INSTITUTO UNIVERSITÁRIO DE LISBOA

PROJETO APLICADO EM CIÊNCIA DE DADOS I

MODELING

LICENCIATURA EM CIÊNCIA DE DADOS

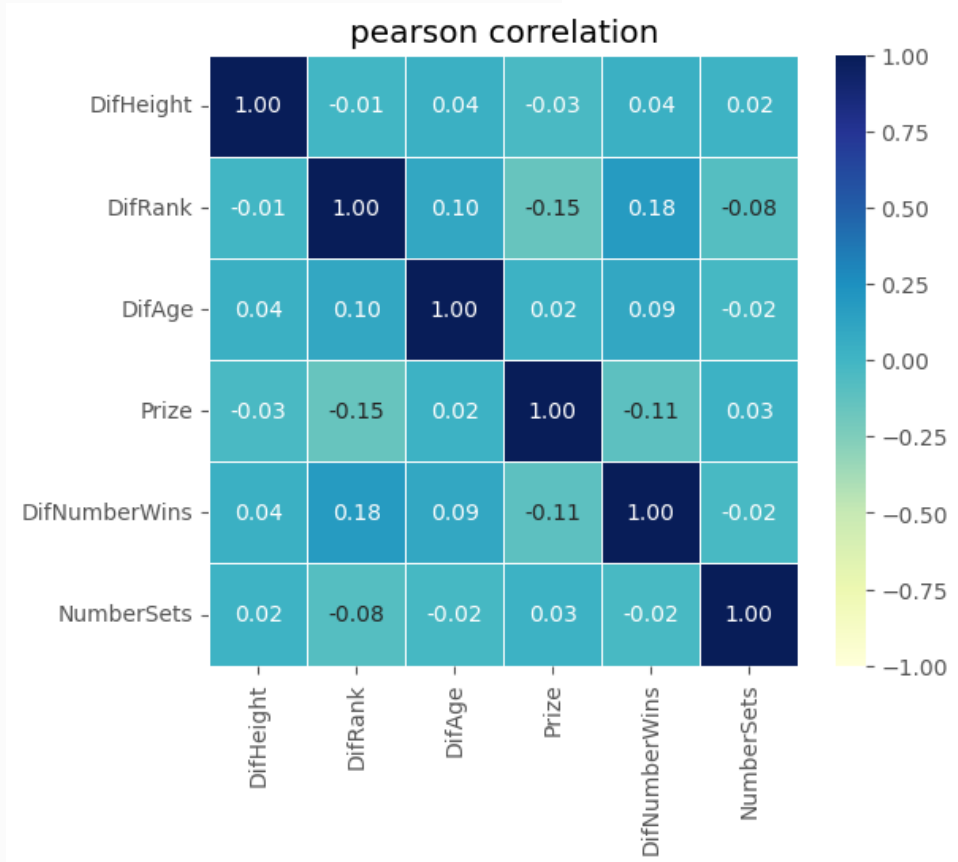
Base de Dados ATP – Brasil

Grupo 5: nº 103303, nº 110451, nº 104716, nº 99239

Docentes: Diana Aldea Mendes e Sérgio Moro

17 de maio de 2023

Correlações entre variáveis corrigida



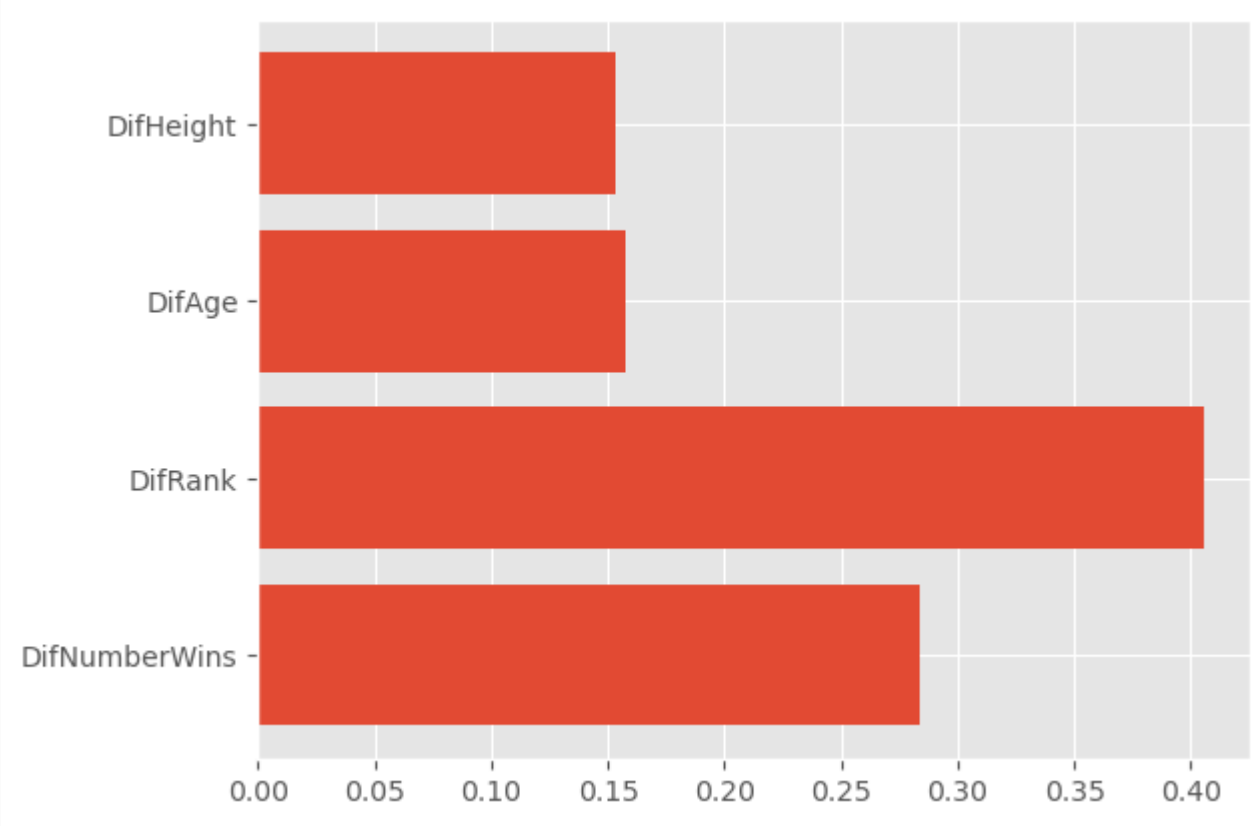
Variáveis numéricas

	NumberSets	L_OR_R_Left-Handed	L_OR_R_Right-Handed	L_OR_R_Opponent_Left-Handed	L_OR_R_Opponent_Right-Handed	Ground_Clay	Ground_Hard
NumberSets	1.000000	0.030467	0.029869	0.034840	0.033865	0.012205	0.006657
L_OR_R_Left-Handed	0.030467	1.000000	0.998751	0.030328	0.030902	0.000641	0.000029
L_OR_R_Right-Handed	0.029869	0.998751	1.000000	0.030709	0.031284	0.000004	0.000583
L_OR_R_Opponent_Left-Handed	0.034840	0.030328	0.030709	1.000000	0.998121	0.029661	0.022897
L_OR_R_Opponent_Right-Handed	0.033865	0.030902	0.031284	0.998121	1.000000	0.030591	0.023787
Ground_Clay	0.012205	0.000641	0.000004	0.029661	0.030591	1.000000	0.949508
Ground_Hard	0.006657	0.000029	0.000583	0.022897	0.023787	0.949508	1.000000

Variáveis categóricas (dummy)

Variáveis consideradas no modelo 1

Variável	Justificação
DifNumberWins	Mais jogos vencidos -> mais experiência no desporto -> menos sets irão ser necessários
DifRank	Maior rank -> mais experiência e perícia -> mais talento no desporto -> menos sets serão precisos
DifAge	Poderá condicionar o nível de experiência, melhor condição física, recuperação de lesões, mais energia, níveis de resistência
DifHeight	Um jogador mais alto poderá cobrir mais terreno, um jogador mais baixo poderá ter mais vantagem na movimentação -> quanto maior a diferença entre ambos -> mais desequilibrado poderá estar o jogo -> menos sets durará



Total = 4 variáveis para a previsão

Gráfico: Feature Importance

Modelo 1 – Gradient Boosting (já apresentado)

MELHOR DE 3

	precision	recall	f1-score	support
1.0	0.000000	0.000000	0.000000	7.000000
2.0	0.658031	1.000000	0.793750	1651.000000
3.0	0.000000	0.000000	0.000000	851.000000
accuracy	0.658031	0.658031	0.658031	0.658031
macro avg	0.219344	0.333333	0.264583	2509.000000
weighted avg	0.433005	0.658031	0.522312	2509.000000

