Bootcamp: Cientista de Dados

Módulo 5: Desafio Final

O propósito deste desafio é colocar em prática todas as etapas de um projeto de ciência de dados. Isso envolve a coleta, a preparação, a análise e a implementação de um algoritmo de regressão linear nos conjuntos de dados de pacientes médicos. O objetivo final é aplicar e consolidar o conhecimento adquirido ao longo do curso.

Objetivos geral do desafio

Exercitar os seguintes conceitos trabalhados no curso:

1. Coleta de dados estruturados;
2. Tratamento, limpeza e processamento de dados;
3. Análise de dados;
4. Visualização de dados;
5. Desenvolvimento de algoritmos de *Machine Learning;*
    a. Regressão linear.
6. Práticas de manipulação de dados.

Enunciado

Você assumiu a função de Cientista de Dados em uma clínica médica altamente respeitada. Sua principal missão é desenvolver um modelo de regressão linear capaz de estimar os níveis de colesterol dos pacientes com base no peso corporal deles. Essa análise é de extrema importância, uma vez

que o colesterol desempenha um papel crítico na avaliação da saúde cardiovascular.

Além disso, aprofundar o conhecimento sobre os pacientes da clínica, incluindo informações como idade, estado civil, nível de educação e renda, permite que a equipe de saúde pública planeje estratégias de prevenção de doenças de maneira mais eficaz. Por meio da análise dos dados demográficos, é possível identificar grupos de risco, personalizar campanhas de conscientização e fornecer orientações adaptadas a estilos de vida saudáveis.

Sua atuação como Cientista de Dados desempenha um papel crítico na obtenção de insights valiosos para a saúde dos pacientes. Isso possibilita o desenvolvimento de iniciativas de prevenção direcionadas e eficazes, que não apenas melhoram a qualidade de vida dos pacientes, mas também reduzem os custos associados ao tratamento de doenças cardiovasculares e aprimoram a eficácia geral dos serviços de saúde.

Atividades do enunciado

1. Coletar e compilar os dados fornecidos, que abrangem informações de pacientes, dados médicos e estados.
2. Realizar uma análise minuciosa dos dados, explorando suas características, distribuições e relações.
3. Identificar e corrigir eventuais problemas ou inconsistências nos dados, garantindo sua integridade.
4. Extrair insights valiosos dos dados, revelando padrões, tendências e informações relevantes para a saúde dos pacientes.

5. Desenvolver um modelo de regressão linear, utilizando as variáveis de peso e colesterol, a fim de estimar os níveis de colesterol com base no peso corporal.

6. Essas atividades são essenciais para melhorar a compreensão da saúde dos pacientes, possibilitar a criação de iniciativas de prevenção direcionadas e promover um atendimento de saúde mais eficaz.

Os alunos deverão desempenhar as seguintes atividades:

1. Coletar dados dos datasets:
    a. dados_pacientes.csv
    b. dados_medicos.csv
    c. estados_brasileiros.csv
2. Analisar os dados coletados;
3. Tratar os dados coletados;
4. Avaliar dados ausentes nas colunas;
5. Criar algoritmo de regressão linear;
6. Responder às questões práticas do desafio.

Dicas do professor

1. Analise com cuidado os dados.
2. Antes de enviar as respostas, verifique se o gabarito está correto.
3. Realize todas as manipulações e junções dos dados antes de responder às questões do desafio.

4. Separe o conjunto de dados em treinamento e teste e avalie os resultados baseado nos dados de teste e predição.
    a. Utilize os seguintes parâmetros: test_size=0.2, random_state=0
5. Atenção no momento de filtrar e corrigir dados (se necessário).
6. Tenha atenção no que pede cada questão.
7. Os dados disponibilizados nos datasets são fictícios. Ou seja, não têm relação com o mundo real.
8. Siga fielmente todos os passos contidos no enunciado das questões.
9. Os datasets utilizados no desafio podem ser obtidos no link:

    https://github.com/ProfLeandroLessa/classroom-datasets/tree/master/CID/Desafio%20Final
10. Abaixo segue as versões utilizadas das bibliotecas neste trabalho.

Atenção! Para garantir a obtenção dos mesmos resultados do projeto, é recomendável o uso das mesmas versões das bibliotecas

```
VERSÕES DAS BIBLIOTECAS UTILIZADAS:
Pandas = 1.5.2
Sklearn = 1.2.0
```

PARA TRATAMENTO DE DADOS

Avalie se será necessário realizar tratamento de dados ausentes nos datasets disponibilizados.

Instruções para correção de dados ausentes:

1. Média arredondada para 2 casas decimais para as variáveis do tipo numéricas;

2. Moda para as variáveis categóricas.

Acredito no potencial de todos vocês!

Bom desafio a todos!