PREVISÃO DE DIABETES UTILIZANDO APRENDIZADO DE MÁQUINA

AUTOR: Rafael Tieppo | RA: 12326525 - [rafael.tieppo525@al.unieduk.com.br](mailto:rafael.tieppo525@al.unieduk.com.br)

INTRODUÇÃO

A diabetes mellitus é uma doença crônica comum caracterizada aos níveis de glicose no sangue altos. Sem um tratamento adequado, ela pode levar a sérias complicações. Dessa forma, identificar a doença no início é muito importante. Este trabalho utiliza do aprendizado de máquina para criar um modelo capaz de prever se a pessoa possui diabetes ou não, com os dados fornecidos por um dataset de dados clínicos, como a idade, glicose, IMC etc. O dataset utilizado foi Pima Indians Diabetes Database, que é um dos mais utilizados para determinado projeto.

DESENVOLVIMENTO TEÓRICO

O algoritmo utilizado foi a Árvore de Decisão, um método baseado em fluxograma, que separa os dados em grupos até atingir uma classificação final. Para um bom desempenho e interpretação da árvore, são utilizadas técnicas como a poda ou a combinação de vários modelos, como o Random Forest. Antes de alimentar o modelo, foram feitas as preliminares dos dados. Valores zero em colunas de glicose, IMC e pressão corrente foram tratados como ausentes, sendo substituídos pela mediana para não formar distorções evitando que o modelo aprenda errado. Após isso, os dados divididos em conjuntos treino e teste. As métricas usadas para avaliar o modelo foram acurácia, precisão, recall e f1-score, extraído da matriz de confusão.

DESENVOLVIMENTO - IMPLEMENTAÇÃO

A aplicação do modelo preditivo para o diabetes seguiu passos metodológicos no desenvolvimento, desde o carregamento e pré-processamento dos dados até o treinamento e validação do algoritmo de Árvore de Decisão em Python. As bibliotecas Python essenciais que atuam nessa história são apresentadas como Pandas para apoiar no processamento de dados, NumPy para operações de ponta, Matplotlib e Seaborn para personalização dos dados, e Scikit-learn para atividades de aprendizado de máquina.

Forma

O conteúdo gerado por IA pode estar incorreto.

Inicialmente, os conjuntos de dados (datasets) de treinamento e teste foram carregados a partir de arquivos CSV. Uma etapa muito importante de pré-processamento envolveu o tratamento de valores zero em colunas onde zero não é um valor clinicamente plausível, como 'GLICOSE', 'PRESSAO\_ARTERIAL', 'ESPESSURA\_PELE', 'INSULINA' e 'IMC'. Esses valores foram substituídos por NaN para indicar dados ausentes, esses valores NaN foram preenchidos com a mediana de cada respectiva coluna, calculada a partir do dataset. Essa abordagem foi aplicada tanto ao dataset de treinamento quanto ao de teste para manter a consistência e evitar que o modelo aprenda padrões errados, achando que zero é um valor válido.

Texto

O conteúdo gerado por IA pode estar incorreto.

A fase de análise exploratória dos dados permitiu uma compreensão inicial das características do dataset. Foram verificados os tipos de dados de cada coluna, visualizadas as primeiras linhas e confirmada a ausência de valores nulos após o tratamento. Também foram geradas estatísticas descritivas e usados gráficos como histogramas, boxplots e matriz de correlação para analisar a distribuição, identificar outliers e observar relações entre as variáveis e o resultado (diabetes). Essas análises ajudaram a guiar os próximos passos da modelagem.

Gráfico, Gráfico de caixa estreita

O conteúdo gerado por IA pode estar incorreto.

Foram retiradas para a preparação dos dados de modelagem, as variáveis preditoras(features) e a variável alvo (‘RESULTADO’) retiradas. As colunas ID\_Paciente e NOME também foram removidos, pois não fornecem dados clínicos para identificação da doença. A coluna SEXO foi convertida para 0 ou 1 (Feminino 0 | masculino 1).

O conjunto de dados foi então dividido em subconjuntos de treinamento e validação.

Para isso, usei a função train\_test\_split da biblioteca Scikit-learn, com a quantidade percentual predefinida que seria de teste por exemplo 20% ou 30% combinação com random\_state que é usado no reproduto de resultados.

Texto

O conteúdo gerado por IA pode estar incorreto.

O modelo de Árvore de Decisão foi criado com a biblioteca Scikit-learn, usando a classe DecisionTreeClassifier. Ele foi configurado com um random\_state para garantir resultados consistentes e treinado com os dados de treino (X\_train, y\_train). Depois, foi usado para fazer previsões com os dados de teste (X\_test).

O desempenho do modelo foi avaliado com métricas comuns de classificação. A acurácia indicou a proporção de acertos gerais. A matriz de confusão mostrou os acertos e erros do modelo em detalhes. A partir dela, foi gerado um relatório com precisão, recall e F1-score para cada classe (diabético e não diabético). Também foi feita a visualização da Árvore de Decisão com a função plot\_tree, facilitando a interpretação das decisões do modelo.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

RESULTADOS

A avaliação do modelo de Árvore de Decisão treinado para a predição de diabetes no conjunto de testes mostrou que o modelo foi realmente promissor. A acurácia global do modelo, que representa a proporção de classificações corretas, atingiu aproximadamente 77.92%. Este valor indica que o modelo foi capaz de prever corretamente o status de diabetes (positivo ou negativo) para uma maioria dos pacientes no conjunto de teste.

O relatório de classificação fornece uma análise detalhada do desempenho.

