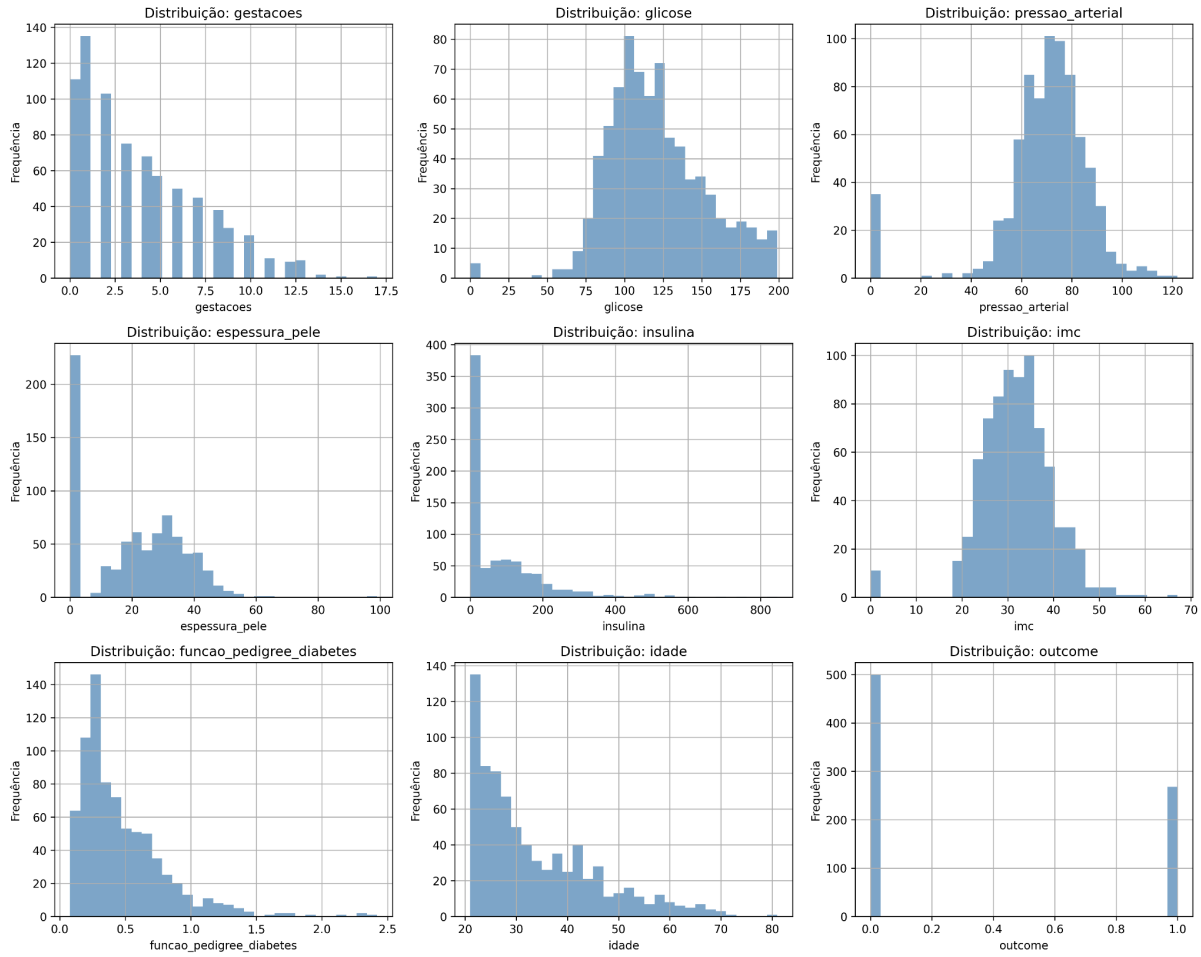
Modelo de Classificação para Diagnóstico de Diabetes Baseado em Redes Neurais

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.

