Relatório - Segunda Avaliação - Aprendizado de Máquina

David Ian[1], **Jeremias Oliveira**[1]

[1]CESAR School
Caixa Postal 77 – 50.030-220 – Recife – PE – Brazil

dipp@cesar.school, jos@cesar.school

Abstract. *This study investigates the sociodemographic and clinical factors associated with diabetes, drawing upon the analysis presented in the article "Comparative Effectiveness of Classification Algorithms in Predicting Diabetes". The referenced article aimed to compare the performance of six machine learning classification algorithms — K-Nearest Neighbors (K-NN), Naive Bayes, Support Vector Machine (SVM), Decision Trees, Random Forest, and Logistic Regression — in predicting diabetes diagnoses. Using the same dataset adopted in the original study, this work presents complementary analyses to explore relevant sociodemographic patterns and formulate hypotheses that may support the development of improved strategies for diabetes detection and prevention.*

Resumo. *Este estudo investiga os fatores sociodemográficos e clínicos associados ao diabetes, com base na análise do artigo "Comparative Effectiveness of Classification Algorithms in Predicting Diabetes". O artigo em questão teve como objetivo principal comparar o desempenho de seis algoritmos de classificação de aprendizado de máquina — K-Nearest Neighbors (K-NN), Naive Bayes, Support Vector Machine (SVM), Árvores de Decisão, Random Forest e Regressão Logística — na predição do diagnóstico de diabetes. Com base no mesmo conjunto de dados utilizado pelo estudo original, este trabalho propõe análises complementares que visam justificar padrões sociodemográficos observados, além de levantar hipóteses que possam contribuir para o aprimoramento das estratégias de detecção e prevenção da doença.*

1. Introdução

O diabetes mellitus representa um dos maiores desafios de saúde pública global da atualidade. Estima-se que, em 2024, cerca de 589 milhões de adultos entre 20 e 79 anos convivam com a doença, o equivalente a 1 em cada 9 adultos no mundo, segundo o IDF Diabetes Atlas (International Diabetes Federation 2024). Em 2021, a enfermidade foi responsável por aproximadamente 6,7 milhões de mortes, evidenciando não apenas sua alta prevalência, mas também seu impacto direto na mortalidade global (Laboratório Lustosa 2022).

O Brasil ocupa atualmente a quinta posição no ranking mundial de incidência de diabetes, com 16,8 milhões de adultos diagnosticados, ficando atrás apenas da China, Índia, Estados Unidos e Paquistão (Biblioteca Virtual em Saúde MS 2021; Agência Brasil 2023). Dados da pesquisa Vigitel Brasil mostram que a prevalência da doença entre adultos brasileiros alcançou 10,2% em 2023, um crescimento considerável

em relação aos 9,1% registrados em 2021 (Agência Brasil 2023), refletindo uma tendência preocupante de aumento contínuo.

Diante desse cenário, torna-se cada vez mais urgente o desenvolvimento de estratégias eficazes para detecção precoce, prevenção e manejo do diabetes. O uso de técnicas de aprendizado de máquina tem se mostrado promissor nesse campo, permitindo identificar padrões em grandes volumes de dados e auxiliar na construção de modelos preditivos robustos. Este estudo, nesse contexto, propõe uma análise focada nos fatores sociodemográficos e clínicos associados ao diabetes, utilizando como base um conjunto de dados amplamente utilizado na literatura e inspirado pelo trabalho de comparação de algoritmos de classificação.

2. Descrição detalhada do Dataset

O conjunto de dados fornecido parece estar relacionado ao diabetes e contém diversas medições biomédicas e características dos pacientes.

Número de registros: 1000 amostras após pré-processamento.

Número de variáveis: 14, incluindo:

1. ID: Identificador único de cada entrada presente no banco de dados.
2. No_Pation: Número de prontuário ou identificação alternativa do paciente.
3. Gender: Gênero biológico do paciente, podendo ser masculino (M) ou feminino (F).
4. AGE: Idade do paciente, expressa em anos completos.
5. Urea: Nível de ureia no sangue, indicado em mg/dL ou mmol/L. Esse parâmetro está associado à função renal e ao metabolismo de proteínas.
6. Cr (Creatinina): Concentração de creatinina no sangue, também relacionada à função dos rins.
7. HbA1c: Valor de hemoglobina glicada, expressa em porcentagem, que reflete os níveis médios de glicose no sangue nos últimos dois a três meses.
8. Chol (Colesterol Total): Quantidade total de colesterol no sangue.
9. TG (Triglicerídeos): Medida da concentração de triglicerídeos, que são um tipo de gordura circulante.
10. HDL: Níveis de lipoproteína de alta densidade, conhecida como colesterol “bom”.
11. LDL: Níveis de lipoproteína de baixa densidade, popularmente chamada de colesterol “ruim”.
12. VLDL: Valores de lipoproteína de densidade muito baixa, relacionados ao transporte de lipídios no sangue.
13. BMI (Índice de Massa Corporal): Relação entre peso e altura do paciente (kg/m²), utilizada como indicador de sobrepeso e obesidade.
14. CLASS: Classe diagnóstica atribuída ao paciente, sendo dividida em três categorias:
    N (Não diabético)
    P (Pré-diabético)
    Y (Diabético)

