INSTITUTO UNIVERSITÁRIO DE LISBOA

Projeto Aplicado em Ciência de Dados I

PREVISÃO DO NÚMERO DE SETS EM JOGOS DO ATP TOUR

Grupo 9, CDB2

Francisco Rodrigues 105427

Margarida Carvalho 104765

Maria Margarida Pereira 105877

Simão Fonseca 105251

Docentes:

Diana Mendes e Sérgio Moro

Maio de 2023

CONTRIBUIÇÃO PRÁTICA DO PROBLEMA

- Problema de classificação
- OBJETIVO: prever o número de sets
- APLICABILIDADE: apostas desportivas

BASE DE DADOS INICAL – CHINA

NULLS

- Número de instâncias: 26 357
- Número de variáveis: 15

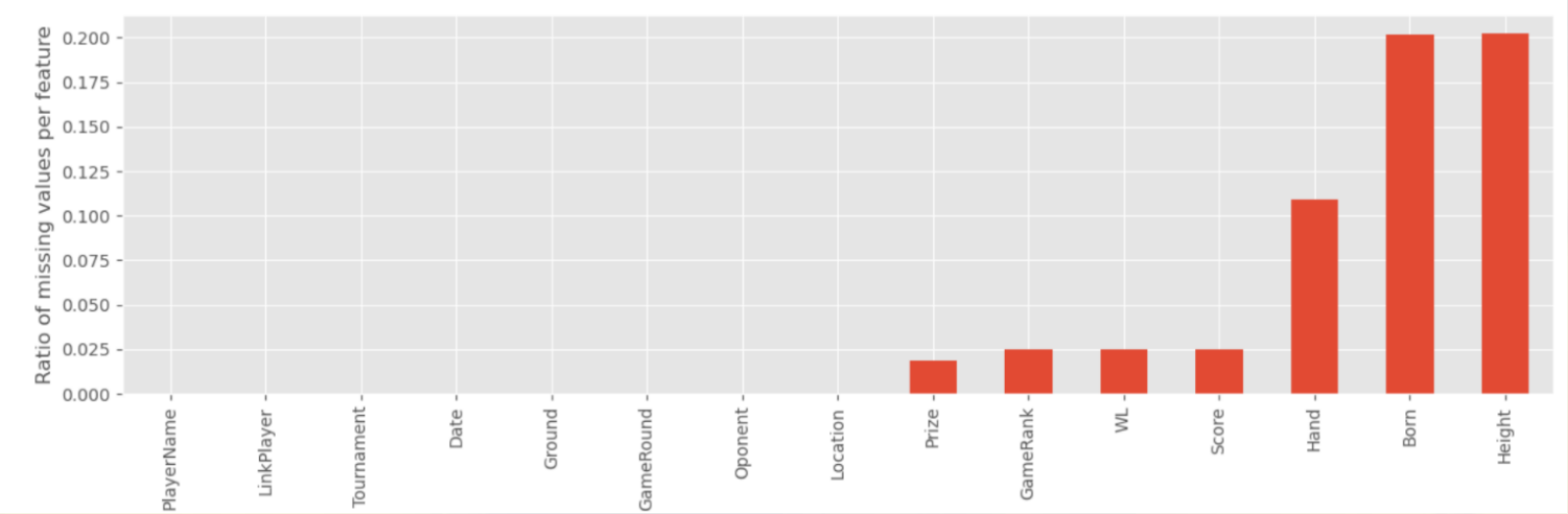


Figura 1 - Percentagem de valores omissos por variável

Variables	Number of nulls
PlayerName	0
Born	5314
Height	5341
Hand	2882
LinkPlayer	0
Tournament	0
Date	0
Ground	0
Prize	485
GameRound	0
GameRank	650
Oponent	0
WL	650
Score	652
Location	0

Tabela 1 - Número de valores nulos para cada variável

BASE DE DADOS INICAL – CHINA

VALORES ÚNICOS

Variables	Number of unique values
Location	1
WL	2
Ground	3
Hand	7
GameRound	11
Height	22
Prize	55
Tournament	118
Date	380
Born	851
PlayerName	1663
LinkPlayer	1663
GameRank	1849
Oponent	2013
Score	2086

