# *Projeto Aplicado a Ciência de Dados I*
# *Ténis na Bélgica*

## Grupo 4:


ALLAN KARDEC, Nº 103380, TURMA CDB1
DIOGO FREITAS, Nº104841, TURMA CDB1
JOÃO BOTAS, Nº104782, TURMA CDB1
RICARDO ÂNGELO, Nº 104826, TURMA CDB1

# CRISP-DM metodologia

## Previsão do número de sets em jogos à melhor de 3


Business Understanding

Data Understanding

Data Preparation

Modeling & Evaluation

Conclusões e Deployment

# Business Understanding

- Como funciona o ténis... E do ponto de vista do negócio?

**Prever o nº sets para ajudar...**

- Consumo dos espectadores nas partidas (investir em produtos alimentares e itens desportivos);

- Publicidade e Merchandising (partidas mais longas, melhores jogadores rankeados);

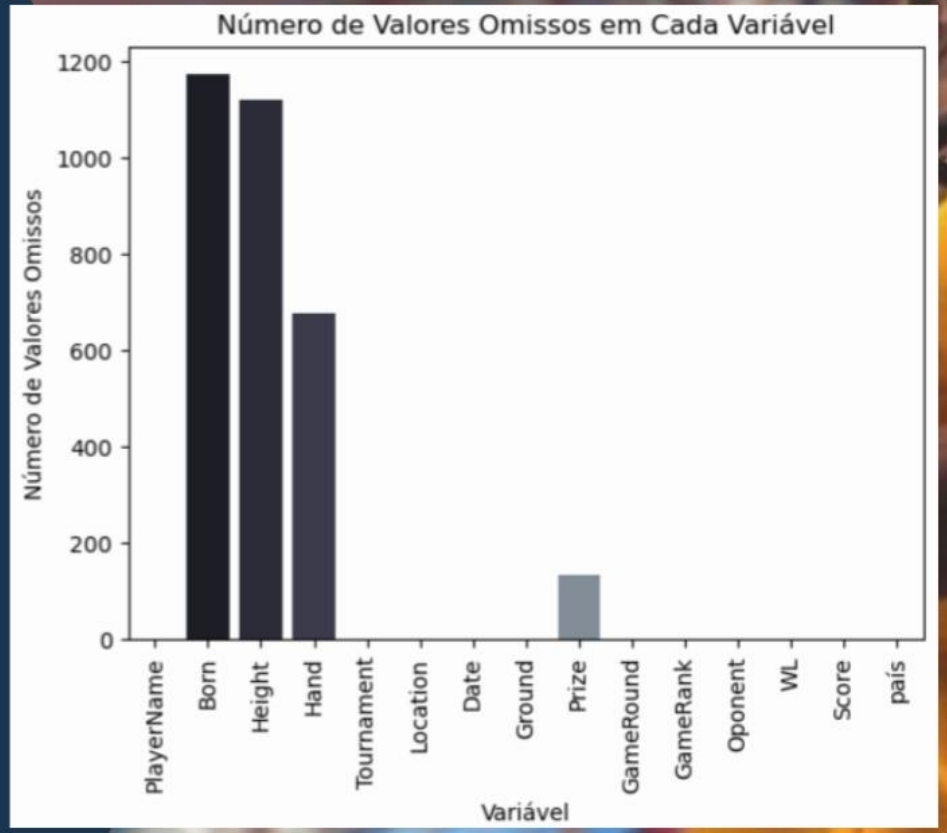
- Casas de apostas (taxa de acerto maior, mais dinheiro envolvido leva a mais interesse);


# Data Understanding

- Perceber/Entender os dados, com conceitos de ténis
    (O que é W/O, RET, DEF)

- Inicialmente os dados relativos à Bélgica correspondiam a 1% da base de dados original
    (10996 linhas e 15 colunas, sem ID:{ })

- Remoção de jogos duplicados por vencedor e perdedor do jogo (W/L);




5982 linhas (54.4% valor inicial)
- Omissos Prize e outras variáveis;

# Data Preparation

- Para aliviar os omissos na base de dados:
  - Webscrapping;
  - Imputação por regressão;

↓

- 4033 linhas e 11 colunas em partidas à melhor de 3

- Omissos:
  quase todos corrigidos, mas **SumTrainer**

  - 0, ambos eram omissos
  - 1, só um dos dois jogadores era omisso
  - 2, ambos tinham treinadores conhecidos