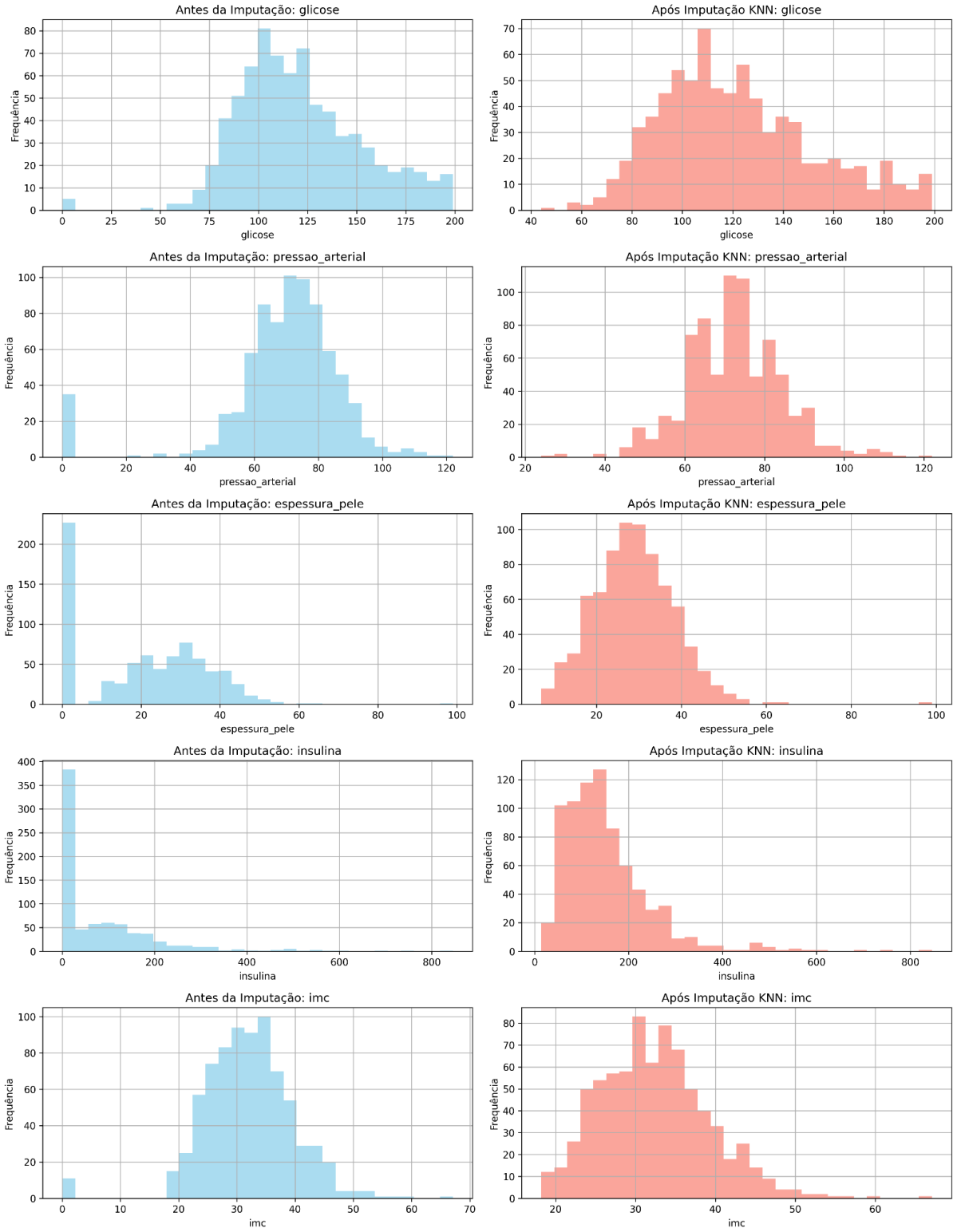


Com a observação dos histogramas acima, é possível percebermos uma presença significativa de valores zero em colunas onde esse valor não é logicamente válido. Esse problema é evidente, nas variáveis relacionadas à glicose, pressão arterial, espessura da pele, insulina e IMC (Índice de Massa Corporal),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

