Identificação de Diabetes com Base Histórico Médico de Paciente: Investigação para Diferentes Classificadores

Vinicius Faber Zamarchi[1]

[1]IFPR - Instituto Federal do Paraná - Campus Palmas

viniciuszamarchi@gmail.com

Resumo. *A diabetes é um desequilíbrio metabólico caracterizado pela presença de níveis elevados de glicose no sangue que é resultado da falta de resposta adequada à insulina ou da produção insuficiente pelo organismo. Quando não é adequadamente gerenciado ou diagnosticado em tempo hábil, a diabetes pode representar um risco para órgãos vitais, podendo levar à óbito. Este estudo foca em abordagens de classificação para identificar a presença ou ausência de diabetes, considerando o histórico médico do paciente. Serão utilizadas técnicas de aprendizado supervisionado, onde as amostras obtidas no conjunto de dados Diabetes Prediction são utilizadas para estabelecer padrões e classificar os dados. O objetivo é testar diferentes parametrizações para os classificadores e determinar qual é mais eficaz na predição do diagnóstico.*

Abstract. *Diabetes is a metabolic disorder characterized by high levels of blood glucose due to inadequate response to insulin or insufficient production by the body. When not properly managed or diagnosed in a timely manner, diabetes can pose a risk to vital organs, potentially leading to fatality. This study focuses on classification approaches to identify the presence or absence of diabetes, taking into account the patient's medical history. Supervised learning techniques will be employed, where samples from the Diabetes Prediction dataset are used to establish patterns and classify the data. The objective is to test different parameterizations for the classifiers and determine which one is most effective in predicting the diagnosis.*

1. Introdução

A diabetes mellitus é uma doença crônica caracterizada por altos níveis de glicose no sangue, resultantes da falta de produção ou do uso inadequado de insulina. De acordo com a Federação Internacional de Diabetes (IDF), estima-se que 463 milhões de pessoas em todo o mundo viviam com diabetes em 2019, e esse número deve chegar a 700 milhões até 2045 [Sun et al., 2022]. No Brasil, a prevalência de diabetes tem aumentado significativamente, afetando cerca de 9,4% da população adulta, de acordo com a Pesquisa Nacional de Saúde de 2019 [Malta et al., 2022].

Para ser possível alcançar o objetivo deste trabalho, é fundamental considerar as diretrizes estabelecidas pela Sociedade Brasileira de Diabetes (SBD). Essa organização tem desempenhado um papel crucial no desenvolvimento de abordagens interdisciplinares para o controle e prevenção eficazes dessa doença complexa. Ao promover a colaboração entre profissionais de diferentes áreas da saúde, a SBD tem contribuído para a melhoria da gestão da diabetes.

A classificação emprega algoritmos computacionais para a detecção automática de alvos com base na análise de padrões, sendo categorizada como não-supervisionada ou supervisionada. Nesta pesquisa, serão utilizadas abordagens de aprendizado supervisionado, que dependem da definição das categorias e do fornecimento de amostras pelo usuário. Essas amostras são utilizadas para estabelecer padrões e, por consequência, realizar a classificação dos dados de acordo com sua categoria [Andrade et al., 2014].

Nesse contexto, este artigo busca não apenas examinar diferentes modelos de classificação e seu desempenho na identificação da presença ou ausência de diabetes em indivíduos, mas também determinar qual classificador apresenta a maior taxa de acerto na predição do diagnóstico, levando em consideração o histórico médico do paciente. Além disso, será realizada uma análise comparativa dos resultados obtidos entre diferentes parametrizações dos classificadores, a fim de identificar possíveis impactos dessa etapa no desempenho dos classificadores. Ao considerar as orientações da SBD, espera-se contribuir para aprimorar o manejo da diabetes e fornecer subsídios relevantes para a prática clínica.

2. Fundamentação Teórica

2.1. Métrica de Avaliação

Uma das principais métricas utilizadas na avaliação de desempenhos de classificadores é a acurácia, pois ela é a mensuração de quão próximas as amostras estão do valor alvo, que seria o resultado correto [Pasin et al., 2018]. Portanto, ela será o parâmetro central de comparação dos resultados obtidos neste projeto.

