

Ténis nos EUA

Grupo 10

Projeto Aplicado em Ciência de Dados I



Introdução

Objetivo: Criar um modelo preditivo para prever o número de *sets* necessários num jogo de ténis profissional.

País atribuído: Estados Unidos - grande representativa no circuito ATP

Metodologia: CRISP-DM



A stylized illustration of a tennis player in mid-swing, serving a ball. The player is wearing a teal shirt, white pants, and a white headband. The ball is frozen in motion at the top of the racket. The background features green bushes and a light orange sky with small dark stars.

01

Business Understanding

Business Understanding

- Ténis nos EUA
- Importância da Previsão de Número de *Sets*:
 - Apostas desportivas: definição de odds e estratégias de risco
 - Transmissão televisivas: maior duração → mais receitas
 - Organização de torneios: ajuda no planeamento dos horários e campos
 - *Merchandising e consumo*: mais *sets* → maior volume de vendas

Data Understanding

02



Fonte e estrutura da base de dados

- Dados do site ATP Tour (1973-2022);
- Cerca de 1,3 milhões de regtos com 15 variáveis;
- Dimensões das variáveis: Jogadores, Torneios e Jogos

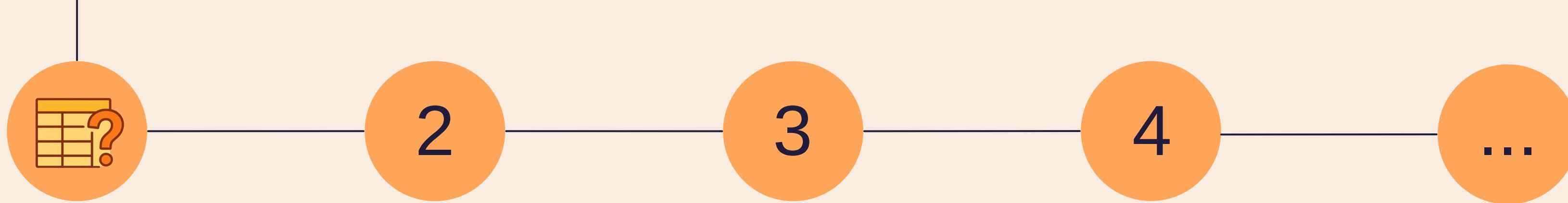
Problemas identificados



Problemas identificados

(Born; Hand; Height;
primary keys)

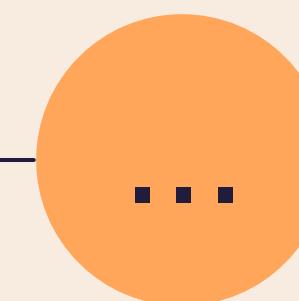
dados em falta



Problemas identificados

(Born; Hand; Height;
primary keys)

dados em falta



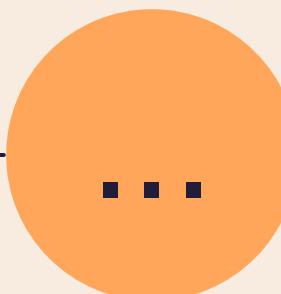
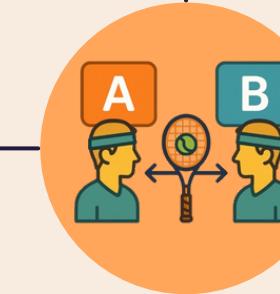
registos duplicados

(2830 documentos)

Problemas identificados

(Born; Hand; Height;
primary keys)

dados em falta



registos duplicados
(2830 documentos)

player A → player B
player B → player A
jogos espelhados

Problemas identificados

(Born; Hand; Height;
primary keys)

dados em falta

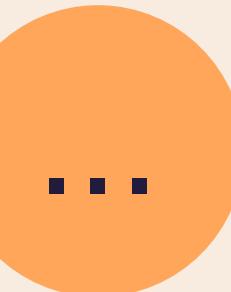


registos duplicados
(2830 documentos)

player A → player B
player B → player A
jogos espelhados



oponentes sem dados
(12731 nomes distintos)



Problemas identificados

(Born; Hand; Height;
primary keys)

dados em falta



registos duplicados
(2830 documentos)

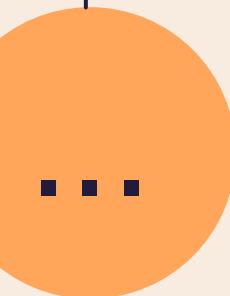
player A → player B
player B → player A
jogos espelhados



oponentes sem dados
(12731 nomes distintos)



entre outros...



