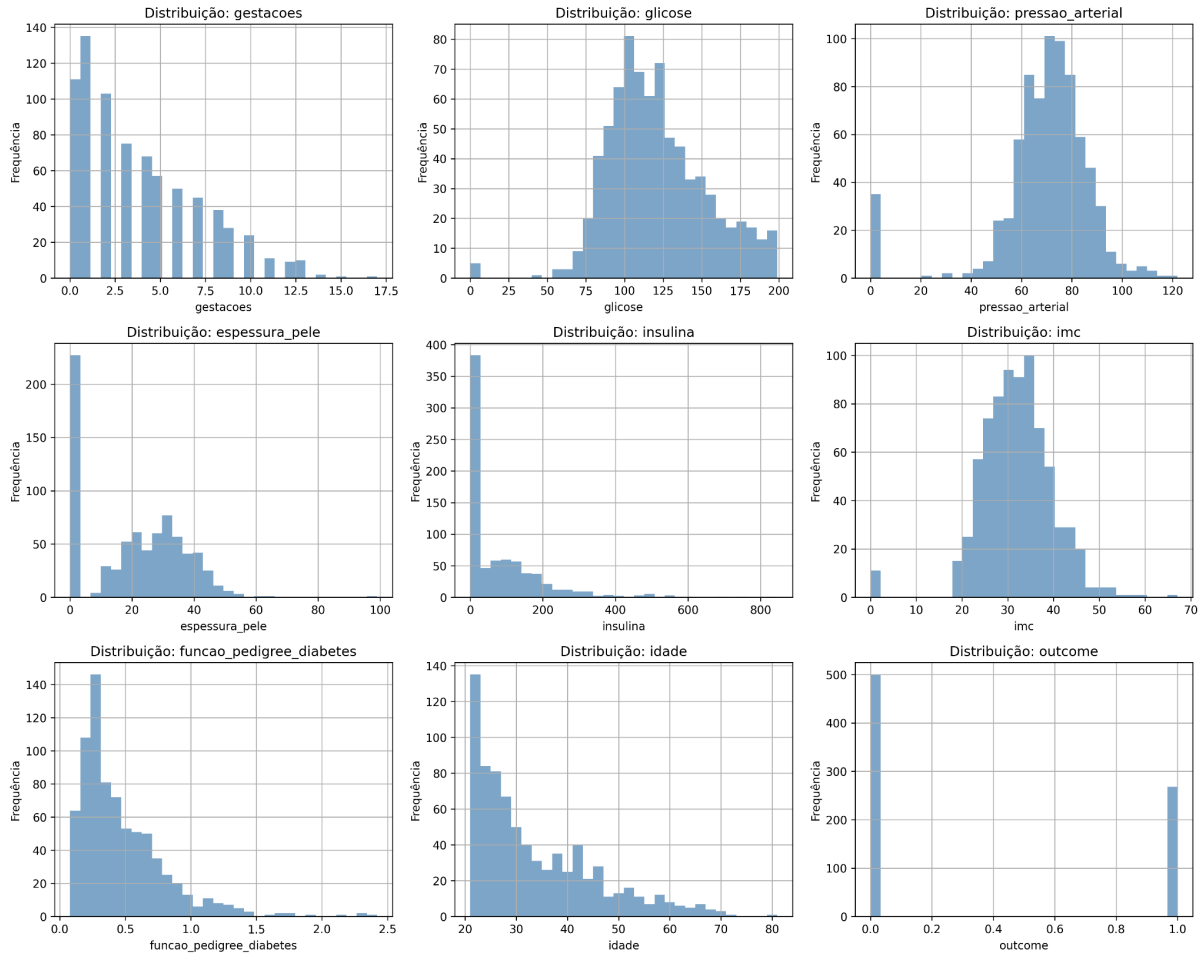
Modelo de Classificação para Diagnóstico de Diabetes Baseado em Redes Neurais

Autor: João Gabriel Galvão de Carvalho Argento
Disciplina: Redes Neurais
Instituição: Universidade do Estado do Rio de Janeiro (UERJ)

Este relatório apresenta os resultados obtidos a partir da aplicação de um modelo de rede neural do tipo *Multilayer Perceptron* (MLP) para a tarefa de classificação binária visando diagnosticar a presença ou ausência de diabetes em pacientes. O estudo foi conduzido utilizando um conjunto de dados disponibilizado pelo Instituto Nacional de Diabetes e Doenças Digestivas e Renais (NIDDK), composto exclusivamente por mulheres de ascendência indígena Pima, com idade mínima de 21 anos. Cada instância no conjunto representa um paciente, caracterizado por diversas medidas diagnósticas relevantes para a detecção da doença. Além da avaliação do desempenho do modelo MLP, também são abordados aspectos da análise exploratória de dados realizada previamente, com o intuito de fornecer contexto sobre as variáveis utilizadas e suas distribuições.

