



Bélgica - Previsão do nº de sets

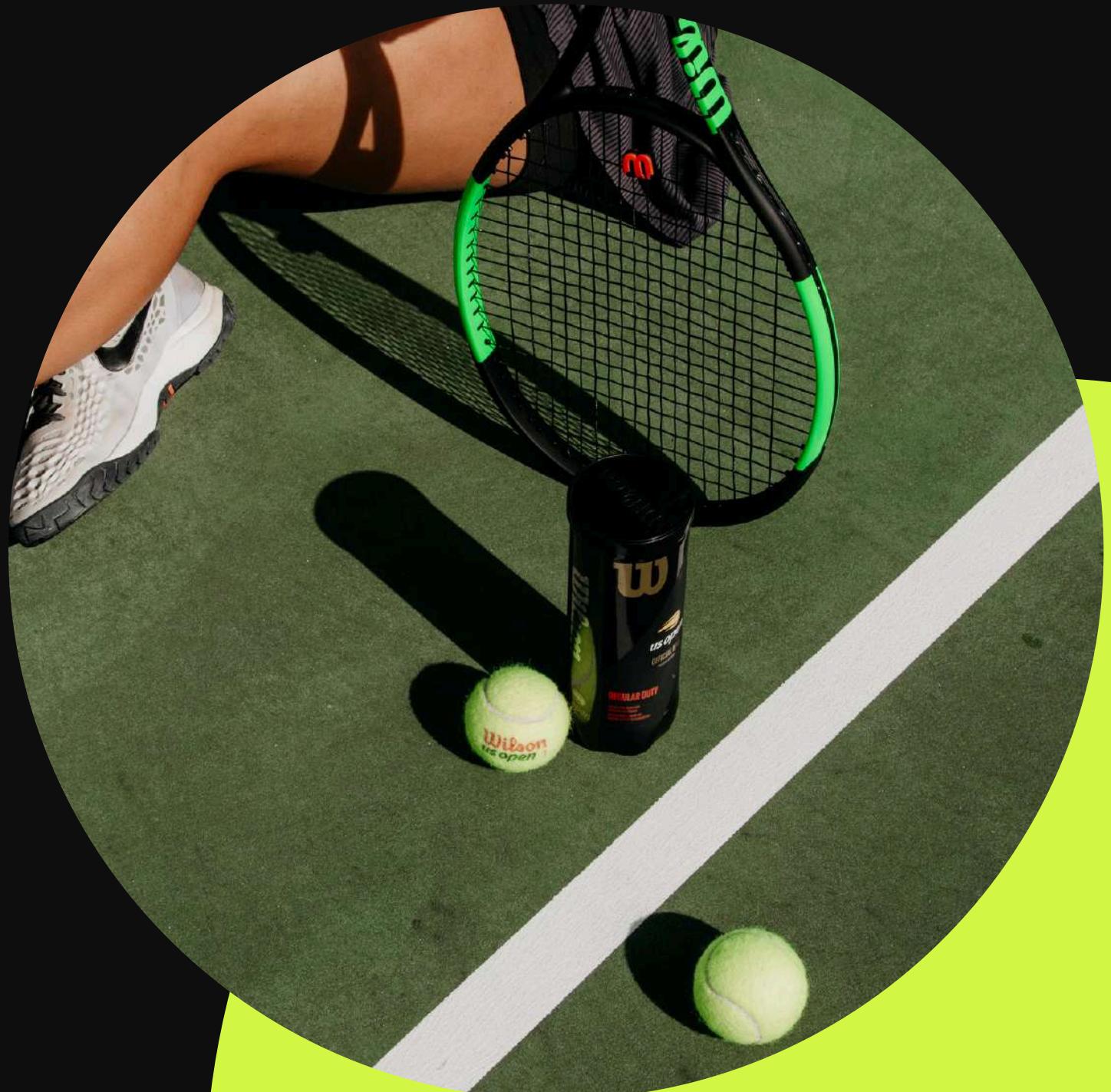
G4 - TURMA CDA2 -
APRESENTAÇÃO FINAL

PROJETO APLICADO A CIÊNCIA DE DADOS I

Business Understanding

Ténis na Bélgica:

- Não é conhecida por sediar os mais célebres torneios de ténis, sendo os mais conhecidos a *European Open Antwerp* e o Torneio *Ethias*;
- A maioria dos torneios são jogados em solo duro;
- No ténis masculino, destacamos *David Goffin* e *Zizou Bergs*.



Business Understanding

Aplicações práticas do modelo:

- **Apostas Desportivas:**
 - Apoia decisões mais informadas em “*set betting*” e apostas múltiplas;
 - Ajuda a identificar padrões;
 - Útil para casas de apostas na gestão do risco e ajuste de *odds* em tempo real.
- **Estratégia de Jogo:**
 - Suporta treinadores e jogadores na preparação tática;
 - Ajuda a prever o número de *sets* e a planear o esforço físico.



Data Preparation

Variável *Hand* e *Hand_Op*

- *Web scraping* para diminuir os valores omissos;
- Criação das variáveis *Hand*, *Backhand*, *Hand_Op* e *Backhand_op* para dividir a informação presente na variável inicial *Hand*;
- Criação da variável *Hand_vs* que representa a combinação de mãos entre *PlayerName* e *Oponent*.

Exemplo:

Hand = Right-Handed

Hand_Op = Left-Handed

Hand_vs = Right-Handed vs Left-Handed



Data Preparation

Variável *Born* e *Born_Op*

- *Web scraping* para diminuir os valores omissos;
- Separação de *strings* inconsistentes em 4 partes (*born_1*, *born_2*, *born_3*, *born_4*);
- Identificação da cidade e país com base em ficheiros auxiliares, resultando em *Born_city* e *Born_Country*, e equivalente para o oponente.

