Análise de Banco de Dados para Identificação Precoce de Diabetes

Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, MG, Brasil
Instituto de Ciências Exatas e Informática

Bruno Braga Guimarães[1], Kaio Henrique Lúcio e Santos[2], Rafael Pereira Vilefort[3], Thais ANgelica Costa Lara[4], Victor Monteiro Martinelli Grataroli[5]

[1] brunobragagalves@gmail.com, [2] kaio.khls.pkm@gmail.com, [3] rafaelvilefort@gmail.com, [4] thaiscostalara@gmail.com, [5] coldvoid77@gmail.com

ABSTRACT

O nosso estudo aborda uma análise de um conjunto de dados sobre diabetes, começando pela contextualização do problema da diabetes como uma doença crônica de alta prevalência e impacto na saúde pública. O objetivo principal foi prever a ocorrência de diabetes utilizando modelos de aprendizado de máquina. A metodologia incluiu a importação e pré-processamento dos dados com o uso de bibliotecas como *Pandas*, *Numpy* e *Seaborn* para análise exploratória. Diversos algoritmos de aprendizado de máquina foram aplicados, incluindo Regressão Logística, Máquina de Vetor de Suporte (SVM), K-Vizinhos Mais Próximos (K-NN) e *Random Forest*. Os resultados demonstraram que o modelo de Random Forest obteve o melhor desempenho, destacando-se como a abordagem mais eficaz na previsão de diabetes neste conjunto de dados.

KEYWORDS

Diabetes, Análise de dados, Visualização de dados, Random Forest, Modelo de aprendizado de máquina

1 Introdução

A diabetes é uma doença crônica prevalente, caracterizada por altos níveis de glicose no sangue. De acordo com a Federação Internacional de Diabetes (IDF), a diabetes é uma das principais causas de morte globalmente, destacando a importância da prevenção e do manejo eficaz da doença . Este trabalho busca analisar dados relacionados à diabetes para desenvolver modelos de aprendizado de máquina capazes de prever e analisar os riscos da doença em pacientes, visando intervenções preventivas mais direcionadas.

O controle eficaz dos níveis de glicose é um desafio na gestão da diabetes, levando a complicações graves, como doenças cardiovasculares, insuficiência renal e neuropatia . A análise de dados pode oferecer *insights* valiosos para entender e mitigar esses problemas. Modelos preditivos baseados em aprendizado de máquina têm o potencial de identificar padrões e fatores de risco que podem não ser evidentes através de métodos tradicionais.

O objetivo principal deste estudo é aplicar diferentes modelos de aprendizado de máquina, como Regressão Logística, Máquina de Vetor de Suporte (SVM), K-Vizinhos Mais Próximos (K-NN) e Random Forest, para prever a ocorrência de diabetes. Através da comparação do desempenho desses modelos, buscamos identificar a abordagem mais eficaz para auxiliar na previsão e gestão da doença, contribuindo assim para a implementação de estratégias preventivas mais eficazes.

2 Materiais e Métodos

2.1 Descrição da base de dados

Tabela 1 – Descrição de atributos da base de dados

Atributo	Descrição do atributo	Valores
Glucose **(Glicose)**	A glicose é um indicador importante no estudo do diabetes, pois níveis elevados de glicose no sangue (hiperglicemia) são característicos da doença.	Numérico Max= 183 Min= 85
BloodPressure **(Pressão Sanguínea)**	A pressão sanguínea também é relevante, pois o diabetes pode aumentar o risco de doenças cardiovasculares, e a pressão sanguínea elevada é um fator de risco para essas doenças.	Numérico Max= 72 Min= 40
SkinThickness **(Espessura da pele)**	Embora menos comum, a espessura da pele pode ser um indicador relevante em pesquisas sobre diabetes, especialmente em relação à resistência à insulina e à obesidade.	Numérico Max= 35 Min= 0
Insulin **(Insulina)**	A insulina é central no diabetes, pois a doença é caracterizada por uma deficiência na produção ou na ação da insulina no organismo.	Numérico Max= 168 Min= 0
Age **(Idade)**	A idade é um fator importante, pois o diabetes tipo 2 é mais comum em adultos mais velhos, embora também possa ocorrer em jovens, especialmente devido a fatores como obesidade e estilo de vida sedentário.	Numérico Max= 50 Min= 21
Outcome **(Resultado)**	Se refere ao resultado da pesquisa em relação à presença ou ausência de diabetes em indivíduos, com base nas outras variáveis como pressão sanguínea, espessura da pele, glicose, insulina e idade.	Numérico Max= 1 Min= 0

