O ciclo de análise de dados

Um roteiro completo para resolver problemas do dia a dia

Análise Macro

Índice

Bem vindo(a)!

Seja bem vindo(a) ao ebook **O ciclo de análise de dados** produzido pela **Análise Macro**.

A demanda por análise de dados é cada vez mais crescente nas empresas, no governo e na academia. Existem muitos problemas que podem ser resolvidos usando a união correta de técnicas, habilidades, dados e ferramentas no dia a dia dos profissionais da área. Nesse ebook você vai aprender sobre um roteiro separado em etapas que agilizam o processo de analisar dados, facilitando a entrega de soluções baseadas em dados. Aprenda a metodologia de trabalho, as aplicações e exemplos de uso de análise de dados no mundo contemporâneo.

Você pode ler a primeira versão deste ebook online, em PDF ou no formato epub.

1 Introdução

Resolver problemas é uma tarefa central para quem trabalha na área de Dados, especialmente em análise de dados.

O papel do analista é utilizar suas habilidades em estatística, matemática, programação e outras, além da expertise da área, para resolver um problema utilizando dados.

Mas qual problema o Fernando, um analista de dados, vai resolver?

Definir o problema a ser resolvido e os objetivos da análise de dados é o primeiro passo fundamental para desenvolver um trabalho bem sucedido.

Sem saber o que é necessário resolver é difícil que qualquer solução desenvolvida atinja e resolva o problema.

É preciso muita sorte para produzir boas análises de dados sem se guiar por um propósito claro.

Portanto, o primeiro passo é utilizar uma metodologia com processos e etapas bem definidas para analisar dados.

Neste ebook vamos descrever que metodologia para analisar dados é essa, quais são as etapas gerais e como elas funcionam para resolver problemas reais.

Sem uma metodologia de trabalho é seguro dizer que o analista de dados está perdido numa selva de ferramentas, modelos e dados, lutando para sobreviver e tentando qualquer coisa a todo momento e a qualquer custo.

Ao seguir uma metodologia de trabalho, o analista estará guiado por uma bússola, o que diminui as chances de se perder no caminho e garante consistência de resultados no longo prazo.

2 O ciclo

O que chamamos de ciclo de análise de dados é uma metodologia de trabalho para otimizar e guiar o processo de analisar dados, desde a definição do problema a ser resolvido até a implementação da solução baseada em dados.

É um ciclo porque, na prática, resolver problemas com dados não é como caminhar em linha reta do ponto A até o ponto B.

O dia a dia de análise de dados é cheio de idas e vindas, tentativas e erros, pois há muitas pedras no caminho e algumas são difíceis de ultrapassar.

Algumas, dentre várias, dessas pedras no caminho de um analista de dados são:

- Dados indisponíveis;
- Dados incorretos;
- Dados ausentes;
- Objetivos e problemas indefinidos;
- Entre outros.

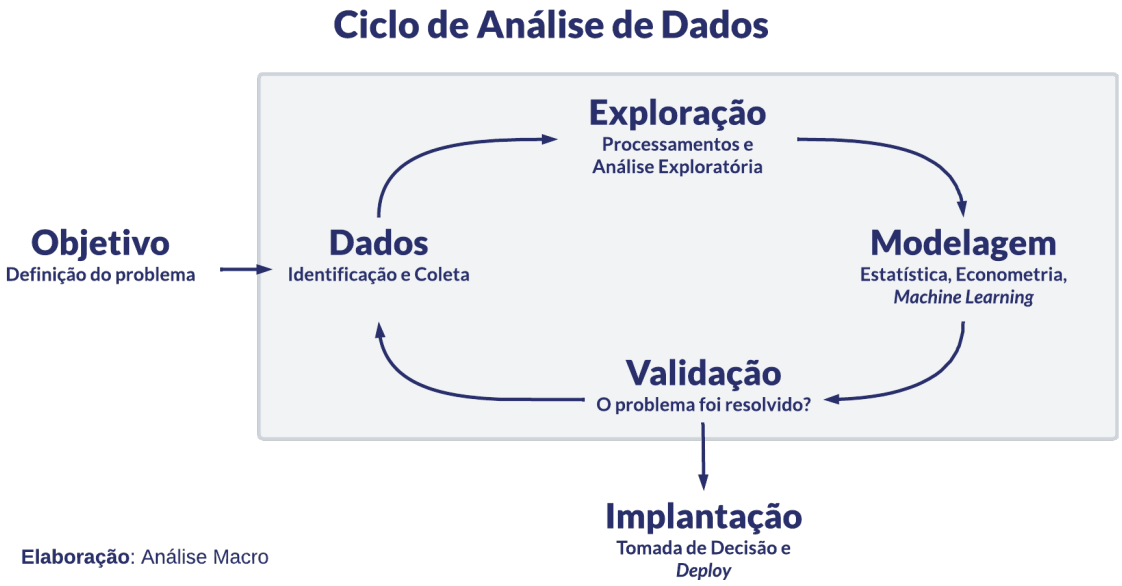
Alguns destes obstáculos para analisar dados podem ser melhor contornados se houver uma visão clara do caminho a ser percorrido.

