Spot-Checking aplicado ao diagnóstico de Diabetes

Vithor Barros Pileco
326674
vithorpileco@gmail.com

João Pedro Licks Corso
337569
jp.corso123@gmail.com

Angelo Fernandes Oliveira
550162
angelo.fernandes.oliveira@gmail.com

***Abstract*—O diabetes é uma doença crônica com graves complicações, tornando o diagnóstico precoce fundamental para a saúde pública. O Aprendizado de Máquina (AM) apresenta grande potencial para auxiliar nesta tarefa, mas a escolha de um modelo ideal não é trivial. Este trabalho realiza uma análise comparativa (spot-checking) de múltiplos algoritmos de AM para identificar os modelos mais promissores para a classificação de diabetes, utilizando o dataset DiaBD. A avaliação foi conduzida por meio de validação cruzada estratificada repetida (10x10), com foco prioritário no F2-Score para penalizar mais fortemente os Falsos Negativos, alinhando-se aos riscos clínicos do problema. Os resultados indicam que os modelos do tipo Regressão Logística (F2-Score médio de 53,0%) e o Naive Bayes (51,6%) e Rede Neural (47%) mostram-se os mais promissores.**

***Index Terms*—Aprendizado de Máquina, Diabetes, Spot-Checking**

I. Introdução

O diabetes é uma doença crônica que ocorre quando o pâncreas não produz insulina suficiente ou quando o corpo não utiliza eficazmente a insulina que produz, resultando em níveis elevados de glicose no sangue. Se não tratada ou mal controlada, a condição pode causar sérias complicações vasculares, levando à cegueira, insuficiência renal, amputações de membros e aumentando significativamente o risco de infarto, AVC e morte [1].

Essa doença se apresenta em dois tipos principais. No Tipo 1, o pâncreas perde a capacidade de produzir insulina, e sua causa e prevenção são desconhecidas. Já o Tipo 2, muito mais comum (95% dos casos [1]), ocorre quando o corpo não utiliza a insulina e está fortemente ligado a fatores de risco, como obesidade e sedentarismo. O diagnóstico precoce é essencial para evitar os efeitos graves desse tipo de diabetes. [1].

O Aprendizado de Máquina (AM) surge como uma abordagem de grande potencial para a triagem e diagnóstico na área da saúde [4]. Modelos de AM consistem em algoritmos estatísticos que visam identificar ou prever resultados através de padrões de dados coletados previamente. Aplicados ao diabetes, esses modelos podem analisar dados de pacientes para auxiliar no diagnóstico precoce da doença.

Contudo, a escolha do modelo de AM mais adequado para um problema específico não é trivial. Dado o teorema No Free Lunch [2] que postula que nenhum modelo AM consegue performar melhor, na média, que outro quando considerados todos os problemas de otimização possíveis. Tendo em vista isso, o processo de Spot-Checking [3] surge para obter uma indicação de quais modelos possuem um viés indutivo que melhor se encaixa na estrutura do problema através de testes rápidos aplicados a diferentes modelos.

Considerando a importância fundamental do diagnóstico precoce do diabetes e o desafio de selecionar um modelo de AM eficaz, este trabalho propõe aplicar a técnica de spot-checking no dataset DiaBD [5]. Com o objetivo de identificar quais são os melhores modelos para a classificação binária do diagnóstico de Diabetes.

Os resultados mostram que os modelos com viés probabilístico e linear como Naive Bayes e Regressão Logística obtiveram o melhor desempenho quando considerado o número de Falsos Negativos. Além disso, é importante ressaltar a grande interpretabilidade desses modelos, o que é muito relevante no âmbito saúde, onde é necessário entender o processo de tomada de decisão dos modelos. Por fim, um modelo baseado em rede neural também obteve um resultado satisfatório.

II. Metodologia

A. Dataset

O dataset DiaBD [5] consiste em 5.288 registros de pacientes coletados em 63 localidades urbanas, semi-urbanas e rurais de Bangladesh. A coleta de dados ocorreu entre 10 de abril de 2021 e 27 de maio de 2024, sendo realizada por agentes comunitários de saúde.

Esse conjunto de dados engloba 14 atributos de 3 distintas áreas: informações demográficas (idade, gênero), medidas clínicas (pressão arterial sistólica e diastólica, níveis de glicose, pulso, altura, peso, IMC) e histórico de saúde (histórico familiar de diabetes e hipertensão, ocorrência de AVC ou doenças cardiovasculares). A classificação de diabetes de cada instância foi validada por três grupos independentes de médicos.

Esse dataset foi obtido para o presente estudo através do download de um arquivo CSV disponibilizado pelos autores no repositório Mendeley Data [6].

B. Pergunta de Pesquisa

PP: Quais os modelos de AM, após um processo de spot-checking [3] aplicado ao dataset DiaBD [5], apresentam o desempenho preditivo mais promissor, priorizando a minimização de falsos negativos, para a classificação binária de pacientes diabéticos e não-diabéticos?