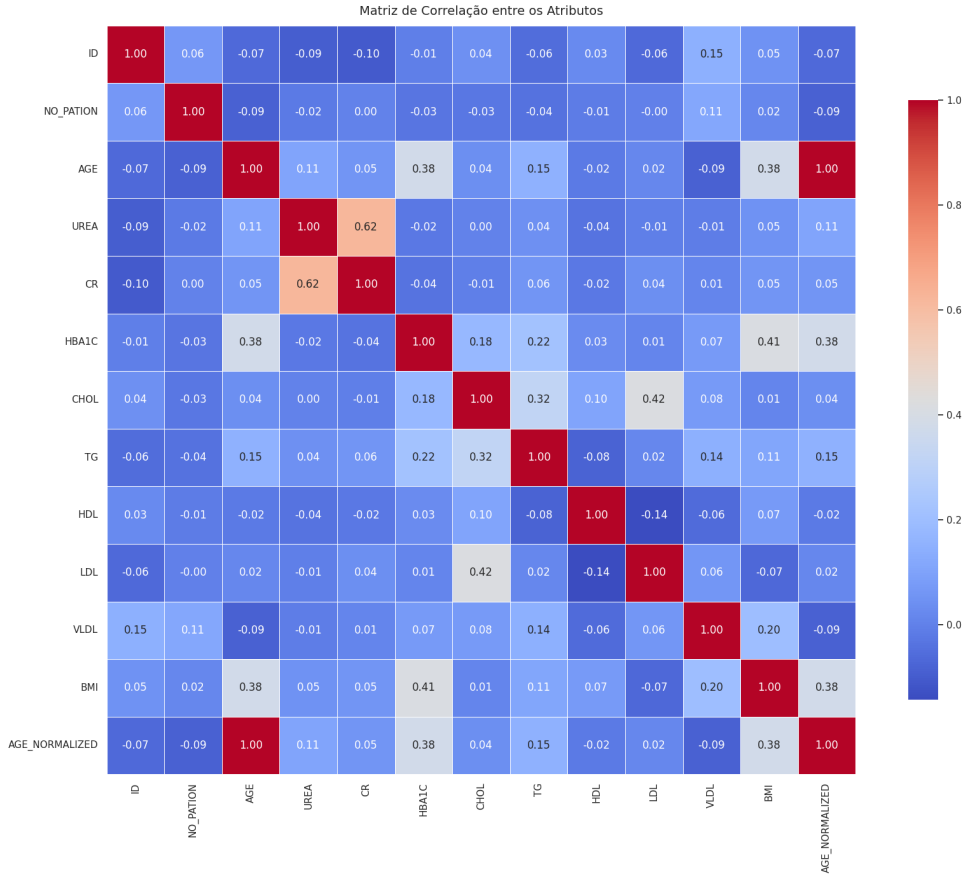

Figura 1. Matriz de Correlação entre os Atributos

2.1. Visualização de Correlações

A matriz de correlação apresentada acima exibe a relação linear entre as variáveis numéricas do dataset, com base no coeficiente de correlação de Pearson. Os valores variam entre -1 e 1, onde: +1 indica correlação perfeitamente positiva;

0 indica ausência de correlação linear;

-1 indica correlação perfeitamente negativa.

As cores foram representadas no gradiente de azul (negativa) a vermelho (positiva), com anotações numéricas para facilitar a interpretação. Principais observações: Urea e Creatinina (CR) apresentaram uma correlação fortemente positiva de 0.62, o que é esperado, visto que ambas são substâncias excretadas pelos rins e frequentemente utilizadas como marcadores de função renal.

HbA1c e BMI mostraram uma correlação de 0.41, o que pode indicar que indivíduos com maior índice de massa corporal tendem a apresentar piores níveis de controle glicêmico.