Exemplo:

Born = “Manhattan, New York, USA”;
Born_Country = “USA”;
Born_City = “New York”

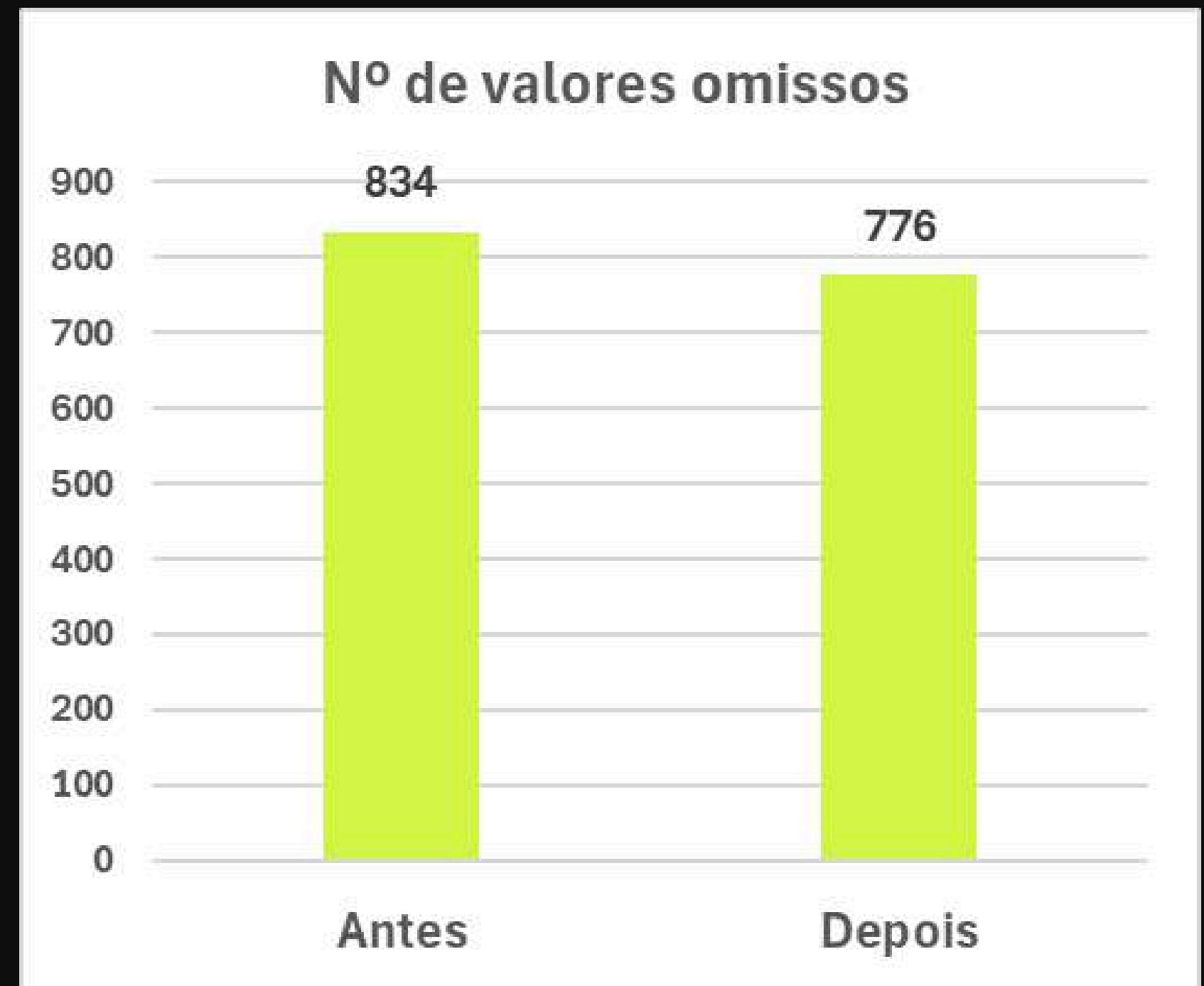


Data Preparation

Variável *Height* e *Height_Op*

Para tratar valores omissos:

- Web scraping no site ATP para os *PlayerNames*.
- Web scraping no site *tennisexplorer* para os *Oponents*.



Data Preparation

Criação de variáveis - Idade

- Criação da variável *DOB* (data de nascimento) para ambos os jogadores, através da recolha de informação via *web scraping*.
- A partir desta variável criou-se:
 - *Age* e *Age_Op*
 - *Age_Gap*
 - *Age_Difference*

Exemplo:

***Age* = 25, *Age_Op* = 28**

***Age_Gap* = 3, *Age_Difference* = -3**



Data Preparation

Variável GameRank

- Apenas o *oponent* tinha informações sobre o *GameRank* → Fizemos *web scraping* para obter o *rank* do *PlayerName* → Criação da variável *Rank_Player*.
- A partir daqui criaram-se duas variáveis:
 - *Difference_ranks_Gap*
 - *Difference_ranks*

Exemplo:

***Rank_Player* = 58, *GameRank* = 30**

***Difference_ranks_Gap* = 28, *Difference_ranks*= -28**



Data Preparation

Variável *Weight* e *IMC*

- Recolhemos o peso dos jogadores via *web scraping*, criando as variáveis: *Weight* e *Weight_Op*.
- Para os jogadores com peso e alturas disponíveis, calculámos o IMC, criando duas novas variáveis: *IMC* e *IMC_Op*.
- Através das duas anteriores criámos a variável *IMC_abs*.

Exemplo:

$\text{IMC} = 23.38$, $\text{IMC_Op} = 23.67$
 $\text{IMC_abs} = 0.29$



Data Preparation

Indicadores de forma e histórico

- Criamos uma função para calcular a percentagem acumulada de vitórias por ronda até à data do jogo, armazenada posteriormente em:
 - *Percentagem_Vitorias_PlayerName*
 - *Percentagem_Vitorias_Oponent*
- Através destas, criaram-se duas variáveis comparativas:
 - *Percentagem_Victory_Abs*
 - *Percentagem_Victory_diff*



Exemplo:

***Percentagem_Vitorias_PlayerName* = 0.7, *Percentagem_Vitorias_Oponent* = 0.25**
***Percentagem_Victory_Abs* = 0.45, *Percentagem_Victory_diff* = -0.45**