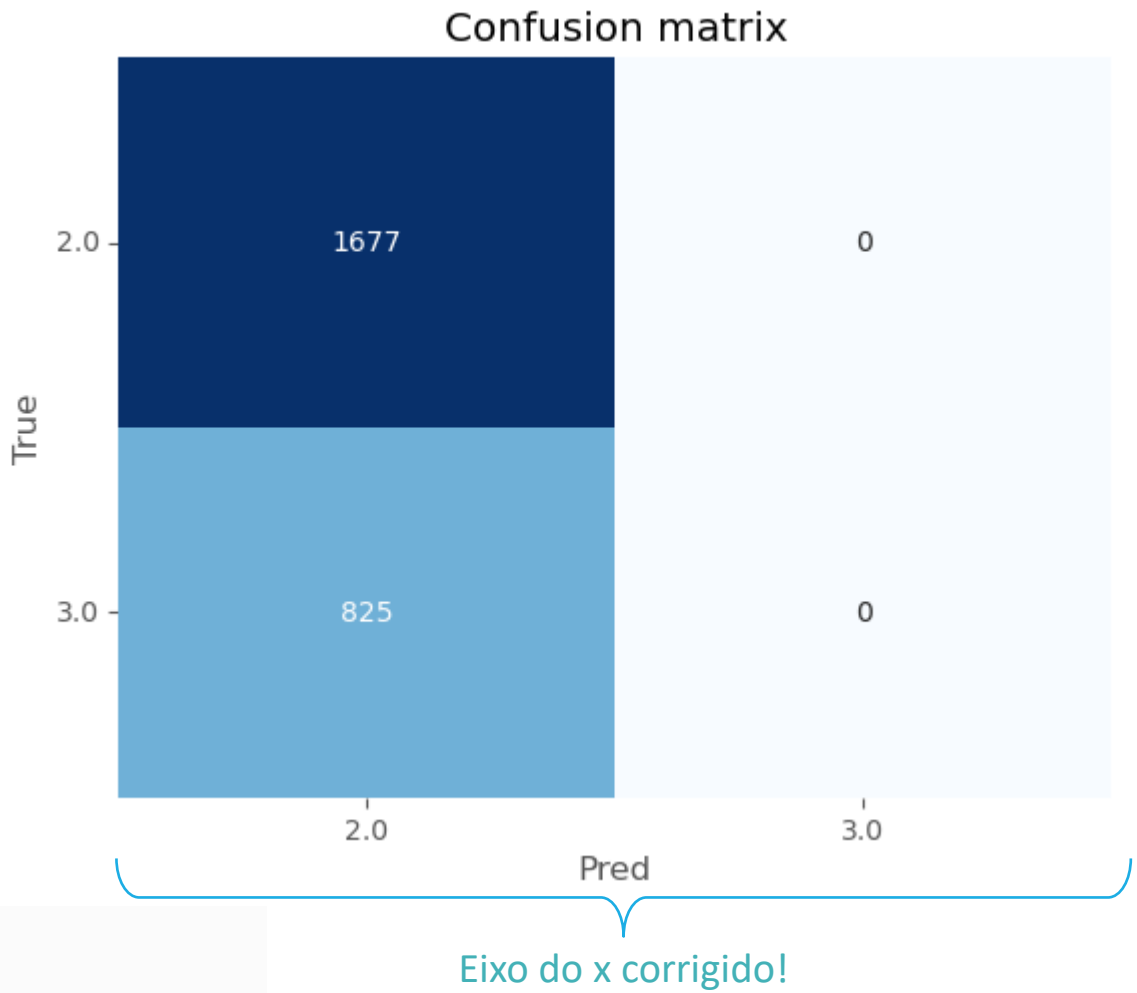
MELHOR DE 5

	precision	recall	f1-score	support
4.0	0.666667	0.200000	0.307692	10.000000
5.0	0.272727	0.750000	0.400000	4.000000
accuracy	0.357143	0.357143	0.357143	0.357143
macro avg	0.469697	0.475000	0.353846	14.000000
weighted avg	0.554113	0.357143	0.334066	14.000000

Primeira tentativa de melhoramento do modelo 1- tirar os sets de 1

MELHOR DE 3

	precision	recall	f1-score	support
2.0	0.670264	1.000000	0.802584	1677.000000
3.0	0.000000	0.000000	0.000000	825.000000
accuracy	0.670264	0.670264	0.670264	0.670264
macro avg	0.335132	0.500000	0.401292	2502.000000
weighted avg	0.449254	0.670264	0.537943	2502.000000



Eixo do x corrigido!

Observações:

- A performance não melhorou significativamente;
- O modelo continua a prever apenas sets de 2

Outras tentativas de melhoramento deste modelo em progresso

Tentativa 2: Retirar a variável altura da base de dados.

Tentativa 3: Contabilizar o total de jogos vencidos para todos os países ao invés de ter apenas a contagem dos jogos vencidos no Brasil.