Para a classe 0 (não diabéticos), o modelo atingiu uma precisão de 0.81, que significa que 81% das previsões para não diabetes estavam corretas. O recall para esta classe foi de 0.87. O F1-score, que é a média harmônica da precisão e do recall, foi de 0.84 para a classe 0. Para a classe 1 (diabéticos), a precisão foi de 0.70. O recall para a classe 1 foi de 0.59. O F1-score para a classe 1 foi de 0.64. Essas métricas sugerem que o modelo teve um desempenho relativamente melhor na identificação de pacientes não diabéticos em comparação com pacientes diabéticos, o que é comum em conjuntos de dados onde pode haver um desequilíbrio entre as classes.

Texto

O conteúdo gerado por IA pode estar incorreto.

A visualização da árvore de decisão permitiu analisar as regras de classificação do modelo. As variáveis glicose, IMC e idade tiveram destaque na predição da diabetes. O primeiro nó da árvore frequentemente utilizava um limiar de glicose, separando pacientes com base nos níveis glicêmicos, o que está alinhado com o conhecimento médico.

As análises exploratórias, como histogramas e boxplots, mostraram diferenças claras entre diabéticos e não diabéticos, indicando que pacientes diabéticos tendem a ter valores mais altos dessas variáveis. A matriz de correlação revelou relações moderadas, como entre idade e glicose, que foram consideradas pelo modelo na definição das regras de decisão

Texto

O conteúdo gerado por IA pode estar incorreto.

CONCLUSÃO

O estudo acima tornou possível fazer uso de um modelo de Árvore de Decisão para a predição de diabetes a partir de um conjunto de dados clínicos. O modelo desenvolvido alcançou uma acurácia de aproximadamente 77.92% no conjunto de teste, indicando uma boa capacidade de distinguir entre pacientes diabéticos e não diabéticos. As métricas de desempenho como precisão, recall e F1-score revelaram que o modelo apresentou um desempenho ligeiramente superior na identificação de indivíduos não diabéticos em comparação com os diabéticos. O recall para a classe diabética foi de 0.59, sugerindo que uma proporção considerável de pacientes diabéticos não foi identificada pelo modelo, o que representa um ponto de atenção importante para futuras melhorias, especialmente considerando as implicações clínicas de falsos negativos no diagnóstico de diabetes.

Variáveis como glicose, IMC e idade foram identificadas como variáveis preditivas importantes do modelo, que confirma o conhecimento médico já existente sobre a doença. A interpretabilidade da Árvore de Decisão permitiu uma visualização clara das regras de classificação aprendidas, o que é uma vantagem significativa em aplicações médicas onde a transparência do modelo é desejável. O pré-processamento dos dados, incluindo a imputação de valores ausentes pela mediana, foi uma etapa fundamental para garantir a qualidade dos dados de entrada para o modelo.

Embora o modelo apresente um desempenho inicial promissor, existem oportunidades para otimização, como a exploração de algoritmos mais complexos (por exemplo, Random Forests ou Gradient Boosting), a aplicação de técnicas de balanceamento de classes para melhorar a detecção da classe minoritária, e a engenharia de novas features que possam capturar informações adicionais relevantes para a predição da diabetes.

CONSIDERAÇÕES FINAIS

Este trabalho contribui para a área de aplicação do aprendizado de máquina na saúde ao apresentar um modelo de Árvore de Decisão para a predição de diabetes. A metodologia empregada, desde o pré-processamento dos dados até a avaliação do modelo, demonstrou a capacidade da técnica em extrair padrões relevantes de dados clínicos. Os resultados obtidos, com uma acurácia de aproximadamente 77.92%, são promissores e estão alinhados com outros estudos que utilizam abordagens semelhantes no mesmo conjunto de dados. A identificação de variáveis como glicose, IMC e idade como fatores importantes na predição reforça a validade clínica do modelo. A interpretabilidade da Árvore de Decisão é uma vantagem significativa, permitindo que profissionais de saúde possam compreender a lógica por trás das predições, o que pode facilitar a aceitação e a integração de tais ferramentas na prática clínica.

Apesar dos resultados positivos, é essencial reconhecer as limitações do estudo e apontar caminhos para trabalhos futuros. O desempenho na detecção de pacientes diabéticos indica que há espaço para melhorias significativas, visto que falsos negativos podem ter consequências sérias. Futuras investigações poderiam explorar o uso de algoritmos de ensemble, como Random Forest ou Gradient Boosting, que frequentemente superam o desempenho de árvores de decisão únicas.

Outra abordagem promissora seria ampliar e diversificar a base de dados, incorporando novas variáveis preditoras que pudessem tornar os modelos mais robustos e aplicáveis a diferentes realidades. A validação externa do modelo também seria essencial para garantir que ele funcione bem em diferentes populações, aumentando sua confiabilidade no mundo real.

Em resumo, este estudo representa um ponto de partida importante. Com pesquisas futuras, é possível refinar e expandir essas estratégias, sempre com o objetivo de desenvolver ferramentas mais precisas e confiáveis para ajudar no diagnóstico e no tratamento da diabetes.

A aplicação de técnicas de reamostragem para lidar com o desbalanceamento de classes, como SMOTE (Synthetic Minority Over-sampling Technique), também poderia ser benéfica. Além disso, a incorporação de um conjunto de dados mais amplo e diversificado, possivelmente incluindo novas variáveis preditoras, poderia levar a modelos mais robustos e generalizáveis. A validação externa do modelo em diferentes populações também seria um passo importante para confirmar sua aplicabilidade em cenários do mundo real. Em suma, este estudo serve como um ponto de partida sólido, e as futuras pesquisas podem refinar e expandir as abordagens aqui apresentadas, com o objetivo final visando fornecer ferramentas cada vez mais precisas e confiáveis para o auxílio no diagnóstico e manejo da diabetes.

REFERÊNCIAS BIBLIOGRÁFICAS

Pima Indians Diabetes Database.

<https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>.