**UNIVERSIDADE DE CAXIAS DO SUL**
**ÁREA DO CONHECIMENTO DE CIÊNCIAS EXATAS E ENGENHARIAS**

**GUILHERME GABRIELLI FRACALOSSI**

# DETECÇÃO DE DIABETES COM APRENDIZADO DE MÁQUINA E EMBEDDINGS PARA VARIÁVEIS CATEGÓRICAS

**BENTO GONÇALVES**

**2025**

**GUILHERME GABRIELLI FRACALOSSI**

# DETECÇÃO DE DIABETES COM APRENDIZADO DE MÁQUINA E EMBEDDINGS PARA VARIÁVEIS CATEGÓRICAS

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de Bacharel em Ciência da Computação na Área do Conhecimento de Ciências Exatas e Engenharias, Campus Universitário da Região dos Vinhedos, da Universidade de Caxias do Sul.

Orientador: Prof. Dr. André Gustavo Adami

**BENTO GONÇALVES**

**2025**

**GUILHERME GABRIELLI FRACALOSSI**

# DETECÇÃO DE DIABETES COM APRENDIZADO DE MÁQUINA E EMBEDDINGS PARA VARIÁVEIS CATEGÓRICAS

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de Bacharel em Ciência da Computação na Área do Conhecimento de Ciências Exatas e Engenharias, Campus Universitário da Região dos Vinhedos, da Universidade de Caxias do Sul.

**Aprovado em 10/07/2025**

**BANCA EXAMINADORA**

_____

Prof. Dr. André Gustavo Adami
Universidade de Caxias do Sul - UCS

_____

Prof. Dra. Helena Graziottin Ribeiro
Universidade de Caxias do Sul - UCS

_____

Prof. Dra. Scheila de Ávila e Silva
Universidade de Caxias do Sul - UCS

*Dedico este trabalho àqueles que acreditam que os dados podem transformar vidas—pois, de fato, podem.*

# AGRADECIMENTOS

Agradeço profundamente aos meus pais, por criarem um ambiente acolhedor e inspirador, que sempre incentivou a dedicação acadêmica e o crescimento pessoal. Sem seu apoio constante, minha jornada não teria sido possível.

Expresso minha gratidão especial aos meus amigos e colegas, que construíram comigo uma relação fundamentada em parceria, confiança e solidariedade. Sua disponibilidade e espírito prestativo nos momentos mais desafiadores fortaleceram minha jornada acadêmica, demonstrando serem pessoas excepcionais.

Por fim, manifesto minha gratidão ao meu coordenador, Professor Dr. André Gustavo Adami, cuja orientação foi fundamental para o êxito deste trabalho. Sua precisão nas orientações e capacidade de identificar pontos críticos com clareza foram decisivos para o resultado final.

*"O avanço da ciência depende de novas técnicas, novos descobrimentos e novas ideias, provavelmente nessa ordem"*

***Sydney Brenner***

# RESUMO

O *diabetes mellitus* representa um problema de saúde pública global e no Brasil, caracterizado por alta prevalência e subdiagnóstico. A detecção precoce é fundamental, e o aprendizado de máquina emerge como uma ferramenta para essa finalidade. Contudo, a aplicação de aprendizado de máquina a dados de saúde enfrenta desafios, especialmente no tratamento de variáveis categóricas. Métodos tradicionais, como a codificação *one-hot*, podem gerar representações de alta dimensionalidade e esparsas, limitando a capacidade dos modelos de capturar relações complexas entre os preditores. Neste contexto, este trabalho investigou o uso de *embeddings* — uma técnica que aprende representações vetoriais densas e de baixa dimensionalidade — como uma alternativa para a codificação dessas variáveis. O objetivo foi desenvolver e avaliar modelos para a detecção de *diabetes mellitus* utilizando dados do *National Health and Nutrition Examination Survey* (NHANES), verificando se essa abordagem melhora o desempenho preditivo em comparação com métodos convencionais. A metodologia consistiu na implementação de uma arquitetura de rede neural com camadas de *embedding* integradas. O desempenho dessa abordagem foi comparado ao de modelos de referência, como XGBoost e Regressão Logística, treinados com dados processados via codificação *one-hot*. Os resultados indicaram que a inclusão de variáveis categóricas melhorou o desempenho de todos os classificadores. A abordagem com *embeddings* gerou um espaço de características mais compacto e alcançou, com o modelo XGBoost, o melhor desempenho individual (F1-Score de 0,573), embora a diferença em relação à codificação *one-hot* não tenha se mostrado estatisticamente significativa. O melhor desempenho geral foi obtido por um modelo de conjunto (*ensemble*) com *Soft Voting*, que atingiu um F1-Score de 0,583. Os resultados mostram que os *embeddings* são uma alternativa viável e compacta para a representação de variáveis categóricas, e que a combinação de representações densas com algoritmos de conjunto representa uma estratégia mais eficaz para a detecção de diabetes no conjunto de dados analisado.

**Palavras-chave**: Diabetes. Aprendizado de Máquina. *Embeddings*. Variáveis Categóricas. Detecção.

# LISTA DE FIGURAS

# LISTA DE TABELAS

# LISTA DE QUADROS

# LISTA DE ABREVIATURAS E SIGLAS

**AB** *Adaptive Boosting*

**ADASYN** Adaptive Synthetic Sampling

**ADA** American Diabetes Association

**AM** Aprendizado de Máquina

**AUC** Area Under the Curve

**AdaBoost** Adaptive Boosting

**BRFSS** Behavioral Risk Factor Surveillance System

**BRF** Balanced Random Forest

**BiLSTM** Bidirectional Long Short-Term Memory

**CDC** Centers for Disease Control and Prevention

**CNN** Convolutional Neural Network

**DM1** Diabetes Mellitus tipo 1

**DM2** Diabetes Mellitus tipo 2

**DMG** Diabetes Mellitus Gestacional

**DM** Diabetes Mellitus

**DT** Decision Tree

**ELSA-Brasil** Estudo Longitudinal de Saúde do Adulto – Brasil

**ENN** Edited Nearest Neighbors

**FINDRISC** Finnish Diabetes Risk Score

**FN** Falso Negativo

**FPR** False Positive Rate

**FP** Falso Positivo

**GB** Gradient Boosting

**GPJ** Glicose Plasmática em Jejum

**HDL** High-Density Lipoprotein

**HbA1c** Hemoglobina Glicada

**IMC** Índice de Massa Corporal

**IQR** Interquartile Range

**KNN** K-Nearest Neighbors

**LDA** Linear Discriminant Analysis

**LIME** Local Interpretable Model-Agnostic Explanations

**LR** Logistic Regression

**LSTM** Long Short-Term Memory

**LightGBM** Light Gradient-Boosting Machine

**MEC** Mobile Examination Center

**MLP** Multilayer Perceptron

**NB** Naive Bayes

**NCHS** National Center for Health Statistics

**NHANES** National Health and Nutrition Examination Survey

**NLP** Processamento de Linguagem Natural

**PCA** *Principal Component Analysis*

**PES** Prontuários Eletrônicos de Saúde

**PIDD** Pima Indian Diabetes Dataset

**RF** Random Forest

**RNA** Redes Neurais Artificiais

**RNN** Recurrent Neural Network

**ROC** Receiver Operating Characteristic

**ROS** Random Oversampling

**SHAP** SHapley Additive exPlanations

**SMOTE-N** SMOTE for Nominal features

**SMOTEENN** Combination of SMOTE and Edited Nearest Neighbors

**SMOTE** Synthetic Minority Over-sampling Technique

**SUS** Sistema Único de Saúde

**SVM** Support Vector Machine

**TNR** True Negative Rate

**TOTG** Teste Oral de Tolerância à Glicose

**TPR** True Positive Rate

**t-SNE** *t-distributed Stochastic Neighbor Embedding*

**VN** Verdadeiro Negativo

**VPP** Valor Preditivo Positivo

**VP** Verdadeiro Positivo

**Vigitel** Sistema de Vigilância de Fatores de Risco e Proteção para Doenças Crônicas por Inquérito Telefônico

**WEM** Word Embedding Model

**XGBoost** Extreme Gradient Boosting

# SUMÁRIO

# 1 INTRODUÇÃO

O Diabetes Mellitus (DM) é uma doença crônica caracterizada pela incapacidade do organismo de produzir ou utilizar adequadamente a insulina, resultando em elevados níveis de glicose no sangue. Trata-se de um problema de saúde pública crescente, especialmente no Brasil, onde mudanças socioeconômicas, urbanização e estilos de vida sedentários têm contribuído para o aumento da prevalência da condição. Segundo dados do MINISTÉRIO DA SAÚDE (2023), 10,2% da população adulta brasileira relata ter o diagnóstico da doença. Contudo, a dimensão real do problema é ainda maior, uma vez que se estima que cerca de 50% dos casos permaneçam sem diagnóstico (SCHMIDT *et al.*, 2014). Além do grande impacto individual, o diabetes representa um significativo desafio para o sistema de saúde, acarretando custos elevados e exigindo ações estratégicas de prevenção e diagnóstico.

Nesse contexto, ferramentas que auxiliem na identificação do diabetes tornam-se fundamentais para ações preventivas mais eficazes, redução dos custos assistenciais e melhoria da qualidade de vida dos indivíduos afetados. A utilização de modelos preditivos baseados em aprendizado de máquina se destaca como uma estratégia promissora, pois permite identificar padrões nos dados que podem antecipar o diagnóstico da doença mesmo em estágios iniciais (OLIVEIRA *et al.*, 2024; TANIM *et al.*, 2025). Além disso, esses modelos possibilitam o monitoramento contínuo de fatores de risco, apoiando decisões clínicas mais assertivas e personalizadas (ZHOU *et al.*, 2024; CICHOSZ; OLESEN; JENSEN, 2024).

Um dos principais desafios na aplicação de ferramentas de aprendizado de máquina na área da saúde está relacionado à complexidade intrínseca dos dados utilizados. Esses dados são, em sua maioria, heterogêneos — combinando variáveis numéricas, categóricas, temporais e textuais — e frequentemente apresentam distribuição desbalanceada entre as classes (CHOWDHURY; AYON; HOSSAIN, 2023). Um exemplo dessa estrutura é a base de dados *Behavioral Risk Factor Surveillance System* (BRFSS), amplamente utilizada para predição de diabetes, que em suas versões comumente analisadas pode conter mais de 20 variáveis categóricas para apenas uma numérica (REN, 2024). Além disso, a predominância de atributos categóricos dificulta a modelagem eficiente, pois seu uso direto pode induzir os modelos a erros, exigindo técnicas de pré-processamento específicas como o *one-hot encoding*, que transforma o atributo em uma representação numérica para viabilizar o treinamento (CHOWDHURY; AYON; HOSSAIN, 2023). Tais fatores demandam abordagens mais robustas e técnicas específicas de pré-processamento, seleção de atributos e ajuste de modelos.

Para superar essas limitações, uma abordagem que tem ganhado destaque é o uso de *embeddings* categóricos (GOODFELLOW; BENGIO; COURVILLE, 2016; DAHOUDA; JOE, 2021a; HANCOCK; KHOSHGOFTAAR, 2020). Essencialmente, esta técnica busca representar cada categoria de uma variável como um vetor de números reais em um espaço multidimensional. O diferen-

cial desta abordagem é que a representação vetorial não é pré-definida, mas, sim, aprendida durante o treinamento do próprio modelo de aprendizado de máquina. O objetivo é ajustar os vetores de modo que categorias com impacto semelhante no resultado final sejam posicionadas próximas umas das outras nesse espaço vetorial, revelando, assim, propriedades intrínsecas das variáveis (GUO; BERKHAHN, 2016). Dessa forma, os algoritmos de aprendizado conseguem capturar relações sutis entre diferentes características dos dados, que seriam ignoradas por métodos mais tradicionais, como o *one-hot encoding* (RUSSAC; CAELEN; HE-GUELTON, 2018). Esta técnica pode ser especialmente vantajosa quando aplicada em bases de dados complexas e amplas, compostas majoritariamente por dados categóricos, como é comum em cenários de saúde.

Diante disso, o presente trabalho visa desenvolver e avaliar uma nova abordagem para detecção de diabetes, utilizando *embeddings* categóricos. Com essa proposta, busca-se verificar se essa técnica melhora significativamente o desempenho dos modelos preditivos existentes, oferecendo um mecanismo mais robusto para a identificação do diabetes na população. Para avaliar o sistema, será utilizada uma base de dados real.

## 1.1 OBJETIVOS

Este trabalho tem por objetivo desenvolver e avaliar uma nova abordagem utilizando aprendizado de máquinas baseado em *embeddings* para a detecção de diabetes, visando melhorar o desempenho dos modelos preditivos para essa tarefa. O sistema de detecção é avaliado utilizando uma base de dados real.

Para alcançar esse objetivo principal, serão desenvolvidos os seguintes objetivos específicos:

1. Analisar e explorar o conjunto de dados sobre diabetes, identificando a natureza das variáveis;
2. Implementar *embeddings* categóricos para representar variáveis categóricas dos dados;
3. Avaliar comparativamente diferentes algoritmos de aprendizado de máquina na detecção de diabetes com e sem a aplicação de *embeddings* categóricos;
4. Identificar as principais variáveis e *embeddings* que influenciam a detecção do diabetes.

## 1.2 ESTRUTURA DO TRABALHO

O presente trabalho está organizado da seguinte forma:

- O Capítulo 2 apresenta a fundamentação teórica abordando conceitos sobre diabetes, métodos de detecção atuais, aprendizado de máquina, abordagens para tratamento de variáveis categóricas, características de dados dessa natureza e trabalhos relacionados;

- O Capítulo 3 descreve a proposta de solução, abordando o pré-processamento dos dados, a implementação dos *embeddings* categóricos e a arquitetura do modelo de aprendizado de máquina;
- O Capítulo 4 apresenta os resultados obtidos, discutindo o desempenho dos modelos propostos e realizando uma comparação dos resultados;
- O Capítulo 5 conclui o trabalho, destacando as contribuições realizadas, limitações identificadas e propostas para trabalhos futuros.

# 2 REFERENCIAL TEÓRICO

Este capítulo estabelece a fundamentação teórica para a detecção de DM por meio de modelos de aprendizado de máquina, com ênfase no tratamento de variáveis categóricas via *embeddings*. A exposição inicia-se com a contextualização do problema de saúde, abordando os conceitos clínicos do DM na Seção 2.1 e os métodos de diagnóstico convencionais na Seção 2.2. Em seguida, a Seção 2.3 formaliza a detecção da doença como uma tarefa de classificação computacional e detalha as métricas de avaliação de desempenho empregadas. As seções 2.4 e 2.5 revisam a aplicação de aprendizado de máquina neste domínio, descrevendo o fluxo de trabalho e classificadores documentados na literatura. A Seção 2.6 aprofunda as dificuldades mais relevantes, como o desbalanceamento de classes e a interpretabilidade. Por fim, a Seção 2.7 dedica-se à questão central, analisando as técnicas de codificação para variáveis categóricas, desde os métodos tradicionais até a abordagem baseada em *embeddings*.

## 2.1 DIABETES MELLITUS

O DM é um grupo heterogêneo de distúrbios metabólicos caracterizado por hiperglicemia crônica, resultante de defeitos na secreção de insulina, na sua ação, ou em ambos (ADA, 2023). Trata-se de uma condição crônica que, se não gerenciada adequadamente, pode levar a complicações graves e multissistêmicas, reduzindo significativamente a qualidade e a expectativa de vida dos indivíduos afetados (KHATIB, 2006; Sociedade Brasileira de Diabetes, 2019).

A maioria dos casos de DM pode ser classificada em dois tipos principais (ADA, 2023). O Diabetes Mellitus tipo 1 (DM1) é caracterizado pela destruição autoimune das células beta pancreáticas, resultando em deficiência absoluta de insulina. O Diabetes Mellitus tipo 2 (DM2) corresponde à vasta maioria dos casos, compreendendo cerca de 90% do total globalmente (ELIASCHEWITZ *et al.*, 2015), e resulta de uma perda progressiva da secreção adequada de insulina pelas células beta, frequentemente no contexto de resistência insulínica pré-existente. Outras formas específicas, embora menos comuns, incluem o Diabetes Mellitus Gestacional (DMG), diagnosticado durante a gravidez, e formas mais raras decorrentes de causas genéticas, doenças do pâncreas exócrino ou induzidas por medicamentos.

As consequências do DM não controlado são vastas e severas, incluindo complicações microvasculares (retinopatia, nefropatia, neuropatia) e macrovasculares (doença arterial coronariana, doença cerebrovascular, doença arterial periférica) (KHATIB, 2006). Essas complicações aumentam o risco de cegueira, insuficiência renal, amputações de membros inferiores, infarto do miocárdio e acidente vascular cerebral, contribuindo significativamente para a morbimortalidade associada à doença.

No Brasil, o DM é um relevante problema de saúde pública. Dados do sistema de mo-

nitoramento Sistema de Vigilância de Fatores de Risco e Proteção para Doenças Crônicas por Inquérito Telefônico (Vigitel) de 2023 indicam que 10,2% da população adulta brasileira referem diagnóstico médico prévio de diabetes (MINISTÉRIO DA SAÚDE, 2023). Esta prevalência demonstra uma tendência de aumento com o avançar da idade e é significativamente maior entre os indivíduos com menor nível de escolaridade (MINISTÉRIO DA SAÚDE, 2023). Contudo, estudos epidemiológicos de base populacional que empregaram métodos diagnósticos mais acurados, como o Teste Oral de Tolerância à Glicose (TOTG) ou a Hemoglobina Glicada (HbA1c), como o Estudo Longitudinal de Saúde do Adulto – Brasil (ELSA-Brasil), sugerem que a prevalência real da doença seja substancialmente maior (SCHMIDT *et al.*, 2014). Estima-se que uma parcela considerável dos casos, possivelmente próxima a 50%, permaneça sem diagnóstico no país (SCHMIDT *et al.*, 2014; ELIASCHEWITZ *et al.*, 2015). Além da doença estabelecida, a prevalência de pré-diabetes – uma condição intermediária de risco elevado para o desenvolvimento futuro de DM2 – também é alta na população brasileira, representando uma janela de oportunidade crucial para intervenções preventivas (ELIASCHEWITZ *et al.*, 2015).

