Modelling
Ténis na Bélgica

Grupo 4:

ALLAN KARDEC, Nº 103380, TURMA CDB1
DIOGO FREITAS, Nº104841, TURMA CDB1
JOÃO BOTAS, Nº104782, TURMA CDB1
RICARDO ÂNGELO, Nº 104826, TURMA CDB1

Data Preparation to Modelling

4ª fase do CRISP-DM

Features selecionadas

Gráfico de correlação e análise preliminar

1° modelo e comentários

Conclusão e Planeamento

Features selecionadas

Na base de dados utilizada temos:

- 4188 linhas

- 18 colunas

Nota: Rank do jogador e oponente não adicionado por causa de problemas com omissos (temos 15 features, por enquanto + target)

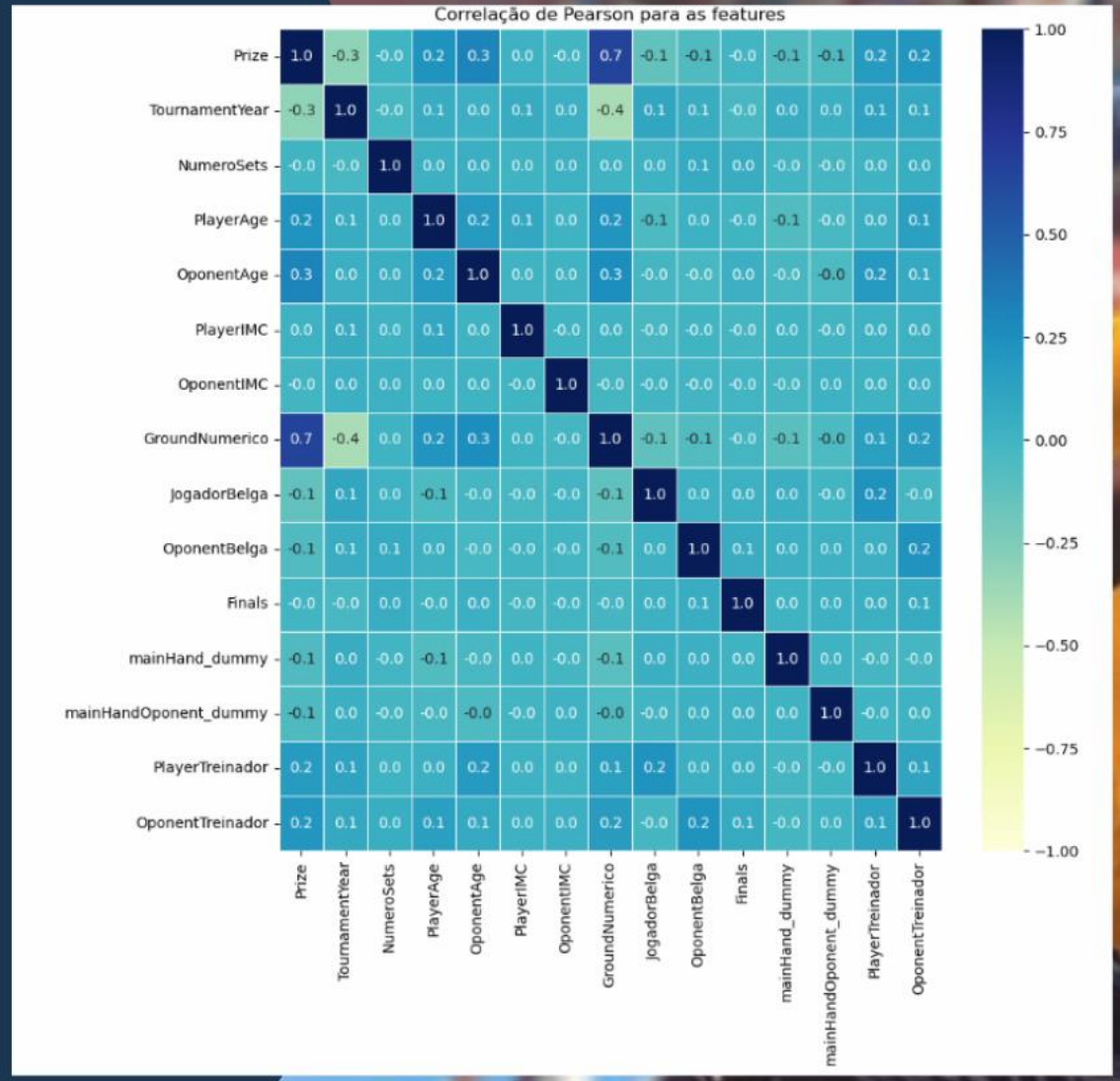
Feature	Explicação do porquê da sua criação/utilização
IMC Jogador	Procura-se classificar a aptidão física do jogador
IMC Oponente	Procura-se classificar a aptidão física do oponente
Belga Jogador	Fator casa deve influenciar a performance do jogador, aumentando a sua performance e, consequentemente, contribuir para jogos mais curtos
Belga Oponente	Fator casa deve influenciar a performance do oponente, aumentando a sua performance e, consequentemente, contribuir para jogos mais curtos
MainHand Jogador	Para diferentes jogadores, associados a diferentes mãos dominantes estão diferentes probabilidades em relação ao número de sets
MainHand Oponente	Para diferentes oponentes, associados a diferentes mãos dominantes estão diferentes probabilidades em relação ao número de sets
Idade Jogador	Para diferentes jogadores, a diferentes idades estão associadas diferentes probabilidades em relação ao número de sets (idade no jogo)
Idade Oponente	Para diferentes oponentes, a diferentes idades estão associadas diferentes probabilidades em relação ao número de sets (idade no jogo)
Treinador Jogador	Jogadores com treinador conhecido, em princípio, têm maior chance de participar em jogos com menor número de sets
Treinador Oponente	Oponentes com treinador conhecido, em princípio, têm maior chance de participar em jogos com menor número de sets
Rank Jogador	Quanto maior a diferença entre os ranks dos participantes (jogador e oponente), maior a chance de a partida ser mais curto
Rank Oponente	Quanto maior a diferença entre os ranks dos participantes (jogador e oponente), maior a chance de o jogo ser mais curto
Torneio Nome	Torneios diferentes têm diferentes probabilidades de total de sets, por partida
Ano Torneio	Torneios em datas diferentes podem estar associados a diferentes números de sets, por partida
Piso Numérico	A diferentes pisos estão associadas diferentes probabilidades em relação ao número de sets
Prémio	A jogos com prémios maiores estão associados jogadores com melhores rankings, o que deve de proporcionar um jogo intenso e, grande parte das vezes, jogos com maior número de sets
Nome com Final	Jogos que são finais (quartos-de-final, meias-finais e finais, por exemplo) têm maior chance de serem jogos com maior número de sets
Número de Sets	**Variável target**

