

**UNIVERSIDADE FEDERAL DE UBERLÂNDIA INSTITUTO DE  
BIOTECNOLOGIA  
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA**

**INTELIGÊNCIA ARTIFICIAL PARA PREDIÇÃO DE DIABETES MELLITUS TIPO  
2 COM BASE EM DADOS SOCIODEMOGRÁFICOS E ESTILO DE VIDA**

**CRISTINA RIBEIRO DA MOTA E SILVA CÔRTEZ**

**PATOS DE MINAS  
2025**

**CRISTINA RIBEIRO DA MOTA E SILVA CÔRTEZ**

**INTELIGÊNCIA ARTIFICIAL PARA PREDIÇÃO DE DIABETES MELLITUS TIPO  
2 COM BASE EM DADOS SOCIODEMOGRÁFICOS E ESTILO DE VIDA**

Dissertação apresentada ao Programa de Pós-Graduação em Biotecnologia do Instituto de Biotecnologia da Universidade Federal de Uberlândia, campus Patos de Minas como requisito parcial para a obtenção do título de Mestre em Biotecnologia.

**Área de concentração: Biociências.**

**Orientador: Prof. Dr. Laurence Rodrigues do  
Amaral**

**PATOS DE MINAS**

**2025**

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU  
com dados informados pelo(a) próprio(a) autor(a).

C882 Côrtes, Cristina Ribeiro da Mota e Silva, 1979-  
2026 Inteligência Artificial Para Predição de Diabetes Mellitus Tipo II  
Com Base em Dados Sociodemográficos E Estilo de vida [recurso  
eletrônico] / Cristina Ribeiro da Mota e Silva Côrtes. - 2026.

Orientador: Laurence Rodrigues do Amaral.  
Dissertação (Mestrado) - Universidade Federal de Uberlândia,  
Pós-graduação em Biotecnologia.

Modo de acesso: Internet.

DOI <http://doi.org/10.14393/ufu.di.2026.25>

Inclui bibliografia.

1. Biotecnologia. I. Amaral, Laurence Rodrigues do, 1978-,  
(Orient.). II. Universidade Federal de Uberlândia. Pós-graduação  
em Biotecnologia. III. Título.

CDU: 60

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:

Gizele Cristine Nunes do Couto - CRB6/2091

Nelson Marcos Ferreira - CRB6/3074



## **ATA DE DEFESA - PÓS-GRADUAÇÃO**

|                               |                                                                                                                                |                 |       |                       |       |
|-------------------------------|--------------------------------------------------------------------------------------------------------------------------------|-----------------|-------|-----------------------|-------|
| Programa de Pós-Graduação em: | Biotecnologia                                                                                                                  |                 |       |                       |       |
| Defesa de:                    | Dissertação de Mestrado Acadêmico                                                                                              |                 |       |                       |       |
| Data:                         | 15/12/2025                                                                                                                     | Hora de início: | 14:00 | Hora de encerramento: | 16:10 |
| Matrícula do Discente:        | 42312BTC008                                                                                                                    |                 |       |                       |       |
| Nome do Discente:             | Cristina Ribeiro da Mota e Silva Côrtes                                                                                        |                 |       |                       |       |
| Título do Trabalho:           | <b>Inteligência Artificial para Predição de Diabetes Mellitus Tipo II com Base em Dados Sociodemográficos e Estilo de Vida</b> |                 |       |                       |       |
| Área de concentração:         | Biociências                                                                                                                    |                 |       |                       |       |
| Linha de pesquisa:            | Bioinformática e Biologia Molecular aplicada à genômica, transcriptômica e proteômica                                          |                 |       |                       |       |

Reuniu-se, na sala 325 bloco Alfa, a Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Biotecnologia, composta pelos seguintes Professores Doutores: Matheus de Souza Gomes (IBTEC - UFU), Kelen Cristina Estavanate de Castro (UNIPAM) e Laurence Rodrigues do Amaral (FACOM - UFU), presidente da banca.

Iniciando os trabalhos, o presidente da banca, Dr. Laurence Rodrigues do Amaral, apresentou a Comissão Examinadora e a candidata, agradeceu a presença dos membros. Em seguida, concedeu a palavra ao discente para a exposição do seu trabalho. A duração da apresentação seguiu as normas do Programa de Pós-Graduação em Biotecnologia.

A seguir o senhor(a) presidente concedeu a palavra, pela ordem sucessivamente, aos(às) examinadores(as), que passaram a arguir o(a) candidato(a). Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o(a) candidato(a):

*Aprovado.*

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente

ata que após lida e achada conforme foi assinada pela Banca Examinadora.



Documento assinado eletronicamente por **Laurence Rodrigues do Amaral, Professor(a) do Magistério Superior**, em 15/12/2025, às 16:11, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Matheus de Souza Gomes, Professor(a) do Magistério Superior**, em 15/12/2025, às 16:39, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Kelen Cristina Estavanate de Castro, Usuário Externo**, em 21/12/2025, às 10:13, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site [https://www.sei.ufu.br/sei/controlador\\_externo.php?acao=documento\\_conferir&id\\_orgao\\_acesso\\_externo=0](https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0), informando o código verificador **6936770** e o código CRC **02F83A39**.

**Referência:** Processo nº 23117.088423/2025-06

SEI nº 6936770

**Resumo:** As doenças crônicas não transmissíveis, especialmente o Diabetes Mellitus tipo 2 (DM2), representam um dos maiores desafios da saúde pública mundial, devido à sua alta prevalência e impacto econômico e social. Diante disso, este estudo aplicou técnicas de Inteligência Artificial (IA), com foco em Aprendizado de Máquina (AM), para prever riscos de DM2 a partir de dados sociodemográficos, clínicos e de estilo de vida. Foram utilizados dois conjuntos de dados públicos do *Behavioral Risk Factor Surveillance System* (BRFSS), contendo 250.360 registros e 21 variáveis. O algoritmo J48 (C4.5) foi aplicado por meio do software Weka 3.8.6, utilizando validação cruzada (*10-fold cross-validation*). O modelo alcançou acurácia média de 83,85% na classificação binária e 88,84% na multiclasse. A seleção de atributos identificou seis variáveis mais relevantes: hipertensão arterial, colesterol elevado, doença cardíaca, consumo excessivo de álcool, autoavaliação da saúde e dificuldade para caminhar. Os resultados demonstram o potencial das técnicas de IA na identificação precoce e prevenção do DM2, reforçando a importância da integração entre dados clínicos, nutricionais e comportamentais na construção de modelos preditivos. Conclui-se que a aplicação de IA na nutrição de precisão pode otimizar o monitoramento e o diagnóstico preventivo, reduzindo custos e promovendo qualidade de vida.

**Palavras-chave:** Aprendizado de Máquina; Nutrição de Precisão; Doenças crônicas não transmissíveis.

**Abstract:** Non-communicable chronic diseases, especially type 2 Diabetes Mellitus (T2DM), represent one of the greatest global public health challenges due to their high prevalence and their economic and social impact. In this context, this study applied Artificial Intelligence (AI) techniques, with an emphasis on Machine Learning (ML), to predict T2DM risk based on sociodemographic, clinical, and lifestyle data. Two public datasets from the Behavioral Risk Factor Surveillance System (BRFSS) were used, comprising 250,360 records and 21 variables. The J48 (C4.5) algorithm was implemented using Weka 3.8.6 software with 10-fold cross-validation. The model achieved an average accuracy of 83.85% for binary classification and 88.84% for multiclass classification. Feature selection identified six most relevant variables: hypertension, high cholesterol, heart disease, excessive alcohol consumption, self-rated health, and difficulty walking. The results demonstrate the potential of AI techniques for the early identification and prevention of T2DM, reinforcing the importance of integrating clinical, nutritional, and behavioral data in the development of predictive models. It is concluded that the application of AI in precision nutrition can optimize monitoring and preventive diagnosis, reducing costs and promoting quality of life.

**Keywords:** Machine Learning; Precision Nutrition; Chronic Non-communicable Diseases.

## Lista de Figuras

|                                                                                                                                                                                                                                                                            |    |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Figura 1: Distribuição das Classes no Dataset Multiclasse do BRFSS (C0, C1 e C2).....                                                                                                                                                                                      | 24 |
| Figura 2 - Distribuição das Classes no Conjunto de Dados Binário (BRFSS).....                                                                                                                                                                                              | 25 |
| Figura 3 - Distribuição das classes C0 e C1 no dataset binário. ....                                                                                                                                                                                                       | 28 |
| Figura 4 - Heatmap da matriz de confusão do modelo binário (C0 e C1).....                                                                                                                                                                                                  | 31 |
| Figura 5 - Ranking dos atributos selecionados pelo método GainRatioAttributeEval + Ranker no dataset binário, evidenciando a importância relativa das variáveis para a predição entre C0 (ausência de diabetes/pré-diabetes) e C1 (presença de diabetes/pré-diabetes)..... | 33 |
| Figura 6: Árvore de decisão gerada pelo algoritmo J48 utilizando os seis atributos selecionados (HighBP, GenHlth, HighChol, HvyAlcoholConsump, HeartDiseaseorAttack e DiffWalk) no dataset binário .....                                                                   | 36 |

## Sumário

|                                                                                                          |    |
|----------------------------------------------------------------------------------------------------------|----|
| 1. INTRODUÇÃO .....                                                                                      | 7  |
| 2. FUNDAMENTAÇÃO TEÓRICA .....                                                                           | 8  |
| 2.1 Panorama das Doenças Crônicas Não Transmissíveis (DCNT).....                                         | 8  |
| 2.2 Diabetes Mellitus: Conceitos e Classificação .....                                                   | 9  |
| 2.3 Critérios diagnósticos.....                                                                          | 11 |
| 2.4 Fatores de risco clínicos, sociodemográficos e comportamentais para o Diabetes Mellitus tipo 2 ..... | 12 |
| 2.5 Bases de dados e Big Data em saúde (BRFSS).....                                                      | 14 |
| 3. OBJETIVO GERAL .....                                                                                  | 21 |
| 3.1 Objetivos Específicos.....                                                                           | 22 |
| 4. MATERIAL E MÉTODOS .....                                                                              | 22 |
| 4.1 Tipo de estudo .....                                                                                 | 22 |
| 4.2 Fonte dos dados.....                                                                                 | 22 |
| 4.3 Definição das classes.....                                                                           | 23 |
| 4.3.1 Dataset multiclasse.....                                                                           | 23 |
| 4.3.2 Dataset binário.....                                                                               | 23 |
| 4.4 Pré-processamento dos dados.....                                                                     | 23 |
| 4.5 Seleção de Atributos.....                                                                            | 25 |
| 4.6 Classificador utilizado.....                                                                         | 26 |
| 4.7 Validação do modelo.....                                                                             | 26 |
| 4.8 Construção dos modelos.....                                                                          | 27 |
| 4.9 Software e ambiente computacional.....                                                               | 27 |
| 5. RESULTADOS E DISCUSSÃO .....                                                                          | 27 |
| 5.1 Descrição inicial dos dados e impacto do desbalanceamento .....                                      | 27 |
| 5.2 Modelo Multiclasse .....                                                                             | 29 |
| 5.3 Modelo Binário e Desempenho da Classificação.....                                                    | 30 |
| 5.4 Seleção de Atributos e Interpretação Clínica .....                                                   | 32 |
| 5.5 Árvore de Decisão com Atributos Seleccionados .....                                                  | 35 |
| REFERÊNCIAS .....                                                                                        | 39 |

## 1. INTRODUÇÃO

O acúmulo de evidências que ligam a alimentação a doenças, particularmente as Doenças Crônicas Não Transmissíveis (DCNT), incluindo o Diabetes Mellitus tipo 2 (DM2), tem obrigado os especialistas a focar a nutrição preventiva como forma de controlar e prevenir o avanço dessas doenças (Blumberg, 1997).

As DCNT representam um grande desafio para a saúde pública global. Conforme dados da Organização Mundial da Saúde (OMS), em 2016, essas doenças foram responsáveis por aproximadamente 41 milhões das 57 milhões de mortes no mundo, sendo 15 milhões dessas mortes prematuras, ou seja, de pessoas com idade entre 30 e 70 anos. A prevalência dessas patologias foi maior em países em desenvolvimento e subdesenvolvidos, nos quais ocorreram 78% das mortes relacionadas às DCNT, sendo 85% destas ocorrências prematuras (WHO, 2018). Dados mais recentes indicam que, em 2019, as DCNT responderam por 73,6% das mortes globais (WHO, 2021).

No Brasil, o Ministério da Saúde informa que, apesar da tendência de redução nas mortes prematuras, passando de 52,2% em 2000 para 43,7% em 2019, esses números ainda são considerados elevados e implicam em consequências econômicas importantes, como o aumento nos gastos com saúde e impacto negativo na qualidade de vida dos indivíduos (Ministério da Saúde, 2021). Assim, mesmo com a redução observada, a mortalidade prematura por DCNT permanece como um problema de alta relevância sanitária e econômica, pois mantém uma carga de custos assistenciais e perdas de produtividade que justificam o fortalecimento de estratégias de prevenção, diagnóstico oportuno e controle dessas condições.

Diante desse cenário, o Ministério da Saúde recomenda ações de enfrentamento baseadas na vigilância contínua das DCNT e de seus fatores de risco (Ministério da Saúde, 2021). Considerando o impacto dessas doenças nos âmbitos produtivo, social, familiar e econômico, o Ministério enfatiza que a análise dos padrões de mortalidade prematura e de sua distribuição populacional pode subsidiar o planejamento, a priorização e o monitoramento de estratégias de prevenção e cuidado (Ministério da Saúde, 2021).