A carga econômica imposta pelo DM é substancial, impactando tanto os indivíduos quanto os sistemas de saúde. Os custos incluem gastos diretos com medicamentos, consultas, hospitalizações e manejo das complicações, e custos indiretos relacionados à perda de produtividade, incapacidade e mortalidade prematura (ELIASCHEWITZ *et al.*, 2015; BAHIA *et al.*, 2011). No Brasil, o Sistema Único de Saúde (SUS) arca com uma parcela significativa desses custos, tornando o controle e a prevenção do diabetes prioridades estratégicas de saúde pública (ELIASCHEWITZ *et al.*, 2015; MALTA *et al.*, 2014). A alta prevalência da doença, especialmente entre grupos socioeconomicamente vulneráveis, idosos e populações não brancas, como demonstrado no ELSA-Brasil (SCHMIDT *et al.*, 2014), agrava as iniquidades em saúde e representa um desafio adicional para o desenvolvimento do país.

## 2.2 DIAGNÓSTICO DO DIABETES

A detecção e o diagnóstico do DM envolvem uma combinação de avaliação clínica e testes laboratoriais. Tradicionalmente, a suspeita clínica surge a partir da identificação de fatores de risco ou da presença de sintomas clássicos de hiperglicemia. Os fatores de risco bem estabelecidos para DM2 incluem idade avançada (tipicamente acima de 35-45 anos), sobrepeso ou obesidade (especialmente obesidade central), histórico familiar de diabetes, pertencimento a certas etnias (como afrodescendentes, hispânicos, indígenas, asiáticos), histórico de diabetes gestacional, sedentarismo, hipertensão arterial, dislipidemia (High-Density Lipoprotein (HDL) baixo e/ou triglicerídeos elevados) e presença de outras condições associadas à resistência insulínica, como a síndrome dos ovários policísticos (ADA, 2023; KHATIB, 2006).

O rastreamento em indivíduos assintomáticos, mas com fatores de risco, é uma estratégia fundamental para a detecção precoce. A ADA (2023) recomenda o rastreamento em adultos com sobrepeso/obesidade que possuam um ou mais fatores de risco adicionais, e em todos os

adultos a partir dos 35 anos, com repetição a cada 3 anos se os resultados forem normais.

Os critérios laboratoriais para o diagnóstico do DM são padronizados internacionalmente (ADA, 2023; Sociedade Brasileira de Diabetes, 2019). Estes incluem a medição da Glicose Plasmática em Jejum (GPJ) com resultado $\geq 126$ mg/dL, a avaliação da glicemia duas horas após a sobrecarga oral de 75g de glicose durante o TOTG com resultado $\geq 200$ mg/dL, ou a dosagem da HbA1c com resultado $\geq 6,5\%$. Adicionalmente, a presença de sintomas clássicos de hiperglicemia – como poliúria, polidipsia, polifagia e perda de peso inexplicada – associada a uma glicemia casual (medida a qualquer hora do dia, independentemente do horário da última refeição) $\geq 200$ mg/dL também estabelece o diagnóstico. Na ausência de hiperglicemia inequívoca, o diagnóstico requer a confirmação por meio de dois resultados de testes anormais, seja da mesma amostra ou de duas amostras distintas coletadas em momentos diferentes (ADA, 2023).

Cada um desses testes diagnósticos – GPJ, TOTG e HbA1c – possui particularidades que influenciam sua aplicabilidade. A GPJ é frequentemente empregada pela conveniência do jejum de 8 horas, embora possa não detectar hiperglicemia pós-prandial isolada (ADA, 2023). O TOTG, apesar de ser considerado mais sensível para identificar alterações precoces na tolerância à glicose, é um procedimento mais complexo, oneroso e com menor reprodutibilidade (ADA, 2023; KHATIB, 2006). A HbA1c reflete a média glicêmica dos últimos dois a três meses, dispensa o jejum e exibe menor variabilidade intraindividual; no entanto, seu custo pode ser superior, a disponibilidade restrita em algumas localidades e os resultados podem ser afetados por condições que interferem na sobrevida das hemácias (como anemias e hemoglobinopatias) ou no processo de glicação (ADA, 2023). A seleção do teste mais adequado deve ponderar o quadro clínico, os recursos disponíveis e as características do indivíduo, pois nenhum método isolado é universalmente preferível (ADA, 2023).

Uma ferramenta complementar e de baixo custo para a triagem inicial de risco são os escores de risco validados, como o *Finnish Diabetes Risk Score* (FINDRISC) ou o escore de risco da American Diabetes Association (ADA) (KHATIB, 2006). Baseados em informações clínicas e demográficas simples, obtidas por questionários, eles estimam a probabilidade de um indivíduo ter diabetes não diagnosticado ou desenvolver a doença no futuro. Indivíduos com pontuação elevada nesses escores devem ser encaminhados para testes laboratoriais confirmatórios (ADA, 2023). Esses escores são úteis para o rastreamento populacional, mas, por serem geralmente baseados em modelos de regressão com um conjunto limitado de variáveis, não capturam toda a complexidade das interações entre múltiplos fatores de risco (KHATIB, 2006).

Apesar da disponibilidade de estratégias de rastreamento e diagnóstico, uma parcela significativa de indivíduos com DM no Brasil e no mundo permanece sem diagnóstico por vários anos (SCHMIDT *et al.*, 2014; ELIASCHEWITZ *et al.*, 2015). Isso se deve, em grande parte, à natureza frequentemente assintomática ou oligossintomática do DM2 em suas fases iniciais. O retardo no diagnóstico está diretamente associado a um maior risco de desenvolvimento de complicações

crônicas no momento da identificação da doença, impactando negativamente o prognóstico e a qualidade de vida (KHATIB, 2006). Portanto, a otimização das estratégias de detecção precoce e rastreamento, incluindo o potencial uso de novas ferramentas como modelos de Aprendizado de Máquina, permanece uma área de grande importância clínica e de saúde pública.

## 2.3 TAREFA DE DETECÇÃO AUTOMÁTICA DE DIABETES

A tarefa de detecção de diabetes pode ser formulada como um problema de classificação binária. O objetivo é desenvolver um modelo matemático (classificador) capaz de aprender um mapeamento a partir de um conjunto de características (variáveis) de entrada de um indivíduo para uma variável de saída que representa a classe de interesse (presença ou ausência de diabetes).

Formalmente, seja $\mathcal{X}$ o espaço das características de entrada, no qual cada instância $\mathbf{x} \in \mathcal{X}$ é representada por um vetor $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ composto por $d$ atributos referentes a um indivíduo (e.g., idade, Índice de Massa Corporal (IMC), pressão arterial). Seja $\mathcal{Y}$ o espaço das classes de saída. No contexto da detecção de diabetes, trabalha-se tipicamente com um problema de classificação binária, de modo que $\mathcal{Y} = \{0, 1\}$, onde $Y = 1$ representa a classe "diabetes" (positivo) e $Y = 0$ representa a classe "não diabetes" (negativo).

Dado um conjunto de dados de treinamento $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, onde $\mathbf{x}_i \in \mathcal{X}$ e $y_i \in \mathcal{Y}$ são as características e a classe verdadeira do $i$-ésimo indivíduo, o objetivo é aprender uma função (modelo) $h : \mathcal{X} \to \mathcal{Y}$ que permita generalizar para dados não vistos. A função $h(\mathbf{x})$ produz uma predição $\hat{y}$ para uma nova instância $\mathbf{x}$.

### 2.3.1 Métricas de Desempenho

A avaliação do desempenho de um modelo classificador é crucial, especialmente em aplicações médicas com consequências significativas. A ferramenta fundamental para essa avaliação é a **matriz de confusão** (WEBB; COPSEY, 2011). Para um problema de classificação binária, a matriz de confusão (Quadro 1) sumariza os resultados das predições do modelo em um conjunto de dados (teste ou validação), comparando os valores preditos com os valores verdadeiros em quatro categorias (BISHOP, 2006):

- **Verdadeiro Positivo (VP)**: número de instâncias da classe positiva ($Y = 1$, diabetes) que foram corretamente classificadas como positivas ($\hat{y} = 1$).
- **Falso Negativo (FN)**: número de instâncias da classe positiva ($Y = 1$) que foram incorretamente classificadas como negativas ($\hat{y} = 0$). Este é conhecido como Erro Tipo II.

- **Falso Positivo (FP)**: número de instâncias da classe negativa ($Y = 0$, não diabetes) que foram incorretamente classificadas como positivas ($\hat{y} = 1$). Este é conhecido como Erro Tipo I.

- **Verdadeiro Negativo (VN)**: número de instâncias da classe negativa ($Y = 0$) que foram corretamente classificadas como negativas ($\hat{y} = 0$).

Quadro 1 – Estrutura da Matriz de Confusão para Classificação Binária.

| | **Valor Predito** | |
|---|---|---|
| | **Diabetes (1)** | **Não Diabetes (0)** |
| **Diabetes (1)** | VP | FN |
| **Não Diabetes (0)** | FP | VN |

Fonte: Adaptado de (BISHOP, 2006).

A partir da matriz de confusão, diversas métricas de desempenho podem ser calculadas para avaliar diferentes aspectos do classificador (BISHOP, 2006; WEBB; COPSEY, 2011; GOODFELLOW; BENGIO; COURVILLE, 2016):

- **Acurácia:** mede a proporção geral de predições corretas em relação ao total de instâncias.

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

  Embora seja uma métrica intuitiva, a acurácia pode ser enganosa em conjuntos de dados desbalanceados, onde uma classe é muito mais frequente que a outra. Nesses casos, um modelo trivial que sempre prevê a classe majoritária pode atingir alta acurácia, mas falhar em identificar a classe minoritária (frequentemente a de maior interesse, como a presença de doença) (KAVAKIOTIS *et al.*, 2017; CHOWDHURY; AYON; HOSSAIN, 2023).

- **Precisão ou Valor Preditivo Positivo (VPP):** mede a proporção de instâncias classificadas como positivas que são, de fato, positivas. Indica a confiabilidade das predições positivas feitas pelo modelo.

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

- ***Recall* (Sensibilidade ou Taxa de Verdadeiros Positivos - *True Positive Rate* (TPR)):** mede a proporção de todas as instâncias verdadeiramente positivas que foram corretamente identificadas pelo modelo. É uma das principais métricas da capacidade do modelo de detectar a classe de interesse.

$$\textit{Recall} = \frac{VP}{VP + FN}$$

  No diagnóstico médico, como na detecção de diabetes, um alto *Recall* para a classe positiva é frequentemente prioritário. Falhar em detectar um caso real da doença (FN alto)

geralmente acarreta consequências mais severas do que classificar erroneamente um indivíduo saudável como doente (FP alto) (CHOWDHURY; AYON; HOSSAIN, 2023; ASHISHA *et al.*, 2024).

- **Especificidade (*Specificity*) ou Taxa de Verdadeiros Negativos (*True Negative Rate* (TNR)):** mede a proporção de todas as instâncias verdadeiramente negativas que foram corretamente identificadas pelo modelo.

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

- **F1-Score:** é a média harmônica da Precisão e do Recall. Fornece uma métrica única que busca balancear ambas, sendo particularmente útil quando há um desequilíbrio entre as classes ou quando tanto Falsos Positivos quanto Falsos Negativos são custosos.

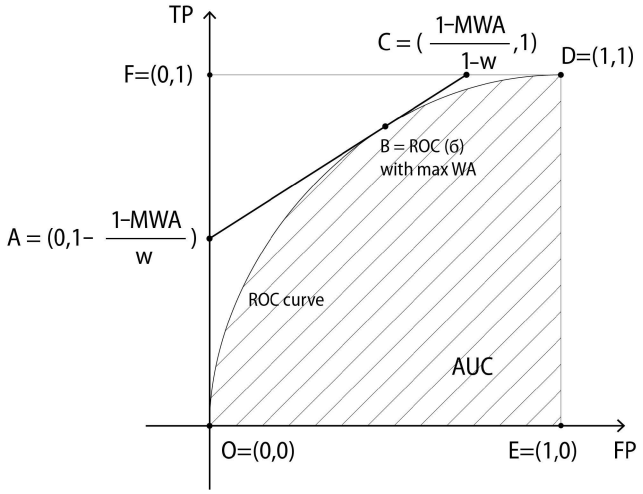
$$\textit{F1-Score} = 2 \times \frac{\text{Precisão} \times \textit{Recall}}{\text{Precisão} + \textit{Recall}} = \frac{2 \times VP}{2 \times VP + FP + FN}$$

- **Área sob a Curva (*Area Under the Curve* (AUC)) *Receiver Operating Characteristic* (ROC) :** a curva ROC é um gráfico que ilustra a capacidade diagnóstica de um classificador binário à medida que seu limiar de discriminação varia. Ela plota a TPR (Sensibilidade) no eixo Y contra a *False Positive Rate* (FPR) no eixo X, onde

$$FPR = 1 - \text{Especificidade} = \frac{FP}{VN + FP}.$$

A AUC representa a área sob esta curva e varia de 0 a 1. Um valor de 0.5 corresponde a um classificador que performa de maneira aleatória, enquanto um valor de 1.0 indica um classificador perfeito.

Figura 1 – Exemplo de curva ROC




Fonte: Le e Nguyen (2022)

A seleção das métricas mais relevantes para avaliação depende intrinsecamente dos objetivos da aplicação e das características do conjunto de dados. No contexto da predição de diabetes, que frequentemente envolve dados desbalanceados, métricas como *Recall* (para a classe de diabetes), F1-*Score* e AUC são consideradas mais informativas e preferíveis à acurácia simples (CHOWDHURY; AYON; HOSSAIN, 2023; ASHISHA *et al.*, 2024).

## 2.4 DETECÇÃO DE DIABETES BASEADA EM APRENDIZADO DE MÁQUINA

Com a digitalização dos cuidados de saúde e a ampla disponibilidade de grandes volumes de dados (*Big Data*), as abordagens baseadas em AM têm emergido como ferramentas promissoras para complementar e, potencialmente, aprimorar a detecção e predição do risco de DM (KAVAKIOTIS *et al.*, 2017). Diferentemente dos escores de risco (discutidos na Seção 2.2), que se baseiam em um conjunto limitado de variáveis e pressupostos lineares, os algoritmos de AM são projetados para aprender padrões complexos, relações não lineares e interações sutis entre um vasto número de variáveis presentes em grandes conjuntos de dados (ASHISHA *et al.*, 2024; LI; PENG; PENG, 2024). Esta capacidade pode se traduzir em modelos com maior acurácia preditiva e melhor poder discriminatório.

A pesquisa em detecção de diabetes utilizando AM é viabilizada pela existência de grandes conjuntos de dados públicos, com destaque para os levantamentos do Centers for Disease Control and Prevention (CDC) dos EUA. O National Health and Nutrition Examination Survey (NHANES) é um conjunto de dados rico e heterogênea, que combina entrevistas com exames físicos e laboratoriais (Centers for Disease Control and Prevention (CDC); National Center for Health Statistics (NCHS), 2017; DINH *et al.*, 2019; ADLER, 2021). Contudo, apesar de sua qualidade, o NHANES frequentemente apresenta desafios como desbalanceamento de classes e a necessidade de tratamento cuidadoso de dados faltantes e da heterogeneidade das variáveis (VANGEEPURAM *et al.*, 2021).

De forma complementar, o BRFSS oferece um volume de dados massivo, coletado por meio de levantamentos telefônicos anuais (CHOWDHURY; AYON; HOSSAIN, 2023; LI; PENG; PENG, 2024). Seus principais desafios incluem o desbalanceamento de classes e a dependência de dados autodeclarados, o que pode introduzir viés de relato (REN, 2024; CHOWDHURY; AYON; HOSSAIN, 2023). Conjuntos de dados adicionais, como o clássico Pima Indian Diabetes Dataset (PIDD) (DIABETES; DIGESTIVE; DISEASES, 2025; RABIE *et al.*, 2022) ou o *Diabetes 130-US hospitals* (CLORE JOHN, 2014), também são utilizados, embora com escopo mais restrito.

As características de entrada ($x$) podem abranger a vasta gama de dados mencionada anteriormente: demográficos (idade, sexo, raça), socioeconômicos (educação, renda), antropométricos (IMC), histórico médico e familiar, hábitos de vida (tabagismo, álcool, dieta, atividade física) e, quando disponíveis, dados laboratoriais (MINISTÉRIO DA SAÚDE, 2023; CHOWDHURY;

AYON; HOSSAIN, 2023; ASHISHA *et al.*, 2024; REN, 2024). Uma vantagem significativa do AM reside na sua capacidade de construir modelos preditivos utilizando dados facilmente coletáveis (e.g., por questionários), permitindo a triagem de risco em larga escala antes da necessidade de exames confirmatórios (ASHISHA *et al.*, 2024; LI; PENG; PENG, 2024; REN, 2024).

## 2.5 ABORDAGENS NO APRENDIZADO DE MÁQUINA PARA DETECÇÃO DE DIABETES

Os métodos utilizados para a detecção de diabetes seguem a abordagem tradicional de aprendizado de máquina, fundamentando-se em etapas clássicas (WEBB; COPSEY, 2011) como mostrado na Figura 2: pré-processamento, seleção/extração de características e classificação.

Figura 2 – Modelo tradicional de aprendizado de máquina para detecção de diabetes.


Fonte: Adaptado de Webb e Copsey (2011)

### 2.5.1 Pré-processamento de Dados

Dados brutos raramente estão prontos para serem utilizados diretamente em modelos de AM, exigindo tratamento para garantir a qualidade e a consistência das informações (ASHISHA *et al.*, 2024). Esta etapa envolve as seguintes operações, quando necessário:

- Limpeza de Dados: tratar dados faltantes (seja por remoção de amostras/características ou por técnicas de imputação, como a média, mediana (ASHISHA *et al.*, 2024), ou algoritmos mais sofisticados, como *bagged trees* (ADLER, 2021)).
- Detecção e Tratamento de *Outliers*: identificar e lidar com valores extremos que podem distorcer o aprendizado do modelo. Técnicas como o método do Interquartile Range (IQR) (ASHISHA *et al.*, 2024) podem ser empregadas.
- Codificação de Variáveis Categóricas: transformar variáveis categóricas em representações numéricas, utilizando métodos como *one-hot encoding* (CHOWDHURY; AYON; HOSSAIN, 2023), *label encoding* ou, como foi proposto, *embeddings*. Esta etapa será detalhada na Seção 2.7.
- Normalização/Padronização de Características Numéricas: colocar as características numéricas em uma escala comum (e.g., entre 0 e 1, ou com média 0 e desvio padrão 1) para evitar que características com maiores magnitudes dominem o processo de aprendizado, especialmente em algoritmos baseados em distância ou gradiente (REN, 2024).

