarefa de classificar as espécies.

6) Validação cruzada

Na Validação Cruzada *K-Fold* (*K-Fold Cross-Validation*), as "dobras" são as subpartições ou subconjuntos nos quais o seu conjunto de dados de treinamento (ou o conjunto de desenvolvimento que estamos usando para a validação cruzada) é dividido.

O processo inicia-se com o particionamento do conjunto de dados, que contém um número total de N amostras (ou instâncias). Este conjunto de N amostras é então segmentado em K subconjuntos distintos, denominados "dobras". Cada uma dessas K dobras possui um tamanho aproximadamente igual, contendo cerca de N/K amostras. É fundamental que essas dobras sejam mutuamente exclusivas (disjuntas) e coletivamente exaustivas, significando que cada amostra do conjunto original pertence a exatamente uma dobra, e todas as amostras são utilizadas no processo. A atribuição das amostras às dobras é tipicamente realizada de forma aleatória para mitigar vieses, conforme brevemente descrito por Wang et al. (2025), a Figura 19 a seguir elucida o entendimento.

Figura 19 – Demonstração da validação cruzada.



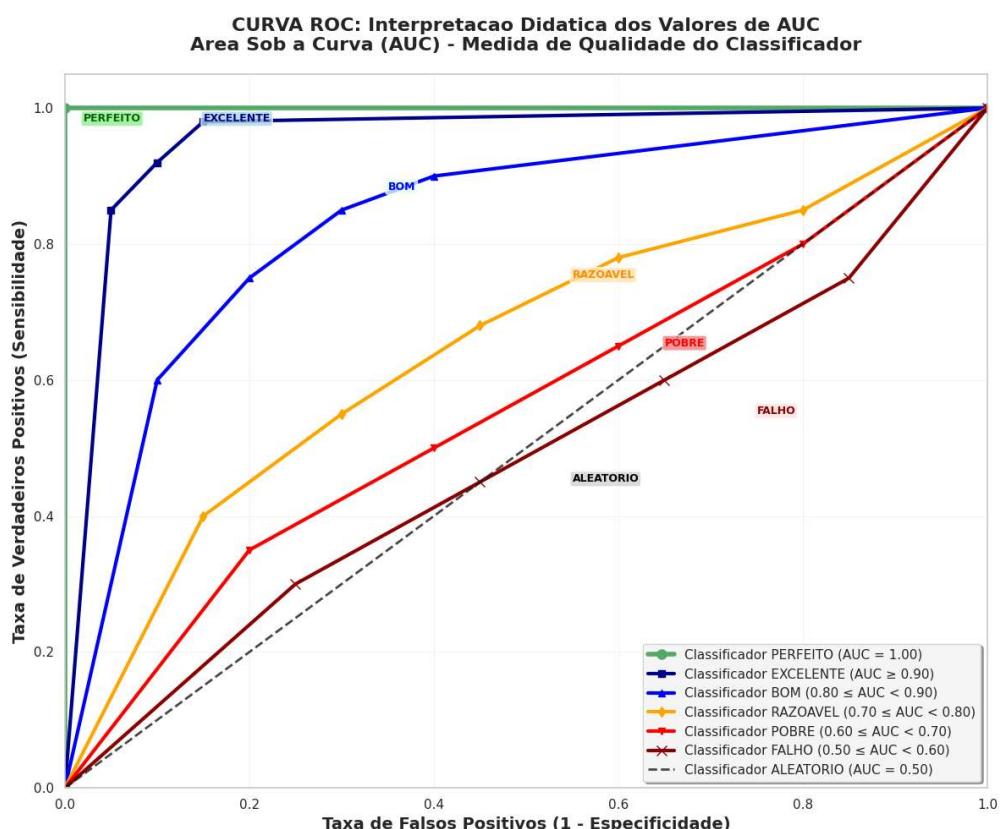
Fonte: elaborado pelo autor.

7) AUC ROC

Enquanto alguns modelos simples produzem uma classificação discreta (por exemplo, "Sim" ou "Não"), muitos dos algoritmos mais poderosos, como Regressão Logística, Redes Neurais e Gradient Boosting, produzem uma pontuação contínua ou uma probabilidade. Esta pontuação representa a confiança do modelo de que uma instância pertence à classe positiva. A análise ROC é projetada precisamente para estes classificadores de pontuação.

Para converter uma pontuação de probabilidade em uma classificação discreta, um limiar de decisão deve ser aplicado. Um limiar comum é 0.5: pontuações acima de 0.5 são classificadas como positivas, e as abaixo como negativas. A ideia central da curva ROC é avaliar o desempenho do modelo não em um único limiar arbitrário, mas em todos os limiares possíveis, de 0 a 1 (LEE; KIM, 2022). A Figura 20 abaixo demonstra e exemplifica diversos classificadores para melhor entendimento.

Figura 20 – Curva ROC.



Fonte: elaborado pelo autor, adaptado de (LEE; KIM, 2022).

- **Interpretando o Gráfico ROC**
- **A Linha Diagonal ($y=x$):** Representa um classificador sem poder discriminativo, equivalente a um palpite aleatório. Um modelo que adivinha aleatoriamente a classe positiva $p\%$ das vezes terá uma TVP e uma TFP de $p\%$, caindo, portanto, nesta linha. Qualquer classificador útil deve ter uma curva acima desta diagonal.
- **O Ponto Ideal (0, 1):** Este ponto no canto superior esquerdo representa um classificador perfeito, com 100% de sensibilidade (TVP=1) e 0% de alarmes falsos (TFP=0). Quanto mais próxima uma curva estiver deste ponto, melhor o seu desempenho.
- **Os Pontos (0, 0) e (1, 1):** Representam os classificadores triviais que sempre preveem a classe negativa e sempre preveem a classe positiva, respectivamente. A curva ROC para qualquer classificador não trivial conecta estes dois pontos.

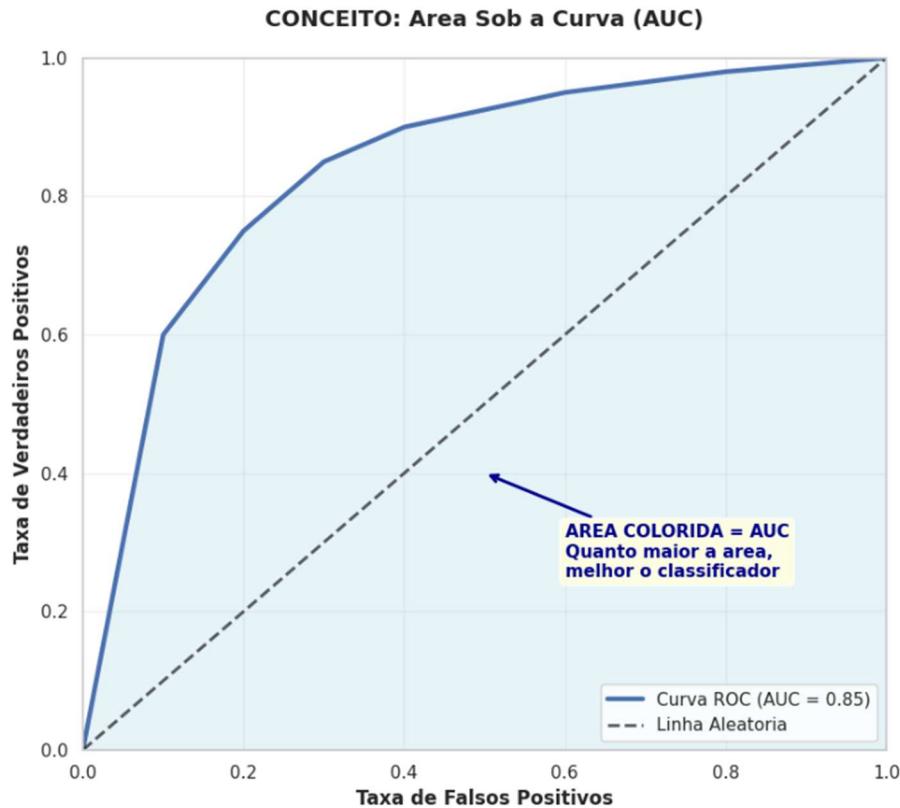
Embora a curva ROC seja altamente informativa, comparar modelos pode ser difícil quando as suas curvas se cruzam. Isso motiva a necessidade de um único valor escalar que resuma o desempenho geral em todos os limiares. A Área Sob a Curva (AUC) é esta métrica de resumo. A tabela 11 e a Figura 21 demonstram a sua forma de medição.

Tabela 11 – Interpretações de AUC.

Área Sob a Curva (AUC)	Interpretação
0.90 ≤ AUC	Excelente
0.80 ≤ AUC < 0.90	Bom
0.70 ≤ AUC < 0.80	Razoável
0.60 ≤ AUC < 0.70	Pobre
0.50 ≤ AUC < 0.60	Falho

Fonte: (LEE; KIM, 2022).

Figura 21 – Área sob a curva AUC.



Fonte: elaborado pelo autor.

- **Um guia prático para valores de AUC**
- **AUC = 1.0:** Representa um classificador perfeito que alcança uma separação perfeita das classes.
- **AUC = 0.5:** Representa um classificador sem capacidade discriminativa, equivalente a um palpite aleatório.
- **AUC < 0.5:** Representa um classificador que tem um desempenho pior que o aleatório. No entanto, tal classificador pode ser tornado útil simplesmente invertendo as suas previsões, o que resultaria em uma AUC de 1-AUC

2.6 INDIANS PIMA DATASET

O *Pima Indians Diabetes Dataset* é um recurso canônico no campo de aprendizado de máquina, servindo como um benchmark clássico para algoritmos de classificação em inúmeros estudos (AKMEŞE, 2022). O dataset apresenta uma característica de desenho fundamental: sua amostra é exclusivamente composta por mulheres de ascendência Pima com idade mínima de 21 anos (KAHN, 2017). Tal restrição não é acidental, mas reflete uma decisão metodológica estratégica, fundamentada em ao menos duas justificativas principais.

A primeira justificativa reside na busca por homogeneidade amostral para o controle de variáveis de confusão. Ao focar em um único sexo, os pesquisadores eliminaram potenciais vieses decorrentes de diferenças metabólicas, hormonais e de distribuição de gordura corporal entre homens e mulheres. Essa homogeneidade confere maior consistência às análises, permitindo que os algoritmos avaliem o impacto das variáveis preditivas sem a interferência do sexo como covariável.

A segunda justificativa, de ordem prática e clínica, está relacionada à relevância dos atributos selecionados. O conjunto de dados inclui o "número de gestações" como um fator preditivo chave, uma variável intrinsecamente ligada ao risco de desenvolvimento de diabetes em mulheres, seja pelo histórico de diabetes gestacional ou pelo impacto metabólico de múltiplas gestações. A escolha por uma coorte exclusivamente feminina, portanto, garantiu a validade e a aplicabilidade de todos os atributos para 100% dos indivíduos no estudo.

É crucial contextualizar que este subconjunto de dados foi extraído de um estudo longitudinal muito mais amplo. Curiosamente, a pesquisa original, iniciada pelo *National Institute of Diabetes and Digestive and Kidney Diseases* (NIDDK) no início dos anos 1960, tinha como foco inicial a investigação da artrite na comunidade Pima. Contudo, os pesquisadores rapidamente descobriram que os habitantes locais apresentavam uma das mais altas taxas de diabetes já registradas, o que redirecionou o foco do estudo. Em 1965, o projeto havia se tornado um estudo longitudinal sobre diabetes, no qual foi solicitado que todos os residentes da área de estudo com mais de cinco anos participassem. A

transição desses dados para a comunidade de aprendizado de máquina foi formalizada em 1990 através da doação feita por Vincent Sigillito, pesquisador do Laboratório de Física Aplicada da Universidade Johns Hopkins. A jornada desses dados, desde sua coleta em uma comunidade indígena até sua padronização como um recurso para aprendizado de máquina, levanta importantes questões éticas e de procedência, conforme discutido por Bock (2020).

Portanto, a composição do *Pima Indians Diabetes Dataset* reflete uma decisão deliberada de desenho de pesquisa, visando otimizar a coerência dos atributos e a robustez dos modelos preditivos disponíveis à época de sua compilação (c. 1990). A exclusão de homens não representa uma falha, mas uma estratégia para reduzir a heterogeneidade e aprimorar a análise de fatores de risco específicos do sexo feminino, como a gestação, no desenvolvimento do diabetes tipo 2.

3 MATERIAIS E MÉTODOS

Este capítulo detalha a metodologia empregada para o desenvolvimento do estudo, desde as ferramentas tecnológicas utilizadas até o delineamento das etapas de pré-processamento, modelagem e avaliação dos algoritmos de aprendizado de máquina.

3.1 FERRAMENTAS & TECNOLOGIAS

Para assegurar rastreabilidade e reproduzibilidade de todas as etapas, utilizou-se um conjunto integrado de ferramentas de código aberto. Esse ecossistema viabilizou a condução interativa dos experimentos, o registro completo do fluxo de trabalho e a comunicação visual dos resultados sem dependências proprietárias. Python, apontado no referencial teórico como a linguagem predominante em ciência de dados, serviu de núcleo para orquestrar bibliotecas consolidadas de manipulação numérica, aprendizado de máquina e visualização. Sua sintaxe de alto nível, aliada à vasta oferta de pacotes especializados (por exemplo, Pandas, NumPy e Scikit-learn), permitiu prototipação rápida e ajustes iterativos dos modelos preditivos, reduzindo barreiras entre código experimental e produção de resultados.

Uma das principais ideias desse processo é facilitar a revisão por pares e execução em diferentes plataformas. Dessa forma, a escolha por ferramentas abertas e padronizadas reforçou a transparência metodológica exigida em pesquisas acadêmicas contemporâneas.

I. Ambiente de desenvolvimento

- **Jupyter Notebook** — ambiente interativo que reúne código, texto descritivo e saídas gráficas em um único documento, facilitando documentação, revisão e compartilhamento dos experimentos.
- **Python 3.12** — versão estável da linguagem que oferece ampla compatibilidade com bibliotecas científicas e recursos atualizados de tipagem e desempenho.

II. Bibliotecas de análise de dados

- **Pandas 2.0+** — estrutura de dados DataFrame para carregamento, limpeza, agregação e junção de tabelas; forneceu funções de filtragem e transformação essenciais na etapa de pré-processamento.
- **NumPy** — base numérica que opera com matrizes n-dimensionais, possibilitando vetorização e ganho expressivo de desempenho em cálculos de álgebra linear.
- **Matplotlib** — biblioteca de baixo nível utilizada para gerar figuras estáticas de qualidade de publicação, ajustando eixos, rótulos e estilos conforme padrões acadêmicos.
- **Seaborn** — camada de alto nível sobre o Matplotlib, adotada para gráficos estatísticos (boxplots, heatmaps, distribuições) com paletas padronizadas e menor necessidade de configuração manual.

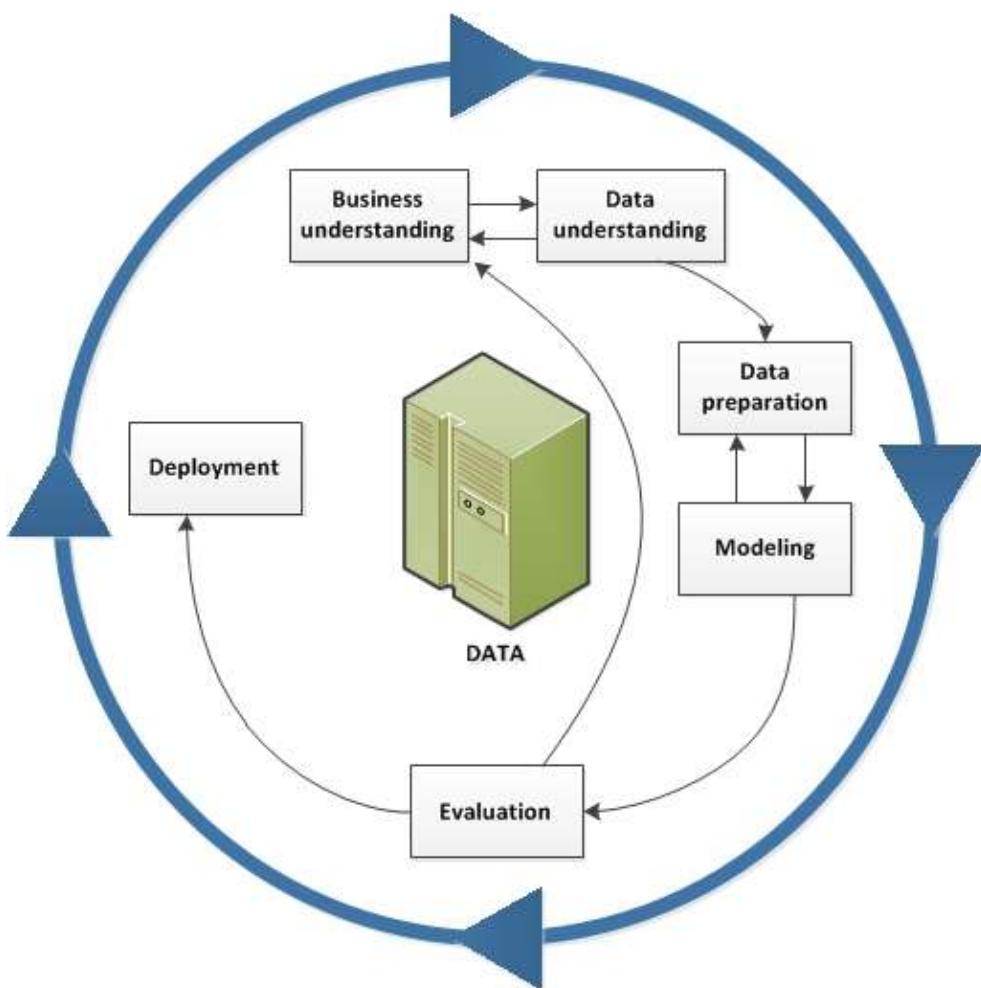
III. Bibliotecas de Aprendizado de Máquina (*Machine Learning*)

- **Scikit-learn** — coleção consolidada de algoritmos de classificação e regressão, validação cruzada, seleção de atributos e métricas de desempenho; permitiu construir *pipelines* completos de modelagem com poucas linhas de código.
- **Joblib** — ferramenta empregada para serializar e restaurar modelos treinados, garantindo persistência eficiente em disco e reutilização em sessões posteriores sem perda de desempenho.

3.2 PROCESSO CRISP-DM ADAPTADO PARA ANÁLISE ACADÊMICA

A condução deste estudo seguiu a metodologia **CRISP-DM** (*Cross-Industry Standard Process for Data Mining*), amplamente empregado em projetos de ciência de dados. Por se tratar de uma investigação acadêmica, o ciclo original — composto por seis fases — foi adaptado com um momento de documentação formal, garantindo rastreabilidade e reproduzibilidade. A Figura X sintetiza o fluxo operacional, enquanto as subseções a seguir detalham cada estágio, conforme proposto pela IBM (2021) e a Tabela X na sequência conforme adaptação do autor.

Figura 22 – Ciclo de vida da mineração de dados.



Fonte: (IBM, 2021)

Tabela 12 – Processo CRISP-DM adaptado.