Tabela 2 - Número de valores únicos para cada variável

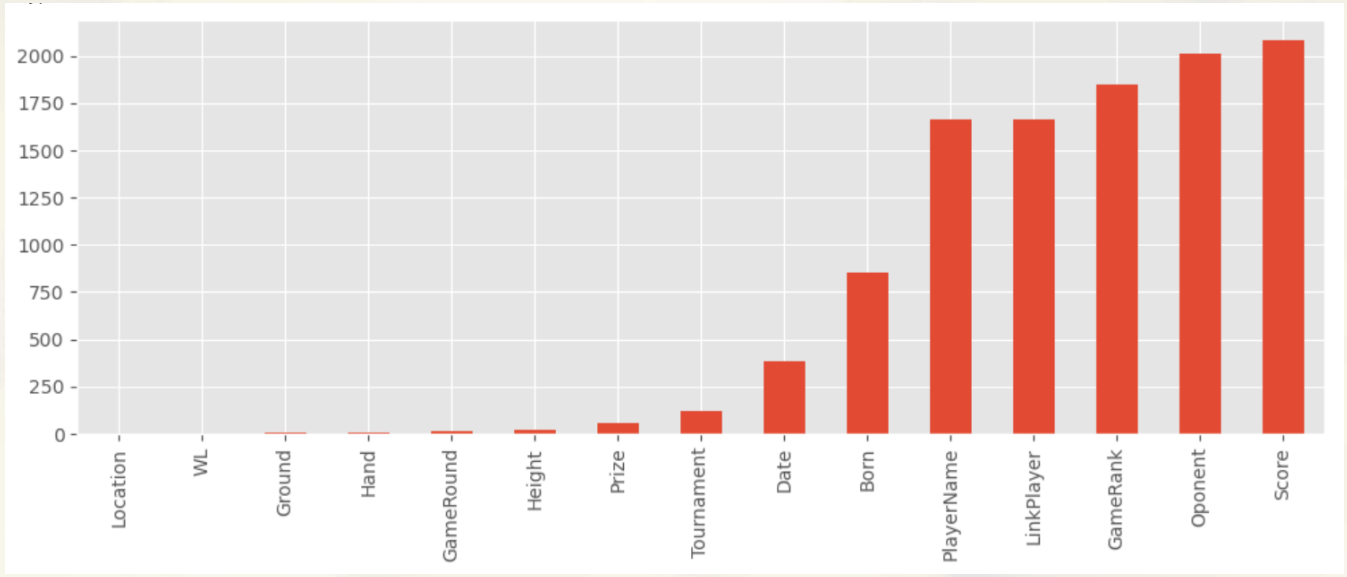


Figura 2 - Valores únicos para cada variável, por ordem crescente

BASE DE DADOS FINAL – CHINA

COM TODAS AS VARIÁVEIS

- Número de instâncias: 12 708
- Número de variáveis: 65

BASE DE DADOS FINAL – CHINA
FILTRADA

- Número de instâncias: 12 708
- Número de variáveis: 16

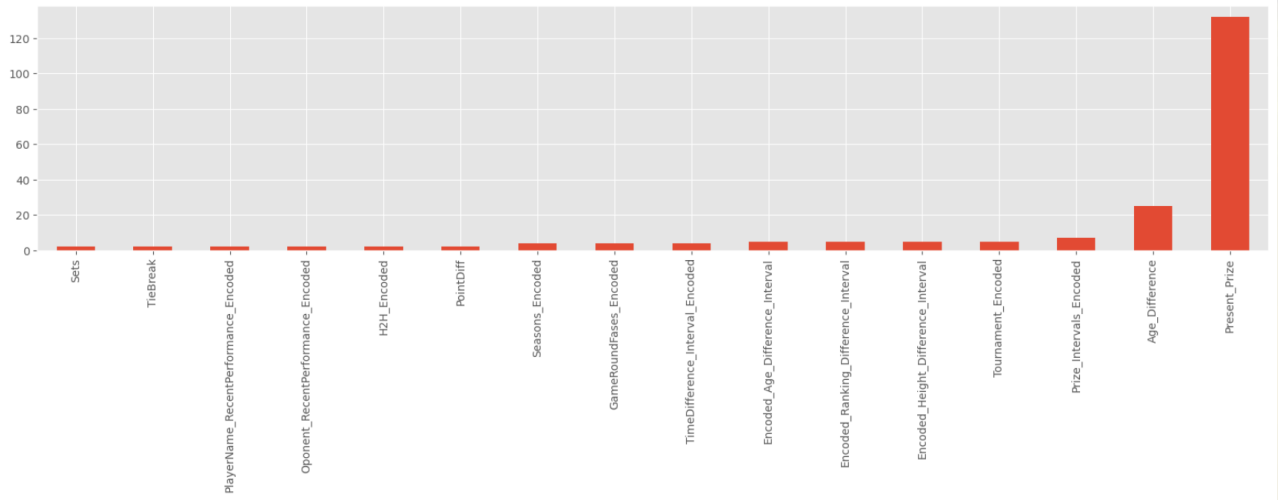


Figura 3 - Valores únicos para cada variável, por ordem crescente

Variables	Number of unique values
Sets	2
Seasons_Encoded	4
GameRoundFases_Encoded	4
Encoded_Age_Difference_Interval	5
Encoded_Ranking_Difference_Interval	5
Encoded_Height_Difference_Interval	5
TieBreak	2
Tournament_Encoded	5
PlayerName_RecentPerformance_Encoded	2
Oponent_RecentPerformance_Encoded	2
H2H_Encoded	2
TimeDifference_Interval_Encoded	4
Present_Prize	132
Prize_Intervals_Encoded	7
PointDiff	2
Age_Difference	25