TABLE I: DICIONÁRIO DE DADOS DAS FEATURES

Feature (Nome no Dataset)	Tipo	Descrição
Idade (Age)	Inteiro	Idade do paciente em anos.
Gênero (Gender)	Categórico	Masculino, Feminino.
Freq. Cardíaca (Pulse Rate)	Float	Número de vezes que o coração bate por minuto (bpm).
PAS (Systolic BP)	Float	A pressão exercida pelo sangue contra as paredes das artérias durante cada batimento. Valores típicos para adultos são $<$ 120 mmHg.
PAD (Diastolic BP)	Float	A pressão exercida pelo sangue contra as paredes das artérias quando o coração está em repouso. Valores típicos para adultos são $<$ 80 mmHg.
Hipertensão (Hypertension)	Booleano	Hipertensão, ou pressão alta, é definida como pressão arterial consistentemente $\geq$ 140/90 mm Hg. (Não=0 ou Sim=1)
Hipertensão Familiar (Family Hypertension)	Booleano	Ter um ou ambos os pais com hipertensão. (Não=0 ou Sim=1)
Diabetes Familiar (Family Diabetes)	Booleano	Ter um pai, irmão ou outro parente próximo com diabetes tipo 2. (Não=0 ou Sim=1)
Glicose (Glucose)	Float	A quantidade de glicose (açúcar) no sangue no período medido.
IMC (BMI)	Float	O IMC é uma medida da gordura corporal de uma pessoa baseada na altura e peso. IMC = (peso em kg / (altura em m)^2)
Altura (Height)	Float	Altura (Metro).
Peso (Weight)	Float	Peso (KG).
AVC (Stroke)	Booleano	Afetado por AVC (Acidente Vascular Cerebral). (Não=0 ou Sim=1)
DCV (CVD)	Booleano	Ter doença cardiovascular (Cardiovascular Disease). (Não=0 ou Sim=1)
Diabético (Diabetic)	Booleano	Variável de classe (Não=Não diabético ou Sim=Diabético).

TABLE II: Estatísticas das Variáveis Contínuas e Categóricas

Variável (Contínua)	Média $\pm$ DP	Med.	Unid.
Idade (Age)	45.75 $\pm$ 13.42	45.0	anos
Freq. Card. (Pulse)	76.63 $\pm$ 12.23	76.0	bpm
PA Sistólica (Sys BP)	134.00 $\pm$ 22.23	130.0	mmHg
PA Diast. (Dia BP)	82.23 $\pm$ 12.48	81.0	mmHg
Glicose (Glucose)	7.56 $\pm$ 2.94	6.93	mmol/L
Altura (Height)	1.55 $\pm$ 0.08	1.55	m
Peso (Weight)	53.64 $\pm$ 10.08	53.0	kg
IMC (BMI)	22.47 $\pm$ 8.82	21.9	kg/m^2

Variável (Categórica)	Categoria	N
Gênero (Gender)	Feminino	3752
	Masculino	1536
Hist. Fam. Diabetes (Family Hist. Db)	Não	5119
	Sim	169
Hipertenso (Hypertensive)	Não	4701
	Sim	587
Hist. Fam. Hipertensão (Family Hist. HTN)	Não	5110
	Sim	178
Doença Cardiovascular (CVD)	Não	5228
	Sim	60
AVC (Stroke)	Não	5268
	Sim	20
Diabético (Diabetic)	Não	4946
	Sim	342

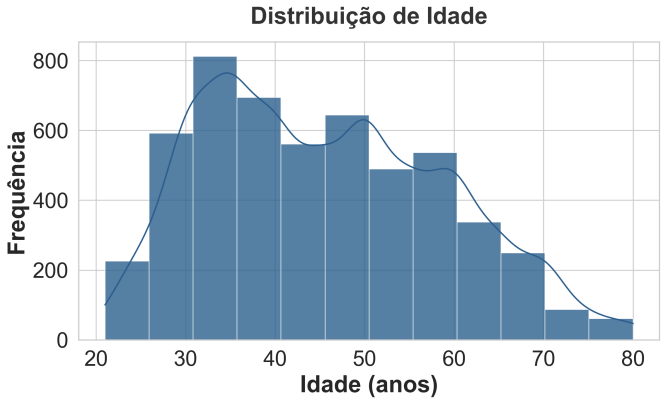

Fig. 1: Histograma por idade

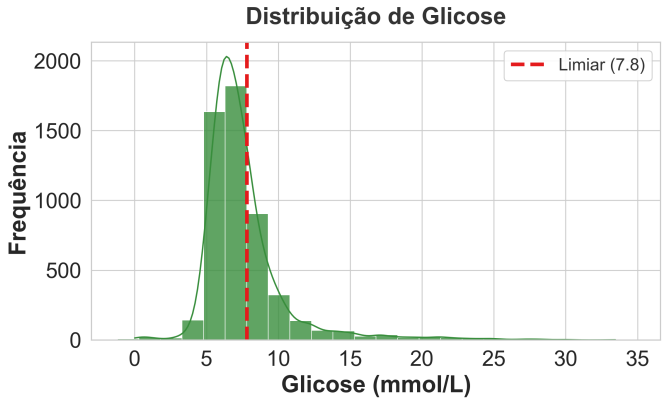

Fig. 2: Histograma por glicose

C. Análise exploratória

Inicialmente, a Tabela I apresenta o dicionário de dados do dataset, descrevendo cada feature utilizada.

