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 Rede Neural, um método baseado na estrutura e funcionamento do cérebro humano, capazes de aprender padrões complexos a partir dos dados. O MLP é composto por camadas de neurônios interconectados: uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Cada conexão possui um peso ajustado durante o treinamento para minimizar os erros de previsão. Funções de ativação, como a ReLU (Rectified Linear Unit) utilizada neste trabalho, permite o modelo aprender relações complexas. O otimizador 'adam' foi empregado para ajustar eficientemente os pesos da rede. Para evitar o overfitting e melhorar a generalização, técnicas como a regularização (parâmetro alpha) e a parada antecipada (early stopping) foram consideradas. Além disso, o pré-processamento dos dados é fundamental; neste caso, a normalização dos dados utilizando StandardScaler foi aplicada para garantir que todas as variáveis tivessem a mesma escala, o que é importante para o desempenho de redes neurais. O desbalanceamento entre as classes (diabéticos e não diabéticos) foi tratado com a técnica SMOTE (Synthetic Minority Over-sampling Technique), que cria exemplos sintéticos da classe minoritária para equilibrar o conjunto de dados de treinamento.

DESENVOLVIMENTO – IMPLEMENTAÇÃO

A aplicação do modelo preditivo para diabetes seguiu passos metodológicos no desenvolvimento, desde o carregamento e pré-processamento dos dados até o treinamento e validação do algoritmo de Rede Neural MLP em Python. As bibliotecas Python essenciais que atuam nessa história incluíram Pandas para manipulação e análise de dados, NumPy para operações numéricas eficientes, Scikit-learn para as funcionalidades de aprendizado de máquina (como MLPClassifier para o modelo, StandardScaler para normalização) e Imbalanced-learn ( imblearn ) para a aplicação da técnica SMOTE.

Forma

O conteúdo gerado por IA pode estar incorreto.

Inicialmente, os conjuntos de dados de treinamento e teste foram carregados a partir de arquivos CSV. Uma etapa crítica de pré-processamento envolveu o tratamento de valores zero em colunas onde tal valor não é clinicamente plausível, como 'GLICOSE', 'PRESSAO\_ARTERIAL', 'ESPESSURA\_PELE', 'INSULINA' e 'IMC'. Estes zeros foram interpretados como dados ausentes, sendo substituídos por NaN (Not a Number) e, subsequentemente, preenchidos com a mediana da respectiva coluna, calculada a partir do conjunto de treinamento. Esta abordagem foi aplicada consistentemente aos dados de treino e teste para evitar a introdução de vieses e garantir que o modelo não aprendesse padrões incorretos baseados em valores inválidos. As variáveis preditoras (features) e a variável alvo ('RESULTADO') foram separadas. Colunas não informativas para a predição clínica, como 'ID\_PACIENTE' e 'NOME', foram removidas. A variável categórica 'SEXO' foi transformada em representação numérica (0 para Feminino, 1 para Masculino) para ser utilizada pelo modelo.

Texto

O conteúdo gerado por IA pode estar incorreto.

Antes de alimentar a rede neural, as features numéricas passaram por um processo de normalização utilizando a classe StandardScaler do Scikit-learn. Este passo ajusta os dados para terem média zero e desvio padrão unitário, o que é essencial para o bom funcionamento de algoritmos sensíveis à escala, como as redes neurais. O StandardScaler foi ajustado apenas nos dados de treinamento e depois aplicado (transform) tanto no treino quanto no teste para evitar vazamento de informação do conjunto de teste. Considerando o potencial desbalanceamento entre pacientes diabéticos e não diabéticos no dataset, a técnica SMOTE foi aplicada ao conjunto de treinamento normalizado. O SMOTE gerou exemplos sintéticos da classe minoritária, resultando em um conjunto de treinamento rebalanceado, o que ajuda o modelo a aprender melhor os padrões de ambas as classes. O modelo de Rede Neural MLP foi instanciado utilizando a classe MLPClassifier do Scikit-learn. A arquitetura definida consistiu em duas camadas ocultas, cada uma com 200 neurônios Outros hiperparâmetros configurados incluíram o solver 'adam', a função de ativação 'relu', um random\_state para reprodutibilidade, o parâmetro de regularização L2 alpha=0.1 , a tolerância tol=1e-8 e a ativação da parada antecipada para prevenir overfitting e otimizar o tempo de treinamento. O modelo foi então treinado utilizando o conjunto de dados de treinamento normalizado e rebalanceado. Após o treinamento, o modelo foi utilizado para realizar previsões no conjunto de teste normalizado.

Texto

O conteúdo gerado por IA pode estar incorreto.