2.2 Etapas de Pré-Processamento

O conjunto de dados foi limpo e revisto, removendo valores inconsistentes, como zeros em algumas instâncias presentes nas classes: *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin* e BMI. Além disso, foi feita uma análise de diabetes e categorização de características como idade, índice de massa corporal, número de gravidezes, níveis de glicose e pressão sanguínea. Para a imputação de dados ausentes, utilizamos a técnica de imputação pela média (mean imputation). Este método substitui os valores ausentes pela média dos valores existentes para aquele atributo específico. Nenhum hiperparâmetro específico é necessário para a imputação pela média

```
# Contando quantidade de instâncias
np.unique(df['Outcome'], return_counts=True)
sns.countplot(x = df['Outcome']);
```

Figura 1 – No código foi utilizado funções como “.unique” do Pandas e “.countplot” do Matplotib.

Para identificação de um desbalanceamento, utilizamos as bibliotecas Pandas e Matplotlib, e descobrimos através da contagem das instâncias da classe, que a maior parte das pessoas não têm diabetes, portanto a base de dados é desbalanceada.

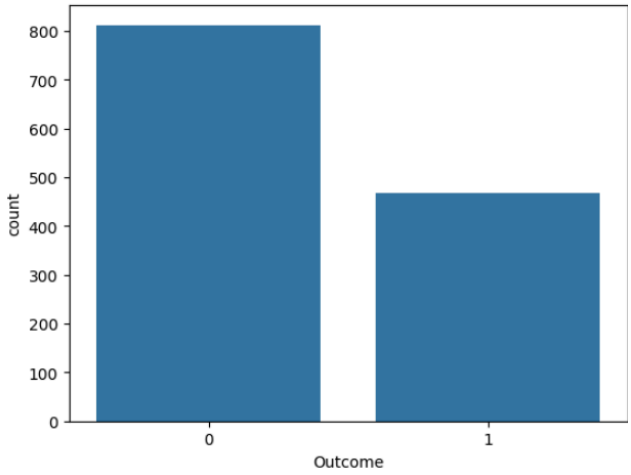

Figura 2 – A quantidade de pessoas sem diabetes(Representado por "0") é maior do que a quantidade de pessoas com diabetes(Representado por "1").

A base de dados escolhida já havia sido dividida em conjuntos de treinamento e teste não sendo necessário um método de separação

2.3 Descrição dos métodos utilizados

Utilizamos o algoritmo a *Decision Tree*(Árvore de Decisão), que é um modelo que divide o conjunto de dados em subconjuntos menores com base em características específicas dos dados, de forma a prever a classe ou valor alvo, e no caso do nosso estudo, essa previsão se trata de pessoas com diabetes a partir de alguns atributos. O hiperparâmetro utilizado se encontra em criterion='*entropy*', o qual indica que a entropia será usada como critério para avaliar a qualidade das divisões da árvore de decisão. Outra opção comum é criterion='*gini*', que usa o índice Gini como critério, mas não usamos.

Para aperfeiçoar a precisão, também utilizamos o *Random Forest*(Floresta Aleatória) que é um algoritmo que utiliza múltiplas árvores de decisão (daí o termo "floresta") para melhorar a precisão do modelo.

Cada árvore na floresta é treinada de forma independente com um subconjunto aleatório dos dados e das características. O resultado final é obtido por meio da votação das árvores individuais.