Nesse contexto, produzir conhecimento aplicável à tomada de decisão em relação ao DM2 torna-se fundamental, sobretudo para apoiar o rastreamento de risco, a estratificação de pacientes e o direcionamento de intervenções. Métodos de Inteligência Artificial (IA) podem contribuir nesse processo ao permitir a construção de modelos preditivos capazes de identificar indicadores associados ao prognóstico do Diabetes Mellitus. Esses modelos extraem padrões e

relações a partir de grandes conjuntos de dados, inclusive aqueles que são de difícil mensuração e interpretação por abordagens tradicionais (Silva et al., 2021).

As ferramentas de Aprendizagem de Máquina (AM) são de grande importância nesse processo, pois contribuem para a intervenção precoce contra o DM2, reduzindo danos à saúde dos indivíduos com predisposição à doença e minimizando custos para a família e para a sociedade, uma vez que essa doença está entre as principais causas de mortes por DCNT no mundo (Gowdak, 2025).

Diante do crescente impacto das DCNT, em especial a complexidade do DM2, observa-se um aumento expressivo do volume de dados clínicos e nutricionais disponíveis na prática e na pesquisa. Entretanto, a multiplicidade de variáveis coletadas nem sempre se traduz em informação útil, pois muitos atributos podem ser redundantes, pouco informativos ou inconsistentes entre estudos, o que dificulta a identificação de marcadores prioritários e compromete a interpretabilidade e o desempenho de modelos preditivos. Essa lacuna, a ausência de uma definição/seleção mais robusta dos atributos mais relevantes para caracterizar e prever a condição metabólica no DM2, motiva esta pesquisa, que propõe aplicar métodos de IA, com ênfase em AM, para apoiar a seleção e análise de atributos associados à doença. Assim, busca-se identificar padrões e potenciais biomarcadores, fortalecendo a prática nutricional baseada em evidências e aprimorando estratégias de prevenção e intervenção.

## **2. FUNDAMENTAÇÃO TEÓRICA**

### **2.1 Panorama das Doenças Crônicas Não Transmissíveis (DCNT)**

As DCNT são específicas de um dos principais desafios de saúde pública da atualidade. Segundo estimativas da Organização Mundial da Saúde (OMS), essas condições foram responsáveis por aproximadamente 70% de todas as mortes globais em 2021, incluindo doenças cardiovasculares, neoplasias, doenças respiratórias crônicas e o Diabetes Mellitus (WHO, 2025). Além do impacto na mortalidade, as DCNT também respondem por uma parcela significativa dos *Disability-Adjusted Life Years* (DALYs), refletindo o impacto prolongado que essas condições exercem sobre a qualidade de vida, a produtividade e os custos assistenciais.

Entre essas condições, o Diabetes Mellitus destaca-se pelo crescimento acelerado e pelo impacto de longo prazo. A edição mais recente do IDF Diabetes Atlas estima que cerca de 589 milhões de adultos entre 20 e 79 anos vivem com diabetes em todo o mundo, dessa forma os gastos globais com diabetes já ultrapassaram US\$ 1 trilhão de dólares, tornando a doença um dos maiores componentes do custo de saúde global (International Diabetes Federation, 2025a).

A situação na Região da América do Sul e Central (SACA) segue a tendência mundial. Segundo o *International Diabetes Federation* (IDF), as estimativas regionais foram construídas a partir de 29 fontes de dados provenientes de 15 países, no entanto, apenas os países, Brasil, Peru, Panamá e Bolívia, realizaram análises epidemiológicos recentes nos últimos cinco anos (International Diabetes Federation, 2025b). As análises realizadas indicam que o número de pessoas com diabetes na região deverá aumentar 46% até 2050, atingindo 52 milhões de indivíduos. Em 2024, a região SACA registrou aproximadamente US\$ 81 bilhões em gastos relacionados ao diabetes, o que corresponde a cerca de 8% do total mundial (IDF, 2025).

No Brasil, a carga da DCNT também é expressiva. Em 2019, cerca de 41,8% das mortes entre adultos de 30 a 69 anos foram atribuídas às DCNT, conforme dados oficiais do Ministério da Saúde (Governo federal do Brasil, 2021). O VIGITEL, sistema de vigilância que monitora fatores de risco desde 2006, identifica tabagismo, consumo alimentar inadequado, uso nocivo de álcool e inatividade física como os principais comportamentos associados ao desenvolvimento das DCNT entre adultos residentes nas capitais brasileiras (Ministério da Saúde, 2022). Esses fatores comportamentais, amplamente difundidos na população, são indicados para o aumento da prevalência de obesidade, hipertensão e dislipidemias, estabelecendo o cenário ideal para a progressão de doenças metabólicas.

Dados mais recentes do Painel de Monitoramento da Mortalidade Prematura por DCNT reforçam a gravidade do problema no país. Somente no último ano com dados completos, foram registrados 198.277 óbitos prematuros associados à DCNT, concentrados principalmente em faixas etárias economicamente produtivas. Esse número expressivo demonstra que as DCNT continuam sendo responsáveis por grande parte da mortalidade evitável no Brasil, evidenciando a importância de estratégias contínuas de vigilância, promoção da saúde e intervenções preventivas capazes de reduzir essa carga (Sistema de Informações sobre Mortalidade, 2025).

Dentro desse panorama, o DM2 assume papel de destaque. Trata-se de uma condição altamente prevalente, de progressão silenciosa e fortemente influenciada por fatores clínicos, comportamentais e sociodemográficos. Diante do aumento contínuo do número de casos e dos altos custos associados às complicações, torna-se essencial adotar abordagens inovadoras que permitam identificar precocemente indivíduos em risco e apoiar ações de prevenção. Entre essas estratégias, as ferramentas de IA e Aprendizado de Máquina são apresentadas especialmente promissoras, ao possibilitar a análise de grandes volumes de dados, a identificação de padrões complexos e o desenvolvimento de modelos preditivos eficientes que podem auxiliar na tomada de decisão em saúde coletiva e clínica.

## **2.2 Diabetes Mellitus: Conceitos e Classificação**

O Diabetes Mellitus (DM) é um distúrbio metabólico caracterizado por hiperglicemia persistente decorrente de defeitos na secreção ou ação da insulina. A insulina, produzida pelas células  $\beta$  pancreáticas, regula a entrada de glicose nas células, e sua deficiência absoluta ou relativa resulta em

acúmulo de glicose na circulação (SBEM, 2025).

Entre os diferentes tipos de diabetes, o DM2 é o mais prevalente, correspondendo a cerca de 90% dos casos. Essa condição apresenta caráter multifatorial e envolve, principalmente, resistência à insulina e redução progressiva da capacidade secretória das células  $\beta$ . Em estágios iniciais, pode ocorrer hiperinsulinemia compensatória, mas, com o avanço da doença, instala-se hiperglicemia sustentada (Basak; Laskar, 2024; Singh; Gholipourmalekabadi; Shafikhani, 2024).

Diversos fatores contribuem para a fisiopatologia do DM2, incluindo alterações metabólicas, depósito excessivo de gordura, disfunção hepática, inflamação de baixo grau e envelhecimento. Esses mecanismos favorecem a evolução gradual da doença e seu caráter inicialmente silencioso, reforçando a importância da detecção precoce. A progressão do DM2 está associada a maior risco de complicações cardiovasculares, renais e neurológicas, especialmente quando não identificada ou tratada adequadamente (Gieroba; Kryska; Sroka-Bartnicka, 2025; Vybhavi et al., 2025).

De acordo com diretrizes internacionais, o DM é classificado em quatro grupos principais: diabetes tipo 1 (autoimune, com deficiência absoluta de insulina), diabetes tipo 2 (resistência insulínica e disfunção secretória), diabetes mellitus gestacional (intolerância à glicose diagnosticada na gestação) e outros tipos específicos associados a causas monogênicas, doenças pancreáticas ou uso de determinados medicamentos (Li; Jiang; Sun, 2022; Yameny, 2024). Para sintetizar os principais parâmetros adotados para o diagnóstico do DM2, a Tabela 1 apresenta os critérios recomendados pela ADA e pela SBD.

Tabela 1 - Critérios diagnósticos segundo ADA/SBD

| <b>Critérios</b>                     | <b>Normal</b> | <b>Pré-diabetes</b> | <b>Diabetes Mellitus (DM)</b> |
|--------------------------------------|---------------|---------------------|-------------------------------|
| Glicemia de jejum (mg/dL)            | Menor que 100 | 100 - 125           | $\geq 126$                    |
| Glicemia ao acaso (mg/dL) + sintomas | -             | -                   | $\geq 200$                    |
| Glicemia de 1 horas no TTGO (mg/dL)  | Menor que 155 | 155-208             | $\geq 209$                    |
| Glicemia de 2 horas no TTGO (mg/dL)  | Menor que 140 | 140 - 199           | $\geq 200$                    |
| HbA1c (%)                            | Menor que 5,7 | 5,7 - 6,4           | $\geq 6,5$                    |

Fonte: Adaptado de diretrizes da Sociedade Brasileira de Diabetes (SBD, 2023).

O DM2 destaca-se por sua forte associação com fatores modificáveis, como obesidade, sedentarismo, padrões alimentares inadequados e consumo excessivo de álcool, além de determinantes sociodemográficos como idade, nível educacional e renda. Por ser uma condição de instalação gradual e frequentemente assintomática, muitos casos não são diagnosticados precocemente, o que evidencia a necessidade de ferramentas capazes de identificar indivíduos sob maior risco antes da manifestação clínica da doença (Mizukami; Kudoh, 2022).

Nesse contexto, a compreensão dos mecanismos gerais do DM2 e sua classificação fundamenta

a aplicação de modelos de IA voltados à predição da doença. Tais modelos podem integrar variáveis clínicas, comportamentais e sociodemográficas para detectar padrões associados ao risco metabólico e apoiar a tomada de decisão em saúde.

### 2.3 Critérios diagnósticos

O diagnóstico do Diabetes Mellitus é estabelecido por meio de parâmetros laboratoriais padronizados internacionalmente, conforme recomendações da *American Diabetes Association* (ADA) e da Organização Mundial da Saúde (OMS). Esses critérios baseiam-se em três exames considerados fundamentais: glicemia plasmática de jejum  $\geq 126$  mg/dL, glicemia  $\geq 200$  mg/dL duas horas após o teste oral de tolerância à glicose (TOTG, 75 g) e hemoglobina glicada (HbA1c)  $\geq 6,5\%$ . Esses métodos são amplamente utilizados na prática clínica por refletirem diferentes aspectos do metabolismo da glicose. Enquanto o TOTG e a glicemia de jejum quantificam diretamente a resposta glicêmica aguda, a HbA1c expressa a média das concentrações glicêmicas nos últimos dois a três meses, apresentando menor variabilidade diária (Duong et al., 2023).

Embora o TOTG seja tradicionalmente considerado o padrão-ouro por sua sensibilidade superior para identificar intolerância à glicose e alterações metabólicas precoces, seu uso é limitado pela maior complexidade operacional e pelo tempo necessário para sua execução. Em contrapartida, a glicemia de jejum e a HbA1c apresentam maior praticidade e aderência em cenários populacionais. Na análise comparativa realizada por (Duong et al., 2023), a glicemia de jejum  $\geq 126$  mg/dL apresentou especificidade de 0,98, configurando-se como o exame mais eficaz para confirmação diagnóstica, enquanto a combinação entre HbA1c  $\geq 6,5\%$  e glicemia de jejum aumentou a sensibilidade, reduzindo a possibilidade de casos não detectados quando apenas um indicador é utilizado.

Além desses exames, algumas diretrizes incluem como critério diagnóstico a presença de sintomas clássicos de hiperglicemia associados a uma glicemia plasmática casual  $\geq 200$  mg/dL (Rodacki et al., 2024). Entretanto, esse critério tem sido discutido na literatura. Em sua avaliação crítica, Mirzazada et al., (2021), observaram que a glicemia aleatória  $\geq 200$  mg/dL apresentou especificidade de apenas 69% para diagnóstico de diabetes, enquanto o ponto de corte de  $\geq 250$  mg/dL mostrou especificidade de 100%, não sendo observado em indivíduos com metabolismo normal ou pré-diabetes. Considerando que sintomas como poliúria, polidipsia, perda de peso e fadiga não são exclusivos do diabetes, os autores sugerem que o valor  $\geq 250$  mg/dL possa representar um limiar mais adequado quando sintomas estão presentes.

No contexto da gestação, os critérios diagnósticos também apresentam variações

metodológicas entre diferentes organismos internacionais. Embora esse tema não constitua o foco deste estudo, (Moon; Jang, 2022), destacam que diretrizes divergentes baseadas em estratégias de um ou dois passos refletem a busca contínua por parâmetros capazes de equilibrar sensibilidade diagnóstica e prevenção de desfechos adversos maternos e neonatais.

Em síntese, os critérios diagnósticos consolidados pela ADA e OMS representam o padrão de referência para identificação do Diabetes Mellitus, oferecendo uma estrutura robusta e amplamente validada para diferenciar normoglicemia, pré-diabetes e diabetes. Esses parâmetros constituem a base conceitual utilizada na literatura científica e são essenciais para a compreensão da epidemiologia e da evolução clínica da doença.