### 2.5.2 Seleção/Extração de Características

A seleção de características é uma etapa fundamental no desenvolvimento de modelos de aprendizado de máquina, cujo objetivo é identificar o subconjunto de variáveis mais relevantes para a tarefa de predição. Este processo não visa apenas reduzir a dimensionalidade e o custo computacional, mas também mitigar o sobreajuste (*overfitting*) e, em muitos casos, melhorar a interpretabilidade do modelo final (BISHOP, 2006). Na literatura sobre detecção de diabetes, são empregados métodos como testes estatísticos (Qui-quadrado (REN, 2024)) e eliminação recursiva de características (ADLER, 2021).

Uma análise dos estudos que utilizam os conjuntos de dados NHANES e BRFSS revela um grupo de características em comum. Fatores demográficos e físicos como idade, gênero, raça/etnia e Índice de Massa Corporal (IMC) são quase universais (ADLER, 2021; QIN *et al.*, 2022; DINH *et al.*, 2019; CICHOSZ; BENDER; HEJLESEN, 2024). Indicadores de saúde cardiovascular, como pressão arterial e níveis de colesterol, também são recorrentes (ADLER, 2021; REN, 2024; PIAS *et al.*, 2023). O conjunto de dados influencia a disponibilidade de variáveis: estudos baseados no NHANES frequentemente incorporam dados laboratoriais, como marcadores sanguíneos (ADLER, 2021; DINH *et al.*, 2019), enquanto trabalhos com o BRFSS, que é um levantamento telefônico,incluem variáveis de autoavaliação da saúde geral e do estado mental (XIE *et al.*, 2019; REN, 2024; PIAS *et al.*, 2023). Além disso, fatores socioeconômicos, como nível educacional e renda, são frequentemente incluídos, refletindo a complexidade multifatorial do risco de diabetes (ADLER, 2021; QIN *et al.*, 2022; CICHOSZ; BENDER; HEJLESEN, 2024).

### 2.5.3 Classificadores

Diversos algoritmos de aprendizado de máquina têm sido aplicados na predição de diabetes (Quadro 2), com a escolha do modelo frequentemente ditada pelas características do conjunto de dados, pela necessidade de interpretabilidade e pelo desempenho em métricas específicas (KAVAKIOTIS *et al.*, 2017). Alguns dos principais classificadores utilizados no problema de detecção incluem:

- **Regressão Logística (LR):** É um modelo linear fundamental para classificação binária, que estima a probabilidade de uma ocorrência com base em uma combinação linear das variáveis de entrada por meio da função logística (HOSMER; LEMESHOW; STURDIVANT, 2013). Em estudos com dados desbalanceados, como o BRFSS, a regressão logística serve como um *baseline* robusto. Chowdhury, Ayon e Hossain (2023), por exemplo, alcançaram um *recall* de 70.7% para a classe diabética ao combinar Logistic Regression (LR) com a técnica de subamostragem Edited Nearest Neighbors (ENN). Contudo, o desempenho pode ser desigual entre subgrupos demográficos; Pias *et al.* (2023) observaram que a mesma arquitetura apresentou um *recall* de apenas 30% para a faixa etária de 30 a 34 anos, evidenciando um viés de idade que pode ser crítico em aplicações de saúde pública.

Quadro 2 – Trabalhos sobre detecção de DM com AM.

| **Artigo** | **conjunto de dados** | **Codificação de Variáveis Categóricas** | **Modelos Avaliados** | **Melhores Resultados** |
|---|---|---|---|---|
| Dinh *et al.* (2019) | NHANES (1999–2014) | Codificação numérica (não especificada) | LR, Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GB), Word Embedding Model (WEM) | Extreme Gradient Boosting (XGBoost): AUC 86.2% (s/ lab); 95.7% (c/ lab) |
| Xie *et al.* (2019) | BRFSS 2014 | – | SVM, Naive Bayes (NB), LR, Redes Neurais Artificiais (RNA), Decision Tree (DT), RF | RNA: AUC 0.79 |
| Adler (2021) | NHANES (2011–2018) | Dummy variables | LR, Linear Discriminant Analysis (LDA), SVM, NB, Multilayer Perceptron (MLP), DT, RF, Boostings | *Diabetes*: SVM AUC 0.92; *Não-diag.*: ElasticNet AUC 0.94 |
| Vangeepuram *et al.* (2021) | NHANES (05-16, jovens) | - | 10 algoritmos de ML (Naive Bayes citado) | ML supera diretriz clínica |
| Rabie *et al.* (2022) | PIDD | *Embeddings* (Keras) | Bidirectional Long Short-Term Memory (BiLSTM), Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), modelos clássicos | BiLSTM: Acurácia 93.07% |
| Qin *et al.* (2022) | NHANES (1999–2020) | One-hot encoding | CatBoost, XGBoost, RF, LR, SVM | CatBoost: AUC 0.83 |
| Chowdhury, Ayon e Hossain (2023) | BRFSS 2021 | One-hot encoding | LR, RF, GB, Adaptive Boosting (AdaBoost), Voting | ENN+GB: Recall 71.7% |
| Ren (2024) | BRFSS 2015 (Kaggle) | – | NB, DT, RF, LR, GB, LDA, CatBoost | CatBoost: Acurácia 86.6% |
| Cichosz, Bender e Hejlesen (2024) | NHANES (2005–2018) | – | RF, AdaBoost, RUSBoost, LogitBoost, MLP | MLP: AUC 0.806 |
| Pias *et al.* (2023) | BRFSS 2021 | One-hot encoding | LR, MLP, NB, *Adaptive Boosting* (AB), RF, K-Nearest Neighbors (KNN) | Enhanced-DP (LR) melhora Recall em jovens |
| Ashisha *et al.* (2024) | BRFSS e PIDD | - | RF, GB, Light Gradient-Boosting Machine (LightGBM), DT | RF (Acurácia BRFSS: 92%, PIDD: 94%) |

Fonte: O Autor (2025).

- **Support Vector Machine (SVM):** Este algoritmo busca o hiperplano que maximiza a margem de separação entre as classes (CORTES; VAPNIK, 1995). Utiliza funções de *kernel* para mapear os dados para espaços de maior dimensão, permitindo a modelagem de relações não lineares. Adler (2021) identificaram a SVM com *kernel* linear como o modelo de melhor desempenho para a predição geral de diabetes no conjunto de dados NHANES, atingindo um AUC de 0.923. Em contrapartida, no estudo de Qin *et al.* (2022), a SVM foi superada por modelos de *ensemble* como o CatBoost, alcançando o menor AUC (0.74) entre os cinco modelos testados, o que indica que sua eficácia pode variar dependendo da composição dos dados e dos outros modelos na comparação.

- **Métodos de Ensemble:** Combinam múltiplos modelos para obter predições mais precisas e estáveis do que as de um modelo individual (GOODFELLOW; BENGIO; COURVILLE, 2016).

  - *Bagging:* O **Random Forest (RF)** é o principal exemplo, construindo um conjunto de árvores de decisão em subamostras dos dados e agregando suas predições por votação majoritária (BREIMAN, 2001). Este método demonstrou alta performance em diversos estudos; Ashisha *et al.* (2024) relataram uma acurácia de 94% no PIDD e 92% no BRFSS ao combinar RF com a seleção de características Boruta e sobreamostragem. Cichosz, Bender e Hejlesen (2024) também utilizaram RF, que apresentou um AUC de 0.786, um resultado competitivo, mas inferior ao da rede neural no mesmo estudo.
  - *Boosting:* Técnicas como AdaBoost, GB, XGBoost, LightGBM e CatBoost, treinam modelos sequencialmente, onde cada novo modelo foca em corrigir os erros do anterior (FRIEDMAN, 2001). Implementações otimizadas de *gradient boosting* frequentemente lideram os resultados de desempenho. Qin *et al.* (2022) encontraram no CatBoost o melhor desempenho (AUC de 0.83) em dados do NHANES, superando RF, LR e SVM. Similarmente, Ren (2024) concluíram que o CatBoost era o classificador mais adequado para o conjunto de dados BRFSS, com 86.6% de acurácia.

- **RNA e Aprendizado Profundo:** Modelos inspirados no cérebro humano que são capazes de aprender representações complexas e hierárquicas dos dados (LECUN; BENGIO; HINTON, 2015). Em uma análise comparativa no conjunto de dados NHANES, Cichosz, Bender e Hejlesen (2024) demonstraram que um modelo de rede neural, embora com desempenho marginalmente superior, alcançou o AUC mais alto (0.806) entre os cinco modelos testados. Da mesma forma, Xie *et al.* (2019) identificaram a rede neural como o modelo com melhor performance (AUC 0.79) em dados do BRFSS. No entanto, a performance de RNA não é universalmente superior; no estudo de Adler (2021), uma rede neural densa não conseguiu aprender padrões úteis, sendo superada por modelos lineares, o que ressalta a sensibilidade dessas arquiteturas à configuração e aos dados.

A seleção e otimização do classificador mais adequado para a detecção de DM envolve a experimentação com vários algoritmos e a avaliação de seu desempenho utilizando métricas apropriadas.

É importante ressaltar que foram priorizados estudos que demonstram rigor metodológico. Trabalhos com resultados excessivamente otimistas, que possuem indícios de problemas como *data leakage* (vazamento de dados), por exemplo, pela aplicação de técnicas de reamostragem antes da divisão dos dados em conjuntos de treino e teste, não foram incluídos na análise comparativa detalhada.

No que diz respeito ao tratamento de variáveis categóricas, a técnica de *one-hot encoding* (ou a criação de variáveis *dummy*, conceitualmente similar) é a abordagem mais reportada (QIN *et al.*, 2022; CHOWDHURY; AYON; HOSSAIN, 2023; ADLER, 2021; PIAS *et al.*, 2023). Alguns estudos não detalham explicitamente o método de codificação, mencionando apenas que as variáveis foram preparadas para os modelos de AM (CICHOSZ; BENDER; HEJLESEN, 2024; VANGEEPURAM *et al.*, 2021; DINH *et al.*, 2019; XIE *et al.*, 2019; REN, 2024). Este cenário evidencia uma lacuna na literatura: a exploração de técnicas de codificação mais avançadas para dados tabulares de saúde é limitada.

Uma exceção é o trabalho de Rabie *et al.* (2022), que empregou *embeddings* aprendidos por uma camada Keras para representar as características do conjunto de dados PIDD antes de alimentar um modelo BiLSTM. Este estudo demonstrou que a combinação de *embeddings* com uma arquitetura de RNA apropriada pode levar a um desempenho superior (acurácia de 93.07%) em comparação com abordagens de AM mais clássicas no mesmo conjunto de dados.

Esta revisão indica que a exploração de técnicas de codificação mais sofisticadas para dados tabulares de saúde é limitada. A maioria dos estudos ainda recorre a métodos de codificação tradicionais, e a aplicação de *embeddings* neste contexto específico permanece relativamente incipiente, com o trabalho de Rabie *et al.* (2022) no PIDD sinalizando um potencial significativo.

## 2.6 PRINCIPAIS DESAFIOS NA DETECÇÃO DE DIABETES COM APRENDIZADO DE MÁQUINA

A aplicação de técnicas de AM para a detecção e predição do risco de diabetes, embora promissora devido à capacidade desses algoritmos de identificar padrões complexos em grandes volumes de dados (KAVAKIOTIS *et al.*, 2017; ASHISHA *et al.*, 2024), não está isenta de desafios. A natureza dos dados de saúde, as limitações intrínsecas dos algoritmos e a necessidade de confiabilidade clínica impõem obstáculos que precisam ser considerados e abordados. Essa análise é crucial para garantir modelos de AM robustos, clinicamente úteis e eticamente responsáveis (CHOWDHURY; AYON; HOSSAIN, 2023; MEHRABI *et al.*, 2021).

### 2.6.1 Variáveis Categóricas

Um desafio fundamental na aplicação de AM a dados de saúde reside no tratamento adequado das **variáveis categóricas**. Diferentemente das variáveis numéricas, que representam quantidades mensuráveis e ordenadas (como idade, IMC ou níveis de glicose), as variáveis categóricas assumem valores que pertencem a um conjunto finito e discreto de categorias ou grupos, sem uma magnitude numérica intrínseca (JAMES *et al.*, 2014). Por exemplo, 'sexo' (masculino/-feminino), 'raça/cor' (branca, preta, parda, etc.), 'status de tabagismo' (fumante, ex-fumante, nunca fumou) ou 'possui plano de saúde' (sim/não) são variáveis categóricas.

Dados provenientes de fontes comuns na área da saúde, como os levantamentos populacionais (e.g., BRFSS (CHOWDHURY; AYON; HOSSAIN, 2023; REN, 2024)) ou os Prontuários Eletrônicos de Saúde (PES), são caracteristicamente heterogêneos, contendo uma mistura substancial dessas variáveis categóricas juntamente com as numéricas (KAVAKIOTIS *et al.*, 2017).

O principal desafio imposto por essas variáveis advém do fato de que a maioria dos algoritmos de AM, particularmente modelos lineares (como LR) e aqueles que dependem de cálculos de distância ou similaridade (como KNN e SVM), foram desenvolvidos para operar com dados de entrada em formato numérico (KAVAKIOTIS *et al.*, 2017; JAMES *et al.*, 2014). A presença de categorias textuais ou nominais impede a aplicação direta desses algoritmos. Portanto, torna-se indispensável transformar as variáveis categóricas em uma representação numérica antes de alimentar os modelos.

A complexidade dessa transformação aumenta ao considerarmos os diferentes tipos de variáveis categóricas:

- *Variáveis Nominais:* as categorias não possuem uma ordem ou hierarquia inerente. Exemplos incluem 'sexo', 'raça/cor', 'estado civil'. Qualquer ordenação atribuída seria arbitrária.
- *Variáveis Ordinais:* existe uma ordem lógica ou classificação entre as categorias. Exemplos incluem 'nível educacional' (ensino fundamental, médio, superior), 'faixa de renda' (baixa, média, alta) ou 'autoavaliação da saúde' (excelente, boa, regular, ruim).

A escolha da técnica de codificação numérica deve, idealmente, levar em conta essa distinção para preservar a informação contida na variável original sem introduzir artefatos ou relações espúrias que possam confundir o algoritmo de aprendizado (JAMES *et al.*, 2014).

As abordagens tradicionais para codificar variáveis categóricas, como a codificação *one-hot* (criação de variáveis *dummy*) ou a codificação por rótulos (*label encoding*), embora amplamente utilizadas, apresentam limitações significativas, especialmente quando se lida com variáveis de alta cardinalidade (muitas categorias distintas) ou quando as relações semânticas entre as categorias são importantes (GUO; BERKHAHN, 2016; GERON, 2019). Essas limitações podem

resultar em espaços de características de alta dimensionalidade e esparsos, ou na introdução de uma ordem artificial indesejada.

A inadequação dessas técnicas em certos cenários motiva a exploração de métodos de representação mais avançados, como os *embeddings*, que buscam aprender representações vetoriais densas e de baixa dimensionalidade capazes de capturar similaridades latentes entre as categorias (GUO; BERKHAHN, 2016; RABIE *et al.*, 2022). As diversas estratégias para representar variáveis categóricas em modelos de AM para detecção de diabetes, incluindo as técnicas tradicionais e a abordagem baseada em *embeddings* explorada neste trabalho, serão discutidas em detalhe na Seção 2.7.

### 2.6.2 Desbalanceamento de Classes

Em muitos problemas de classificação médica, incluindo a detecção de diabetes, o conjunto de dados é comumente desbalanceado. Isso significa que o número de instâncias pertencentes a uma classe (por exemplo, indivíduos não diabéticos, a classe majoritária) é substancialmente maior do que o número de instâncias pertencentes à outra classe (indivíduos diabéticos, a classe minoritária) (CHOWDHURY; AYON; HOSSAIN, 2023). No conjunto de dados BRFSS 2021 utilizado por (CHOWDHURY; AYON; HOSSAIN, 2023), por exemplo, a classe diabética representava apenas cerca de 18% dos dados após o pré-processamento focado em indivíduos com 40 anos ou mais.

O desbalanceamento de classes representa um desafio significativo para os algoritmos de AM, pois muitos deles são projetados assumindo uma distribuição de classes relativamente equilibrada. Na presença de desbalanceamento, os modelos tendem a focar excessivamente na classe majoritária (que contribui mais para a função de perda global, como a acurácia) e a negligenciar ou classificar incorretamente as instâncias da classe minoritária (KAVAKIOTIS *et al.*, 2017; CHOWDHURY; AYON; HOSSAIN, 2023). Isso resulta em modelos com alta acurácia geral, mas com baixo desempenho na identificação da classe de interesse (baixo *Recall* para a classe diabética). Em um contexto médico, onde a falha em detectar a doença (FN) pode ter consequências graves, este comportamento é inaceitável (CHOWDHURY; AYON; HOSSAIN, 2023).

Diversas estratégias foram desenvolvidas na literatura para mitigar os efeitos adversos do desbalanceamento de classes. Essas abordagens podem ser categorizadas da seguinte forma:

- **Nível de Dados:** modificar a distribuição do conjunto de treinamento para torná-lo mais equilibrado.
  - *Undersampling:* remover aleatoriamente instâncias da classe majoritária. Pode levar à perda de informação útil. Exemplos incluem Random Undersampling, ENN (CHOWDHURY; AYON; HOSSAIN, 2023; ALEJO *et al.*, 2010) e NearMiss.

  - *Oversampling:* replicar aleatoriamente instâncias da classe minoritária. Pode levar a *overfitting*. Exemplos incluem Random Oversampling (ROS) (ASHISHA *et al.*, 2024).
  - *Geração Sintética:* criar novas instâncias sintéticas da classe minoritária. O método mais conhecido é o *Synthetic Minority Over-sampling Technique* (SMOTE) (CHAWLA *et al.*, 2002) e suas variantes como *SMOTE for Nominal features* (SMOTE-N) (para dados nominais), *Adaptive Synthetic Sampling* (ADASYN) (LI; PENG; PENG, 2024), Borderline-SMOTE.
  - *Métodos Híbridos:* combinam oversampling e undersampling, como SMOTE-Tomek e *Combination of SMOTE and Edited Nearest Neighbors* (SMOTEENN) (CHOWDHURY; AYON; HOSSAIN, 2023; BATISTA; PRATI; MONARD, 2004).