AGE e HbA1c também tiveram uma correlação moderada (0.38), sugerindo que a idade pode influenciar no acúmulo de glicose no sangue ao longo do tempo.

As variáveis CHOL (colesterol total), TG (triglicerídeos) e LDL apresentaram correlações entre si de forma esperada, como por exemplo:

CHOL e LDL: 0.42

CHOL e TG: 0.32

LDL e BMI: 0.20

Essas relações reforçam o vínculo fisiológico entre perfil lipídico e composição corporal. Correlação da variável AGE_NORMALIZED Como esperado, AGE_NORMALIZED possui correlação perfeita com AGE (1.00) e valores idênticos com as mesmas variáveis correlacionadas, apenas transformadas por uma escala estatística padrão (StandardScaler). Isso comprova a integridade da transformação.

2.2. Distribuição de Classes

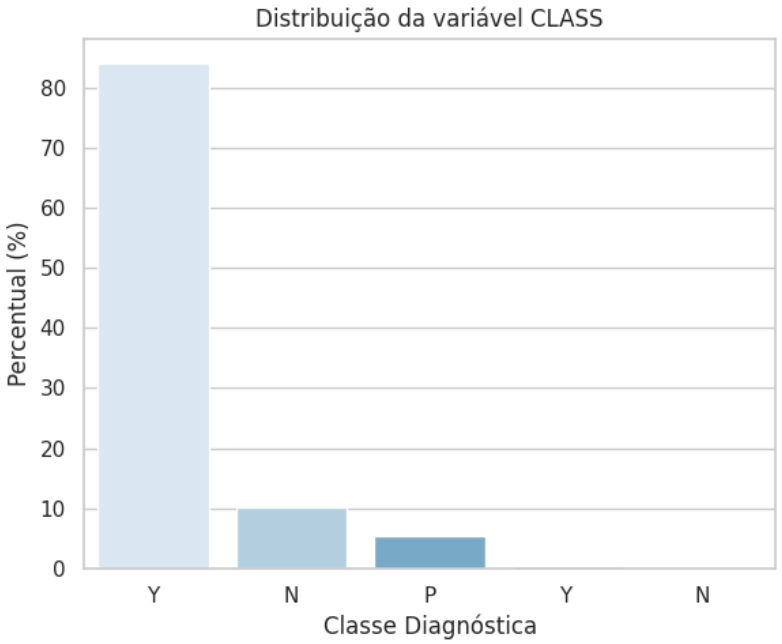

Figura 2. Distribuição da variável CLASS

A variável CLASS é a variável alvo (target) do dataset e representa o diagnóstico do paciente em relação ao diabetes. Ela é categórica e está dividida em três possíveis classes: N: Paciente não diabético

P: Paciente pré-diabético

Y: Paciente diabético

Observações da distribuição: A classe Y (diabético) representa a grande maioria dos registros, com aproximadamente 84% dos dados.

A classe N (não diabético) aparece em torno de 10% dos casos.

A classe P (pré-diabético) é a menos representada, com apenas cerca de 6

Implicações: Trata-se de um conjunto de dados altamente desbalanceado, o que pode comprometer a performance de algoritmos de classificação, principalmente em relação às classes minoritárias (P e N).

Algoritmos sem tratamento prévio podem favorecer a classe majoritária (Y), resultando em alta acurácia, mas baixa sensibilidade para detectar pré-diabetes ou ausência da doença.

Será necessário aplicar técnicas de balanceamento, como:

Oversampling (ex: SMOTE) das classes minoritárias;

Undersampling da classe majoritária;

Ou utilizar algoritmos robustos a desbalanceamento (ex: XGBoost, Random Forest com class_weight='balanced').

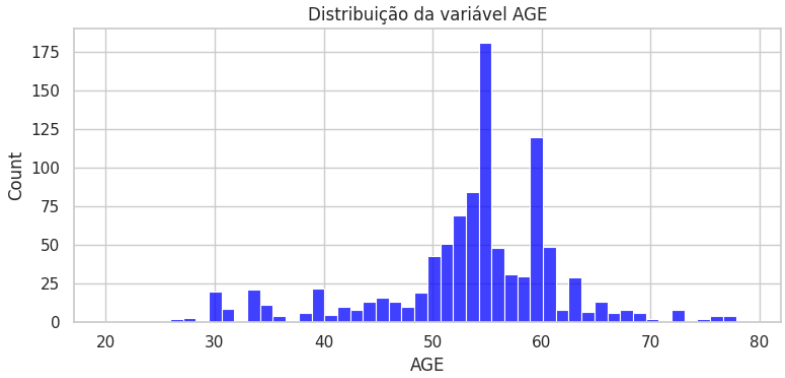

Figura 3. Distribuição da variável AGE

O gráfico acima apresenta a distribuição da idade dos pacientes registrados no dataset. A variável AGE é uma variável numérica contínua que representa a idade dos indivíduos em anos completos. Análise da Distribuição: Observa-se uma concentração significativa de pacientes entre os 50 e 60 anos, com pico próximo dos 55 anos, indicando uma predominância de indivíduos de meia-idade no conjunto de dados.