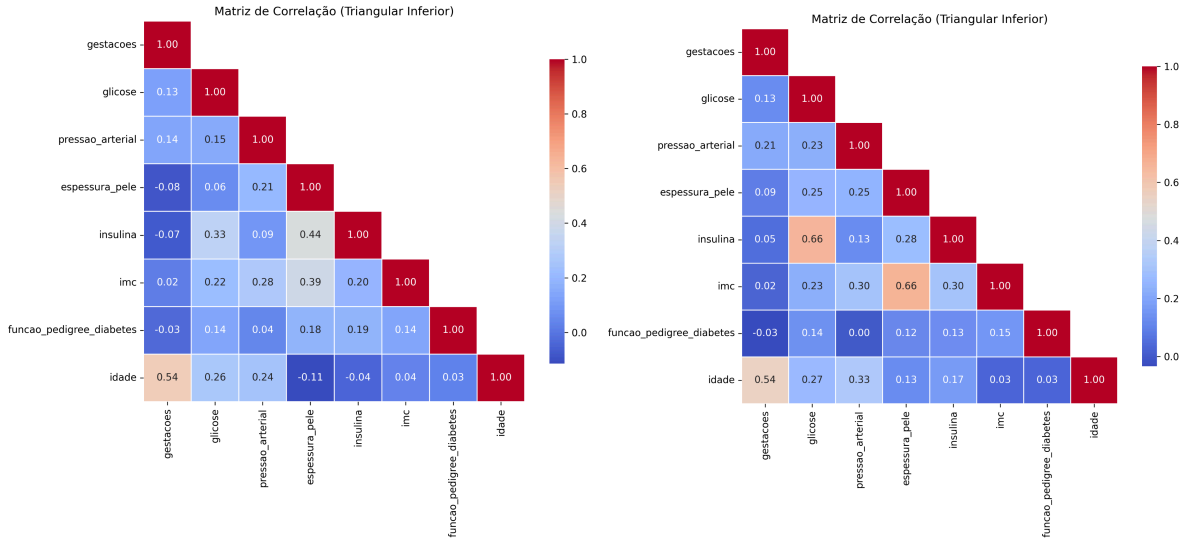
A seguir, apresentam-se os histogramas comparando a distribuição das variáveis antes e depois do processo de imputação. Essa visualização permite avaliar o impacto direto da correção dos valores ruidosos na estrutura dos dados.



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, foi avaliada a relação entre as variáveis explicativas por meio de uma matriz de correlação, comparando os resultados da Base 0 (à direita, sem imputação) e da Base 1 (à esquerda, com imputação KNN).