2.4 Métricas de avaliação de qualidade

Para avaliar o desempenho desses modelos, utilizamos as métricas de acurácia, precisão, *recall* e *F1-score.* No código, a acurácia foi calculada utilizando a função *accuracy_score* do *sklearn.metrics.*

A matriz de confusão foi gerada e visualizada utilizando a função *confusion_matrix*, também do *sklearn.metrics* e o ConfusionMatrix do *yellowbrick.classifier.*

O relatório de classificação foi gerado utilizando a função *classification_report*, também do *sklearn.metrics* que fornece as métricas de precisão, *recall e F1-Score.*

3. Resultados e discussões

A Árvore de Decisão apresentou uma acurácia de 0.72. Observamos que o modelo tem uma precisão maior para a classe "Não Diabético" em comparação com a classe "Diabético". A revocação para a classe "Diabético" é menor, indicando que o modelo tem dificuldade em identificar todos os casos positivos de diabetes. Isso resulta em um F1-score mais baixo para a classe "Diabético".

0

Predicted Class

	precision	recall	f1-score	support
0	0.86	0.73	0.79	111
1	0.52	0.71	0.60	45
accuracy			0.72	156
macro avg	0.69	0.72	0.69	156
weighted avg	0.76	0.72	0.73	156

Figura 3 – Relatório de classificação para Árvore de Decisão.

A Floresta Aleatória apresentou uma acurácia de 0.81. Comparado com a Árvore de Decisão, este modelo teve um melhor desempenho geral, com uma precisão e revocação mais equilibradas entre as classes. O F1-score para a classe "Diabético" foi superior, indicando uma melhoria na capacidade do modelo de identificar corretamente os casos de diabetes.

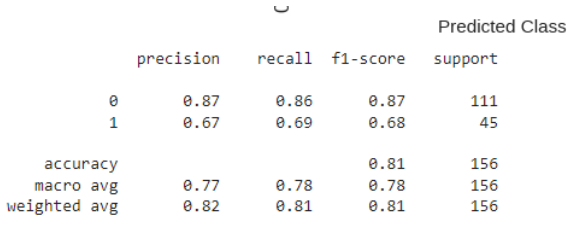

Figura 4 – Relatório de classificação para *Random Forest*.

Comparando os algoritmos, a Floresta Aleatória superou a Árvore de Decisão em todas as métricas, especialmente na precisão e F1-score para a classe "Diabético". A menor variabilidade nos resultados da Floresta Aleatória (menor desvio padrão) indica maior consistência do modelo.

Algoritmo	**Árvore de Decisão**	***Random Forest***
Precisão (Não Diabético)	0.87 (0.02)	0.87 (0.01)
Precisão (Diabético)	0.55 (0.03)	0.67 (0.02)
***Recall* (Não Diabético)**	0.77 (0.03)	0.86 (0.02)
***Recall* (Diabético)**	0.71 (0.04)	0.69 (0.03)
***F1-score* (Não Diabético)**	0.81 (0.02)	0.87 (0.01)
***F1-score* (Diabético)**	0.62 (0.03)	0.68 (0.02)
Acurácia	0.75	0.81
Desvio Padrão	0.03	0.02

Figura 5 – Comparação das métricas de desempenho entre Árvore de Decisão e Floresta Aleatória

4. Considerações finais

A escolha de um modelo para aplicações práticas pode depender de requisitos mais específicos, como a necessidade de um recall mais alto para garantir que todos os casos de diabetes sejam identificados, ou uma precisão mais alta para reduzir o número de falsos positivos. No presente estudo, o modelo *Random Forest* proporcionou um equilíbrio favorável entre estas métricas, tornando-o uma escolha recomendada quando comparado ao *Decision Tree.*

5. Utilização do GPT

Devido a sua vasta utilidade, o Chat-GPT foi utilizado para revisão e correção do código, bem como enriquecimento do texto, este redigido primariamente pelos integrantes.

6. Código desenvolvido

https://colab.research.google.com/drive/1Am7AL9L4b3OmVzbLy92KW8zB9qkCNz9H?usp=sharing

REFERENCES

International Diabetes Federation. IDF Diabetes Atlas, 9th ed. 2019. Available at: **https://www.diabetesatlas.org**

American Diabetes Association. Complications of Diabetes. Available at: **https://www.diabetes.org/diabetes/complications**