## **2.4 Fatores de risco clínicos, sociodemográficos e comportamentais para o Diabetes Mellitus tipo 2**

O risco de desenvolvimento e progressão do DM2 resulta de um conjunto de fatores clínicos, sociodemográficos e comportamentais que se inter-relacionam e influenciam diretamente o estado metabólico dos indivíduos. A literatura recente destaca que variáveis pessoais, como idade, condição socioeconômica, escolaridade, presença de comorbidades e hábitos de vida, exercem impacto significativo tanto sobre o controle glicêmico quanto sobre a probabilidade de complicações da doença. Em um estudo multicêntrico realizado no Sri Lanka com mais de dois mil pacientes, a inadequação do controle glicêmico esteve associada a fatores como idade mais jovem no diagnóstico, menor adesão ao tratamento, uso de insulina, consumo diário de álcool e pertencimento a grupos étnicos minoritários, evidenciando a importância dos determinantes sociais e comportamentais no manejo do DM2 (Zoysa et al., 2025).

Resultados semelhantes foram observados em uma investigação conduzida na Arábia Saudita com 980 indivíduos diabéticos, na qual a ocorrência de complicações esteve fortemente associada a características sociodemográficas, como idade mais elevada, menor escolaridade, renda reduzida e situação de desemprego, sinalizando que vulnerabilidades sociais podem intensificar a progressão da doença. Nesse mesmo estudo, fatores clínicos apresentaram impacto expressivo, especialmente a presença de hipertensão, hiperlipidemia, obesidade, doença cardiovascular, doença renal e maior tempo desde o diagnóstico, que se mostraram consistentemente associados a maior risco de complicações metabólicas (AlOmeir et al., 2025).

Os fatores comportamentais também desempenham papel fundamental. Padrões alimentares inadequados caracterizados por elevado consumo de bebidas açucaradas, carboidratos refinados, produtos industrializados e gorduras saturadas estiveram associados a

maior ocorrência de complicações, enquanto dieta rica em frutas, vegetais e leite mostrou efeito protetor. Outros comportamentos de risco, como tabagismo, uso de dispositivos eletrônicos para fumar, sedentarismo e privação de sono, também foram associados a pior prognóstico metabólico, reforçando a relevância dos hábitos de vida na evolução do DM2 (AlOmeir et al., 2025). Esses achados são consistentes com evidências da literatura, que indicam que fatores modificáveis relacionados ao estilo de vida como dieta, atividade física e padrões de sono representam componentes centrais na prevenção e no controle da doença.

No contexto da IA aplicada à saúde, revisões sistemáticas demonstram que atributos clínicos, sociodemográficos e comportamentais são amplamente utilizados como variáveis preditoras em modelos destinados a estimar risco de DM2 ou de complicações relacionadas, devido à sua forte capacidade de discriminação e à relevância fisiológica desses fatores (Aparicio et al., 2021). A análise de preditores utilizada em diversos modelos confirma que a combinação entre marcadores clínicos (como IMC, pressão arterial e perfil lipídico), determinantes sociodemográficos (como escolaridade, renda e idade) e comportamentos de saúde (como dieta, tabagismo e atividade física) aumenta de forma substancial a acurácia dos algoritmos. Além disso, estudos em IA destacam que abordagens capazes de integrar múltiplos domínios de dados clínicos, comportamentais e sociais oferecem melhor desempenho na identificação de padrões associados ao DM2 e possibilitam interpretações mais alinhadas à prática clínica (Esteve et al., 2019).

Assim, a literatura demonstra de forma consistente que fatores clínicos, sociodemográficos e comportamentais constituem componentes essenciais na determinação do risco para DM2 e desempenham papel central em modelos preditivos. A incorporação simultânea desses elementos permite capturar a complexidade da doença, contribuindo para estratégias mais eficazes de prevenção, diagnóstico precoce e intervenção personalizada. A Tabela 2 apresenta uma síntese dos fatores clínicos, comportamentais e sociodemográficos associados ao risco para DM2, conforme discutido anteriormente.

Tabela 2 - Principais fatores clínicos, comportamentais e sociodemográficos associados ao risco de DM2

| <b>Categoria</b>                    | <b>Fatores de risco</b>                   | <b>Evidências associadas</b>                                |
|-------------------------------------|-------------------------------------------|-------------------------------------------------------------|
| Clínicos e Metabólicos              | Hipertensão arterial                      | Aumenta resistência à insulina e risco cardiovascular.      |
|                                     | Colesterol elevado / dislipidemias        | Contribui para inflamação e piora do metabolismo glicídico. |
|                                     | Obesidade / IMC elevado                   | Forte preditor de resistência à insulina e DM2.             |
|                                     | Doença cardiovascular                     | Compartilha vias fisiopatológicas com DM2.                  |
|                                     | Doença renal                              | Relacionada ao agravamento de complicações.                 |
|                                     | Maior tempo de diagnóstico                | Relacionado à pior evolução metabólica.                     |
| Comportamentais / Estilo de Vida    | Dieta rica em açúcares e ultraprocessados | Aumenta obesidade, inflamação e hiperglicemia.              |
|                                     | Baixo consumo de frutas e vegetais        | Reduz fibras e antioxidantes protetores.                    |
|                                     | Álcool em excesso                         | Relacionado à desregulação metabólica.                      |
|                                     | Sedentarismo                              | Reduz sensibilidade à insulina.                             |
|                                     | Tabagismo / vaping                        | Aumenta estresse oxidativo e inflamação.                    |
| Percepção de Saúde / Funcionalidade | Privação de sono                          | Prejudica controle hormonal e glicêmico.                    |
|                                     | Autoavaliação ruim da saúde               | Preditora de pior autocuidado e maior risco metabólico.     |
|                                     | Dificuldade para caminhar                 | Indicador de fragilidade e maior risco para DM2.            |

Fonte: Elaborado pela autora, com base em Zoysa et al. (2025), AlOmeir et al. (2025), Aparicio et al. (2021), Esteva et al. (2019) e demais estudos citados na seção 2.3.

## 2.5 Bases de dados e Big Data em saúde (BRFSS)

O avanço das tecnologias digitais e a expansão da capacidade de armazenamento e processamento têm impulsionado o uso de Big Data em saúde, possibilitando análises de larga escala orientadas por técnicas de mineração de dados e aprendizado de máquina. Diferentemente dos bancos clínicos tradicionais, os grandes repositórios populacionais reúnem informações extensas, heterogêneas e continuamente atualizadas, favorecendo a investigação de fatores de risco, o monitoramento de doenças crônicas e o desenvolvimento de modelos

preditivos aplicáveis à saúde pública. Conforme discutido por (Wu et al., 2021), esses bancos caracterizam-se por alta dimensionalidade, grande volume amostral e variabilidade temporal, exigindo metodologias robustas de pré-processamento, seleção de atributos e modelagem estatística para que seu potencial analítico seja plenamente aproveitado.

Nesse contexto, um dos sistemas mais amplamente utilizados em epidemiologia comportamental e vigilância de doenças crônicas nos Estados Unidos é o *Behavioral Risk Factor Surveillance System* (BRFSS). Coordenado pelo *Centers for Disease Control and Prevention* (CDC), o BRFSS constitui o maior inquérito telefônico contínuo de saúde do mundo, reunindo dados autorrelatados de mais de 400 mil adultos por ano, obtidos por amostragem aleatória estratificada. Segundo (Wu, 2024), a abrangência do BRFSS e sua capacidade de captar múltiplas dimensões associadas ao estilo de vida incluindo comportamentos de risco, condições crônicas autorreferidas, fatores sociodemográficos e acesso a serviços tornam-no uma fonte altamente relevante para análises preditivas sobre diabetes e outras doenças de interesse populacional.

O BRFSS se destaca entre bancos de Big Data em saúde pelo volume e diversidade dos dados coletados, são mais de 300 variáveis que incluem informações sobre idade, escolaridade, renda, hábitos alimentares, prática de atividade física, tabagismo, consumo de álcool, histórico de hipertensão, colesterol elevado, doenças cardiovasculares, condições respiratórias e indicadores de saúde mental (Hu et al., 2025). Esse carácter multidimensional permite capturar a natureza multifatorial de doenças como o DM2, cuja etiologia envolve fatores biológicos, comportamentais e ambientais.

Outro aspecto central é o potencial do BRFSS para estudos de modelagem preditiva. A literatura mostra que diferentes algoritmos de Aprendizado de Máquina têm sido aplicados com sucesso para identificar padrões associados ao risco de diabetes utilizando esses dados. Estudos citados por Wu, (2024), demonstram desempenhos elevados de modelos como *Random Forest*, *Decision Trees* e redes neurais na classificação de indivíduos com diabetes, com acurácia variando entre 80% e 96% dependendo das variáveis selecionadas e do ano da base utilizada.

Modelos baseados em *Random Forest* têm se destacado por sua capacidade de lidar com dados de alta dimensionalidade, reduzir sobreajuste e explorar interações complexas entre variáveis, características comuns em bases populacionais extensas. Além disso, a estrutura do BRFSS favorece análises replicáveis e comparações temporais, visto que o levantamento segue protocolos padronizados e consiste em uma fonte pública e acessível de dados de vigilância. A inclusão recorrente de módulos temáticos (módulos core, opcionais e estaduais) permite

investigar tendências emergentes e o impacto de mudanças comportamentais ao longo do tempo. Para doenças crônicas como o diabetes, cuja prevalência está fortemente associada a comportamentos modificáveis, esse tipo de monitoramento contínuo é especialmente relevante.

Por fim, embora apresente vantagens como grande representatividade e volume amostral, o BRFSS também possui limitações inerentes a dados autorrelatados, incluindo viés de memória, subdiagnóstico e ausência de medidas laboratoriais diretas. Como destacam (Wu et al., 2021), esses desafios reforçam a importância de técnicas avançadas de limpeza, imputação e seleção de atributos, etapas essenciais na construção de modelos robustos a partir de Big Data em saúde. Mesmo assim, sua capacidade de integrar fatores demográficos, comportamentais e clínicos o torna uma plataforma valiosa para estudos preditivos focados na identificação precoce do risco de diabetes em nível populacional.

## **2.6 Inteligência Artificial e Aprendizado de Máquina na saúde**

Os avanços recentes em IA têm transformado o cenário da saúde ao permitir análises complexas de grandes volumes de dados clínicos, epidemiológicos e comportamentais, ampliando a capacidade de detecção precoce, diagnóstico e tomada de decisão. Na literatura, a IA compreende um conjunto de técnicas computacionais capazes de aprender padrões a partir de dados, sendo o Aprendizado de Máquina (AM) uma de suas principais vertentes, responsável por desenvolver modelos que identificam relações não lineares e auxiliam na predição de eventos clínicos. Segundo Maheshwari et al., (2025), a incorporação de IA na prática médica tem ocorrido em múltiplas frentes, desde apoio ao diagnóstico até monitoramento remoto por dispositivos vestíveis, integrando dados heterogêneos provenientes de prontuários eletrônicos, sensores contínuos de glicose e imagens médicas.

No campo das doenças crônicas, o diabetes tem se destacado como uma das áreas com maior produção científica em IA, devido à disponibilidade de grandes bases de dados e à multiplicidade de fatores de risco envolvidos na progressão da doença. Modelos supervisionados, como *Support Vector Machines* (SVM), *Random Forest* (RF), *k-Nearest Neighbors* (kNN), *Logistic Regression*, bem como arquiteturas de *Deep Learning*, têm sido amplamente aplicados na predição, detecção e manejo do DM2. Uma síntese das evidências apresentada por Olisah; Smith; Smith, (2022), demonstra que modelos de Aprendizado de Máquina aplicados ao diabetes têm alcançado acurácias que variam amplamente, de aproximadamente 80% a valores superiores a 95%, dependendo da qualidade dos dados utilizados, das estratégias de seleção de atributos e dos métodos empregados para lidar com

valores ausentes. Os autores destacam que etapas de pré-processamento, incluindo imputação de dados faltantes e análise de correlação para seleção de variáveis, influenciam diretamente o desempenho dos algoritmos, sendo fundamentais para aumentar a robustez e a precisão dos modelos preditivos.

Além da predição diagnóstica, a IA tem sido empregada em aplicações relacionadas ao tratamento e ao acompanhamento clínico. Modelos baseados em redes neurais têm demonstrado capacidade de prever tendências glicêmicas futuras, auxiliar na titulação de insulina e otimizar sistemas de pâncreas artificial, integrando dados gerados por sensores contínuos de glicose. Maheshwari et al., (2025), relatam que o uso de IA tem ampliado a precisão na identificação de hipoglicemia iminente e melhorado a tomada de decisão terapêutica por meio de sistemas inteligentes que interpretam padrões glicêmicos e fatores comportamentais.

A literatura também aponta para o potencial da IA na detecção de complicações associadas ao diabetes. Métodos de Visão Computacional e Redes Convolucionais (CNNs) têm alcançado desempenho comparável ao de especialistas na identificação de retinopatia diabética, nefropatia e neuropatia, o que representa um avanço significativo no rastreamento dessas complicações em larga escala. Em paralelo, Ma et al., (2025), destacam que sistemas de IA aplicados ao suporte à decisão clínica têm ampliado a capacidade de analisar grandes volumes de dados provenientes de histórico médico, resultados de exames e informações de acompanhamento, favorecendo avaliações mais precisas e auxiliando profissionais de saúde na identificação de riscos e na escolha de intervenções mais adequadas.