Dessa forma, o ciclo de análise de dados é como um mapa que o analista pode utilizar para pegar um problema, analisar os dados e entregar uma solução.

Entender a fundo o ciclo de análise de dados é fundamental para conseguir entregar soluções e informações a partir de dados.

Portanto, um analista de dados deve ser capaz de mapear mentalmente, dado um contexto, essas etapas para desenvolver uma solução a partir de dados.

Vamos dar uma olhada nas etapas?



2.1 Objetivo

É a primeira etapa de um projeto de análise de dados, onde há um contexto/situação na área de atuação do analista de onde surge um problema a ser resolvido.

É papel do analista de dados, com apoio de outros atores envolvidos, identificar esse problema de forma clara para prosseguir com uma solução analítica de dados com determinados objetivos.

Por exemplo:

Você é analista de dados na Netflix e o setor que monitora o engajamento do usuário (tempo de uso, nº de títulos assistidos, etc.) no serviço de streaming percebe uma queda em várias métricas, o que pode ser um prenúncio de cancelamento de assinaturas.

Nesse caso o problema é a queda de engajamento e o objetivo poderia ser aumentar o engajamento com vistas a evitar cancelamento de assinaturas.

Nessa etapa é fundamental a expertise de negócio para definir o problema e os objetivos do projeto de análise de dados, além de ser importante habilidades de comunicação interpessoal para contato com outras pessoas técnicas e não-técnicas.

2.2 Dados

É a segunda etapa de um projeto de análise de dados, onde o objetivo é, a partir de um problema definido, identificar quais dados podem ser úteis para o desenvolvimento de uma solução.

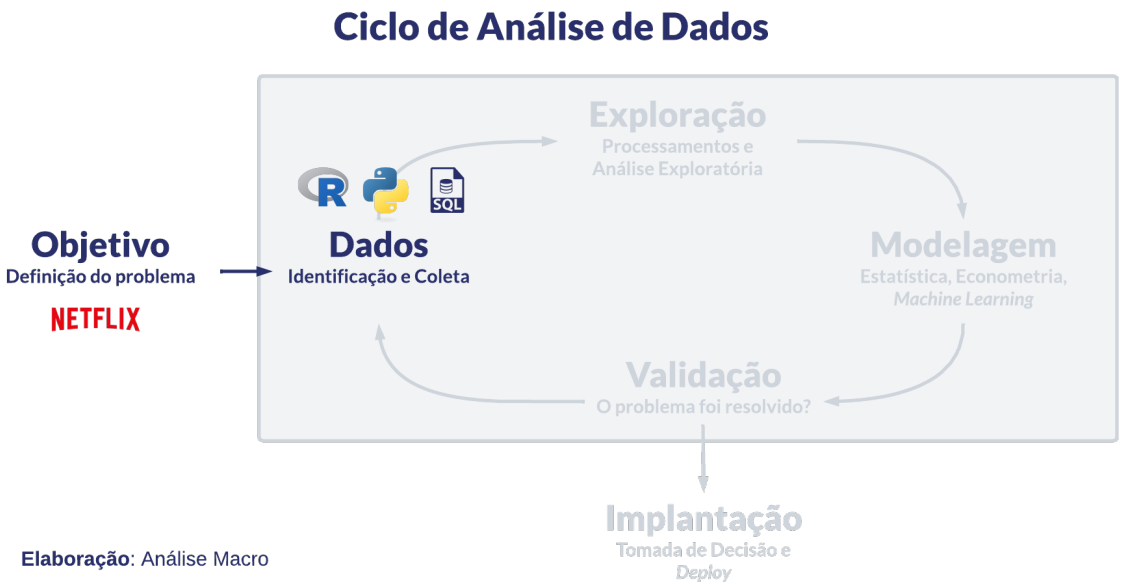
Os dados podem ser disponibilizados internamente ou externamente, portanto essa etapa também compreende os procedimentos de coleta dos dados necessários.