Data Preparation

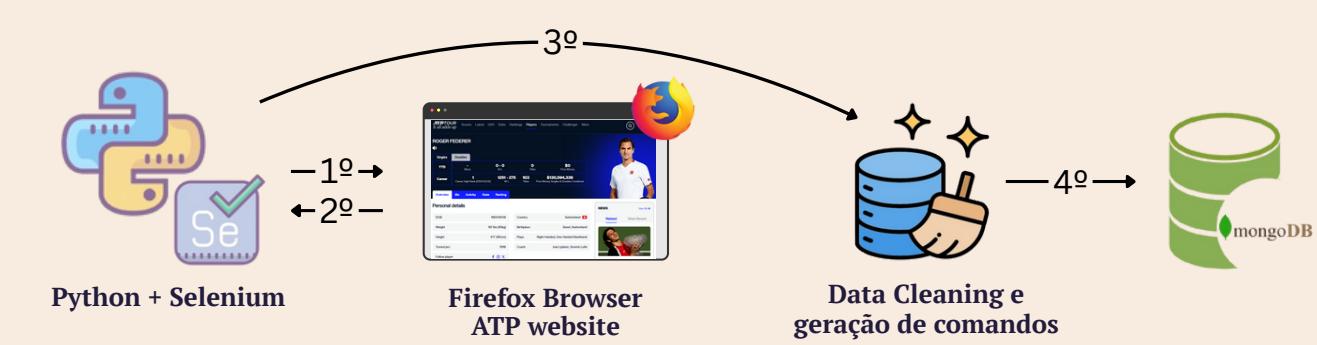
03



Etapas



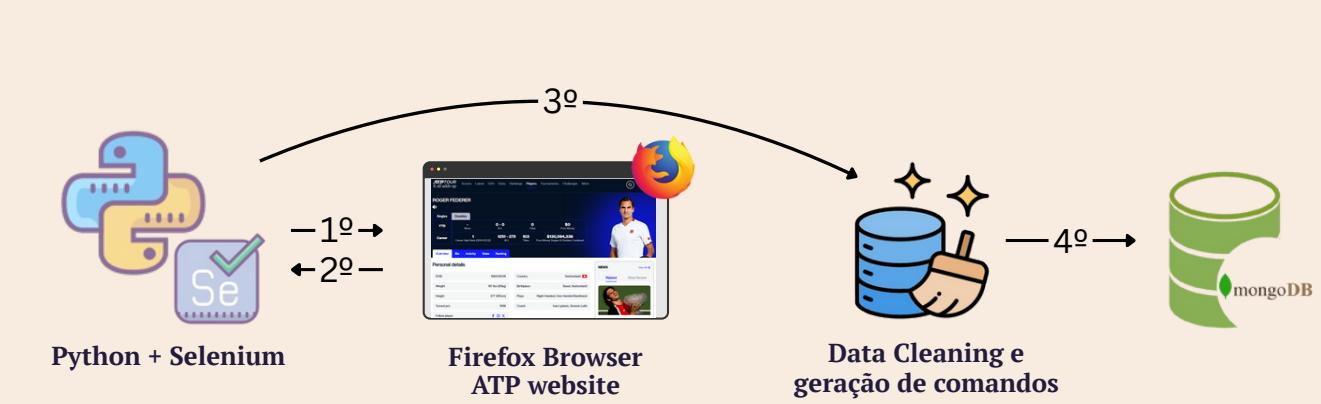
Recolha de dados (web scraping)
*dados de todos os jogadores e posteriormente
datas de nascimento + ranks*



Etapas



Recolha de dados (web scraping)
*dados de todos os jogadores e posteriormente
datas de nascimento + ranks*



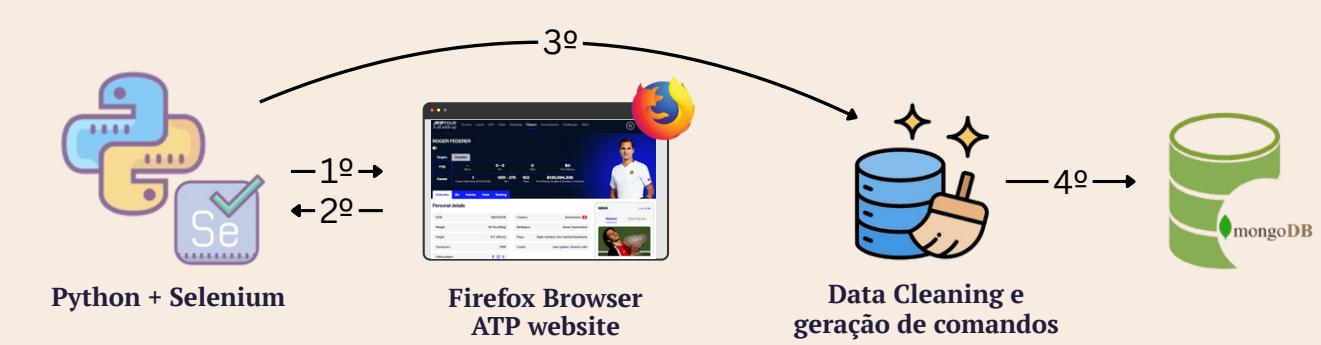
Criação de Chaves Primárias
o MatchId resolve problema dos jogos espelhados

BornCountryCode	TournamentCountryCode	TournamentId	MatchId
ES	IT	000b750bc715670198d700eae86bc71c	000b750bc715670198d700eae86bc71c_Roundof64_bye_g076
HR	JP	000cea2d9229b5c919fca319557a9c7a	000cea2d9229b5c919fca319557a9c7a_Finals_a385_m759
ZA	JP	000cea2d9229b5c919fca319557a9c7a	000cea2d9229b5c919fca319557a9c7a_Finals_a385_m759

Etapas



Recolha de dados (web scraping)
*dados de todos os jogadores e posteriormente
datas de nascimento + ranks*



Criação de Chaves Primárias
o MatchId resolve problema dos jogos espelhados

BornCountryCode	TournamentCountryCode	TournamentId	MatchId
ES	IT	000b750bc715670198d700eae86bc71c	000b750bc715670198d700eae86bc71c_Roundof64_bye_g076
HR	JP	000cea2d9229b5c919fca319557a9c7a	000cea2d9229b5c919fca319557a9c7a_Finals_a385_m759
ZA	JP	000cea2d9229b5c919fca319557a9c7a	000cea2d9229b5c919fca319557a9c7a_Finals_a385_m759