No contexto específico da predição de diabetes, modelos computacionais vêm sendo utilizados para analisar atributos comportamentais e sociodemográficos, promovendo maior eficiência em estratégias de prevenção. Ponnaboina e Veeraswamy (2025), apresentam um sistema baseado em técnicas de Aprendizado de Máquina, complementado por IA generativa, para predição de diabetes e oferta de recomendações personalizadas. O modelo utiliza parâmetros clínicos simples, como glicose, pressão arterial, IMC e idade, processados por um classificador SVM treinado com dados públicos. Após a predição, um modelo generativo (Google Gemini) produz orientações individualizadas sobre estilo de vida e manejo da saúde. Essa integração entre AM e IA generativa permite combinar estimativas de risco em tempo real com aconselhamento automatizado, ampliando o potencial de uso do sistema em contextos de prevenção e educação em saúde.

Apesar dos avanços, desafios persistem na integração da IA à saúde. Aspectos como qualidade dos dados, variabilidade entre populações, necessidade de transparência dos modelos

e exigências éticas relacionadas à privacidade são frequentemente discutidos na literatura. (Maheshwari et al., 2025), ressaltam que a adoção clínica de modelos preditivos depende não apenas da performance, mas também da interpretabilidade e da capacidade de generalização para diferentes contextos reais.

Os estudos revisados convergem ao demonstrar que a IA e o AM desempenham papel central no avanço do monitoramento e da predição de doenças crônicas, oferecendo ferramentas capazes de identificar padrões complexos, reduzir o subdiagnóstico e aprimorar estratégias preventivas. No caso do diabetes, esse potencial é ainda mais evidente, dada a robustez dos modelos desenvolvidos e a capacidade dessas tecnologias de transformar a avaliação e o manejo dos fatores de risco em saúde populacional.

## **2.7 Aprendizado de Máquina aplicado especificamente ao Diabetes**

A aplicação de técnicas de Aprendizado de Máquina ao Diabetes Mellitus tem avançado de forma significativa nos últimos anos, impulsionada pela disponibilidade de grandes conjuntos de dados clínicos e pela necessidade de métodos mais precisos para triagem, diagnóstico precoce, manejo terapêutico e predição de desfechos. O AM destaca-se como uma abordagem capaz de identificar padrões complexos e não lineares relacionados ao risco de diabetes, superando limitações de métodos tradicionais baseados exclusivamente em medidas isoladas de glicemia ou em modelos estatísticos linearmente paramétricos.

No campo do diagnóstico e da estratificação de risco, modelos supervisionados vêm sendo empregados para integrar variáveis clínicas, laboratoriais e demográficas, permitindo estimativas acuradas de probabilidade de diabetes. Estudos realizados por Agliata et al., (2023), desenvolveram um modelo robusto baseado em AM para predição binária de diabetes utilizando bancos de dados amplos e heterogêneos, como NHANES, MIMIC-III e MIMIC-IV. O modelo apresentou desempenho elevado, com acurácia superior a 86% e AUC de 0,934, indicando alta capacidade discriminativa entre indivíduos com e sem diabetes. A análise de importância de atributos revelou maior contribuição de variáveis como idade, IMC, triglicerídeos, pressão arterial sistólica, circunferência da cintura, glicemia e HbA1c, refletindo a natureza multifatorial da doença.

A avaliação do impacto da remoção progressiva dessas variáveis reforça a necessidade de combinações multivariadas para predições precisas. No estudo de Agliata et al. (2023), a exclusão de atributos altamente informativos resultou em quedas substanciais na acurácia e no AUC, evidenciando que abordagens simplificadas comprometem sensibilidade e

especificidade. Esses achados corroboram o caráter complexo do diabetes, cuja predição depende da integração simultânea de aspectos clínicos, antropométricos e metabólicos. Além do diagnóstico, técnicas de AM têm sido amplamente estudadas para estimar o risco futuro de desenvolvimento de diabetes em populações saudáveis ou com pré-diabetes. Abordagens baseadas em dados transversais e séries temporais possibilitam modelar padrões dinâmicos de evolução glicêmica, utilizando redes neurais recorrentes e modelos de regressão avançada para prever incidência ao longo dos anos e integrar fatores comportamentais, metabólicos e sociodemográficos.

Outro eixo relevante diz respeito à previsão glicêmica e ao manejo terapêutico de indivíduos já diagnosticados. Estudos como os de Afsaneh et al., (2022); Dritsas; Trigka, (2022), demonstram que modelos de AM têm sido efetivos na previsão de valores futuros de glicemia a partir de sinais provenientes de sensores contínuos de glicose, taxas de infusão de insulina, ingestão alimentar e variáveis de estilo de vida. Redes neurais profundas (LSTM, GRU) mostram desempenho promissor ao capturar dependências temporais e oscilações glicêmicas, auxiliando no ajuste de doses de insulina e no desenvolvimento de sistemas de pâncreas artificial. Tais modelos podem antecipar episódios de hipo e hiperglicemia, aumentando a segurança terapêutica.

O AM também tem sido aplicado na detecção automatizada de complicações do diabetes. Métodos de visão computacional aplicados a imagens de retina e exames de fundo de olho apresentam elevado desempenho na identificação de retinopatia diabética, neuropatia periférica e doença renal incipiente (Sushith et al., 2025; Yang et al., 2025). Essas ferramentas são fundamentais para triagem em larga escala, permitindo identificar precocemente indivíduos com risco elevado de complicações e otimizar fluxos de encaminhamento (Mohanty et al., 2023).

Por fim, vários estudos integram AM a sistemas de apoio à decisão clínica, oferecendo estimativas personalizadas de risco e recomendações automatizadas de manejo, como no modelo proposto por Ponnaboina; Veeraswamy. (2025), embora tais ferramentas não substituam o julgamento clínico, complementam a prática assistencial e ampliam o alcance de intervenções estratégicas em populações numerosas e heterogêneas.

Em síntese, o AM aplicado ao diabetes constitui um campo em rápida evolução, com forte potencial para diagnóstico, predição e monitoramento contínuo. A combinação de múltiplas variáveis e o uso de modelos não lineares proporcionam maior acurácia e generalização. Apesar de desafios relacionados à qualidade dos dados, representatividade

populacional e interpretabilidade, o AM tem potencial significativo para transformar estratégias de vigilância, prevenção e cuidado em diabetes, contribuindo para uma medicina mais personalizada e eficiente.

## 2.8 Seleção de atributos (*Feature Selection*)

A seleção de atributos (*Feature Selection*, FS) constitui uma etapa fundamental na construção de modelos de Aprendizado de Máquina (AM), especialmente em contextos de alta dimensionalidade e variáveis fortemente correlacionadas, como em dados clínicos, biomédicos e epidemiológicos. Em termos formais, FS refere-se ao processo de identificar e remover atributos irrelevantes, redundantes ou ruidosos, com o objetivo de melhorar a acurácia preditiva, reduzir o custo computacional, mitigar a maldição da dimensionalidade e aumentar a interpretabilidade dos modelos (Noroozi; Orooji; Erfannia, 2023). Esse processo torna-se ainda mais crítico em bases heterogêneas, onde a redundância estrutural e a correlação interna entre variáveis podem comprometer a estabilidade dos classificadores e inflacionar o risco de overfitting.

De acordo com os estudos realizados por (Bommert et al., 2020; Dokeroglu; Deniz; Kiziloğlu, 2022), os métodos de FS são tradicionalmente categorizados em três classes principais: filtro, *wrapper* e *embedded*. Métodos filtro operam independentemente de qualquer algoritmo de indução e utilizam métricas estatísticas ou informacionais, como *Information Gain*, *Gain Ratio*, *Symmetrical Uncertainty*, *Relief* e *Correlation-based Feature Selection* (CFS), para avaliar relevância e redundância de atributos. Apresentam menor custo computacional e melhor capacidade de generalização, sendo adequados para dados de larga escala (Noroozi; Orooji; Erfannia, 2023). Em contraste, métodos *wrapper* avaliam subconjuntos de atributos com base no desempenho de um classificador, permitindo capturar interações complexas, porém com maior custo computacional. Métodos *embedded* integram a seleção ao processo de aprendizado, como ocorre em modelos com regularização L1/Lasso e árvores de decisão, resultando em soluções mais compactas e alinhadas ao algoritmo utilizado (Dokeroglu; Deniz; Kiziloğlu, 2022).

Revisões recentes indicam que a FS evoluiu substancialmente, incorporando novos paradigmas teóricos, critérios probabilísticos e mecanismos avançados de mensuração de relevância e redundância entre variáveis. Estudos de survey apontam a transição de métodos essencialmente estatísticos e univariados para arcabouços robustos baseados em dependência mútua, medidas de estabilidade, aprendizado probabilístico e estratégias híbridas que agregam múltiplos critérios de avaliação (Ahadzadeh et al., 2023). Além disso, revisões contemporâneas

ênfatizam que o crescimento exponencial do volume e da dimensionalidade dos dados impulsionou o desenvolvimento de abordagens capazes de operar em espaços de busca combinatorial extremamente amplos, especialmente no contexto de big data biomédico e bioinformático (Mostafa et al., 2023).

Paralelamente, diversos autores destacam que a incorporação de algoritmos evolucionários e metaheurísticos como *Particle Swarm Optimization*, *Genetic Algorithms*, *Ant Lion Optimizer*, *Whale Optimization Algorithm* e outras estratégias inspiradas na natureza tornou-se um marco importante na evolução metodológica da FS. Esses métodos oferecem maior capacidade exploratória e permitem lidar com a explosão combinatória inerente à seleção de subconjuntos ótimos em cenários de alta dimensionalidade (Dokeroglu; Deniz; Kiziloz, 2022; Mostafa et al., 2023; Noroozi; Orooji; Erfannia, 2023). Tais abordagens apresentam desempenho especialmente relevante em domínios biomédicos, onde o número de atributos frequentemente supera o número de amostras, exigindo estratégias capazes de evitar overfitting, maximizar acurácia e reduzir redundância estrutural entre variáveis.

Do ponto de vista teórico, *surveys* recentes ressaltam a consolidação de estruturas formais para a classificação dos métodos de FS, abrangendo abordagens filtro, *wrapper*, *embedded* e híbridas, além de novos modelos de avaliação baseados na estabilidade e na robustez estatística dos subconjuntos selecionados (Ahadzadeh et al., 2023; Dokeroglu; Deniz; Kiziloz, 2022; Mostafa et al., 2023). Pesquisas atuais reforçam a necessidade de métodos que conciliem parcimônia, capacidade preditiva e elevada eficiência computacional, considerando que os problemas de seleção de atributos são NP-difíceis e altamente sensíveis à dimensionalidade.

Assim, os estudos recentes fornecem um panorama consolidado sobre o avanço da área, evidenciando que o desenvolvimento contemporâneo de FS está cada vez mais associado a critérios probabilísticos, medidas de dependência não linear, técnicas de aprendizado profundo e algoritmos bioinspirados todos essenciais para lidar com a complexidade crescente das bases de dados modernas.

### **3. OBJETIVO GERAL**

Aplicar técnicas de Inteligência Artificial, com ênfase em Aprendizado de Máquina, para prever Diabetes Mellitus tipo 2 a partir de dados sociodemográficos, clínicos e de estilo de vida (BRFSS), identificando e selecionando os atributos mais relevantes, com vistas a subsidiar

a tomada de decisão do profissional de Nutrição e apoiar estratégias de prevenção e detecção precoce do DM2.

### 3.1 Objetivos Específicos

- Selecionar dataset adequado (BRFSS)
- Realizar pré processamento dos dados
- Construir modelos de árvore de decisão (J48- Weka)
- Avaliar métricas: acurácia
- Identificar os atributos mais relevantes para risco de DM 2

## 4. MATERIAL E MÉTODOS

### 4.1 Tipo de estudo

Trata-se de um estudo quantitativo, de natureza aplicada, baseado em técnicas de Mineração de Dados e Aprendizado de Máquina supervisionado, com o objetivo de desenvolver e avaliar modelos preditivos para classificação do risco de Diabetes Mellitus a partir de atributos clínicos, comportamentais e sociodemográficos.

### 4.2 Fonte dos dados

Foram utilizados dois conjuntos de dados públicos provenientes do sistema *Behavioral Risk Factor Surveillance System* (BRFSS), mantido pelo *Centers for Disease Control and Prevention* (CDC). O BRFSS é considerado o maior sistema contínuo de vigilância de fatores de risco comportamentais do mundo, coletando anualmente informações autorreferidas de adultos residentes nos Estados Unidos sobre condições de saúde, estilo de vida, uso de serviços e comportamentos de risco.

Os datasets utilizados neste estudo foram:

- 1\_diabetes\_012\_health\_indicators\_BRFSS2015\_3\_classes\_N\_P\_D\_v1

(problema multiclasse — C0, C1, C2)

- 3\_diabetes\_binary\_health\_indicators\_BRFSS2015\_2\_classes\_N\_PD

(problema binário — C0, C1)

A base completa inclui 253.680 registros e 21 atributos preditores, além do atributo-alvo correspondente à classificação glicêmica. Os atributos compreendem variáveis sociodemográficas (idade, sexo, escolaridade, renda), comportamentos de saúde (atividade

física, tabagismo, consumo de frutas e vegetais), condições clínicas autorreferidas (hipertensão, colesterol elevado, histórico de AVC e doença cardíaca) e indicadores de saúde percebida.