No exemplo de queda de engajamento de usuários da Netflix, o analista de dados poderia coletar internamente dados históricos de tempo de uso, horas assistidas, categorias e temas de títulos assistidos, atores/diretores do título, dados socioeconômicos como região, idioma, gênero e etc. sobre os usuários.

Externamente o analista de dados poderia coletar dados dos players concorrentes do mercado, se houver suspeitas que o engajamento está sendo direcionado para outros serviços de streaming.

Nessa etapa já é necessário habilidades técnicas de programação, consultas a bancos de dados, APIs e outras para que os dados possam ser disponibilizados para análise.

Ferramentas comuns utilizadas nessa etapa são as linguagens de programação R e Python e a linguagem de consulta SQL.



2.3 Exploração

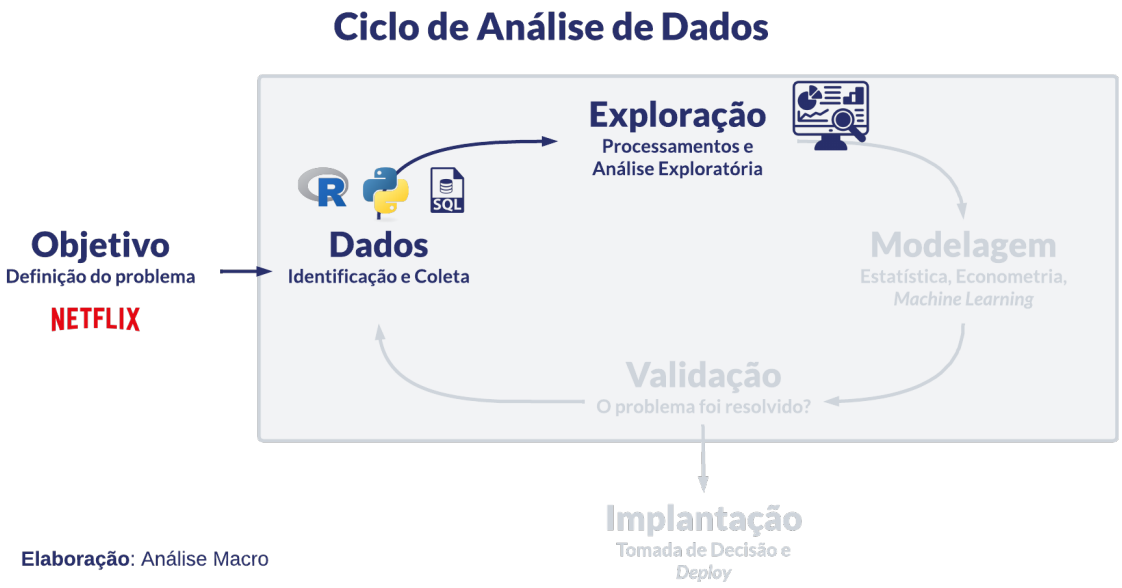
Nessa etapa da análise de dados o objetivo é compreender o que está acontecendo ou aconteceu com os dados, identificar padrões, relações e anomalias que possam servir de sinal para a escolha de uma solução do problema.

Os dados precisam estar organizados para que possam ser analisados, portanto é necessário transformar os dados brutos coletados previamente para construir uma Tabela Analítica Base (ABT, no inglês), que servirá para realizar a análise exploratória dos dados, desenvolver modelos preditivos ou construir produtos de dados como relatórios e dashboards.

No exemplo anterior da Netflix, o analista de dados poderia fazer as limpezas e cruzamentos de tabelas de dados necessárias, analisar a distribuição das variáveis, identificar a variável "alvo" (aquela que é utilizada para modelos preditivos, por exemplo), detectar valores ausentes, verificar valores extremos ou outliers, analisar correlações e autocorrelações dos dados, identificar tendências e sazonalidades, dentre outras análises que podem ser úteis.