Análise Exploratória de Dados

Para iniciar a análise das características das variáveis numéricas presentes na base de dados, a distribuição de cada uma delas foi visualizada por meio de histogramas. Essa abordagem inicial permite identificar padrões, assimetrias e possíveis anomalias nos dados, além de oferecer uma visão geral sobre a variabilidade das variáveis utilizadas no modelo.

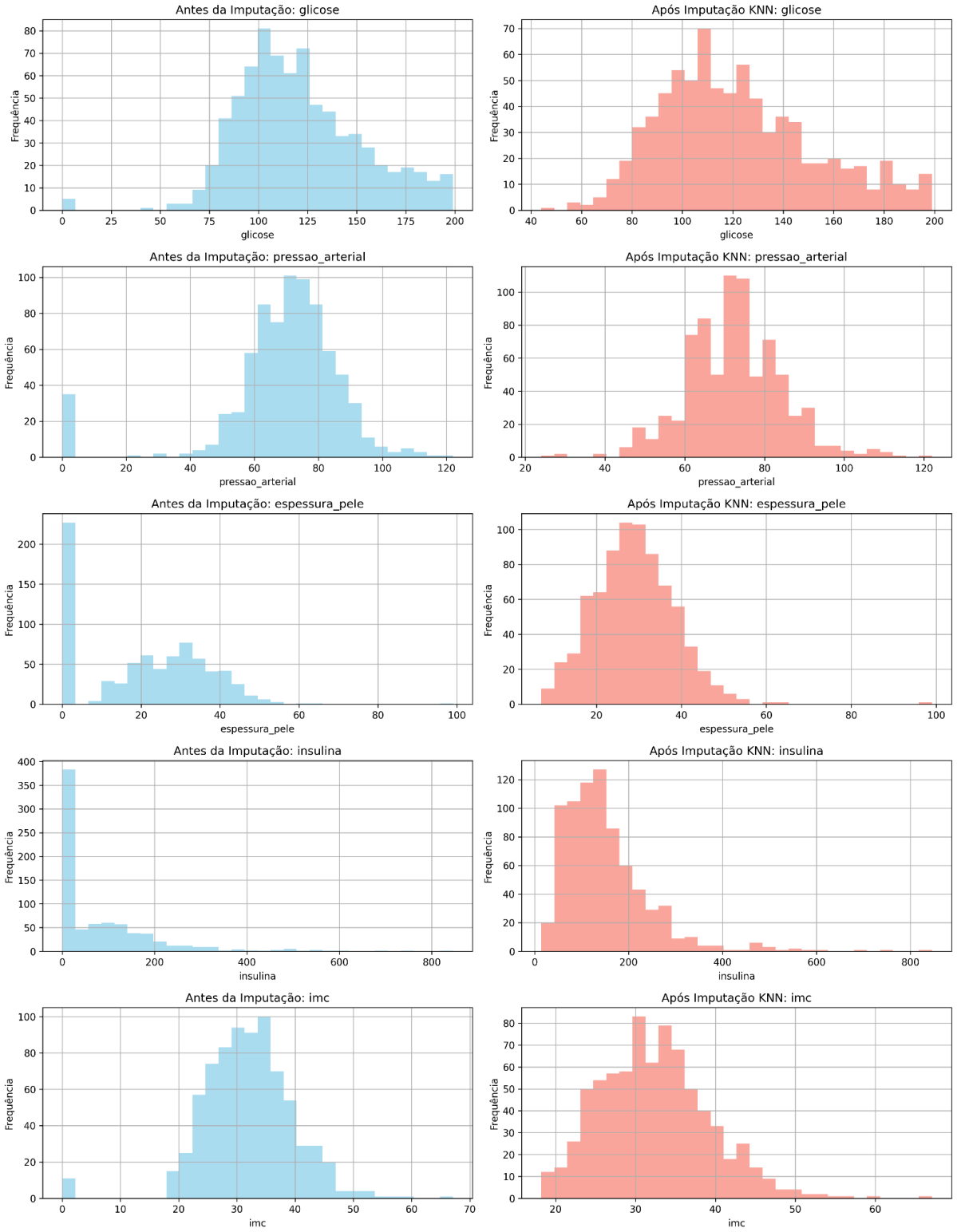


Com a observação dos histogramas acima, é possível percebermos uma presença significativa de valores zero em colunas onde esse valor não é logicamente válido. Esse problema é evidente, nas variáveis relacionadas à glicose, pressão arterial, espessura da pele, insulina e IMC (Índice de Massa Corporal), nas quais um valor igual a zero é fisiologicamente improvável ou impossível.

Com um total de 768 observações, a base de dados possui uma proporção considerável dos registros onde os zeros funcionam como ruído. Destacam-se, por