Fase	Objetivo	Principais atividades executadas
Compreensão do problema	Definir metas e critérios de sucesso	Clara formulação da pergunta de pesquisa: <i>qual é o risco de diabetes em novos pacientes?</i> • Seleção das métricas-chave (<i>AUC, F1-Score, Recall</i>) e critérios de performance.
Compreensão dos dados	Avaliar qualidade e estrutura do conjunto informacional	Carregamento do <i>Pima Indians Diabetes Dataset</i> (768 registros). Estatísticas descritivas (média, mediana, desvio-padrão). Inspeção de valores ausentes, zeros implausíveis e análise de correlação. Geração de histogramas, <i>boxplots</i> e gráficos para análise exploratória de dados.
Preparação dos dados	Transformar o conjunto bruto em formato apto à modelagem	Análise comparativa de técnicas para tratamento de valores zero. Divisão estratificada em treino (60%), validação (20%) e teste (20%). Remoção de outliers do conjunto de treino via critério IQR (1.5). Balanceamento da classe minoritária com a técnica SMOTE. Normalização final dos dados.
Modelagem	Construir e ajustar algoritmos preditivos	Implementação de 9 algoritmos de classificação com parâmetros pré-definidos para estabelecimento de uma linha de base comparativa..
Avaliação	Medir desempenho e verificar generalização	Cálculo de métricas (<i>AUC, F1-Score, Recall</i> , entre outras) para todos os modelos. Seleção dos 4 melhores com base em critérios de AUC e Recall. Otimização de <i>threshold</i> para os modelos selecionados. Validação qualitativa final com perfis sintéticos.
Documentação	Garantir reproduzibilidade e auditoria	Jupyter Notebooks comentados passo a passo, exportados em PDF para anexos. Registro de parâmetros e ambiente de execução (<i>requirements.txt</i>).

Fonte: elaborado pelo autor.

3.3 ANÁLISE EXPLORATÓRIA DE DADOS (AED)

Foram analisados os 768 registros do conjunto *Pima Indians Diabetes Dataset* para compreender a estrutura, as distribuições e as relações entre variáveis.

Inicialmente, verificaram-se os tipos de dados, a contagem de valores não nulos e as estatísticas descritivas (média, mediana, desvio-padrão). Em seguida, geraram-se histogramas de cada *feature* e um gráfico de barras da variável *Outcome* para avaliar o desbalanceamento de classes (65,1% não diabéticos / 34,9% diabéticos). Calculou-se uma matriz de correlação e plotou-se um *heatmap* para quantificar as relações lineares, complementado por gráficos de dispersão entre *Glucose* e outras variáveis de relevância clínica, como *BMI*, *Age* e *Insulin*.

Valores zero em *Glucose*, *BloodPressure* e *BMI* foram tratados como potenciais dados ausentes, por não serem fisiologicamente plausíveis; registrou-se a frequência desses zeros para embasar a etapa subsequente de pré-processamento.

As informações extraídas nesta fase subsidiaram as decisões metodológicas da etapa de pré-processamento.

3.4 ETAPAS DO PRÉ-PROCESSAMENTO DE DADOS

O conjunto de dados bruto foi submetido a um rigoroso pipeline de pré-processamento, aplicado sequencialmente para garantir a qualidade e a integridade dos dados antes da modelagem.

3.4.1 Tratamento de valores zero em variáveis fisiológicas

Inicialmente, a análise exploratória de dados (AED) identificou um achado crítico: a presença de valores iguais a zero em múltiplas variáveis fisiológicas. Foi necessário contextualizar esses valores, pois enquanto um zero é válido para a variável *Pregnancies*, indicando ausência de gestações anteriores, o mesmo não se aplica a outras medições. Valores zero para *Glucose*, *BloodPressure* e *BMI* são biologicamente ou fisicamente implausíveis e incompatíveis com a vida,

sendo, portanto, interpretados como dados ausentes mascarados. Da mesma forma, valores zero para *SkinThickness* e *Insulin* foram considerados indicativos de dados não mensurados.

Diante da presença destes zeros implausíveis, foi conduzida uma análise metodológica para determinar a melhor estratégia de tratamento. Compararam-se cinco abordagens distintas:

- a) Utilização dos dados brutos, sem alteração;
- b) Exclusão completa dos registros que continham valores zero;
- c) Imputação pela mediana;
- d) Imputação por k-Nearest Neighbors ($k=5$);
- e) Imputação por regressão linear.

Cada estratégia foi aplicada sobre uma cópia do conjunto de treino. Em seguida, para avaliar o impacto de cada abordagem, foi realizada uma análise dupla. A avaliação qualitativa consistiu na inspeção visual da dispersão dos valores e da forma da distribuição, por meio de histogramas, *boxplots* e gráficos de dispersão, para verificar a coerência dos dados gerados por cada técnica.

Para complementar esta análise visual com uma justificativa quantitativa, um modelo de *Random Forest* foi treinado rapidamente em cada dataset pré-processado, utilizando os hiperparâmetros: *criterion='gini'*, *n_estimators=100*, *max_depth=10* e *random_state=42*. As métricas de AUC, F1-Score e Recall foram calculados para se ter uma estimativa inicial da capacidade preditiva. Esse processo completo permitiu uma avaliação tanto do impacto na distribuição dos dados quanto do potencial preditivo de cada abordagem.

3.4.2 Divisão e estratificação dos dados

O conjunto de dados foi dividido em subconjuntos de treinamento (60%), validação (20%) e teste (20%), utilizando amostragem estratificada para preservar a proporção de classes em todos eles. Para garantir a reproduzibilidade, o processo foi executado com um estado aleatório fixo (*random_state=42*).

3.4.3 Remoção de outliers

A detecção e remoção de outliers foi realizada exclusivamente no conjunto de treinamento, utilizando o critério do Intervalo Interquartil (IQR) com fator 1,5. Para cada variável numérica, calcularam-se o primeiro quartil (Q_1) e o terceiro quartil (Q_3); o IQR foi obtido por:

$$(IQR = Q_3 - Q_1).$$

Observações com valores inferiores a

$$\text{Limite Inferior} = Q_1 - 1,5 \cdot IQR$$

Ou superiores a

$$\text{Limite Superior} = Q_3 + 1,5 \cdot IQR$$

Foram consideradas outliers e removidas.

3.4.4 Balanceamento das classes

Para corrigir o desbalanceamento de classes no conjunto de treinamento, foi utilizada a técnica SMOTE (*Synthetic Minority Over-sampling Technique*), escolhida por sua capacidade de criar amostras sintéticas em vez de apenas duplicá-las. A técnica de *Random Oversampling* foi descartada previamente com base na revisão de literatura, devido ao seu elevado risco de causar *overfitting*. O SMOTE foi aplicado para igualar o número de amostras da classe minoritária ao da classe majoritária, resultando em um conjunto de treinamento final perfeitamente balanceado.

3.4.5 Normalização dos dados

Como etapa final, os dados foram normalizados pela técnica de Padronização (*Standardization*) (Z-score). O *Scaler* foi ajustado (*fit*) exclusivamente com os dados de treinamento e, em seguida, foi usado para transformar (*transform*) os três conjuntos (treino, validação e teste).

3.5 MODELAGEM E AVALIAÇÃO

3.5.1 Treinamento dos modelos de classificação

Foram implementados e treinados nove algoritmos de classificação supervisionada. Para garantir uma comparação de base justa, os modelos foram instanciados com os seguintes hiperparâmetros, definidos com base em artigos de referência e padrões da biblioteca *Scikit-learn* (Python):

Tabela 13 – Modelos e parâmetros utilizados no treinamento.

Modelo	Hiperparâmetros Utilizados
Random Forest	<i>criterion='gini', n_estimators=100, max_depth=10, random_state=42</i>
Gradient Boosting	<i>criterion='friedman_mse', learning_rate=0.1, max_depth=3, n_estimators=100, subsample=1.0, random_state=42</i>
XGBoost	<i>booster='gbtree', learning_rate=0.3, max_depth=6, n_estimators=100, random_state=42</i>
Decision Tree	<i>criterion='gini', max_depth=10, min_samples_leaf=1, min_samples_split=2, random_state=42</i>
AdaBoost	<i>learning_rate=1, n_estimators=100, random_state=42</i>
SVM	<i>C=1.0, cache_size=200, coef0=0.0, kernel='rbf', max_iter=-1, random_state=42</i>
Logistic Regression	<i>C=1.0, max_iter=1000, tol=0.0001, random_state=42</i>
kNN	<i>leaf_size=30, metric='minkowski', n_neighbors=5</i>
Naive Bayes	<i>var_smoothing=1e-09</i>

Fonte: (AKMEŞE, 2022).

3.5.2 Métricas de processo de avaliação

O desempenho dos modelos foi mensurado por um conjunto de métricas, incluindo Acurácia, Precisão, Especificidade, F1-Score e, com especial atenção, a Área Sob a Curva ROC (AUC) e o Recall (Sensibilidade). O processo de avaliação e seleção foi estruturado em três estágios:

- I. Seleção por performance: Neste estágio inicial, o critério primário para a seleção dos modelos foi um AUC superior a 0.80 nos conjuntos de validação e teste. O *Recall* e a robustez na validação cruzada de 5 *folds* (*CV-5*) foram usados como critérios secundários.
- II. Otimização de *threshold*: Os modelos que avançaram foram submetidos a uma otimização do limiar de classificação, buscando o ponto que maximizasse o F1-Score como métrica de equilíbrio entre Precisão e Recall.
- III. Validação qualitativa: Como critério final e decisivo, os modelos otimizados foram avaliados contra perfis clínicos sintéticos para analisar a plausibilidade e a coerência de suas predições.

4 RESULTADOS E DISCUSSÕES

Este capítulo apresenta os resultados obtidos ao longo do estudo, desde a análise exploratória inicial dos dados até a avaliação final do modelo preditivo de diabetes. Cada seção discute os achados, contextualizando-os no âmbito do problema e justificando as decisões metodológicas tomadas nas etapas subsequentes do trabalho.

4.1 ANÁLISE EXPLORATÓRIA DE DADOS

A primeira etapa da investigação consistiu em uma Análise Exploratória de Dados (AED) para compreender a estrutura, as distribuições e as relações presentes no conjunto de dados. O *dataset* é composto por 768 registros e 9 variáveis (features), onde 8 são variáveis preditoras e 1 é a variável-alvo, *Outcome*, que indica o diagnóstico de diabetes (0 para não diabético, 1 para diabético).

Tabela 14 – Descrição das variáveis do conjunto de dados/dataset.

Variável	Tipo	Not null	Valores Únicos	Exemplo
Pregnancies	int64	768	17	6
Glucose	int64	768	136	148
BloodPressure	int64	768	47	72
SkinThickness	int64	768	51	35
Insulin	int64	768	186	0
BMI	float64	768	248	33.6
DiabetesPedigreeFunction	float64	768	517	0.627
Age	int64	768	52	50
Outcome	int64	768	2	1

Fonte: elaborado pelo autor

Verificou-se a análise da integridade de dados, isso inclui a análise de valores que, embora numericamente válidos, são semanticamente ou biologicamente implausíveis. Neste conjunto de dados foi investigada a ocorrência de valores iguais a zero que podem representar tanto uma medição real quanto um dado ausente ou mascarado. A Tabela 13 a seguir reúne os principais achados.

Tabela 15 - Análise de ocorrência de valores zero por variável

Variável	Qty. de Zeros	(%)	Status e Justificativa
Pregnancies	111	14.5	● Normal: Valor zero indica ausência de gestações anteriores.
Glucose	5	0.7	● Suspeito: Níveis de glicose zero são biologicamente inviáveis.
BloodPressure	35	4.6	● Suspeito: Pressão arterial zero é incompatível com a vida.
SkinThickness	227	29.6	● Suspeito: Espessura da dobra cutânea zero é improvável.
Insulin	374	48.7	● Suspeito: Pode representar um valor não medido ou muito baixo.
BMI	11	1.4	● Suspeito: IMC zero é fisicamente impossível.
Outcome	500	65.1	● Normal: Valor zero representa a classe "não diabético".

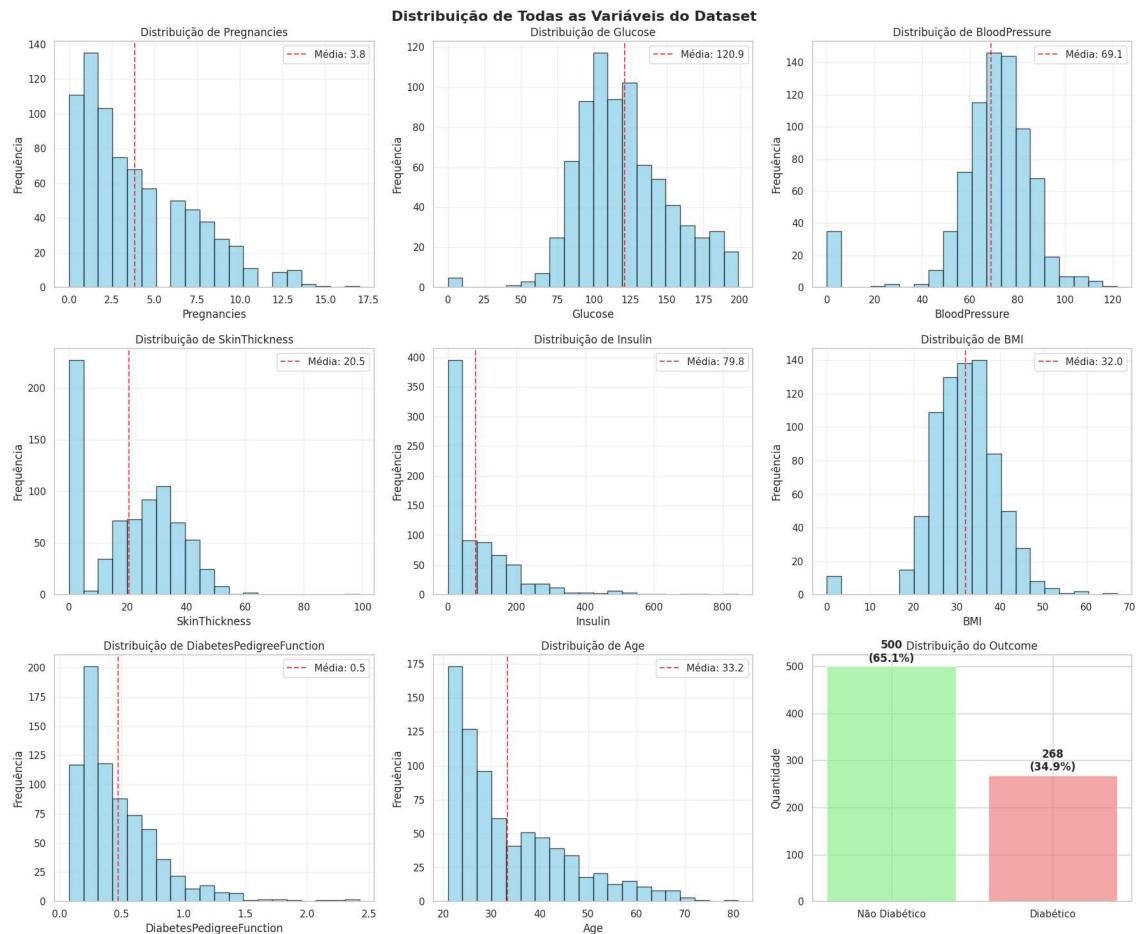
Fonte: elaborado pelo autor.

A análise revelou um achado crítico: variáveis fisiológicas fundamentais, como Glucose, BloodPressure, SkinThickness, Insulin e BMI, contavam com uma quantidade significativa de valores zero. A presença desses zeros foi um forte indicativo de que foram utilizados para preencher dados ausentes (nulos) durante a coleta ou tabulação. Esses valores não representavam medições reais e, quando não tratados, podiam distorcer severamente as análises estatísticas e o desempenho dos modelos de machine learning.

Ficou evidente, portanto, a necessidade de uma etapa de pré-tratamento para esses dados.

A Figura 21 a seguir exibe as distribuições de frequência de todas as variáveis. Esta análise visual é um passo fundamental da exploração de dados para se observar a forma, a tendência central e a dispersão geral de cada atributo do conjunto de dados.

Figura 23 - Histogramas de frequência e distribuição de classes para todas as variáveis do dataset.

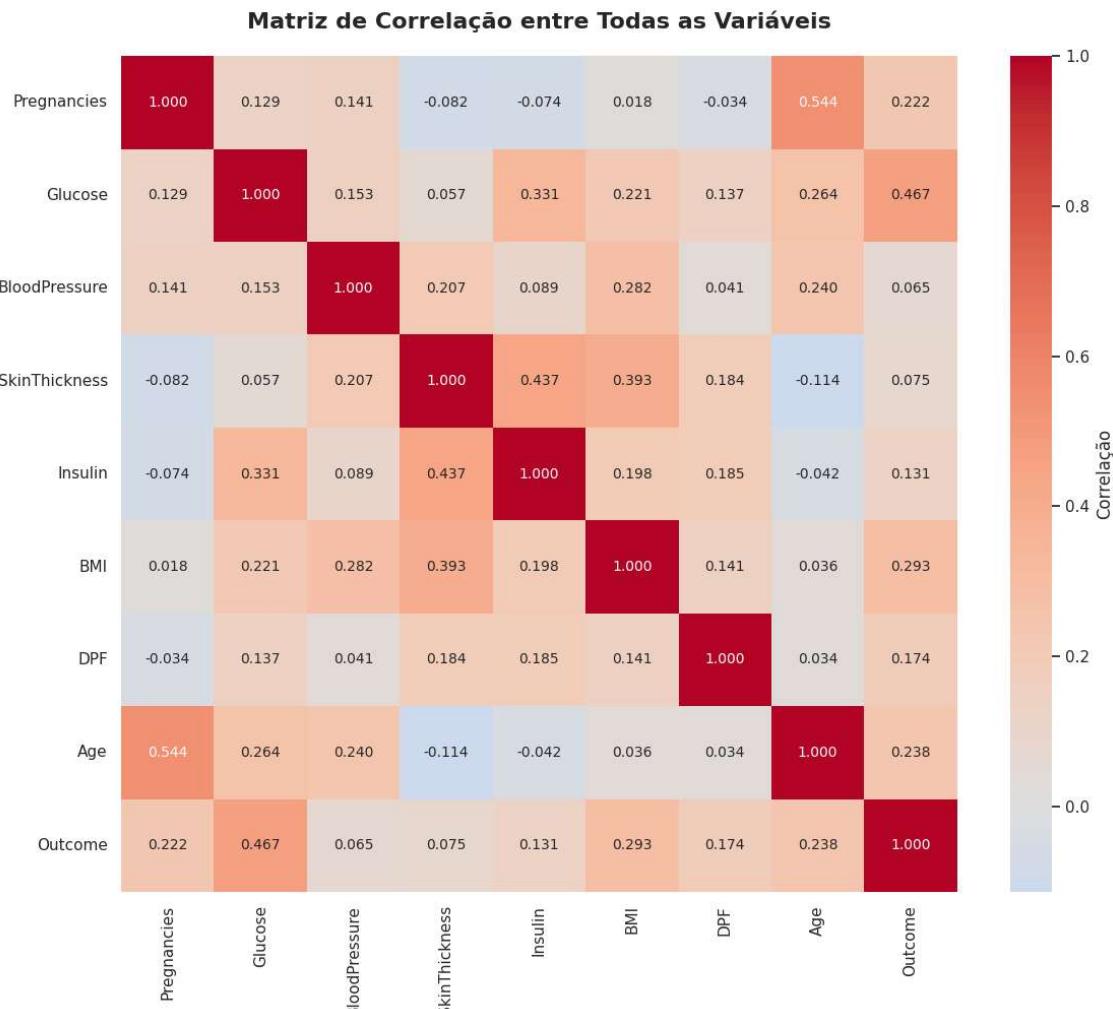


Fonte: elaborado pelo autor.

A análise das distribuições revela perfis distintos: variáveis como Glucose, BloodPressure e BMI são aproximadamente simétricas, enquanto Age, Insulin, Pregnancies e outras demonstram forte assimetria à direita. O achado mais relevante, contudo, é o desbalanceamento da variável-alvo Outcome, que possui 65,1% de casos não diabéticos (n=500) contra 34,9% de diabéticos (n=268). Essa desproporção é um fator crítico que deve ser considerado na etapa de pré-processamento e balanceamento da modelagem.

Para uma visão macro das inter-relações lineares foi gerada uma matriz de correlação, conforme a Figura 21 a seguir.

Figura 24 – Matriz de correlação entre todas as variáveis.