A Tabela II destaca os valores estatísticos das features numéricas e as frequências das categóricas. É possível notar um desbalanceamento no atributo gênero, com mulheres compondo 71% das amostras, uma limitação já reconhecida

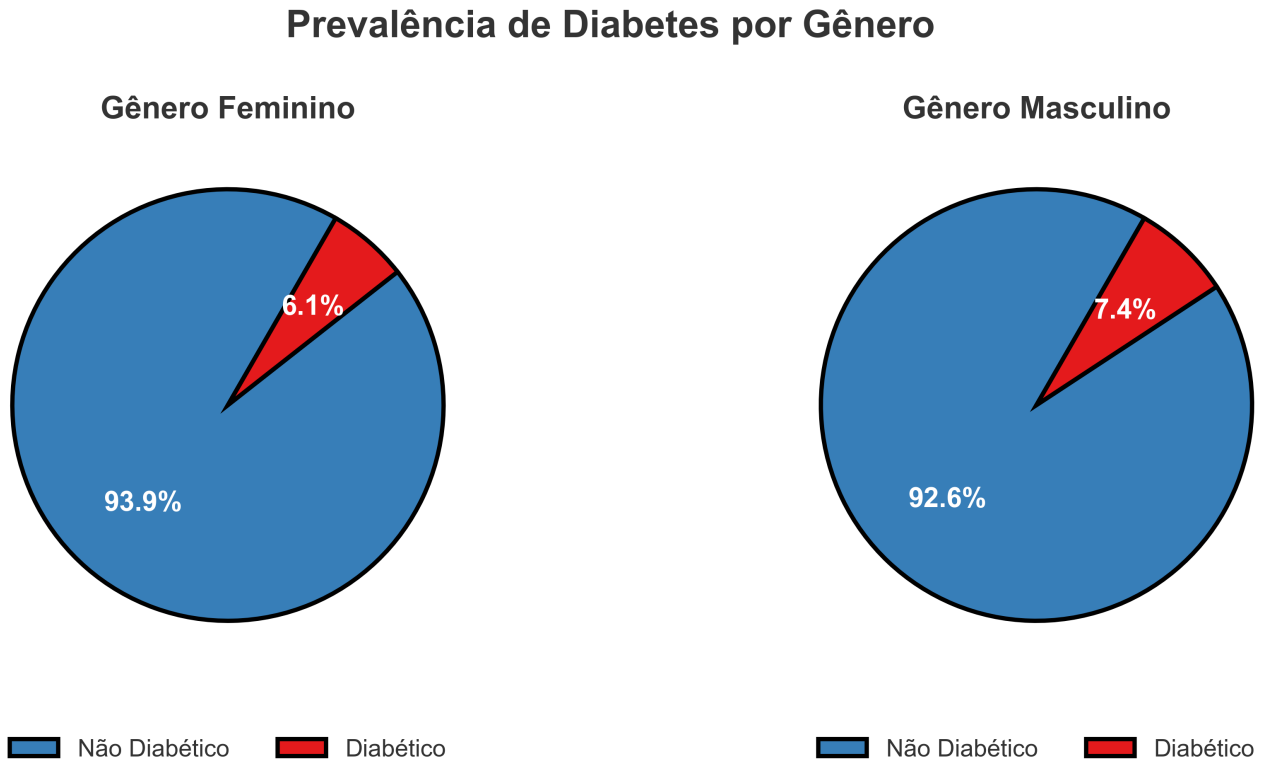

Fig. 3: Distribuição da classe diabetic por gênero

pelos autores do dataset [5].

A Figura 1 indica uma boa variedade de faixas etárias. Já a Figura 2 mostra uma distribuição assimétrica à direita da feature glicose, onde valores acima de 7,8 são indícios de diabetes ou pré-diabetes [7].

Apesar da predominância feminina no dataset, a Figura 3 demonstra que a classe alvo apresenta uma distribuição similar entre os gêneros.

A Figura 4 apresenta a matriz de correlação dos atributos do dataset, revelando associações tanto esperadas quanto anômalas. Entre as esperadas, destacam-se a relação entre as pressões diastólica e sistólica, bem como a de ambas com a condição de hipertensão.

