ATP Players

Apresentação Semanal [1]

UC | Projeto Aplicado a Ciência de Dados I
Docentes | Diana Mendes **&** Sérgio Moro

Grupo 2

André Silvestre	Nº104532
Diogo Catarino	Nº104745
Francisco Gomes	Nº104944
Rita Matos	Nº104936

CDB1

CRISP-DM

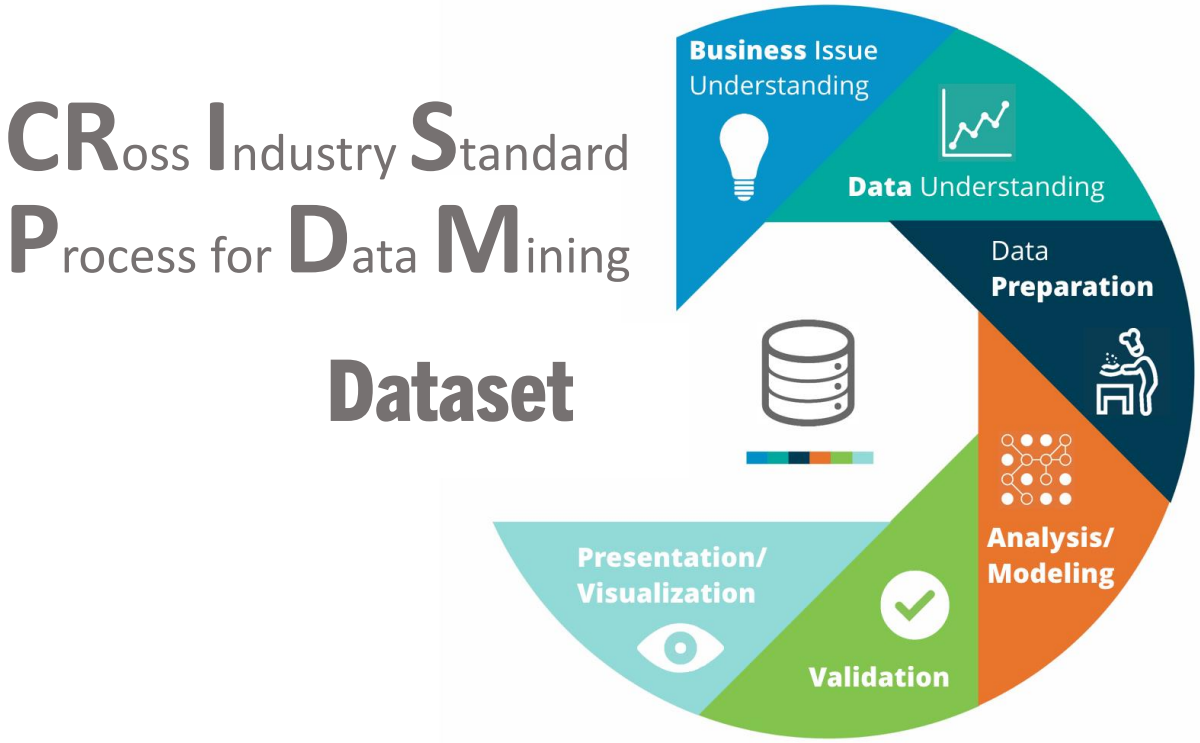

Figura 1 | Ciclo da Metodologia CRISP-DM.

Fonte: Adaptado de Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*. O'Reilly.