2.2. Classificadores

Este trabalho utiliza como base a aplicação de aprendizado de máquina para identificação de diabetes em indivíduos. Os classificadores escolhidos são: K-Vizinhos Mais Próximos (KNN), Árvore de Decisão e Rede Neural Artificial (RNA), o intuito é identificar qual deles é mais adequado para ser utilizado no *Diabetes Prediction dataset*.

2.2.1. K-Vizinhos Mais Próximos (KNN)

O KNN ou K-Vizinhos Mais Próximos, localiza um ponto em um espaço e determina a classificação com base em seus vizinhos (k). A classificação é realizada medindo a distância entre o determinado ponto e seus vizinhos, onde k representa o número de vizinhos que ele realiza o cálculo de distância. [Tan et al., 2005].

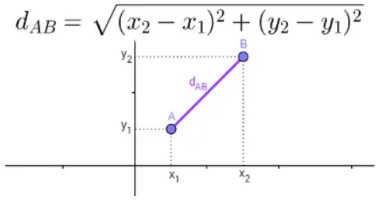

Figura 1. Exemplo – KNN
[Azank, 2019]

2.2.2. Árvore de Decisão

Um classificador de árvore de decisão divide um conjunto de dados em subconjuntos menores com base em características relevantes. Ele seleciona o melhor atributo para dividir, criando nós e ramos na árvore. Cada ramo representa um valor possível para o atributo selecionado. Esse processo é realizado de forma recursiva até que sejam alcançados os nós terminais (folhas), onde as instâncias são classificadas com base nas regras definidas na árvore [Monard and Baranauskas, 2003].

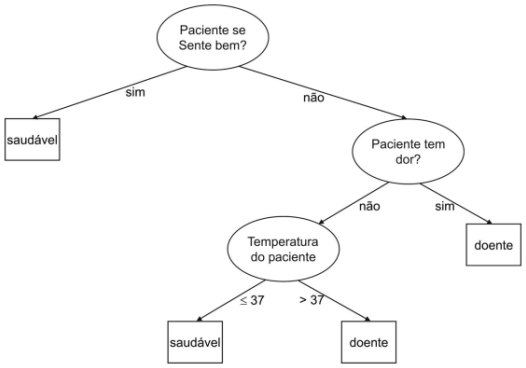

Figura 2. Exemplo - Árvore de Decisão
[Monard and Baranauskas, 2003]

2.2.3. Rede Neural Artificial (RNA)

Uma rede neural consiste basicamente em camadas interconectadas de neurônios, onde cada conexão possui um peso que determina sua relevância. Os neurônios realizam um cálculo ponderado de entradas e aplicam uma função de ativação para gerar uma saída. Este processo é repetido nas camadas subsequentes até que a resposta final seja obtida na camada de saída. Durante o treinamento, os pesos das conexões são ajustados iterativamente para minimizar a diferença entre as saídas previstas e as saídas desejadas. Isso permite que a rede aprenda a identificar padrões nos dados e faça previsões ou classificações com base nesses padrões [Haykin, 2001].

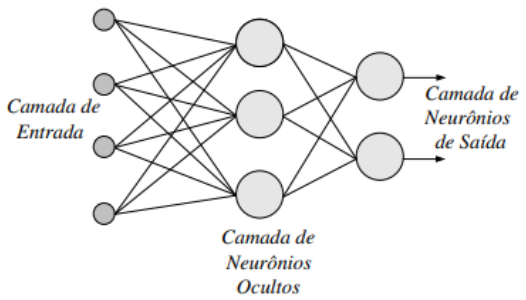

Figura 3. Exemplo - Rede Neural Artificial
[Matsunaga, 2012]

3. Metodologia

Nesta pesquisa, os dados utilizados foram extraídos do *Diabetes Prediction dataset*, obtido na plataforma online *Kaggle*, que é um local onde são reunidos diversos materiais relacionados a aprendizado de máquina. Em seguida, os dados foram carregados e trabalhados por meio do *Google Colaboratory*, uma plataforma online vinculada a nuvem onde é possível desenvolver códigos de machine learning na linguagem *Python*.

A base de dados *Diabetes Prediction* possui 100.000 tuplas e está dividida em 9 colunas, sendo elas: *age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level*,*blood glucose level* e a coluna alvo *diabetes*. As colunas *gender* e *smoking history* possuem dados do tipo *string*, enquanto as demais possuem valores numéricos.

