INSTITUTO UNIVERSITÁRIO DE LISBOA

PROJETO APLICADO EM CIÊNCIA DE DADOS I

APRESENTAÇÃO FINAL

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

2 de junho de 2023

CONTRIBUIÇÃO PRÁTICA DO PROBLEMA DA PREVISÃO DE SETS

Serviços de consultoria em análise desportiva

Possíveis utilizações:

- Análise do desempenho de jogadores.
- Recomendações estratégicas.
- Treino personalizado.

- Melhoria da performance do jogo.
- Promoção de melhores resultados

LIMPEZA DOS DADOS PARA O BRASIL

Conjunto de dados inicial para o Brasil

37367 instâncias

24 variáveis

Conjunto de dados final para o Brasil

19208 instâncias

42 variáveis

VALORES OMISSOS PARA A BASE DE DADOS INICIAL – ATP BRASIL

RankPlayer	BornCity	BornCountry	Height
2620	11305	6615	11432

RankOpponent	Prize	Score
3511	367	2

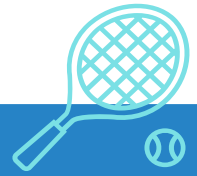
VALORES OMISSOS PARA A BASE DE DADOS FINAL – ATP BRASIL

Variável	Número de omissos
Born	4666
BornCity	4647
BornCountry	40
Height	4358
Hand	2975
L_OR_R	9
RankOpponent	622
BornOpponent	9357
BornCityOpponent	9344
BornCountryOpponent	3208

Variável	Número de omissos
BirthdayOpponent	723
AgeOpponent	723
HeightOpponent	8984
HandOpponent	6788
L_OR_R_Opponent	3026
Score	2
DifRank	622
DifAge	723
DifHeight	10303

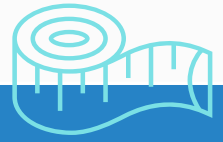
Total = 19 variáveis com valores omissos

VARIÁVEIS INTERESSANTES



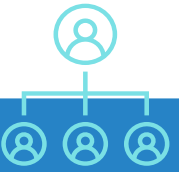
DifNumGamesByYear

- Diferença de número de jogos por ano entre o jogador principal e o oponente.
- O número de jogos por ano é determinante para o cansaço dos jogadores e, consequentemente, afetar a duração dos jogos.



DifHeights

- Diferença de altura entre o jogador principal e o oponente.
- Jogadores mais altos têm vantagem no serviço, porque têm maior capacidade de impulsão, mas podem vir a ser mais lentos.
- Jogadores mais baixos são geralmente mais ágeis e rápidos nos seus movimentos, mas têm um poder inferior de serviço.



DifRank

- Quanto maior for a diferença de ranks entre os jogadores, mais desequilibrado será o jogo.
- Jogadores no topo são mais conceituados, têm mais experiência, mais *skills* e mais técnica que o seu oponente, ou seja, à partida, haverá menor número de sets e o jogo tem menor duração.

MELHOR MODELO ENCONTRADO

Variáveis escolhidas:

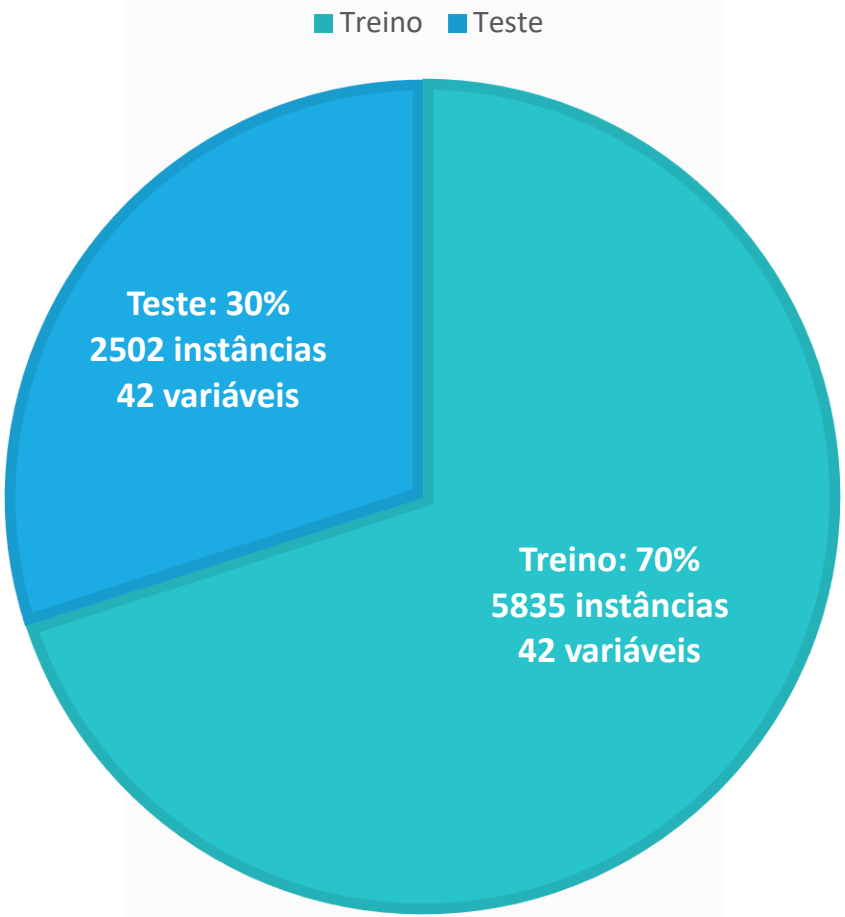
- Ground_Hard
- Ground_Clay
- DifHands
- Prize
- DifHeight
- DifAge
- DifRank
- DifNumberWins

Modelo	Sets	Accuracy	Precision	Recall	F1-score	Support	AUC
XGBoost	2	0.62	0.67	0.85	0.75	1670	0.52
	3		0.35	0.17	0.22	832	

Observações:

- Utiliza-se o algoritmo XGBoost;
- Sensibilidade do modelo aos sets de 3 (não prevê a moda);
- *Accuracy* alta (em consideração com os requisitos da sensibilidade);
- AUC perto de 0.5, que indica aleatoriedade alta.

DIVISÃO DO DATASET PARA O BRASIL



CONCLUSÕES FINAIS

- A aplicação da metodologia CRISP-DM permitiu desenvolver um modelo analítico para a análise de dados no contexto do ténis;
- Ao compreender as variáveis e sua relação com os resultados do jogo, podemos fornecer insights valiosos para os treinadores e jogadores;
- O objetivo de negócio de maximizar o desempenho dos jogadores e criar uma proposta satisfatória para a Confederação Brasileira de Ténis foi alcançado;
- Em termos da utilidade do modelo, pode-se dizer que a meta de previsão dos sets não foi cumprida da forma desejada;
- As conclusões tiradas a partir da previsão podem não vir a ser as mais alinhadas com a realidade mas podem providenciar linhas de ação;
- Sugestões de melhoria: experimentar variáveis novas, como por exemplo a força média dos serviços. Estudar outros modelos possíveis como SVM, algoritmos de Deep Learning como Multilayer Perceptron (MLP), entre outros. Como trabalho futuro, também poderia ser explorada a metodologia SCRUM.