Nessa etapa são fundamentais conhecimentos e habilidades em estatística, programação e visualização de dados.

As principais ferramentas utilizadas para essas análises são linguagens de programação como R e Python, pacotes de tratamento e exploração de dados como tidyverse e pandas e pacotes de visualização de dados como ggplot2 e matplotlib.



2.4 Modelagem

Nessa etapa o objetivo é levantar e experimentar possíveis soluções baseadas em dados para o problema identificado previamente, podendo ser:

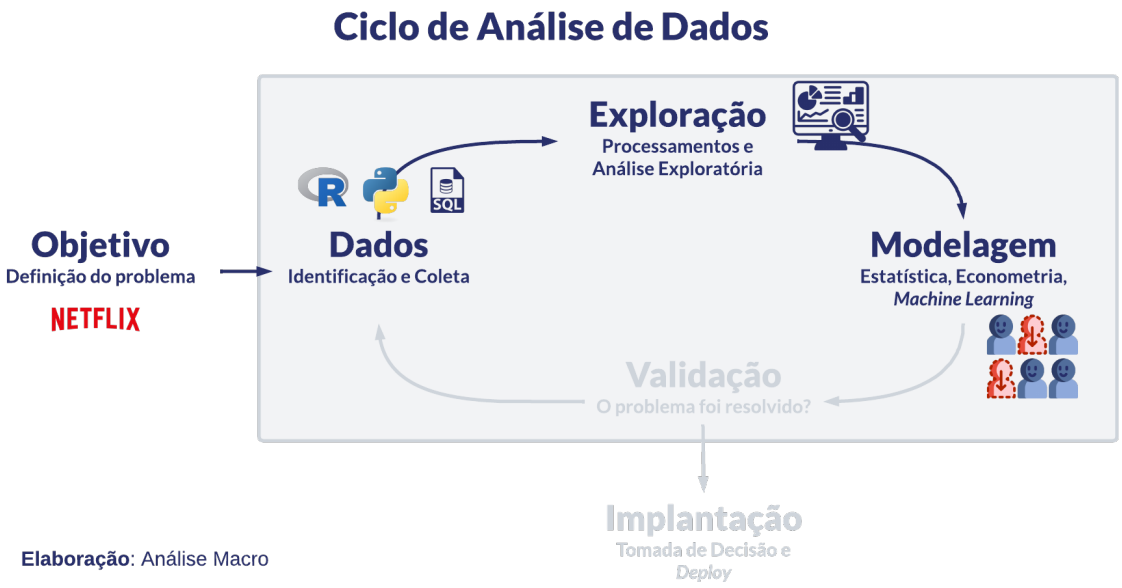
1. Simples consultas SQL para agregar e sumarizar dados e informações;
2. Análises estatísticas como testes de hipótese, análise de regressão e outras;
3. Modelos econométricos para explicar relações, produzir inferências ou previsões;
4. Modelos preditivos com técnicas de machine learning.

A técnica escolhida depende diretamente da definição do problema e dos dados escolhidos, além de ser preferível, a depender do contexto, técnicas/soluções simples e rápidas.

No mundo real o tempo custa dinheiro e implementar algoritmos complexos e avançados em produção gera uma fatura no final do mês que precisa ser paga.

No exemplo anterior da Netflix, o analista poderia focar, por exemplo, em uma solução de redução de Churn, identificando o perfil de usuários que cancelaram a assinatura e prevendo a probabilidade de ocorrer o cancelamento (risco de evasão), o que possibilita a tomada de decisão para minimizar essa evasão de usuários.

Em outras palavras, poderiam ser empregados modelos supervisionados de classificação, usando técnicas de machine learning.



Elaboração: Análise Macro

Nessa etapa é fundamental o conhecimento de uma ampla gama de técnicas estatísticas, econométricas e de machine learning, domínio de algoritmos e pacotes computacionais para implementar essas técnicas com linguagens de programação, como o R e o Python e, dependendo do contexto, conhecimento de ferramentas para processamento de Big Data.

2.5 Validação

Nessa etapa o objetivo é avaliar se a solução analítica baseada em dados é capaz de resolver o problema, podendo ser analisadas as métricas de acurácia de modelos, os resultados estatísticos e econométricos de testes ou ainda o feedback do usuário / stakeholder em caso de soluções simples, como entrega de informações e insights em relatórios/dashboards.