Data Preparation

Indicadores de forma e histórico

- Calculamos a percentagem de vitórias nas 5 partidas anteriores a cada jogo, armazenadas em:
 - *Recent_Form_Player*
 - *Recent_Form_Oponent*
- A partir daqui, criaram-se duas variáveis comparativas:
 - *Abs_Recent_Form*
 - *Diff_Recent_Form*



Exemplo:

Recent_Form_Player = 0.4, *Recent_Form_Oponent* = 0.8
Abs_Recent_Form = 0.4, *Diff_Recent_Form* = -0.4

Data Preparation

Indicadores de forma e histórico

- Criamos uma fórmula para calcular a percentagem de sets vencidos nos último 5 jogos e foram criadas as seguintes variáveis:
 - *Per_Win_Sets*
 - *Per_Win_Sets_Oponent*
- A partir destas criaram-se variáveis comparativas:
 - *Per_Win_Sets_abs*
 - *Per_Win_Sets_diff*



Exemplo:

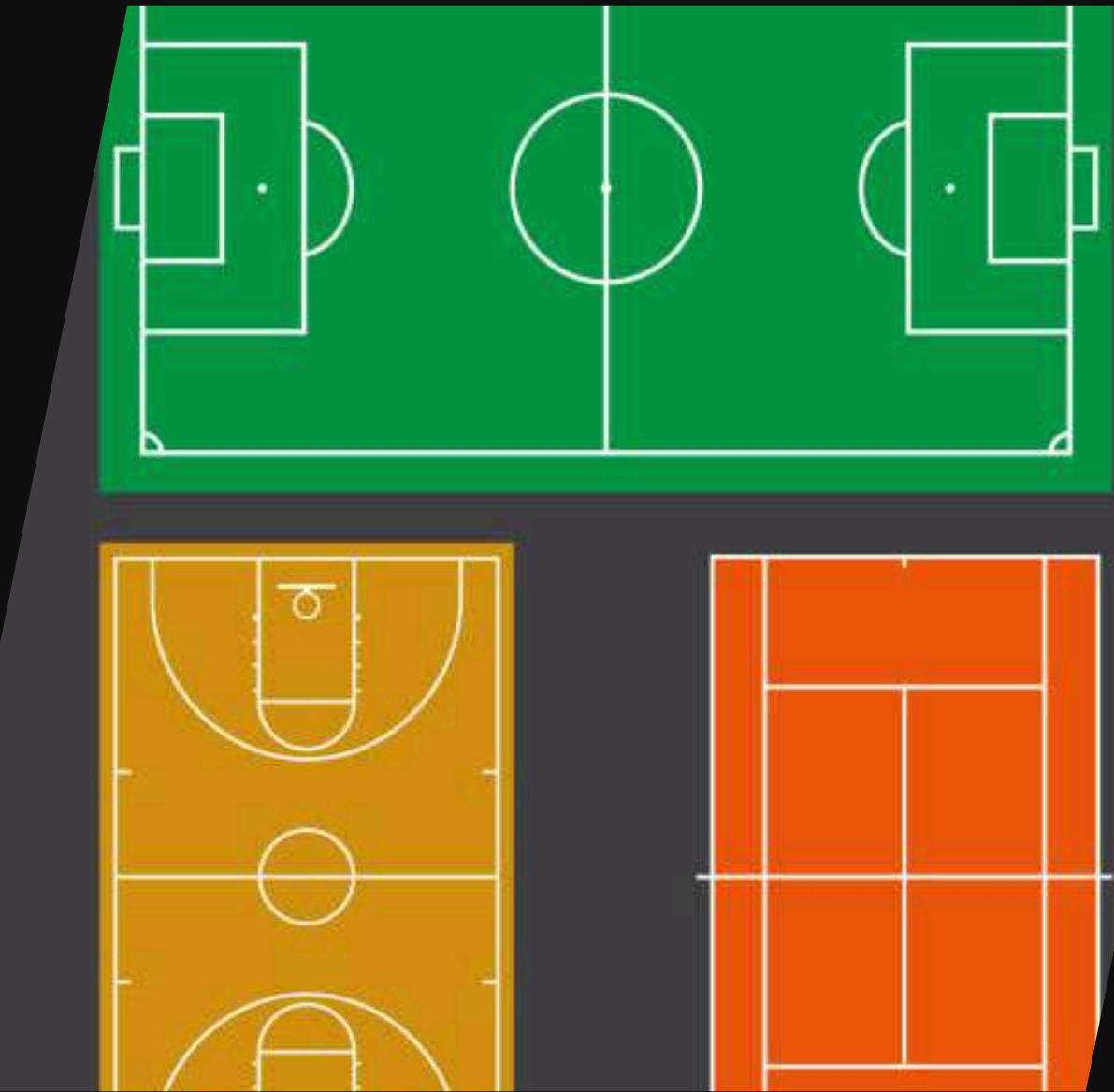
$\text{Per_Win_Sets} = 0.8$, $\text{Per_Win_Sets_Oponent} = 0.455$
 $\text{Per_Win_Sets_abs} = 0.345$, $\text{Per_Win_Sets_diff} = -0.345$



Data Preparation

Indicadores de forma e histórico

- Criámos a variável percentagem de vitórias por tipo de piso e obtivemos as seguintes variáveis:
 - ***Ground_Wins***
 - ***Ground_Wins_Op***
- A partir daqui criaram-se variáveis comparativas:
 - ***Ground_Wins_Abs***
 - ***Ground_Wins_Diff***



Exemplo:

Ground_Wins = 0.29, ***Ground_Wins_Op*** = 0.286
Ground_Wins_Abs = 0.4 , ***Ground_Wins_Diff*** = -0.4



Data Preparation

Indicadores de forma e histórico

- Criámos a variável relativa ao confronto direto (*H2H*), onde foram atribuídos pesos consoante o número de confrontos anteriores.
 - ***H2H***
 - ***H2H_Op***
- A partir destas criaram-se variáveis comparativas:
 - ***H2H_Diff***
 - ***H2H_Abs***
- O número total de confrontos foi armazenado na variável:
N_Games



Exemplo:

$$H2H = 0.233, H2H_Op = 0.116$$

$$H2H_Abs = 0.117; H2H_Diff = -0.117$$



Data Preparation

Partidas inválidas

- **Partidas não realizadas:** removemos partidas cujo *Oponent* era “bye”;
- **Partidas duplicadas:** removemos partidas realizadas entre os mesmos jogadores, na mesma ronda, no mesmo torneio e na mesma data, deixando apenas uma partida;
- **Partidas à melhor de 5 sets:** removemos, para uniformizar as previsões (só 2 ou 3 sets).