Por outro lado, chama a atenção a correlação inesperadamente forte (0,97) entre possuir histórico familiar de diabetes e de hipertensão, o que pode indicar falhas na coleta ou pré-processamento dos dados. Curiosamente, esta mesma correlação consta no mapa presente no artigo original do dataset, mas foi interpretada erroneamente pelos autores [5]. Adicionalmente, nota-se a ausência de correlação entre as condições atuais (hipertensão e diabetes) e seus respectivos históricos, um fato que também causa estranheza.

Em relação à classe alvo, observa-se uma forte correlação com os níveis de glicose, um resultado esperado [7], e com a condição de hipertensão, explicável pela predisposição de diabéticos desenvolverem esta comorbidade [8].

D. Pré-processamento

O tratamento inicial de ruídos, duplicatas e valores ausentes foi realizado na origem dos dados [5].

Foi adicionado um novo atributo 'pulse_pressure' que foi obtido através da diferença dos atributos 'systolic_bp' e 'diastolic_bp' [9]. Subsequentemente, estas colunas originais, juntamente com 'height' e 'weight', foram removidas para mitigar a multicolinearidade evidenciada na Figura 4.

Posteriormente, as variáveis categóricas 'gender' e 'diabetic' passaram por um processo de codificação binária. Para as variáveis numéricas ('pulse_rate', 'pulse_pressure', 'glucose', 'bmi'), foi aplicada a padronização via *z-score*. A padronização foi realizada após a divisão dos conjuntos de treinamento e teste, o escalonador foi ajustado apenas com os dados de treinamento e então aplicado aos dados de teste, para evitar vazamento de dados (*data leakage*).

Por fim, utilizou-se a técnica SMOTE para realizar o balanceamento da classe alvo ('diabetic'), aplicada a cada conjunto de treinamento dos folds.

E. Modelos utilizados

Foram avaliados 6 modelos distintos, com diferentes viés indutivos. A implementação desses modelos foi através da biblioteca scikit-learn, versão 1.7.2. Para qualquer hiperparâmetro que não tenha sido detalhado nessa seção, considera-se que foi utilizado seu valor padrão na biblioteca [10].