### **4.3 Definição das classes**

#### **4.3.1 Dataset multiclasse**

O atributo-alvo Diabetes\_012 apresenta três categorias:

- C0: Não diabético
- C1: Pré-Diabetes
- C2: Diabetes Mellitus

#### **4.3.2 Dataset binário**

No dataset binário, Pré-Diabetes e Diabetes foram agrupados, resultando em:

- C0: Não diabético
- C1: Pré-Diabetes ou Diabetes Mellitus

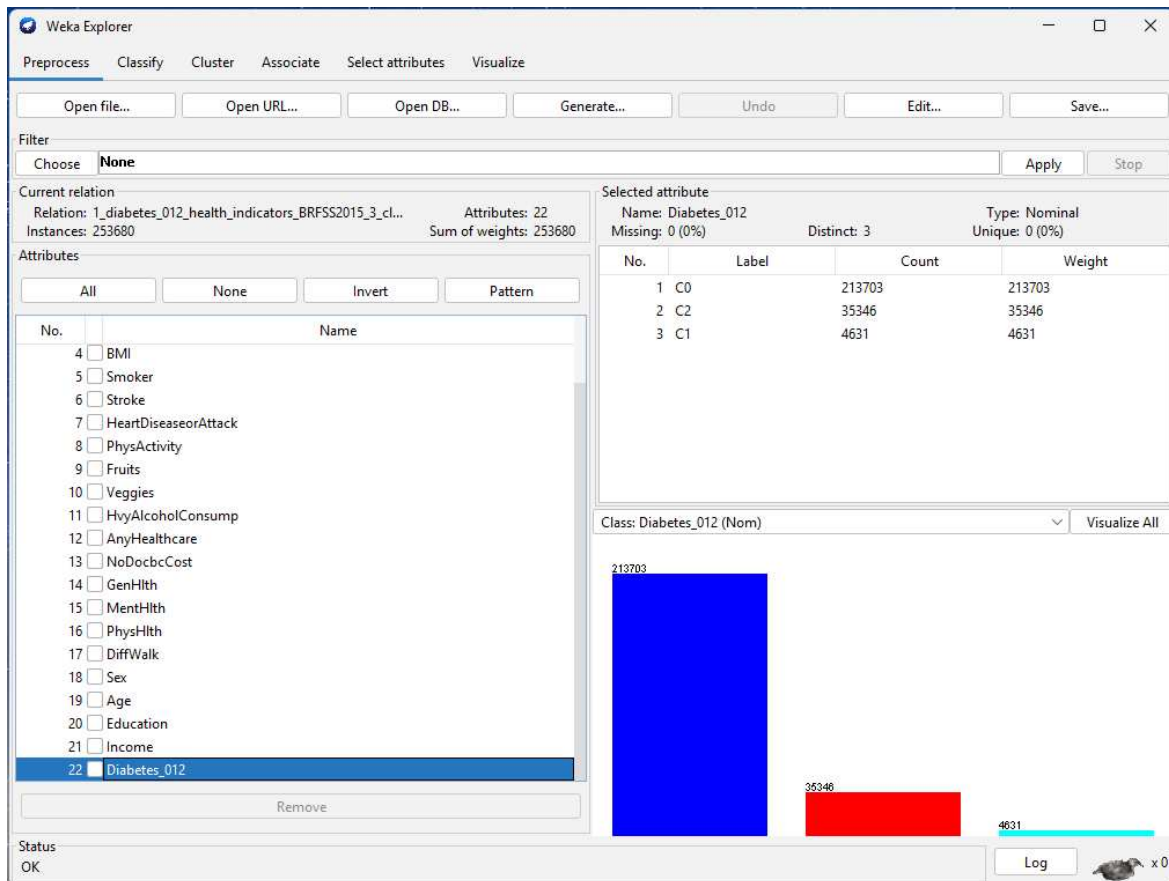
### **4.4 Pré-processamento dos dados**

O pré-processamento consistiu nas seguintes etapas:

- Carregamento dos dados no software Weka 3.8.6, formato CSV.
- Verificação de valores ausentes, constatando-se ausência de *missing values* nos atributos.
- Padronização dos nomes das variáveis, mantendo consistência entre os dois datasets.
- Análise exploratória da distribuição das classes, identificando forte desbalanceamento (Figuras 1 e 2).
- Manutenção dos dados originais sem técnicas de balanceamento, justificando-se a decisão por permitir avaliar o desempenho genuíno do classificador frente a bases reais desbalanceadas, característica comum em estudos epidemiológicos populacionais.

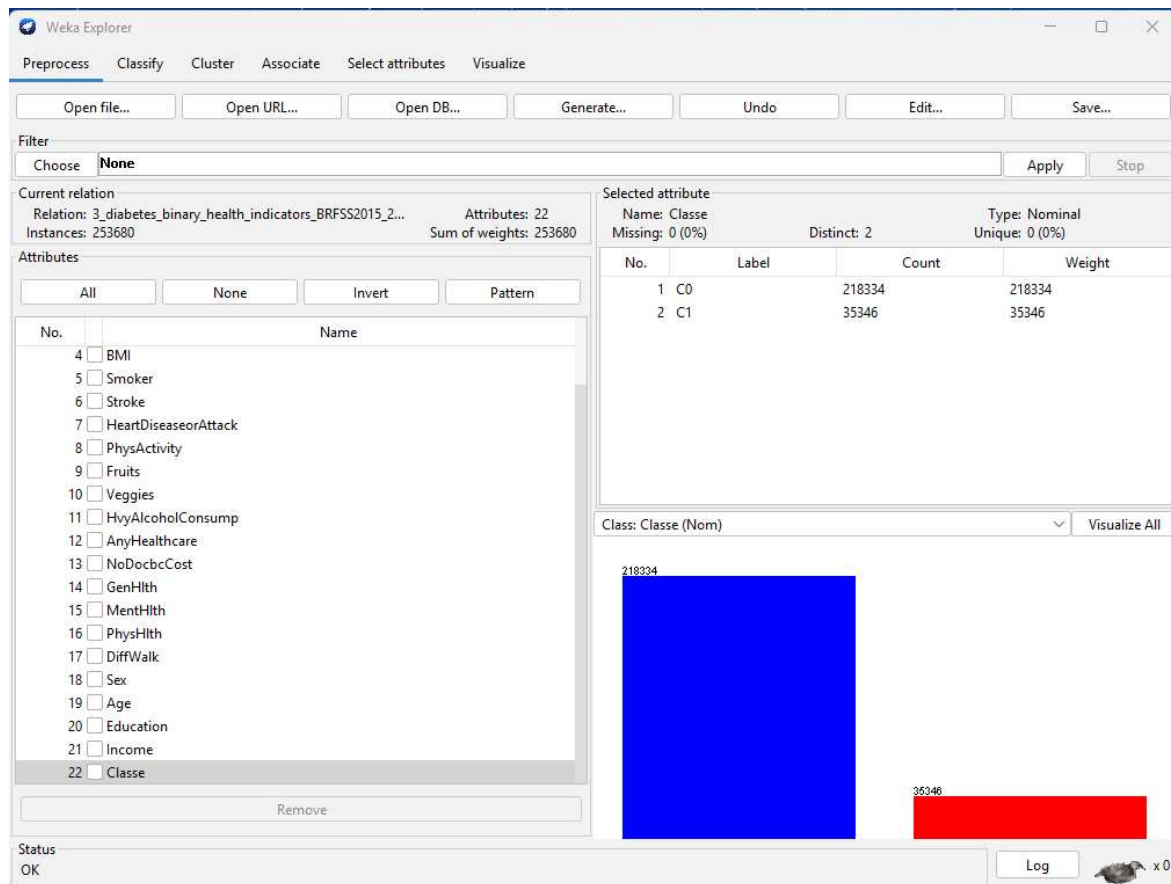
A etapa de pré-processamento respeitou a integridade original da base BRFSS, sem remoção de instâncias ou transformação dos dados.

Figura 1: Distribuição das Classes no Dataset Multiclasse do BRFSS (C0, C1 e C2).



Tela do Weka Explorer (aba Preprocess) mostrando o conjunto de dados 1\_diabetes\_012\_health\_indicators\_BR... (253.680 instâncias; 22 atributos) e a distribuição do atributo de classe Diabetes\_012 (nominal; 3 categorias). Observa-se desbalanceamento entre as classes, com C0 = 213.703 (azul), C2 = 35.346 (vermelho) e C1 = 4.631 (verde água) instâncias, conforme o histograma apresentado.

Figura 2: Distribuição das Classes no Conjunto de Dados Binário (BRFSS)



Tela do Weka Explorer (aba Preprocess) apresentando o conjunto de dados

3\_diabetes\_binary\_health\_indicators\_BR... (253.680 instâncias; 22 atributos) e a distribuição da variável de classe Classe (nominal; 2 categorias). Observa-se desbalanceamento entre as classes, com C0 = 218.334 (azul) e C1 = 35.346 (vermelho) instâncias, conforme o histograma.

#### 4.5 Seleção de Atributos

Para avaliar a relevância dos atributos e reduzir a complexidade dos modelos, foi aplicado o método GainRatioAttributeEval, associado ao algoritmo de busca Ranker, disponível no Weka. Essa abordagem avalia a capacidade de cada atributo em minimizar a entropia e melhorar a separação das classes. Os atributos foram ranqueados segundo seu ganho de informação, resultando em uma lista ordenada de relevância.

Posteriormente, os seis atributos com maior contribuição discriminativa foram selecionados para construção de uma segunda árvore reduzida (modelo 2), visando:

- aumentar a interpretabilidade,
- reduzir complexidade estrutural,
- manter desempenho preditivo adequado.

#### 4.6 Classificador utilizado

O classificador escolhido foi o J48, implementação do algoritmo C4.5, amplamente usado em problemas de classificação clínica por sua interpretabilidade.

Os parâmetros utilizados foram:

- Use Laplace correction: True
- Confidence factor for pruning: 0.25
- Minimum number of instances per leaf: 2
- Pruning: Enabled (reduced-error pruning)

Para ambos os *datasets* (multiclasse e binário), o modelo foi construído sob as mesmas condições experimentais.

A escolha do J48 foi fundamentada em:

- Alta interpretabilidade clínica,
- Facilidade de visualização das regras de decisão,
- Desempenho consistente em bases epidemiológicas,
- Compatibilidade com seleção de atributos baseada em ganho de informação.

#### 4.7 Validação do modelo

A validação foi realizada por meio de *10-fold cross-validation*, técnica que divide o conjunto de dados em 10 subconjuntos estratificados, garantindo representatividade proporcional das classes em cada divisão.

As métricas avaliadas incluíram:

- Acurácia
- Erro Médio Absoluto (MAE)
- Coeficiente Kappa
- Recall por classe
- Precisão
- F1-score
- AUC (quando aplicável)
- Matthews Correlation Coefficient (MCC), métrica recomendada para bases desbalanceadas

As matrizes de confusão foram analisadas detalhadamente para quantificar os erros de classificação por classe.

#### **4.8 Construção dos modelos**

Foram desenvolvidos dois modelos principais:

Modelo 1 – Árvore completa

- Construída com todos os 21 atributos preditores.

Modelo 2 – Árvore reduzida

- Utilizou apenas os seis atributos mais relevantes segundo o GainRatio.

Para cada modelo foram gerados:

- Árvore de decisão completa,
- Métricas de desempenho,
- Matriz de confusão,
- Análise comparativa entre desempenho e interpretabilidade.

#### **4.9 Software e ambiente computacional**

Todas as análises foram conduzidas no software Weka 3.8.6, executado em ambiente Windows 11. O Weka foi escolhido por:

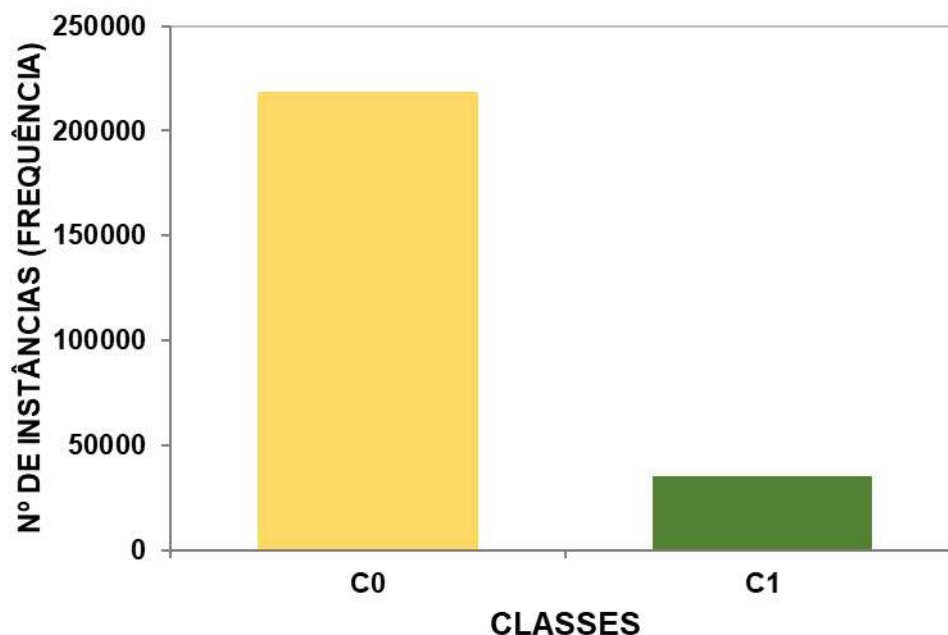
- Ampla aceitação acadêmica,
- Interface gráfica intuitiva,
- Implementação estável dos algoritmos de decisão,
- Suporte nativo a validação cruzada e seleção de atributos.