Utilizando a função *applymap*, do *Pandas*, será realizada a conversão dos valores *string* das colunas *gender* e *smoking history* para valores numéricos. Após a conversão, os valores de *gender* passam de *female, male* e *other* para 0.0, 1.0 e 2.0, enquanto os valores de *smoking history* irão de *no Info, never, former, current, not current* e *ever* pra 0.0, 1.0, 2.0, 3.0, 4.0 e 5.0 respectivamente.

```
# Converter os valores de string para num
converte = {'Female': 0.0, 'Male':1.0, 'Other':2.0}
df = df.applymap(lambda s: converte.get(s) if s in converte else s)
converte = {'No Info': 0.0, 'never':1.0, 'former':2.0, 'current':3.0, 'not current':4.0, 'ever':5.0}
df = df.applymap(lambda s: converte.get(s) if s in converte else s)
```

Figura 4. Conversão de Dados de String para Num
Fonte: O autor

Finalizada a conversão, as variáveis *x* e *y* serão criadas de forma que, *x* terá todas as colunas do dataset armazenadas com excessão de *diabetes*, enquanto *y* terá apenas a coluna alvo (*diabetes*) armazenada.

```
# Definição de parâmetros
x = df.drop('diabetes', axis=1)
y = df['diabetes']
x.shape, y.shape
```

Figura 5. Declaração de Variáveis *x* e *y*
Fonte: O autor

Com as variáveis criadas, será feita a normalização dos dados através da função *MinMaxScaler* da *Scikit Learn*, de modo a limitar a variância dos dados a valores entre 0 e 1. A formatação será aplicada ao *dataset* utilizando a função *fit.transform*.

```
# Normalização dos dados
scaler = MinMaxScaler()
x = pd.DataFrame(scaler.fit_transform(x), columns=x.columns.values)
x.head()
```

Figura 6. Normalização dos Dados
Fonte: O autor

Após a normalização dos dados ser executada, será feita a separação dos conjuntos de treino e teste utilizando a função *train test split*, sendo definidos em 75% para o conjunto de treino o restantes para o conjunto de teste.

```
# Dividindo os Conjuntos entre treino 75% e teste 25%
x_train, x_test, y_train, y_test = train_test_split(x, y, stratify=y, train_size=0.75)
x_test, x_valid, y_test, y_valid = train_test_split(x_test, y_test, stratify=y_test, train_size=0.25)
x_train.shape, x_test.shape, x_valid.shape, y_train.shape, y_test.shape, y_valid.shape
```

Figura 7. Separação de Conjuntos de Treino e Teste
Fonte: O autor

A próxima etapa consiste na execução dos classificadores, serão realizadas três rodadas de testes para cada classificador onde a cada rodada será alterada a parametrização de cada um. Para o classificador *KNN* será alterado o número de vizinhos (k), sendo utilizado 3 na primeira rodada, 5 na segunda e 8 na terceira. Para a árvore de decisão, será alterada a profundidade máxima da árvore, sendo 3, 5 e 8 respectivamente para cada rodada. Por último, na Rede Neural Artificial os parâmetros modificados serão o número de iterações e o número de camadas, sendo definidos da seguinte forma: Para a primeira rodada, a camada de entrada terá o valor de 100, a camada intermediária 75, camada de saída 50 e 300 iterações. Segunda rodada 50 para entrada, 25 para intermediária, 10 para saída e 1000 iterações. Na terceira rodada, serão 10 na entrada, 5 na intermediária, 3 na saída e 10000 iterações. Assim serão executados os classificadores e após isso, a acurácia de cada teste será coletada e comparada para que seja possível identificar qual possui o melhor desempenho[1].

4. Resultados e Discussões

Após a execução das três rodadas de teste com o classificador *KNN*, observou-se que a configuração de $K = 5$ utilizada na segunda rodada obteve o melhor resultado, atingindo uma acurácia de 95,95%. Em seguida, na primeira execução, utilizando $K = 3$, foi obtida uma acurácia de 95,87%. Por fim, na terceira execução, com $K = 8$, a acurácia foi de 95,85%.