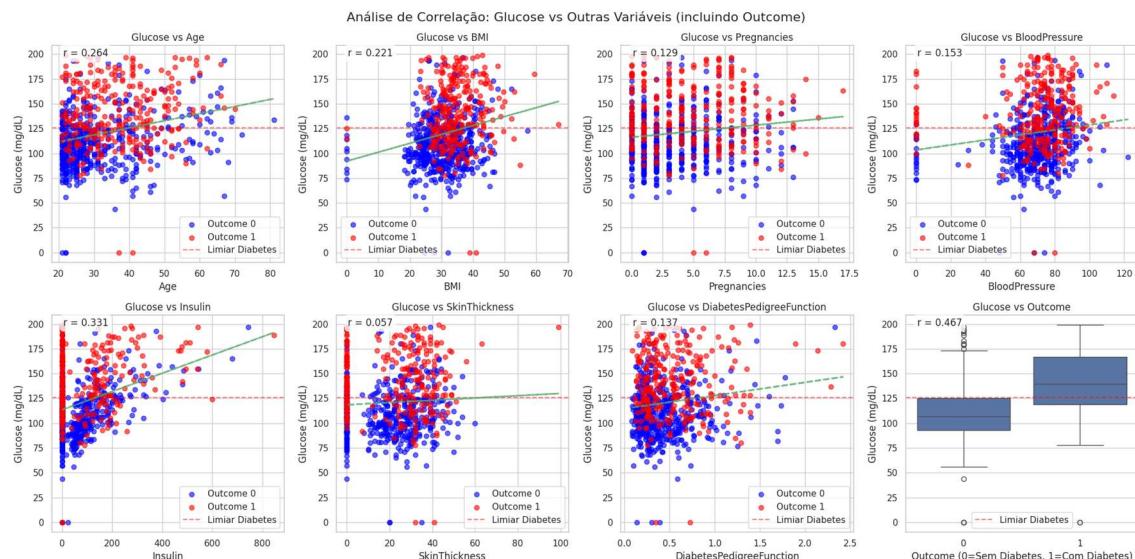


Fonte: elaborado pelo autor.

A matriz revela que a variável Glucose possui a mais forte correlação com a variável alvo *Outcome* ($r=0.467$), seguida por BMI ($r=0.293$), Age ($r=0.238$) e *Pregnancies* ($r=0.222$). O que está de acordo com o perfil de risco citado pela SBD (2024). A Figura 2 oferece uma visualização focada, ordenando as variáveis pela força de sua correlação com *Glucose*, confirmando que o *Outcome* é o mais fortemente associado.

Para além dos coeficientes, a análise dos gráficos de dispersão (Figura 23) a seguir permite uma inspeção visual dessas relações. Os gráficos mostram a Glicose em relação a outras variáveis chave. Nota-se que, embora exista uma tendência positiva (a linha de regressão é ascendente) em variáveis como *Age*, *BMI*, *Insulin* e *Pregnancies*, há uma dispersão considerável dos pontos. Isso indica que, embora a correlação exista, ela não é fortemente linear e a variabilidade individual é alta, reforçando a necessidade de modelos mais complexos que o linear para capturar tais padrões. Os pontos em vermelho (diabéticos) tendem a se concentrar em níveis mais altos de glicose em todos os gráficos.

Figura 25 - Análise de correlação/dispersão - glicose vs. outras variáveis.

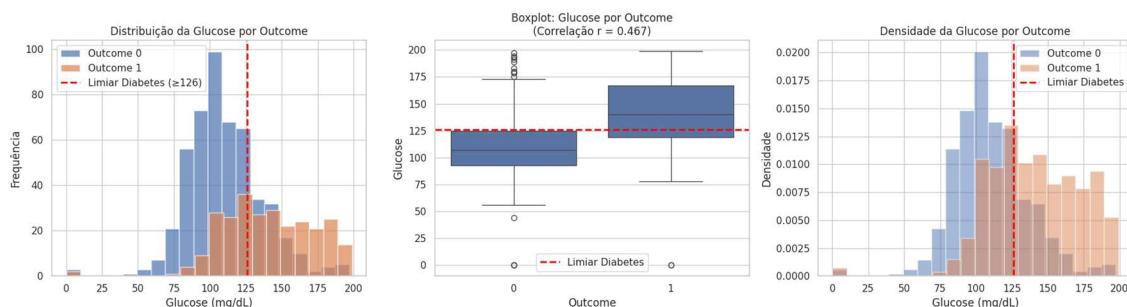


Fonte: elaborado pelo autor.

Adicionalmente, a análise visual destes gráficos de dispersão torna evidente o problema dos valores nulos implausíveis, previamente apontado na Tabela 2. Nos gráficos "Glucose vs Insulin" e "Glucose vs SkinThickness", é notável a grande concentração de pontos de dados alinhados verticalmente sobre o valor zero no eixo horizontal. Essa aglomeração representa os numerosos casos em que Insulin e SkinThickness foram registrados como zero, reforçando visualmente que estes são dados anômalos e sublinhando a necessidade imperativa de seu tratamento na etapa de pré-processamento.

A análise da variável *Glucose*, ilustrada na Figura 24, confirmou sua relevância. Histogramas e gráfico de densidade evidenciaram uma separação clara entre os grupos diabético e não diabético. O grupo com diabetes apresentou valores mais elevados, com distribuição deslocada à direita. O boxplot reforçou esse padrão, mostrando a mediana da glicose acima de 126 mg/dL para diabéticos e bem abaixo desse valor para não diabéticos.

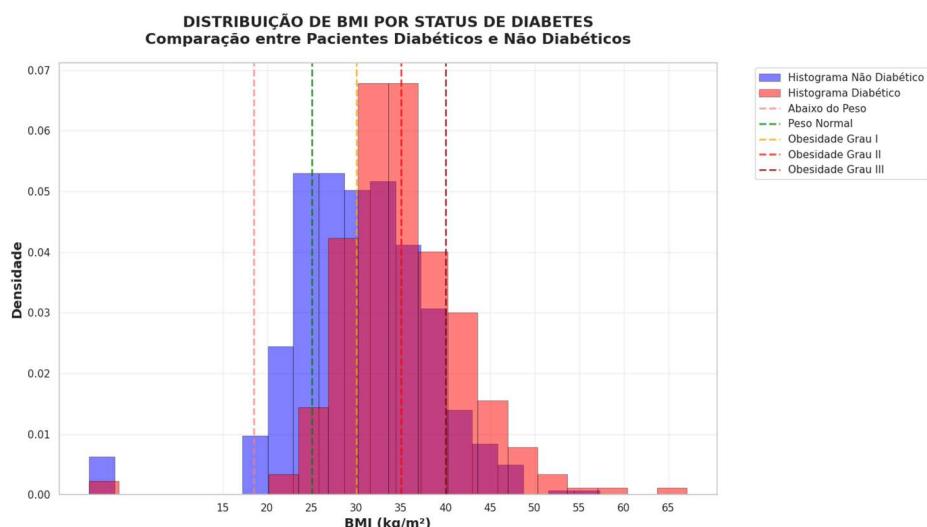
Figura 26 - Análise da Distribuição da glicose por desfecho/outcome.



Fonte: elaborado pelo autor.

De forma similar, o *BMI* (IMC) demonstrou ser um fator distintivo. A Figura 3 compara a distribuição de densidade do IMC entre os grupos. Observa-se que a distribuição para pacientes diabéticos (em vermelho) está deslocada para a direita, indicando uma tendência a valores de IMC mais altos neste grupo.

Figura 27 – Distribuição de *BMI*/IMC por status de diabetes.



Fonte: elaborado pelo autor.

A análise estatística corrobora esta visualização. O grupo diabético apresentou uma média de IMC de 35,14 kg/m², enquanto o grupo não diabético teve uma média de 30,30 kg/m² (ainda alto, sendo considerado como obesidade grau I). Ao categorizar os pacientes segundo as faixas de IMC, a relação entre obesidade e diabetes torna-se ainda mais evidente, como mostra a Tabela 14. A taxa de diabetes cresce progressivamente com o aumento do IMC, partindo de 6,9% em indivíduos com peso normal e alcançando 56,1% em pacientes com Obesidade Grau III.

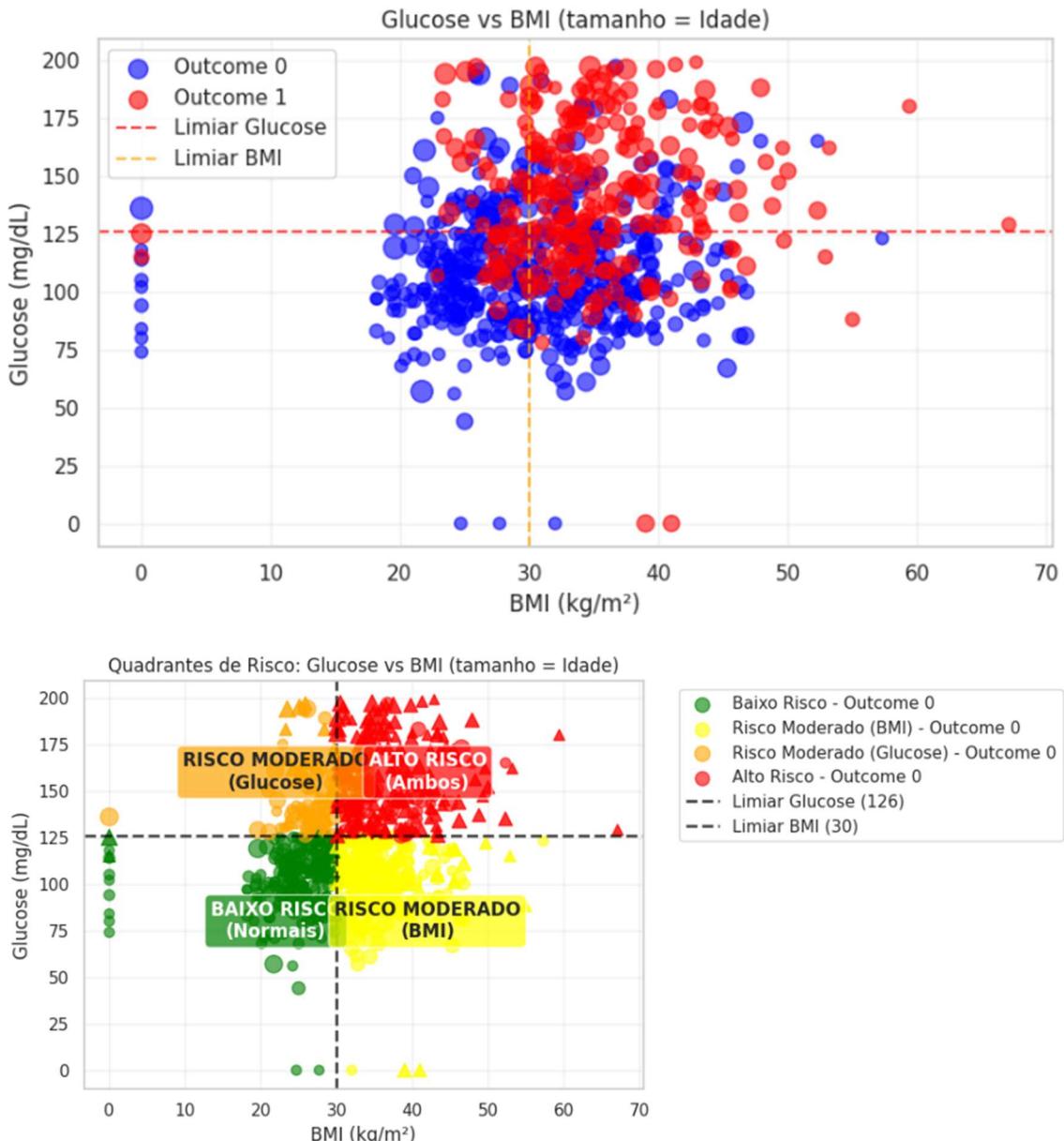
Tabela 16 – Taxa de Diabetes por classificação de BMI/IMC

Classificação	IMC	Total de Pacientes	Diabéticos	Taxa de Diabetes (%)
Abaixo do Peso	< 18.5	15	2	13.3
Peso Normal	18.5 - 24.9	102	7	6.9
Sobrepeso	25.0 - 29.9	179	40	22.3
Obesidade Grau I	30.0 - 34.9	224	101	45.1
Obesidade Grau II	35.0 - 39.9	150	63	42.0
Obesidade Grau III	≥ 40.0	98	55	56.1

Fonte: elaborado pelo autor.

Para entender o efeito combinado dos principais fatores de risco, foi realizada uma análise de quadrantes utilizando os limiares clínicos para Glucose (≥ 126 mg/dL) e BMI (≥ 30 kg/m²), conforme a Figura 26. Os pacientes foram divididos em quatro grupos de risco: Baixo Risco (ambos os marcadores normais), Risco Moderado (apenas um marcador alterado) e Alto Risco (ambos os marcadores alterados). A taxa de diabetes foi de 11,6% no grupo de baixo risco, subiu para cerca de 26-30% com um fator de risco presente e atingiu 70,4% no grupo de alto risco, demonstrando um efeito sinérgico entre obesidade e hiperglicemia.

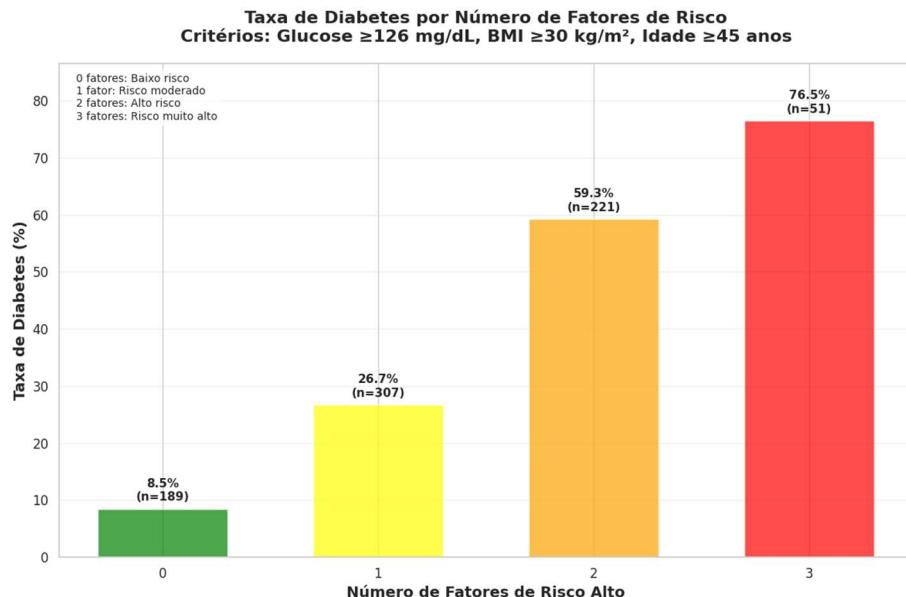
Figura 28 – Quadrantes de risco glicose vs IMC e taxa de diabetes associada.



Fonte: elaborado pelo autor

A taxa de diabetes foi de 11,6% no grupo de baixo risco, subiu para cerca de 26-30% com um fator de risco presente e atingiu 70,4% no grupo de alto risco, demonstrando um efeito sinérgico entre obesidade e hiperglicemia. Conforme a Figura 27.

Figura 29 – Taxa de diabetes por número de fatores de risco.



Fonte: elaborado pelo autor.

Principais achados da Análise Exploratória de Dados (AED):

- I. **Dados ausentes mascarados:** Foi identificada uma quantidade crítica de valores zero implausíveis em variáveis fisiológicas importantes como Glicose, Pressão Arterial, Espessura da Dobra Cutânea e IMC, indicando a presença de dados ausentes que exigiram de tratamento.
- II. **Desbalanceamento de classes:** Constatou-se um desbalanceamento significativo na variável-alvo, onde 65,1% da amostra corresponde a pacientes não diabéticos e 34,9% a diabéticos.
- III. **Principais fatores de correlação:** A Glicose se destacou como a variável de maior correlação com o diagnóstico de diabetes ($r=0,467$), seguida pelo IMC ($r=0,293$) e pela Idade ($r=0,238$).
- IV. **Efeito sinérgico de risco:** A análise evidenciou um forte efeito sinérgico, no qual a combinação de Glicose (≥ 126 mg/dL) e IMC (≥ 30 kg/m 2) elevados eleva a taxa de diabetes para 70,4%.

4.2 PRÉ-PROCESSAMENTO DO CONJUNTO DE DADOS

4.2.1 Análise comparativa das técnicas de pré-tratamento de dados

Reconhecendo que métricas de performance, isoladamente, podem ser insuficientes para validar um modelo de aplicação em saúde, foi estabelecida uma metodologia de teste baseada em perfis de pacientes. Foram criados seis perfis sintéticos, projetados para representar um espectro contínuo de risco para diabetes, desde um indivíduo saudável até um com diagnóstico severo.

Estes perfis, detalhados na Tabela 15, serviram como um teste de estresse e de coerência para todos os modelos finais gerados a partir das diferentes estratégias de pré-tratamento. A capacidade de um modelo estratificar corretamente o risco entre estes perfis foi utilizada como o critério decisivo para a seleção do pipeline final.

Tabela 17 - Perfis de risco sintéticos utilizados para análise de plausibilidade dos modelos.

Perfil de Risco	Pregnancies	Glucose	Blood Pressure	Skin Thickness	BMI	DPF	Age
Diabético Severo	4	200	100	50	40.0	1.50	65
Diabético Moderado	2	160	85	35	32.0	0.80	50
Pré-Diabético Alto	4	124	78	30	31.0	0.60	45
Pré-Diabético Baixo	2	115	75	25	30.0	0.40	40
Risco Baixo	0	99	70	22	29.0	0.30	35
Saudável	0	85	65	18	22.0	0.15	25

Fonte: elaborado pelo autor.

Para avaliar o impacto do tratamento de valores nulos e atípicos, notadamente os valores iguais a zero, cinco abordagens distintas foram investigadas, utilizando a variável *SkinThickness* como principal referência na análise por ser a feature com maior incidência de zeros a serem tratados. As estratégias foram: **(1) utilização dos dados brutos, (2) exclusão de features e observações, (3) imputação pela mediana, (4) imputação via K-Nearest Neighbors (KNN), e (5) imputação por Regressão Linear.**

Observou-se que a estratégia de pré-tratamento exerceu uma influência profunda sobre a distribuição dos dados e a importância das features. A abordagem de imputação por **K-Nearest Neighbors (KNN)**, por exemplo, demonstrou ser eficaz na criação de um conjunto de dados homogêneo. A análise dos boxplots (Figura 30) indica que a imputação por KNN gerou uma distribuição de dados plausível, eliminando os valores zero sem criar outliers artificiais, o KNN também teve uma boa homogeneidade em *Skinthickness* comparado aos outros modelos, conforme (Figura 29). Contudo, ao comparar a distribuição e a importância das features com o conjunto de dados original, não foram observadas alterações significativas que justificassem a complexidade computacional e a introdução de valores sintéticos por este método. A Glucose permaneceu como a feature de maior importância, seguida por BMI e *SkinThickness*, em uma ordem muito similar à original.

A imputação pela **mediana** foi analisada como uma alternativa mais simples. No entanto, esta abordagem foi preterida por sua tendência a distorcer a distribuição original das variáveis. Essa distorção é claramente visível nos histogramas de densidade (Figura 28), que exibem um pico artificial e proeminente no valor da mediana para as features imputadas. Os gráficos de dispersão para *SkinThickness* (Figura 29) reforçam essa observação, mostrando os valores imputados formando uma linha horizontal constante, o que reduz a variância natural do conjunto de dados.