No exemplo anterior da Netflix, o analista poderia analisar a acurácia de diferentes modelos usando amostras de treino/teste, validação cruzada, além de verificar a importância das variáveis utilizadas.

O analista também deve ser capaz de fazer escolhas e tomar decisões sem que isso prejudique ou deturpe os resultados encontrados.

Nesta etapa é fundamental o conhecimento em amostragem de dados, interpretação estatística e programação usando linguagens como R e Python.

Elaboração: Análise Macro

2.6 Implantação

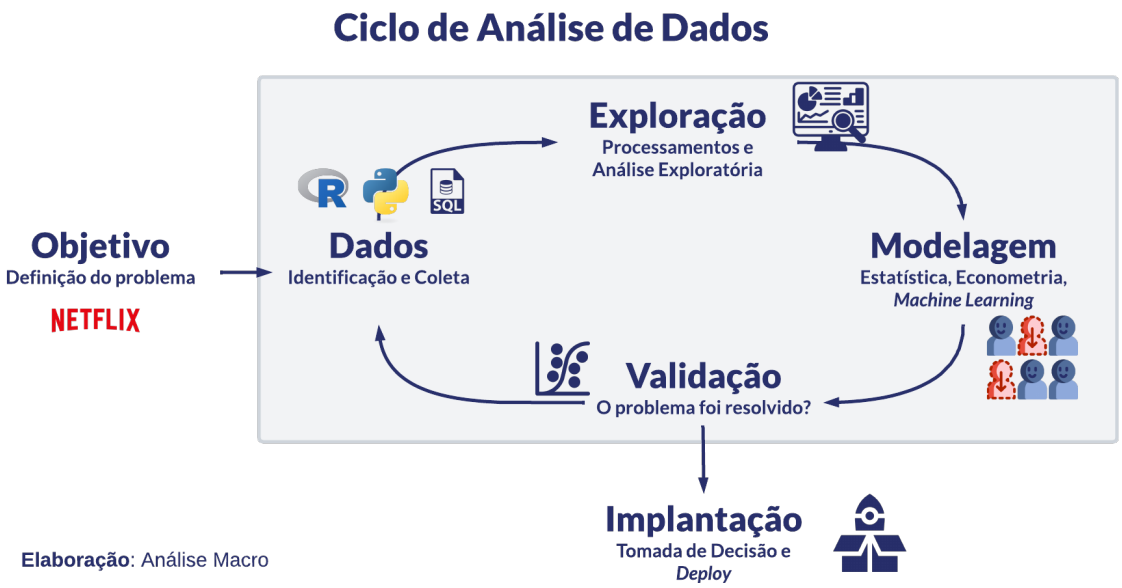
Na última etapa do ciclo de análise de dados o objetivo é comunicar os resultados do trabalho para os stakeholders e usuários, permitindo a tomada de decisão baseada em dados.

Isso pode se traduzir na implementação em ambiente de produção de um modelo preditivo, um sistema de recomendação, uma dashboard ou relatório automatizado, dentre outras possibilidades.

No exemplo anterior da Netflix, o analista poderia elaborar uma apresentação para os tomadores de decisão da companhia, permitindo a elaboração de estratégias para reter os usuários que possuem alta probabilidade de Churn.

O modelo de classificação poderia, adicionalmente, ser implementado em produção para, por exemplo, automaticamente recomendar títulos ou oferecer descontos para usuários com probabilidade de evasão.

Nesta etapa é fundamental habilidades não técnicas de comunicação interpessoal, apresentação e argumentação, além de habilidades técnicas de infraestrutura e serviços de Cloud e deploy de modelos.



Elaboração: Análise Macro

3 Conclusão

O ciclo de análise de dados é vasto e complexo, mas ao mesmo tempo é uma metodologia poderosa para solucionar problemas usando dados.

O profissional que atua ou deseja atuar na área de dados precisa de diversas habilidades e conhecimentos técnicos e não técnicos, de uma ponta até a outra do ciclo, para agregar valor em uma empresa.

Neste ebook apresentamos uma visão geral sobre o processo de análise de dados, exemplos de aplicações e uso e as habilidades e ferramentas necessárias para trabalhar na área.