Bootcamp: Arquiteto(a) Big Data

Desafio Prático

Módulo 1º: Fundamentos de Big Data

Objetivos de Ensino

Exercitar os seguintes conceitos trabalhados no Módulo:

1. Coleta de dados.
2. Analisar e realizar tratamento de dados.
3. Criar visualização de dados.
4. Implementar algoritmo de Machine Learning.
5. Analisar resultados obtidos.
6. Conhecimento teórico ministrado nas vídeo aulas.

Enunciado

Como Arquiteto de Big Data em uma equipe de saúde, você faz parte de um projeto cujo objetivo é identificar padrões e riscos de saúde em pacientes, utilizando suas informações demográficas e biomarcadores específicos. Neste estudo, nossa análise concentra-se nas variáveis de peso e colesterol, com a finalidade de avaliar o risco de desenvolvimento de problemas cardiovasculares.

Para atingir tal propósito, optamos por aplicar o algoritmo de agrupamento k-means aos dados de peso e colesterol, segmentando os pacientes em três grupos distintos. O intuito é discernir padrões nesses dados que possam indicar o risco de um paciente desenvolver doenças cardiovasculares. Identificar esses riscos precocemente é fundamental para possibilitar intervenções adequadas e tratamentos personalizados, visando prevenir complicações graves.

Os três clusters identificados neste estudo são:

Baixo Risco: este grupo engloba pacientes cujas características de peso e colesterol estão associadas a um risco relativamente baixo de desenvolvimento de problemas cardiovasculares. Podem ter um perfil de peso saudável e níveis de colesterol dentro da faixa normal.

Risco Moderado: aqui, encontram-se pacientes que apresentam algumas características indicativas de um risco moderado de problemas cardiovasculares, com uma combinação de fatores como peso ligeiramente acima do ideal e níveis de colesterol elevados, mas ainda dentro de limites considerados moderados.

Risco Alto: já neste grupo, abrange pacientes com características que sugerem um risco significativamente elevado de desenvolvimento de problemas cardiovasculares, como excesso de peso ou obesidade e níveis de colesterol muito acima dos limites recomendados.

Através dessa análise, buscamos fornecer informações valiosas para a equipe de saúde, embasando decisões mais assertivas em relação aos pacientes e implementando medidas preventivas e personalizadas. Assim, almejamos promover uma melhor qualidade de vida e reduzir os riscos cardiovasculares.

ATENÇÃO PARA TRATAMENTO DE DADOS

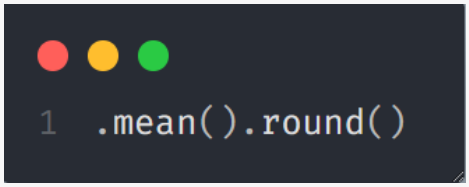
Avalie se será necessário realizar tratamento de dados ausentes nos datasets disponibilizados.

Instruções para correção de dados ausentes

Exclua todos os dados ausentes da base de dados clínicos

Para as base de estados e dados de paciente utilize:

- Dados numéricos: média arredondada



- Dados categóricos: moda da variável analisada.

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

```
Versões utilizadas no trabalho:
pandas: 1.5.2
seaborn: 0.12.1
matplotlib: 3.6.2
sklearn: 1.2.0
plotly: 5.11.0
```

É crucial reconhecer que a linguagem de programação Python e suas bibliotecas associadas estão em constante evolução. Como resultado, pode ocorrer que funções ou métodos específicos, que costumavam estar disponíveis em versões anteriores, deixem de existir ou passem a ser implementados de maneira diferente em versões mais recentes.

Essas atualizações são realizadas para melhorar a eficiência, corrigir erros e fornecer novos recursos aos desenvolvedores. No entanto, essa dinâmica de mudança também pode criar desafios, especialmente quando se trabalha com código legado ou ao compartilhar código com outros membros da equipe. Portanto, é de extrema importância que os alunos estejam cientes dessas mudanças e estejam dispostos a se adaptar a elas.

Acredito no seu potencial.

Bom desafio!

Atividades

Para essa atividade, os alunos deverão criar um algoritmo de clusterização para agrupar os dados de pacientes baseado em seu peso e colesterol...

Criar um projeto no google drive.

1. Coletar e inserir os seguintes arquivos:

a. dados_clinicos.zip.

b. estado_regiao.zip

c. dados_pacientes.zip

Os arquivos se encontram em: https://leandrolessa.com.br/datasets/

2. Analisar os dados coletados.
3. Avaliar a relação entre as variáveis.
4. Criar algoritmo de clusterização k-means.

a. Utilize k-means = 3, Random state = 42, i_init=10

5. Responder as questões teóricas e práticas do trabalho.

Dicas do professor:

1. Antes de prosseguir com qualquer análise ou desenvolvimento de algoritmos de machine learning, é essencial compreender profundamente a base de dados e realizar uma verificação minuciosa dos dados disponíveis.
2. Elimine os dados duplicados caso necessário.

3. Corrija dados ausentes caso necessário.
4. Realize o tratamento de dados antes de responderem as questões.
5. Realize a integração dos dados através dos identificadores de cada uma das bases.
6. Após realizar a integração dos dados ordene o dataframe pelo id_cliente
7. Analisem bem o gráfico gerado e a disponibilização dos dados.
8. Antes de enviar as respostas verifiquem se o gabarito está correto.
9. Tenham atenção no que pede cada questão.
10. Para obter mais orientações sobre como criar um gráfico de dispersão solicitado em uma questão, siga estas dicas adicionais:

a. https://leandrolessa.com.br/tutoriais/grafico-de-dispersao-como-criar-e-analisar-na-pratica/

11. Dicas para realizar integração de dados:

a. https://leandrolessa.com.br/tutoriais/integracao-de-dados-descubra-os-4-tipos-de-joins-essenciais/