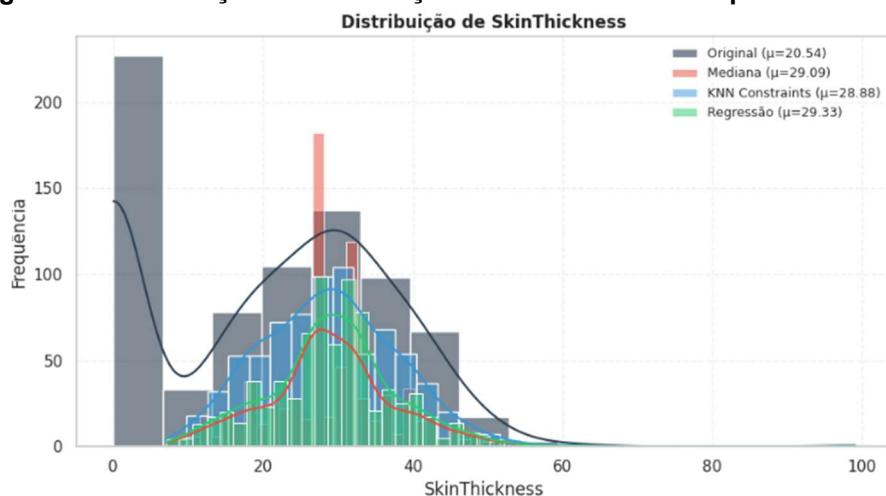
A imputação por **Regressão Linear**, uma técnica mais sofisticada que a mediana, também foi avaliada. Semelhante ao KNN, este método gerou distribuições de dados visualmente coerentes, conforme visto nos histogramas (Figura 28), gráficos de dispersão (Figura 29) e *boxplots* (Figura 30). No entanto, assim como o KNN, a análise de importância das features revelou um padrão

quase idêntico ao dos dados originais, com a Glucose e *SkinThickness* nas primeiras posições. Concluiu-se que o esforço para implementar e gerar dados por regressão não se traduziu em uma melhoria ou mudança estrutural significativa no *dataset* que justificasse sua adoção.

A quarta abordagem, de **exclusão de dados**, consistiu na remoção das colunas *Insulin* (29,6% de valores zero) e *SkinThickness* (48,7% de valores zero), seguida pela exclusão das linhas que ainda continham zeros. Esta técnica, embora simples, levou a uma redução substancial do volume amostral, diminuindo o poder estatístico do *dataset*. A perda de informação foi evidente, com as duas features desaparecendo dos ranqueamentos de importância e alterando o peso relativo das demais, o que justificou seu descarte.

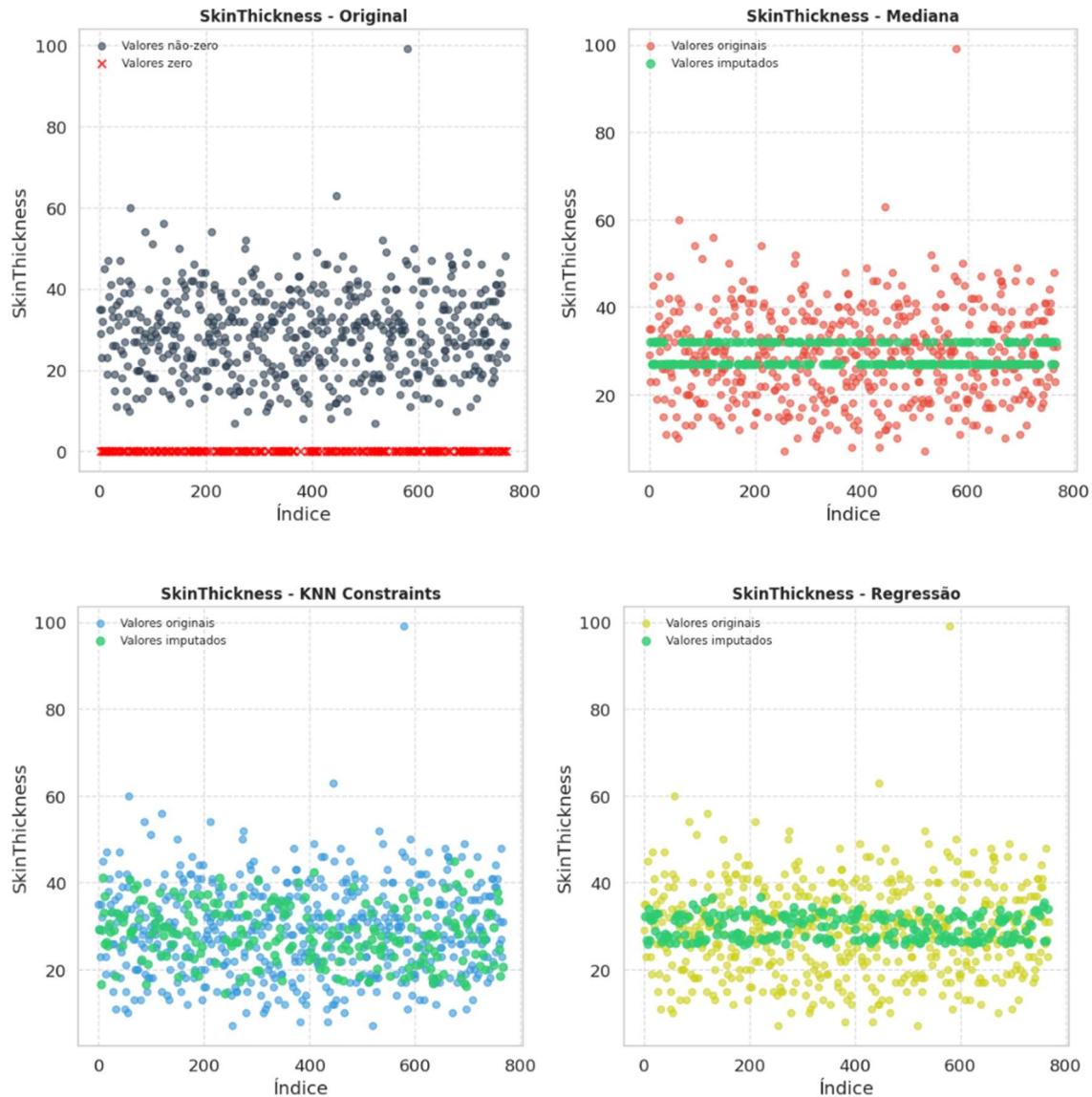
Diante deste cenário, optou-se por prosseguir com os **dados brutos**. A decisão fundamenta-se na premissa de preservar a máxima integridade e fidelidade dos dados originais. A análise de importância das features no *dataset* original já apontava claramente Glucose e BMI como os preditores mais relevantes. Portanto, concluiu-se que as abordagens de pré-tratamento, embora visualmente interessantes, não ofereceram uma vantagem prática que superasse o risco de introduzir artefatos ou a perda de informação.

Figura 30 – Diferenças na distribuição de acordo com cada processamento.



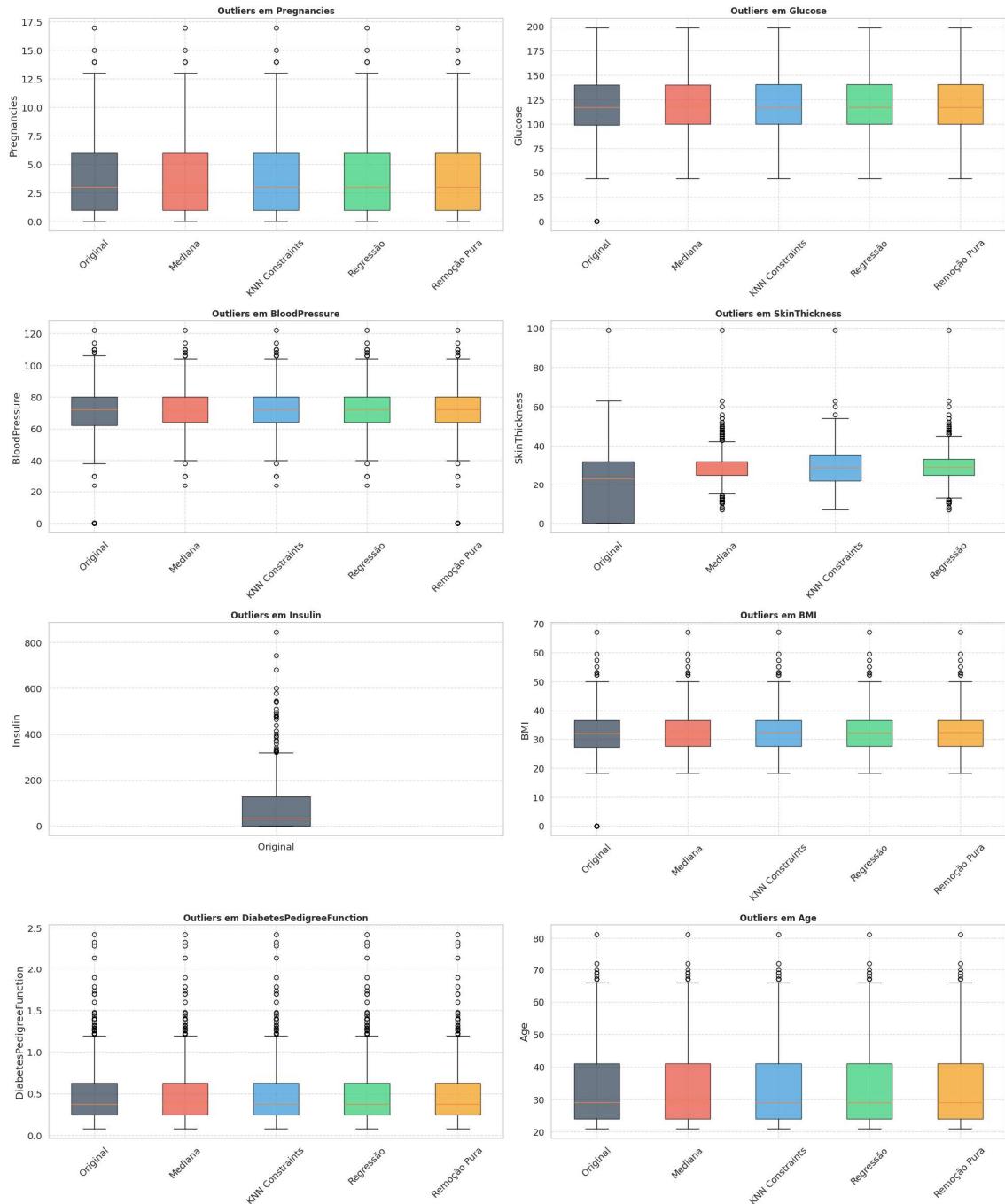
Fonte: elaborado pelo autor.

Figura 31 – Dispersão dos pontos inseridos de acordo com cada processamento.



Fonte: elaborado pelo autor.

Figura 32 – Boxplot de acordo com cada processamento



Fonte: elaborado pelo autor

Principais achados das estratégias de tratamento de dados:

- I. **Distorção por imputação simples:** A imputação pela mediana foi descartada por distorcer a distribuição dos dados, criando picos artificiais que reduzem a variância e podem ocultar padrões relevantes.
- II. **Ineficiência de métodos complexos:** As imputações por KNN e Regressão Linear, apesar de sofisticadas, não ofereceram melhorias que justificassem sua complexidade e o risco de gerar dados sintéticos.
- III. **Perda de poder estatístico por exclusão:** A remoção de colunas e linhas com dados zero, embora uma abordagem direta, causou uma perda crítica de informação e poder estatístico, enfraquecendo os modelos.
- IV. **Eficácia da abordagem robusta:** A combinação de dados brutos com um modelo robusto (Random Forest) foi a estratégia superior, lidando eficazmente com o ruído do dataset sem a necessidade de manipulação artificial dos dados.

4.2.2 Estratificação (Treino 80%, validação 20%, teste 20%)

Para a construção e avaliação rigorosa dos modelos preditivos, uma etapa metodológica fundamental foi a divisão do conjunto de dados brutos em subconjuntos independentes de treinamento, validação e teste. Este processo é crucial para garantir que a performance do modelo seja avaliada de forma imparcial, em dados que não foram utilizados durante sua fase de aprendizado, prevenindo assim o *overfitting* e o vazamento de dados (*data leakage*).

Adotou-se uma estratégia de divisão na proporção de **60% para treinamento, 20% para validação e 20% para teste**. Esta proporção foi escolhida por representar um equilíbrio robusto, alocando a maior parte dos dados para o aprendizado do modelo, enquanto reserva porções estatisticamente significativas para o ajuste de hiperparâmetros (validação) e para a avaliação final de sua capacidade de generalização (teste).

O processo foi executado em duas etapas sequenciais, utilizando a técnica de amostragem estratificada. A estratificação garante que a proporção original entre as classes da variável-alvo (34,9% de diabéticos e 65,1% de não

diabéticos) seja preservada em todos os subconjuntos. A metodologia foi a seguinte:

- **Isolamento do conjunto de teste:** Primeiramente, 20% do conjunto de dados total (154 amostras) foi separado para formar o conjunto de teste. Este conjunto permaneceu "intocável" durante todas as fases de treinamento e ajuste, sendo utilizado exclusivamente para a avaliação final de performance dos modelos já finalizados.
- **Subdivisão para treinamento e validação:** Os 80% restantes dos dados (614 amostras) foram então subdivididos, alocando-se 75% deste montante para o conjunto de treinamento (resultando em 60% do total, ou 460 amostras) e os 25% restantes para o conjunto de validação (resultando em 20% do total, ou 154 amostras).

Para garantir a reproduzibilidade dos resultados, o processo de divisão foi executado com um estado aleatório fixo (*random_state=42*). O resultado desta divisão estratificada está sumarizado na Tabela 16.

Tabela 18 – Estratificação dos dados.

Conjunto	Nº de Amostras	% do Total	Classe 1 (Diabético)	Classe 0 (Não Diabético)	Proporção (Classe 0 / 1)
Treinamento	460	60,0%	160	300	65,2% / 34,8%
Validação	154	20,0%	54	100	64,9% / 35,1%
Teste	154	20,0%	54	100	64,9% / 35,1%

Fonte: elaborado pelo autor.

Como observado na tabela, a estratificação foi bem-sucedida, mantendo o desbalanceamento de classes de forma consistente entre os três conjuntos.

Esta divisão rigorosa é a fundação que permite uma comparação científica justa entre as diferentes técnicas de tratamento do desbalanceamento de classes (*SMOTE*, *Oversampling* e Reponderação). Todas as técnicas partem de uma base idêntica:

- As manipulações para balanceamento (geração de dados sintéticos ou duplicação) são aplicadas **exclusivamente sobre o conjunto de treinamento (460 amostras)**.
- A técnica de reponderação ajusta os pesos do algoritmo **durante o treinamento** com este mesmo conjunto.
- Todos os modelos resultantes, independentemente da técnica de balanceamento utilizada, são avaliados de forma imparcial utilizando os **mesmos e inalterados conjuntos de validação (154 amostras) e teste (154 amostras)**.

Dessa forma, garante-se que as diferenças de performance observadas entre os modelos são devidas à eficácia de cada técnica de balanceamento, e não a variações casuais na composição dos dados de avaliação.

Principais achados da estratificação dos dados:

- I. **Representatividade das amostras:** A divisão estratificada foi bem-sucedida em replicar a proporção de classes desbalanceada (aproximadamente 65,1% / 34,9%) de forma consistente entre os conjuntos de treinamento, validação e teste, garantindo a representatividade estatística em todas as etapas do projeto.
- II. **Base de avaliação imparcial:** A separação prévia e o isolamento do conjunto de teste (20% da amostra) estabeleceram uma base de avaliação final “intocável” e livre de vazamento de dados (*data leakage*), assegurando uma medição honesta da capacidade de generalização dos modelos.

4.2.3 Remoção de outliers (IQR) e

Após a estratificação dos dados, foi realizado o tratamento de outliers com o objetivo de remover observações extremas do conjunto de treinamento que poderiam impactar negativamente a performance e a capacidade de generalização dos modelos. A remoção foi aplicada exclusivamente sobre este conjunto para evitar o vazamento de informações (*data leakage*) e garantir uma avaliação imparcial nos dados de validação e teste.

A técnica selecionada foi o método padrão do **Intervalo Interquartil (IQR)**. Este método estatístico define os limites de normalidade dos dados utilizando um fator multiplicativo de **1.5** sobre o intervalo entre o primeiro quartil (Q1) e o terceiro quartil (Q3). As fórmulas para a definição dos limites de outliers foram:

$$IQR = Q3 - Q1$$

$$Limite_{inferior} = Q1 - 1.5 \times IQR$$

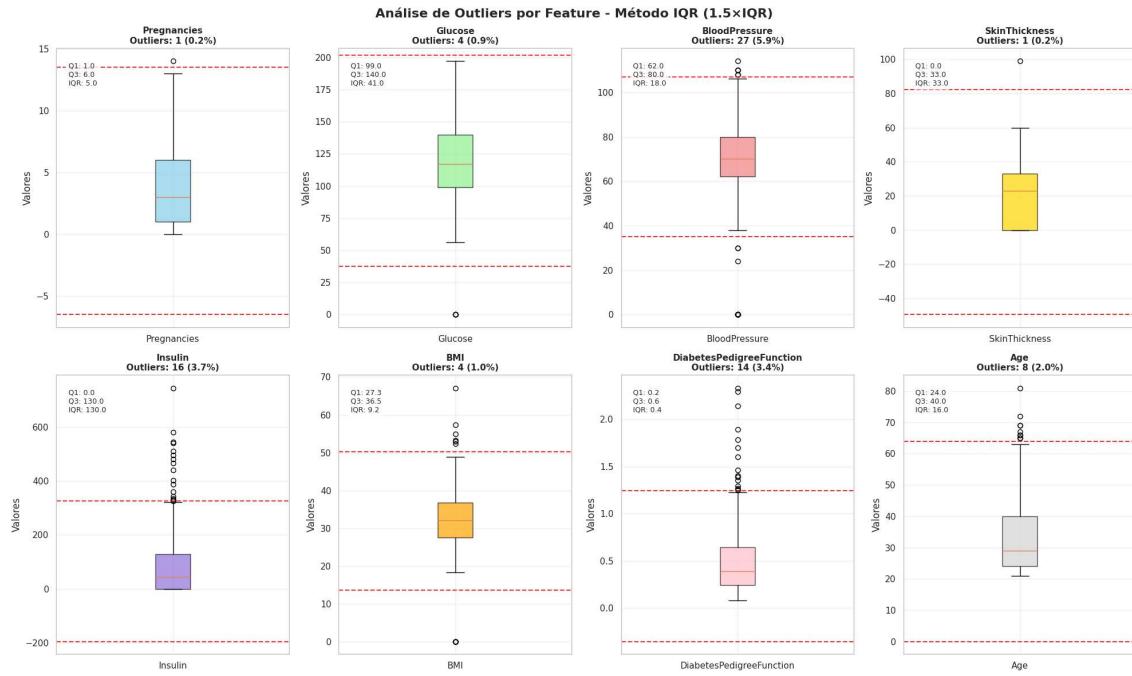
$$Limite_{superior} = Q3 + 1.5 \times IQR$$

A aplicação desta metodologia padrão no conjunto de treinamento resultou na identificação e remoção de 75 outliers, o que corresponde a 16,3% das 460 amostras originais. Consequentemente, o conjunto de dados de treinamento foi reduzido para 385 amostras. A Figura 31 ilustra os limites calculados pelo método e os pontos identificados como outliers em cada variável, enquanto a Figura 32 sumariza quantitativamente o impacto da remoção por feature.