### **5. RESULTADOS E DISCUSSÃO**

#### **5.1 Descrição inicial dos dados e impacto do desbalanceamento**

A análise inicial das bases do BRFSS evidenciou forte desbalanceamento entre as classes, tanto no dataset multiclasse quanto no dataset binário. No dataset binário (3\_diabetes\_binary\_health\_indicators\_BRFSS2015\_2\_classes\_N\_PD), a classe C0 (não diabéticos) apresentou 218.334 registros, enquanto a classe C1 (pré-diabetes + diabetes) contou com 35.346 registros, conforme ilustrado na Figura 3. Essa distribuição corresponde a aproximadamente 86% de registros em C0 e 14% em C1, indicando uma diferença expressiva na quantidade de exemplos disponíveis para o treinamento do modelo.

Figura 3: Distribuição das classes C0 e C1 no dataset binário.



C0 representa indivíduos sem diabetes (barra amarela). C1 representa indivíduos com pré-diabetes ou diabetes (barra verde).

Esse desbalanceamento de aproximadamente 86% (C0) vs. 14% (C1) afeta diretamente o desempenho dos modelos, levando a maior facilidade em identificar indivíduos saudáveis e maior dificuldade na detecção da classe minoritária justamente o grupo de maior interesse clínico. Esse comportamento é consistente com a literatura, modelos supervisionados tendem a favorecer a classe majoritária em datasets reais de saúde, especialmente quando há assimetria marcada nas proporções de classes (Agyemang et al., 2025; Kumar et al., 2022).

Abousaber; Abdallah; El-Ghaish, (2024), destacam que o desbalanceamento é um dos principais obstáculos para a predição do diabetes em bases populacionais, visto que a grande predominância de indivíduos não diabéticos leva a modelos enviesados e com baixa sensibilidade para a classe minoritária. Os autores alertam que esse viés resulta em falsos negativos clinicamente relevantes, atrasando intervenções preventivas essenciais na transição de pré-diabetes para diabetes tipo 2. Além disso, demonstram que, sem técnicas corretivas de balanceamento, algoritmos como Logistic Regression, Decision Trees e SVM exibem quedas substanciais em recall e sensibilidade para classes positivas. Assim, estratégias como SMOTE, ADASYN ou SMOTE-ENN tornam-se fundamentais para restaurar o equilíbrio no processo de aprendizagem.

No contexto deste estudo, o desbalanceamento observado (86% vs. 14%) representa um desafio metodológico já amplamente documentado na literatura e justifica a necessidade de técnicas específicas para evitar que o modelo priorize excessivamente a classe C0, prejudicando a detecção de indivíduos com risco glicêmico.

Os estudos realizados por Zhang et al., (2025), que analisaram um dataset clínico com 100.000 amostras, no qual apenas 8,5% pertenciam à classe diabética. Os autores observaram queda significativa na capacidade dos modelos tradicionais de reconhecer a classe minoritária, mesmo quando a acurácia global permanecia alta. Segundo o estudo, a escassez de exemplos positivos induz um processo de aprendizagem enviesado, no qual o modelo captura predominantemente padrões da classe majoritária e falha na identificação de casos limítrofes ou iniciais, situação diretamente relacionada ao risco de subdiagnóstico.

Esse panorama descrito por Zhang et al. (2025) é totalmente compatível com o observado neste estudo: uma clara tendência dos modelos em reconhecer C0 com facilidade, mas apresentar dificuldade expressiva na classificação da classe C1. Assim, o comportamento identificado nos resultados é coerente com o que se espera de bases clínicas marcadamente desbalanceadas, reforçando a importância de estratégias mais robustas nas etapas seguintes da modelagem.

## **5.2 Modelo Multiclasse**

No dataset multiclasse

(1\_diabetes\_012\_health\_indicators\_BRFSS2015\_3\_classes\_N\_P\_D\_v1), o modelo J48 apresentou acurácia de 88,845%, indicando bom desempenho geral na classificação das três classes. Entretanto, o modelo resultante mostrou-se altamente complexo, com 11.629 nós e 5.815 folhas, conforme indicado pela saída do WEKA.

O desempenho observado no modelo J48 aplicado ao dataset multiclasse reflete limitações amplamente documentadas quando há desbalanceamento entre três ou mais classes. Embora a acurácia global tenha sido elevada (88,845%), a árvore resultante apresentou elevada complexidade estrutural (11.629 nós e 5.815 folhas), indicando que o modelo buscou compensar a dificuldade de separar padrões sobrepostos entre C0, C1 e C2. Segundo Razali et al., (2025) problemas de classificação multiclasse tornam-se substancialmente mais desafiadores quando existe assimetria entre as classes, pois os algoritmos tendem a concentrar o aprendizado nas classes majoritárias, deixando as classes com menor representatividade com desempenho inferior. Esse efeito explica o baixíssimo recall observado na classe intermediária

C1 (0,092), que se mostra mais vulnerável à confusão com classes adjacentes em cenários com fronteiras não lineares. Kumar et al., (2021), reforçam essa interpretação ao demonstrar que a acurácia global é frequentemente enganosa em bases multiclasse desbalanceadas, uma vez que modelos como árvores de decisão favorecem as classes com maior número de exemplos, o que resulta em desempenho severamente reduzido nas classes minoritárias. Os autores destacam ainda que a escassez de exemplos suficientes impede que o modelo aprenda adequadamente as fronteiras de decisão, levando não apenas à queda de sensibilidade para a classe minoritária, mas também à formação de árvores excessivamente profundas e pouco interpretáveis, como tentativa de captar padrões complexos e altamente sobrepostos. Assim, a elevada complexidade da árvore J48 obtida neste estudo não representa robustez do modelo, mas sim uma indicação clara de dificuldade de generalização e tentativa de ajuste aos ruídos e ambiguidades inerentes ao desbalanceamento multiclasse.

Dessa forma, a estudos realizados pelos autores Razali et al. (2025) e Kumar et al. (2022) confirma que o desempenho inferior da classe C1 e a complexidade extrema da árvore resultante são consequências diretas da distribuição desigual entre as classes e da presença de regiões de fronteira difíceis de separar. Esses achados reforçam a necessidade de aplicar métodos de balanceamento como reamostragem, ajuste de pesos ou técnicas híbridas para melhorar o desempenho específico das classes minoritárias e reduzir a complexidade dos modelos em análises subsequentes.

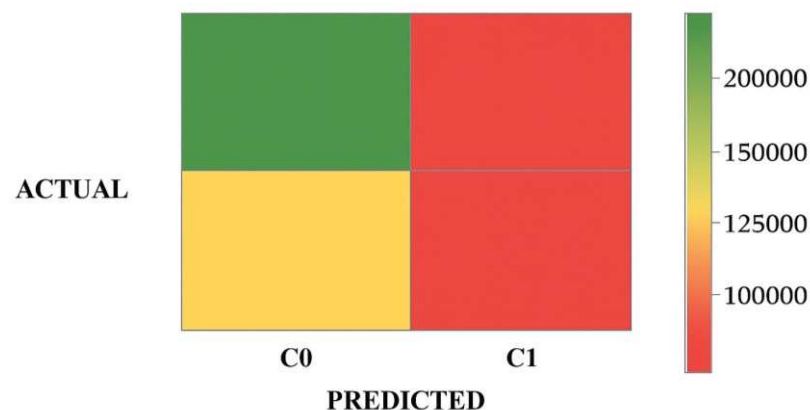
### 5.3 Modelo Binário e Desempenho da Classificação

No dataset binário (3\_diabetes\_binary\_health\_indicators\_BRFSS2015\_2\_classes\_N\_PD), a aplicação do método *10-fold cross-validation* utilizando o classificador J48 resultou em acurácia de 85,613%. As métricas adicionais confirmam o comportamento esperado para bases com forte desbalanceamento, apresentando:

- Acurácia: 85,613%
- Erro médio absoluto (MAE): 0,1974
- Kappa: 0,2215
- Coeficiente de Correlação de Matthews MCC: 0,243
- Recall da classe C0: 0,961
- Recall da classe C1: 0,210

A discrepância entre os recalls evidencia que o modelo possui excelente capacidade de identificar indivíduos sem diabetes/pré-diabetes (C0), enquanto apresenta baixa sensibilidade para a classe C1, classificando corretamente apenas 21% dos indivíduos com alteração glicêmica. Esse padrão é consequência direta da predominância da classe C0 na base, que direciona o processo de aprendizagem do classificador. A Figura 2 ilustra essa assimetria, com maior concentração de acertos em C0 e baixa densidade de classificações corretas em C1, refletindo a dificuldade do modelo em reconhecer adequadamente a classe minoritária.

Figura 4: Heatmap da matriz de confusão do modelo binário (C0 e C1).



C0 corresponde a indivíduos sem diabetes e C1 representa indivíduos com pré-diabetes ou diabetes.

A interpretação desse conjunto de métricas encontra suporte direto na literatura. Chicco; Jurman. (2020), demonstram que a acurácia é uma métrica inadequada para avaliar classificadores treinados em bases desbalanceadas, pois tende a produzir valores artificialmente elevados quando o modelo acerta majoritariamente a classe mais frequente, mesmo apresentando falhas sistemáticas na classe de interesse. Os autores recomendam o uso de métricas mais robustas, como o *Matthews Correlation Coefficient* (MCC), por refletirem simultaneamente o desempenho nas classes positiva e negativa. Segundo os autores, o MCC só atinge valores elevados quando o modelo acerta de forma consistente ambas as classes o que não ocorre em cenários com forte assimetria entre os recalls de C0 e C1, como observado neste estudo. O MCC reduzido (0,243) confirma essa limitação, indicando baixa capacidade discriminativa global apesar da acurácia elevada.

Portanto, os resultados do modelo binário evidenciam que métricas tradicionais de acurácia não são suficientes para avaliar modelos em bases de saúde desbalanceadas e reforçam a importância de empregar métodos de balanceamento ou estratégias de modelagem alternativas

para melhorar a detecção da classe minoritária. Esses achados fundamentam a necessidade de ajustes metodológicos em etapas posteriores, especialmente considerando a relevância clínica da identificação precoce de indivíduos com pré-diabetes e diabetes.

Complementarmente, o estudo conduzido por Chicco; Tötsch; Jurman, (2021), aprofunda essa discussão ao demonstrar que métricas amplamente utilizadas, como *acurácia*, *balanced accuracy*, *F1-score* e *bookmaker informedness*, podem produzir interpretações equivocadas em bases assimétricas, gerando valores elevados mesmo quando o classificador falha de forma consistente na classe minoritária. Os autores mostram que o MCC é a única métrica capaz de refletir simultaneamente as quatro taxas fundamentais da matriz de confusão: sensibilidade, especificidade, valor preditivo positivo e valor preditivo negativo oferecendo uma avaliação mais fidedigna da capacidade discriminativa do modelo. No presente estudo, o MCC de 0,243 confirma essa premissa, indicando baixa correlação entre as predições e os valores reais e reforçando que a *acurácia*, isoladamente, não é adequada para avaliar modelos treinados em bases clínicas desbalanceadas.

Além das limitações já discutidas, a literatura recente também aponta restrições importantes no uso da AUC como métrica primária em bases desbalanceadas. Carrington et al., (2023), destacam que a AUC pode produzir avaliações excessivamente generalistas, pois considera todos os limiares de decisão, incluindo aqueles que não são realistas ou clinicamente úteis. Os autores argumentam que tanto medidas globais, como a própria AUC, quanto métricas pontuais, como *acurácia*, sensibilidade ou valor preditivo positivo, carecem de profundidade para descrever adequadamente o desempenho de classificadores em cenários assimétricos, podendo ocultar falhas graves na detecção da classe minoritária. Essa crítica é coerente com os resultados do presente estudo, no qual o modelo apresentou AUC elevada, mas recall extremamente baixo para C1, indicando que medidas convencionais não capturam de maneira completa o impacto do desbalanceamento na capacidade discriminativa real do classificador.

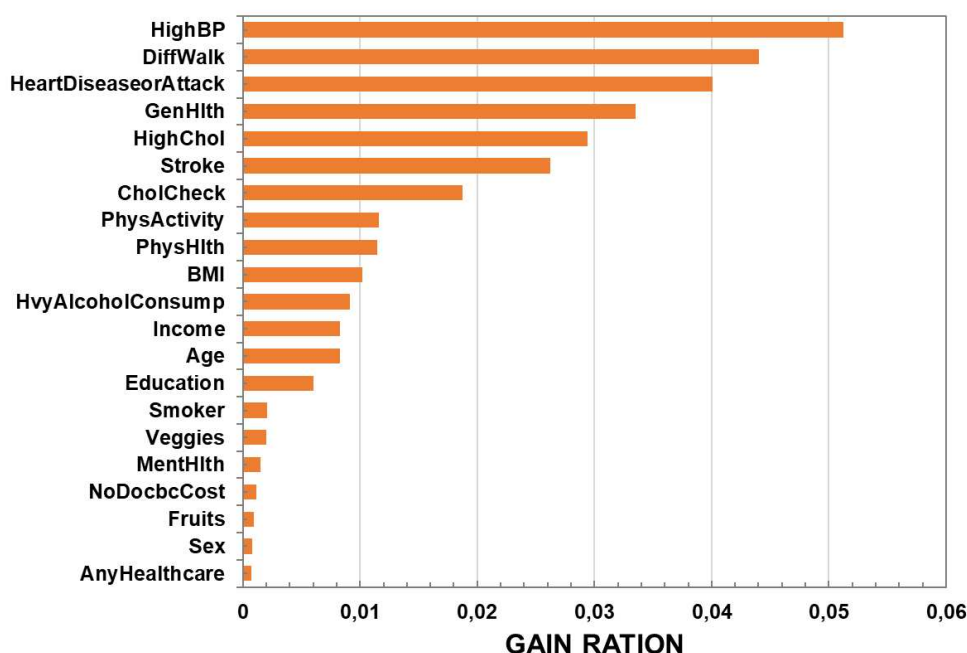
#### **5.4 Seleção de Atributos e Interpretação Clínica**

Para reduzir a complexidade do modelo e aprimorar sua interpretabilidade, aplicou-se o método *GainRatioAttributeEval* associado ao algoritmo de busca *Ranker*, resultando no ranqueamento apresentado na Figura 3. Os atributos com maior contribuição para a separação entre as classes foram *HighBP*, *DiffWalk*, *HeartDiseaseorAttack*, *GenHlth*, *HighChol* e *Stroke*, que exibiram os maiores valores de ganho de informação.