Data Preparation

Date

- **Date**: a variável **Date** (data do torneio), foi separada em data de início (**Start**) e data final do torneio (**End**);
- **Days**: foi criada a variável **Days** (duração, em dias, do torneio).

Exemplo:

Date = “1999.07.05 - 1999.07.11”

Start = “1999.07.05”; **End** = “1999.07.11”

Days = 5

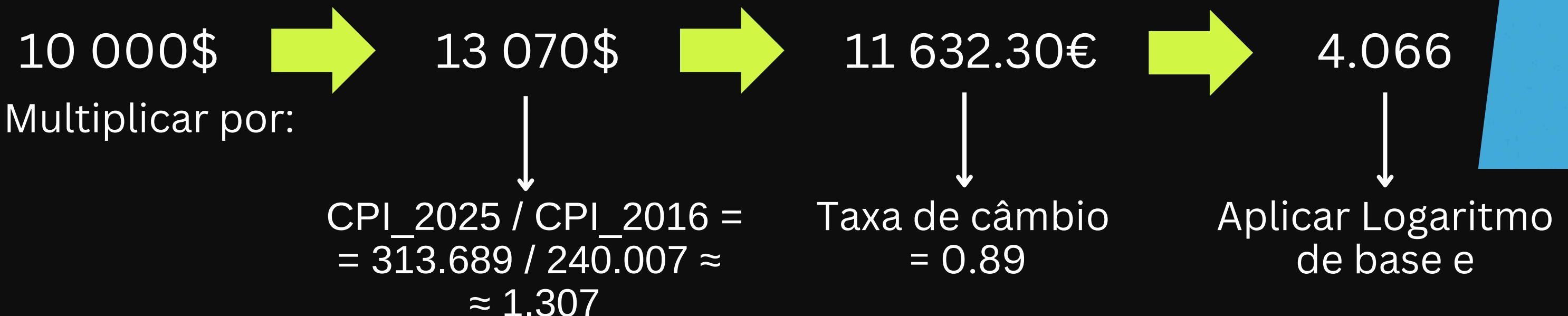


Data Preparation

Prize

- **Prize:** inflacionámos, através de CPI, os valores dos prémios de cada torneio para o valor de 2025;
- Como tínhamos bastantes *outliers*, transformámos o valor em logaritmo de base e.

Exemplo:



Data Preparation

Location

- **Location:** segmentámos a *string* original com base nas vírgulas, criando três colunas auxiliares:
location_1, location_2
- De seguida, fizemos uma correspondência de forma a obtermos a nova variável **Location_City**.

Exemplo:

Location: Arlon, Belgium

location_1: Arlon; **location_2:** Belgium

Location_City: Arlon



Data Preparation

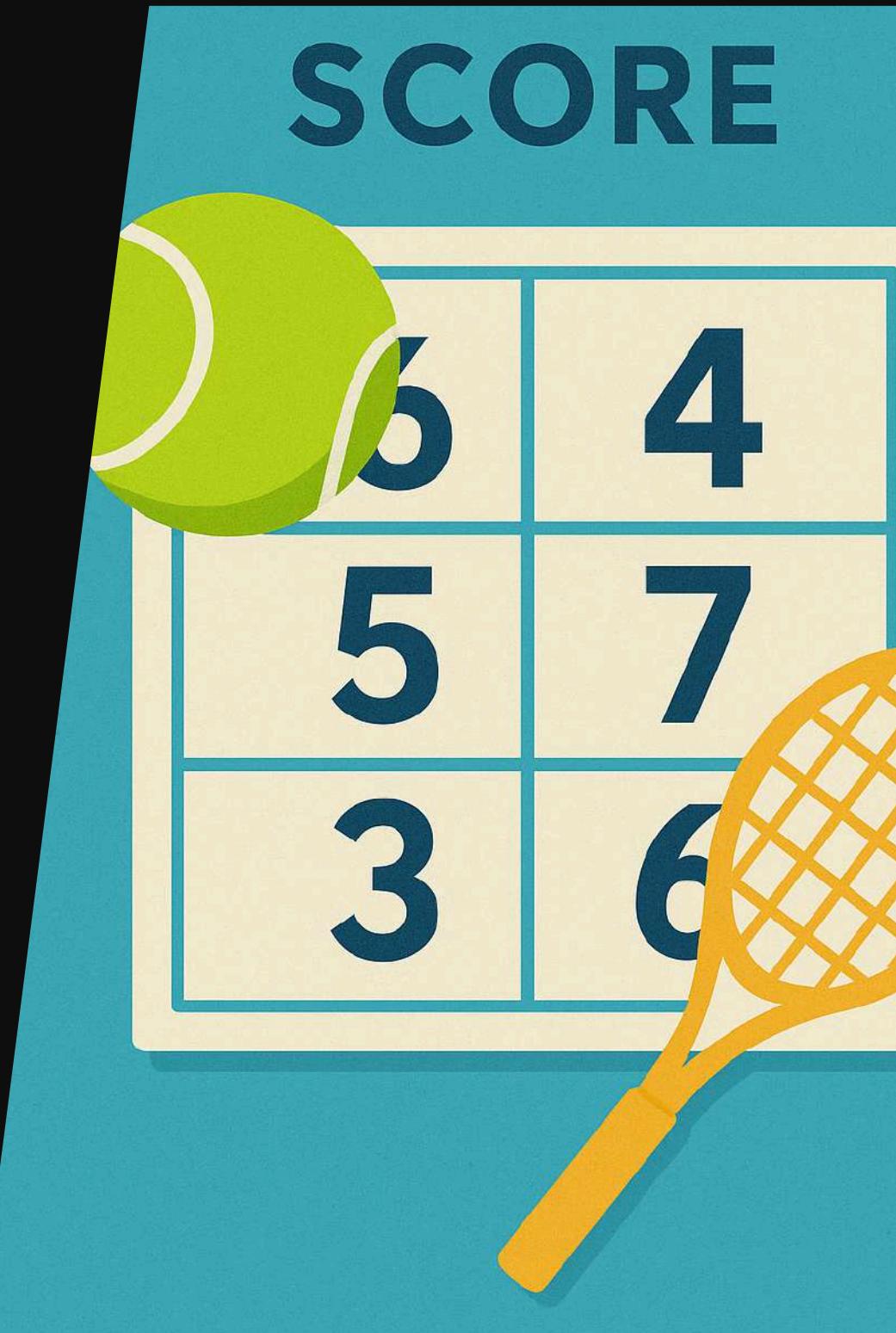
Score

- **Score:** apresenta resultados do tipo: “63 62 62”
- **Sets:** a partir do Score, soma o número de pares de números, separados por espaços.

Exemplo: “63 62 62” → 3 sets

- **Games:** soma todos os algarismos que aparecem na variável Score original.

Exemplo: “63 62 62” → 25 jogos



Escolha de variáveis

VARIÁVEIS EXCLUÍDAS

VARIÁVEIS CATEGÓRICAS COM
MUITOS VALORES ÚNICOS
EX: *BORN_CITY*

VARIÁVEIS INDIVIDUAIS
EX: *RANK_PLAYER*

DATAS

VARIÁVEIS DIFF
EX: *H2H_DIFF*

VARIÁVEIS RELACIONADAS COM
O DESFECHO DO JOGO
EX: *GAMES*

RESUMO - DATA SELECTION

SETS (TARGET)