Embora a remoção de 16,3% dos dados represente uma intervenção significativa, ela foi considerada uma etapa necessária para aumentar a robustez do modelo. O processo foi eficaz na eliminação de valores extremos e biologicamente implausíveis (como Glicose ou Pressão Arterial iguais a zero) que poderiam enviesar o processo de aprendizado. Desta forma, o tratamento resultou em um conjunto de treinamento mais limpo e representativo da

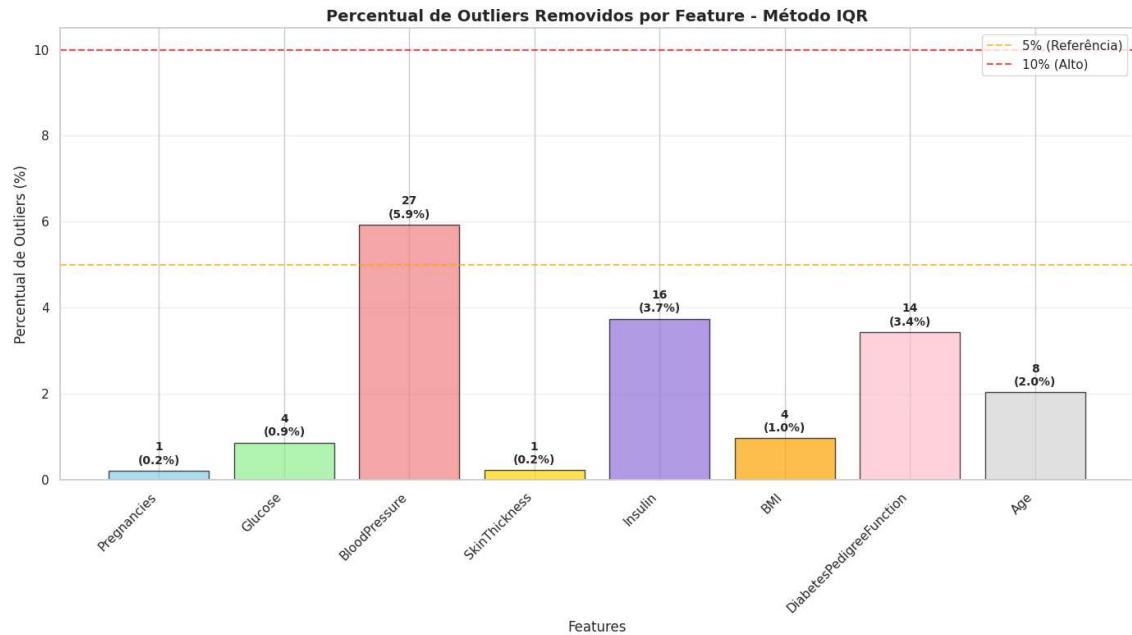
normalidade dos dados, preparando-o adequadamente para a subsequente etapa de balanceamento de classes.

Figura 33 – Boxplot de remoção de outliers com ($1.5 \times IQR$).



Fonte: elaborado pelo autor.

Figura 34 - Percentual de outliers removidos por feature.



Fonte: elaborado pelo autor.

4.2.4 Balanceamento: SMOTE

Após a limpeza dos outliers, o conjunto de treinamento, composto por 427 amostras, ainda apresentava um desbalanceamento de classes (65% Não Diabético / 35% Diabético). Para mitigar o risco de o modelo desenvolver um viés em favor da classe majoritária, foi aplicada a técnica de balanceamento *SMOTE* (*Synthetic Minority Over-sampling TTechnique*).

A técnica SMOTE foi escolhida por sua abordagem sofisticada, que, em vez de simplesmente duplicar amostras da classe minoritária, cria amostras sintéticas. O algoritmo opera selecionando uma amostra da classe minoritária e gerando um novo ponto de dados sintético em algum lugar ao longo do segmento de linha que une essa amostra a um de seus vizinhos mais próximos da mesma classe. Este processo foi aplicado exclusivamente ao conjunto de treinamento para criar um ambiente de aprendizado balanceado, sem contaminar os conjuntos de validação e teste.

A aplicação do SMOTE aumentou o número de amostras da classe minoritária (Diabético) de 149 para 278, igualando-se ao número de amostras da classe majoritária. Isso resultou em um novo conjunto de treinamento, agora com 556 amostras perfeitamente balanceadas (50% / 50%). A análise estatística da Tabela 17 confirma que o SMOTE gerou dados sintéticos que preservaram as características centrais da classe minoritária original.

Tabela 19 - Comparativo Estatístico da Classe Minoritária (Diabético) Antes e Depois do SMOTE

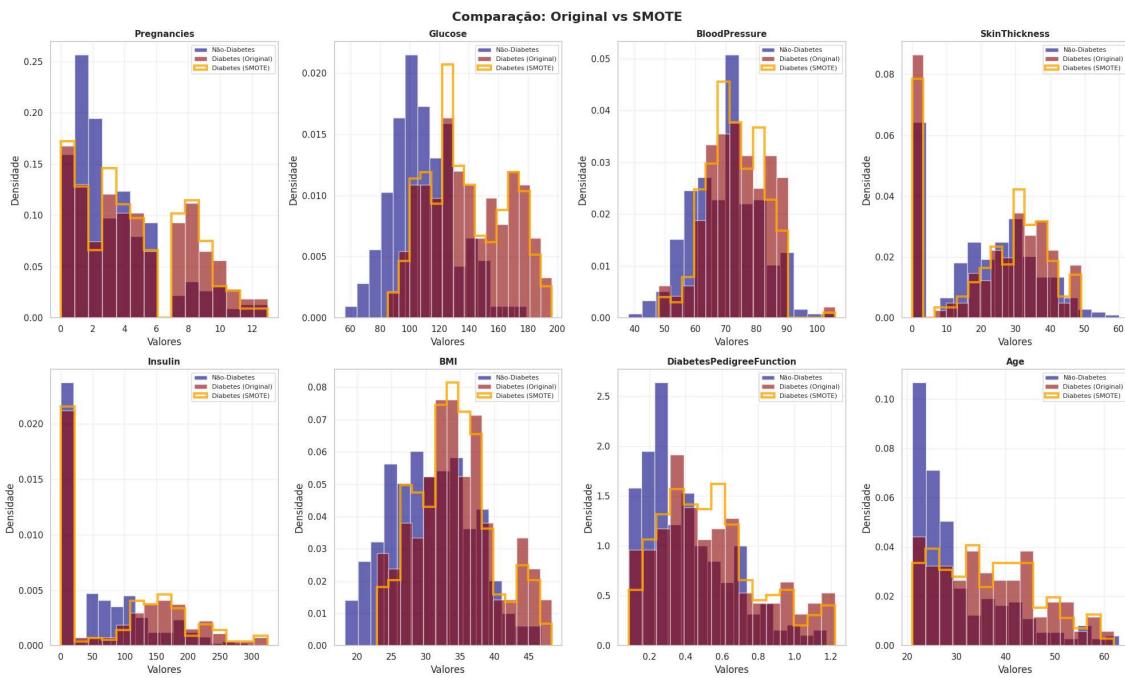
Feature	Média Original	Média SMOTE	Desvio Padrão Original	Desvio Padrão SMOTE
Glucose	140.28	138.56	32.57	27.19
BMI	35.42	34.15	7.55	5.47
Age	36.58	36.67	11.10	10.03
BloodPressure	70.89	73.08	20.45	9.29
Insulin	103.42	78.90	129.17	93.37

Fonte: elaborado pelo autor.

Observa-se na tabela que as médias das principais variáveis permaneceram notavelmente estáveis após a aplicação da técnica, indicando que a tendência central dos dados foi mantida. Houve uma redução geral no desvio padrão, um comportamento esperado do SMOTE, que tende a gerar amostras mais concentradas em torno dos exemplos existentes.

A plausibilidade da distribuição dos novos dados é confirmada visualmente. A Figura 33 demonstra o efeito do balanceamento, onde a distribuição de densidade da classe diabética (em roxo) se equipara à da classe não diabética. Mais importante, mostra que os dados sintéticos gerados (em amarelo) seguem de perto a distribuição dos dados diabéticos originais (em vermelho), confirmando que as novas amostras foram criadas em regiões de alta probabilidade e clinicamente coerentes, sem introduzir padrões artificiais ou ruído excessivo.

Figura 35 - Histogramas mostrando a comparação geral "Antes e Depois" do balanceamento com SMOTE



Fonte: elaborado pelo autor.

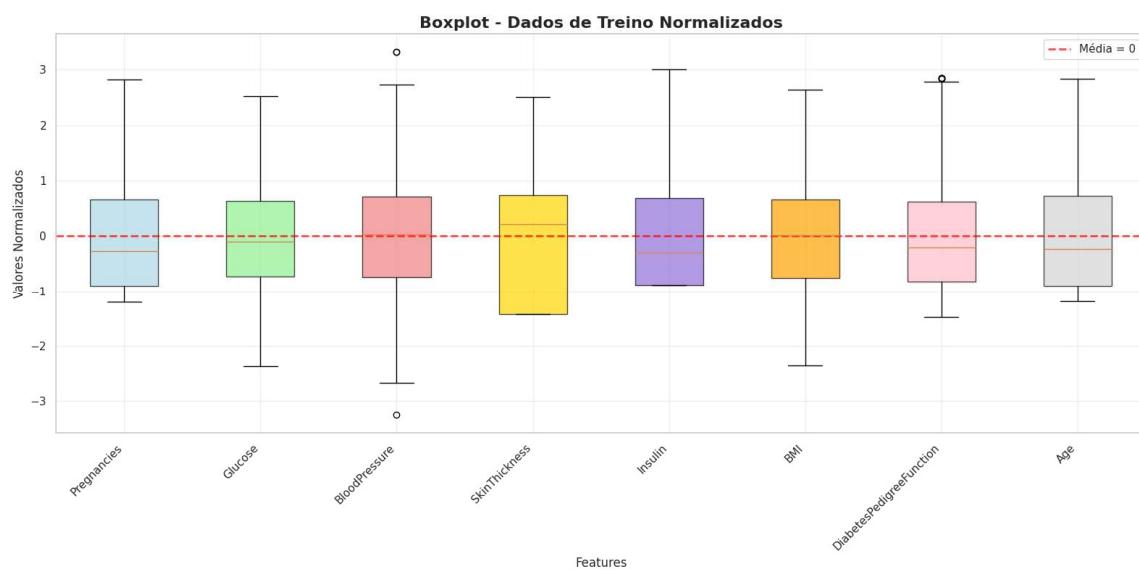
4.2.5 Normalização dos dados

Como etapa final do pré-processamento, a normalização dos dados. O objetivo deste processo é reescalar todas as *features* para uma mesma ordem de grandeza, garantindo que variáveis com escalas numericamente maiores não dominem indevidamente o processo de aprendizado, assegurando assim que cada variável contribua de forma equitativa para a construção do modelo.

Para esta tarefa, foi selecionada a técnica de Padronização (*Standardization*), que transforma os dados para que sigam uma distribuição com média igual a 0 e desvio padrão igual a 1, conforme detalhado na seção de metodologia. É crucial destacar que, para evitar o vazamento de dados (*data leakage*), o cálculo dos parâmetros de normalização foi realizado exclusivamente com base nos dados do conjunto de treinamento. Estes mesmos parâmetros foram então aplicados para transformar os conjuntos de validação e teste, garantindo a consistência e a integridade da avaliação do modelo.

O sucesso do processo de normalização é confirmado visualmente pela Figura 34. A análise do *boxplot* demonstra que todas as features foram efetivamente reescaladas para uma distribuição com média centrada em zero e com variância e amplitude comparáveis. Com esta etapa, o pré-processamento dos dados é concluído, resultando em um conjunto de dados limpo, balanceado e normalizado, pronto para a etapa de modelagem.

Figura 36 – Dados de treino normalizados.



Fonte: elaborado pelo autor.

4.3 TREINAMENTO E MODELAGEM

Com o conjunto de dados devidamente pré-processado – limpo, balanceado e normalizado – esta seção descreve a fase de treinamento, avaliação e seleção dos modelos de aprendizado de máquina. O objetivo central desta etapa foi identificar o algoritmo com a melhor performance e a maior plausibilidade clínica para a predição do risco de diabetes.

A abordagem metodológica foi executada em etapas sequenciais. Inicialmente, foi realizada uma avaliação abrangente de nove distintos algoritmos de classificação, cujos parâmetros já foram descritos na seção de metodologia, para identificar os candidatos com maior potencial preditivo. Com base em métricas de performance, os três modelos de melhor desempenho foram então selecionados para uma fase subsequente de otimização, na qual seus limiares de decisão (*thresholds*) foram ajustados para maximizar a eficácia na detecção de casos positivos. Por fim, a aplicabilidade prática e a coerência dos modelos otimizados foram validadas através da simulação com perfis de pacientes realistas.

As subseções seguintes detalham os resultados e as discussões de cada uma dessas etapas.

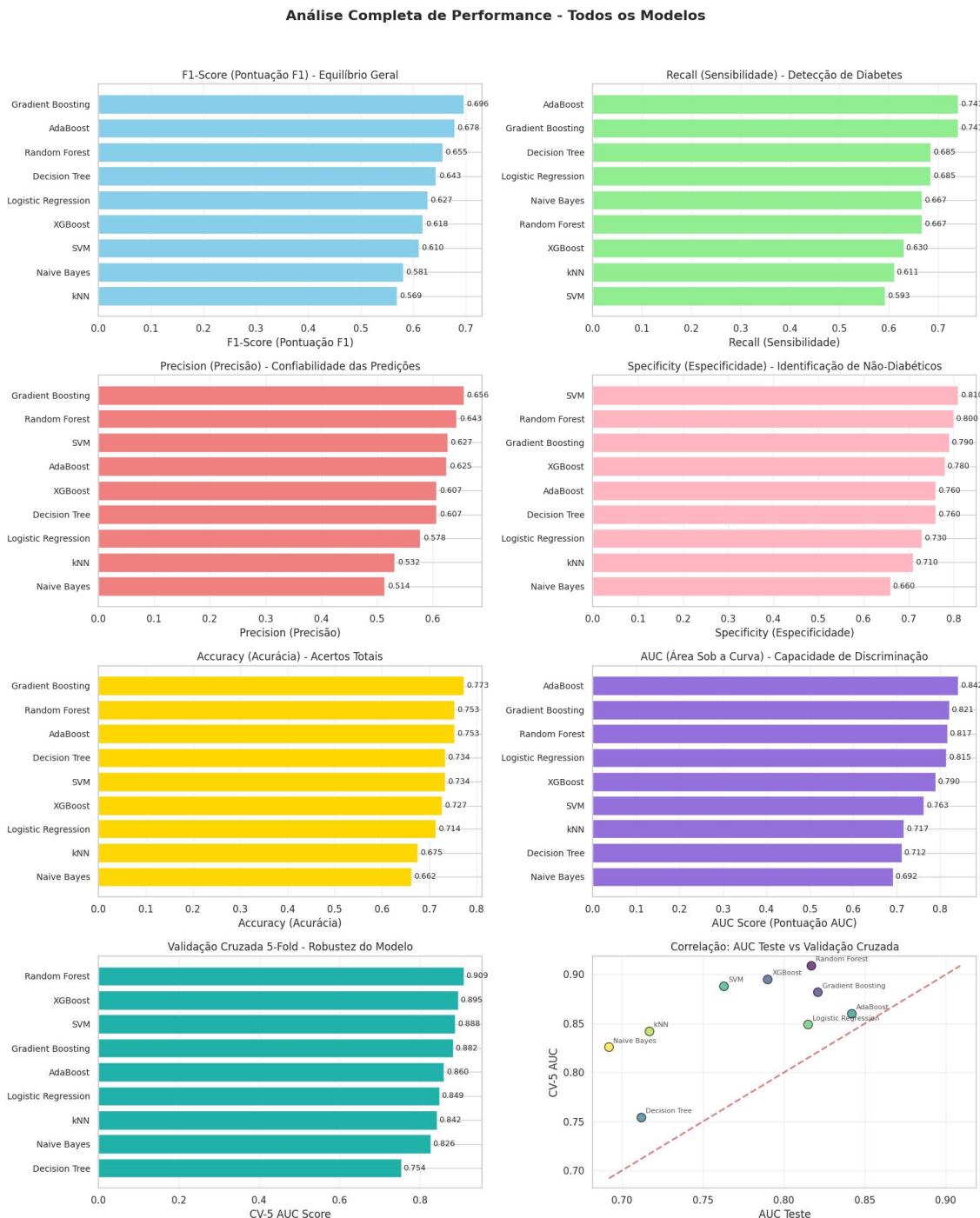
4.3.1 Avaliação e métricas dos modelos

Com os dados já preparados, os nove algoritmos foram treinados e avaliados por um conjunto diversificado de métricas para analisar a performance sob diferentes óticas, como o equilíbrio (F1-Score) e a robustez (CV-5). Conforme a metodologia deste trabalho, foi conferida ênfase especial a duas métricas principais: a AUC (Área Sob a Curva), por medir a capacidade de discriminação geral do modelo, e o Recall (Sensibilidade). A importância do Recall se deve ao contexto clínico, no qual o custo de um falso negativo (não diagnosticar um paciente doente) é consideravelmente mais grave que o de um falso positivo.

Uma visão inicial do desempenho dos modelos é apresentada através de um painel detalhado de métricas na Figura 35. Este panorama avalia os modelos sob as perspectivas de equilíbrio, foco no diagnóstico, confiabilidade das

previsões, capacidade de discriminação e robustez. Uma primeira inspeção destes resultados já aponta para um desempenho superior dos modelos de *ensemble*, notadamente *AdaBoost*, *Gradient Boosting* e *Random Forest*.

Figura 37 – Análise completa da performance de todos os modelos.



Fonte: elaborado pelo autor.

Para uma análise quantitativa precisa, os valores exatos de cada métrica foram compilados e estão apresentados na Tabela 18.

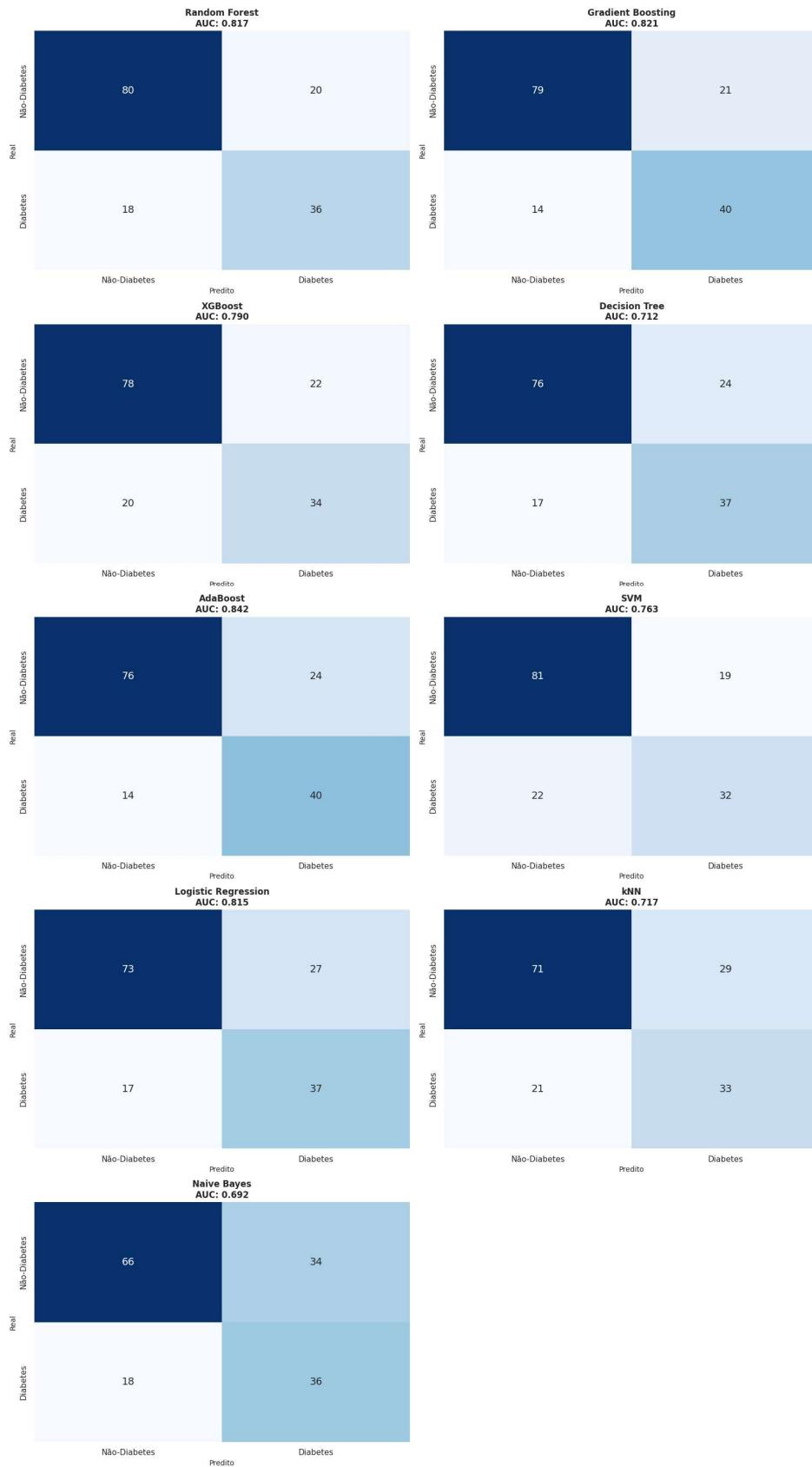
Tabela 20 - Métricas de performance detalhadas dos modelos no conjunto de teste.

