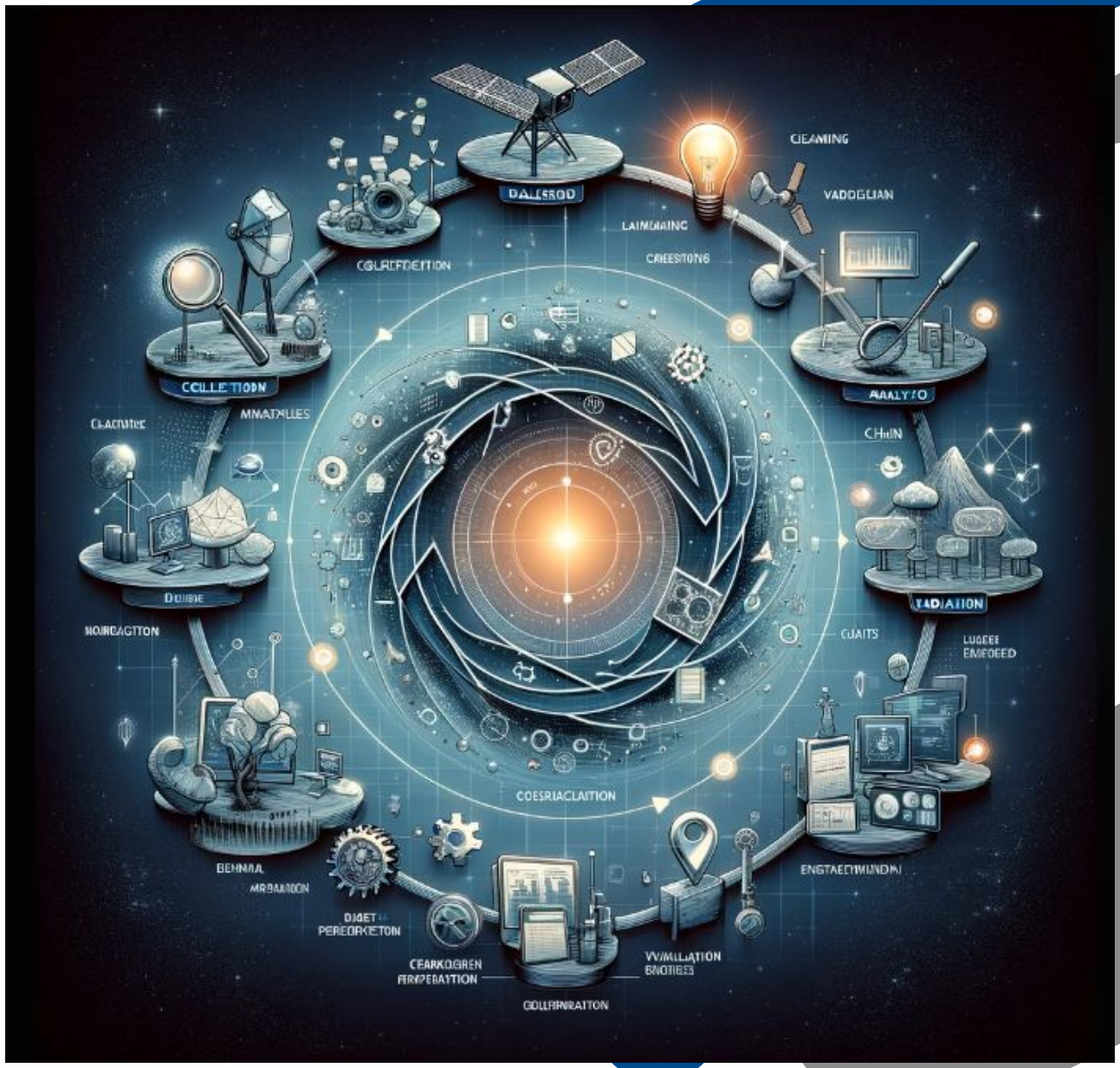
Senac

Fecomércio Sesc

Data Science – Princípios e Técnicas



Abril

2025

Projeto de Ciência de dados

Data Science

Em geral os projetos de data science envolvem as seguintes etapas:

1. Entendimento do negócio: aqui são definidas as perguntas, o objetivo da análise de dados e o plano a ser seguido;

2. Compreensão dos dados: etapa utilizada para coletar e explorar os dados, aumentando a compreensão sobre sua estrutura, atributos e contexto;

Data Science

3. **Preparação dos dados:** após a análise exploratória, inicia-se o processo de limpeza, filtragem, estruturação, redução e integração dos dados;

4. **Modelagem dos dados:** envolve as tarefas de seleção dos dados, definição e construção do modelo;

Data Science

5. **Validação do modelo:** os resultados gerados pelo modelo são avaliados, para verificar se a precisão obtida está satisfatória e coesa;

6. **Utilização do modelo:** após serem validados, os resultados dos modelos são utilizados e monitorados.