1 | Business Understanding

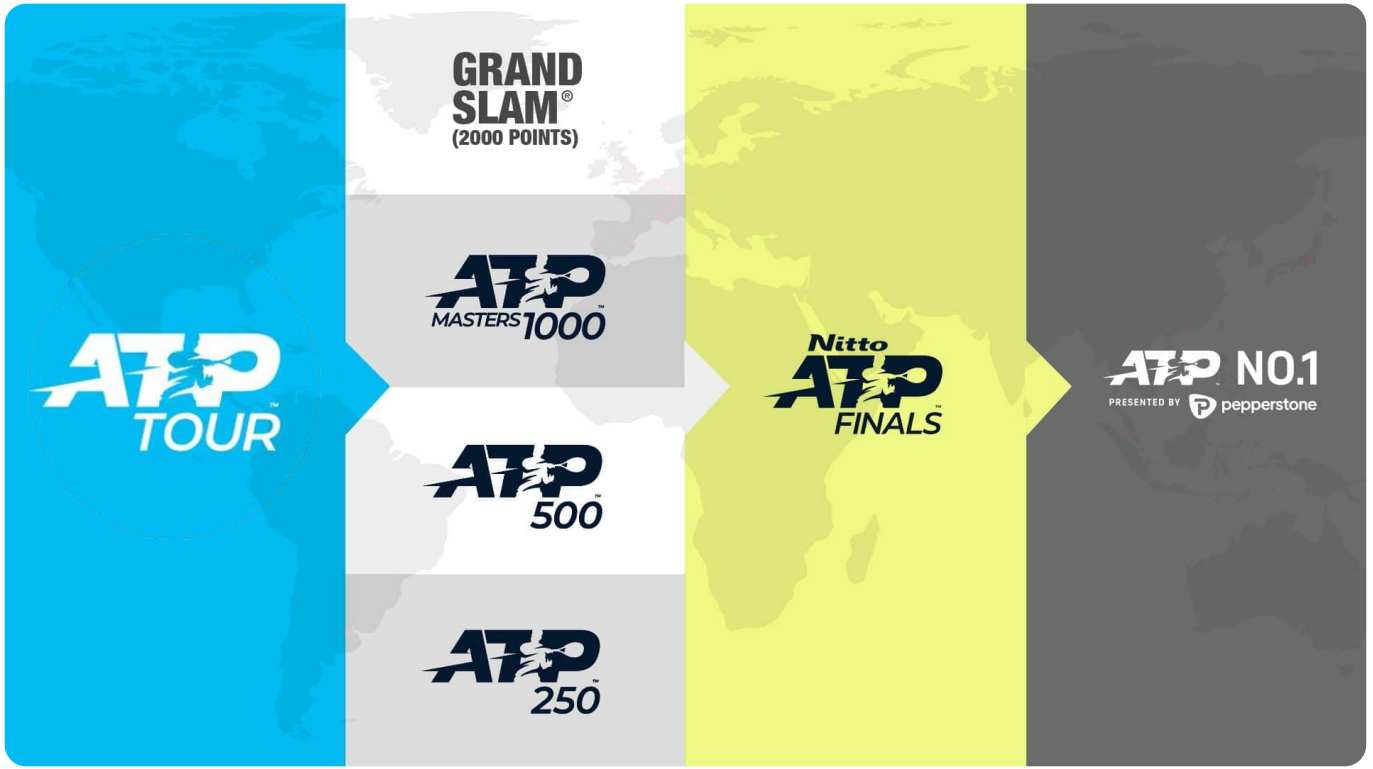

Figura 2 | Diferentes Temporadas até ao Título de *Melhor Tenista Profissional*.

Fonte: *Official Site of Men's Professional Tennis | ATP Tour | Tennis*. (2019). ATP Tour. https://www.atptour.com/en

1 | Business Understanding

Questão Problema

Como podemos prever o **Nº de Sets** que decorreram em jogos da **Suécia**, sabendo aspetos ligados aos jogadores e torneios associados aos jogos?

2 | Data Understanding

Nº	Variável	Descrição	
~~0~~	~~_id~~	~~Identificador único de cada observação~~	
1	PlayerName	Nome do principal jogador da partida	J
2	Born	Cidade e/ou país de onde o jogador é natural	J
3	Height	Altura do jogador (em *cm*)	J
4	Hand	Mão possante do jogador e mãos utilizadas para executar o serviço	J
5	LinkPlayer	Link do perfil do jogador no website do *ATP Tour*	J
6	Tournament	Nome do torneio	J
7	Location	Localização do torneio	J
8	Date	Intervalo de datas em que se realizou o torneio	J
9	Ground	Terreno em que é jogado o torneio (`Hard` , `Grass` , `Clay` , `Carpet`)	J
10	Prize	Valor do prémio na moeda local	J
11	GameRound	Fase do torneio em que o jogo está a ser disputado	J
~~12~~	~~GameRank~~	~~Rank do jogador adversário~~	
13	Opponent	Nome do jogador adversário	J
14	WL	Resultado do jogo (`W` - venceu, `L` - perdeu)	J
15	Score	Distribuição dos resultados dos *sets*	J

Legenda:

- J Jogadores
- J Jogos

Figura 3 | Descrição das Variáveis.

2 | Data Understanding - Problemas Identificados

Jogadores J

Nº	Variável	Descrição
1	PlayerName	Nome do principal jogador da partida
2	Born	Cidade e/ou país de onde o jogador é natural
3	Height	Altura do jogador (em *cm*)
4	Hand	Mão possante do jogador e mãos utilizadas para executar o serviço
5	LinkPlayer	Link do perfil do jogador no website do *ATP Tour*
13	Opponent	Nome do jogador adversário

Existem **9953** jogadores com nomes distintos.

Observam-se alturas anómalas (p.e., **0 cm**, **15 cm**, **510 cm**)

Existem **9960** jogadores com *links* distintos.

Existem valores que indicam situações particulares

- **bye**
- **W/o** (*Without*)

Têm variações heterogêneas para a mesma região de naturalidade.

Têm 7 classes possíveis:

- `Right-Handed, Two-Handed Backhand`
- `Left-Handed, Two-Handed Backhand`
- `Right-Handed, One-Handed Backhand`
- `Left-Handed, One-Handed Backhand`
- `Right-Handed, Unknown Backhand`
- `Left-Handed, Unknown Backhand`
- `Ambidextrous, Two-Handed Backhand`

♦ Não está presente no *dataset* a variável referente à **Data de Nascimento** dos jogadores, que é solicitada no enunciado do projeto.

2 | Data Understanding - Problemas Identificados

Jogos J

Nº	Variável	Descrição
6	Tournament	Nome do torneio
7	Location	Localização do torneio
8	Date	Intervalo de datas em que se realizou o torneio
9	Ground	Terreno em que é jogado o torneio (`Hard` , `Grass` , `Clay` , `Carpet`)
10	Prize	Valor do prémio na moeda local
11	GameRound	Fase do torneio em que o jogo está a ser disputado
14	WL	Resultado do jogo (`W` - venceu, `L` - perdeu)
15	Score	Distribuição dos resultados dos *sets*

Existem terminologias como:
- **TBA** (*To Be Announced*),
- **TBC** (*To Be Confirmed*)
- **TBD** (*To Be Determined*)

Contém carácter especial ◆

Indica a tomada de decisão da partida.

Diferentes designações para o mesmo torneio

Encontra-se em 2 formatos *string*:
- **`aaaa.mm.dd`** (torneio de 1 dia)
- **`aaaa.mm.dd - aaaa.mm.dd`** (torneio de mais de 1 dia)

Possuem vários formatos distintos:
- **`NN (RET)`**
- **`NN NN`**
- **`NN, NN`**
- **…**
- **`NN NN NN-NNN NN`** (Quantos *sets* são?)