Modelo	Accuracy	Precision	Recall	F1-Score	AUC	CV-5	CV-Std
AdaBoost	0.753	0.625	0.741	0.678	0.842	0.86	0.022
Gradient Boosting	0.773	0.656	0.741	0.696	0.821	0.882	0.032
Random Forest	0.753	0.643	0.667	0.655	0.817	0.909	0.022
Logistic Regression	0.714	0.578	0.685	0.627	0.815	0.849	0.023
XGBoost	0.727	0.607	0.63	0.618	0.79	0.895	0.021
SVM	0.734	0.627	0.593	0.61	0.763	0.888	0.028
kNN	0.675	0.532	0.611	0.569	0.717	0.842	0.021
Decision Tree	0.734	0.607	0.685	0.643	0.712	0.754	0.026
Naive Bayes	0.662	0.514	0.667	0.581	0.692	0.826	0.034

Fonte: elaborado pelo autor.

Para entender a natureza dos erros e acertos de cada modelo, uma análise fundamental foi a das **matrizes de confusão** do conjunto de teste, apresentadas na Figura 36, matriz de confusão. Esta ferramenta diagnóstica vai além da acurácia geral, permitindo uma inspeção detalhada do comportamento do classificador ao quantificar não apenas os acertos, mas também os tipos de erros cometidos. Ela detalha os **Verdadeiros Positivos (VP)**, que são os pacientes diabéticos corretamente identificados, e os **Verdadeiros Negativos (VN)**, que são os indivíduos não diabéticos também classificados corretamente. De igual importância, a matriz expõe os erros do modelo: os **Falsos Positivos (FP)**, que representam alertas desnecessários que levariam a exames adicionais, e, crucialmente, os **Falsos Negativos (FN)**, que são os casos de diabetes não detectados e representam o erro de maior custo clínico para este estudo. A Tabela 18 detalha numericamente estes componentes, evidenciando o desempenho prático de cada algoritmo na classificação dos pacientes.

Figura 38 – Matriz de confusão dos modelos treinados.



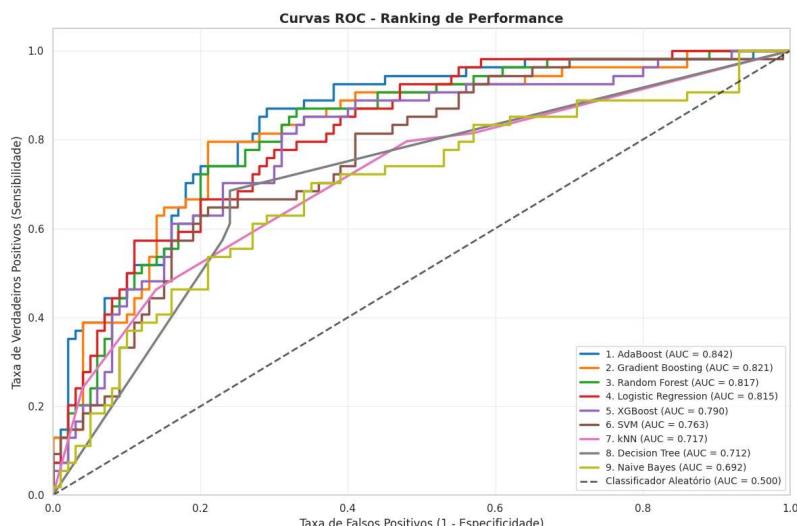
Fonte: elaborado pelo autor.

Tabela 21 – Valores da matriz de confusão.

Modelo	VN	FP	FN	VP	Sens.	Espec.	AUC
AdaBoost	76	24	14	40	0.741	0.76	0.842
Gradient Boosting	79	21	14	40	0.741	0.79	0.821
Random Forest	80	20	18	36	0.667	0.8	0.817
Logistic Regression	73	27	17	37	0.685	0.73	0.815
XGBoost	78	22	20	34	0.63	0.78	0.79
SVM	81	19	22	32	0.593	0.81	0.763
kNN	71	29	21	33	0.611	0.71	0.717
Decision Tree	76	24	17	37	0.685	0.76	0.712
Naive Bayes	66	34	18	36	0.667	0.66	0.692

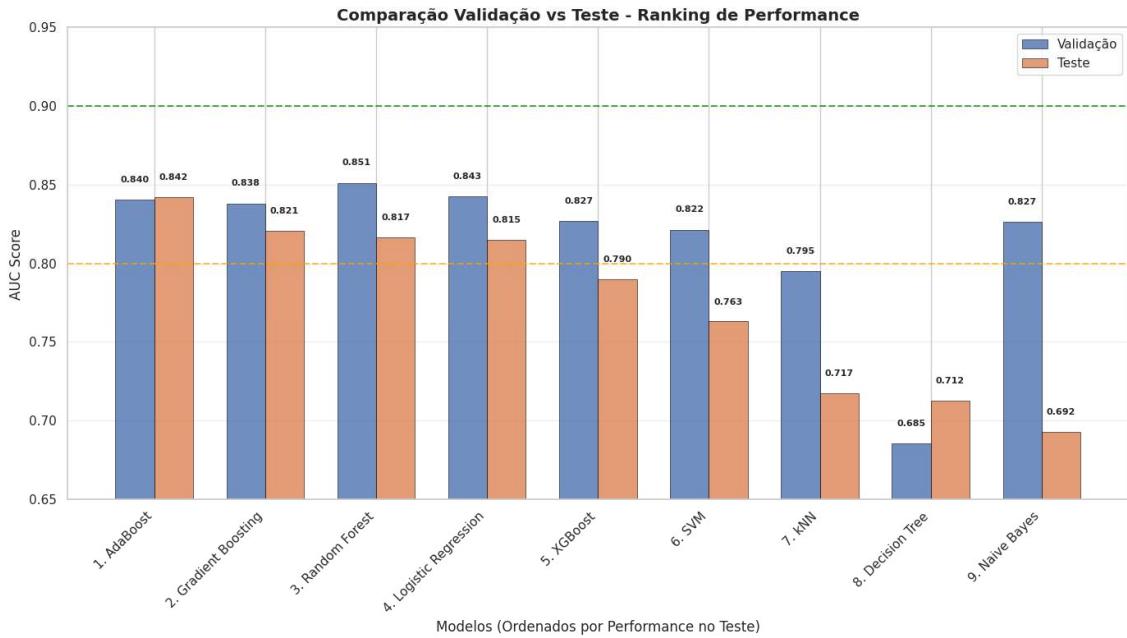
Fonte: elaborado pelo autor.

A capacidade de discriminação geral de cada modelo, ou seja, sua habilidade de distinguir corretamente entre as classes, é melhor visualizada através das **curvas ROC**, apresentadas na Figura 37. Esta figura corrobora o ranking de performance, com AdaBoost (AUC=0.842), Gradient Boosting (AUC=0.821), Random Forest (AUC=0.817) e Logistic Regression (AUC=0.815) se destacando com uma capacidade de discriminação superior. Adicionalmente, a estabilidade e a capacidade de generalização destes modelos foram aferidas comparando-se a performance nos conjuntos de validação e teste, como ilustrado na Figura 38. Notavelmente, os modelos de melhor desempenho demonstraram boa consistência, indicando que não houve superajuste (*overfitting*) significativo.

Figura 39 – Comparação validação vs teste, ranking de performance.

Fonte: elaborado pelo autor

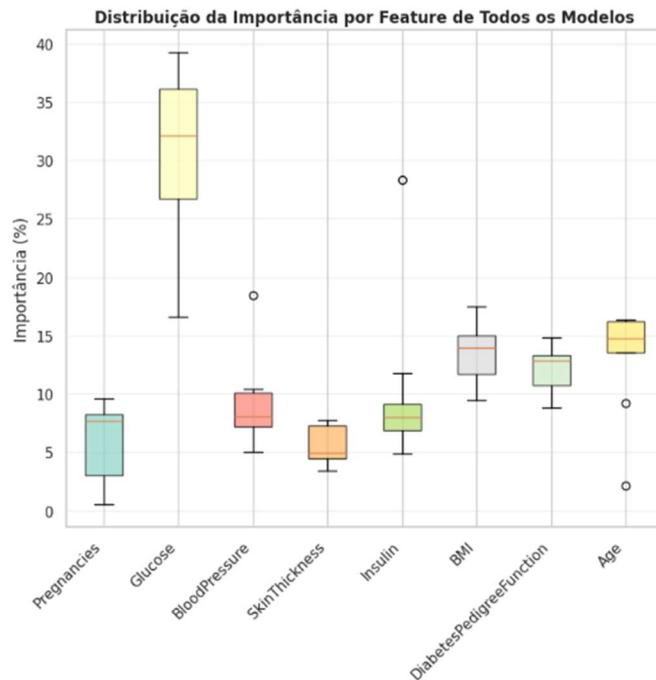
Figura 40 – Comparação validação vs teste, todos os modelos, faixa de corte em 80%.



Fonte: elaborado pelo autor.

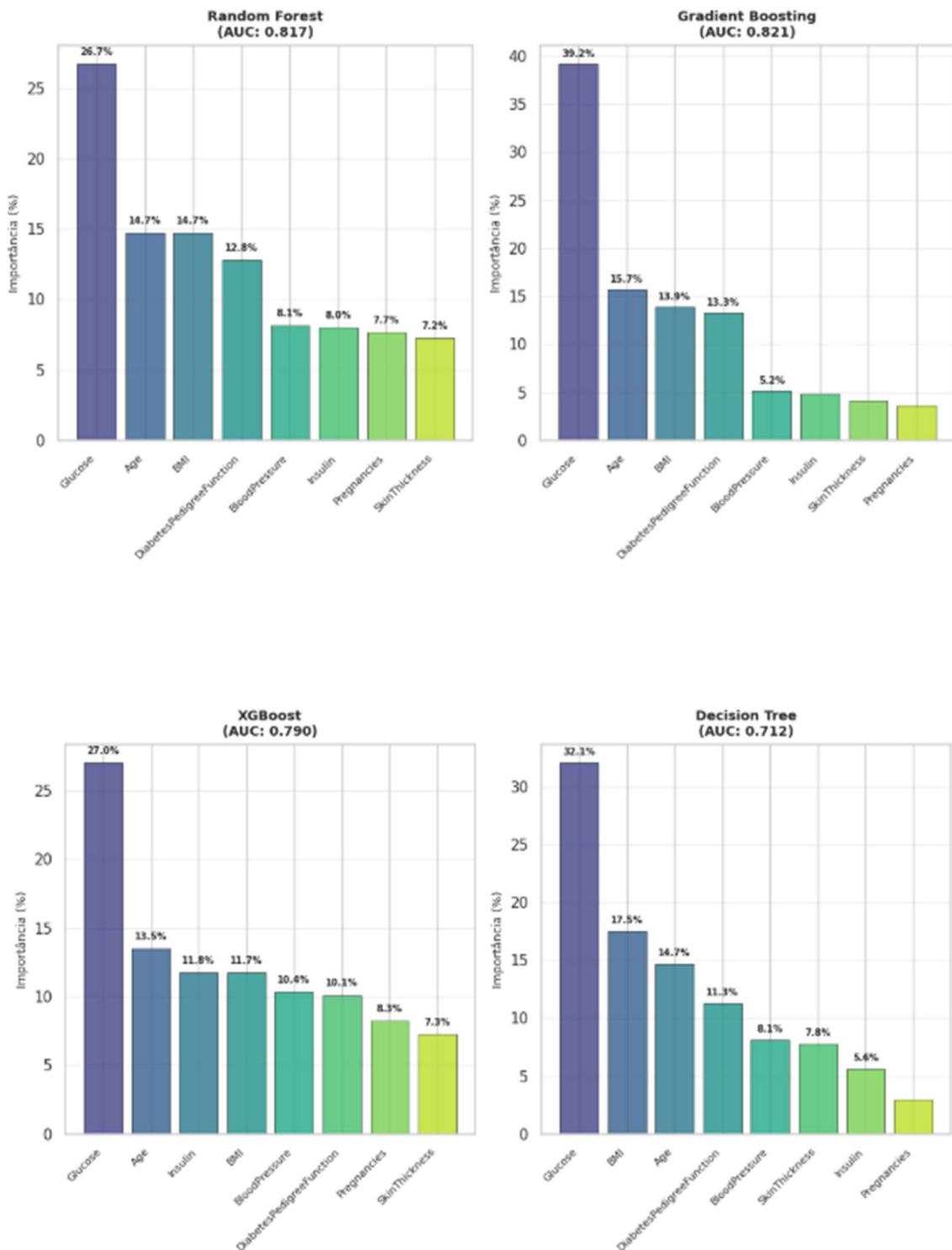
Finalmente, a análise da importância das features (Figura 39 e Figura 40) revelou uma notável consistência. Conforme a Tabela 19, *Glucose* foi inequivocamente identificada como a variável de maior poder preditivo pela vasta maioria dos algoritmos.

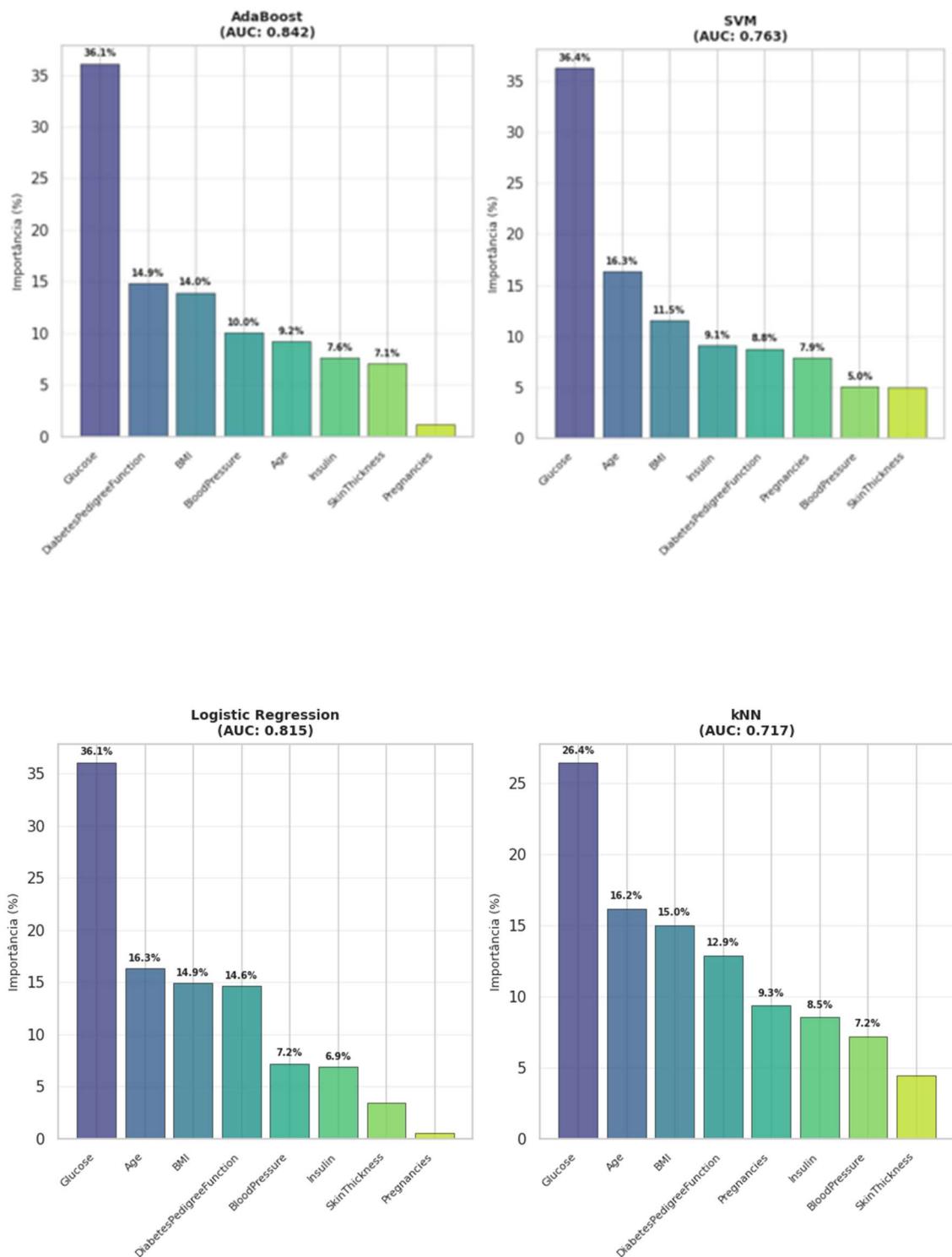
Figura 41 – Distribuição da importância por feature de todos os modelos.

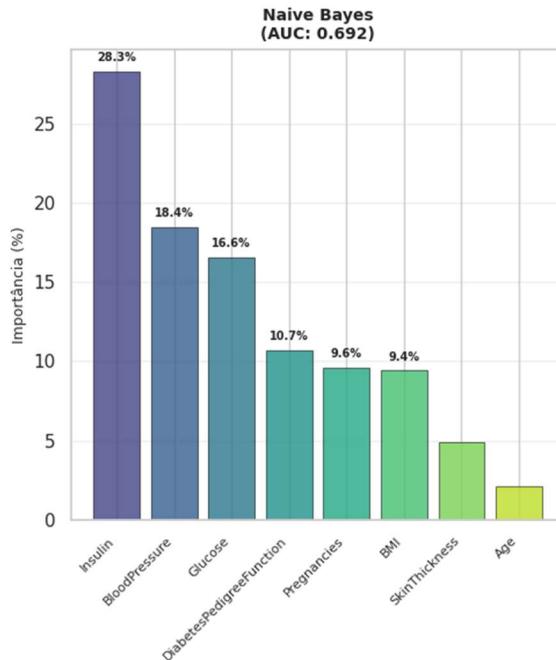


Fonte: elaborado pelo autor.

Figura 42 – Importância das features por cada modelos.







Fonte: elaborado pelo autor.

Tabela 22 – Importância das features por peso.