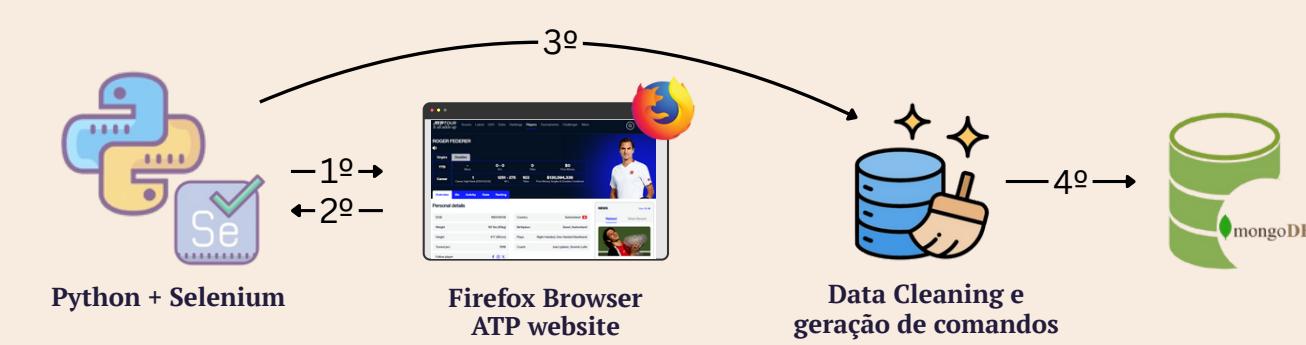
Filtrar jogos realizados nos EUA
amostra do projeto

TournamentCountryCode	Tournament	Location	Date	Ground	Prize
US	U.S.A. F14	Innisbrook, FL, U.S.A.	2013.06.03 - 2013.06.09	Clay	\$10,000
US	U.S.A. F24	Decatur, United States	2015.08.03 - 2015.08.09	Hard	\$15,000
US	U.S.A. F23	U.S.A.	2014.08.04 - 2014.08.10	Hard	\$10,000

Etapas



Recolha de dados (web scraping)
*dados de todos os jogadores e posteriormente
datas de nascimento + ranks*



Criação de Chaves Primárias
o MatchId resolve problema dos jogos espelhados

BornCountryCode	TournamentCountryCode	TournamentId	MatchId
ES	IT	000b750bc715670198d700eae86bc71c	000b750bc715670198d700eae86bc71c_Roundof64_bye_g076
HR	JP	000cea2d9229b5c919fca319557a9c7a	000cea2d9229b5c919fca319557a9c7a_Finals_a385_m759
ZA	JP	000cea2d9229b5c919fca319557a9c7a	000cea2d9229b5c919fca319557a9c7a_Finals_a385_m759



Filtrar jogos realizados nos EUA
amostra do projeto

TournamentCountryCode	Tournament	Location	Date	Ground	Prize
US	U.S.A. F14	Innisbrook, FL, U.S.A.	2013.06.03 - 2013.06.09	Clay	\$10,000
US	U.S.A. F24	Decatur, United States	2015.08.03 - 2015.08.09	Hard	\$15,000
US	U.S.A. F23	U.S.A.	2014.08.04 - 2014.08.10	Hard	\$10,000



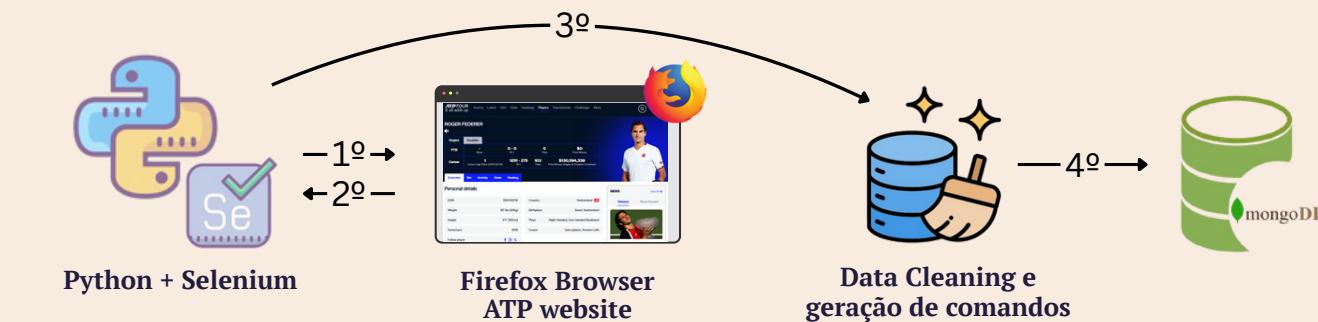
Criação da variável alvo
“number_of_sets”
a partir da variável “score”

Score	number_of_sets
67 63 63 62	4
46 62 64 46 62	5
62 63 62	3

Etapas



Recolha de dados (web scraping)
*dados de todos os jogadores e posteriormente
datas de nascimento + ranks*



Criação de Chaves Primárias
o MatchId resolve problema dos jogos espelhados

BornCountryCode	TournamentCountryCode	TournamentId	MatchId
ES	IT	000b750bc715670198d700eae86bc71c	000b750bc715670198d700eae86bc71c_Roundof64_bye_g076
HR	JP	000cea2d9229b5c919fca319557a9c7a	000cea2d9229b5c919fca319557a9c7a_Finals_a385_m759
ZA	JP	000cea2d9229b5c919fca319557a9c7a	000cea2d9229b5c919fca319557a9c7a_Finals_a385_m759