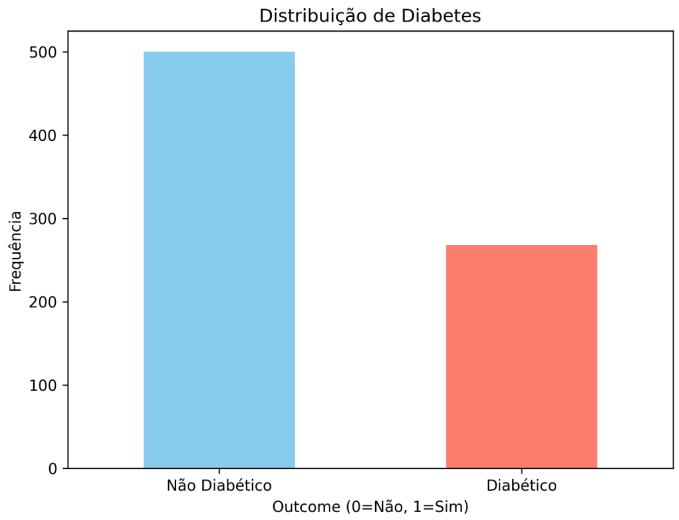


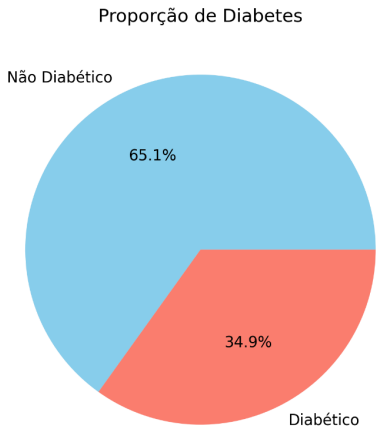
Na análise da matriz de correlação da Base 0, observa-se a ausência de relações fortes entre as variáveis explicativas, exceto pela correlação de 0,54 entre “idade” e “gestacoes” (também presente na Base 1), algo esperado, já que mulheres mais velhas costumam ter maior número de gestações ao longo da vida. Além disso, identifica-se uma correlação moderada de 0,44 entre insulina e espessura da pele, menos comum, mas presente.

Após a aplicação da imputação de dados utilizando o algoritmo KNN (Base 1), verifica-se um aumento significativo em algumas correlações: a relação entre espessura da pele e índice de massa corporal passou de 0,39 para 0,66, enquanto a correlação entre glicose e insulina aumentou de 0,33 para 0,66. Por outro lado, a correlação menos intuitiva observada inicialmente entre insulina e espessura da pele diminuiu, passando de 0,44 para 0,28, o que indica um ajuste na estrutura do conjunto de dados após a imputação. Essas alterações são coerentes do ponto de vista clínico, uma vez que indivíduos com maior espessura de pele tendem a apresentar IMC mais elevado e níveis mais altos de glicose costumam estar associados a maiores concentrações de insulina no sangue — padrões que não eram perceptíveis de forma clara na Base 0.

Esses resultados sugerem que a imputação via KNN contribuiu para enriquecer a estrutura informacional do conjunto de dados, permitindo identificar relações mais consistentes entre variáveis relevantes. As demais variáveis permaneceram com correlações baixas, indicando relativa independência entre os atributos — aspecto positivo para a etapa de modelagem preditiva, pois reduz o risco de multicolinearidade e pode favorecer um melhor desempenho da 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 objetivo principal deste estudo foi construir um modelo de rede neural MLP capaz de alcançar um equilíbrio entre Acurácia e F1-Score. A acurácia foi considerada por indicar de forma direta a proporção de previsões corretas feitas pelo modelo em relação ao total de casos avaliados, sendo útil como medida geral de desempenho. No entanto, em problemas quando o custo de erros é diferente para cada tipo (como neste caso, em que falsos negativos têm impacto mais grave), a acurácia isolada pode ser enganosa. Por isso, também se adotou o F1-Score como métrica central, pois ele representa a média harmônica entre precisão e recall, fornecendo uma avaliação mais equilibrada do modelo ao penalizar fortemente situações em que há desequilíbrio entre esses dois indicadores. O F1-Score é calculado pela fórmula: 2 × (Precisão × Recall) / (Precisão + Recall).

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".

Durante a validação cruzada, foi realizada a busca pela melhor arquitetura com o objetivo de maximizar a acurácia. Para isso, foi utilizada a técnica de *Bayesian Optimization*, escolhida por sua eficiência em encontrar boas combinações de hiperparâmetros com menos tentativas, evitando o custo computacional de buscas exaustivas (como grid search). A estratégia incluiu cinco iterações aleatórias iniciais para explorar o espaço de busca e 25 iterações adicionais guiadas pelo processo bayesiano, que aprende com os resultados anteriores para sugerir novas combinações mais promissoras. Foram testadas arquiteturas com até cinco camadas ocultas e até 50 neurônios por camada, um número limitado propositalmente para equilibrar a capacidade de aprendizado do modelo com o risco de overfitting e os recursos computacionais disponíveis.

