ICDA6 - Projeto - Ciclo de Vida dos Dados: Messi x Cristiano Ronaldo

---

ℹ️ **Autor do projeto:** Paulo André Silveira Júnior. **Prontuário:** SP306347X.

🔄 Ciclo de vida dos dados

🚰 Produção

A produção dos dados usados neste projeto se dá a partir das partidas disputadas por Lionel Messi e Cristiano Ronaldo, dois dos jogadores profissionais de futebol de maior destaque durante os últimos 15 anos desse esporte. Esses dados foram devidamente coletados e tabulados a partir dos registros publicamente disponíveis das partidas (entidades organizadoras dos campeonatos, Wikipedia, etc).

📦 Armazenamento

O armazenamento dos dados é feito através do Kaggle, uma plataforma online voltada para a comunidade de ciência de dados.

- Kaggle - Lionel Messi
- Kaggle - Cristiano Ronaldo

🪄 Transformação

- Foi necessário padronizar o formato dos campos de data do dataset (Exemplo: DD/MMM/AA x MM-DD/AA)
- Foi necessário tratar os espaços em branco (”trim whitespaces”) em determinados campos do dataset
- Foi necessário unir os dois datasets escolhidos e criar uma nova coluna para identificar o jogador relacionado à linha em questão

---

🔍 Análise (baseada no modelo CRISP-DM)

Entendimento do negócio

O contexto de negócio do projeto é o cenário mundial de futebol profissional, um dos esportes mais populares de todos os tempos, e dois dos jogadores de maior destaque nesse âmbito: Lionel Messi e Cristiano Ronaldo. Durante a Copa do Mundo 2022

ambos estiveram envolvidos em eventos de (possível) mudança de filiação ao nível de clube (vide imagens abaixo).

Fonte: ge.globo.com

Fonte: lance.com.br

Num aspecto específico: **gols marcados em partidas oficiais entre clubes**, considerando-o como escopo deste projeto, exploraremos o seguinte problema em relação à ambos os jogadores mencionados: **eu, enquanto gestor de futebol e tomador de decisão, contrataria Messi ou Cristiano Ronaldo para o meu clube?**

Entendimento dos dados

- Número de linhas: 1398 (697 para Lionel Messi e 701 para Cristiano Ronaldo)
- Número de colunas: 13
    - Season (texto)
    - Competition (texto)
    - Matchday (texto)
    - Date (data)
    - Venue (texto)
    - Club (texto)
    - Opponent (texto)
    - Result (tempo. Exemplo: 1:00)
    - Playing_Position (texto)
    - Minute (texto)
    - At_score (tempo. Exemplo: 1:02)
    - Type (texto)
    - Goal_assist (texto)

Preparação de dados

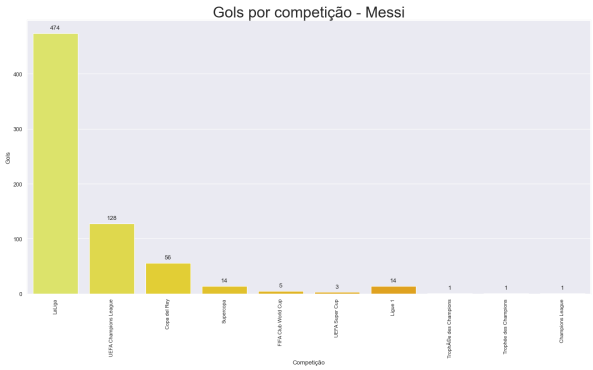
- Dado que estamos trabalhando com *apenas* 13 colunas e 1398 linhas, não optou-se por um recorte amostral dos dados disponíveis.

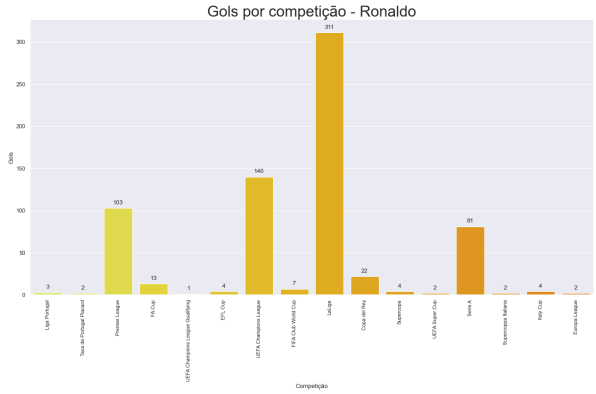
Análise explícita e análise exploratória (EDA)