- **Nível de Algoritmo:** neste nível, modifica-se o algoritmo de aprendizado para considerar os custos de erro de classificação. Uma abordagem comum é o aprendizado sensível a custo, que atribui penalidades maiores para erros na classe minoritária, como falsos negativos, utilizando matrizes de custo ou ajuste do limiar de decisão (ELKAN, 2001; LING; SHENG, 2011).

- **Nível de Ensemble:** algoritmos de ensemble como o RUSBoost e o Balanced Random Forest (BRF) têm se mostrado eficazes para lidar com desbalanceamento. O RUSBoost combina undersampling com boosting, resultando em bom desempenho com menor custo computacional (SEIFFERT *et al.*, 2010). Já o BRF balanceia os dados em cada árvore da floresta, melhorando a detecção da classe minoritária (CHEN; BREIMAN, 2004).

A literatura sobre detecção de diabetes com AM frequentemente emprega abordagens de reamostragem no nível de dados. Ashisha *et al.* (2024) aplicaram ROS. Li, Peng e Peng (2024) compararam ROS, ADASYN, SMOTE e SMOTEENN, identificando SMOTEENN como a mais eficaz. Chowdhury, Ayon e Hossain (2023) avaliaram SMOTE-N, ENN, SMOTE-Tomek e SMOTE-ENN, concluindo que a técnica de undersampling ENN proporcionou o maior ganho no *Recall* da classe diabética para o conjunto de dados BRFSS (focado em >40 anos), embora outras técnicas também tenham mostrado melhorias em relação aos dados originais.

Contudo, é crucial notar que a aplicação de técnicas de reamostragem, especialmente as de oversampling sintético como SMOTE, não é uma "solução mágica". Elas podem introduzir ruído, criar instâncias sintéticas em regiões de sobreposição de classes ou não necessariamente melhorar o desempenho discriminativo geral do modelo (medido por AUC, por exemplo), embora tendem a melhorar o Recall da classe minoritária (CHOWDHURY; AYON; HOSSAIN, 2023). A aplicação dessas técnicas deve ser feita com cautela, apenas no conjunto de treinamento para evitar vazamento de dados (*data leakage*), e seus resultados devem ser avaliados criticamente usando métricas apropriadas para dados desbalanceados (CHOWDHURY; AYON; HOSSAIN, 2023).

### 2.6.3 Interpretabilidade e Confiabilidade

Modelos de AM, especialmente os mais complexos como Gradient Boosting ou Redes Neurais Profundas, são considerados "caixas-pretas" (*black boxes*). Isso significa que, embora possam alcançar alta acurácia preditiva, o processo interno que leva a uma determinada predição pode ser difícil ou impossível de entender para um ser humano (LI; PENG; PENG, 2024; PIAS *et al.*, 2023).

No contexto médico, a falta de interpretabilidade é uma barreira significativa para a adoção clínica. Médicos e pacientes precisam confiar nas predições do modelo e entender quais fatores contribuíram para um diagnóstico ou avaliação de risco (LI; PENG; PENG, 2024). Modelos interpretáveis permitem:

- Validar se o modelo está baseando suas decisões em fatores clinicamente relevantes e não em artefatos ou vieses nos dados.
- Identificar os principais fatores de risco para um determinado paciente, permitindo intervenções personalizadas.
- Explicar as predições aos pacientes, aumentando a adesão ao tratamento e a confiança no sistema.
- Detectar e corrigir possíveis vieses (como o viés relacionado à idade ou a grupos minoritários) no comportamento do modelo.

Técnicas para aumentar a interpretabilidade incluem o uso de modelos inerentemente interpretáveis (como Regressão Logística ou Árvores de Decisão simples) ou a aplicação de métodos *post-hoc* de explicação para modelos complexos, como Local Interpretable Model-Agnostic Explanations (LIME) ou SHapley Additive exPlanations (SHAP) (LI; PENG; PENG, 2024). (LI; PENG; PENG, 2024), por exemplo, utilizaram valores SHAP para analisar a contribuição das *features* em seu modelo *Stacking* para predição de diabetes. A interpretabilidade é, portanto, um requisito cada vez mais importante para o desenvolvimento responsável e a implementação de AM na saúde.

## 2.7 TRATAMENTO DE VARIÁVEIS CATEGÓRICAS

Como abordado na Seção 2.6.1, variáveis categóricas são frequentes em conjuntos de dados do mundo real, especialmente na área da saúde, e representam um desafio particular para algoritmos de AM. A maioria desses algoritmos exige entradas numéricas, tornando necessária a conversão de categorias (como 'sexo', 'nível educacional' ou 'histórico familiar') em uma representação numérica antes da modelagem (KAVAKIOTIS *et al.*, 2017; JAMES *et al.*, 2014). A escolha da técnica de codificação pode impactar significativamente o desempenho e a interpretabilidade do modelo final (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Algumas das técnicas mais

utilizadas incluem *one-hot*, por rótulo e *dummy*. Em contraste com esses métodos, os *embeddings* surgem como uma técnica adaptada do campo de processamento de linguagem natural, que propõe o aprendizado de representações vetoriais densas.

### 2.7.1 Codificação *One-Hot*

A codificação *one-hot*, também conhecida como criação de variáveis *dummy*, é talvez a técnica mais utilizada para lidar com variáveis categóricas nominais (onde não há uma ordem intrínseca entre as categorias) (GUO; BERKHAHN, 2016). O processo consiste em transformar uma única variável categórica com $k$ categorias distintas em $k$ novas variáveis binárias (0 ou 1). Cada nova variável corresponde a uma categoria original, e para cada observação, apenas a variável correspondente à categoria presente terá o valor 1, enquanto todas as outras terão o valor 0 (DAHOUDA; JOE, 2021b). Por exemplo, a codificação *one-hot* da variável 'Tipo Sanguíneo' com as categorias A, B, AB e O resultaria em quatro novas variáveis, conforme ilustrado no Quadro 3.

Quadro 3 – Exemplo de Codificação *One-Hot* para a variável 'Tipo Sanguíneo'.

| **Tipo Sanguíneo (Original)** | **Tipo_A** | **Tipo_B** | **Tipo_AB** | **Tipo_O** |
|---|---|---|---|---|
| A | 1 | 0 | 0 | 0 |
| B | 0 | 1 | 0 | 0 |
| AB | 0 | 0 | 1 | 0 |
| O | 0 | 0 | 0 | 1 |
| A | 1 | 0 | 0 | 0 |

Fonte: O Autor (2025).

Pode-se listar as seguintes desvantagens para a codificação *one-hot*:

- **Alta Dimensionalidade:** se uma variável categórica tiver um grande número de categorias (alta cardinalidade), a codificação *one-hot* pode aumentar drasticamente o número de características (colunas) no conjunto de dados. Isso pode levar à "maldição da dimensionalidade", onde o volume do espaço de características cresce exponencialmente com a dimensão, tornando os dados esparsos, exigindo mais dados para generalização e aumentando a complexidade computacional dos algoritmos (GOODFELLOW; BENGIO; COURVILLE, 2016; GUO; BERKHAHN, 2016).

- **Esparsidade:** a matriz resultante contém uma grande proporção de zeros, o que pode ser ineficiente em termos de armazenamento e processamento para alguns algoritmos (GUO; BERKHAHN, 2016).

- **Perda de Relações Semânticas:** trata cada categoria como completamente independente, ignorando quaisquer similaridades ou relações latentes que possam existir entre elas (GUO; BERKHAHN, 2016). Por exemplo, categorias como 'Cardiologista' e 'Cirurgião Cardíaco' seriam tratadas como igualmente distintas de 'Dermatologista'.

- **Multicolinearidade:** a representação completa (com $k$ colunas para $k$ categorias) introduz colinearidade perfeita (uma coluna pode ser perfeitamente prevista pelas outras). Para alguns modelos (como regressão linear), isso pode ser problemático, sendo comum remover uma das colunas *dummy* (tornando-a a categoria de referência) (KUTNER, 2005).

### 2.7.2 Codificação por Rótulos (*Label Encoding*)

Atribui um número inteiro único a cada categoria de uma variável, iniciando em 0 ou 1 e incrementando sequencialmente (DAHOUDA; JOE, 2021b). Por exemplo, em uma variável ordinal como Nível de Atividade Física, as categorias Sedentário, Levemente Ativo, Moderadamente Ativo e Muito Ativo podem ser mapeadas para os inteiros 0, 1, 2 e 3, respectivamente.

Quadro 4 – Exemplo de Codificação por Rótulos para 'Nível de Atividade Física'.

| **Nível de Atividade Física (Original)** | **Codificado** |
|---|---|
| Sedentário | 0 |
| Levemente Ativo | 1 |
| Moderadamente Ativo | 2 |
| Muito Ativo | 3 |
| Levemente Ativo | 1 |

Fonte: O Autor (2025).

Apesar da simplicidade, essa técnica apresenta desvantagens. A principal delas é a introdução de uma ordem artificial quando aplicada a variáveis nominais. Ao atribuir 0 a Masculino e 1 a Feminino, por exemplo, o método insere uma relação de magnitude que não existe nos dados originais, o que pode levar a interpretações errôneas por parte de alguns algoritmos (GERON, 2019). Adicionalmente, mesmo para variáveis ordinais, a codificação assume que a distância entre categorias adjacentes é uniforme. Isso implica que a diferença entre Levemente Ativo (1) e Moderadamente Ativo (2) seria a mesma que entre Moderadamente Ativo (2) e Muito Ativo (3), uma suposição que pode não refletir a realidade dos dados (GERON, 2019).

Devido a essas limitações, a codificação por rótulos não é recomendada para variáveis nominais em algoritmos que são sensíveis à magnitude ou à ordem das características (BOEHMKE; GREENWELL, 2019).

### 2.7.3 Codificação *Dummy*

A codificação *dummy* é uma variação da codificação *one-hot* que visa especificamente contornar o problema da multicolinearidade perfeita, introduzido quando se cria uma coluna binária para cada categoria de uma variável (JAMES *et al.*, 2014).

A técnica consiste em representar uma variável categórica com $k$ níveis por meio de $k-1$ novas variáveis binárias. Uma das categorias é omitida e se torna a "categoria de referência"(ou *baseline*). Essa categoria é implicitamente representada quando todas as $k-1$ variáveis *dummy* possuem o valor 0 (KUTNER, 2005). A escolha da categoria de referência pode ser arbitrária,

mas frequentemente se utiliza a categoria mais comum para garantir uma base de comparação estável ou uma categoria que represente um estado "padrão"ou de "controle"no contexto do problema (KUHN; JOHNSON, 2019).

Utilizando o mesmo exemplo da Seção 2.7.1, se a categoria 'O' for escolhida como referência para a variável 'Tipo Sanguíneo', a codificação *dummy* resultaria em apenas três novas colunas, como demonstrado no Quadro 5.

Quadro 5 – Exemplo de Codificação Dummy para 'Tipo Sanguíneo'. A categoria 'O' é a de referência.

| **Tipo Sanguíneo (Original)** | **Tipo_A** | **Tipo_B** | **Tipo_AB** |
|---|---|---|---|
| A | 1 | 0 | 0 |
| B | 0 | 1 | 0 |
| AB | 0 | 0 | 1 |
| O | 0 | 0 | 0 |
| A | 1 | 0 | 0 |

Fonte: O Autor (2025).

A principal vantagem desta abordagem é que ela evita a redundância de informações, eliminando a colinearidade perfeita. Isso é essencial para algoritmos de regressão (como a regressão linear e logística), cujos estimadores de coeficientes podem se tornar instáveis ou não identificáveis na presença de colinearidade (JAMES *et al.*, 2014). Além disso, a codificação *dummy* aprimora a interpretabilidade desses modelos. O coeficiente de cada variável *dummy* representa a diferença média no resultado de interesse entre aquela categoria e a categoria de referência, mantendo as outras variáveis constantes (KUTNER, 2005).

Contudo, a codificação *dummy* ainda compartilha desvantagens com a codificação *one-hot*, como o aumento da dimensionalidade e a esparsidade dos dados para variáveis de alta cardinalidade (GUO; BERKHAHN, 2016). A interpretação dos coeficientes do modelo também se torna dependente da categoria escolhida como referência.

### 2.7.4 *Embeddings*

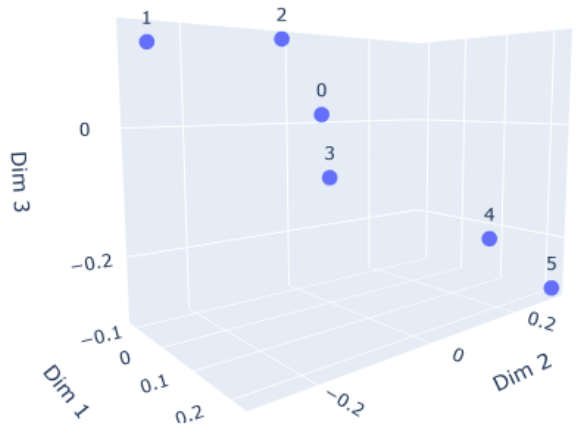
Como alternativa às limitações das codificações tradicionais, os *embeddings* (ou *entity embeddings*) — surgiram como uma técnica importante, impulsionada pelos avanços em RNA (GUO; BERKHAHN, 2016). Esta abordagem propõe o aprendizado de representações vetoriais densas e de baixa dimensionalidade para cada categoria, com o objetivo de capturar suas relações semânticas e funcionais no contexto da tarefa de AM em questão. Originada no campo de Processamento de Linguagem Natural (NLP) para representar palavras (*word embeddings*) (PENNINGTON; SOCHER; MANNING, 2014), a técnica foi estendida com sucesso para dados tabulares (GUO; BERKHAHN, 2016; DAHOUDA; JOE, 2021b).

A ideia central é mapear cada categoria única de uma variável para um vetor de números reais de uma dimensão fixa e relativamente pequena. Formalmente, para uma variável com $k$

categorias, cada categoria é associada a um vetor $\mathbf{v} \in \mathbf{R}^d$. A dimensão do *embedding*, $d$, é um **hiperparâmetro** — uma configuração externa ao modelo cujo valor é definido antes do treinamento e que controla o próprio processo de aprendizado, em contraste com os parâmetros do modelo, que são aprendidos a partir dos dados (GOODFELLOW; BENGIO; COURVILLE, 2016; BISHOP, 2006). Em casos de alta cardinalidade, $d$ é um valor significativamente menor que o número de categorias $k$ (GUO; BERKHAHN, 2016). Diferentemente de técnicas pré-calculadas, no aprendizado supervisionado os vetores de *embedding* são justamente esses parâmetros do modelo, sendo inicializados aleatoriamente e ajustados durante o treinamento.

Essa abordagem pode ser melhor visualizada no seguinte exemplo: um dicionário onde cada palavra (categoria) não é definida por uma descrição textual, mas sim por um conjunto de coordenadas (o vetor de *embedding*) em um mapa multidimensional. Neste mapa, palavras (categorias) com significados ou usos semelhantes tendem a estar localizadas próximas umas das outras.

Figura 3 – Visualização de um *embedding* de uma categoria com 6 classes




Fonte: O Autor (2025).

O aprendizado ocorre dentro de uma arquitetura de rede neural por meio de uma **camada de *embedding*** (*embedding layer*). Essa camada atua como uma **tabela de consulta** (*lookup table*), mantendo uma matriz de pesos onde cada linha é o vetor de *embedding* de uma categoria. Durante o treinamento, os valores categóricos de entrada são utilizados para buscar os vetores correspondentes. Esses vetores, inicialmente aleatórios, são então ajustados via **retropropagação** (*backpropagation*) do erro da tarefa final (e.g., erro na predição de diabetes), otimizando-os para serem informativos para a predição (GOODFELLOW; BENGIO; COURVILLE, 2016; GUO; BERKHAHN, 2016).

A adoção de *embeddings* oferece benefícios substanciais. Primeiramente, resolve o pro-

blema da alta dimensionalidade e esparsidade da codificação *one-hot*, tornando os modelos computacionalmente mais eficientes e menos suscetíveis à "maldição da dimensionalidade" (GUO; BERKHAHN, 2016). A vantagem mais significativa, contudo, é a capacidade de capturar relações semânticas e funcionais latentes. Categorias que influenciam a variável alvo de maneira similar são mapeadas para regiões próximas no espaço vetorial, permitindo que o modelo generalize melhor para categorias raras ou combinações não vistas (GUO; BERKHAHN, 2016). Por exemplo, Guo e Berkhahn (2016) demonstraram que *embeddings* aprendidos para estados alemães em um problema de previsão de vendas reproduziram a geografia real do país, agrupando estados com características econômicas e culturais similares sem que essa informação fosse explicitamente fornecida.

A **dimensionalidade do embedding (**$d$**)** é um hiperparâmetro crucial. Uma dimensão muito baixa pode não ter capacidade expressiva suficiente (subajuste), enquanto uma dimensão excessivamente alta pode levar a sobreajuste (*overfitting*). A escolha ótima geralmente requer experimentação, mas diretrizes empíricas sugerem valores na faixa de $\log_2(|C|)$ até $\min(50, |C|/2)$, onde $|C|$ é o número de categorias (GUO; BERKHAHN, 2016). Adicionalmente, o espaço vetorial resultante pode ser projetado em duas ou três dimensões usando técnicas como *t-distributed Stochastic Neighbor Embedding* (t-SNE) (MAATEN; HINTON, 2008) ou *Principal Component Analysis* (PCA), permitindo a visualização e interpretação das relações aprendidas entre as categorias.

No domínio da detecção de diabetes, onde variáveis categóricas como raça, escolaridade, tabagismo e histórico familiar são preditores importantes, a técnica de *embeddings* oferece um meio promissor de representá-las de forma mais eficaz (YU *et al.*, 2023). Ao codificar essas variáveis em representações densas e ricas, os *embeddings* permitem que modelos de AM, especialmente redes neurais, capturem interações complexas e não lineares entre esses fatores e outras variáveis numéricas (como idade, IMC e pressão arterial), podendo resultar em classificadores mais precisos e robustos.

## 2.8 CONSIDERAÇÕES FINAIS

A análise da literatura revelou que, embora diversos estudos tenham avançado na aplicação de AM para predição de DM, o tratamento de variáveis categóricas frequentemente se baseia em técnicas tradicionais como *one-hot encoding* ou *label encoding*. Tais abordagens, apesar de sua disseminação, podem resultar em representações de alta dimensionalidade e esparsas, ou na introdução de relações ordinais artificiais, limitando potencialmente a capacidade dos modelos. Em contrapartida, a técnica de *embeddings*, que propõe o aprendizado de representações vetoriais densas e semanticamente ricas, emerge como uma alternativa promissora, conforme demonstrado por Guo e Berkhahn (2016) e aplicado com êxito em problemas de saúde, como no estudo de Rabie *et al.* (2022) para o conjunto de dados PIDD.