Tabela 3 - Número de valores únicos para cada variável

IMPORTÂNCIA DAS VARIÁVEIS

Índice da variável	Variável	Importância
13	PointDiff	0.108941
3	Encoded_Ranking_Difference_Interval	0.096790
12	Prize_Intervals_Encoded	0.086076
8	Oponent_RecentPerformance_Encoded	0.066441
10	TimeDifference_Interval_Encoded	0.061902
5	TieBreak	0.061882
9	H2H_Encoded	0.061860
2	Encoded_Age_Difference_Interval	0.060752
6	Tournament_Encoded	0.059704
1	GameRoundFases_Encoded	0.059150
11	Present_Prize	0.056179
4	Encoded_Height_Difference_Interval	0.055886
0	Seasons_Encoded	0.055299
7	PlayerName_RecentPerformance_Encoded	0.055240
14	Age_Difference	0.053898

Tabela 4 - Valores da importância das variáveis para o Random Forest (por ordem decrescente)

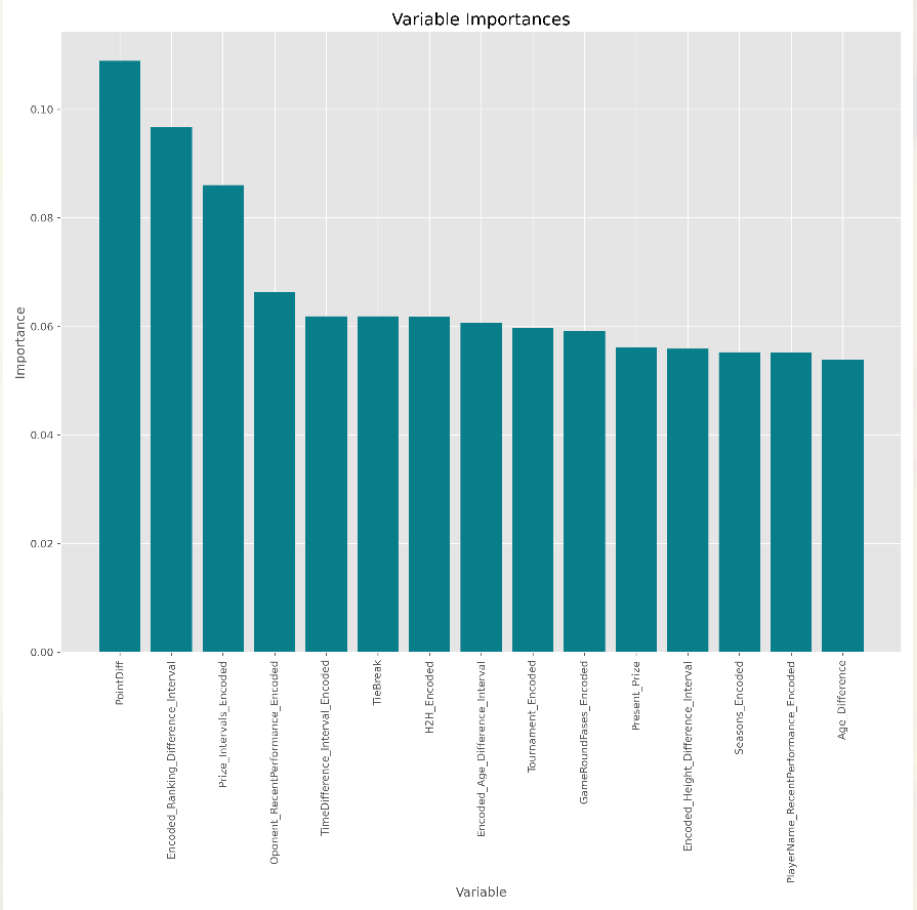
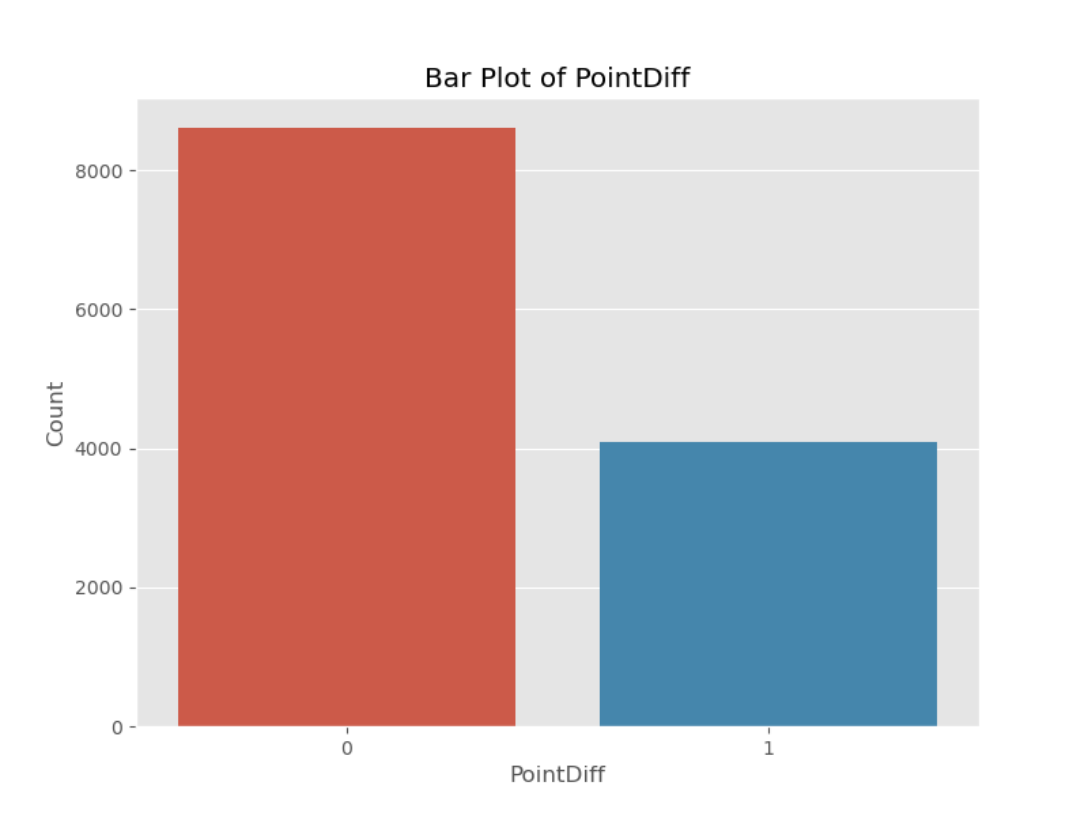