- *HighBP* e *HighChol*: componentes centrais da síndrome metabólica, frequentemente associados ao desenvolvimento de resistência à insulina e diabetes tipo 2.
- *Heart Diseaseor Attack* e *Stroke*: condições cardiovasculares intimamente relacionadas ao descontrole glicêmico e ao processo inflamatório crônico característico do diabetes.
- *GenHlth*: medida subjetiva de saúde geral, amplamente validada como indicador epidemiológico robusto e preditor independente de risco metabólico.
- *DiffWalk*: marcador funcional de limitação motora, comum em indivíduos com complicações musculoesqueléticas ou neuropáticas associadas ao diabetes.

Assim, a seleção automática de atributos resultou em um conjunto clinicamente plausível, reforçando a consistência fisiopatológica das variáveis que melhor discriminam os indivíduos com e sem risco glicêmico.

Figura 5 - Ranking dos atributos selecionados pelo método *GainRatioAttributeEval* + *Ranker* no dataset binário, evidenciando a importância relativa das variáveis para a predição entre C0 (ausência de diabetes/pré-diabetes) e C1 (presença de diabetes/pré-diabetes).



A interpretação clínica dos atributos selecionados é coerente não apenas com os mecanismos fisiopatológicos bem estabelecidos, mas também com achados de estudos recentes que empregaram abordagens de engenharia de atributos para predição de diabetes tipo 2. Ganie; Malik, (2022), reforçam que indicadores metabólicos, cardiovasculares e funcionais são

consistentemente identificados como preditores centrais em modelos de aprendizado de máquina, destacando que variáveis relacionadas ao estado geral de saúde, sintomas metabólicos e fatores de risco tradicionais apresentam maior importância relativa em processos de seleção de atributos.

Segundo os autores, a engenharia de atributos revela que fatores como histórico familiar, fadiga, sede excessiva, idade, peso e comportamentos de risco aparecem repetidamente entre os principais preditores do desfecho glicêmico padrão que converge com os atributos selecionados neste estudo, especialmente *HighBP*, *HighChol*, *HeartDiseaseorAttack* e *Stroke*, todos amplamente reconhecidos como componentes de risco cardiovascular associados ao desenvolvimento e progressão do diabetes. Assim, a seleção automática realizada pelo método *GainRatioAttributeEval* resultou em um conjunto de variáveis clinicamente consistente, alinhado tanto com a literatura epidemiológica quanto com evidências obtidas em estudos de modelagem baseados em dados reais de saúde.

Os achados da seleção de atributos neste estudo também convergem com resultados de investigações recentes que utilizaram modelos de árvore para identificar os preditores clínicos mais relevantes para o diabetes. Sadiq et al., (2025), demonstraram que, em modelos de decisão aplicados a dados bioquímicos e demográficos, os atributos mais importantes para a predição de diabetes foram HbA1c, índice de massa corporal (BMI) e idade, seguidos por marcadores lipídicos como colesterol total, triglicerídeos e VLDL. Os autores também observaram diferenças significativas nesses parâmetros entre indivíduos diabéticos e não diabéticos, destacando o papel de fatores cardiometabólicos semelhantes aos identificados neste trabalho, como pressão arterial elevada e disfunções cardiovasculares. Além disso, a análise de componentes principais apresentada no estudo agrupou variáveis metabólicas e cardiovasculares em clusters fisiologicamente coerentes, reforçando que preditores clínicos relacionados ao metabolismo, inflamação e risco cardiovascular tendem a emergir com maior importância em abordagens de seleção de atributos. Portanto, a lista final de atributos selecionados pelo *GainRatio* neste estudo apresenta elevada plausibilidade clínica e é consistente com o que a literatura recente descreve como os principais determinantes do risco glicêmico.

Estudos realizados por Olisah; Smith; Smith, (2022), demonstraram que, em modelos aplicados aos conjuntos PIMA e LMCH, os atributos com maior relevância estatística para a predição de diabetes foram glicemia, pressão arterial, insulina e idade todos fortemente associados ao risco cardiometabólico. Esses marcadores metabólicos compõem o mesmo eixo

fisiopatológico dos atributos identificados no presente estudo, como *HighBP*, *HighChol*, *HeartDiseaseorAttack* e *Stroke*, reforçando que variáveis relacionadas ao estado inflamatório crônico e ao sistema cardiovascular emergem consistentemente como preditores centrais na classificação do risco glicêmico. Os autores destacam ainda que a etapa de seleção de atributos melhora significativamente o desempenho de classificadores tradicionais, aumentando a acurácia de modelos como Random Forest, SVM e DNN após a redução do espaço de atributos. Dessa forma, a seleção automática obtida neste estudo não apenas reduz a complexidade dos modelos, mas também resulta em um conjunto de atributos clinicamente plausível e alinhado aos principais determinantes epidemiológicos do diabetes descritos na literatura recente.

### 5.5 Árvore de Decisão com Atributos Seleccionados

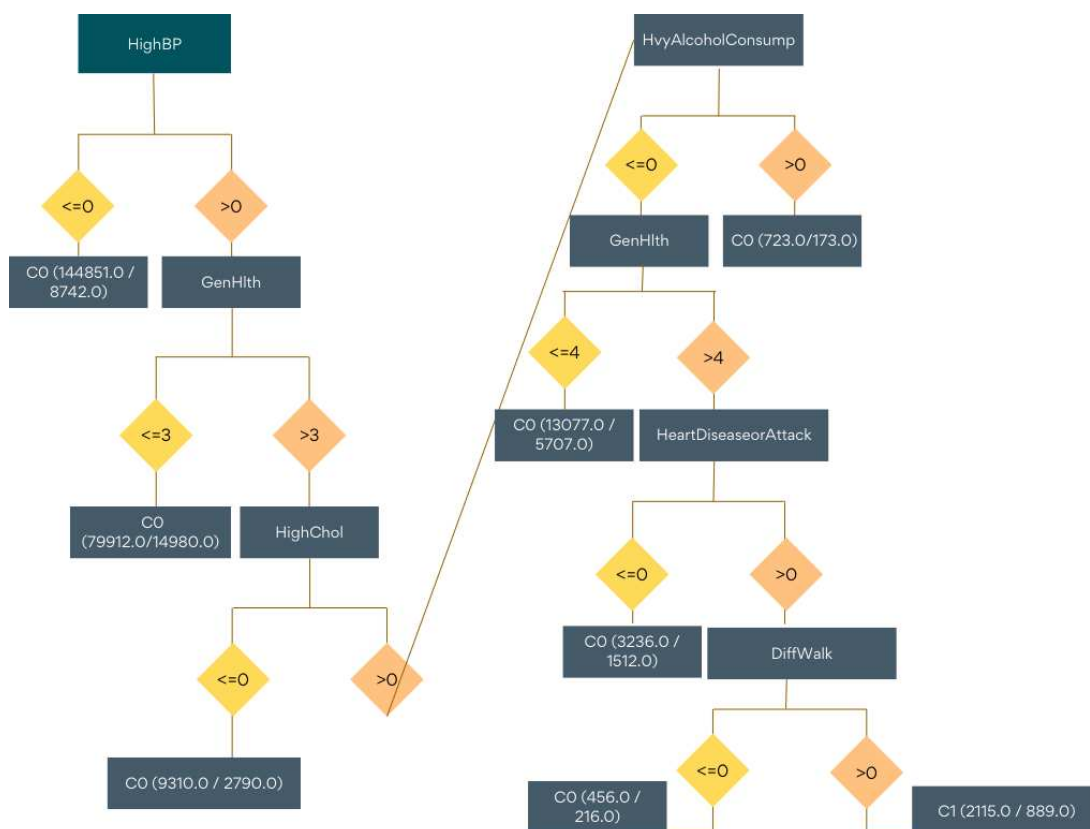
Ao treinar o modelo J48 utilizando apenas os seis atributos seleccionados, observou-se uma redução substancial na complexidade da árvore, tornando-a mais interpretável sem perda significativa de desempenho. A acurácia permaneceu próxima de 83%, indicando que os atributos seleccionados concentravam a maior parte da capacidade preditiva do modelo.

A nova estrutura da árvore (Figura 6) inicia com o atributo *HighBP*, sugerindo que a presença de hipertensão arterial é o fator mais discriminante entre indivíduos com e sem alteração glicêmica. Na sequência, surgem divisões baseadas em *GenHlth*, *HighChol* e *HvyAlcoholConsump*, compondo caminhos de decisão que refletem relações clínicas amplamente descritas em estudos epidemiológicos sobre risco cardiometabólico.

Por outro lado, atributos como *DiffWalk* e *HeartDiseaseorAttack* aparecem em ramificações específicas, indicando que limitações funcionais e histórico de doença cardiovascular contribuem para a identificação de indivíduos com maior probabilidade de pertencer à classe C1.

De modo geral, a árvore reduzida demonstra que o modelo aprende padrões clinicamente plausíveis, reproduzindo a hierarquia de risco metabólico frequentemente observada em estudos populacionais: fatores hemodinâmicos e metabólicos iniciais (pressão arterial e perfil lipídico), seguidos por marcadores de saúde percebida e comorbidades cardiovasculares.

Figura 6: Árvore de decisão gerada pelo algoritmo J48 utilizando os seis atributos seleccionados (*HighBP*, *GenHlth*, *HighChol*, *HvyAlcoholConsump*, *HeartDiseaseorAttack* e *DiffWalk*) no dataset binário



A estrutura reduzida da árvore observada neste estudo também encontra respaldo em achados recentes da literatura. Sadiq et al. (2025) avaliaram modelos de Árvore de Decisão aplicados a indicadores bioquímicos e demográficos e observaram que tais modelos tendem a organizar seus nós de forma coerente com a hierarquia fisiopatológica dos fatores de risco metabólico. No estudo, variáveis como BMI, colesterol, triglicerídeos e especialmente HbA1c foram priorizadas nos primeiros níveis da árvore, refletindo sua relevância clínica na distinção entre indivíduos diabéticos e não diabéticos. Essa organização é semelhante ao padrão encontrado na árvore reduzida deste trabalho, que inicia com *HighBP* e segue com *GenHlth* e *HighChol*, compondo caminhos de decisão que refletem risco cardiometabólico. Os autores demonstraram que, mesmo com estruturas mais compactas, os modelos mantêm elevado poder discriminativo resultado atribuído à capacidade das árvores de incorporar relações clínicas reais entre preditores metabólicos e cardiovasculares. Assim, os achados presentes reforçam que a redução de atributos não compromete o desempenho e ainda melhora a interpretabilidade, mantendo a coerência com o comportamento observado em estudos clínicos baseados em

## Árvores de Decisão.

A consistência da árvore reduzida observada neste estudo também é sustentada por evidências recentes da literatura. (Sadhasivam et al., 2021), demonstraram que a combinação entre técnicas de seleção de atributos e algoritmos de Árvore de Decisão reduz substancialmente a complexidade do modelo, ao mesmo tempo em que preserva ou até aprimora o desempenho preditivo. Os autores mostram que métodos como RFE, Boruta e algoritmos genéticos são capazes de identificar o subconjunto de variáveis mais informativo, resultando em modelos mais rápidos, interpretáveis e clinicamente úteis, característica especialmente vantajosa em aplicações médicas. No estudo, a seleção prévia de atributos reduziu dimensionalidade sem comprometer a acurácia do classificador, revelando árvores que organizam seus nós de decisão a partir de marcadores metabólicos e demográficos relevantes. Esse comportamento é semelhante ao observado na árvore J48 simplificada deste trabalho, que priorizou atributos fisiologicamente plausíveis como *HighBP*, *GenHlth* e *HighChol*, indicando que a redução da complexidade não prejudica a capacidade discriminativa do modelo e reforça sua interpretabilidade clínica.

Evidências adicionais da literatura reforçam o comportamento observado na árvore reduzida obtida neste estudo. Seto et al., (2022), demonstraram que modelos baseados em árvores, como *Gradient Boosting Decision Trees*, apresentam desempenho superior a métodos lineares em bases clínicas amplas devido à capacidade de capturar relações não lineares entre fatores cardiometabólicos. No estudo, variáveis como BMI, pressão sistólica, triglicerídeos, HDL-C, LDL-C e HbA1c apresentaram padrões distintos entre indivíduos diabéticos e não diabéticos, sendo organizadas de forma hierárquica pelos modelos de árvore, comportamento semelhante ao observado no J48 reduzido deste trabalho, que priorizou *HighBP*, *GenHlth* e *HighChol* nos níveis iniciais. Além disso, os autores evidenciaram que modelos baseados em árvores se tornam particularmente eficazes quando aplicados a grandes amostras (acima de 10.000 indivíduos), cenário semelhante ao conjunto utilizado nesta análise. Dessa forma, o desempenho e a estrutura da árvore reduzida obtida neste estudo são compatíveis com o que tem sido documentado para modelos de decisão aplicados a grandes bases de dados de saúde. Em conjunto, esses resultados sustentam as considerações apresentadas na conclusão deste trabalho.