A incorporação de *embeddings* para o tratamento de variáveis categóricas na detecção de DM detém considerável potencial para avanços na prática clínica. Destaca-se, por exemplo, seu impacto em estratégias de triagem em larga escala, onde a melhor capacidade de processar dados heterogêneos pode conduzir a modelos preditivos mais acurados e, consequentemente, à otimização da alocação de recursos diagnósticos. Esta perspectiva fundamenta a investigação, que se propõe a desenvolver e avaliar modelos de AM para detecção de DM utilizando *embeddings* no contexto do conjunto de dados NHANES. O arcabouço teórico aqui consolidado estabelece, assim, a base para a metodologia e as análises experimentais que serão detalhadas nos capítulos subsequentes.

# 3 PROPOSTA DE SOLUÇÃO

Este capítulo descreve a metodologia empregada para o desenvolvimento e a avaliação de modelos de AM para a detecção de DM utilizando dados do NHANES. A Seção 3.1 descreve o conjunto de dados NHANES, a extração dos dados e a seleção das variáveis preditoras. A Seção 3.2 apresenta a arquitetura baseada em uma RNA do tipo MLP com camadas de *embedding* para o tratamento de variáveis categóricas e detalha as etapas de pré-processamento e classificação. Por fim, o capítulo conclui com a descrição das ferramentas e bibliotecas utilizadas na Seção 3.3.

## 3.1 FONTE DE DADOS

O conjunto de dados utilizado neste trabalho é o NHANES, um programa de levantamento conduzido pelo National Center for Health Statistics (NCHS), parte do CDC dos Estados Unidos (Centers for Disease Control and Prevention (CDC); National Center for Health Statistics (NCHS), 2017). O NHANES é desenhado para avaliar o estado de saúde e nutricional de adultos e crianças nos EUA, combinando entrevistas domiciliares com exames físicos e laboratoriais, realizados em Mobile Examination Center (MEC)s (Centers for Disease Control and Prevention (CDC); National Center for Health Statistics (NCHS), 2017). Para este trabalho, o foco está nos dados coletados nos ciclos bienais compreendidos entre 2005 e 2018.

Especificamente para a detecção de diabetes, o NHANES disponibiliza marcadores bioquímicos relevantes. Um destes marcadores é a Hemoglobina Glicada (HbA1c), a qual se destaca por sua ampla cobertura amostral. A HbA1c é medida na maioria dos participantes examinados com 12 anos ou mais, independentemente da condição de jejum, o que resultou em um volume de dados superior ao de outros marcadores, como a GPJ. A coleta de GPJ, por requerer jejum prévio, é realizada apenas em uma subamostra de participantes, o que limitou o número de observações disponíveis. A Tabela 1 ilustra a diferença na dimensão amostral entre esses componentes nos ciclos do NHANES.

A escolha do NHANES fundamentou-se em várias vantagens substanciais. Diferentemente de outros levantamentos como o BRFSS (CHOWDHURY; AYON; HOSSAIN, 2023; REN, 2024), o NHANES inclui resultados de exames laboratoriais objetivos, como a dosagem de GPJ e HbA1c. Isso permitiu a definição da condição de diabetes com base em critérios bioquímicos padronizados, mitigando o viés de autorrelato. Este viés pode levar à subnotificação de casos ou a erros de classificação (VANGEEPURAM *et al.*, 2021; ADLER, 2021). Adicionalmente, o NHANES fornece uma ampla gama de variáveis preditivas, cobrindo aspectos demográficos, socioeconômicos, comportamentais, histórico médico, dados antropométricos e clínicos.

A população de estudo é definida para incluir **adultos com 20 anos ou mais**. São ex-

Tabela 1 – Número de registros em arquivos selecionados dos últimos ciclos do NHANES.

| Ciclo (ano) | Entrevistados (DEMO)* | Glicose em Jejum | HbA1c |
|---|---|---|---|
| 2009–2010 | 10.537 | 3.581 | 7.369 |
| 2011–2012 | 9.756 | 3.239 | 6.549 |
| 2013–2014 | 10.175 | 3.329 | 6.979 |
| 2015–2016 | 9.971 | 3.191 | 6.744 |
| 2017–2018 | 9.254 | 3.036 | 6.401 |
| 2017–Mar 2020† | 15.560 | 5.090 | 10.409 |
| 2021–2022‡ | 8.860 | 3.996 | 7.199 |

Fonte: Adaptado da documentação do NHANES (National Center for Health Statistics, 2024; SHEPHERD *et al.*, 2021).

Nota: *Número de registros no arquivo `DEMO` (todos que completaram a entrevista domiciliar).

†: Ciclo especial de 3,2 anos (2017–2018 + parte de 2019–2020 pré-pandemia).

‡: Ciclo 2021–2022 com amostra reduzida após interrupções pela COVID-19.

cluídas **mulheres grávidas** no momento da coleta, pois a gravidez pode induzir alterações temporárias no metabolismo da glicose. Após a aplicação do pré-processamento (detalhado na Seção 3.2.1), a amostra final para análise é constituída. Assegura-se que cada participante seja representado uma única vez na amostra final, mesmo que tenha participado em múltiplos ciclos do NHANES, através da utilização de identificadores únicos.

### 3.1.1 Extração de dados

Os dados brutos são publicamente disponibilizados pelo CDC em seu website oficial, organizados em ciclos bienais. Para cada ciclo, os dados são segmentados em arquivos específicos por componente. A extração consiste na fusão dos arquivos relevantes dos ciclos de **2005-2006** a **2017-2018**. Para assegurar robustez estatística e um tamanho amostral adequado, a combinação de múltiplos ciclos é uma prática recomendada Johnson *et al.* (2013).

A classificação de um participante como tendo diabetes ($Y = 1$) foi baseada nos critérios diagnósticos recomendados pela ADA (2023), utilizando tanto informações de autorrelato quanto dados laboratoriais disponíveis no NHANES:

1. Autorrelato de Diagnóstico Médico: participantes que responderam "sim" à pergunta sobre já terem sido informados por um médico ou profissional de saúde que tinham diabetes;
2. HbA1c: nível de HbA1c $\geq 6.5\%$ (48 mmol/mol), medido em amostra laboratorial.

Um participante é classificado como tendo diabetes ($Y = 1$) se **pelo menos um** desses critérios for satisfeito. Consequentemente, o rótulo positivo engloba todos os subtipos de DM, sem distinção entre DM1 ou DM2. Participantes que não atendem a nenhum dos critérios descritos são classificados como não tendo diabetes ($Y = 0$). Esta definição combinada

visa capturar tanto os casos já diagnosticados quanto os não diagnosticados identificados pelos critérios laboratoriais.

A escolha das variáveis para o classificador baseia-se na análise da literatura focada na predição de DM utilizando dados do NHANES. O processo visa identificar a interseção de variáveis empregadas por múltiplos trabalhos, estabelecendo um conjunto de preditores que representa o consenso sobre os fatores de risco mais relevantes. O critério de seleção privilegia variáveis que demonstram alta frequência de uso em estudos prévios, o que indica sua forte associação com o desfecho final. As 28 variáveis selecionadas são organizadas em 5 categorias principais: demográficas e socioeconômicas, antropométricas, clínicas e fisiológicas, laboratoriais e estilo de Vida. As variáveis selecionadas são listas a seguir, com os códigos correspondentes no banco de dados do NHANES:

- Fatores Demográficos e Socioeconômicos: características basais e contextuais do indivíduo.
  - Sexo (`RIAGENDR`)
  - Idade (`RIDAGEYR`)
  - Etnia (`RIDRETH1`)
  - Nível Educacional (`DMDEDUC2`)
  - Razão Renda/Pobreza (`INDFMPIR`)
- Antropométricas: indicadores diretos de medidas corporais.
  - Peso (kg) (`BMXWT`)
  - Altura (cm) (`BMXHT`)
  - Índice de Massa Corporal (IMC) (`BMXBMI`)
  - Circunferência da Cintura (`BMXWAIST`)
- Clínicas e Fisiológicas: medições que refletem o estado de saúde.
  - Pressão Arterial Sistólica (`BPXSY1`)
  - Pressão Arterial Diastólica (`BPXDI1`)
- Laboratoriais: análises bioquímicas de amostras biológicas.
  - Colesterol HDL (`LBDHDD`)
  - Colesterol Total (`LBXTC`)
  - Osmolalidade Sanguínea (`LBXSOSSI`)
  - Nitrogênio Ureico no Sangue (`LBXSBU`)

- Estilo de Vida, Dieta e Histórico de Saúde: informações obtidas por meio de questionários e sobre hábitos.
  - Atividade física moderada (`PAQ665 / PAD320`)
  - Diagnóstico de Pressão Alta (`BPQ020`)
  - Em tratamento para Hipertensão (`BPQ050A`)
  - Saúde geral autoavaliada (`HSD010`)
  - Duração do Sono (`SLD012 / SLD010H`)
  - Ingestão de Calorias (`DR1TKCAL`)
  - Ingestão de Carboidratos (`DR1TCARB`)
  - Ingestão de Proteínas (`DR1TPROT`)
  - Ingestão de Gorduras (`DR1TTFAT`)
  - Ingestão de Fibras (`DR1TFIBE`)
  - Ingestão de Açúcar (`DR1TSUGR`)
  - Ingestão de Sódio (`DR1TSODI`)
  - Histórico Familiar de Diabetes (`DIQ170`)

Desse modo, o conjunto de 28 preditores selecionados para o estudo é composto por 7 variáveis categóricas (como sexo e nível educacional) e 21 variáveis numéricas contínuas (como idade, IMC e resultados laboratoriais). Essa distinção é fundamental para a etapa de pré-processamento, na qual cada tipo de variável recebe o tratamento adequado antes de ser utilizada no treinamento dos modelos de classificação.

## 3.2 ARQUITETURA DO SISTEMA DE DETECÇÃO COM *EMBEDDING*

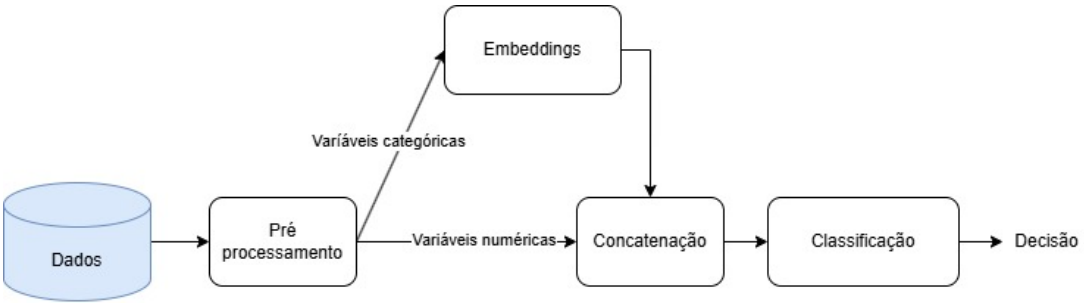
A arquitetura proposta adota uma abordagem híbrida em duas fases para a detecção de diabetes. Na primeira fase, uma RNA do tipo MLP, contendo camadas de *embedding* dedicadas para cada variável categórica, é treinada. O objetivo principal desta fase é aprender representações vetoriais ricas e densas (*embeddings*) para as variáveis categóricas, além de representar a performance do classificador RNA em si. Os pesos dessas camadas são inicializados e ajustados conjuntamente com os demais parâmetros da rede durante o treinamento, por meio do algoritmo de retropropagação de erro (RUMELHART; HINTON; WILLIAMS, 1986).

Uma vez que a rede neural é treinada, os vetores de *embedding* aprendidos são extraídos e utilizados como novas características. Na segunda fase, estas características de *embedding* são concatenadas com as variáveis numéricas originais (previamente normalizadas), criando um conjunto de dados enriquecido. Este novo conjunto de dados serve então como entrada para um

conjunto diversificado de algoritmos de aprendizado de máquina, como XGBoost e SVM, para a tarefa de classificação final.

Esta abordagem permite que modelos clássicos, que não lidam nativamente com a semântica de dados categóricos, se beneficiem das relações complexas capturadas pelas camadas de *embedding* durante o treinamento da rede neural. Para variáveis categóricas binárias, utiliza-se uma codificação numérica simples (0 e 1), enquanto as multiclasse são mapeadas por meio dos *embeddings*. O desempenho desta arquitetura foi comparado com o de modelos treinados utilizando a codificação *One-Hot*.

Figura 4 – Diagrama do Sistema de Detecção de Diabetes Proposto.




Fonte: Autor (2025)

O desempenho dos modelos finais foi avaliado no conjunto de teste utilizando o conjunto de métricas detalhado na Seção 2.3. Dada a natureza do problema e o desbalanceamento, ênfase particular foi dada à Curva ROC, AUC, Acurácia, Precisão, *Recall* (Sensibilidade), Especificidade e F1-Score. A matriz de confusão também foi analisada.

### 3.2.1 Pré-processamento de dados

Os dados brutos extraídos do NHANES passam por um processo de limpeza e tratamento antes da modelagem:

- Tratamento de Dados Faltantes: registros com valores ausentes na variável alvo ou em um conjunto mínimo de preditores críticos (e.g., idade, sexo, IMC) são removidos.
- Tratamento de *Outliers*: para o tratamento de valores atípicos, adota -se uma abordagem baseada em critérios de plausibilidade clínica e fisiológica. Essa escolha metodológica visa remover registros com valores clinicamente implausíveis, que são mais prováveis de decorrer de erros de medição ou de digitação do que de variações biológicas extremas. Dada a ausência de um consenso único na literatura para os limites de exclusão de todas as variáveis, a definição desses limites é fundamentada sobre faixas de valores razoáveis para a população adulta. São considerados válidos os registros com valores dentro das seguintes faixas: altura (130–240 cm), peso (30–450 kg), índice de massa corporal (10–60 kg/m²), circunferência da cintura (50–200 cm), pressão arterial sistólica (50–220 mmHg)

e diastólica (40–140 mmHg), colesterol HDL (10–120 mg/dL), colesterol total (80–500 mg/dL), osmolalidade sanguínea (240–350 mOsm/kg), nitrogênio ureico (3–200 mg/dL) e duração do sono (2–14 horas). Registros que apresentam valores fora desses limites pré-definidos são, portanto, excluídos da análise final.

- Normalização de Variáveis Numéricas: as variáveis numéricas contínuas são padronizadas (normalização Z-score) para que tenham média 0 e desvio padrão 1. Os parâmetros de média e desvio padrão são calculados no conjunto de treinamento e aplicados a todos os conjuntos de dados (validação e teste).

### 3.2.2 Classificação

A etapa de classificação corresponde à segunda fase da arquitetura proposta. Nela, um conjunto de algoritmos de aprendizado de máquina é treinado e avaliado utilizando o conjunto de dados enriquecido, no qual as variáveis categóricas são substituídas por suas respectivas representações de *embedding* extraídas da rede neural. O desempenho desta abordagem foi comparado com o de modelos de referência treinados com a codificação tradicional *one-hot*. O conjunto de classificadores inclui:

- LR: um modelo linear fundamental, utilizado como *baseline* devido à sua simplicidade, rapidez de treinamento e interpretabilidade.
- LDA: um classificador linear que modela a distribuição de cada classe para estimar a probabilidade de pertencimento, servindo como outro *baseline* estatístico robusto.
- SVM: um classificador que busca o hiperplano que maximiza a margem de separação entre as classes.
- GB: um método de *ensemble* que constrói modelos de forma sequencial, onde cada novo modelo corrige os erros do anterior.
- XGBoost: uma implementação otimizada do algoritmo *gradient boosting*, reconhecida por sua eficiência e pela inclusão de técnicas de regularização.
- LightGBM: outra implementação de *gradient boosting* de alta performance, que utiliza uma estratégia de crescimento de árvores baseada em folhas.
- MLP: a mesma arquitetura de rede neural utilizada na primeira fase, agora avaliada como um classificador final sobre os dados enriquecidos para fins de comparação.

Adicionalmente, são construídos modelos de *ensemble* que combinam as predições de múltiplos classificadores individuais. São exploradas duas estratégias de combinação:

- *Hard Voting* (Votação Majoritária): o rótulo final é determinado pela classe que recebe o maior número de votos entre os classificadores base.
- *Soft Voting* (Votação Ponderada por Probabilidades): o rótulo final é atribuído à classe com a maior média de probabilidades preditas entre os classificadores base.

O protocolo de validação inicia-se com uma divisão estratificada do conjunto de dados em **80% para treinamento e 20% para um conjunto de teste (*hold-out*)**. Este último foi mantido isolado e utilizado exclusivamente para a avaliação final. Sobre o conjunto de 80%, foi aplicada a técnica de **validação cruzada k-fold estratificada** (com $k = 5$ folds) para obter uma estimativa de desempenho estável e guiar a otimização de hiperparâmetros. Concluída a validação cruzada, o modelo final foi treinado com a totalidade dos 80% dos dados e sua performance foi aferida no conjunto de teste.

A prevalência de diabetes em populações gerais resulta em conjuntos de dados desbalanceados. Para mitigar o viés do modelo em favor da classe majoritária, os modelos são treinados e avaliados em dois cenários distintos:

1. Sem Balanceamento: Utilizando o conjunto de treinamento com sua distribuição original.
2. Treinamento com dados balanceados: Aplica-se a técnica de sobreamostragem SMOTE (CHAWLA *et al.*, 2002). Esta abordagem gera instâncias sintéticas da classe minoritária. Para evitar o vazamento de dados, a aplicação do SMOTE ocorre exclusivamente no conjunto de treinamento de cada iteração da validação cruzada, após a divisão dos dados (BATISTA; PRATI; MONARD, 2004).

#### 3.2.2.1 Otimização de *threshold*

A otimização do limiar de decisão (*decision threshold*) é empregada como um procedimento complementar para refinar o desempenho dos classificadores. A maioria dos algoritmos de classificação gera uma probabilidade contínua, que por padrão é convertida em uma classe binária utilizando um limiar de 0,5. Contudo, em cenários com dados desbalanceados, este valor padrão é frequentemente subótimo.