Figura 4 - Importância das variáveis para o modelo Random Forest (por ordem decrescente)

PointDiff



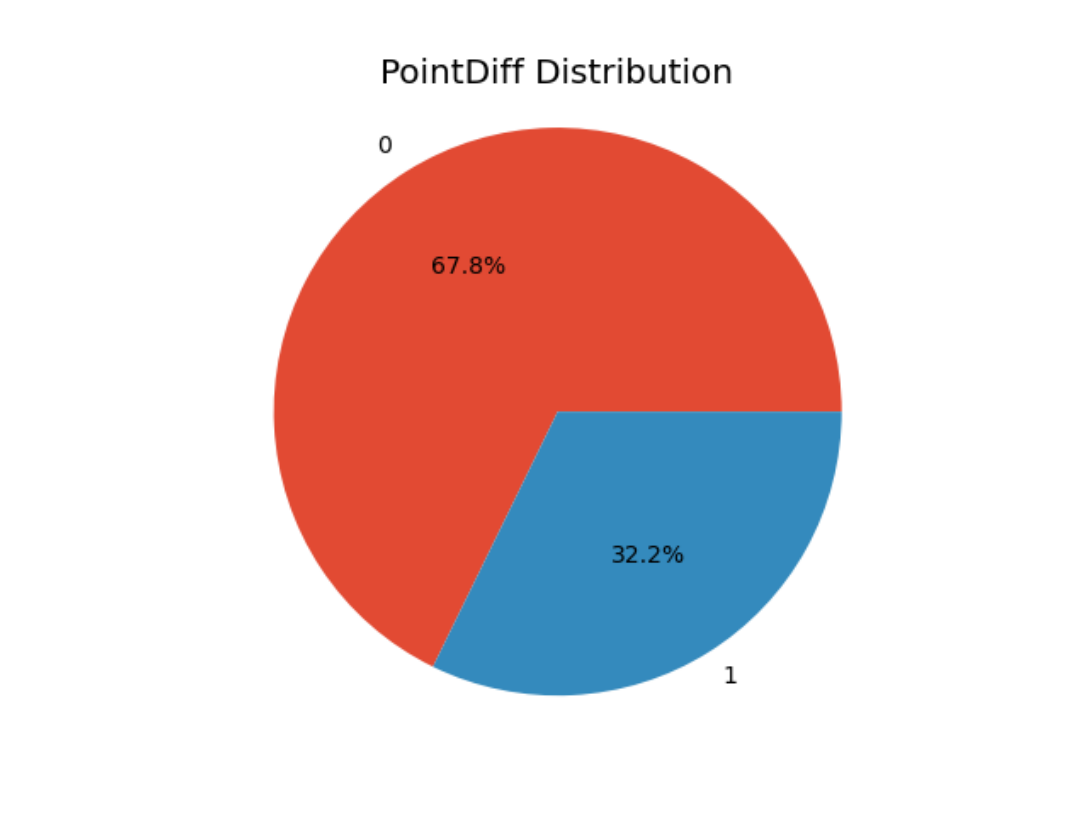
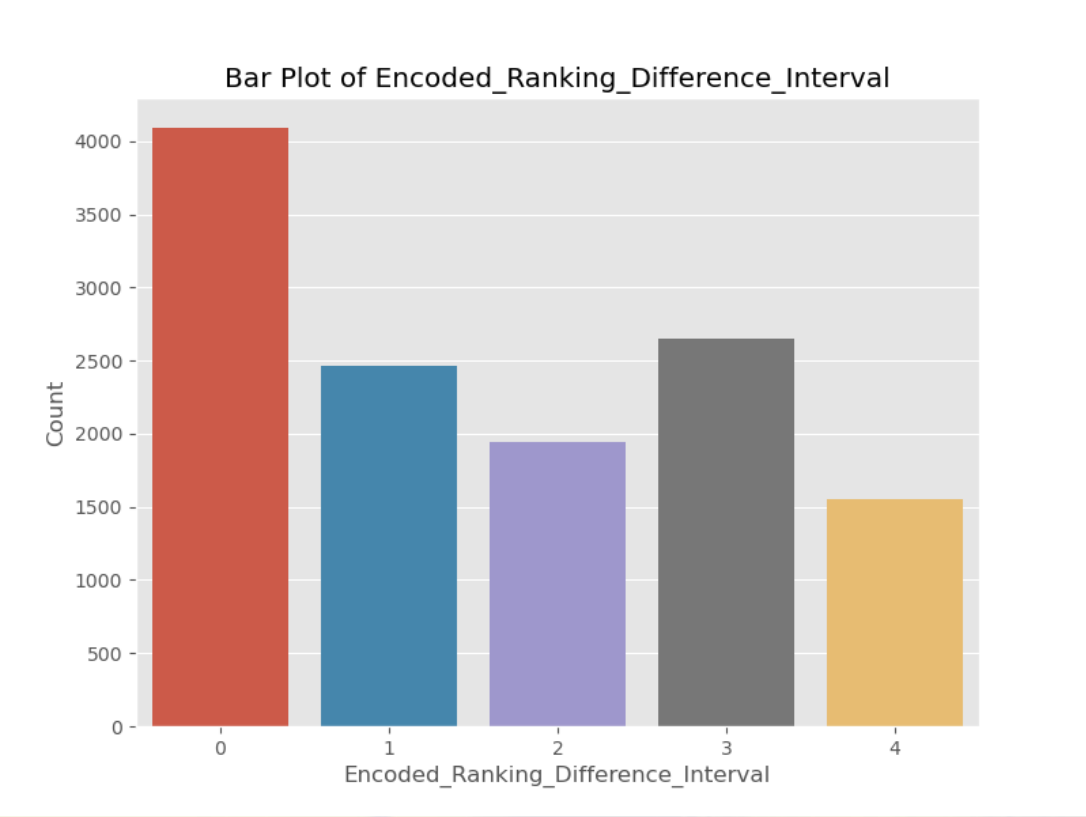