Filtrar jogos realizados nos EUA
amostra do projeto

TournamentCountryCode	Tournament	Location	Date	Ground	Prize
US	U.S.A. F14	Innisbrook, FL, U.S.A.	2013.06.03 - 2013.06.09	Clay	\$10,000
US	U.S.A. F24	Decatur, United States	2015.08.03 - 2015.08.09	Hard	\$15,000
US	U.S.A. F23	U.S.A.	2014.08.04 - 2014.08.10	Hard	\$10,000



Criação da variável alvo
“number_of_sets”
a partir da variável “score”

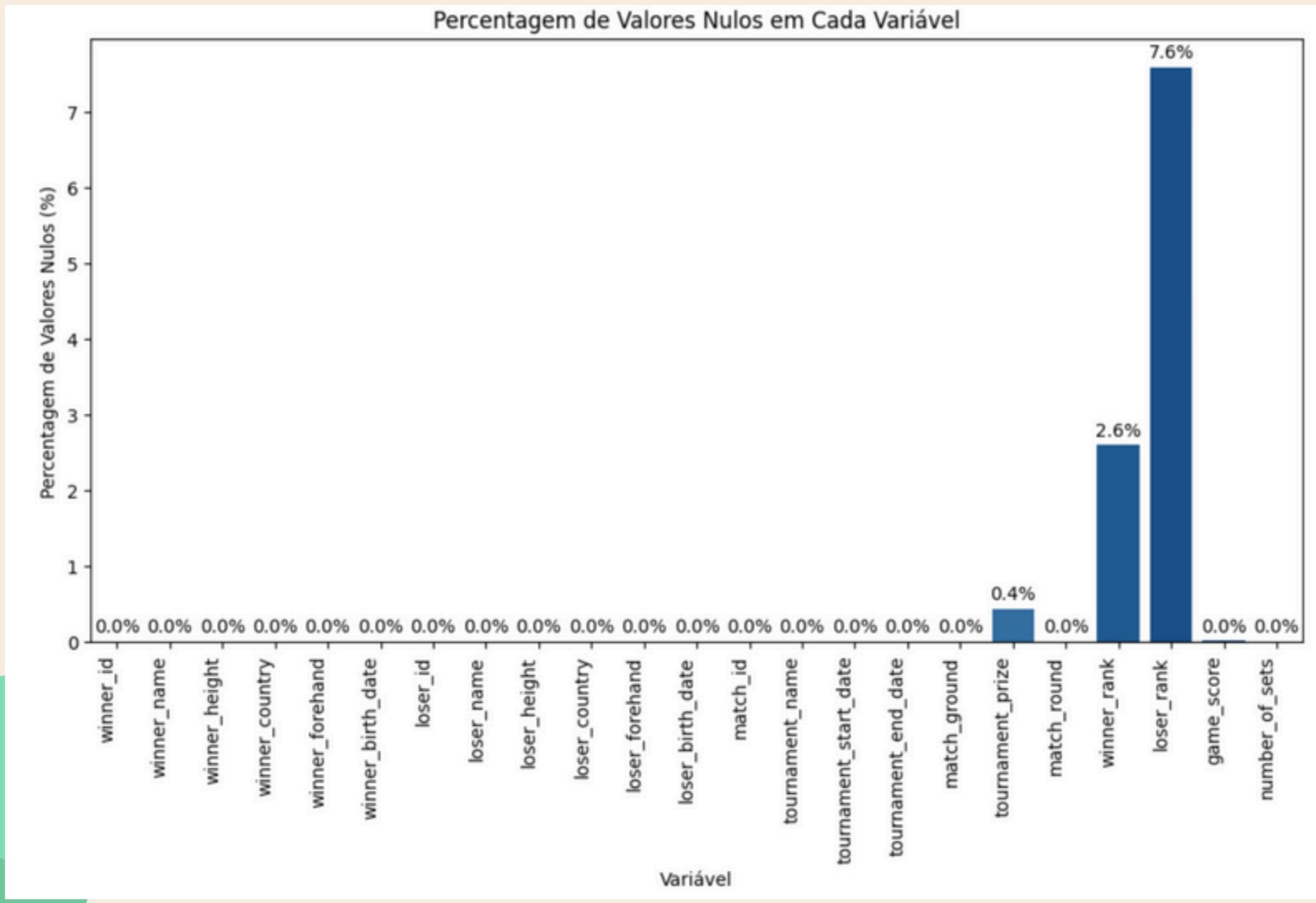
Score	number_of_sets
67 63 63 62	4
46 62 64 46 62	5
62 63 62	3



Exportação em CSV
para usar em Python nas próximas fases

```
winner_id,winner_name,winner_height,winner_country,winner_forehand,winner_birth_date,loser_id,loser_name,loser_height,loser_country,loser_forehand,loser_birth_date  
a092,Andre Agassi,180,United States,Right-Handed,1970/04/29,p024,Mikael Pernfors,173,Sweden  
a092,Andre Agassi,180,United States,Right-Handed,1970/04/29,p012,David Pate,183,United States  
c057,Kevin Curren,185,United States,Right-Handed,1958/03/02,g023,Andres Gomez,193,Ecuador,  
p024,Mikael Pernfors,173,Sweden,Right-Handed,1963/07/16,g039,Jim Grabb,193,United States,R
```

Valores omissos



As variáveis referentes ao *ranking* do jogador são as mais afetadas (7.6% e 2,6%)