A determinação deste limiar ótimo é integrada ao protocolo de validação cruzada. Dentro de cada *fold*, o modelo é treinado na partição de treino e utilizado para prever as probabilidades na partição de validação. A partir dessas probabilidades, é gerada uma curva de precisão-*recall*, e o limiar que maximiza o F1-Score para a classe positiva (diabéticos) é identificado. O limiar final para cada algoritmo é definido como a média aritmética dos limiares ótimos encontrados em todos os *folds*. Este valor médio é, então, utilizado para converter as predições de probabilidade em classificações binárias finais no conjunto de teste.

### 3.2.3 Análise de Interpretabilidade

A interpretabilidade dos modelos de AM é um requisito essencial para sua validação clínica e adoção em contextos de saúde. A fim de superar as limitações de modelos "caixa-preta"(*black box*), esta proposta inclui uma etapa de análise do classificador que apresentar o melhor desempenho preditivo.

A análise é conduzida por meio de duas abordagens complementares. Primeiramente, é examinada a importância global das características, extraída diretamente do modelo treinado. Esta métrica quantifica a contribuição de cada preditor para a performance do classificador, permitindo hierarquizar os fatores de maior impacto. Em seguida, emprega-se a metodologia SHAP (*Shapley Additive exPlanations*) para uma análise mais detalhada. Baseado em conceitos da teoria dos jogos, o SHAP atribui um valor de importância a cada característica para predições individuais, oferecendo explicações tanto locais quanto globais sobre o comportamento do modelo. A análise é realizada sobre o conjunto de teste para garantir que as conclusões se apliquem a dados não vistos durante o treinamento.

## 3.3 AMBIENTE COMPUTACIONAL E FERRAMENTAS

A execução dos experimentos computacionais e o treinamento dos modelos foram realizados em uma estação de trabalho pessoal. O ambiente de hardware utilizado consistiu em um processador Intel Core i5-8400 de 6 núcleos e 6 threads, e 16 GB de memória RAM.

A implementação da proposta é realizada utilizando a linguagem **Python**. As principais bibliotecas incluem:

- Pandas e NumPy para manipulação de dados.
- Scikit-learn para os algoritmos de AM, pré-processamento e validação.
- TensorFlow/Keras para a RNA e as camadas de *embedding*.
- Imbalanced-learn para a técnica SMOTE.
- Matplotlib e Seaborn para visualização de dados.

# 4 RESULTADOS E DISCUSSÃO

Este capítulo apresenta os resultados obtidos a partir da aplicação da solução proposta. A Seção 4.1 apresenta uma análise descritiva detalhada do conjunto de dados final, utilizado para o treinamento e a avaliação dos modelos. A Seção 4.2 detalha o desempenho dos algoritmos de classificação em diferentes cenários experimentais, avaliando o impacto da inclusão de variáveis categóricas e das estratégias de codificação. Por fim, a Seção 4.3 aborda os efeitos das técnicas de balanceamento de classes no desempenho dos modelos.

## 4.1 CARACTERIZAÇÃO DO CONJUNTO DE DADOS

O conjunto de dados NHANES contém 70.190 registros e 33 variáveis. Após a aplicação dos critérios de inclusão e exclusão, que consistem na remoção de participantes com menos de 20 anos, gestantes e registros com dados ausentes na variável alvo ou em preditores essenciais, o número de amostras do conjunto é reduzido a **25.546** participantes.

### 4.1.1 Análise Descritiva

O conjunto de dados final é composto por 21 variáveis numéricas e 7 variáveis categóricas. A idade dos participantes varia de 20 a 85 anos, com uma média de 49,3 anos (desvio padrão de 17,5 anos), indicando uma população adulta com ampla representatividade etária. A distribuição de gênero é equilibrada, com 51,0% de participantes do sexo masculino (13.025) e 49,0% do sexo feminino (12.521). A composição étnico-racial da amostra, apresentada na Tabela 2, reflete a diversidade da população norte-americana, com predominância de indivíduos brancos não hispânicos (45,5%).

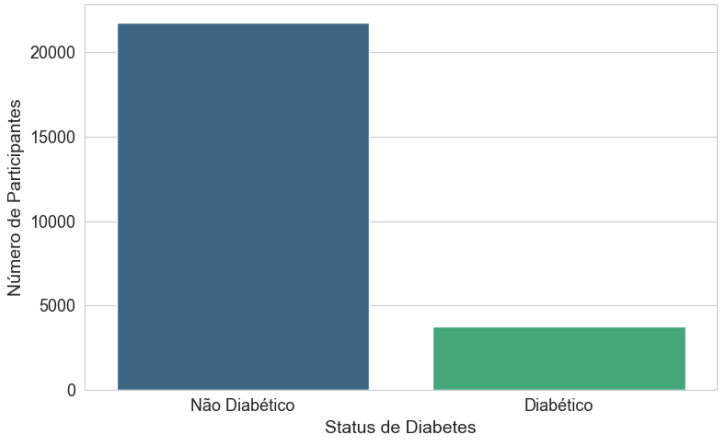
Tabela 2 – Distribuição dos participantes por etnia na amostra final.

| **Etnia** | **Contagem** | **Percentual (%)** |
|---|---|---|
| Branco Não-Hispânico | 11.625 | 45,51 |
| Negro Não-Hispânico | 5.100 | 19,96 |
| Mexicano-Americano | 3.923 | 15,36 |
| Outra Raça | 2.555 | 10,00 |
| Outro Hispânico | 2.343 | 9,17 |
| **Total** | **25.546** | **100,00** |

Fonte: Elaborado pelo autor (2025).

Do total de 25.546 participantes, 3.773 (14,77%) foram classificados como portadores de diabetes, enquanto 21.773 (85,23%) foram classificados como não portadores. Essa proporção configura um cenário de desbalanceamento de classes, onde a classe positiva (diabetes) é minoritária. A Figura 5 ilustra visualmente esta distribuição.

Figura 5 – Distribuição da variável alvo (Diabetes) no conjunto final.




Fonte: Autor (2025).

### 4.1.2 Visualização das Variáveis

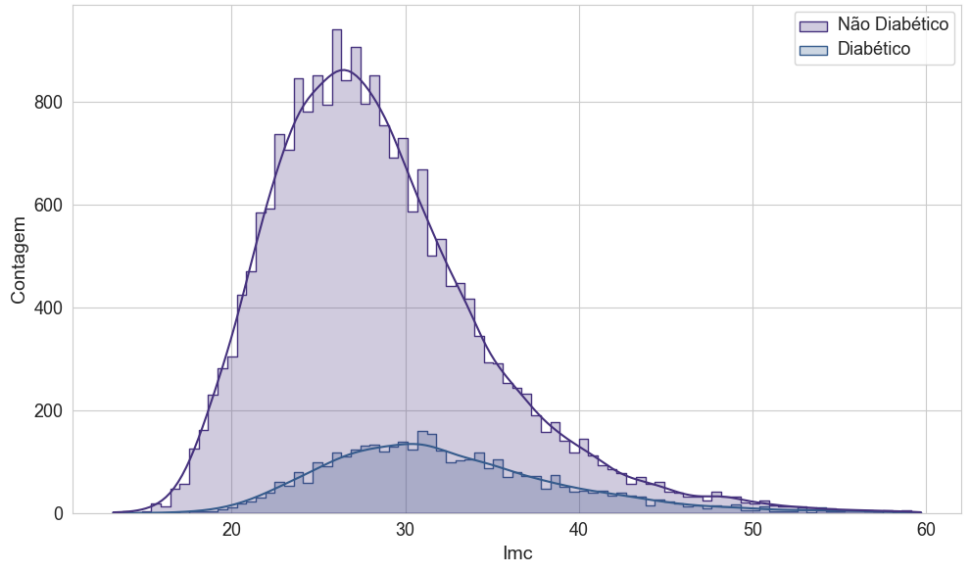
Para compreender as relações entre as variáveis preditoras e a variável alvo, foram geradas visualizações que comparam as distribuições para os grupos com e sem diabetes. A Figura 6 exibe a distribuição do Índice de Massa Corporal (IMC) para ambos os grupos. Observa-se que a mediana e a dispersão dos valores de IMC são mais elevadas no grupo de participantes com diabetes (Y=1) em comparação ao grupo sem diabetes (Y=0). Essa observação está alinhada com a literatura que estabelece a obesidade, frequentemente mensurada pelo IMC, como um dos principais fatores de risco para o desenvolvimento de diabetes tipo 2 (QIN *et al.*, 2022; DINH *et al.*, 2019). Um IMC mais alto geralmente indica maior adiposidade corporal, que está associada à resistência à insulina, um precursor chave da doença (ADA, 2023). No entanto, a sobreposição considerável entre as distribuições nos dois grupos, visível no gráfico, sugere que, embora o IMC seja um indicador relevante, ele isoladamente não possui capacidade discriminatória suficiente para um diagnóstico preciso, justificando a necessidade de modelos multivariados.

## 4.2 DESEMPENHO DOS MODELOS DE CLASSIFICAÇÃO

A avaliação dos modelos foi conduzida em três cenários experimentais para mensurar o impacto das variáveis categóricas e das técnicas de codificação. Os resultados foram avaliados com base em um conjunto de métricas, com destaque para a Área Sob a Curva ROC (AUC) e o F1-Score, por sua robustez em cenários com desbalanceamento de classes.

A arquitetura do MLP é composta por três camadas ocultas com 256, 128 e 64 neurônios, respectivamente, utilizando a função de ativação ReLU, normalização em lote (*Batch Normalization*) e regularização via *Dropout* (com taxas de 0.5, 0.4 e 0.3) para mitigar o sobreajuste. A camada de saída utilizou a função sigmoide para classificação binária, e o modelo foi otimizado com o algoritmo Adam, com taxa de aprendizado de 0.003. A definição destas configurações, assim como a dos hiperparâmetros dos demais classificadores, foi conduzida por meio

Figura 6 – Distribuição do IMC por diagnóstico de diabetes.




Fonte: Autor (2025).

de um processo iterativo de experimentação empírica, no qual os valores foram selecionados mediante a observação dos melhores resultados de desempenho. Para os modelos de *ensemble*, o XGBoost foi configurado com hiperparâmetros como taxa de aprendizado de 0.03, 600 estimadores e profundidade máxima de 6. O *Random Forest* utilizou o critério "gini"para divisão dos nós, e o SVM foi implementado com um *kernel* de função de base radial (RBF).

Para determinar se as diferenças de desempenho observadas entre os modelos são estatisticamente significantes, foi empregado o teste Z para a comparação de duas proporções, aplicado a métricas como acurácia e *recall* (AGRESTI, 2002). Este procedimento avalia se a diferença entre as taxas de sucesso de dois modelos, calculadas sobre o mesmo conjunto de teste, é real ou se pode ser atribuída à variabilidade amostral. O teste opera sob a hipótese nula ($H_0$) de que não existe diferença real no desempenho entre os modelos ($p_1 = p_2$). A análise consiste no cálculo de uma estatística Z (*Z-score*), que quantifica a magnitude da diferença observada em relação à variabilidade esperada caso a hipótese nula seja verdadeira. Adotou-se um nível de significância ($\alpha$) de 0,05 para um teste bilateral. Um valor-p resultante inferior a 0,05, ou um Z-score que exceda os valores críticos de $\pm 1,96$, leva à rejeição da hipótese nula. Tal resultado indica que a superioridade de uma abordagem sobre a outra é estatisticamente significante.

### 4.2.1 Avaliação utilizando somente dados numéricos

Este experimento avalia o poder de discriminação das variáveis numéricas. O objetivo é ter uma referência de desempenho quando as variáveis categóricas forem incluídas. A Tabela 3 apresenta os resultados dos classificadores. Apesar de que o *Gradient Boosting* obteve o melhor desempenho, a diferença do seu resultado é estatisticamente significante em relação ao resultado do XGBoost. Não é possível avaliar se classificadores lineares ou não-lineares, paramétricos e

não-paramétricos e generativos ou discriminativos fornecem um modelo mais apropriado para este problema.

Um padrão comum observado foi o baixo *Recall*, que não ultrapassou 0,45 para nenhum modelo. O *Recall*, ou sensibilidade, é a métrica que quantifica a proporção de casos positivos reais (indivíduos com diabetes) que foram corretamente identificados pelo modelo. Um valor que não excede 0,45 indica que mais da metade dos participantes com diabetes no conjunto de teste foi classificada incorretamente como não diabética, gerando um alto número de falsos negativos. Este resultado evidencia a dificuldade dos algoritmos em aprender a reconhecer a classe minoritária e é uma consequência esperada do desbalanceamento de classes.

Tabela 3 – Resultados dos modelos treinados apenas com variáveis numéricas.

| **Modelo** | **AUC** | **Acurácia** | **Recall** | **Precisão** | **F1-Score** |
|---|---|---|---|---|---|
| MLP | 0,829 | 0,859 | 0,652 | 0,446 | 0,530 |
| *Gradient Boosting* | 0,830 | 0,860 | 0,644 | 0,448 | 0,528 |
| LDA | 0,821 | 0,838 | 0,656 | 0,431 | 0,520 |
| LightGBM | 0,814 | 0,853 | 0,676 | 0,419 | 0,518 |
| Regressão Logística | 0,822 | 0,844 | 0,640 | 0,430 | 0,514 |
| XGBoost | 0,823 | 0,852 | 0,617 | 0,430 | 0,507 |
| Random Forest | 0,817 | 0,842 | 0,620 | 0,419 | 0,500 |

Fonte: Elaborado pelo autor (2025).

### 4.2.2 Avaliação com variáveis categóricas

Neste experimento, as 7 variáveis categóricas foram incorporadas no modelo utilizando dois métodos de codificação: *One-Hot Encoding* e camadas de *embedding*. Na codificação *One-Hot*, a transformação das 7 variáveis categóricas resultou em um aumento da dimensionalidade do conjunto de dados de 28 para 45 preditores. A Tabela 4 mostra o desempenho dos modelos utilizando esta codificação. Pode-se observar um aumento no desempenho de todos os modelos. Por exemplo, o F1-score do *Gradient Boosting* obteve um aumento de 6,6% em relação ao resultado utilizando somente variáveis numéricas. O modelo XGBoost alcançou o maior F1-Score (0,571), com um ganho de 12,6% em relação à sua performance no cenário anterior. Todos os modelos apresentaram melhorias em AUC, *Recall* e F1-Score, o que confirma a importância das informações contextuais para a detecção de diabetes.

Na abordagem com *embeddings*, a representação das variáveis categóricas foi obtida por meio de um processo de duas etapas, que utilizou uma rede neural como ferramenta para engenharia de características. Primeiramente, um modelo MLP foi desenvolvido e treinado com o objetivo principal de classificar os participantes. Durante este treinamento, as camadas de *embedding* e as camadas de classificação (densas) foram otimizadas simultaneamente via retropropagação. Desse modo, os vetores de *embedding* para cada categoria foram ajustados para se tornarem representações numéricas informativas e úteis para a tarefa de predição de diabe-

Tabela 4 – Resultados dos modelos com variáveis numéricas e categóricas codificadas via *One-Hot*.

| **Modelo** | **AUC** | **Acurácia** | **Recall** | **Precisão** | **F1-Score** |
|---|---|---|---|---|---|
| XGBoost | 0,849 | 0,880 | 0,678 | 0,492 | 0,571 |
| LightGBM | 0,843 | 0,882 | 0,693 | 0,478 | 0,566 |
| Gradient Boosting | 0,862 | 0,882 | 0,604 | 0,528 | 0,563 |
| MLP | 0,858 | 0,882 | 0,620 | 0,515 | 0,563 |
| SVM | 0,833 | 0,878 | 0,702 | 0,458 | 0,554 |
| Regressão Logística | 0,837 | 0,873 | 0,674 | 0,465 | 0,550 |
| Random Forest | 0,839 | 0,866 | 0,637 | 0,468 | 0,540 |
| LDA | 0,847 | 0,865 | 0,587 | 0,485 | 0,531 |

Nota: Ordenado por F1-Score.

Fonte: Elaborado pelo autor (2025).

tes. Este modelo neural treinado serviu, portanto, tanto como um classificador autônomo para avaliação quanto como o extrator de características para os demais algoritmos.

Concluído o treinamento da rede neural, os pesos aprendidos das camadas de *embedding* foram extraídos e utilizados para transformar as variáveis categóricas dos conjuntos de dados em suas respectivas representações vetoriais densas. A dimensionalidade de cada vetor foi determinada com base na cardinalidade da variável, seguindo a heurística de $\min\left(50, \left\lceil \frac{|C|}{2} \right\rceil\right)$, onde $|C|$ é o número de categorias únicas (GUO; BERKHAHN, 2016). As dimensões resultantes para cada variável são detalhadas na Tabela 5. Estes vetores foram então concatenados com as 21 variáveis numéricas originais, formando um novo conjunto de dados enriquecido e inteiramente numérico. Foi sobre este conjunto de dados transformado que os demais classificadores, como XGBoost, SVM e Regressão Logística, foram treinados e avaliados.

Tabela 5 – Dimensionalidade dos *embeddings* por variável categórica.

| **Variável Categórica** | **Dimensão do *Embedding*** |
|---|---|
| Gênero | 2 |
| Etnia | 3 |
| Nível Educacional | 3 |
| Diagnóstico de Hipertensão | 2 |
| Atividade Física Moderada | 2 |
| Histórico Familiar de Diabetes | 2 |
| Autoavaliação da Saúde Geral | 3 |

Fonte: Elaborado pelo autor (2025).

A concatenação destes vetores de *embedding* com as 21 variáveis numéricas originais resultou em um espaço de características final com 38 preditores. Este resultado contrasta com a codificação *One-Hot*, que gerou um espaço de 45 preditores. A abordagem com *embeddings* produziu, portanto, uma representação de entrada mais compacta e densa para os classificadores.

A adoção de *embeddings* resultou em variações de desempenho distintas conforme a família do algoritmo utilizado. Para mensurar o impacto direto da troca da estratégia de codi-

Tabela 6 – Resultados dos modelos com variáveis numéricas e categóricas (*Embeddings*).

| **Modelo** | AUC | **Acurácia** | **Recall** | **Precisão** | **F1-Score** |
|---|---|---|---|---|---|
| XGBoost | 0,860 | 0,883 | 0,637 | 0,521 | 0,573 |
| LightGBM | 0,853 | 0,883 | 0,665 | 0,502 | 0,572 |
| Gradient Boosting | 0,853 | 0,882 | 0,650 | 0,502 | 0,567 |
| MLP | 0,830 | 0,880 | 0,729 | 0,453 | 0,559 |
| SVM | 0,833 | 0,879 | 0,711 | 0,457 | 0,557 |
| Random Forest | 0,853 | 0,870 | 0,600 | 0,503 | 0,547 |
| Regressão Logística | 0,839 | 0,873 | 0,657 | 0,468 | 0,547 |
| LDA | 0,842 | 0,865 | 0,625 | 0,473 | 0,538 |

Nota: Ordenado por F1-Score.