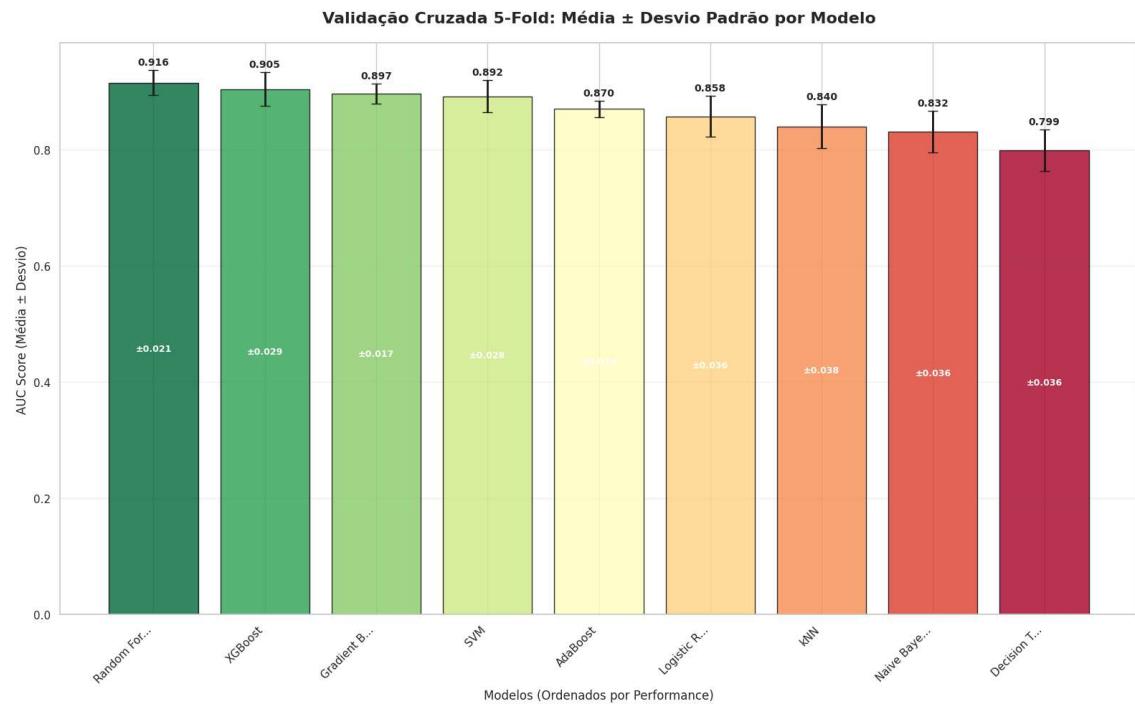
Modelo	1ª Feature	2ª Feature	3ª Feature
AdaBoost	Glucose (36.1%)	DPF (14.9%)	BMI (14.0%)
Gradient Boosting	Glucose (39.2%)	Age (15.7%)	BMI (13.9%)
Random Forest	Glucose (26.7%)	Age (14.7%)	BMI (14.7%)
Logistic Regression	Glucose (36.1%)	Age (16.3%)	BMI (14.9%)
XGBoost	Glucose (27.0%)	Age (13.5%)	Insulin (11.8%)
SVM	Glucose (36.4%)	Age (16.3%)	BMI (11.5%)
kNN	Glucose (26.4%)	Age (16.2%)	BMI (15.0%)
Decision Tree	Glucose (32.1%)	BMI (17.5%)	Age (14.7%)
Naive Bayes	Insulin (28.3%)	BloodPressure (18.4%)	Glucose (16.6%)

Fonte: elaborado pelo autor.

Por fim, a robustez e a estabilidade de cada modelo foram avaliadas através da técnica de validação cruzada de 5 folds (CV-5). A Figura 41 sumariza este desempenho, apresentando a média e o desvio padrão da AUC para cada algoritmo. Os resultados destacaram o Random Forest (AUC CV = 0.916) e o Gradient Boosting (AUC CV = 0.897) como modelos de altíssima performance. Notavelmente, o AdaBoost, apesar de um AUC CV médio ligeiramente menor (0.870), apresentou a maior estabilidade (desvio padrão de ± 0.014), reforçando a confiabilidade dos modelos de *ensemble*. A performance detalhada em cada

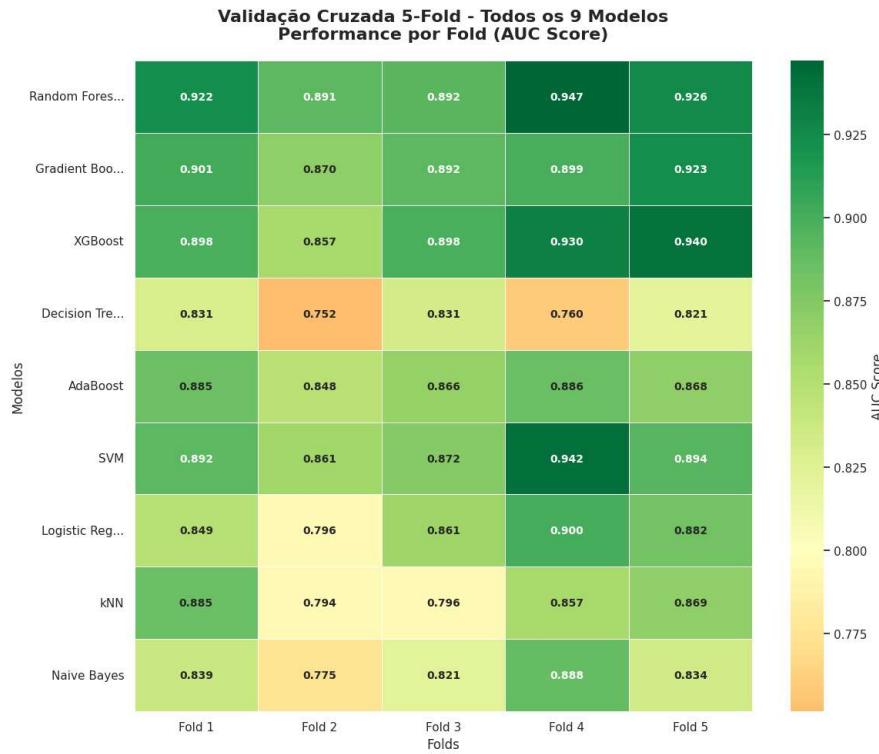
um dos cinco folds, apresentada na Figura 42, confirma visualmente a consistência destes modelos de ponta.

Figura 43 – Validação cruzada 5-Fold, média e desvio padrão.



Fonte: elaborado pelo autor.

Figura 44 – Validação cruzada 5-Fold, performance a cada fold.



Fonte: elaborado pelo autor.

Conforme a metodologia definida, a seleção dos modelos para a fase de otimização foi baseada em um critério primário de AUC superior a 0.80 tanto no conjunto de validação quanto no de teste, utilizando as métricas de *Recall* e a robustez da CV-5 como critérios secundários. Com base nesta análise completa, quatro modelos foram selecionados: *AdaBoost*, *Gradient Boosting*, *Random Forest* e *Logistic Regression*, por demonstrarem a combinação mais robusta de capacidade de discriminação, generalização e eficácia na identificação de casos positivos. Estes quatro modelos avançaram, portanto, para a etapa seguinte de otimização de seus limiares de classificação (*thresholds*), buscando aprimorar ainda mais sua performance para a aplicação prática.

4.3.2 Otimização dos melhores modelos

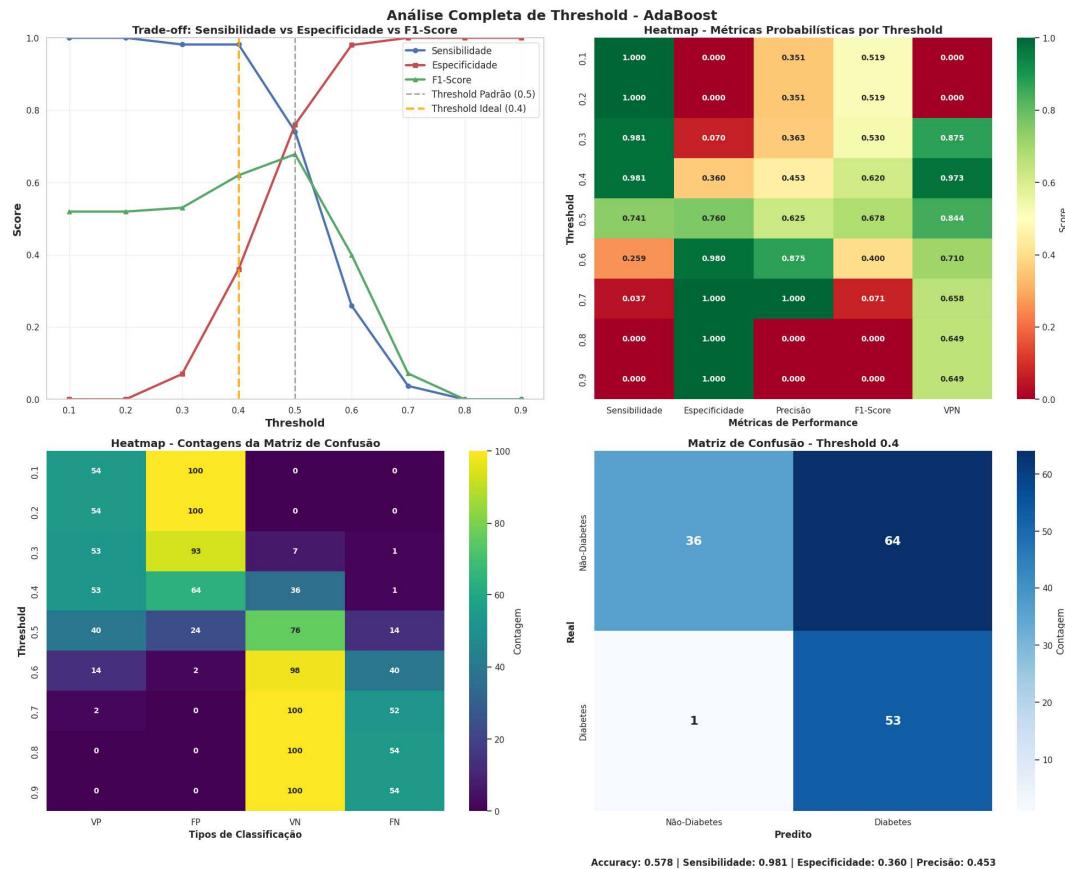
Após a seleção dos quatro algoritmos de melhor performance (*AdaBoost*, *Gradient Boosting*, *Random Forest* e *Logistic Regression*), foi realizada uma etapa de otimização para ajustar o comportamento dos modelos ao contexto clínico do diagnóstico de diabetes. O foco desta fase foi a calibração do limiar de classificação (*threshold*), que é o ponto de corte utilizado para converter a probabilidade predita pelo modelo (um valor entre 0.0 e 1.0) em um diagnóstico binário (Diabético ou Não-Diabético). Por padrão, este limiar é definido como 0.5.

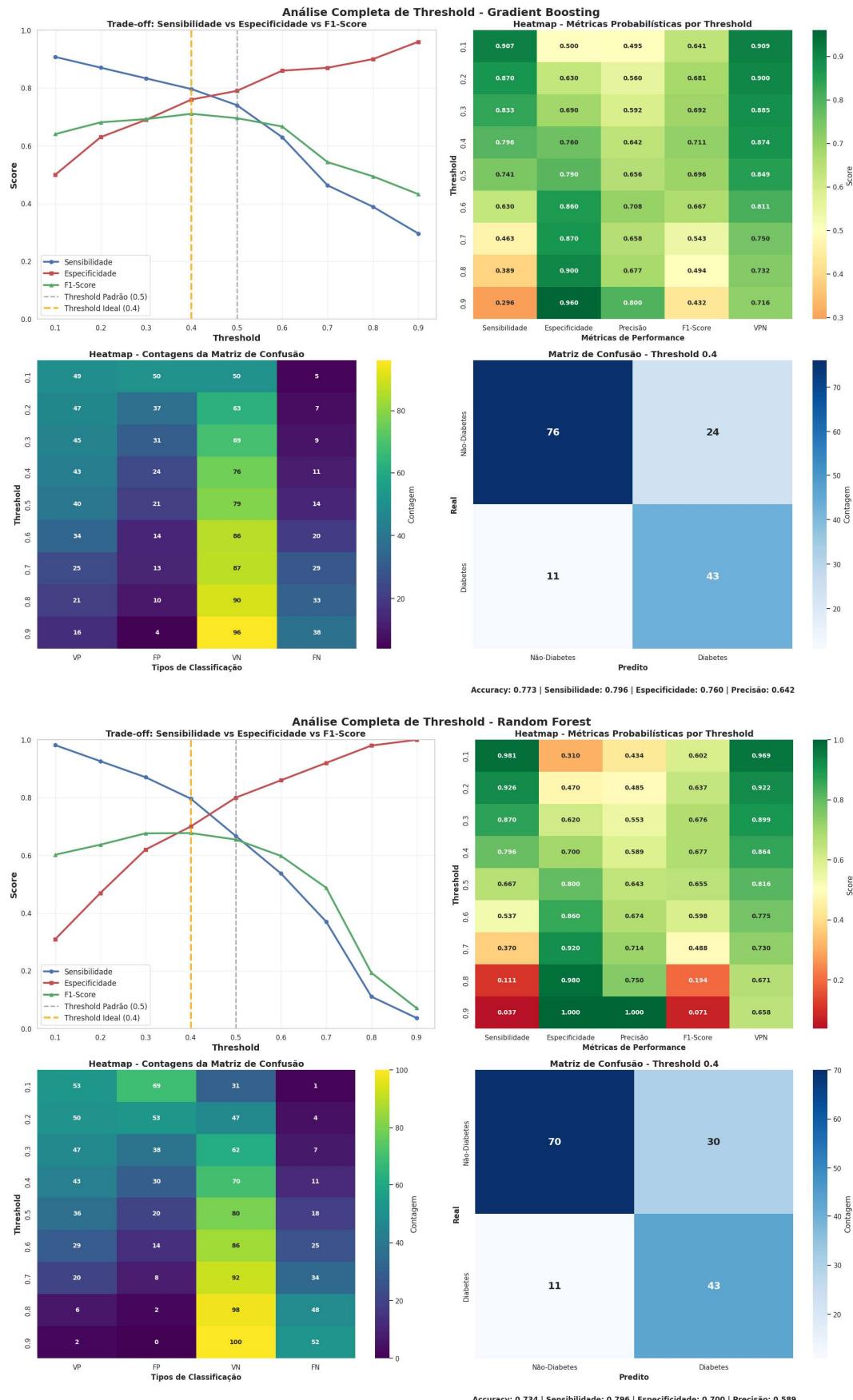
No entanto, um limiar de 0.5 trata os erros de classificação com pesos iguais. Conforme estabelecido, para este estudo, um Falso Negativo (não diagnosticar um paciente doente) é um erro consideravelmente mais grave que um Falso Positivo. Portanto, a otimização buscou encontrar um novo limiar que aumentasse a Sensibilidade (*Recall*), mesmo que isso implicasse em uma ligeira redução da Especificidade. O objetivo foi encontrar o melhor ponto de equilíbrio (*trade-off*) entre essas duas métricas, frequentemente indicado pelo pico da métrica F1-Score.

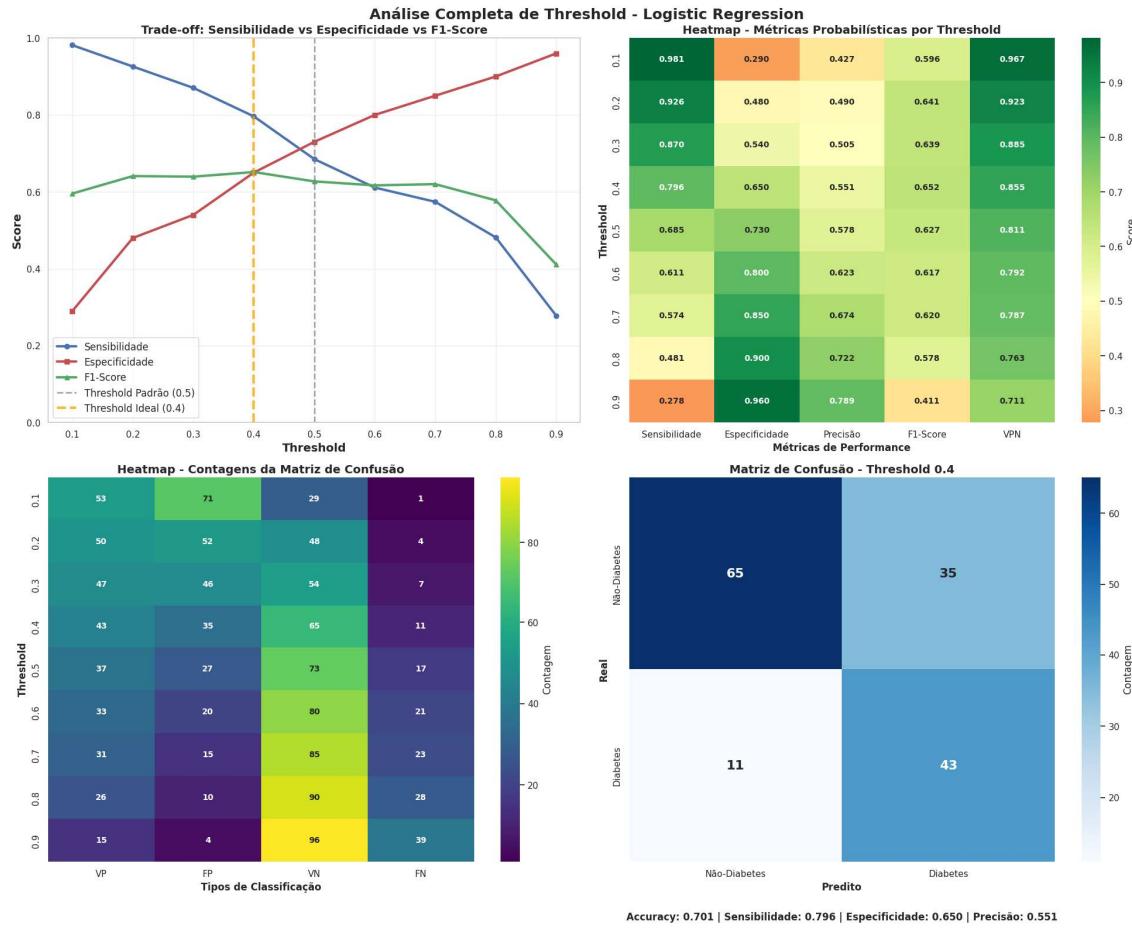
Para cada um dos quatro modelos, foi conduzida uma análise detalhada e consolidada do impacto da variação do *threshold*, como apresentado nos painéis completos na Figura 43. Cada um desses gráficos contém o trade-off, os *heatmaps* de performance e a matriz de confusão final para o limiar ótimo. A análise sistemática destes painéis revelou que um *threshold* de 0.4 representou

um ponto de equilíbrio ideal para a maioria dos modelos, maximizando a capacidade de detecção (Sensibilidade) sem comprometer excessivamente a habilidade de identificar corretamente os casos negativos.

Figura 45 – Análise completa de threshold dos melhores modelos selecionados.





