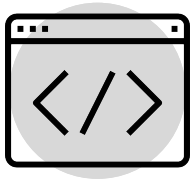
escola britânica de artes criativas & tecnologia

Profissão: Cientista de Dados

escola britânica de artes criativas & tecnologia

BOAS PRÁTICAS

Limpeza e preparação de dados



- Identifique e trate dados ausentes
- Renomeie índices e colunas
- Categorização e dummies
- Compreenda a amostragem
- Junte tabelas

Identifique e trate dados ausentes

- Sempre verifique a presença de dados ausentes em seu conjunto de dados. Isso pode ser feito usando os métodos isna() ou isnull() em Python.
- Antes de remover linhas com dados ausentes usando o método "dropna", certifique-se de que isso não irá distorcer seus dados. Às vezes, a ausência de dados pode ser informativa.
- Ao lidar com dados ausentes, considere várias técnicas como preenchê-los com a média ou a mediana dos dados existentes, substituir os dados ausentes por zero usando o método "fillna", ou usar os métodos "bfill" e "ffill", que preenchem os dados ausentes com o valor seguinte ou anterior, respectivamente.

Identifique e trate dados ausentes

- Ao usar métodos como "dropna" ou "drop_duplicates", lembre-se de que eles não alteram os dados originais a menos que o parâmetro "inplace" seja definido como verdadeiro.
- Sempre verifique a consistência dos dados após o tratamento de dados ausentes ou duplicados.
- Use o método "map" para mapear valores em uma coluna para outros valores. Isso pode ser útil para transformar dados categóricos em numéricos, por exemplo.
- Lembre-se de que o tratamento adequado de dados ausentes e duplicados é essencial para a qualidade do seu modelo de Machine Learning.

Renomeie índices e colunas

- Use o método 'reset_index' para redefinir os índices para seus valores padrão quando necessário. Lembre-se de que isso retorna uma cópia do DataFrame original por padrão.
- Evite alterar os dados originais sempre que possível. Em vez disso, trabalhe com cópias dos dados para preservar os dados originais.
- Ao fazer alterações em um DataFrame, como renomear colunas ou índices, lembre-se de que essas alterações não são aplicadas ao DataFrame original a menos que você passe o parâmetro 'inplace' como True.
- Sempre verifique os dados após fazer alterações para garantir que as alterações foram aplicadas corretamente.

Categorização e dummies

- Ao transformar variáveis numéricas em categóricas, tenha em mente o contexto e o significado dos dados. Por exemplo, ao categorizar o IMC, use faixas que são significativas do ponto de vista médico.
- Ao lidar com variáveis categóricas em uma análise que requer variáveis numéricas, considere a criação de variáveis dummy. Lembre-se de que cada categoria se tornará uma nova coluna, então cuidado com o aumento da dimensionalidade dos dados.

Categorização e dummies

- Ao criar variáveis dummy, use ferramentas que automatizem o processo, como a função get_dummies do pandas. Isso pode economizar muito tempo e evitar erros.
- Após a categorização e a criação de variáveis dummy, lembre-se de concatenar os resultados em um único DataFrame. Isso facilitará a análise subsequente.
- Ao categorizar com base na distribuição dos dados, considere o uso de quartis ou outra medida de posição. Isso pode ajudar a criar categorias que refletem a distribuição dos dados.

Compreenda a amostragem

- Ao trabalhar com grandes conjuntos de dados, a amostragem é uma prática útil. Use métodos como 'head', 'tail' e 'sample' para selecionar subconjuntos de seus dados.
- Ao trabalhar com strings, esteja ciente da diferença entre maiúsculas e minúsculas. Funções como 'upper', 'lower' e 'replace' podem ser úteis para manipular strings.
- Ao usar o método 'sample' para selecionar uma amostra aleatória, considere o uso do parâmetro 'random_state'. Isso permite que você defina uma semente para o gerador de números aleatórios, garantindo que você obtenha os mesmos resultados cada vez que você executa o código.

Junte tabelas

- Ao juntar tabelas, é importante entender qual tipo de junção é mais adequado para o seu conjunto de dados e para a análise que você deseja realizar. Os diferentes tipos de junções (Left, Right, Inner, Outer) oferecem diferentes maneiras de combinar tabelas e podem resultar em diferentes conjuntos de dados finais.

- Ao usar o método 'merge' para realizar uma junção, certifique-se de especificar corretamente a chave comum entre as tabelas. Isso garantirá que os dados sejam combinados corretamente.

Junte tabelas

- O método 'append' é um atalho para o método 'concat' ao longo do eixo das linhas. Ele pode ser útil quando você deseja adicionar rapidamente novas linhas a uma tabela existente.
- Ao usar o método 'concat', lembre-se de que ele junta as tabelas pelo índice, colocando uma tabela embaixo da outra. Se os índices das suas tabelas não são relevantes para a sua análise, este pode ser um método útil para combinar tabelas.

escola britânica de artes criativas & tecnologia

Bons estudos!