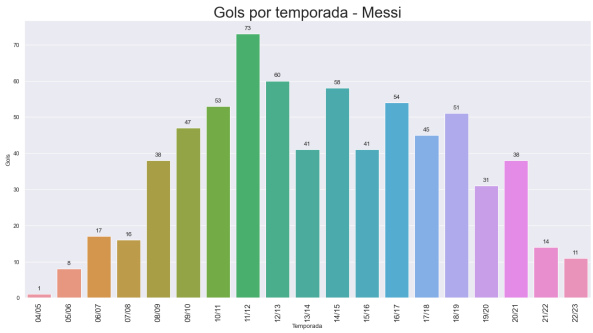
- Análise explícita
  - Junção dos dois datasets
  - Análise de dados faltantes (missing data)
    - 16 valores faltantes na coluna Type (1 Messi e 15 Cristiano Ronaldo)
    - 58 valores faltantes na coluna Playing_Position (Cristiano Ronaldo)
    - 455 valores faltantes na Coluna Goal_assist (213 Messi e 242 Cristiano Ronaldo)
  - Resumo estatístico com valores mais frequentes
    - **Messi**
      - Temporada com maior quantidade de gols marcados: temporada 11-12 com 73 gols
      - Dia com maior quantidade de gols marcados: 03/07/12 com 5 gols
      - Oponente contra quem mais marcou gols: Sevilla FC com 38 gols
      - Minuto de jogo que mais marcou gols: 55 minutos com 13 gols
      - Placar de jogo mais frequente após marcar um gol: 1 a 0 com 98 gols
      - Jogador que mais forneceu assistências para seus gols: Luis Suarez com 48 assistências
    - **Cristiano Ronaldo**
      - Temporada com maior quantidade de gols marcados: temporada 14-15 com 61 gols
      - Dia com maior quantidade de gols marcados: 09/12/15 com 5 gols
      - Oponente contra quem mais marcou gols: Sevilla FC com 27 gols
      - Minuto de jogo que mais marcou gols: 90 minutos com 17 gols

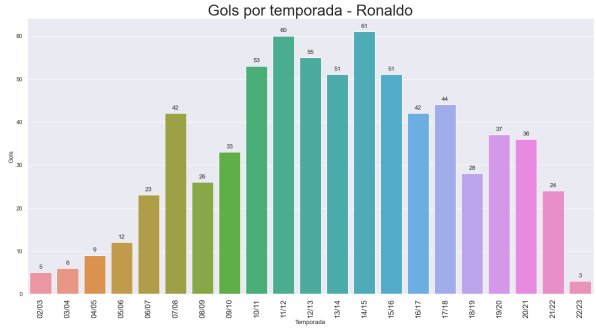
- Placar de jogo mais frequente após marcar um gol: 1 a 0 com 111 gols
- Jogador que mais forneceu assistências para seus gols: Karim Benzema com 44 assistências