Figura 5 – Distribuição dentro da variável PointDiff

Encoded_Ranking_Difference_Interval



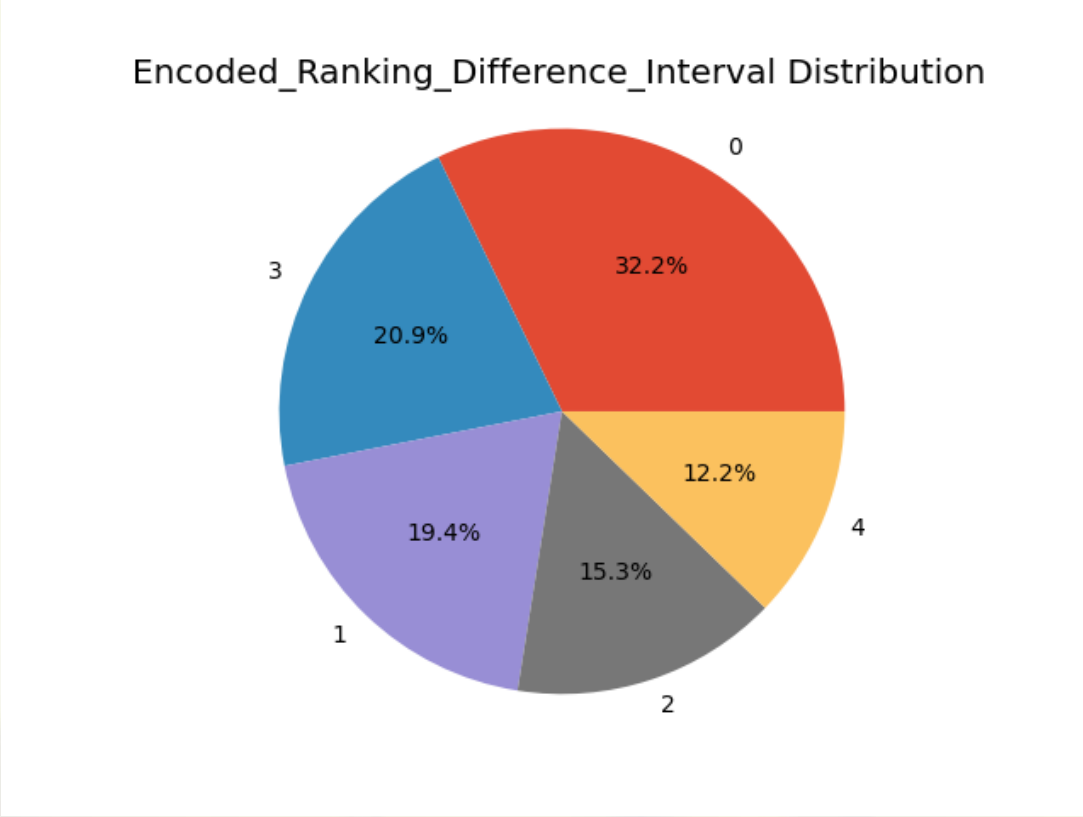
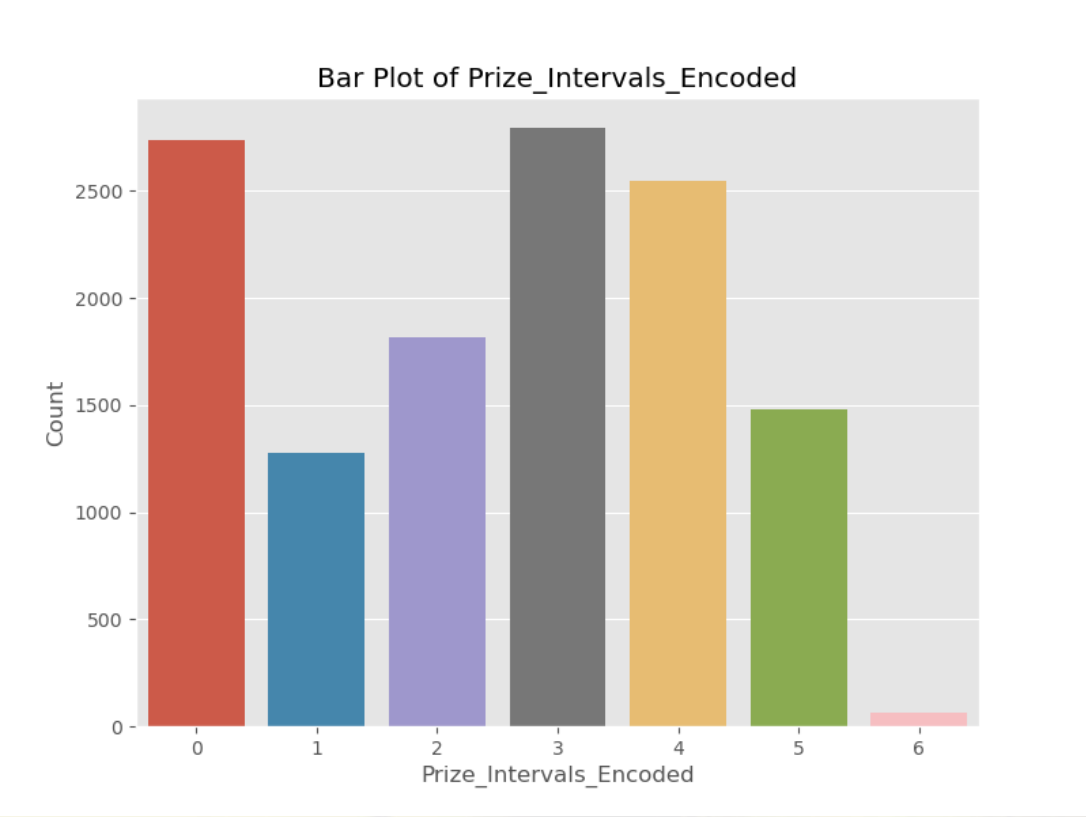