exemplo, as variáveis “insulina” e “espessura da pele”, que apresentam 374 e 227 valores ruidosos (cerca de 48,7% e 29,5% do total), respectivamente. Esse tipo de inconsistência pode comprometer o desempenho de modelos preditivos ao introduzir viés ou mascarar padrões relevantes nos dados.

Diante dessa limitação, foi criada uma nova versão da base, denominada **Base 1** (sendo “Base 0” a forma original, sem imputação), com o objetivo de corrigir os valores inválidos por meio da imputação com o algoritmo *K-Nearest Neighbors* (KNN). Foram considerados os 10 vizinhos mais próximos (*n_neighbors=10*) para estimar os valores ausentes, com base em médias ponderadas das observações semelhantes.

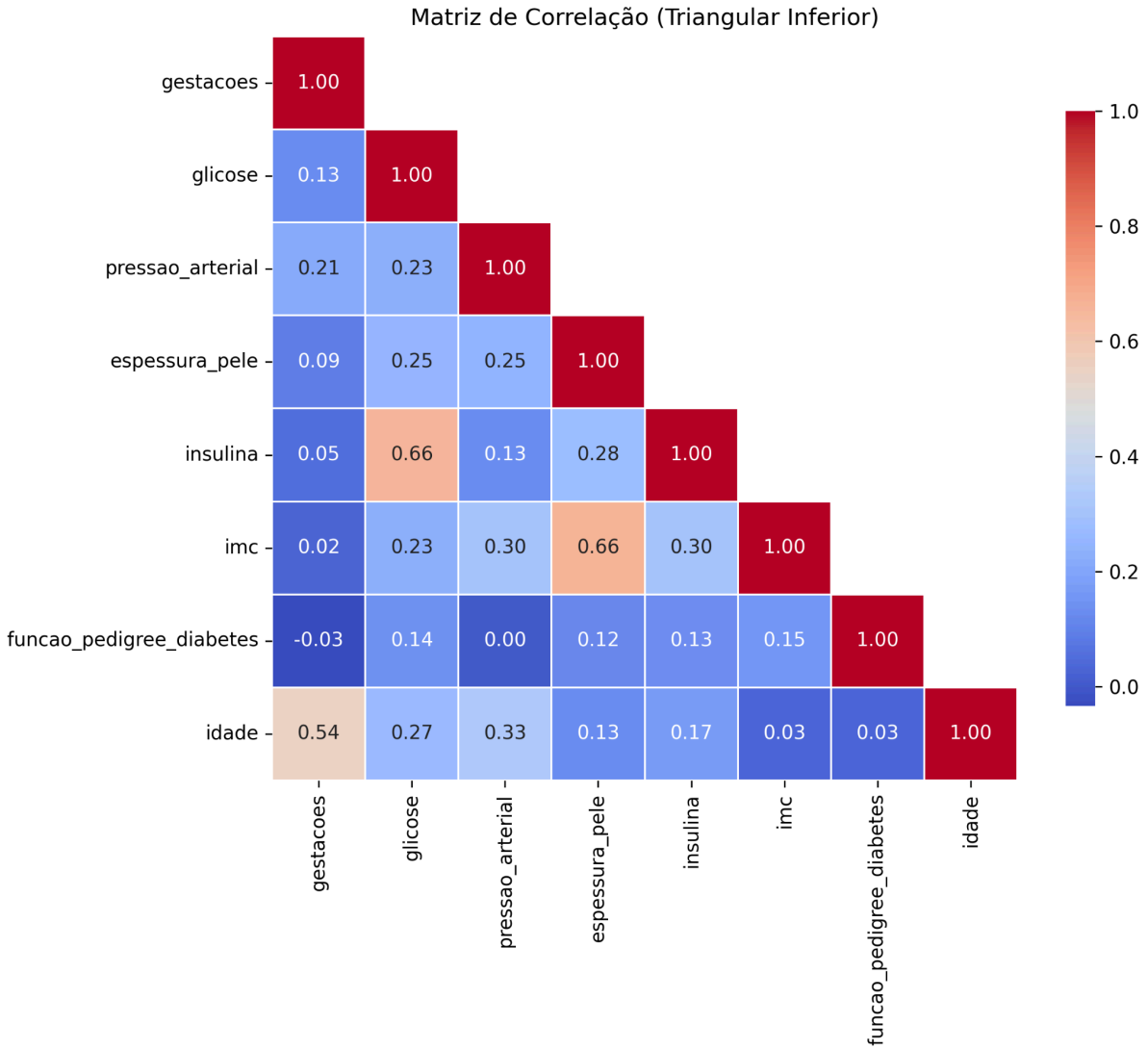
A seguir, apresentam-se os histogramas comparando a distribuição das variáveis antes e depois do processo de imputação. Essa visualização permite avaliar o impacto direto da correção dos valores ruidosos na estrutura dos dados.