Fonte: Elaborado pelo autor (2025).

ficação, a Tabela 7 apresenta o comparativo do F1-Score entre as abordagens *One-Hot* e *Embeddings*. Observa-se que os modelos baseados em árvores de decisão (XGBoost, LightGBM, Gradient Boosting e Random Forest) e o LDA beneficiaram-se da representação densa vetorial. O Random Forest e o LDA registraram os maiores incrementos absolutos, com ganho de 0,70 pontos percentuais (p.p.) cada. Em contrapartida, o MLP e a Regressão Logística apresentaram uma redução marginal no desempenho. O modelo XGBoost manteve a liderança nos resultados, atingindo o F1-Score mais elevado entre os experimentos individuais (57,30%) e uma AUC de 0,860.

Tabela 7 – Comparativo do F1-Score entre codificação *One-Hot* e *Embeddings*.

| **Modelo** | **One-Hot (%)** | **Embeddings (%)** | **Diferença (p.p.)** |
|---|---|---|---|
| XGBoost | 57,10 | 57,30 | +0,20 |
| LightGBM | 56,60 | 57,20 | +0,60 |
| Gradient Boosting | 56,30 | 56,70 | +0,40 |
| MLP | 56,30 | 55,90 | -0,40 |
| SVM | 55,40 | 55,70 | +0,30 |
| Random Forest | 54,00 | 54,70 | +0,70 |
| Regressão Logística | 55,00 | 54,70 | -0,30 |
| LDA | 53,10 | 53,80 | +0,70 |

Fonte: Elaborado pelo autor (2025).

Foi realizado um teste Z de duas proporções comparando o desempenho dos dois melhores modelos sobre o conjunto de teste, que consistiu em 5.109 amostras. Não há evidências estatísticas para se afirmar que a diferença de desempenho entre as duas abordagens de codificação seja estatisticamente significante, podendo ser atribuída à variabilidade amostral.

## 4.3 BALANCEANDO AS CLASSES

Para mitigar o viés induzido pelo desbalanceamento de classes, foi aplicada a técnica de sobreamostragem SMOTE nos dados de treinamento. Esta intervenção alterou a distribuição

original dos dados, resultando em um conjunto de treino perfeitamente balanceado com um total de 34.836 amostras, divididas em 17.418 registros para a classe positiva (diabéticos) e 17.418 para a classe negativa (não diabéticos). É fundamental destacar que a aplicação do SMOTE ocorreu exclusivamente dentro de cada *fold* do processo de validação cruzada, prevenindo o vazamento de dados (*data leakage*) para o conjunto de validação.

Os resultados obtidos após o balanceamento com SMOTE são apresentados na Tabela 8. A análise destes dados permite avaliar se a geração de instâncias sintéticas da classe minoritária foi eficaz em melhorar a capacidade dos modelos de identificar corretamente os casos de diabetes.

Tabela 8 – Resultados dos modelos (balanceado com SMOTE).

| **Modelo** | AUC | **Acurácia** | **Recall** | **Precisão** | **F1-Score** |
|---|---|---|---|---|---|
| LightGBM | 0,884 | 0,853 | 0,681 | 0,502 | 0,578 |
| XGBoost | 0,882 | 0,846 | 0,679 | 0,484 | 0,565 |
| Gradient Boosting | 0,876 | 0,846 | 0,670 | 0,484 | 0,562 |
| Rede Neural | 0,880 | 0,830 | 0,729 | 0,453 | 0,559 |
| LDA | 0,872 | 0,841 | 0,665 | 0,472 | 0,552 |
| Regressão Logística | 0,872 | 0,842 | 0,654 | 0,474 | 0,550 |
| XGBoost | 0,865 | 0,831 | 0,657 | 0,450 | 0,534 |
| SVM | 0,863 | 0,834 | 0,636 | 0,457 | 0,532 |

Nota: Ordenado por F1-Score.

Fonte: Elaborado pelo autor (2025).

## 4.4 APRENDIZAGEM DE CONJUNTO

Com o objetivo de aprimorar a capacidade preditiva, foi investigada a aplicação de técnicas de aprendizagem de conjunto (*ensemble learning*). Esta abordagem busca combinar as predições dos modelos de melhor desempenho individual — XGBoost, MLP e SVM — para obter um classificador final mais robusto e generalizável. Foram avaliados dois métodos de combinação: *Hard Voting*, que agrega as predições por meio de uma votação majoritária, e *Soft Voting*, que calcula a média das probabilidades de predição de cada modelo antes de tomar a decisão final.

Os resultados dos classificadores de conjunto, detalhados na Tabela 9, demonstram que a combinação de modelos resultou em um desempenho superior ao dos modelos individuais na configuração com *embeddings* sem balanceamento. O *Soft Voting* alcançou um F1-Score de 0,583, superando o desempenho do melhor modelo individual, o XGBoost, que obteve um F1-Score de 0,573. Notavelmente, ambos os métodos de *voting* promoveram um aumento no *Recall* para a classe diabética, indicando uma melhor capacidade de identificação de casos positivos, embora com uma ligeira redução na Precisão. O *Soft Voting* demonstrou o melhor equilíbrio entre essas duas métricas, consolidando-se como a abordagem de melhor desempenho geral.

Tabela 9 – Resultados dos modelos de aprendizagem de conjunto.

| Modelo | Acurácia | Precisão | Recall | F1-Score |
|---|---|---|---|---|
| Soft Voting | 0,8562 | 0,5099 | 0,6808 | 0,5831 |
| Hard Voting | 0,8458 | 0,4850 | 0,7060 | 0,5750 |

Nota: As métricas de Precisão, Recall e F1-Score referem-se à classe positiva (diabetes).

Fonte: Elaborado pelo autor (2025).

## 4.5 ANÁLISE DE INTERPRETABILIDADE

A validação de modelos de AM para aplicações clínicas depende não apenas de sua acurácia preditiva, mas também de sua interpretabilidade. Para compreender os fatores que guiam as predições do modelo XGBoost com *embeddings*, que apresentou o melhor desempenho individual, foram conduzidas duas análises de interpretabilidade: a avaliação da importância global das características e a aplicação da metodologia SHAP.

### 4.5.1 Importância Global das Características

A análise da importância global das características, extraída diretamente do modelo XGBoost treinado, quantifica a contribuição de cada preditor para a redução de impureza ao longo de todas as árvores de decisão do conjunto. A Figura 7 apresenta as características mais influentes. A análise revela que as variáveis mais importantes foram o diagnóstico prévio de pressão alta e o uso de medicação para hipertensão, seguidas pela idade. Este resultado reforça a validade clínica do modelo, pois a hipertensão e a idade avançada são fatores de risco bem estabelecidos para o desenvolvimento de DM2.

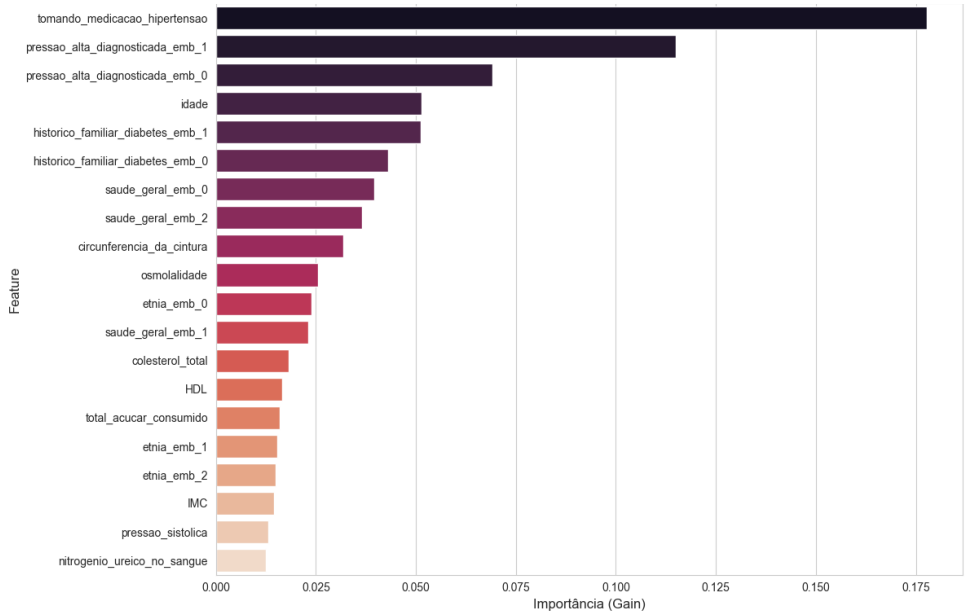
### 4.5.2 Análise com SHAP

Para uma análise mais aprofundada, foi utilizada a metodologia SHAP, que atribui um valor de impacto para cada característica em cada predição individual. A agregação desses valores permite uma visão global robusta do comportamento do modelo, como ilustrado no gráfico de resumo da Figura 8. A análise SHAP identificou a idade, a circunferência da cintura e a osmolalidade sanguínea como os três preditores de maior impacto. Este resultado complementa a análise anterior, destacando a importância de marcadores de obesidade central (circunferência da cintura) e de desequilíbrio metabólico (osmolalidade sanguínea).

## 4.6 COMPARAÇÃO COM TRABALHOS ANTERIORES

A comparação dos resultados obtidos neste estudo com os da literatura exige uma análise das metodologias empregadas, particularmente em trabalhos que reportam métricas de de-

Figura 7 – Importância das características para o modelo XGBoost, ordenada por contribuição.




Fonte: Autor (2025).

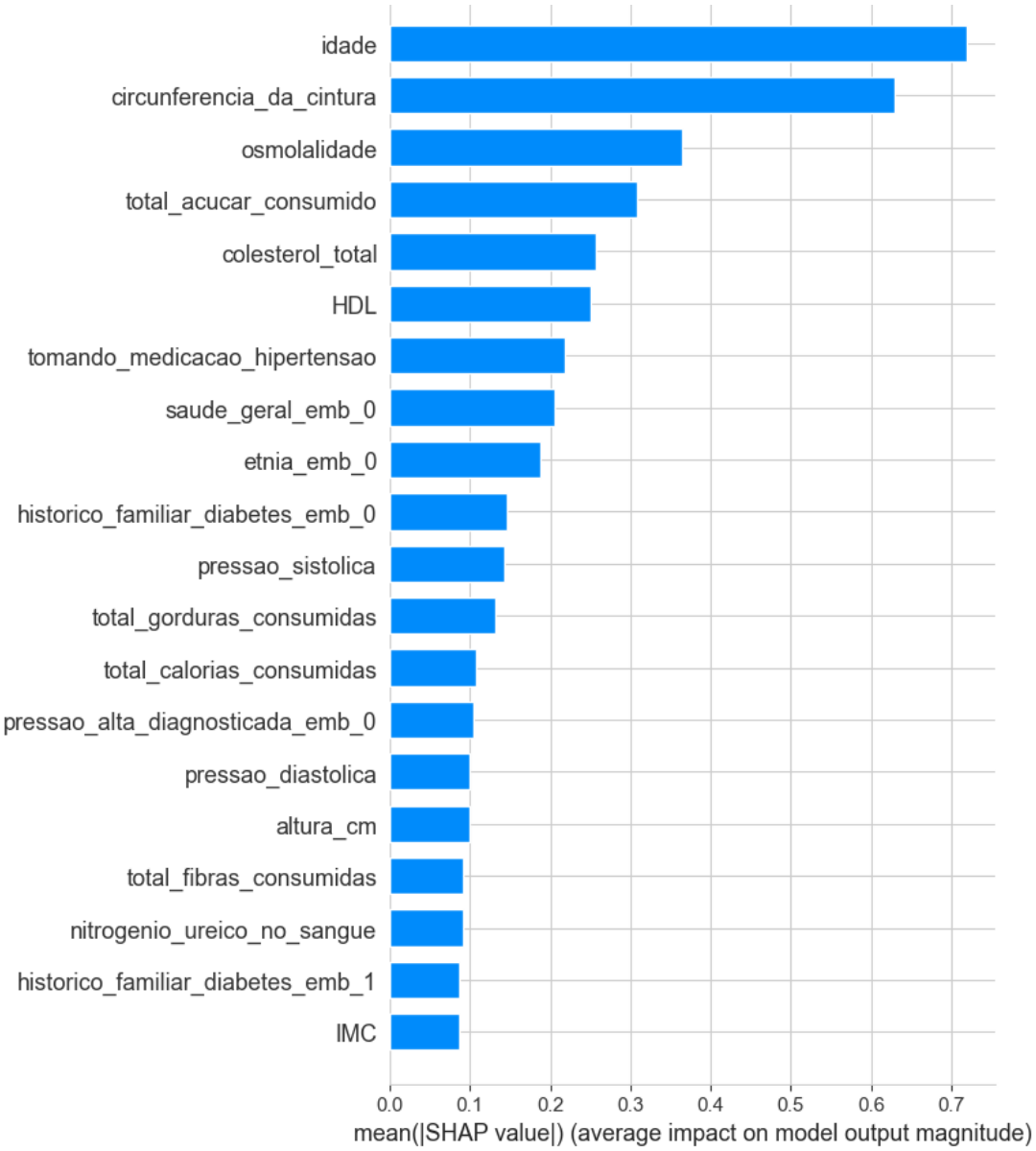
sempenho elevadas. Estudos como os de Dinh *et al.* (2019) e Qin *et al.* (2022), que também utilizaram o conjunto de dados NHANES, apresentam resultados superiores. Contudo, um exame de seus procedimentos experimentais revela falhas metodológicas que comprometem a validade de suas conclusões e a comparabilidade de seus resultados.

Um exemplo é o trabalho de Qin *et al.* (2022), que reportou uma AUC de 0,83. A principal falha metodológica identificada nesse estudo foi a aplicação da técnica de sobreamostragem SMOTE em todo o conjunto de dados antes da sua divisão em subconjuntos de treinamento e teste. Este procedimento constitui um caso de vazamento de dados (*data leakage*), pois permite que informações do conjunto de teste influenciem a criação de amostras sintéticas no conjunto de treinamento. Consequentemente, o modelo é avaliado com dados que contêm informações derivadas de sua própria base de treinamento, resultando em uma estimativa de desempenho excessivamente otimista.

De forma análoga, o estudo de Dinh *et al.* (2019), que alcançou uma AUC de 95,7% em um cenário com dados laboratoriais, também apresenta uma metodologia suscetível a vieses de avaliação. Para investigar a robustez desses achados, a abordagem de reamostragem proposta por Dinh *et al.* (2019) foi replicada, com a aplicação da técnica de subamostragem exclusivamente sobre os dados de treinamento dentro de cada iteração da validação cruzada, evitando a contaminação do conjunto de teste. Os resultados desta replicação, apresentados na Tabela 10, demonstram um desempenho substancialmente inferior ao originalmente reportado.

A análise da Tabela 10 evidencia que, ao corrigir a falha metodológica, o F1-Score do melhor modelo (WEM - Ensemble Ponderado) atinge 0,5055. Este valor é consideravelmente mais baixo que os resultados obtidos no presente trabalho, onde o modelo XGBoost com *embed-*

Figura 8 – Gráfico de resumo SHAP para as características mais influentes no modelo XGBoost.




Fonte: Autor (2025).

Tabela 10 – Resultados da replicação da metodologia de Dinh *et al.* (2019) com aplicação correta da técnica de subamostragem.

| **Modelo** | **Acurácia** | **AUC** | **Precisão** | **Recall** | **F1-Score** |
|---|---|---|---|---|---|
| WEM - Ensemble Ponderado | 0,7376 | 0,8316 | 0,3740 | 0,7795 | 0,5055 |
| XGBoost | 0,7356 | 0,8265 | 0,3717 | 0,7773 | 0,5029 |
| SVM | 0,7298 | 0,8197 | 0,3652 | 0,7727 | 0,4960 |
| Regressão Logística | 0,7215 | 0,8169 | 0,3577 | 0,7773 | 0,4900 |
| Random Forest | 0,7192 | 0,8211 | 0,3558 | 0,7795 | 0,4886 |

Fonte: Elaborado pelo autor (2025).

*dings* alcançou um F1-Score de 0,573 (Tabela 6). A discrepância confirma que o desempenho originalmente reportado por Dinh *et al.* (2019) foi inflado pela aplicação inadequada da técnica de reamostragem.

# 5 CONSIDERAÇÕES FINAIS

A detecção de DM por meio de aprendizado de máquina enfrenta o desafio de representar adequadamente as variáveis categóricas, que são abundantes em bases de dados de saúde. Abordagens de codificação convencionais, como *one-hot encoding*, frequentemente resultam em representações de alta dimensionalidade e esparsas, enquanto a codificação por rótulos pode introduzir relações ordinais artificiais que limitam o desempenho dos modelos.

Nesse contexto, este trabalho investigou a técnica de *embeddings* como uma alternativa para o tratamento dessas variáveis. A abordagem consistiu no aprendizado de representações vetoriais densas e de baixa dimensionalidade, com o objetivo de capturar as relações semânticas latentes entre as categorias no contexto da tarefa de predição. Essa técnica permite que o modelo generalize o conhecimento entre categorias funcionalmente similares, gerando um espaço de características mais compacto e potencialmente mais informativo.

A metodologia executada centrou-se na implementação de uma arquitetura de RNA que integra camadas de *embedding* para tratar as variáveis categóricas do conjunto de dados NHANES. O desempenho desta abordagem foi comparado ao de modelos treinados com codificação *one-hot* e ao de algoritmos de referência, como XGBoost e Regressão Logística. O protocolo experimental assegurou rigor metodológico, com divisão estratificada dos dados e aplicação de técnicas de reamostragem, como o SMOTE, exclusivamente sobre os dados de treinamento para evitar o vazamento de informações.