- **IMC:** indicador geral para identificar excesso de peso e obesidade
  (são todos atletas e as diferenças são notáveis em casos extremos)

$$IMC = \frac{PESO}{ALTURA^2}$$

| Feature | Explicação da Variável |
|---|---|
| *GroundNumerico* | Variável que tem como valor, de forma codificada, o tipo de chão da partida |
| *Prize* | Variável que tem como valor o prémio do torneio |
| *TournamentYear* | Variável que tem como valor o ano em que o torneio ocorreu |
| *DiffIMC* | Variável que tem como valor a subtração do IMC do **Oponent** ao IMC do **PlayerName** |
| *DiffAge* | Variável que tem como valor a subtração da idade do **Oponent** à idade do **PlayerName** |
| *DiffRank* | Variável que tem como valor a subtração do *rank* do **Oponent** ao *rank* do **PlayerName** |
| *QTDBelgas* | Variável tem como valor a quantidade de jogadores belgas na partida |
| *Finals* | Variável que refere se a ronda da partida possui a palavra **Final** |
| *SumMainHand* | Variável que tem como valor a quantidade de jogadores destros na partida |
| *SumTrainer* | Variável que tem como valor a quantidade de jogadores na partida que possuem um treinador |
| **NumeroSets** | **Variável Target** |

# Modeling & Evaluation

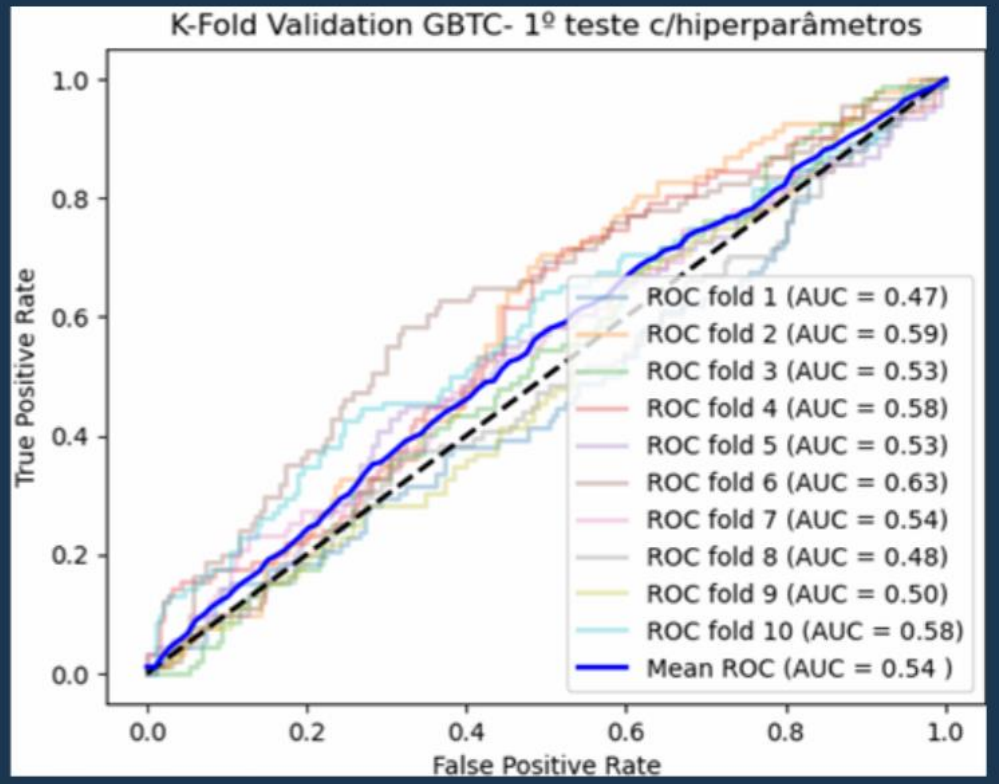
Validação cruzada k-fold partition com k=10:

- Número suficiente de Folds
- Variância reduzida
- Não é tão computacionalmente intensivo
- Maior robustez dos conjuntos (maior representatividade)

**Resultados**

Métricas de classificação do modelo:

- Mean Accuracy - 0.674
- Mean Recall - 0.990
- Mean Precision - 0.677
- Mean F1-Score - 0.804
- Mean ROC (AUC) - 0.54



# Conclusões e Deployment

- Resultados aquém do que esperávamos, no entanto, são dados reais;

Fatores que achamos que tenham sido flagrantes:

- nº sets é difícil de classificar observando para dados de jogadores/torneios;
- a Bélgica ter poucas linhas na base de dados;
- existência de variáveis com pouca correlação, mesmo dos dados originais.

Mas conseguimos...

- Desenvolver as apostas desportivas e marketing
- Análise de estratégias táticas
- Transmissão e comentários ao vivo
- Planeamento de Eventos e Logística
- Complementar Informação