Figura 6 – Distribuição dentro da variável PointDiff

Prize_Intervals_Encoded



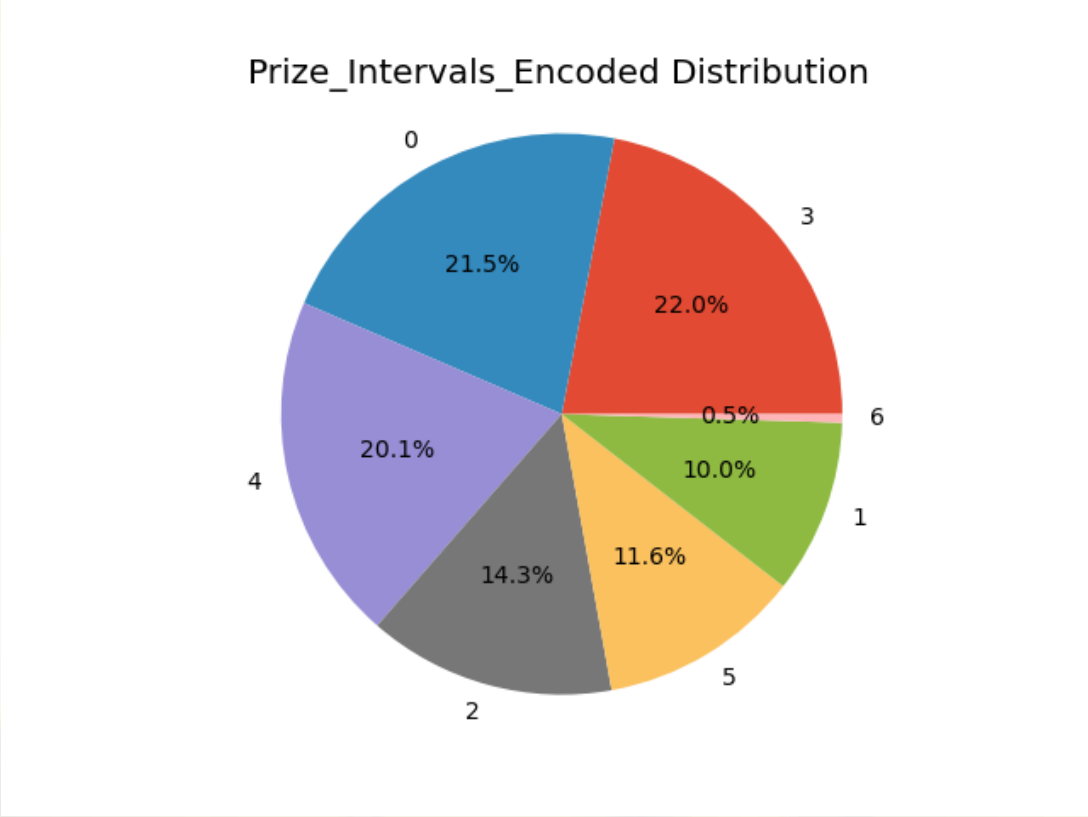


Figura 7 – Distribuição dentro da variável PointDiff

MODELOS

- Divisão entre treino e teste: 80/20
- Modelos feito com Stratified Sampling
- Variáveis usadas nos 2 modelos:
  - Modelo Pré-Jogo: 'Encoded_Ranking_Difference_Interval'; 'Prize_Intervals_Encoded'; 'H2H_Encoded'; 'Oponent_RecentPerformance_Encoded'; 'TimeDifference_Interval_Encoded';
  - Modelo Pós-Primeiro Set: 'PointDiff'; 'TieBreak'; 'Encoded_Ranking_Difference_Interval'; 'Prize_Intervals_Encoded'; 'Oponent_RecentPerformance_Encoded'; 'TimeDifference_Interval_Encoded'; 'H2H_Encoded'

MODELO PRÉ-JOGO

Accuracy (overall correct predictions)	0.72
Auc	0.57
Recall (all 1s predicted right):	0.02
Precision (confidence when predicting a 1)	0.4

Tabela 5 - Métricas para avaliação do modelo Pré-Jogo

	precision	**recall**	**f1-score**	**support**
0	0.73	0.99	0.84	1844
1	0.40	0.02	0.05	698