Antes da imputação, os histogramas evidenciavam picos concentrados em zero nas distribuições de glicose, pressão arterial, espessura da pele, insulina e IMC, um forte indicativo da presença de registros ausentes ou não medidos corretamente. Tais picos destoavam do comportamento fisiológico esperado dessas variáveis em uma população real. Após o processo de imputação via KNN, esses picos foram eliminados e as distribuições tornaram-se mais suaves e realistas, refletindo de maneira

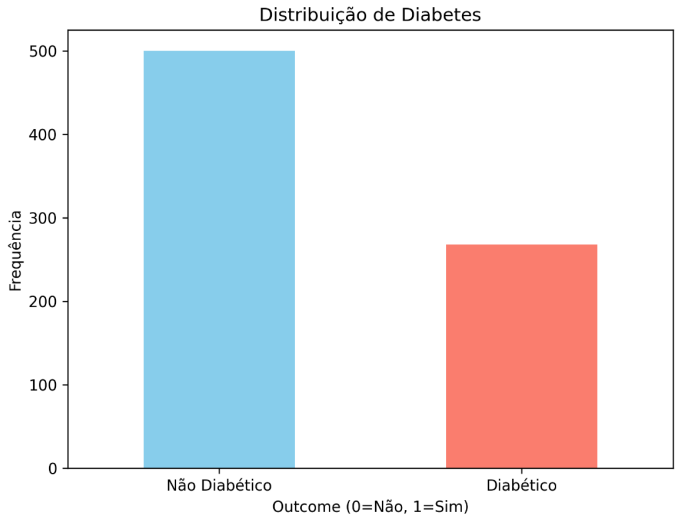
mais adequada a variabilidade natural dos dados biométricos. Com isso, espera-se que o modelo MLP se beneficie de uma representação mais precisa do fenômeno investigado.

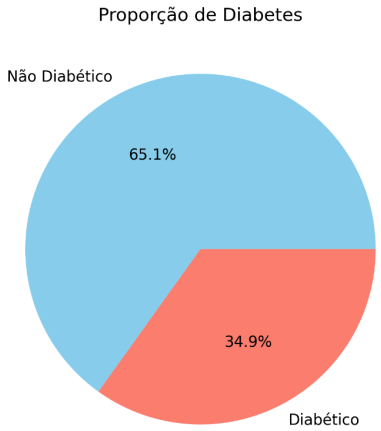
Além da análise univariada, também foi examinada a relação entre as variáveis explicativas por meio de uma matriz de correlação, utilizando a Base 1.