Variável alvo

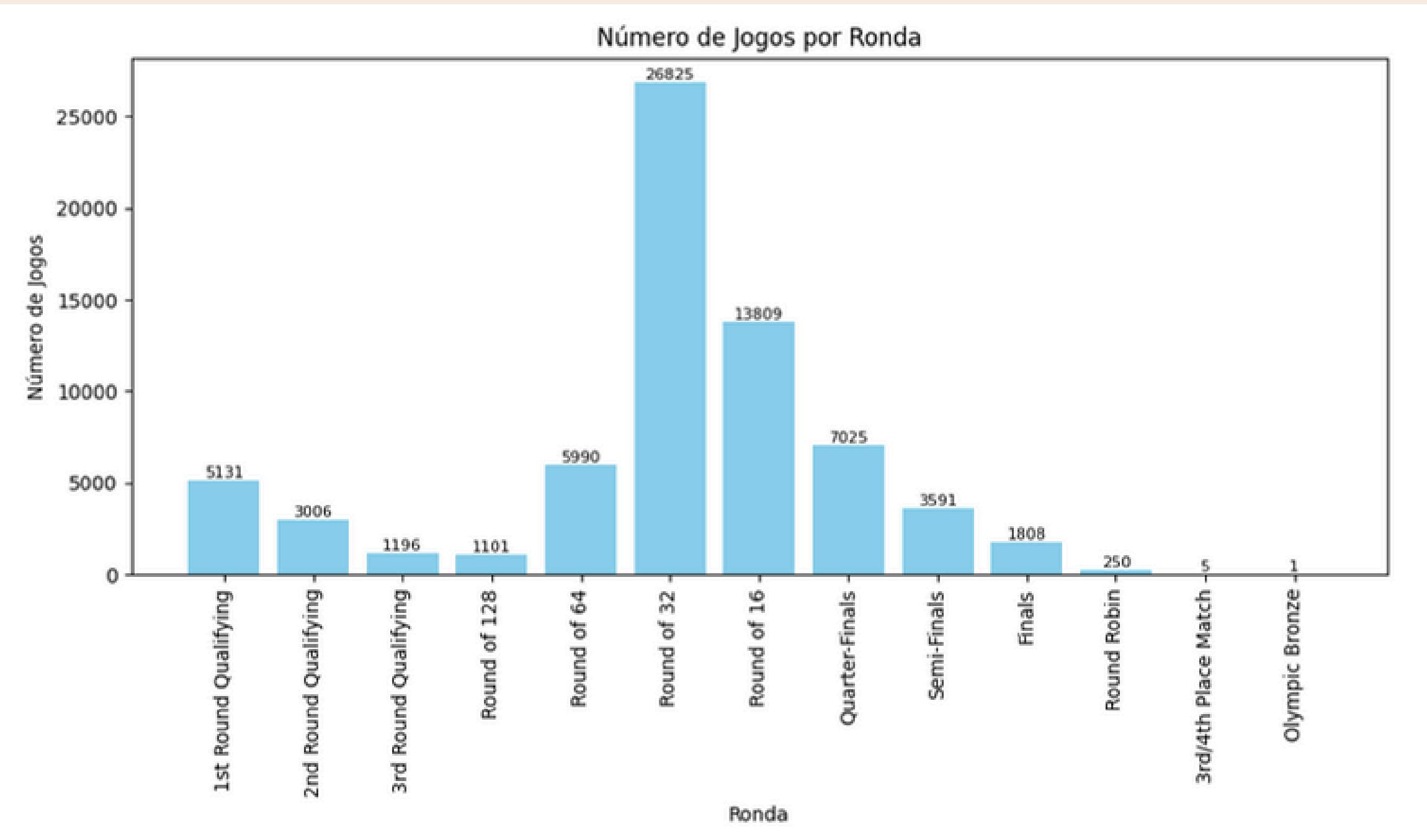
- Exclusões aplicadas para garantir consistência:
 1. Jogos irregulares: “RET”, “W/O”, “DEF”
 2. Jogos à melhor de 5 sets (5651 registos)
- Foco exclusivo em jogos à melhor de 3 sets



Base de dados com 69738 registos

number_of_sets	Frequência Absoluta	Percentagem (%)
2	46753	67.04
3	22985	32.96

Match Round



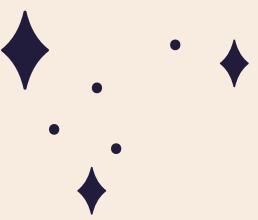
Redução da Dimensionalidade:

- Agrupamento de fases de qualificação:
 - *1st, 2nd, 3rd Round Qualifying*
 - *Round Robin*
 - Nova categoria: "*Qualifiers*"

Exclusões por baixa frequência:

- *3rd/4th Place Match* → 5 jogos
- *Olympic Bronze* → 1 jogo

Criação de novas variáveis



rank_difference - diferença absoluta entre os *ranks*

total_matches_difference - diferença absoluta entre o número de jogos disputados até ao momento