Há uma distribuição assimétrica à esquerda (viés à direita), com poucos registros abaixo de 40 anos e acima de 65 anos.

A presença de picos acentuados em faixas etárias específicas (ex: 55 e 60 anos) pode indicar agrupamentos artificiais ou erros de arredondamento nos dados originais (ex: muitos pacientes com "idade registrada" como 55).

Indivíduos com menos de 30 anos e mais de 70 anos representam uma minoria no dataset.

Implicações para a análise: A predominância de pacientes mais velhos pode ter impacto na interpretação dos dados, já que a idade é um fator de risco importante para o diabetes.

Essa distribuição deve ser considerada na avaliação de correlações com outras variáveis, especialmente glicemia, HbA1c, função renal e diagnóstico (variável CLASS).

Pode ser útil normalizar ou padronizar a variável AGE, especialmente para modelos sensíveis à escala.

3. Análise exploratória e Pré-processamento

3.1. Pré-processamento

- **Valores Ausentes:** Não foram detectados valores ausentes em nenhuma das colunas dos datasets.
- **Valores Duplicados:** Não foram detectados valores duplicados em nenhuma das colunas dos datasets

3.1.1. Normalização

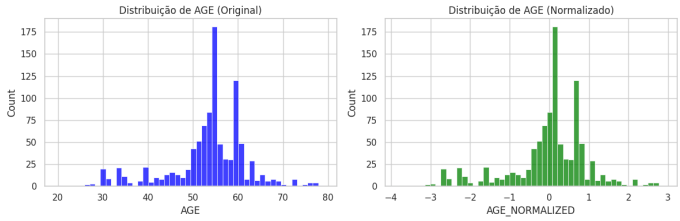

Figura 4. Normalização da variável AGE

Diante da constatação de uma concentração significativa de valores na faixa etária entre 50 e 60 anos, foi realizada a normalização da variável AGE utilizando a técnica de padronização (StandardScaler), com o objetivo de adequar os dados a uma escala com média zero e desvio padrão um.

No entanto, poucas diferenças visuais foram observadas na distribuição da variável, uma vez que os dados originais já apresentavam uma forma aproximadamente normal. A transformação alterou a escala dos valores, mas preservou o formato da curva de distribuição.

Essa etapa é importante principalmente em contextos de modelagem preditiva, onde algoritmos de aprendizado supervisionado são sensíveis à magnitude dos atributos.

3.1.2. Balanceamento de Classes com SMOTEENN

Diante da constatação de que a variável CLASS apresentava um desbalanceamento significativo — com predominância expressiva de pacientes classificados como diabéticos (Y) em relação aos grupos de pré-diabéticos (P) e não diabéticos (N) —, optou-se pela aplicação da técnica SMOTEENN para reequilibrar a distribuição das classes. O SMOTEENN combina dois métodos: SMOTE (Synthetic Minority Over-sampling Technique): realiza o oversampling da(s) classe(s) minoritária(s), gerando novas amostras sintéticas com base nos vizinhos mais próximos;

ENN (Edited Nearest Neighbors): executa o undersampling seletivo da classe majoritária, removendo amostras ruidosas ou ambíguas.

Essa abordagem híbrida é especialmente eficaz para reduzir viés nos algoritmos de aprendizado supervisionado, garantindo que modelos de classificação consigam aprender