- **Regressão Logística:**

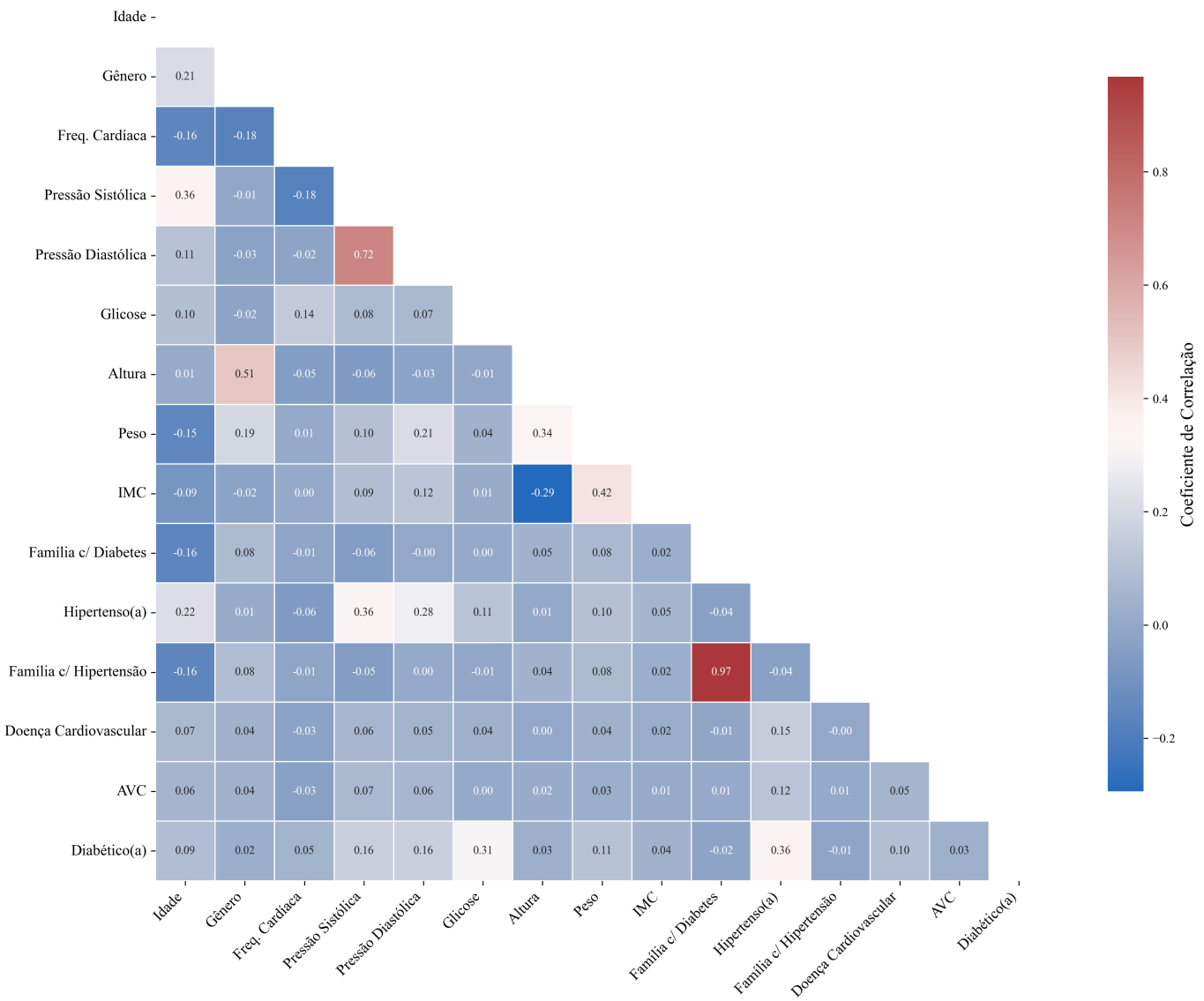

Fig. 4: Mapa de correlação do dataset DiaBD

TABLE III: Resultados dos Modelos

Modelo	**Acurácia**		**F2-Score**		**Precisão**		**Recall**		**AUC**	
	Média	DP	Média	DP	Média	DP	Média	DP	Média	DP
Árvore de Decisão	88%	0.01	32%	0.04	23.5%	0.03	35.7%	0.05	63.3%	0.03
GaussianNB	84.4%	0.04	51.6%	0.04	25.6%	0.04	70%	0.05	81.6%	0.03
KNN-3	85.4%	0.01	38%	0.04	21.5%	0.02	47.4%	0.06	71%	0.03
KNN-5	83.6%	0.01	40%	0.04	20.5%	0.02	53%	0.06	73.5%	0.03
KNN-7	82.4%	0.01	41.5%	0.03	20%	0.02	57%	0.05	75%	0.03
Regressão Logística	86%	0.01	53%	0.03	27%	0.02	70%	0.05	84.9%	0.02
MLP	87.8%	0.02	47%	0.04	29%	0.04	57%	0.08	81%	0.02
Random Forest	92%	0.007	38.7%	0.05	39%	0.05	38.8%	0.05	84.2%	0.02

Modelo estatístico linear que combina os atributos em uma função logística para estimar a probabilidade da classe-alvo. Assume que a fronteira de decisão é aproximadamente linear no espaço das features. Foi utilizado a regularização L2 com regularização C=1 e solver lbfgs

- **Árvore de Decisão:**
  Modelo hierárquico que particiona o espaço de dados por regras sequenciais, buscando maximizar ganho de