Data Science

Também vimos que: Os dados utilizados estão normalmente “sujos”

Muitas vezes, as bases contêm dados incompletos, inconsistentes, corrompidos, duplicados ou em formatos inadequados, entre outros problemas. Portanto, é essencial a intervenção de um profissional capacitado para tratar esses dados antes de iniciar a análise propriamente dita.

Data Science

A crescente geração de dados amplifica a necessidade de limpeza de dados eficiente, visto que a má qualidade dos dados implica custos significativos para as empresas, chegando a US$ 12,9 milhões anualmente, conforme indicações da Gartner.

Data Science

Além do impacto financeiro, dados imprecisos demandam dos cientistas de dados até 60% do seu tempo em tarefas de limpeza e organização, ressaltando a importância de adotar ferramentas e estratégias proativas para assegurar a qualidade dos dados desde sua coleta.

Data Science

Mas como resolvemos esse problema?

Com o processo de limpeza de dados!

Data Science

O que é Limpeza de dados?

A limpeza de dados, essencial no gerenciamento de qualidade, envolve identificar e corrigir erros ou inconsistências nos dados para garantir sua precisão, consistência e utilidade.

Data Science

Esse processo é crucial pois o uso de dados brutos, sem tratamento, pode levar a decisões baseadas em informações imprecisas, prejudicando a eficácia operacional e estratégica das organizações.

Data Science

Mas por que não podemos usar dados brutos em vez de gastar tanto tempo na limpeza de dados?

Data Science

- **Entradas com erros ortográficos:** Erros de digitação e ortografia podem levar a erros de categorização.
- **Formatos inconsistentes:** Datas, números ou categorias podem ser representados de forma diferente no mesmo conjunto de dados.
- **Valores discrepantes e erros:** Entradas incomuns ou erradas podem levar a análises imprecisas.

Data Science

- **Registros duplicados:** Dados redundantes podem levar a estatísticas e conclusões imprecisas.
- **Valores nulos ou ausentes:** Dados incompletos podem levar a lacunas na análise e a insights imprecisos e/ou limitados.
- **Dados imprecisos:** Informações incorretas ou desatualizadas podem levar a decisões imprecisas.

Data Science

- **Unidades não padronizadas:** Diferentes unidades de medida podem criar problemas de inconsistência de dados, especialmente ao comparar ou agregar dados.
- **Dados incompatíveis:** Dados conflitantes de fontes diferentes podem causar discrepâncias integração de dados e análise.

Data Science

Principais técnicas de limpeza de dados

Removendo duplicatas:

Como fazer: Utilize algoritmos para identificar e remover linhas duplicadas com base em atributos vitais selecionados.

Data Science

Tratamento de dados ausentes:

Como fazer: As opções incluem imputação, exclusão ou uso de algoritmos que podem lidar com valores ausentes. A imputação pode usar estratégias baseadas em média, mediana ou modelo, como k-NN.

Data Science

Corrigindo dados incorretos:

Como fazer: verificações de consistência e revisão manual, se necessário. Ferramentas de preparação de dados podem ajudar na correspondência de padrões e correções.

Data Science

Tratamento de valores discrepantes:

Como fazer: identifique valores discrepantes por meio de métodos estatísticos, como pontuação Z ou IQR, e decida se deseja limitá-los, transformá-los ou removê-los.

Data Science

Normalizando Dados:

Como fazer: aplique técnicas como dimensionamento mínimo-máximo, normalização de pontuação Z ou transformações de log.

Data Science

Validando a consistência dos dados:

Como fazer: Crie regras de validação para verificar relacionamentos e consistência entre atributos.

Data Science

Transformando Dados:

Como fazer: usar transformações de dados como codificação de dados categóricos ou criação de termos de interação com base em necessidades analíticas.

Data Science

Mais detalhes:

https://www.astera.com/pt/type/blog/data-cleansing/

Um pouco de estatística

Data Science

Nos últimos anos, houve um crescimento exponencial no volume de dados gerados pela humanidade, o que gerou uma demanda crescente por profissionais capazes de extrair informações e tomar decisões fundamentadas com base nesses dados.

Data Science

Um aspecto crucial ao trabalhar com dados é a habilidade de descrevê-los, resumi-los e representá-los visualmente. A estatística descritiva é uma ferramenta essencial nesse processo, utilizando duas abordagens principais:

Data Science

- **A abordagem quantitativa**, que descreve e resume os dados numericamente.