A correlação mais forte observada foi entre espessura da pele e IMC e entre glicose e insulina, ambos com um coeficiente de 0,66. Isso indica que indivíduos com maior espessura de pele tendem a apresentar um IMC mais elevado e indivíduos com níveis mais altos de glicose estão associados a maiores concentrações de insulina no sangue. Uma terceira correlação relevante foi entre idade e número de gestações (0,54), o que é esperado, considerando que mulheres mais velhas tendem a ter tido mais gestações ao longo da vida. As demais variáveis apresentaram correlações baixas entre si, o que sugere relativa independência entre os atributos. Isso é desejável do ponto de vista de modelagem preditiva, pois reduz o risco de multicolinearidade e pode contribuir para um melhor desempenho do modelo de rede neural.

Após as análises das variáveis explicativas do modelo, também foi avaliado a proporção de pessoas com e sem diabetes na base.





Conforme apresentado anteriormente, do total de participantes, **500 indivíduos (65,1%) não possuem diabetes**, enquanto os **268 restantes (34,9%) foram diagnosticados com a condição**. Apesar de relativamente balanceado, o conjunto de dados apresenta mais casos negativos que positivos, o que pode influenciar o desempenho do modelo. Essa diferença deve ser considerada na modelagem, especialmente na escolha das métricas de avaliação e na possível aplicação de técnicas de balanceamento de classes.

Multilayer Perceptron - MLP

Metodologia

Inicialmente, as variáveis explicativas (características do paciente) foram separadas da variável alvo ('outcome', que indica presença ou ausência de diabetes) e normalizadas com *MinMaxScaler*, garantindo que todas as características ficassem na mesma escala (entre 0 e 1).

Após a normalização, os dados foram divididos em conjuntos de treinamento (70%) e teste (30%) com a função *train_test_split* e os pesos das classes foram calculados com *compute_class_weight* (opção 'balanced'), atribuindo maior peso às classes minoritárias.

O treinamento utilizou validação cruzada K-Fold com 5 folds (*n_splits=5*), dividindo os dados em 5 partes para treinar e validar o modelo 5 vezes, cada vez com uma partição diferente para validação. Em cada fold, um modelo MLP sequencial foi construído, compilado com *BinaryCrossentropy*, otimizador *Adam* e treinado com *Early Stopping* para evitar o overfitting. O processo monitorou a *val_loss* e interrompeu o treinamento se ela não melhorasse após 15 épocas (*patience=15*), restaurando os melhores pesos. Ao final de cada fold, as métricas (acurácia, precisão, recall) foram armazenadas e o modelo com menor *val_loss* foi selecionado como o "melhor modelo".

Após a validação cruzada, a média das métricas foi calculada para estimar o desempenho geral do modelo. Como o threshold padrão de 0.5 (default) pode não ser o mais adequado, o melhor modelo, identificado na etapa anterior, foi avaliado com diferentes thresholds variando de 0.3 a 0.7, selecionando aquele que apresentou o maior F1-score médio. Como estamos lidando com um modelo de detecção de diabetes, a decisão de maximizar o F1-score se deve pelo objetivo de construir um modelo que minimize os falsos negativos, ou seja, que minimize a ocorrência de classificar erroneamente um paciente com diabetes como saudável. Os falsos negativos podem ter consequências

graves, já que o indivíduo deixaria de receber acompanhamento e tratamento adequado, aumentando o risco de complicações.

Na tabela a seguir, são apresentadas 6 combinações diferentes de arquiteturas da rede neural MLP e suas respectivas funções de ativação, com o objetivo de avaliar e identificar a configuração que resulta no maior F1-score. Cada combinação foi submetida ao mesmo processo de validação cruzada descrito anteriormente e os resultados mostrados correspondem às médias das métricas obtidas nos folds, utilizando o melhor threshold de decisão, ou seja, aquele que maximizou o F1-score em cada caso.
Além disso, é possível notar a presença de 4 diferentes funções de ativação, *ReLU*, *Swish*, *ELU* e *Sigmoid*. A *ReLU* zera valores negativos e mantém os positivos, sendo rápida e eficiente, mas pode desativar neurônios durante o treino. A *Swish* é uma função suave que permite a passagem de valores negativos de forma controlada, facilitando o aprendizado contínuo. A *ELU* também aceita valores negativos, o que ajuda a estabilizar os gradientes e acelerar a convergência. Já na camada de saída, foi usada a Sigmoid em todos os testes, pois retorna valores entre 0 e 1, interpretados como probabilidade da classe positiva, ideal para problemas de classificação binária, como a detecção de diabetes.