ABS_RECENT_FORM

AGE_GAP

BACKHAND_VS

DAYS

DIFFERENCE_RANKS_GAP

GAMEROUND

GROUND

GROUND_WINS_ABS

H2H_ABS

HAND_VS

IMC_ABS

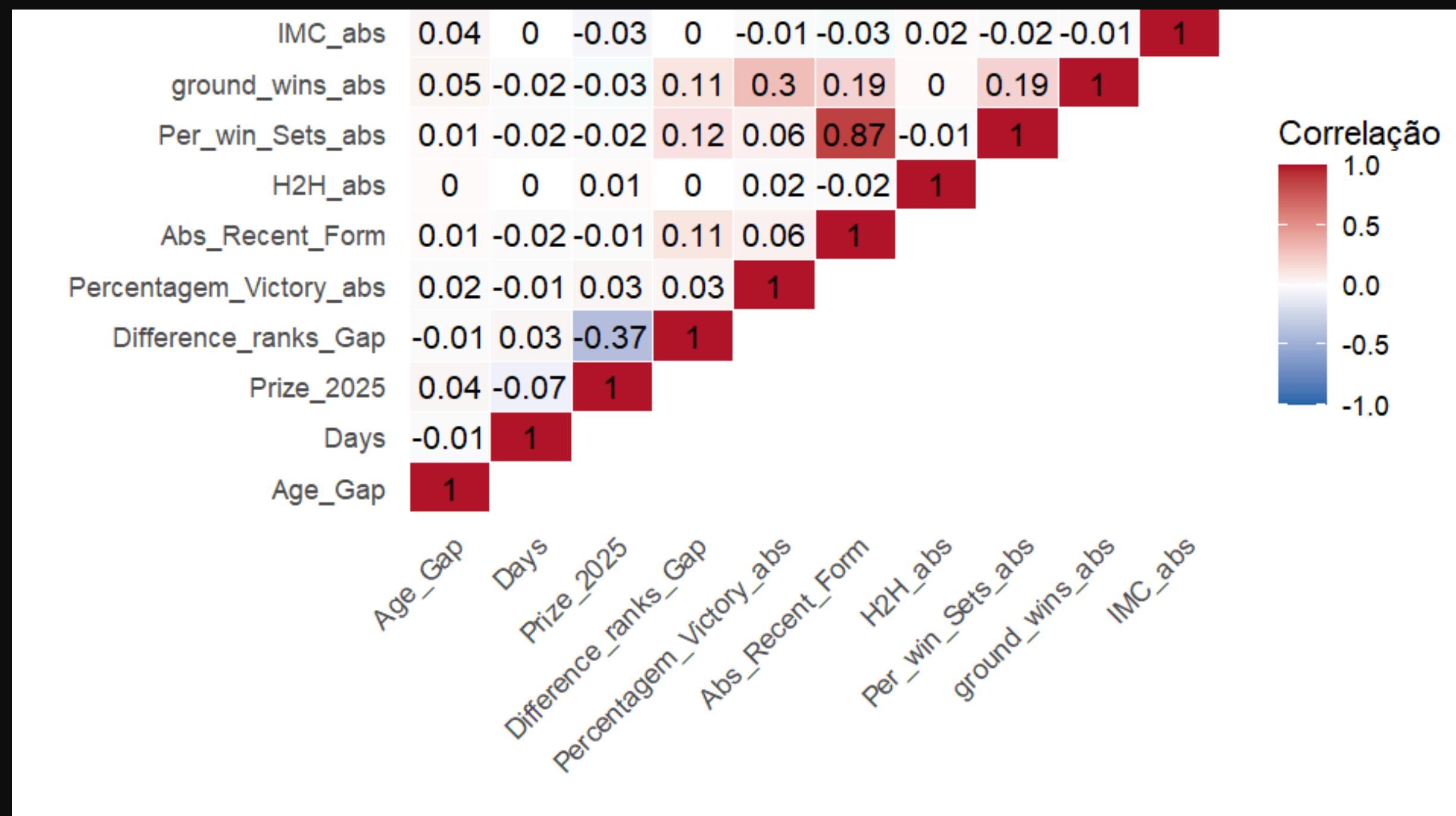
PER_WIN_SETS_ABS

PERCENTAGEM_VICTORY_ABS

PRIZE_2025

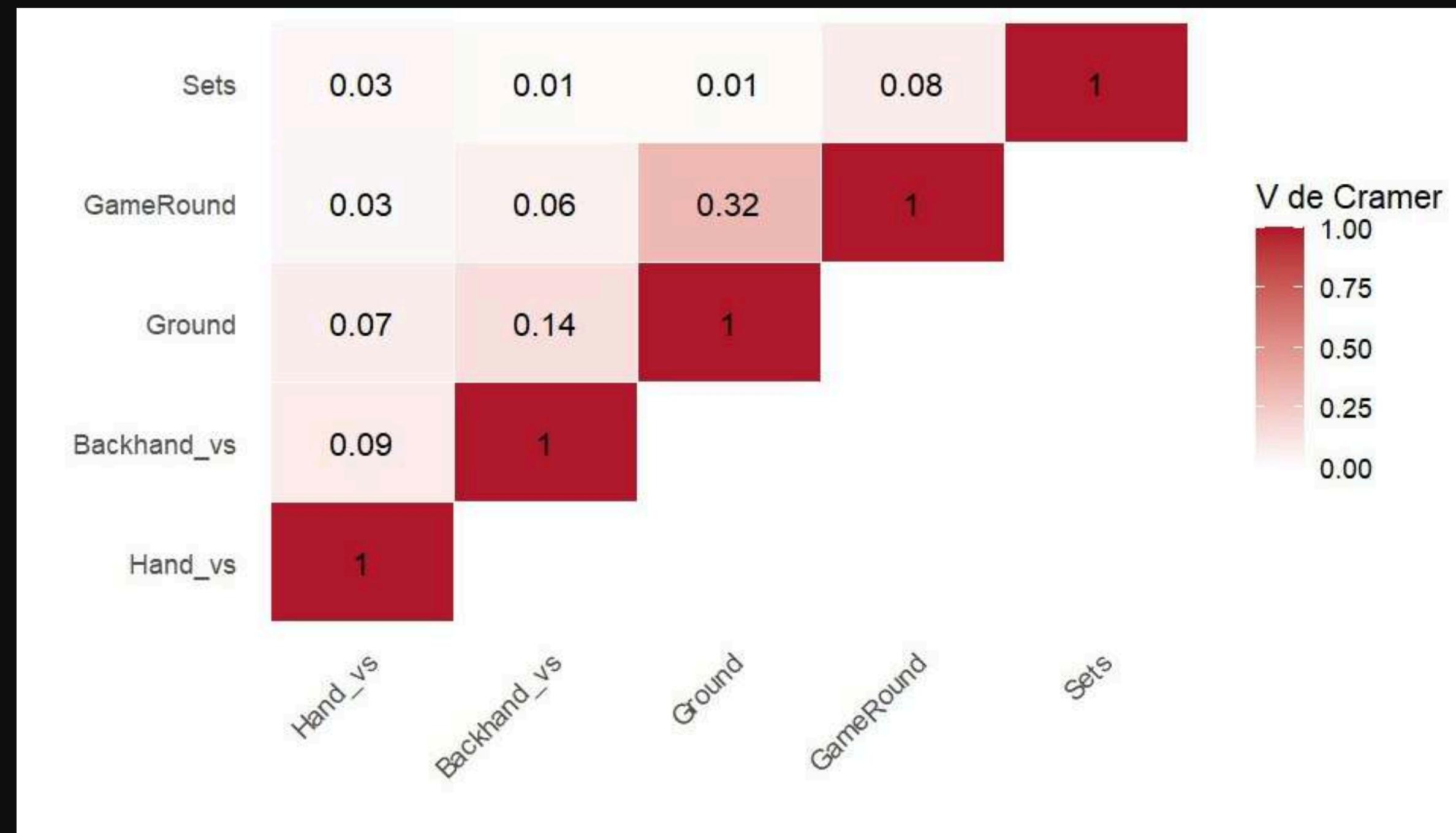
Correlações Variáveis Numéricas

R de Pearson



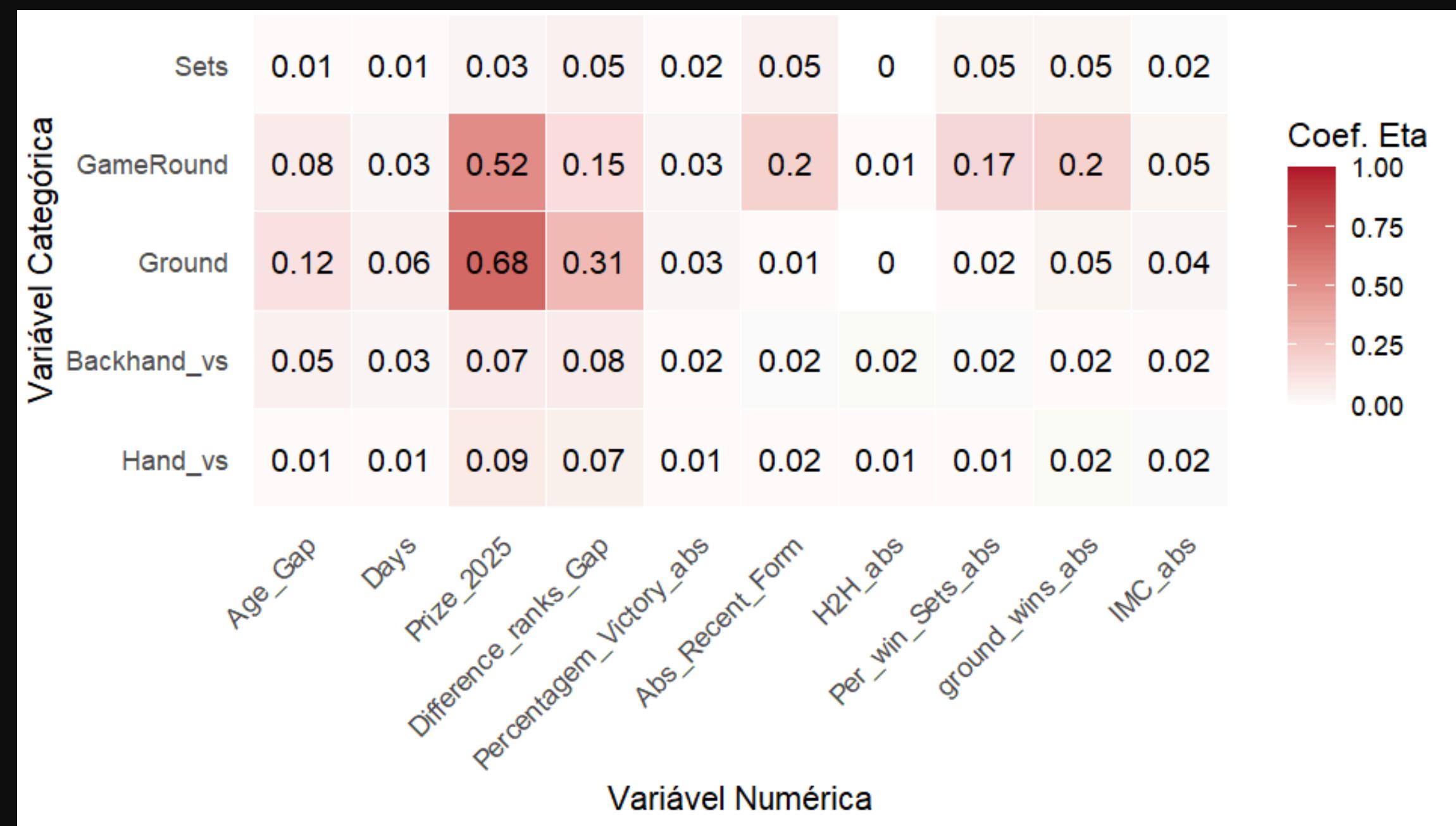
Correlações Variáveis Categóricas

V de Cramer



Correlações Categóricas vs Numéricas

Coeficiente ETA



Multicolinearidade

Critério VIF (Valor Crítico: > 10)

IMC_ABS	HAND_VS	BACKHAND_VS	AGE_GAP
1.00	1.01	1.02	1.01
DAY	PRIZE_2025	GROUND	GAMEROUND
1.00	1.87	1.32	1.05
DIFFERENCE_RANKS_GAP	PERCENTAGEM_VICTORY_ABS	ABS_RECENT_FORM	H2H_ABS
1.14	1.05	1.97	1.00
PER_WIN_SETS_ABS		GROUND_WINS_ABS	
1.96		1.09	

Data Preparation

Tratamento dos omissos e equilíbrio do dataset

- **Variáveis categóricas** → moda
- **Variáveis numéricas** → média
- **Problema:** algumas variáveis tinham muitos valores omissos e foram imputados muitos dados.
- **2 sets = 3866 (68.4%);**