- **A abordagem visual**, que ilustra os dados por meio de gráficos e visualizações.

Data Science

Na análise quantitativa, destacamos:

- A **tendência central** informa sobre os centros dos dados. Medidas úteis incluem a média, mediana e moda.

- A **variabilidade** informa sobre a dispersão dos dados. Medidas úteis incluem variância e desvio padrão.

Data Science

- A **correlação** (ou **variabilidade conjunta**) informa sobre a relação entre um par de variáveis em um conjunto de dados. Medidas úteis incluem a covariância e o coeficiente de correlação.

As medidas de tendência central

Data Science

Antes de começarmos a estudar o princípios de estatísticas baixe o seu projeto e crie um banco sqlite localmente.

Data Science

Média:

A média aritmética, ou simplesmente média, de um conjunto de valores é a medida de centro encontrada somando todos os valores do conjunto e dividindo pelo número de valores. Assim:

$$Média = \frac{soma\ dos\ valores}{total\ de\ observações}$$

Data Science

Exemplo 1: Determine a média de um dos atributos no banco.

Data Science

Obs: Existem outras médias, porém cada uma delas é usada em situações específicas. A saber:

- **Média Ponderada**: Você deve usar uma média ponderada quando deseja atribuir mais importância a alguns números em um conjunto de dados do que a outros. Isso é útil em cenários onde um evento pode ter vários resultados positivos ou negativos, e a magnitude desses resultados varia.

Data Science

- **Média Harmônica:** A média harmônica é calculada como o número de valores dividido pela soma do inverso de cada valor. É apropriada quando os dados representam grandezas que são inversamente proporcionais, como taxas.

Data Science

- **Média Geométrica:** A média geométrica é calculada como a raiz N-ésima do produto de todos os valores, onde N é o número de valores. É útil quando os dados estão em uma escala multiplicativa, como em situações envolvendo crescimento ou taxa de variação entre diferentes unidades de medida.

Mediana

A **mediana** da amostra é o elemento central de um conjunto de dados ordenado (crescente ou decrescente). Se o número de elementos n do conjunto de dados for ímpar, então a mediana é o valor na posição do meio. Se n for par, então a mediana é a média aritmética dos dois valores no meio

Data Science

Exemplo 2: Encontre a mediana de um dos atributos no banco.

Data Science

Importante: A principal diferença entre o comportamento da média e da mediana está relacionada aos valores extremos (outliers) do conjunto de dados. De uma maneira geral:

Data Science

- Se você adicionar um valor discrepante maior do que a média em um conjunto de dados, a média aumentará, mas o valor da mediana vai sofrer pouca influência.

- Se você remover um valor discrepante de um conjunto de dados, a média diminuirá, mas o valor da mediana vai sofrer pouca influência.

Data Science

Agora vamos dividir o conjunto original em subconjuntos com características específicas e observar o que acontece com a média. Esse processo, conhecido como **estratificação**, consiste em separar os dados em subgrupos (ou estratos) que compartilham características semelhantes.

Data Science

Ao analisar cada estrato individualmente, podemos identificar variações internas e entender como cada subgrupo contribui para o comportamento geral dos dados, resultando em análises mais detalhadas e representativas.

Data Science

Exemplo 3: A média é mantida quando analisamos tipos de cartões diferentes?

Moda

A **moda** da amostra é o valor no conjunto de dados que ocorre com mais frequência. Se não houver um único valor desse tipo, o conjunto será multimodal, pois possui vários valores modais.

Data Science

Exemplo 4:

Medidas de Localização

O percentil p da amostra é o elemento no conjunto de dados tal que p% dos elementos no conjunto de dados são menores ou iguais a esse valor. Além disso, (100 − p)% dos elementos são maiores ou iguais a esse valor. Se houver dois desses elementos no conjunto de dados, o percentil p da amostra é a média aritmética deles.

Data Science

- O primeiro quartil Q1 é o percentil 25 da amostra. Ele divide aproximadamente 25% dos menores itens do restante do conjunto de dados.

- O segundo quartil Q2 é o percentil 50 da amostra, também conhecido como a mediana. Aproximadamente 25% dos itens situam-se entre o primeiro e o segundo quartis, e outros 25% entre o segundo e o terceiro quartil.

Data Science

- O terceiro quartil Q3 é o percentil 75 da amostra. Ele divide aproximadamente 25% dos maiores itens do restante do conjunto de dados.

Data Science

Exemplo 6:

Dúvidas?

Marco Mialaret, MSc

Telefone:
81 98160 7018

E-mail:
marcomialaret@gmail.com