Tabela 6 - Métricas para avaliação do modelo Pré-Jogo para cada classe

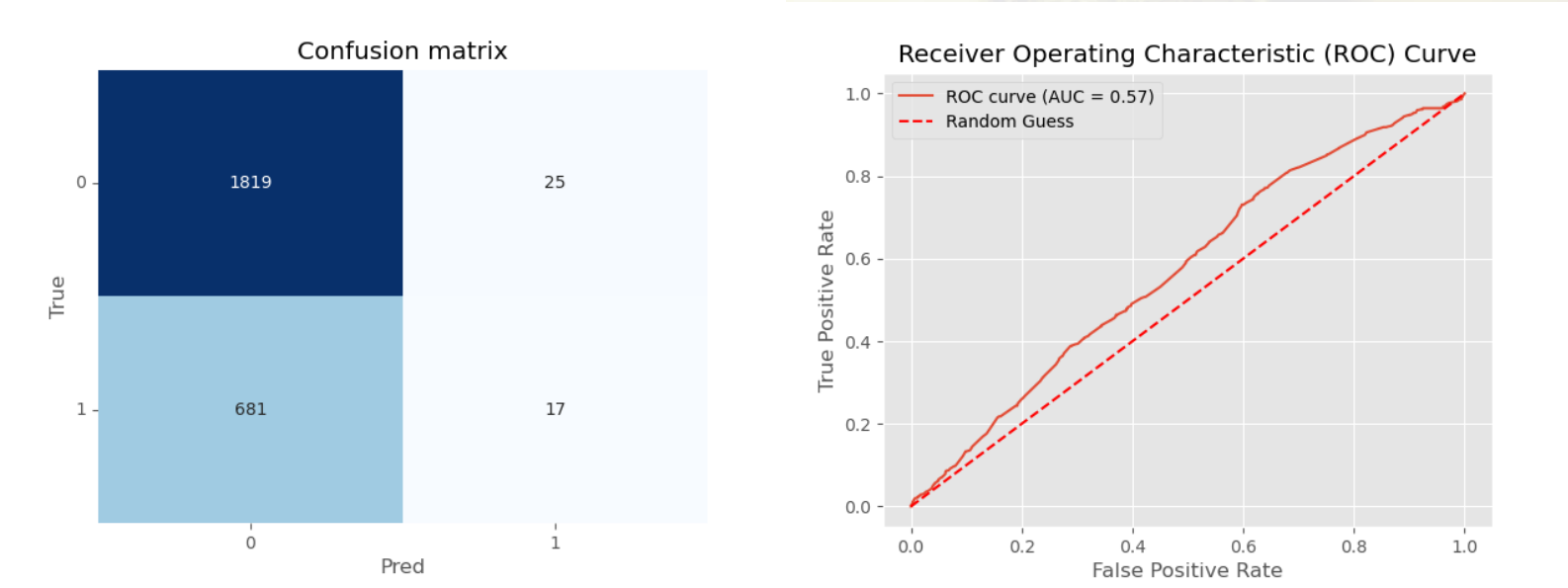


Figura 26 - Matriz de confusão e curva ROC do modelo Pré-Jogo

MODELO PÓS-PRIMEIRO SET

Accuracy (overall correct predictions)	0.71
AUC	0.58
Recall (all 1s predicted right):	0.06
Precision (confidence when predicting a 1)	0.33

Tabela 7 - Métricas para avaliação do modelo Pós-Primeiro Set

	precision	**recall**	**f1-score**	**support**
0	0.73	0.96	0.83	1844
1	0.33	0.06	0.10	698