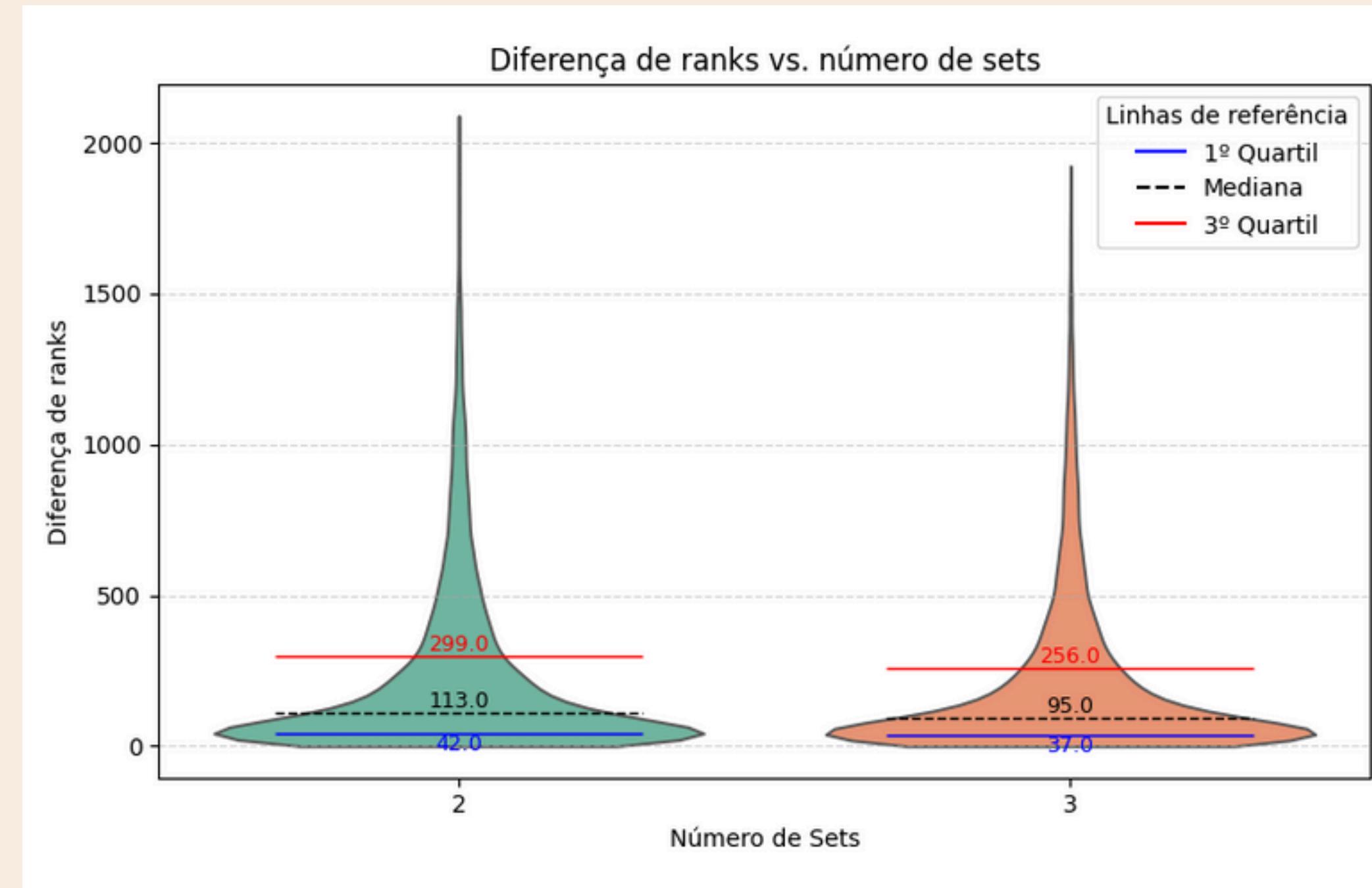
ratio_difference - diferença absoluta entre o rácio de vitórias

height_difference - diferença absoluta entre as alturas

age_difference - diferença absoluta entre as idades

avg_sets_difference - diferença asboluta entre o número médio de *sets* dos últimos 5 jogos

rank_difference vs. number_of_sets



Modelling

04



Construção da amostra treino-teste

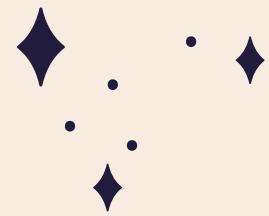


80 % treino (36.598 jogos)

20% teste (9150 jogos)

Devido à discrepância entre as classes da variável alvo, onde os jogos com 2 sets predominavam na base de dados foi adotada uma técnica de balanceamento que igualou o número de observações de cada classe - *Undersampling*

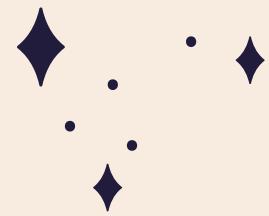
Cross-Validation



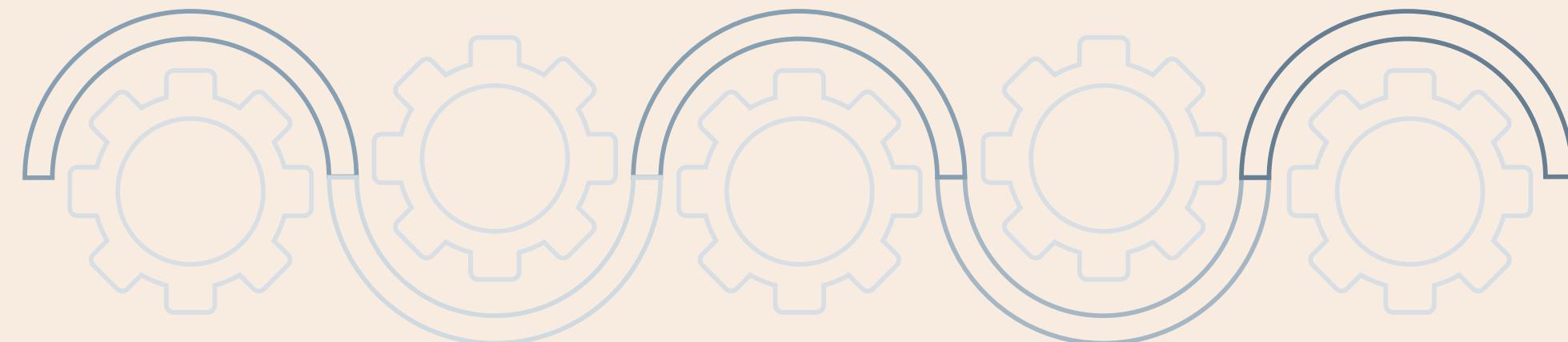
Além da avaliação no conjunto de teste, foi também implementada a técnica de validação cruzada com 10 *folds*, de forma a obter uma estimativa mais robusta e generalizável da performance dos modelos.