RESULTADOS

A avaliação do modelo de Rede Neural MLP treinado para a predição de diabetes, realizada no conjunto de testes normalizado, demonstrou um desempenho bastante promissor. A acurácia global do modelo, que mede a proporção geral de classificações corretas (tanto para diabéticos quanto para não diabéticos), alcançou aproximadamente 79.87%. Este valor sugere que o modelo foi capaz de prever corretamente o status de diabetes para uma parcela significativa dos pacientes no conjunto de teste, superando ligeiramente a acurácia obtida pelo modelo de Árvore de Decisão no estudo anterior (que foi de 77.92%). Embora o script fornecido calcule diretamente a acurácia, uma análise mais aprofundada normalmente envolveria métricas como precisão, recall e F1-score, derivadas da matriz de confusão, para entender melhor como o modelo performa em cada classe específica (diabéticos e não diabéticos). No contexto da diabetes, o recall para a classe positiva (diabéticos) é particularmente importante, pois indica a capacidade do modelo de identificar corretamente os pacientes que realmente têm a doença, minimizando falsos negativos. A precisão, por sua vez, mede a proporção de previsões positivas que estavam corretas. O F1-score combina ambas as métricas. Com uma acurácia próxima de 80%, espera-se que o modelo MLP apresente um balanço razoável entre essas métricas, embora uma análise detalhada do relatório de classificação fosse necessária para confirmar essa suposição e identificar possíveis vieses na previsão de uma classe sobre a outra. O código implementado inclui um sistema interativo que permite ao usuário inserir dados de um novo paciente (nome, sexo, idade, glicose, pressão arterial, espessura da pele, insulina, IMC), com algumas validações das entradas do paciente, verificando se tal entrada é válida no input. Esses dados são pré-processados (conversão de sexo, normalização com o scaler já treinado) e então alimentados ao modelo MLP treinado. O sistema retorna a probabilidade estimada de o paciente ser diabético e o diagnóstico final, indicando se o paciente é classificado como diabético ou não diabético. Esta funcionalidade demonstra uma aplicação prática potencial do modelo desenvolvido.

Texto

O conteúdo gerado por IA pode estar incorreto.

Texto

O conteúdo gerado por IA pode estar incorreto.

Texto

O conteúdo gerado por IA pode estar incorreto.

Texto

O conteúdo gerado por IA pode estar incorreto.

CONCLUSÃO

O estudo acima tornou possível fazer uso de um modelo de Rede Neural para a predição de diabetes a partir de um conjunto de dados clínicos. O modelo desenvolvido alcançou uma acurácia de aproximadamente 79.87% no conjunto de teste, indicando uma capacidade robusta de distinguir entre pacientes diabéticos e não diabéticos, e apresentando uma melhoria em relação ao modelo de Árvore de Decisão avaliado anteriormente. A utilização de técnicas de pré-processamento adequadas, como o tratamento de valores ausentes/inválidos, a normalização dos dados com StandardScaler, e o tratamento do desbalanceamento de classes com SMOTE, foram etapas fundamentais para alcançar este desempenho. A Rede Neural MLP, configurada com duas camadas ocultas, função de ativação ReLU e otimizador Adam, mostrou-se capaz de aprender padrões complexos nos dados. A implementação de parada antecipada e regularização ajudou a mitigar o risco de sobreajuste. Embora a interpretabilidade das Redes Neurais seja geralmente menor em comparação com Árvores de Decisão, o ganho em acurácia pode ser vantajoso em cenários onde o desempenho preditivo é prioritário. A falta de um relatório de classificação detalhado no script limita a análise completa do desempenho por classe, mas a acurácia geral obtida é um indicador positivo.

CONSIDERAÇÕES FINAIS

Este trabalho contribui para a aplicação de aprendizado de máquina na área da saúde, especificamente na predição de diabetes, utilizando um modelo de Rede Neural MLP. A metodologia empregada, desde o pré-processamento cuidadoso dos dados até a configuração e treinamento da rede neural, resultou em um modelo com desempenho promissor, evidenciado pela acurácia de quase 80% no conjunto de teste. Este resultado está em linha ou supera muitos estudos que utilizam abordagens de aprendizado de máquina no mesmo conjunto de dados Pima Indians. A escolha da Rede Neural MLP permitiu capturar relações não-lineares nos dados, potencialmente levando a uma maior precisão preditiva em comparação com modelos lineares ou árvores de decisão simples. A inclusão de etapas como normalização e tratamento de desbalanceamento com SMOTE foi crucial para otimizar o treinamento e a performance do modelo. O sistema interativo de predição desenvolvido ao final do script ilustra como tal modelo poderia ser integrado em uma ferramenta de auxílio ao diagnóstico. Apesar dos resultados encorajadores, existem limitações e oportunidades para trabalhos futuros. A avaliação do modelo poderia ser aprofundada com a análise da matriz de confusão, precisão, recall e F1-score por classe, além da curva ROC, para um entendimento mais completo de seus pontos fortes e fracos, especialmente na identificação da classe minoritária (diabéticos). A otimização dos hiper parâmetros da rede neural (número de camadas, neurônios por camada, taxa de aprendizado, etc.) através de técnicas como Grid Search ou Random Search poderia levar a melhorias adicionais no desempenho. A exploração de arquiteturas de redes neurais mais avançadas ou a combinação com outras técnicas de aprendizado de máquina (ensembles) também são caminhos promissores. Finalmente, a validação do modelo em conjuntos de dados externos, provenientes de populações diferentes, seria essencial para avaliar sua generalização e aplicabilidade clínica em cenários do mundo real. Este estudo estabelece uma base sólida, e pesquisas futuras podem refinar a abordagem para fornecer ferramentas preditivas ainda mais precisas e confiáveis para o manejo da diabetes.

REFERÊNCIAS BIBLIOGRÁFICAS

Pima Indians Diabetes Database.

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