Combinação	Arquitetura da Rede	Funções de Ativação	Threshold	Falsos Negativos	F1-Score	Recall	Precisão	Acurácia
1	12-8-1	relu-relu-sigmoid	0.4	5.6	0.7163	0.8478	0.6238	0.7692
2	16-8-4-1	relu-relu-relu-sigmoid	0.4	4.8	0.7267	0.8703	0.6259	0.7748
3	18-9-1	swish-swish-sigmoid	0.4	7.4	0.6787	0.7974	0.5948	0.7413
4	24-12-6-1	swish-swish-swish-sigmoid	0.4	7.4	0.6819	0.7971	0.6012	0.7451
5	20-10-1	elu-elu-sigmoid	0.5	10.8	0.6815	0.7040	0.6645	0.7787
6	32-16-8-1	elu-elu-elu-sigmoid	0.5	11	0.6768	0.6990	0.6609	0.7749

Como pode ser observado nos resultados, a **Combinação 2** apresentou o melhor desempenho em termos de F1-score, alcançando **72,67%**. Essa configuração possui **quatro camadas**, com arquitetura **16-8-4-1** e funções de ativação **relu-relu-relu-sigmoid**.
Para aprofundar a avaliação, a **Combinação 2** foi testada com diferentes **batch sizes**, que definem quantas amostras o modelo processa antes de atualizar seus pesos. Foram usados os valores **48, 40, 32, 24, 16 e 8** para identificar o batch size que oferece o melhor desempenho.

Batch Size	Threshold	Falsos Negativos	F1-Score	Recall	Precisão	Acurácia
48	0.4	7.4	0.7026	0.7971	0.6323	0.7712
40	0.5	10.2	0.6943	0.7216	0.6737	0.7842
32	0.4	4.8	0.7267	0.8703	0.6259	0.7748
24	0.4	7.6	0.7036	0.7921	0.6382	0.7712
16	0.4	7.6	0.7105	0.7952	0.6480	0.7767

8	0.4	8.2	0.7181	0.7738	0.6735	0.7934

Conforme os resultados, o maior F1-score (**72,67%**) foi obtido com **batch size igual a 32**, que também apresentou a menor média de falsos negativos (**4,8**). Além disso, essa configuração alcançou uma acurácia de **77,48%**, com threshold ideal de **0,4**, precisão de **62,59%** e recall de **87,03%**.

Por fim, o melhor modelo, após as análises, foi avaliado no conjunto de teste. Na tabela a seguir, são apresentados os resultados de suas métricas, a saber, acurácia, precisão, recall, F1-score e a proporção de falsos negativos em relação ao total deste conjunto.

Falsos Negativos	**F1-Score**	**Recall**	**Precisão**	**Acurácia**
0.0519	0.7302	0.8519	0.6389	0.7792

Na avaliação com o conjunto de teste, a rede neural MLP demonstrou um desempenho alinhado ao seu objetivo principal de minimizar falsos negativos. O modelo alcançou um **recall (sensibilidade) de 85,19%**, o que resultou em apenas 12 erros de falso negativo (correspondente a 5,19% do total do conjunto). As demais métricas de performance registraram uma acurácia de **77,92%**, precisão de **63,89%** e um F1-score de **73,02%**.

Conclusão

Após o tratamento criterioso do conjunto de dados, a construção da rede neural MLP para o diagnóstico de diabetes foi realizada. A correção de inconsistências na base de dados contou com o uso do algoritmo KNN para imputar dados considerados incoerentes, ou seja, dados que registravam zeros em colunas nas quais um valor igual a zero é fisiologicamente improvável ou impossível. Além disso, a otimização do modelo com o objetivo de minimizar a ocorrência de falsos negativos foi considerada como um requisito crítico (considerando o contexto de aplicação clínica), uma vez que classificar erroneamente um paciente com diabetes como saudável representa um erro grave a ser mitigado.

Na avaliação com o conjunto de teste, os indicadores validam a eficácia do modelo como uma ferramenta de triagem, na qual a alta sensibilidade para identificar casos positivos é priorizada, aceitando-se como contrapartida uma precisão menor (63,89%), cujo impacto pode ser mitigado por exames confirmatórios.