- **3 sets = 1781 (31.6%).**
- **Equilíbrio do dataset** → *undersampling*
- Realizar *oversampling* criava ainda mais dados artificiais.



Modelo 1

Regressão logística

		Previsto	
		2	3
Real	2	1041	740
	3	901	880

Accuracy = 0.5394

Crossvalidation: k = 5

Modelo 2

Árvore de Decisão

		Previsto	
		2	3
Real	2	904	877
	3	813	968

Accuracy = 0.5256

Crossvalidation: k = 5

Modelo 3

Random Forest

		Previsto	
		2	3
Real	2	922	859
	3	766	1015

Accuracy = 0.5438

Crossvalidation: k = 5

Modelo 4

Gradient Boosting

		Previsto	
		2	3
Real	2	920	861
	3	779	1002

Accuracy = 0.5396

Crossvalidation: k = 5

Evaluation

Accuracy Comparison

	Regressão Logistica	Árvore de Decisão	<i>Random Forest</i>	<i>Gradient Boosting</i>
Accuracy	0.5394	0.5256	0.5438	0.5396

Conclusão:

Todos os modelos têm performance semelhante, sendo o ***Random Forest*** o melhor.

Nenhum dos modelos é muito melhor que o acaso (**accuracy = 0.5** em classificações binárias), logo, a sua utilização não cumpre os objetivos propostos.