Tabela 8 - Métricas para avaliação do modelo Pós-Primeiro Set para cada classe

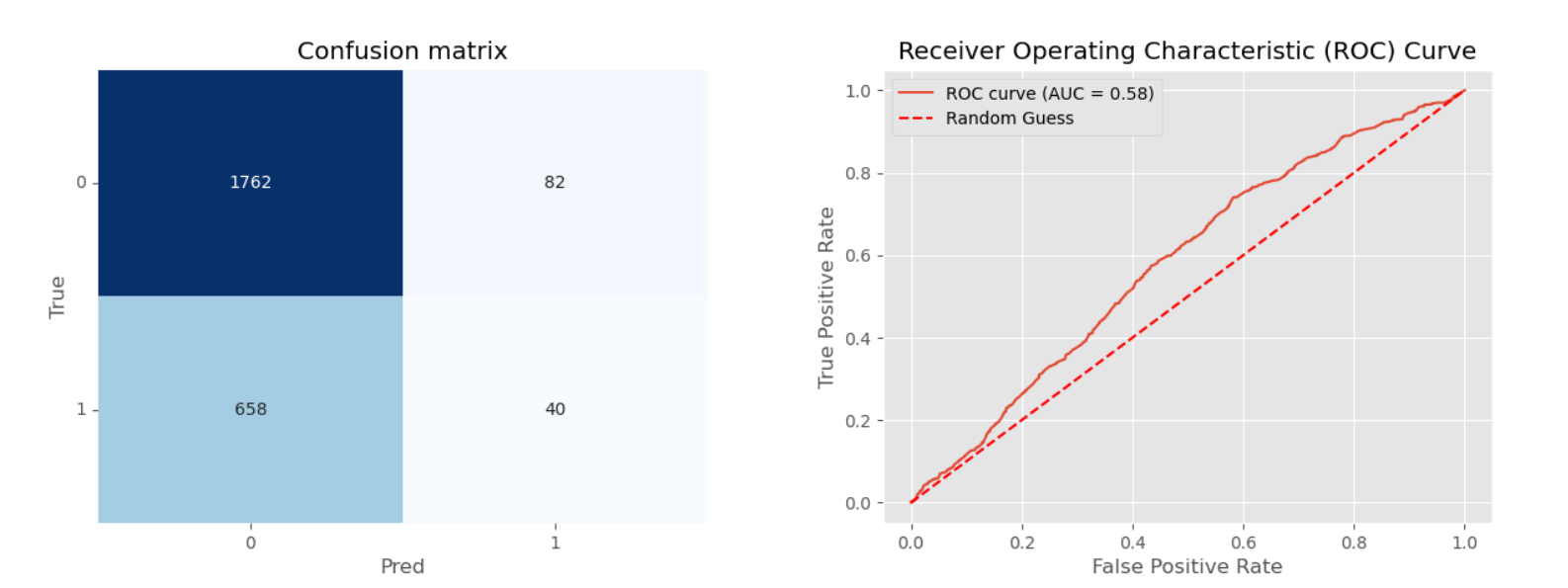


Figura 27 - Matriz de confusão e curva ROC do modelo Pós-Primeiro Set

CONCLUSÃO

- Os modelos alcançados aliados ao conhecimento de ténis e espírito crítico de cada pessoa, formam uma ferramenta útil para os entusiastas das apostas desportivas;

OBRIGADO

Grupo 9, CDB2

Francisco Rodrigues 105427

Margarida Carvalho 104765

Maria Margarida Pereira 105877

Simão Fonseca 105251