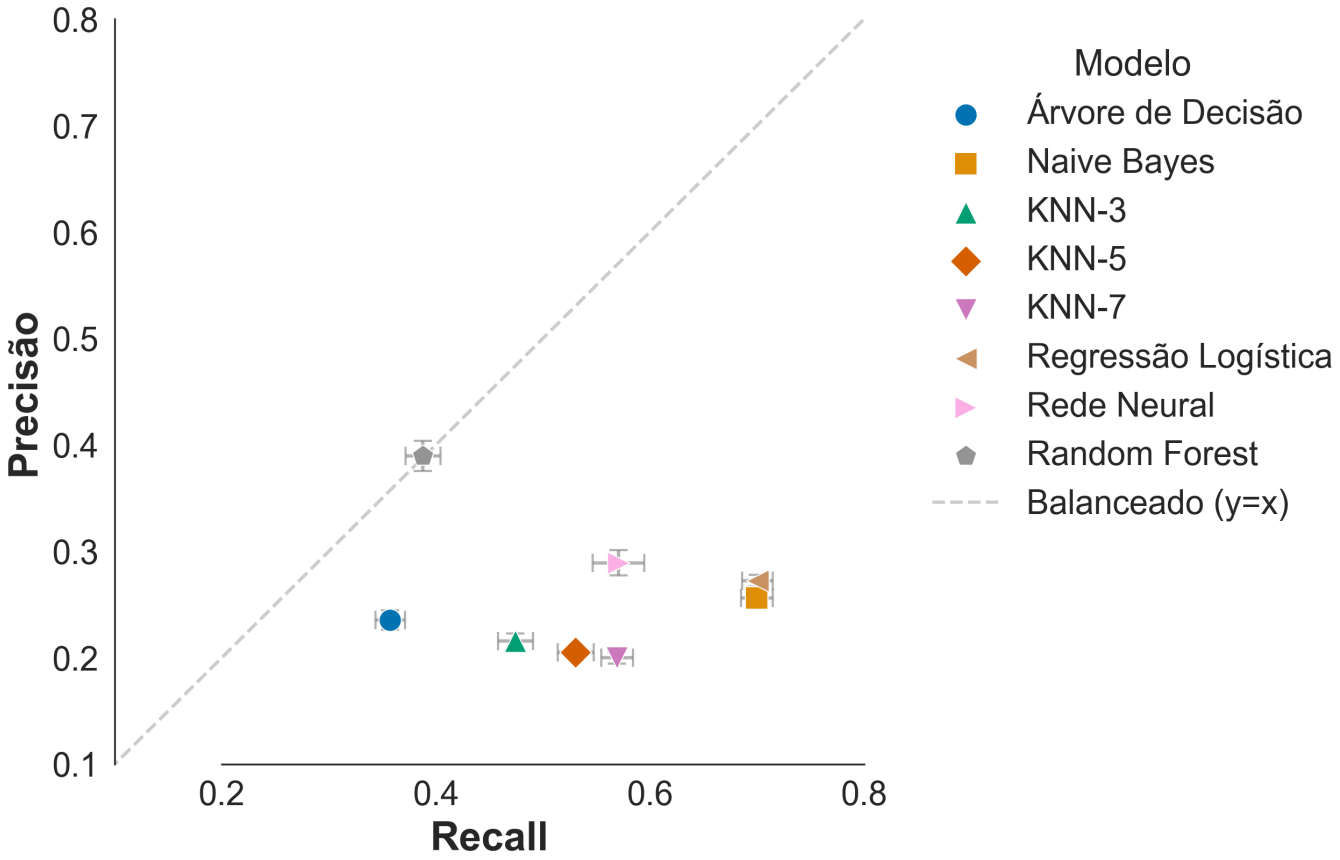

Fig. 5: Precisão/Recall dos modelos usados

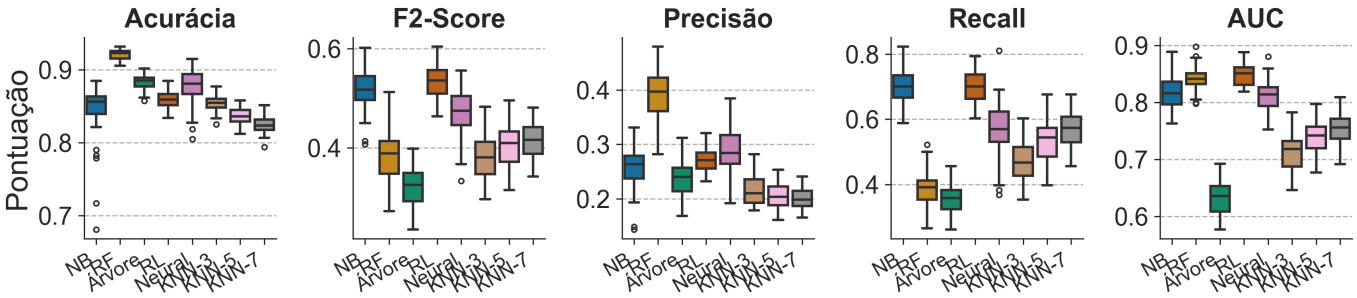

Fig. 6: Box Plot métricas dos modelos. NB (Naive Bayes); RL (Regressão Logística); RF (Random Forest); Árvore (Árvore de Decisão); Neural (Rede Neural)

informação. Para mitigar o overfitting, aplicou-se a poda por complexidade de custo (CCP), utilizando validação cruzada para selecionar o parâmetro ccp_alpha ideal dentre cinco candidatos amostrados do caminho de poda.

- **Random Forest:**
  Ensemble baseado em bagging que utiliza votação majoritária de múltiplas árvores de decisão, reduzindo a variância típica de árvores individuais. Foi utilizado um modelo com 100 árvores e parâmetros padrões da biblioteca sklearn.
- **Naive Bayes:**
  Modelo probabilístico baseado no teorema de Bayes com a suposição de independência condicional entre os atributos. Foi utilizada a variante Gaussiana (GNB), que assume que os atributos contínuos, condicionais à classe, seguem uma distribuição normal.
- **k-Nearest Neighbors (KNN):**
  Modelo baseado em instâncias que classifica uma amostra com base na votação majoritária de seus vizinhos mais próximos. Foram utilizados três variações k=3, 5, 7 com pesos uniformes, utilizando a métrica de distância euclidiana.