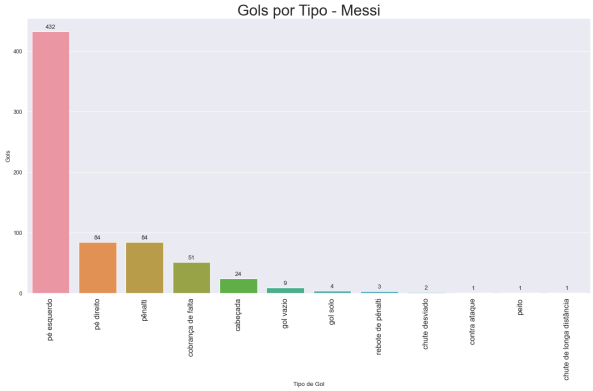
- Análise exploratória

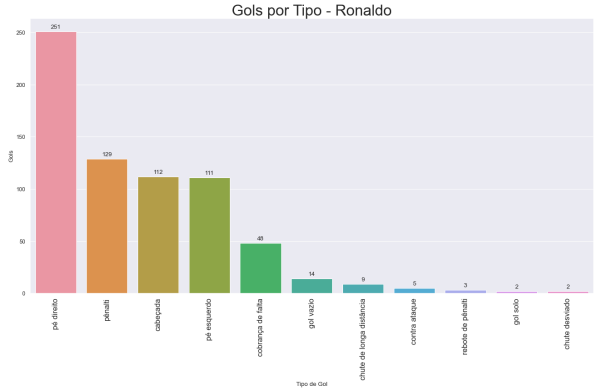






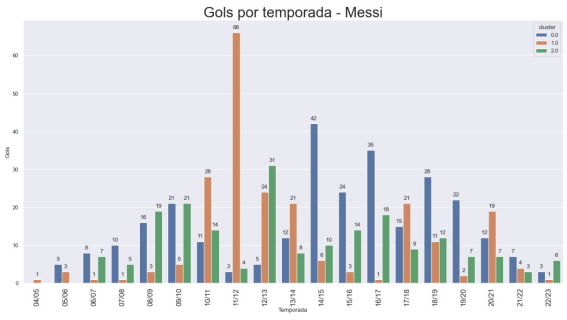


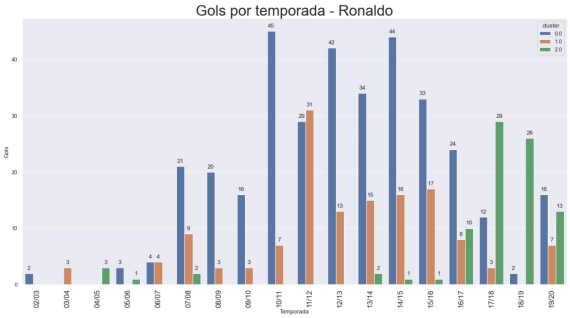


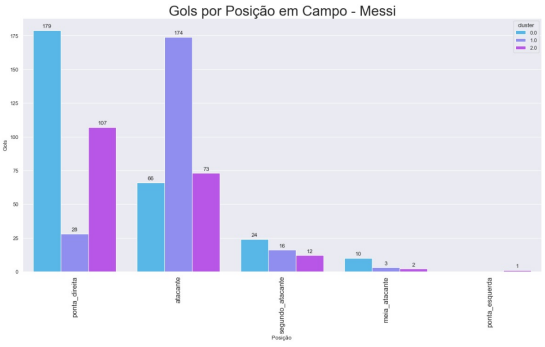


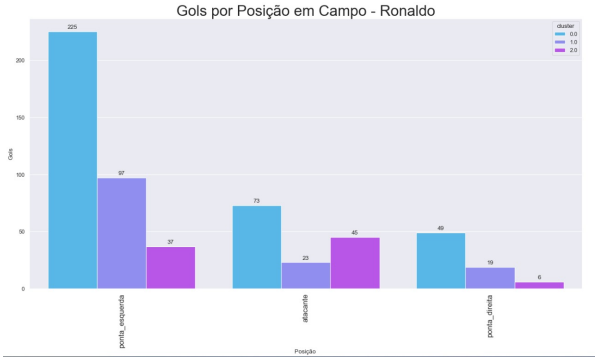
🤖 Modelagem (análise implícita)

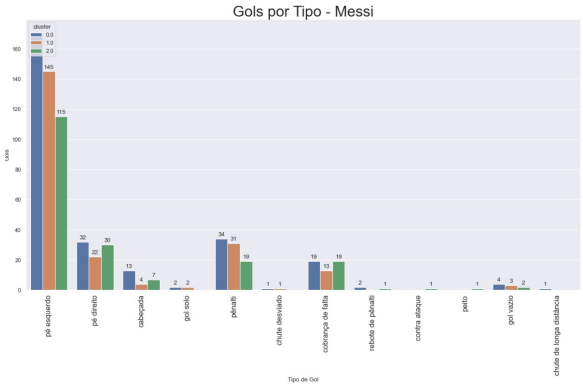
- Clustering (K-modes, variação do K-means mais adequado para trabalhar com variáveis categóricas)

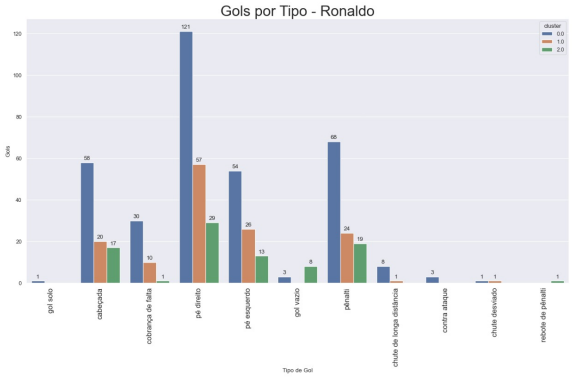












- Dashboard (R Shiny) [PENDENTE]

📝 Avaliação

- Realizamos uma tentativa de avaliar a pureza dos clusters, mas o resultado do cálculo foi 1. Isso acontece por conta da utilização do mesmo conjunto de dados da

construção do cluster e no cálculo da pureza. Isso poderia ser mitigado através de: 1. aumentando o conjunto de dados (exemplo: gerando dados através de simulações no FIFA) ou dividindo o conjunto de dados para realização do cálculo.

☁️ Implantação

Para o projeto em questão temos, em linhas gerais, duas propostas para seguir com a implantação:

- Geração de relatórios *ad hoc* para consumo dos tomadores de decisão
- Avaliação e revisão dos resultados de negócio
- Integração com plataforma de analytics (caso exista) ou implantação do arcabouço de ferramentas de analytics
    - **Exemplo:** cron job para captura periódica dos dados atualizados + pipeline de ETL → output em arquivos .csv armazenados no S3 para serem consumidos via Athena (serviços da Amazon Web Services) e Looker como ferramenta de data visualization

---

🗑️ Descarte

Considerando que estamos trabalhando sob a Open Data Commons Open Database License (ODbL) v1.0, a política de descarte dos dados usados no projeto está majoritariamente restrita pela capacidade de armazenamento disponível para a sua manutenção ao longo do tempo.