Os resultados demonstraram que a inclusão de variáveis categóricas, independentemente do método de codificação, aprimorou o desempenho de todos os modelos em comparação ao cenário que utilizava apenas dados numéricos, confirmando sua relevância preditiva. A abordagem com *embeddings* produziu um espaço de características mais compacto (38 preditores) em relação ao *one-hot encoding* (45 preditores). O modelo XGBoost, quando alimentado com as características geradas pelos *embeddings*, alcançou o melhor desempenho entre os classificadores individuais, com um F1-Score de 0,573. Contudo, essa performance foi apenas marginalmente superior à obtida com *one-hot encoding* (F1-Score de 0,571), e um teste estatístico Z indicou que a diferença não era estatisticamente significativa. O melhor resultado geral foi alcançado por um modelo de fusão de múltiplos classificadores (*ensemble*) com *Soft Voting*, que obteve um F1-Score de 0,583.

A contribuição deste estudo reside também na contextualização de seus resultados frente à literatura. A análise de trabalhos anteriores, como os de Dinh *et al.* (2019) e Qin *et al.* (2022), revelou falhas metodológicas, como o vazamento de dados decorrente da aplicação inadequada de técnicas de reamostragem, que levaram a estimativas de desempenho excessivamente otimistas. Ao replicar uma dessas metodologias de forma correta, demonstrou-se que os resulta-

dos aqui obtidos, embora numericamente inferiores, representam um referencial mais robusto e realista para a tarefa de detecção de DM com o conjunto de dados NHANES.

Conclui-se que os *embeddings* constituem uma técnica viável para a representação de variáveis categóricas em dados de saúde, oferecendo uma alternativa mais compacta ao *one-hot encoding* e alcançando um desempenho preditivo competitivo. Embora a superioridade sobre os métodos tradicionais não tenha sido estatisticamente conclusiva para os modelos baseados em árvores, a combinação de representações densas com algoritmos de conjunto mostrou-se a estratégia mais promissora, estabelecendo um novo patamar de desempenho para a detecção de diabetes neste importante conjunto de dados de saúde pública.

## 5.1 TRABALHOS FUTUROS

Para dar continuidade a esta pesquisa, sugere-se a exploração de arquiteturas de aprendizado profundo desenvolvidas especificamente para dados tabulares, especialmente modelos baseados na arquitetura *Transformer*, que podem ser mais eficazes na captura de interações complexas entre as variáveis. Adicionalmente, a aplicação da metodologia em conjuntos de dados longitudinais permitiria a transição de um modelo de detecção para um de predição do risco de desenvolvimento futuro de diabetes. Outra linha de investigação consiste em uma análise mais aprofundada do espaço vetorial dos *embeddings* aprendidos, utilizando técnicas de visualização como t-SNE ou PCA, para interpretar as relações semânticas que o modelo identificou entre as diferentes categorias das variáveis preditoras.

# REFERÊNCIAS


ADA. 2. diagnosis and classification of diabetes: Standards of care in diabetes—2024. **Diabetes Care**, v. 47, n. Supplement_1, p. S20–S42, 12 2023.

ADLER, A. **Using Machine Learning Techniques to Identify Key Risk Factors for Diabetes and Undiagnosed Diabetes**. 2021. Disponível em: <https://arxiv.org/abs/2105.09379>.

AGRESTI, A. **Categorical Data Analysis**. Wiley, 2002. ISSN 1940-6347. ISBN 9780471249689. Disponível em: <http://dx.doi.org/10.1002/0471249688>.

ALEJO, R. *et al.* Edited nearest neighbor rule for improving neural networks classifications. In: ____. **Advances in Neural Networks - ISNN 2010**. Springer Berlin Heidelberg, 2010. p. 303–310. ISBN 9783642132780. Disponível em: <http://dx.doi.org/10.1007/978-3-642-13278-0_39>.

ASHISHA, G. R. *et al.* Random oversampling-based diabetes classification via machine learning algorithms. **International Journal of Computational Intelligence Systems**, Springer Science and Business Media LLC, v. 17, n. 1, nov. 2024. ISSN 1875-6883. Disponível em: <http://dx.doi.org/10.1007/s44196-024-00678-3>.

BAHIA, L. R. *et al.* The costs of type 2 diabetes mellitus outpatient care in the brazilian public health system. **Value in Health**, Elsevier, v. 14, n. 5, p. S137–S140, Jul 2011. ISSN 1098-3015. Disponível em: <https://doi.org/10.1016/j.jval.2011.05.009>.

BATISTA, G. E. A. P. A.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. **SIGKDD Explor. Newsl.**, Association for Computing Machinery, New York, NY, USA, v. 6, n. 1, p. 20–29, jun. 2004. ISSN 1931-0145. Disponível em: <https://doi.org/10.1145/1007730.1007735>.

BISHOP, C. M. **Pattern Recognition and Machine Learning (Information Science and Statistics)**. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN 0387310738.

BOEHMKE, B.; GREENWELL, B. **Hands-On Machine Learning with R**. Chapman and Hall/CRC, 2019. ISBN 9780367816377. Disponível em: <http://dx.doi.org/10.1201/9780367816377>.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde e Ambiente. Departamento de Análise Epidemiológica e Vigilância de Doenças Não Transmissíveis. **Vigitel Brasil 2023**: vigilância de fatores de risco e proteção para doenças crônicas por inquérito telefônico: estimativas sobre frequência e distribuição sociodemográfica de fatores de risco e proteção para doenças crônicas nas capitais dos 26 estados brasileiros e no distrito federal em 2023. Brasília, DF: Ministério da Saúde, 2023. ISBN 978-65-5993-476-8. Disponível em: <http://bvsms.saude.gov.br/bvs/publicacoes/vigitel_brasil_2023.pdf>.

BREIMAN, L. Random forests. **Machine Learning**, Springer Science and Business Media LLC, v. 45, n. 1, p. 5–32, 2001. ISSN 0885-6125. Disponível em: <http://dx.doi.org/10.1023/A:1010933404324>.

Centers for Disease Control and Prevention (CDC); National Center for Health Statistics (NCHS). **National Health and Nutrition Examination Survey, 2017-2018**. 2017. <https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/DIQ_J.htm>. Accessed: 2025-04-24.

CHAWLA, N. V. *et al.* Smote: Synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, AI Access Foundation, v. 16, p. 321–357, jun. 2002. ISSN 1076-9757. Disponível em: <http://dx.doi.org/10.1613/jair.953>.

CHEN, C.; BREIMAN, L. Using random forest to learn imbalanced data. **University of California, Berkeley**, 01 2004. Disponível em: <https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf>.

CHOWDHURY, M. M.; AYON, R. S.; HOSSAIN, M. S. Diabetes diagnosis through machine learning: Investigating algorithms and data augmentation for class imbalanced brfss dataset. **medRxiv**, Cold Spring Harbor Laboratory Press, 2023. Disponível em: <https://www.medrxiv.org/content/early/2023/10/19/2023.10.18.23292250>.

CICHOSZ, S. L.; BENDER, C.; HEJLESEN, O. A comparative analysis of machine learning models for the detection of undiagnosed diabetes patients. **Diabetology**, MDPI AG, v. 5, n. 1, p. 1–11, jan. 2024. ISSN 2673-4540. Disponível em: <http://dx.doi.org/10.3390/diabetology5010001>.

CICHOSZ, S. L.; OLESEN, S. S.; JENSEN, M. H. Explainable machine-learning models to predict weekly risk of hyperglycemia, hypoglycemia, and glycemic variability in patients with type 1 diabetes based on continuous glucose monitoring. **Journal of Diabetes Science and Technology**, SAGE Publications Sage CA: Los Angeles, CA, 2024. Disponível em: <https://doi.org/10.1177/19322968241286907>.

CLORE JOHN, C. **Diabetes 130-US Hospitals for Years 1999-2008**. 2014. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5230J.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, Springer Science and Business Media LLC, v. 20, n. 3, p. 273–297, set. 1995. ISSN 1573-0565. Disponível em: <http://dx.doi.org/10.1007/BF00994018>.

DAHOUDA, M. K.; JOE, I. A deep-learned embedding technique for categorical features encoding. **IEEE Access**, v. 9, p. 114381–114391, 2021.

____. A deep-learned embedding technique for categorical features encoding. **IEEE Access**, v. 9, p. 114381–114391, 2021.

DIABETES, N. I. of; DIGESTIVE; DISEASES, K. **Pima Indians Diabetes Database**. 2025. <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>. Acessado em: 24 de abril de 2025.

DINH, A. *et al.* A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. **BMC Medical Informatics and Decision Making**, Springer Science and Business Media LLC, v. 19, n. 1, nov. 2019. ISSN 1472-6947. Disponível em: <http://dx.doi.org/10.1186/s12911-019-0918-5>.

ELIASCHEWITZ, F. *et al.* Type 2 diabetes in Brazil: epidemiology and management. **Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy**, Informa UK Limited, p. 17–28, January 2015. ISSN 1178-7007. Disponível em: <http://dx.doi.org/10.2147/DMSO.S72542>.

ELKAN, C. The foundations of cost-sensitive learning. In: . San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. (IJCAI'01), p. 973–978. ISBN 1558608125.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 29, n. 5, out. 2001. ISSN 0090-5364. Disponível em: <http://dx.doi.org/10.1214/aos/1013203451>.

GERON, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**. 2nd. ed. [S.l.]: O'Reilly Media, Inc., 2019. ISBN 1492032646.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. Cambridge, MA: MIT Press, 2016. Disponível em: <https://www.deeplearningbook.org/>.

GUO, C.; BERKHAHN, F. **Entity Embeddings of Categorical Variables**. arXiv, 2016. Disponível em: <https://arxiv.org/abs/1604.06737>.

HANCOCK, J. T.; KHOSHGOFTAAR, T. M. Survey on categorical data for neural networks. **Journal of big data**, Springer, v. 7, n. 1, p. 28, 2020.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. Springer, 2009. (Springer series in statistics). ISBN 9780387848846. Disponível em: <https://books.google.com.br/books?id=eBSgoAEACAAJ>.

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied Logistic Regression**. 3. ed. Hoboken, NJ: John Wiley & Sons, 2013. 528 p. ISBN 978-0470582473.

JAMES, G. *et al.* **An Introduction to Statistical Learning: with Applications in R**. [S.l.]: Springer Publishing Company, Incorporated, 2014. ISBN 1461471370.

JOHNSON, C. L. *et al.* National health and nutrition examination survey: analytic guidelines, 1999-2010. **Vital and health statistics. Series 2, Data evaluation and methods research**, National Center for Health Statistics, n. 161, p. 1–24, set. 2013. ISSN 0083-2057. All material appearing in this report is in the public domain and may be reproduced or copied without permission; citation as to source is appreciated.

KAVAKIOTIS, I. *et al.* Machine learning and data mining methods in diabetes research. **Computational and Structural Biotechnology Journal**, v. 15, p. 104–116, 2017. ISSN 2001-0370. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2001037016300733>.

KHATIB, O. M. N. **Guidelines for the Prevention, Management and Care of Diabetes Mellitus**. [S.l.]: World Health Organization, 2006. (32). ISBN 978-92-9021-404-5.

KUHN, M.; JOHNSON, K. **Feature Engineering and Selection: A Practical Approach for Predictive Models**. Chapman and Hall/CRC, 2019. ISBN 9781315108230. Disponível em: <http://dx.doi.org/10.1201/9781315108230>.

KUTNER, M. **Applied Linear Statistical Models**. McGraw-Hill Irwin, 2005. (McGrwa-Hill international edition). ISBN 9780071122214. Disponível em: <https://books.google.com.br/books?id=0xqCAAAACAAJ>.

LE, P. B.; NGUYEN, Z. T. Roc curves, loss functions, and distorted probabilities in binary classification. **Mathematics**, v. 10, n. 9, 2022. ISSN 2227-7390. Disponível em: <https://www.mdpi.com/2227-7390/10/9/1410>.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, Springer Science and Business Media LLC, v. 521, n. 7553, p. 436–444, maio 2015. ISSN 1476-4687. Disponível em: <http://dx.doi.org/10.1038/nature14539>.

LI, W.; PENG, Y.; PENG, K. Diabetes prediction model based on ga-xgboost and stacking ensemble algorithm. **PLOS ONE**, Public Library of Science (PLoS), v. 19, n. 9, p. e0311222, set. 2024. ISSN 1932-6203. Disponível em: <http://dx.doi.org/10.1371/journal.pone.0311222>.

LING, C. X.; SHENG, V. S. Cost-sensitive learning and the class imbalance problem. In: SAMMUT, C.; WEBB, G. I. (Ed.). **Encyclopedia of Machine Learning**. Springer, 2011. p. 231–235. ISBN 978-0-387-30768-8. Disponível em: <https://doi.org/10.1007/978-0-387-30164-8_142>.

MAATEN, L. van der; HINTON, G. Visualizing data using t-sne. **Journal of Machine Learning Research**, v. 9, n. 86, p. 2579–2605, 2008. Disponível em: <http://jmlr.org/papers/v9/vandermaaten08a.html>.

MALTA, D. C. *et al.* Mortalidade por doenças crônicas não transmissíveis no brasil e suas regiões, 2000 a 2011. **Epidemiologia e Serviços de Saúde**, v. 23, p. 599–608, 12 2014.

MEHRABI, N. *et al.* A survey on bias and fairness in machine learning. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 54, n. 6, jul. 2021. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3457607>.

National Center for Health Statistics. **NHANES Survey Methods and Analytic Guidelines - Data, Documentation, Codebooks, SAS Code**. [S.l.], 2024. Disponível em: <https://wwwn.cdc.gov/nchs/nhanes/analyticguidelines.aspx>.

OLIVEIRA, F. *et al.* Machine learning-based diabetes detection using photoplethysmography signal features. In: **Anais do XXIV Simpósio Brasileiro de Computação Aplicada à Saúde**. Porto Alegre, RS, Brasil: SBC, 2024. p. 94–105. ISSN 2763-8952. Disponível em: <https://sol.sbc.org.br/index.php/sbcas/article/view/28809>.

PENNINGTON, J.; SOCHER, R.; MANNING, C. GloVe: Global vectors for word representation. In: MOSCHITTI, A.; PANG, B.; DAELEMANS, W. (Ed.). **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1532–1543. Disponível em: <https://aclanthology.org/D14-1162/>.

PIAS, T. S. *et al.* Enhancing fairness and accuracy in diagnosing type 2 diabetes in young population. Cold Spring Harbor Laboratory, maio 2023. Disponível em: <http://dx.doi.org/10.1101/2023.05.02.23289405>.

QIN, Y. *et al.* Machine learning models for data-driven prediction of diabetes by lifestyle type. **International Journal of Environmental Research and Public Health**, MDPI AG, v. 19, n. 22, p. 15027, nov. 2022. ISSN 1660-4601. Disponível em: <http://dx.doi.org/10.3390/ijerph192215027>.

RABIE, O. *et al.* A decision support system for diagnosing diabetes using deep neural network. **Frontiers in Public Health**, Volume 10 - 2022, 2022. ISSN 2296-2565. Disponível em: <https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2022.861062>.

REN, X. Predictions of diabetes through machine learning models based on the health indicators dataset. **Applied and Computational Engineering**, EWA Publishing, v. 32, n. 1, p. 216–222, jan. 2024. ISSN 2755-273X. Disponível em: <http://dx.doi.org/10.54254/2755-2721/32/20230214>.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **Nature**, Springer Science and Business Media LLC, v. 323, n. 6088, p. 533–536, out. 1986. ISSN 1476-4687. Disponível em: <http://dx.doi.org/10.1038/323533a0>.

RUSSAC, Y.; CAELEN, O.; HE-GUELTON, L. Embeddings of categorical variables for sequential data in fraud context. In: ____. **The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018)**. Springer International Publishing, 2018. p. 542–552. ISBN 9783319746906. Disponível em: <http://dx.doi.org/10.1007/978-3-319-74690-6_53>.

SCHMIDT, M. I. *et al.* High prevalence of diabetes and intermediate hyperglycemia – the brazilian longitudinal study of adult health (elsa-brasil). **Diabetology And Metabolic Syndrome**, Springer Science and Business Media LLC, v. 6, n. 1, nov. 2014. ISSN 1758-5996. Disponível em: <http://dx.doi.org/10.1186/1758-5996-6-123>.

SEIFFERT, C. *et al.* Rusboost: A hybrid approach to alleviating class imbalance. **IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans**, v. 40, n. 1, p. 185–197, 2010.

SHEPHERD, B. *et al.* **Survey Methodology for the 2017–March 2020 National Health and Nutrition Examination Survey Prepandemic Data**. [S.l.], 2021. Disponível em: <https://www.cdc.gov/nchs/data/series/sr_02/sr02-200.pdf>.

Sociedade Brasileira de Diabetes. **Diretrizes da Sociedade Brasileira de Diabetes 2019-2020**. São Paulo: Sociedade Brasileira de Diabetes, 2019. Distribuição exclusiva da Sociedade Brasileira de Diabetes. Direitos exclusivos para a língua portuguesa. Disponível em: <https://portaldeboaspraticas.iff.fiocruz.br/biblioteca/diretrizes-da-sociedade-brasileira-de-diabetes-2019-2020/>.

TANIM, S. A. *et al.* Explainable deep learning for diabetes diagnosis with deepnetx2. **Biomedical Signal Processing and Control**, v. 99, p. 106902, 2025. ISSN 1746-8094. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1746809424009601>.

VANGEEPURAM, N. *et al.* Predicting youth diabetes risk using nhanes data and machine learning. **Scientific Reports**, Springer Science and Business Media LLC, v. 11, n. 1, maio 2021. ISSN 2045-2322. Disponível em: <http://dx.doi.org/10.1038/s41598-021-90406-0>.

WEBB, A.; COPSEY, K. **Statistical Pattern Recognition**. Wiley, 2011. ISBN 9781119961406. Disponível em: <https://books.google.com.br/books?id=gb79v56894oC>.

XIE, Z. *et al.* Building risk prediction models for type 2 diabetes using machine learning techniques. **Preventing Chronic Disease**, Centers for Disease Control and Prevention (CDC), v. 16, set. 2019. ISSN 1545-1151. Disponível em: <http://dx.doi.org/10.5888/pcd16.190109>.

YU, Z. *et al.* Dmnet: A personalized risk assessment framework for elderly people with type 2 diabetes. **IEEE Journal of Biomedical and Health Informatics**, Institute of Electrical and Electronics Engineers (IEEE), v. 27, n. 3, p. 1558–1568, mar. 2023. ISSN 2168-2208. Disponível em: <http://dx.doi.org/10.1109/JBHI.2022.3233622>.

ZHOU, Z. *et al.* Glumarker: A novel predictive modeling of glycemic control through digital biomarkers. In: **2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)**. [S.l.: s.n.], 2024. p. 1–7.