Para as camadas internas do modelo, foram avaliadas as funções de ativação ReLU, ELU e Swish, selecionadas por sua reconhecida capacidade de mitigar o problema do desvanecimento do gradiente, ou seja, quando o sinal de correção, essencial para o aprendizado da rede neural, torna-se progressivamente mais fraco ao ser propagado para as camadas iniciais, dificultando ou até mesmo impedindo que estas aprendam com os dados. A ReLU foi incluída como um padrão de eficiência computacional, enquanto ELU e Swish foram testadas como alternativas que podem promover maior estabilidade e rapidez ao treinamento. Na camada de saída, foi utilizada exclusivamente a função Sigmoid, por ser ideal para a tarefa de classificação binária, ao converter a saída do modelo em um valor entre 0 e 1, representando a probabilidade de diagnóstico positivo para diabetes.

Após a validação cruzada, foi calculada a média das métricas para estimar o desempenho geral do modelo. Como o threshold padrão de 0.5 pode não ser o mais adequado, o melhor modelo foi avaliado com thresholds variando de 0.3 a 0.7, selecionando-se aquele que apresentou o maior F1-score médio. Como o objetivo é detectar casos de diabetes, existe preocupação especial em minimizar os falsos negativos, pois classificar erroneamente um paciente doente como saudável pode trazer graves consequências. Embora o recall seja a métrica que atua diretamente nessa redução, o F1-score foi escolhido por equilibrar recall e precisão, evitando que o modelo foque apenas em aumentar o recall à custa de piorar outras métricas como precisão e acurácia. Dessa forma, o modelo final busca manter um desempenho mais equilibrado e confiável.

Resultados

Após concluir o processo de otimização, a tabela abaixo contém as cinco melhores arquiteturas identificadas, acompanhadas de suas respectivas acurácias médias obtidas durante a validação cruzada.

Rank	Arquitetura da Rede	Funções de Ativação	Acurácia
1	43-9-15-22-31-1	relu-relu-elu-relu-relu-sigmoid	0.7841
2	48-11-23-22-31-1	elu-elu-relu-swish-relu-sigmoid	0.7786

3	51-21-23-25-37-1	relu-relu-swish-swish-relu-sigmoid	0.7786
4	51-25-22-15-37-1	relu-relu-swish-swish-relu-sigmoid	0.7786
5	49-15-19-18-39-1	relu-relu-swish-swish-relu-sigmoid	0.7767

Como pode ser observado nos resultados, a maior acurácia média atingida no conjunto de validação foi de 78,41%, com uma arquitetura de cinco camadas e funções de ativação relu-relu-elu-relu-relu-sigmoid.
Para aprofundar a avaliação, a arquitetura otimizada encontrada foi testada com diferentes **thresholds**, variando de 0.3 a 0.7, ainda na validação cruzada, selecionando-se aquele que apresentou o maior F1-score médio.

Threshold	**Acurácia**	**Precisão**	**Recall**	**F1-Score**
0.3	0.7636	0.6100	0.8798	0.7191
0.4	0.7897	0.6577	0.8155	0.7263
0.5	0.8027	0.7108	0.7180	0.7124
0.6	0.8046	0.7647	0.6285	0.6869
0.7	0.8046	0.8293	0.5473	0.6551