Metodologia: Este processo consiste em dividir o conjunto de treino em 10 subconjuntos (*folds*) de igual dimensão: em cada iteração, o modelo é treinado com 9 desses subconjuntos e testado no subconjunto restante, repetindo-se o processo 10 vezes para garantir que cada *fold* é usado como teste exatamente uma vez.

Algoritmos



Foram testados quatro algoritmos: **Regressão Logística, KNN, Árvore de Decisão e Random Forest**. No entanto, destacamos o Random Forest, que apresentou o melhor desempenho global.



Evaluation

05



Métricas de classificação

accuracy – indica a proporção de observações corretamente classificadas em relação ao total de observações avaliadas

precision - indica a proporção de observações corretamente classificadas das que são previstas como positivas

recall - indica a proporção de observações corretamente classificadas dos casos positivos

F1-score – representa o equilíbrio entre a *precision* e o *recall*, sendo especialmente útil em problemas com classes desbalanceadas

Curva ROC

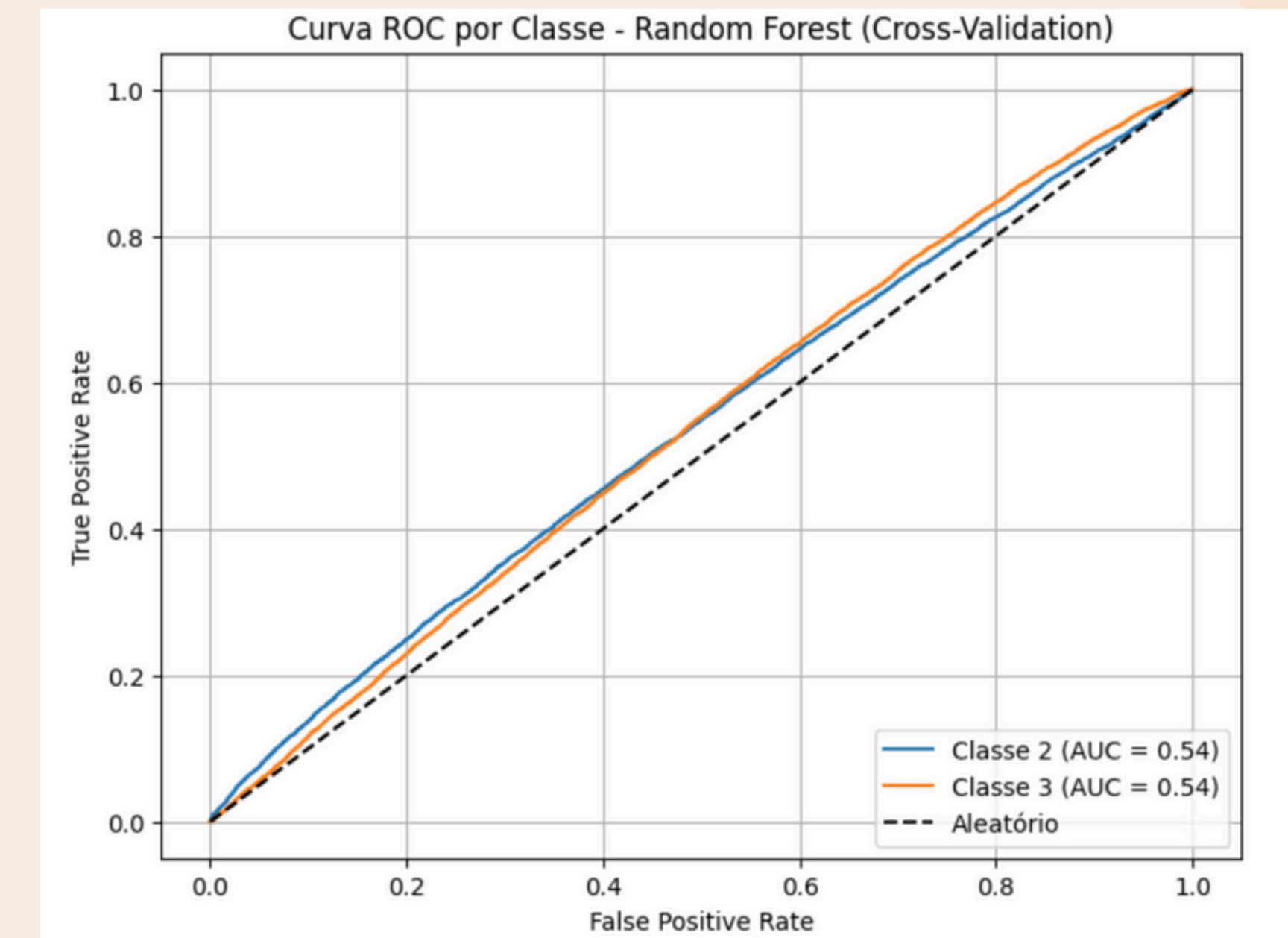
O que é ? - representa graficamente a taxa de verdadeiros positivos (*recall*) em função da taxa de falsos positivos, para diferentes limiares de decisão (*threshold*).

Limiariaes de decisão - Os limiares de decisão referem-se aos valores que determinam o ponto a partir do qual uma observação é atribuída a uma determinada classe.