Modelo 2 – Utilização de variáveis dummy “L_OR_R “

Variáveis usadas
DifNumberWins
DifRank
DifAge
DifHeight
L_OR_R_player-right-handed
L_OR_R_player-left-handed
L_OR_R_oponent-right-handed
L_OR_R_oponent-left-handed

MELHOR DE 3

	precision	recall	f1-score	support
2.0	0.67	0.65	0.66	1685
3.0	0.31	0.33	0.32	817
accuracy			0.55	2502
macro avg	0.49	0.49	0.49	2502
weighted avg	0.55	0.55	0.55	2502

Razão de utilizar a variável L_OR_R:
Quanto mais confortável o jogador está a utilizar determinada mão -> com mais precisão irá jogar -> poderá ter mais vantagem jogo -> menos sets

Observações:
- A performance do modelo está ligeiramente mais baixa.
- Este modelo prevê sets de 3!

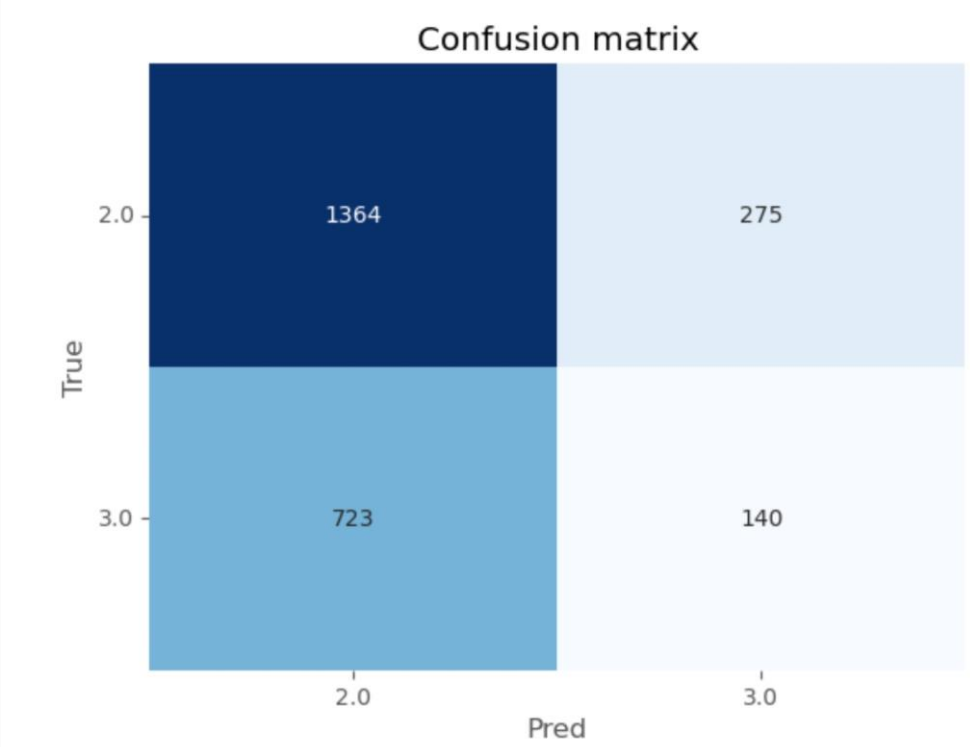
Tentativa de melhoramento do modelo
2- Algortimo XGBoost

Variáveis usadas (iguais ao modelo anterior)
DifNumberWins
DifRank
DifAge
DifHeight
L_OR_R_player-right-handed
L_OR_R_player-left-handed
L_OR_R_oponent-right-handed
L_OR_R_oponent-left-handed

Observações:

- A performance do modelo está mais alta.
- O modelo já não prevê a moda.

MELHOR DE 3



Accuracy 0.601

Modelo 2 – Utilização de variáveis dummy “Ground”

Variáveis usadas
DifNumberWins
DifRank
DifAge
DifHeight
Ground-clay
Ground-hard

Razão de utilizar a variável “Ground”:
Pode haver jogadores que se sintam mais confortáveis a jogar num dado piso e, nesses casos, o número de sets pode ser menor, e/ou podem ter mais prática num tipo de piso em comparação com o outro -> menos sets

MELHOR DE 3

	precision	recall	f1-score	support
2.0	0.68	0.64	0.66	1680
3.0	0.34	0.39	0.36	822
accuracy			0.56	2502
macro avg	0.51	0.51	0.51	2502
weighted avg	0.57	0.56	0.56	2502

Observações:
- O modelo também prevê sets de 3