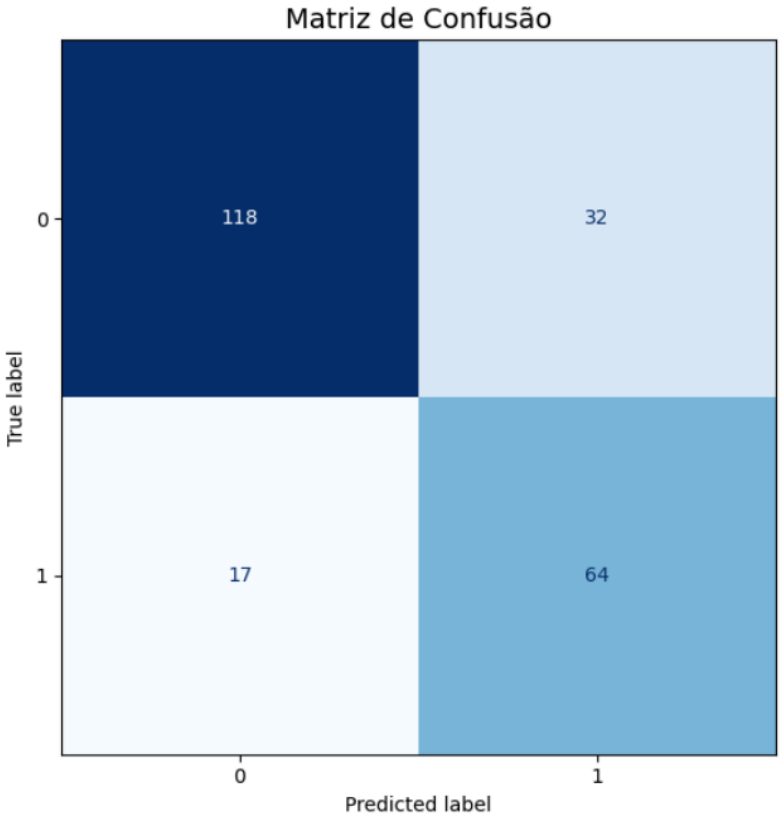
Conforme os resultados, o maior F1-score (**72,63%**) foi obtido com **threshold de 0,4**, que também apresentou o segundo melhor recall (**81,55%**). Além disso, essa configuração alcançou uma acurácia de **78,97%** e precisão de **65,77%**.

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 e F1-score.

Acurácia	**Precisão**	**Recall**	**F1-Score**
0.7879	0.6667	0.7901	0.7232

Conforme os resultados do conjunto de teste, a rede neural MLP atingiu o objetivo de apresentar um desempenho equilibrado em todas as métricas, especialmente ao buscar aumentar a acurácia e o recall sem comprometer as demais. O modelo alcançou uma acurácia de **78,79%** e um recall (sensibilidade) de **79,01%**, reforçando a importância de reduzir os falsos negativos, ou seja, evitar que pacientes com diabetes sejam incorretamente classificados como saudáveis. Além disso, obteve uma precisão de **66,67%** e um F1-score de **72,32%**, indicando que o modelo conseguiu manter bons resultados gerais, conciliando capacidade de identificar corretamente os casos positivos com uma taxa controlada de falsos positivos.

A seguir, apresenta-se a matriz de confusão obtida a partir da avaliação do modelo no conjunto de teste, permitindo visualizar de forma detalhada os acertos e erros do modelo nas classificações.



A matriz de confusão revelou que, entre os casos positivos, o modelo classificou corretamente 118 pacientes com diabetes (verdadeiros positivos) e cometeu 17 erros ao classificá-los como saudáveis (falsos negativos). Entre os casos negativos, identificou corretamente 64 pacientes como saudáveis (verdadeiros negativos), mas classificou 32 de forma incorreta como diabéticos (falsos positivos). Esses resultados indicam que o modelo foi relativamente eficaz em identificar corretamente os casos de diabetes, mantendo o número de falsos negativos relativamente baixo, algo fundamental para reduzir o risco de deixar de diagnosticar pacientes que realmente têm a doença, ao custo de um aumento moderado no número de falsos positivos, o que, embora gere maior quantidade de alertas, é preferível neste contexto por garantir maior segurança no rastreamento da condição.

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.

O requisito fundamental do projeto era a minimização de falsos negativos, dado o risco clínico associado a não diagnosticar um paciente. A estratégia, portanto, foi otimizar o modelo para um recall elevado, gerenciando o trade-off com as demais métricas. Buscou-se um equilíbrio que favorecesse a sensibilidade do modelo sem degradar excessivamente a acurácia geral.

Os resultados da avaliação confirmam o sucesso dessa abordagem. A precisão de 66,67%, embora moderada, é um resultado esperado e aceitável para uma ferramenta de triagem clínica. Nesse

contexto, o custo de um falso negativo é muito superior ao de um falso positivo, o qual pode ser facilmente identificado e descartado em exames confirmatórios subsequentes.