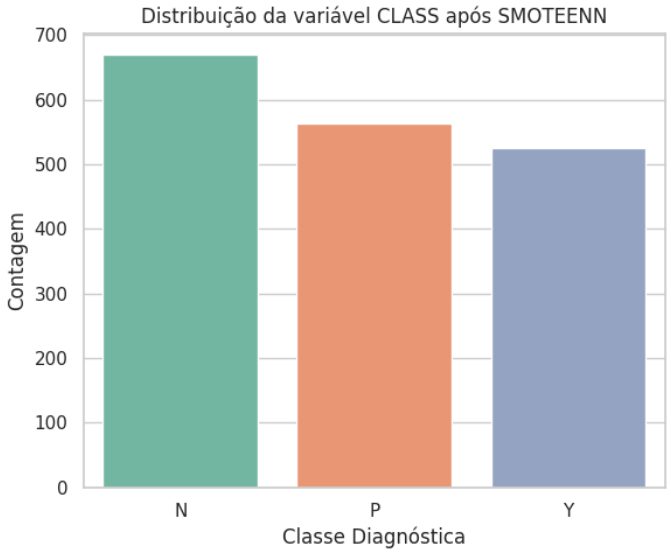

Figura 5. Balanceamento de Classes com SMOTEENN

de forma mais equilibrada entre as diferentes categorias. Após o balanceamento, observa-se na Figura 7 uma distribuição mais proporcional entre as classes N, P e Y, o que contribui diretamente para o aumento da sensibilidade e da acurácia balanceada nos modelos preditivos que serão posteriormente utilizados.

3.1.3. Metodologia

A metodologia adotada neste estudo combina etapas clássicas de análise exploratória de dados com técnicas modernas de aprendizado de máquina para prever diagnósticos de diabetes com base em características clínicas e sociodemográficas.

Inicialmente, o dataset foi importado diretamente de uma fonte pública, seguido da padronização dos nomes de colunas e checagem de valores ausentes e duplicados. Não foram identificados dados faltantes nem linhas duplicadas. Em seguida, aplicou-se a detecção de outliers com base no intervalo interquartílico (IQR), e observou-se a presença de valores extremos em variáveis como triglicerídeos (TG), creatinina (CR) e colesterol total (CHOL).

A variável alvo (CLASS) foi analisada e revelou forte desbalanceamento entre as classes. Para lidar com esse viés, foi empregada a técnica de balanceamento híbrida SMOTEENN, que combina oversampling das classes minoritárias com undersampling das amostras ruidosas da classe majoritária.

Por fim, foram aplicados modelos de classificação supervisionada — incluindo Regressão Logística, K-Nearest Neighbors, Random Forest e Support Vector Machine — com validação cruzada, a fim de comparar seus desempenhos na predição dos diagnósticos.

4. Sugestões para intervenção

- Campanhas preventivas voltadas para adultos acima de 45 anos com foco em alimentação e atividade física;

- Monitoramento contínuo da HbA1c em populações com alto IMC;

- Estratégias educacionais segmentadas por gênero e idade, considerando variações no perfil lipídico e pressão arterial;

- Adoção de modelos preditivos em unidades de saúde pública para triagem automatizada.

5. Conclusão

A partir da aplicação de diferentes algoritmos de classificação sobre um dataset clínico de pacientes com suspeita de diabetes, foi possível identificar a Regressão Logística como o modelo mais eficaz, alcançando alto desempenho mesmo em um cenário de dados desbalanceados.

A análise reforça a relevância de variáveis como HbA1c, idade e índice de massa corporal no diagnóstico da doença, indicando caminhos claros para ações de saúde pública. O uso de aprendizado de máquina se mostrou uma ferramenta promissora para triagem e diagnóstico precoce de diabetes, sobretudo quando combinado com estratégias adequadas de balanceamento e pré-processamento dos dados.

Referências

[Agência Brasil 2023] Agência Brasil (2023). Número de adultos com diabetes sobe para 10,2% no brasil, aponta vigitel. https://agenciabrasil.ebc.com.br/saude/noticia/2023-11/mais-de-10-dos-brasileiros-vivem-com-diabetes. Acesso em: 4 jun. 2025.

[Biblioteca Virtual em Saúde MS 2021] Biblioteca Virtual em Saúde MS (2021). Diabetes mellitus – brasil é o quinto país em incidência de diabetes no mundo. https://bvsms.saude.gov.br/26-6-dia-nacional-do-diabetes-4/#:~:text=O%20Brasil%20%C3%A9%20o%205%C2%BA,chega%20a%2021%2C5%20milh%C3%B5es. Acesso em: 4 jun. 2025.

[International Diabetes Federation 2024] International Diabetes Federation (2024). Idf diabetes atlas – 2024 update. https://diabetesatlas.org/. Acesso em: 4 jun. 2025.

[Laboratório Lustosa 2022] Laboratório Lustosa (2022). Diabetes: mais de 6,7 milhões de mortes no mundo em 2021. https://www.lustosa.com.br/wp-content/uploads/2022_06_Diabetes.pdf. Acesso em: 4 jun. 2025.