Gráfico de correlação e análise preliminar

- Não parece haver multicolinearidade

- Baixa correlação com o alvo

- O Rank do Oponente e do Player não se encontram nesta matriz de correlação, mas serão adicionados futuramente.



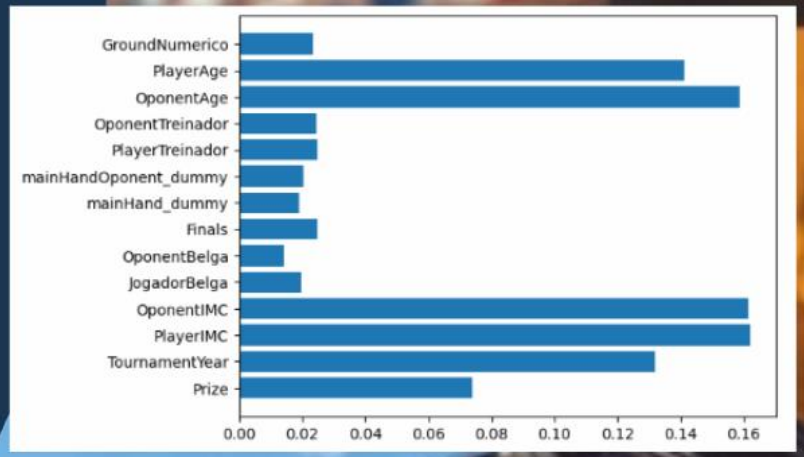
1° modelo

- técnica de partição: 70 treino/ 30 teste (k-fold partition no futuro...)

- Modelo de classificação utilizado: RandomForestClassifier do scikit-learn

- não muito bons, parece ter sido um modelo aleatório, apesar de faltar o rank dos jogadores

- Métricas de performance do modelo
  - accuraccy: 0.65
  - precision: 0.85
  - recall: 0.70
  - auc: 0.55 (Quase aleatório)



Conclusão:

- Features a utilizar
- Correlações
- Análise 1º modelo

Para a próxima semana...

- Ver novos tipos de modelos
- Rank dos jogadores para adicionar
- Ver outras técnicas de partição