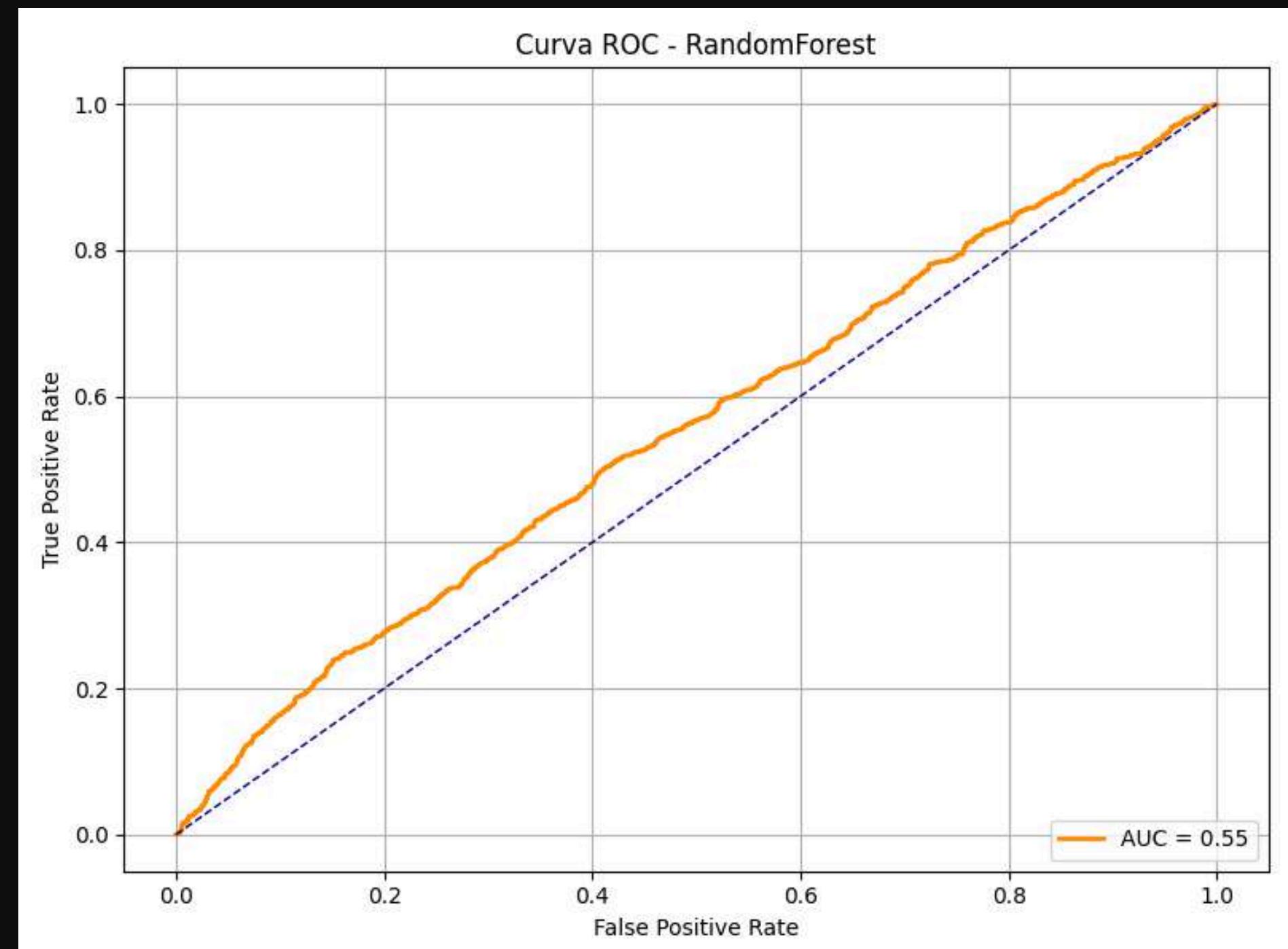
Evaluation

ROC Curve - Best Model

A curva representa a capacidade do modelo em distinguir entre classes, comparando:

- Taxa de acerto em 2 ***Sets*** (**y**);
- Taxa de erro em 2 ***Sets*** (**x**).

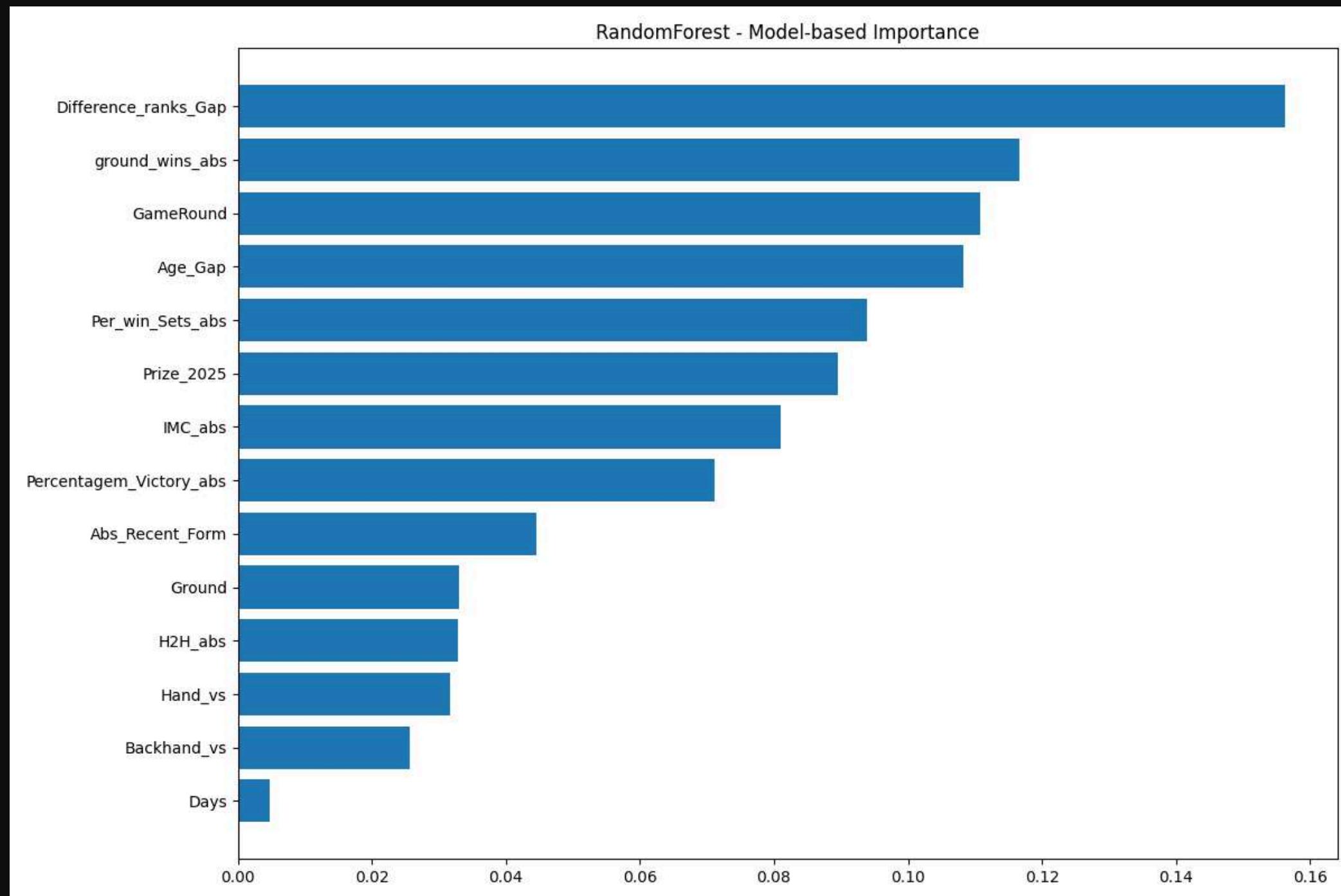
AUC: 0.55



Evaluation

Variable Importance - Best Model

- Valores mais elevados traduzem variáveis que separam melhor os dados.



Fim!