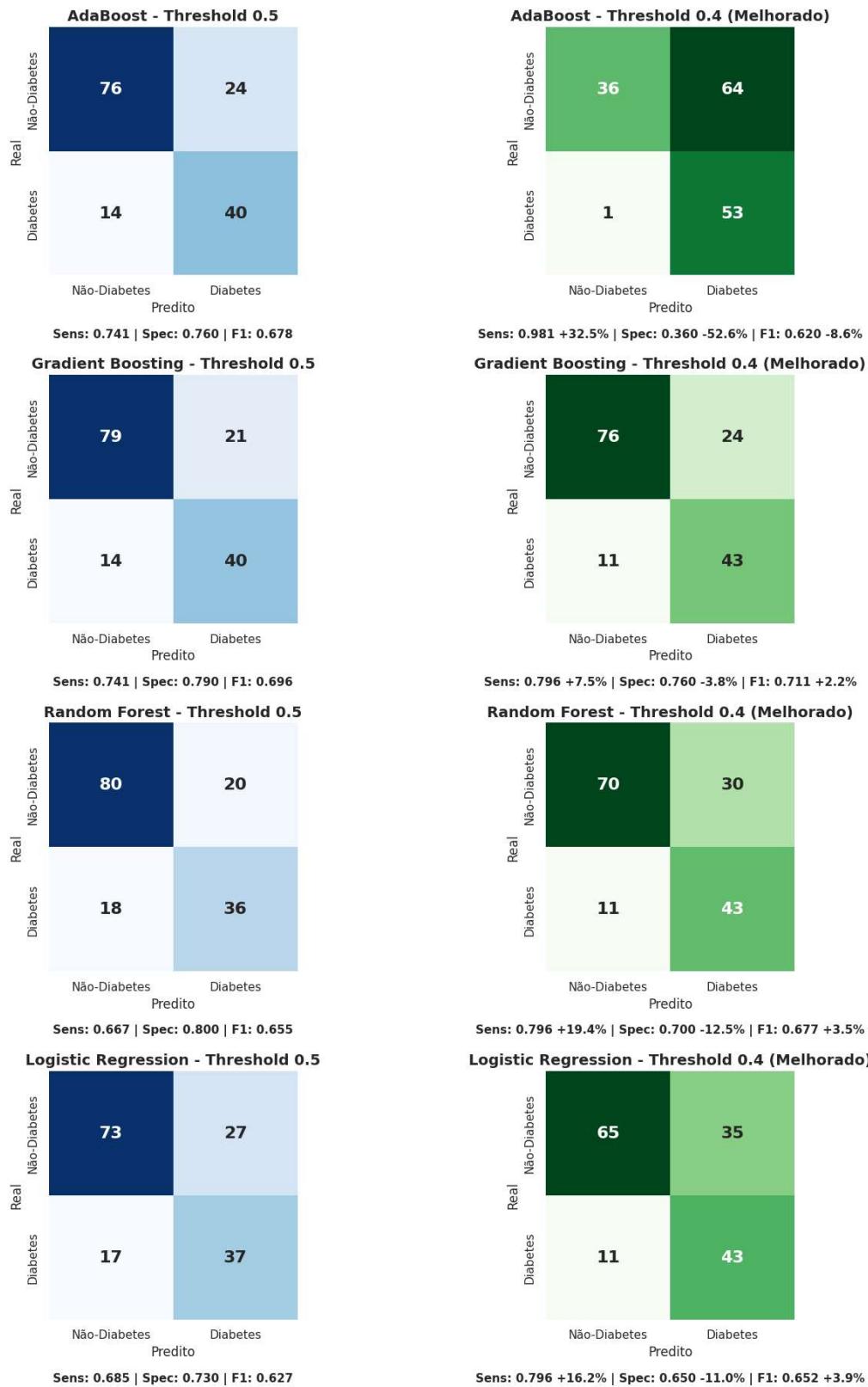


Fonte: elaborado pelo autor.

Para demonstrar o impacto prático da otimização do limiar, a seguir será apresentada a comparação entre as matrizes de confusão geradas com o *threshold* padrão (0.5) e o novo limiar de 0.4, Figura 44. Foi observado que a recalibração do ponto de corte trouxe melhorias de performance para todos os quatro modelos finalistas (*recall//sensibilidade*) e uma leve deterioração em três modelos, e uma piora drástica no Adabost neste quesito. A melhoria mais significativa, e que atende ao objetivo primário da otimização, é a notável redução no número de Falsos Negativos em todos os casos, o que se traduz em uma maior e mais segura capacidade de detecção da doença.

Figura 46 - Diferença da matriz de confusão com threshold 0,5 e 0,4.

Comparação: Matrizes de Confusão - Threshold 0.5 vs 0.4



Fonte: elaborado pelo autor.

4.3.3 Aplicação em perfis reais

A avaliação final e qualitativa dos quatro modelos otimizados foi realizada através da sua aplicação em um conjunto de perfis sintéticos de pacientes, permitindo uma análise de seu comportamento em cenários clínicos, conforme Figura 45.

A análise de perfis revelou comportamentos distintos que não eram visíveis apenas pelas métricas e *benchmarks*. Os modelos *Gradient Boosting* e *AdaBoost* demonstraram ser inadequados para a aplicação prática. O primeiro, por subestimar drasticamente o risco do perfil "Diabético Severo" (probabilidade de 1,3%), e o segundo por superestimar o risco dos perfis não diabéticos, atribuindo probabilidades excessivamente altas e não representando bem a graduação de risco.

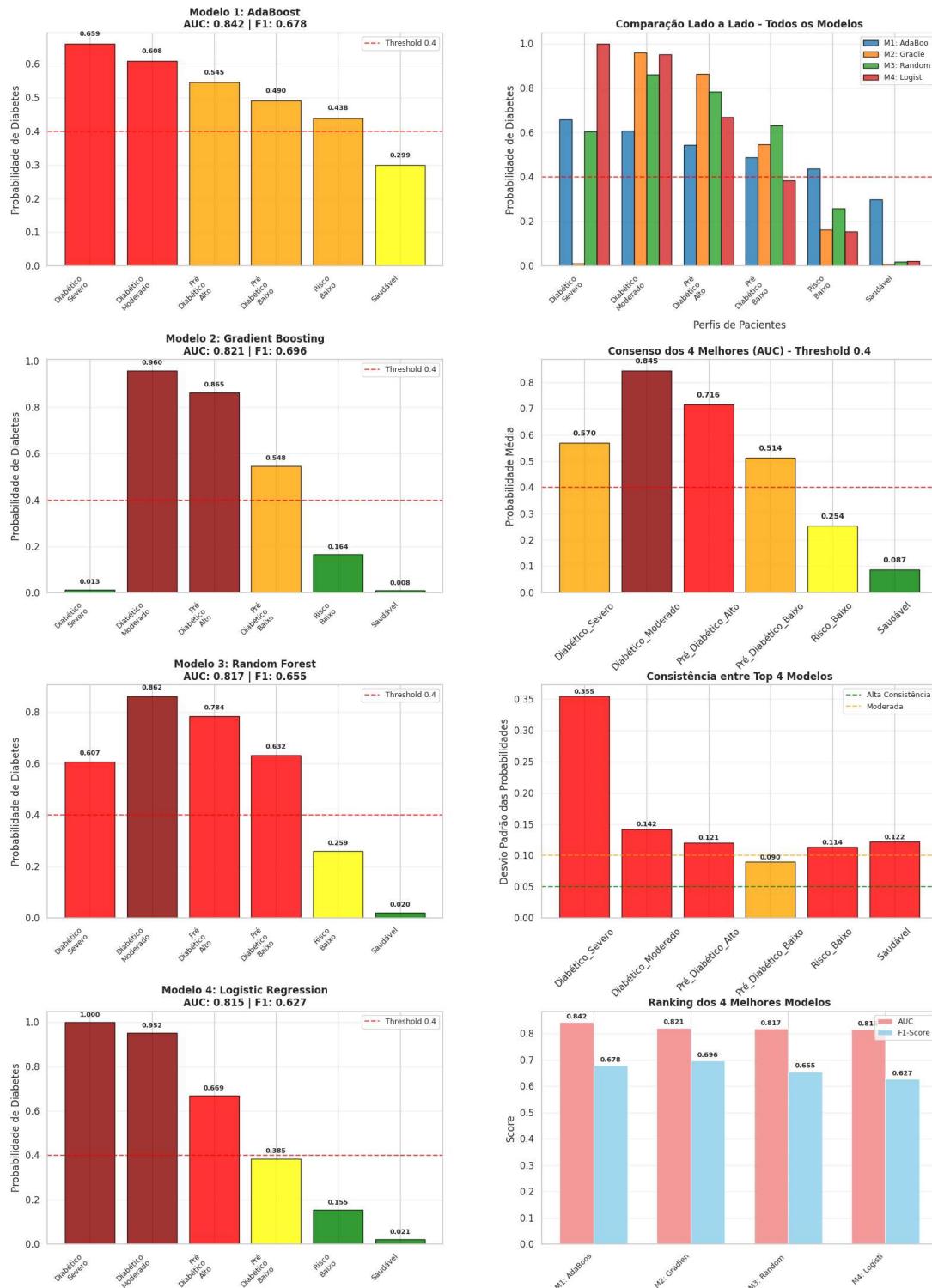
O modelo de Regressão Logística de fato se destacou por apresentar a estratificação de risco mais coesa e coerente, com uma clara e monotônica redução da probabilidade à medida que o risco diminuía. Embora suas métricas quantitativas, como o *F1-Score* e *recall*, não tenham liderado o ranking na etapa anterior, sua performance qualitativa nesta análise foi considerada excelente e sugere que, com diferentes parametrizações ou pré-processamentos, seu potencial poderia ser ainda maior.

No entanto, o modelo *Random Forest* emergiu como a escolha superior e definitiva deste estudo. Ele conseguiu unir o melhor dos dois parâmetros *qualitativo/quantitativo*: apresentou métricas quantitativas de ponta, superiores às da Regressão Logística, e simultaneamente demonstrou um comportamento qualitativo igualmente robusto e seguro, com uma graduação de risco clara e intuitiva entre os perfis.

Portanto, esta etapa de validação qualitativa foi indispensável. Ela permitiu desqualificar modelos com comportamentos de risco (*Gradient Boosting*) e realizar uma comparação aprofundada entre os finalistas. Com base na combinação superior de performance quantitativa e, crucialmente, de um comportamento preditivo seguro e plausível, o modelo Random Forest foi selecionado como o melhor o algoritmo para este estudo, com boa avaliação quantitativa e qualitativa.

Figura 47 – Análise dos modelos candidatos finais.

Análise dos 4 Melhores Modelos SMOTE - Threshold 0.4 - Selecionados por AUC



Fonte: elaborado pelo autor.

5 CONCLUSÕES E RECOMENDAÇÕES

Este trabalho alcançou seu objetivo geral de investigar e desenvolver um modelo de aprendizado de máquina para a predição de risco de diabetes tipo 2, demonstrando com sucesso a aplicabilidade prática da ciência de dados como uma ferramenta de apoio à detecção precoce e à tomada de decisão em saúde. Os objetivos específicos, que guiaram desde a análise exploratória até a avaliação final, foram cumpridos, resultando em um modelo preditivo validado e em aprendizados metodológicos cruciais.

É fundamental ressaltar que o conjunto de dados utilizado, *PIMA Indians Diabetes*, reflete o perfil de uma tribo indígena específica com padrões de prevalência de diabetes historicamente elevados. Portanto, os resultados aqui obtidos, embora metodologicamente válidos, não são diretamente generalizáveis para uma população mais ampla, especialmente a brasileira, que possui características demográficas, genéticas e de estilo de vida distintas.

A investigação metodológica, em resposta aos objetivos de tratamento e preparação dos dados, demonstrou que abordagens complexas de pré-processamento podem introduzir ruídos e valores artificiais sem garantir um resultado superior. Concluiu-se que a estratégia mais eficaz consistiu na utilização dos dados brutos em conjunto com um algoritmo robusto, que se mostrou capaz de lidar com as imperfeições inerentes à amostra sem a introdução de artefatos ou a perda de informação.

Além de toda a análise quantitativa que levou à seleção dos modelos com melhor desempenho, a principal contribuição deste trabalho foi a demonstração da indispensabilidade da análise qualitativa. Verificou-se que as métricas de performance, por si só, podem ser enganosas. A simulação com perfis de pacientes revelou-se uma ferramenta de validação crucial, capaz de desqualificar modelos de alto risco que, de outra forma, seriam considerados promissores. Esse processo garantiu a seleção do modelo *Random Forest* como uma solução não apenas performática, mas verdadeiramente confiável para uma aplicação na área da saúde.

Com base nos achados e visando a evolução e aplicabilidade do projeto, recomendam-se os seguintes trabalhos futuros:

- **Pesquisa e criação de um conjunto de dados nacional:** A principal recomendação é o desenvolvimento de um novo conjunto de dados, uma triagem clínica abrangente, com perfis clínicos e demográficos de pacientes do Brasil. Isso é fundamental para treinar e validar modelos que refletem com maior fidelidade a realidade e os fatores de risco da população brasileira.
- **Validação clínica do modelo atual:** Realizar a validação do modelo Random Forest otimizado com dados de pacientes reais, em um ambiente clínico, para aferir seu desempenho prático e sua utilidade como ferramenta de triagem.
- **Otimização de hiperparâmetros:** Conduzir uma busca aprofundada por hiperparâmetros para os modelos de melhor performance qualitativa (Random Forest e Regressão Logística), visando refinar ainda mais os resultados.
- **Estudo para Rastreamento Populacional:** Utilizar o futuro conjunto de dados brasileiro, proposto na primeira recomendação, para desenvolver e validar um estudo focado na aplicação do modelo como uma ferramenta de baixo custo para o rastreamento em massa da população.

Em suma, o estudo cumpre seus objetivos ao entregar não apenas um modelo preditivo, mas ao estabelecer a primazia da validação de plausibilidade clínica sobre as métricas puramente estatísticas, um passo essencial na construção de soluções de inteligência artificial responsáveis para a saúde.

REFERÊNCIAS

ALTERAR O NEGRITO PARA O TITULO

AKMEŞE, Ömer Faruk. **Diagnosing diabetes with machine learning techniques.** Hittite Journal of Science and Engineering, Çorum, v. 9, n. 1, p. 09-18, 2022. DOI: <https://doi.org/10.17350/HJSE19030000250>.

AMERICAN DIABETES ASSOCIATION. 2. **Diagnosis and Classification of Diabetes.** Diabetes Care, v. 47, supl. 1, p. S20–S29, 2024. DOI: <https://doi.org/10.2337/dc24-S002>.

ANTHROPIC. Introducing Claude 4. **Blog Anthropic**, São Francisco, 22 maio 2025. Disponível em: <https://www.anthropic.com/news/clause-4>. Acesso em: 28 maio 2025.

BAZON, Mariana Resende; PEREIRA, Marcelo Marques. **Diabetes gestacional autorreferido – uma análise da Pesquisa Nacional de Saúde.** Cadernos de Saúde Coletiva, Rio de Janeiro, v. 31, n. 3, e330043, 2023. DOI: <https://doi.org/10.1590/1414-462X202331030043>.

BERGSTRA, J.; BENGIO, Y. **Random search for hyper-parameter optimization.** Journal of Machine Learning Research, Cambridge, MA, v. 13, p. 281-305, 2012. DOI: <https://doi.org/10.5555/2188385.2188395>.

BOCK, Eric. **Medical History Matters in Era of Big Data.** NIH Record, [S.I.], v. LXXII, n. 15, p. 3, 24 jul. 2020. Disponível em: <https://nihrecord.nih.gov/sites/nihrecord/files/pdf/2020/NIH-Record-2020-7-24.pdf>. Acesso em: 5 jun. 2025.

CHAWLA, Nitesh V. et al. SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, 2002. DOI: <https://doi.org/10.1613/jair.953>.

CHOI, R. Y. *et al.* Introduction to Machine Learning, Neural Networks, and Deep Learning. **Translational Vision Science & Technology**, v. 9, n. 2, art. 14, p. 1-12, 2020. DOI: <https://doi.org/10.1167/tvst.9.2.14>.

FONSECA, F. R. *et al.* O impacto da inteligência artificial na interpretação de exames de imagem em diagnóstico médico. **Brazilian Journal of Health Review**, Uberlândia, v. 7, n. 3, p. e69808, 2024. DOI: <https://doi.org/10.34119/bjhrv7n3-132>.

GEEKSFORGEEKS. What Is Data Science? Definition, Skills, Applications, Projects, and More. **GeeksforGeeks**, [s.d.]. Disponível em: <https://www.geeksforgeeks.org/data-science/>. Acesso em: 27 maio 2025

GOOGLE. Dividindo os dados de maneira correta. **Google Developers: Machine Learning Crash Course**, [S.I.], [s.d.]. Disponível em: <https://developers.google.com/machine-learning/crash-course/overfitting/dividing-datasets?hl=pt-br>. Acesso em: 27 maio 2025.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**: data mining, inference, and prediction. 2. ed. New York: Springer, 2009. DOI: <https://doi.org/10.1007/978-0-387-84858-7>.

HE, R.; CAO, J.; TAN, T. Generative artificial intelligence: a historical perspective. **National Science Review**, v. 12, n. 5, nwaf050, 2025. DOI: <https://doi.org/10.1093/nsr/nwaf050>.

HUTTER, F.; KOTTHOFF, L.; VANSCHOREN, J. (org.). **Automated machine learning**: methods, systems, challenges. Cham: Springer, 2019. DOI: <https://doi.org/10.1007/978-3-030-05318-5>.

IBM. Aprendizado supervisionado. **IBM**, [S.I.], [s.d.]. Disponível em: <https://www.ibm.com/br-pt/topics/supervised-learning>. Acesso em: 27 maio 2025.

IBM. Visão geral da ajuda do CRISP-DM. **IBM SPSS Modeler**, 17 ago. 2021. Disponível em: <https://www.ibm.com/docs/pt-br/spss-modeler/saas?topic=dm-crisp-help-overview>. Acesso em: 27 maio 2025.

IGUENFER, Fouzi. IQR Rule for Outliers Detection. **Medium**, 20 jul. 2020. Disponível em: <https://medium.com/@fz.iquenfer/iqr-rule-for-outliers-detection-6e9fcacf2099>. Acesso em: 27 maio 2025.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: trends, perspectives, and prospects. **Science**, v. 349, n. 6245, p. 255, 2015. DOI: <https://doi.org/10.1126/science.aaa8415>.

KAHN, Michael. **Diabetes**. Irvine, CA: UCI Machine Learning Repository, 2017. Disponível em: <https://doi.org/10.24432/C5T59G>.

KUHN, Max; JOHNSON, Kjell. **Feature Engineering and Selection: A Practical Approach for Predictive Models**. New York: Chapman and Hall/CRC, 2019. 310 p. DOI: <https://doi.org/10.1201/9781315108230>.

LARTEY, Clement et al. **Effective Outlier Detection for Ensuring Data Quality in Flotation Data Modelling Using Machine Learning (ML) Algorithms**. **Minerals**, Basel, v. 14, n. 9, p. 925, set. 2024. DOI: <https://doi.org/10.3390/min14090925>.

LEE, Sang-Hwan; KIM, Young-Tae. **Artificial intelligence in healthcare: past, present and future**. *Korean Journal of Anesthesiology*, v. 75, n. 1, p. 25–36, 2022. <https://doi.org/10.4097/kja.21209>

MITCHELL, Tom M. **Machine Learning**. New York: McGraw-Hill, 1997. 432 p. ISBN 0070428077.

OPENAI. Improving language understanding with unsupervised learning. **Blog OpenAI**, San Francisco, 11 jun. 2018. Disponível em: <https://openai.com/index/language-unsupervised/>. Acesso em: 28 maio 2025.

OPENAI. Language models are few-shot learners. **Blog OpenAI**, San Francisco, 28 maio 2020. Disponível em: <https://openai.com/index/language-models-are-few-shot-learners/>. Acesso em: 28 maio 2025.

PAIXÃO, G. M. D. M. et al. Machine Learning na Medicina: Revisão e Aplicabilidade. **Arquivos Brasileiros de Cardiologia**, São Paulo, v. 118, n. 1, p. 95–102, 2022. DOI: <https://doi.org/10.36660/abc.20200596>.

PARASCHIV, Eugen. Synthetic Minority Over-sampling Technique (SMOTE). **Baeldung**, 21 fev. 2024. Disponível em: <https://www.baeldung.com/cs/synthetic-minority-oversampling-technique>. Acesso em: 27 maio 2025.

SOCIEDADE BRASILEIRA DE DIABETES. Diagnóstico de Diabetes Mellitus e Pré-Diabetes. [S.I.], 09 jul. 2024. DOI: <https://doi.org/10.29327/5412848.2024-1>.

STANEK, James McCaffrey. **Test Run - Deep Neural Network Training. MSDN Magazine**, set. 2017. Disponível em: <https://learn.microsoft.com/pt-br/archive/msdn-magazine/2017/september/test-run-deep-neural-network-training>. Acesso em: 2 jun. 2025.

TURING, A. M. **Computing Machinery and Intelligence**. Mind, v. 59, n. 236, p. 433–460, 1950. DOI: <https://doi.org/10.1093/mind/LIX.236.433>.

WANG, Lin et al. **Predicting isolated impaired glucose tolerance without oral glucose tolerance test using machine learning in Chinese Han men**. Frontiers in Endocrinology, Lausanne, v. 16, e1514397, 24 abr. 2025. DOI: <https://doi.org/10.3389/fendo.2025.1514397>.