



Microsoft Power BI Para Business Intelligence e Data Science

Microsoft Power BI Para Business Intelligence e Data Science

O Que Exatamente Fazemos em Limpeza e Manipulação de Dados?

A limpeza e manipulação de dados são etapas cruciais no processo de análise e modelagem de dados. Essas atividades envolvem a organização, transformação e remoção de erros ou inconsistências nos dados para garantir que eles estejam prontos para serem utilizados em análises, visualizações ou aplicação de modelos de aprendizado de máquina (Machine Learning). Algumas das principais tarefas de limpeza e manipulação de dados incluem:

Remoção de dados duplicados: Eliminar registros duplicados que podem distorcer a análise.

Tratamento de valores ausentes: Substituir, remover ou estimar valores ausentes nos dados, usando métodos como média, mediana, interpolação ou outros algoritmos.

Correção de erros de digitação e inconsistências: Identificar e corrigir erros de digitação, formatação e padronização dos dados.

Conversão de tipos de dados: Transformar variáveis em tipos de dados apropriados, como numérico, categórico ou textual.

Renomeação e reorganização de colunas: Ajustar os nomes das colunas para facilitar a compreensão e organizá-las de acordo com a necessidade da análise.

Filtragem e seleção de dados: Extrair subconjuntos específicos de dados com base em critérios pré-determinados, como faixas de valores ou categorias.

Discretização e binning: Converter variáveis contínuas em categorias ou agrupar dados em intervalos específicos para análise.

Normalização e padronização: Ajustar a escala dos valores numéricos para facilitar a comparação e melhorar o desempenho de modelos de aprendizado de máquina.

Transformação de variáveis: Criar novas variáveis a partir de outras existentes ou aplicar transformações matemáticas para simplificar análises ou melhorar a interpretação dos dados.

Deteção e tratamento de outliers: Identificar e tratar valores extremos que podem afetar a análise ou a modelagem.

Codificação de variáveis categóricas: Converter variáveis categóricas em formatos numéricos, como codificação one-hot ou ordinal, para serem utilizadas em modelos de aprendizado de máquina.

Essas etapas podem variar de acordo com o contexto e os objetivos da análise, e as ferramentas utilizadas para limpeza e manipulação de dados podem incluir linguagens de programação como Python, R e SQL, bem como software específico para análise de dados, como Excel, Power BI ou Tableau.

Todas essas técnicas são estudadas em detalhes nos cursos aqui na DSA e neste capítulo vamos trazer uma breve introdução ao tema abordando 3 das técnicas mencionadas acima.