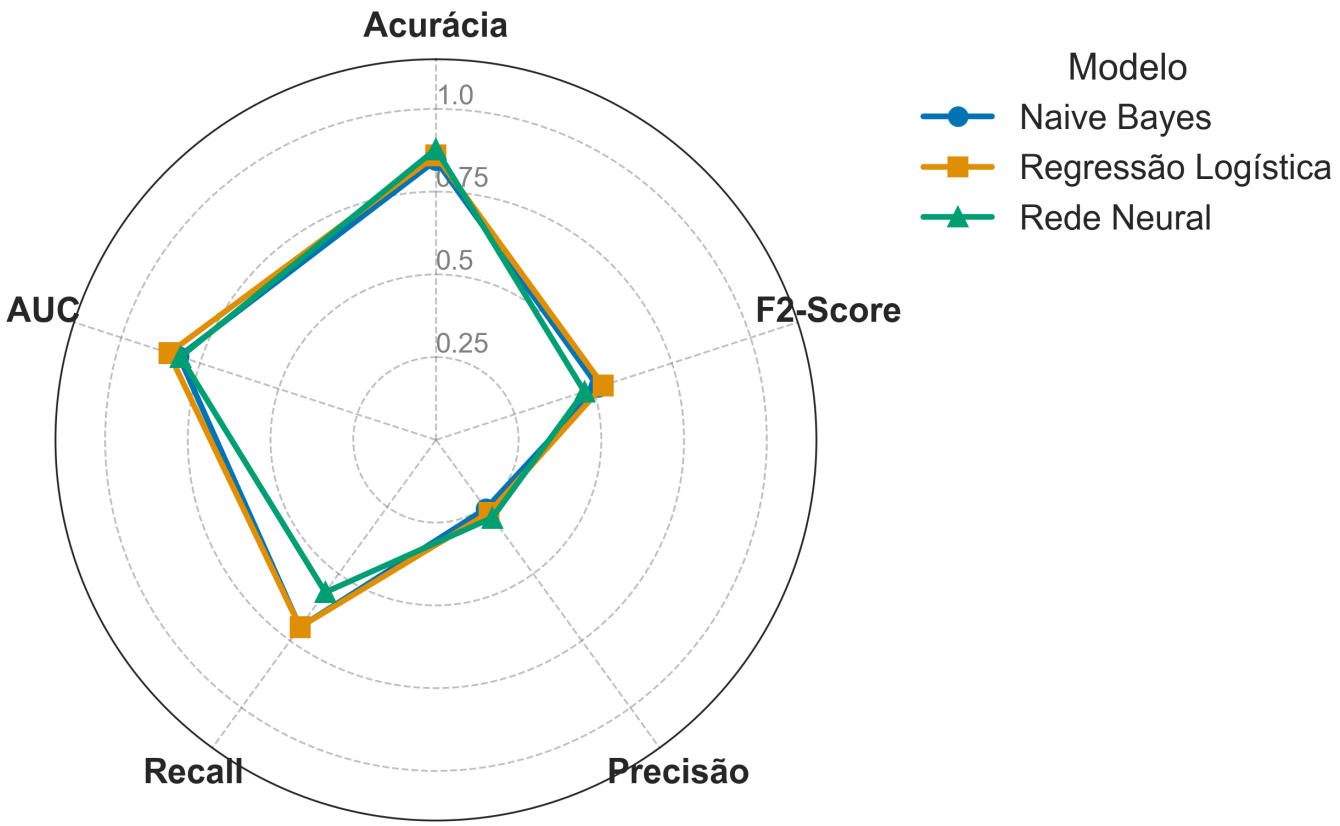

Fig. 7: Radar comparando atributos dos 3 melhores modelos

- **Rede Neural Artificial:** Modelo computacional composto por camadas de neurônios interconectados com funções de ativação não lineares. São modelos muito utilizados para fronteiras de decisões complexas, porém possuem uma baixa interpretabilidade. Foi utilizado um Perceptron de Múltiplas Camadas (MLP), a arquitetura consistiu em 3 camadas ocultas, com 100 neurônios por camada, e função de ativação RELU. O modelo foi treinado com o otimizador Stochastic Gradient Descent (SGD) por 1000 épocas, com taxa de aprendizado de 0,001 e tamanho de batch automático.

F. Divisão dos Dados

Foi adotado a validação cruzada estratificada com repetição (Repeated Stratified K-Fold), dividindo o dataset em 10 folds. A estratificação assegura que cada fold mantenha a mesma distribuição da classe alvo. Essa estratégia foi escolhida para garantir uma validação robusta dos modelos, testando-os repetidamente em diferentes partições dos dados. Esse processo foi repetido 10 vezes.

G. Métricas de Avaliação

Para uma avaliação do desempenho dos modelos, foram utilizadas as métricas descritas a seguir. Estas métricas baseiam-se nos quatro possíveis resultados de uma classificação binária: Verdadeiro Positivo (VP), Falso Positivo (FP), Verdadeiro Negativo (VN) e Falso Negativo (FN).

- **AUC (Area Under the Curve):** Corresponde à área sob a curva ROC (Receiver Operating Characteristic). A curva ROC é determinada pela relação entre a taxa de verdadeiros positivos (VP / (VP + FN)) e a taxa de falsos positivos (FP / (FP + VN)). A AUC foi adotada por sua capacidade de informar a performance geral do modelo.
- **Precisão:** Mede a proporção de instâncias classificadas como positivas que eram de fato positivas. É definida como:

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{1}$$

- **Recall:** Mede a proporção de instâncias positivas reais que foram corretamente identificadas pelo modelo. É definida como:

$$\text{Recall} = \frac{VP}{VP + FN} \tag{2}$$

- **F2-Score:** Corresponde à média harmônica ponderada entre Precisão e Recall:

$$F_\beta\text{-Score} = (1 + \beta^2) \cdot \frac{\text{Precisão} \cdot \text{Recall}}{(\beta^2 \cdot \text{Precisão}) + \text{Recall}} \tag{3}$$

No presente trabalho, adotou-se o F2-Score ($\beta = 2$), que atribui maior peso ao Recall. Priorizando modelos que possuam uma confiabilidade maior no diagnóstico

negativo, visto que um falso positivo pode ser descartado com exames mais aprofundados.

