IEEE
Computational
Intelligence
Society

Universidade de Brasília
IEEE Student Branch

Avaliação Prática

IEEE Computational Intelligence Society | Universidade de Brasília
IEEE Student Branch

Recado ao Candidato

Toda a equipe do CIS gostaria de te parabenizar por ter chegado até aqui! Sabemos que processos seletivos podem ser desafiadores, e sua participação já demonstra seu interesse e dedicação. Independentemente do resultado desta avaliação, esperamos manter contato para futuras oportunidades de colaboração.

De todos os membros do CIS,
Boa sorte!

Instruções da Avaliação

Este exame tem como objetivo avaliar suas competências em inteligência artificial e ciência de dados, abrangendo conceitos de Machine Learning, Deep Learning, Estatística e Processamento de Dados.

Todas as respostas devem ser enviadas para o seguinte e-mail:
pscisunb2024@gmail.com

Prazo de Entrega

Todas as respostas devem ser enviadas até o dia 08/04 para o e-mail indicado. Certifique-se de revisar suas respostas antes do envio.

Observações

☑️ Todas as soluções devem ser implementadas em **Python**, utilizando bibliotecas como Pandas, NumPy, Matplotlib, Seaborn e Scikit-Learn.
☑️ As respostas devem ser organizadas em um notebook do **Google Colab** (.ipynb).
☑️ O notebook deve conter todas as **soluções, explicações e visualizações necessárias.**
☑️ Todas as células do código devem ser executadas e os outputs devem estar visíveis no arquivo enviado. **Respostas sem os resultados das execuções não serão consideradas.**
☑️ O envio de arquivos incompletos ou que não sigam as diretrizes estabelecidas pode resultar em **penalizações na pontuação final.**

Recomendação: Utilize os recursos do Google Colab, como células de texto para documentar suas respostas e gráficos interativos para melhorar a visualização dos dados.

IEEE Computational Intelligence Society

Universidade de Brasília
IEEE Student Branch

Introdução

A crescente incidência de **diabetes mellitus** tem se tornado uma preocupação global, exigindo estratégias cada vez mais eficazes para diagnóstico precoce e prevenção da doença. Neste contexto, a análise de dados clínicos desempenha um papel fundamental na identificação de padrões e fatores de risco associados ao desenvolvimento do diabetes.

O conjunto de dados utilizado neste desafio foi originalmente coletado pelo **National Institute of Diabetes and Digestive and Kidney Diseases** e tem como objetivo prever, com base em medições diagnósticas, se um paciente possui diabetes. Para isso, o dataset contém informações de pacientes do **sexo feminino, com idade mínima de 21 anos e pertencentes à etnia indígena Pima.**

Os dados incluem variáveis como **níveis de glicose, pressão arterial, índice de massa corporal (IMC), número de gestações e histórico familiar de diabetes.** A variável **Outcome** indica se a paciente foi diagnosticada com **diabetes (1)** ou **não (0)**.

Por meio deste desafio, você deverá explorar esse conjunto de dados para extrair insights relevantes, visualizar padrões estatísticos e desenvolver modelos preditivos para a detecção da doença.

Questões

1. Existem valores faltantes ou outliers no dataset? Se sim, como você abordaria o tratamento dessas inconsistências? Explique as técnicas que utilizaria para lidar com essas questões.
2. Como as principais variáveis (Glucose, BloodPressure, BMI, etc.) estão distribuídas? Utilize histogramas e boxplots para representar visualmente essas distribuições e analise as características de cada uma.
3. Existe uma correlação entre a idade dos indivíduos e a presença de diabetes? Realize uma análise estatística (como teste de correlação) e utilize gráficos (como scatter plot ou boxplot) para ilustrar essa relação.
4. Quais variáveis apresentam maior correlação com a presença de diabetes? Quais variáveis parecem ser as mais indicativas da presença de diabetes?
5. Existe uma relação entre o IMC dos pacientes e o diagnóstico de diabetes? Compare os valores médios de IMC entre os grupos com e sem diabetes, e analise a diferença estatisticamente.
6. Existe um valor específico de glicose que pode ser considerado crítico para o diagnóstico de diabetes? Utilize gráficos de dispersão e cálculos estatísticos para investigar esse ponto e definir um limite crítico, se possível.
7. Treine um modelo de árvore de decisão para prever a presença de diabetes com base nas variáveis do dataset. Qual foi a acurácia obtida? Discuta os resultados e possíveis melhorias para o modelo.

8. A variável DiabetesPedigreeFunction está relacionada à presença de diabetes? Pacientes com histórico familiar de diabetes apresentam maior risco? Realize uma análise exploratória e estatística para verificar essa relação.
9. Pacientes com mais de 50 anos têm taxas de diabetes mais altas do que pacientes mais jovens? Utilize estatísticas descritivas e gráficos comparativos para demonstrar as diferenças entre esses dois grupos etários.
10. Utilize regressão logística para estimar a probabilidade de um paciente ser diagnosticado com diabetes. Quais variáveis são mais influentes no modelo e como elas impactam a probabilidade de diagnóstico?
11. Quais técnicas de feature engineering podem ser aplicadas para melhorar a previsão do diagnóstico de diabetes utilizando modelos de aprendizado de máquina? Experimente transformar variáveis existentes, criar novas variáveis a partir de combinações ou interações e utilize técnicas como encoding, normalização ou transformação de características. Avalie o impacto dessas mudanças no desempenho de um modelo de aprendizado de máquina (por exemplo, Random Forest ou XGBoost).

Critérios de Avaliação

Explicação escrita: Clareza na justificativa das abordagens adotadas.
Qualidade do código: Código limpo, organizado e bem estruturado.
Desenvolvimento dos algoritmos: Correção e eficiência das soluções propostas.