KNN		
Rodada	**K**	**Acurácia**
1°	3	95,87%
2°	5	95,95%
3°	8	95,85%

Figura 8. Resultados de Execuções - KNN
Fonte: O autor

A variação nos resultados de acurácia para o classificador árvore de decisão foi mínima. Na primeira execução, com uma profundidade de 3, obteve-se o melhor desempenho, alcançando uma acurácia de 97,13%. Na segunda execução, utilizando uma profundidade de 5, obteve-se uma acurácia de 97,10%. Por fim, na terceira execução, com uma profundidade de 8, a acurácia foi de 97,08%.

[1]Caderno completo com os testes: https://tinyurl.com/diabetes-pred

Árvore de Decisão		
Rodada	Profundidade Max	Acurácia
1°	3	97,13%
2°	5	97,10%
3°	8	97,08%

Figura 9. Resultados de Execuções - Árvore de Decisão
Fonte: O autor

Por fim, após a execução das três rodadas de teste com o classificador RNA, os seguintes resultados foram obtidos: o melhor resultado foi obtido durante a segunda execução, atingindo uma acurácia de 97,12%, seguido pela terceira execução que chegou a 96,91% e por último a primeira execução, obtendo 96,46% de acurácia.

Rede Neural Artificial					
Rodada	Entrada	Intermerdiária	Saída	Iterações	Acurácia
1°	100	75	50	300	96,46%
2°	50	25	10	1000	97,12%
3°	10	5	3	10000	96,91%

Figura 10. Resultados de Execuções - Rede Neural Artificial
Fonte: O autor

5. Conclusão

Embora o classificador KNN tenha apresentado resultados consistentes, em nenhuma de suas execuções obteve a acurácia acima de 96%. Já o classificador Rede Neural Artificial (RNA) apresentou desempenho melhor em comparação ao KNN, onde sua segunda execução atingiu um resultado similar ao melhor resultado obtido dentre todas as execuções.

Portanto, com base nos resultados apresentados, podemos concluir que o classificador árvore de decisão obteve o melhor desempenho em termos de acurácia. Durante as três execuções de teste, a variação nos resultados de acurácia foi mínima, mas o classificador alcançou valores mais altos em comparação aos outros dois (KNN e RNA).

Referências

[Andrade et al., 2014] Andrade, A. d., Francisco, C. N., and Almeida, C. M. (2014). Desempenho de classificadores paramétrico e não paramétrico na classificação da fisionomia vegetal. *Revista Brasileira de Cartografia*, 66(2):349–363.

[Azank, 2019] Azank, F. (2019). Modelos de predição | knn.

[Haykin, 2001] Haykin, S. (2001). *Redes neurais: princípios e prática*. Bookman Editora.

[Malta et al., 2022] Malta, D. C., Bernal, R. T. I., Sá, A. C. M. G. N. d., Silva, T. M. R. d., Iser, B. P. M., Duncan, B. B., and Schimdt, M. I. (2022). Diabetes autorreferido e fatores associados na população adulta brasileira: Pesquisa nacional de saúde, 2019. *Ciência & Saúde Coletiva*, 27(7):2643–2653.

[Matsunaga, 2012] Matsunaga, V. Y. (2012). Curso de redes neurais utilizando o matlab. *Belém do Pará*.

[Monard and Baranauskas, 2003] Monard, M. C. and Baranauskas, J. A. (2003). Indução de regras e árvores de decisão. *Sistemas Inteligentes-fundamentos e aplicações*, 1:115–139.

[Pasin et al., 2018] Pasin, M., Rodrigues, R., Schmidt, L., and Machado, R. (2018). Avaliação experimental da acurácia e da precisão de tecnologias de comunicação visando auto-localização em redes veiculares. In *Anais Estendidos do VIII Simpósio Brasileiro de Engenharia de Sistemas Computacionais*, Porto Alegre, RS, Brasil. SBC.

[Sun et al., 2022] Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B. B., Stein, C., Basit, A., Chan, J. C. N., Mbanya, J. C., Pavkov, M. E., Ramachandaran, A., Wild, S. H., James, S., Herman, W. H., Zhang, P., Bommer, C., Kuo, S., Boyko, E. J., and Magliano, D. J. (2022). IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Research and Clinical Practice*, 183:109119.

[Tan et al., 2005] Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, first edition edition.