H. Spot-Checking

A seleção dos modelos de melhor desempenho foi baseada na análise estatística de suas métricas, utilizando intervalos de confiança de 95% para determinar a significância das diferenças observadas. A métrica prioritária para comparação foi o F2-Score. Essa métrica foi escolhida por atribuir o dobro do peso ao recall em relação à precisão. Isso se deve ao grande risco que o não tratamento de diabétes pode causar, assim prioriza-se a diminuição de falsos negativos.

III. Resultados

A média e desvio padrão das métricas obtidas estão sumarizados na Tabela III.

A variabilidade e a distribuição estatística desses resultados são detalhadas visualmente através de gráficos de box-plot na Figura 6. Esta figura apresenta os intervalos de confiança de 95% para cada métrica.

Adicionalmente, a relação entre Precisão e Recall nos diferentes modelos é apresentada na Figura 5.

Por fim, um gráfico radar (Figura 7) é utilizado para apresentar um perfil comparativo dos três modelos que obtiveram os maiores valores médios na métrica F2-Score: Naive Bayes, Regressão Logística e Rede Neural.

IV. Discussão

Os resultados apresentados (Tabela III e Figura 6) indicam que os modelos Naive Bayes, Regressão Logística e Rede Neural alcançaram os valores médios mais altos para o F2-Score.

Conforme detalhado na Figura 5, embora o Random Forest tenha obtido a maior precision entre todos os algoritmos, seu desempenho em recall foi um dos piores. Dado que o objetivo do problema é minimizar falsos negativos (maximizando o recall) e que o F2-Score penaliza fortemente essa falha, o Random Forest foi descartado da seleção final.

Após compilá-los na figura 7, os modelos escolhidos para a segunda etapa do trabalho foram: Rede Neural, pois, por mais que não se destaque em nenhuma métrica, tem um resultado equilibrado, o terceiro maior F2-Score e um estilo de modelo altamente otimizável. Regressão logística, pelo desempenho geral positivo em todas as métricas, e o Naive Bayes, que tem um desempenho praticamente igual em matéria de AUC e F2, ficando em segundo lugar.

A. Próximas Etapas

Na segunda etapa, será realizada a otimização de hiperparâmetros dos três modelos selecionados.

Referências

[1] World Health Organization, "Diabetes" World Health Organization, Nov. 14, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes [Accessed: Oct. 25, 2025].

[2] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization" *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, April 1997, doi: 10.1109/4235.585893.

[3] Machine Learning Mistery, "Why you should be Spot-Checking Algorithms on your Machine Learning Problems", Aug. 16, 2020. [Online]. Available: https://machinelearningmastery.com/why-you-should-be-spot-checking-algorithms-on-your-machine-learning-problems [Accessed: Oct. 25, 2025].

[4] H. Habehh and S. Gohel, "Machine Learning in Healthcare," *Curr. Genomics*, vol. 22, no. 4, pp. 291–300, Dec. 2021, doi: 10.2174/1389202922666210705124359.

[5] T. T. Prama, M. J. Rahman, M. Zaman, F. Sarker, and K. A. Mamun, "DiaBD: A diabetes dataset for enhanced risk analysis and research in Bangladesh," *Data in Brief*, vol. 61, p. 111746, 2025, doi: 10.1016/j.dib.2025.111746.

[6] Data Mendeley, "DiaBD: A Diabetes Dataset for Enhanced Risk Analysis and Research in Bangladesh", Jun. 11, 2025. [Online]. Available: https://data.mendeley.com/datasets/m8cgwxs9s6/2 [Accessed: Nov. 02, 2025].

[7] Diabetes Australia, "Blood glucose level range", [Online]. Available: https://www.diabetesaustralia.com.au/managing-diabetes/blood-glucose-range/ [Accessed: Nov. 08, 2025].

[8] Petrie JR, Guzik TJ, Touyz RM. Diabetes, Hypertension, and Cardiovascular Disease: Clinical Insights and Vascular Mechanisms. Can J Cardiol. 2018 May;34(5):575-584. doi: 10.1016/j.cjca.2017.12.005. Epub 2017 Dec 11. PMID: 29459239; PMCID: PMC5953551.

[9] Healthline, "Pulse Pressure Calculation Explained ", Abril 24, 2023 [Online]. Available: https://www.healthline.com/health/pulse-pressure [Accessed: Nov. 09, 2025].

[10] Scikit, "Scikit Learn API 1.7.2 ",[Online]. Available: https://scikit-learn.org/1.7/api/index.html [Accessed: Nov. 16, 2025].