## 6. CONCLUSÃO

Os resultados deste estudo demonstram que, no contexto de bases clínicas desbalanceadas, nem sempre as árvores de decisão com maior acurácia são as mais adequadas para fins de interpretação clínica ou detecção de classes minoritárias. A utilização do algoritmo J48 mostrou-se uma abordagem eficiente para a classificação do risco de Diabetes Mellitus tipo 2 a partir dos indicadores do BRFSS, especialmente quando combinada com validação cruzada e técnicas de seleção de atributos. Esses procedimentos permitiram a construção de modelos preditivos com desempenho satisfatório, ao mesmo tempo em que aprimoraram a capacidade de generalização e reduziram a complexidade desnecessária das estruturas geradas.

A identificação de fatores fortemente associados ao risco glicêmico como hipertensão arterial, colesterol elevado, limitações funcionais e histórico de doença cardiovascular reforça a relevância de avaliar esses indicadores de forma integrada na prática clínica, nutricional e epidemiológica. A redução da árvore para um conjunto enxuto de atributos clinicamente plausíveis confirmou que a interpretabilidade pode ser preservada sem prejuízo substancial ao desempenho, aspecto fundamental para a futura utilização de modelos de apoio à decisão em ambientes de saúde.

Considerando as limitações inerentes ao desbalanceamento da base, recomenda-se que estudos futuros explorem técnicas adicionais de IA, como métodos ensemble, estratégias de balanceamento avançadas e abordagens híbridas que integrem aprendizado supervisionado, conhecimento clínico e análises de risco. A adoção dessas técnicas pode aprimorar a sensibilidade para a classe minoritária e viabilizar a implementação desses modelos em sistemas reais de triagem e acompanhamento de pacientes.

Dessa forma, a combinação de mineração de dados com conhecimento clínico apresenta grande potencial para fortalecer estratégias de prevenção e controle do Diabetes Mellitus, permitindo intervenções mais precoces e custo-efetivas. Modelos preditivos bem calibrados poderão fornecer estimativas confiáveis de risco glicêmico, especialmente em indivíduos assintomáticos, contribuindo para a identificação antecipada de populações vulneráveis e para o desenvolvimento de políticas públicas e práticas clínicas mais assertivas.

## REFERÊNCIAS

ABOUSABER, Inam; ABDALLAH, Haitham F.; EL-GHAISH, Hany. Robust predictive framework for diabetes classification using optimized machine learning on imbalanced datasets. *Frontiers in Artificial Intelligence*, v. 7, 2024.

AFSANEH, Elaheh *et al.* Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review. *Diabetology and Metabolic SyndromeBioMed Central Ltd*, , 1 dez. 2022.

AGLIATA, Antonio *et al.* Machine Learning as a Support for the Diagnosis of Type 2 Diabetes. *International Journal of Molecular Sciences*, v. 24, n. 7, 1 abr. 2023.

AGYEMANG, Edmund Fosu *et al.* Addressing Class Imbalance Problem in Health Data Classification: Practical Application From an Oversampling Viewpoint. *Applied Computational Intelligence and Soft Computing*, v. 2025, n. 1, 26 jan. 2025.

AHADZADEH, Behrouz *et al.* SFE: A Simple, Fast, and Efficient Feature Selection Algorithm for High-Dimensional Data. *IEEE Transactions on Evolutionary Computation*, v. 27, n. 6, p. 1896–1911, 1 dez. 2023.

ALOMEIR, Othman *et al.* Prevalence and impact of sociodemographic factors, comorbidities, and lifestyle on diabetes complications among patients with type 2 diabetes in Riyadh. *Scientific Reports*, v. 15, n. 1, 1 dez. 2025.

APARICIO, Luis Fregoso *et al.* Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. *Diabetology and Metabolic SyndromeBioMed Central Ltd*, , 1 dez. 2021.

BASAK, Mrinmoy; LASKAR, Moksood Ahmed. PATHOPHYSIOLOGY, LIFESTYLE INTERVENTION AND COMPLICATIONS OF TYPE-2 DIABETES: A REVIEW. *Journal of Applied Pharmaceutical ResearchCreative Pharma Assent*, , 30 jun. 2024.

BLUMBERG, J. Nutritional needs of seniors. *J Am Coll Nutr*, 1997.

BOMMERT, Andrea *et al.* Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics and Data Analysis*, v. 143, 1 mar. 2020.

CARRINGTON, Andre M. *et al.* Deep ROC Analysis and AUC as Balanced Average Accuracy, for Improved Classifier Selection, Audit and Explanation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 45, n. 1, p. 329–341, 1 jan. 2023.

CHICCO, Davide; JURMAN, Giuseppe. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, v. 21, n. 1, 2 jan. 2020.

CHICCO, Davide; TÖTSCH, Niklas; JURMAN, Giuseppe. The matthews correlation coefficient (Mcc) is more reliable than balanced accuracy, bookmaker informedness, and

markedness in two-class confusion matrix evaluation. *BioData Mining*, v. 14, p. 1–22, 2021.

DOKEROGLU, Tansel; DENIZ, Ayça; KIZILOZ, Hakan Ezgi. A comprehensive survey on recent metaheuristics for feature selection. *Neurocomputing* Elsevier B.V., , 14 jul. 2022.

DRITSAS, Elias; TRIGKA, Maria. Data-Driven Machine-Learning Methods for Diabetes Risk Prediction. *Sensors*, v. 22, n. 14, 1 jul. 2022.

DUONG, Khanh N. C. *et al.* Comparison of diagnostic accuracy for diabetes diagnosis: A systematic review and network meta-analysis. *Frontiers in Medicine* Frontiers Media S.A., , 24 jan. 2023.

ESTEVA, Andre *et al.* A guide to deep learning in healthcare. *Nature Medicine* Nature Publishing Group, , 1 jan. 2019.

GANIE, Shahid Mohammad; MALIK, Majid Bashir. An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators. *Healthcare Analytics*, v. 2, 1 nov. 2022.

GIEROBA, Barbara; KRYSKA, Adrianna; SROKA-BARTNICKA, Anna. Type 2 diabetes mellitus – conventional therapies and future perspectives in innovative treatment. *Biochemistry and Biophysics Reports* Elsevier B.V., , 1 jun. 2025.

GOVERNO FEDERAL DO BRASIL. Cenário das doenças crônicas não transmissíveis. 2021.

GOWDAK, Marcia Maria Godoy. Diretriz da Sociedade Brasileira de Diabetes. SOCESP, 2025.

HU, Shengying *et al.* Risk factors for progression to type 2 diabetes in prediabetes: a systematic review and meta-analysis. *BMC Public Health*, v. 25, n. 1, p. 1220, 31 mar. 2025.

INTERNATIONAL DIABETES FEDERATION. The Diabetes Atlas. Disponível em: <<https://diabetesatlas.org/>>. Acesso em: 18 nov. 2025a.

INTERNATIONAL DIABETES FEDERATION. South and Central America. 2025b.

KUMAR, Mukesh *et al.* A Comparative Performance Assessment of Optimized Multilevel Ensemble Learning Model with Existing Classifier Models. *Big Data*, v. 10, n. 5, p. 371–387, 2021.

KUMAR, Vinod *et al.* Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques. *Healthcare*, v. 10, n. 7, p. 1293, 13 jul. 2022.

LI, Li Mei; JIANG, Bei Ge; SUN, Liang Liang. HNF1A : From Monogenic Diabetes to Type 2 Diabetes and Gestational Diabetes Mellitus. *Frontiers in Endocrinology*, v. 13, 1 mar. 2022.

MA, Shizhan *et al.* Artificial intelligence and medical-engineering integration in diabetes management: Advances, opportunities, and challenges. *Healthcare and Rehabilitation*, v. 1, n.

1, p. 100006, jan. 2025.

MAHESHWARI, Shivangi *et al.* Artificial intelligence for diabetes management – a review. *Journal of Diabetes, Metabolic Disorders & Control*, v. 12, n. 1, p. 24–32, 2025.

MINISTÉRIO DA SAÚDE. Vigilância de fatores de risco e proteção para doenças crônicas por inquérito telefônico. Brasília: [S.n.].

MIRZAZADA, V. A. *et al.* Diabetes Mellitus: Once More About Diagnostics. *International Journal of Diabetes & Metabolic Syndrome*, 2021.

MOHANTY, Cheena *et al.* Using Deep Learning Architectures for Detection and Classification of Diabetic Retinopathy. *Sensors*, v. 23, n. 12, p. 5726, 19 jun. 2023.

MOON, Joon Ho; JANG, Hak Chul. Gestational Diabetes Mellitus: Diagnostic Approaches and Maternal-Offspring Complications. *Diabetes and Metabolism Journal* Korean Diabetes Association, , 1 jan. 2022.

MOSTAFA, Reham R. *et al.* ST-AL: a hybridized search based metaheuristic computational algorithm towards optimization of high dimensional industrial datasets. *Soft Computing*, v. 27, n. 18, p. 13553–13581, 1 set. 2023.

NOROOZI, Zeinab; OROOJI, Azam; ERFANNIA, Leila. Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. *Scientific Reports*, v. 13, n. 1, 1 dez. 2023.

OLISAH, Chollette C.; SMITH, Lyndon; SMITH, Melvyn. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Computer Methods and Programs in Biomedicine*, v. 220, 1 jun. 2022.

PONNABOINA, Abhishek; VEERASWAMY, D. Diabetes Prediction and Recommendations Using Machine Learning and Generative AI. *International Journal of Scientific Research in Engineering and Management*, 2025.

RAZALI, Muhammad Nazim *et al.* Optimizing Multiclass Classification Using Convolutional Neural Networks with Class Weights and Early Stopping for Imbalanced Datasets. *Electronics (Switzerland)*, v. 14, n. 4, 1 fev. 2025.

RODAKCI, Melanie *et al.* Diagnóstico de diabetes mellitus. *In: Diretriz da Sociedade Brasileira de Diabetes*. [S.l.]: Conectando Pessoas, 2024.

SADHASIVAM, Jayakumar *et al.* Diabetes disease prediction using decision tree for feature selection. *In: IOP Publishing Ltd*, 23 jul. 2021.

SADIQ, Idris Zubairu *et al.* Data-driven diabetes mellitus prediction and management: a comparative evaluation of decision tree classifier and artificial neural network models along with statistical analysis. *Scientific Reports*, v. 15, n. 1, 1 dez. 2025.

SBEM. O que é Diabetes? Disponível em: <<https://www.endocrino.org.br/o-que-e-diabetes/>>. Acesso em: 24 nov. 2025.

SETO, Hiroe *et al.* Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. *Scientific Reports*, v. 12, n. 1, 1 dez. 2022.

SINGH, Raj; GHOLIPOURMALEKABADI, Mazaher; SHAFIKHANI, Sasha H. Animal models for type 1 and type 2 diabetes: advantages and limitations. *Frontiers in Endocrinology*Frontiers Media SA, , 2024.

SISTEMA DE INFORMAÇÕES SOBRE MORTALIDADE. Painel de monitoramento da mortalidade prematura (30 a 69 anos) por DCNT. Disponível em: <<https://svs.aids.gov.br/daent/centrais-de-conteudos/paineis-de-monitoramento/mortalidade/dcnt/>>. Acesso em: 18 nov. 2025.

SUSHITH, Mishmala *et al.* A hybrid deep learning framework for early detection of diabetic retinopathy using retinal fundus images. *Scientific Reports*, v. 15, n. 1, p. 15166, 30 abr. 2025.

VYBHAVI, V. S. Jyothi *et al.* New Insights into the Pathophysiology of Type 2 Diabetes: A Review Article. *Journal of Pharmacy and Bioallied Sciences*, v. 17, n. Suppl 2, p. S1070–S1072, jun. 2025.

WHO. Noncommunicable diseases. Disponível em: <<https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases>>. Acesso em: 18 nov. 2025.

WU, Shiyong. Multi-factors correlation to diabetes using Machine Learning: findings from BRFSS. *Transactions on Materials, Biotechnology and Life Sciences*, 2024.

WU, Wen Tao *et al.* Data mining in clinical big data: the frequently used databases, steps, and methodological models. *Military Medical Research*BioMed Central Ltd, , 1 dez. 2021.

YAMENY, Ahmed Abdelhalim. Diabetes Mellitus Overview 2024. *Journal of Bioscience and Applied Research*Society of Pathological Biochemistry and Hematology, , 1 set. 2024.

YANG, Qianhui *et al.* Use of artificial intelligence with retinal imaging in screening for diabetes-associated complications: systematic review. *eClinicalMedicine*, v. 81, p. 103089, mar. 2025.

ZHANG, Wangyouchen *et al.* Enhancing diabetes risk prediction through focal active learning and machine learning models. *PLOS ONE*, v. 20, n. 7 July, 1 jul. 2025.

ZOYSA, Warsha De *et al.* Behavioral and demographic associations of optimal glycemic control among patients with type 2 diabetes mellitus in Sri Lanka: a multicenter study. *BMC Endocrine Disorders*, v. 25, n. 1, 1 dez. 2025.