AUC - O valor da área sob a curva (AUC) quantifica esta performance: um valor de 0.5 indica uma classificação equivalente a uma aleatória, enquanto um valor de 1.0 representa uma classificação perfeita.

Avaliação - Random Forest

	Original			10-Folds		
	Precision	Recall	F1-score	Precision	Recall	F1-score
2 sets	0.53	0.46	0.49	0.53	0.46	0.49
3 sets	0.52	0.59	0.55	0.52	0.60	0.56
Accuracy	0.53			0.53		



Deployment

05



Aplicações Práticas



Apoio à Decisão em Casas de Apostas

Os modelos permitem gerar uma estimativa da duração do jogo que pode ser útil para:

- Definição de *odds* de apostas mais ajustadas
- Criar mercados de apostas mais competitivos



Gestão e Planeamento de Torneios

A previsão do número de sets contribui para estimar a duração dos encontros, útil na:

- Distribuição eficiente dos campos disponíveis
- Planeamento dos horários dos jogos
- Redução de atrasos e sobreposição de partidas

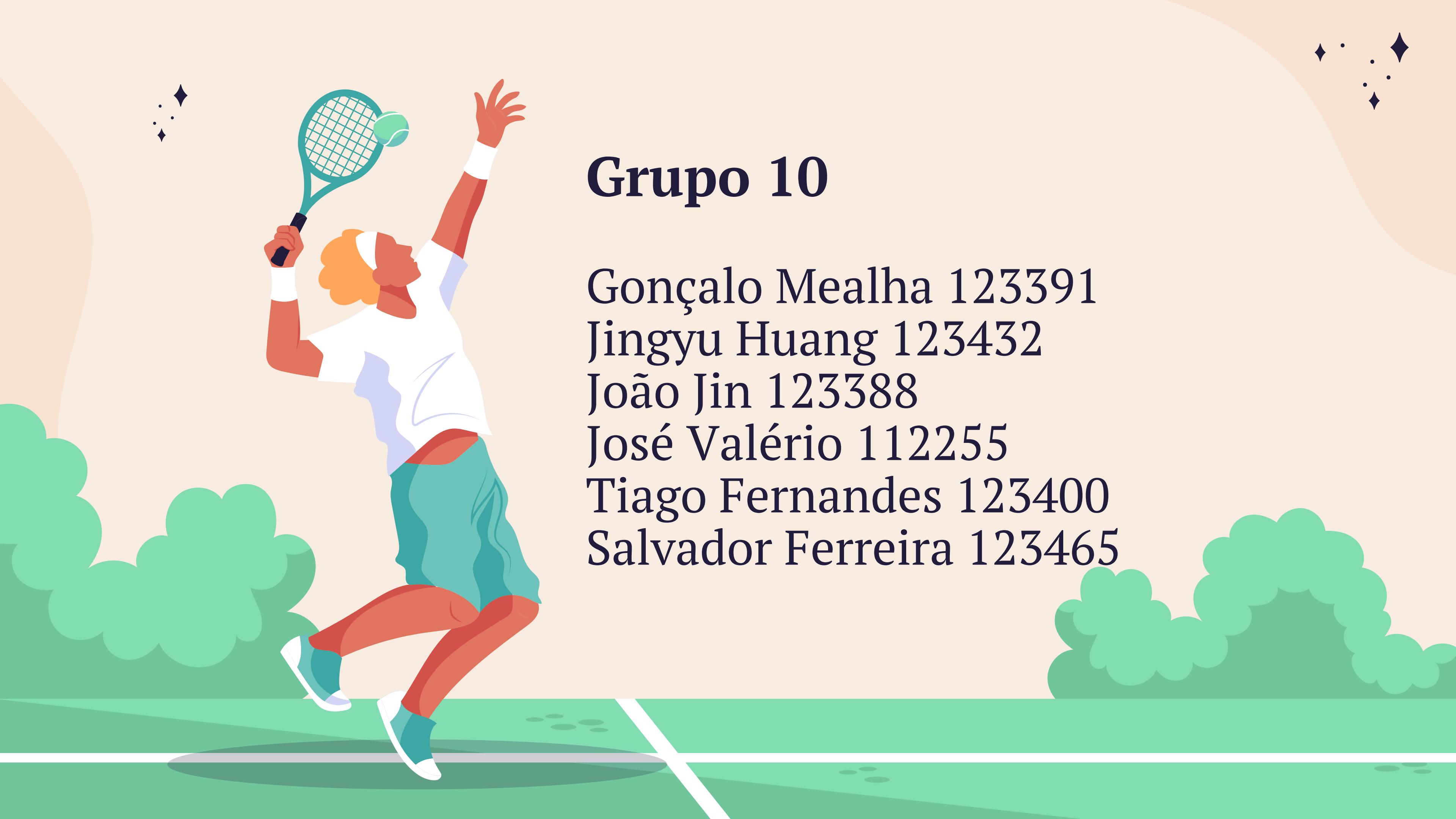
Considerações finais

Desafio atual

A previsão do número de *sets* ainda apresenta dificuldades com os dados disponíveis.

Possíveis melhorias:

- Integração de novas variáveis, como:
 - Taxa de serviços bem-sucedidos
 - Histórico de confrontos diretos
 - Condições meteorológicas durante os jogos

A colorful illustration of a tennis player in action. The player, wearing a white shirt, blue shorts, and a red cap, is captured in the middle of a serve. A green tennis ball is shown in the air near the racket. The background features stylized green bushes and a light orange sky with small black stars.

Grupo 10

Gonçalo Mealha 123391
Jingyu Huang 123432
João Jin 123388
José Valério 112255
Tiago Fernandes 123400
Salvador Ferreira 123465



FIM!