2º Ano da Licenciatura de Ciência de Dados, 2023-2024

Unidade Curricular de Projeto Aplicado em Ciência de Dados I

Docente: Sérgio Moro

INSTITUTO UNIVERSITÁRIO DE LISBOA

Afonso Lourenço | Nº111487 | CDB2

Afonso Pereira | Nº111134 | CD-PL-B1

Afonso Santos | Nº111431 | CDB2

Diogo Santos | 111386 | CDB2

Guilherme Simões | Nº111604 | CDB2

Índice

Business Understanding

O objetivo deste projeto é prever o número de sets para conclusão de um jogo de ténis profissional (ranking ATP) nos torneios realizados na Índia, com o intuito de auxiliar os treinadores, jogadores e equipas técnicas a otimizarem os seus treinos e planos de jogo para melhorarem as suas performances durante os jogos. Esta previsão também pode ser importante para os apostadores na medida em que podem ajudá-los a ter mais apostas bem sucedidas e por sua vez maior lucro. Também poderá ser importante para as empresas televisivas, que podem estimar através do número de sets o tempo necessário que precisam de reservar para um determinado jogo.

Para esta análise, vai-se considerar a história dos jogos ocorridos na Índia, tendo em conta vários fatores, como a mão dominante, o seu país de nascimento e o tipo de superfície de jogo, bem como ainda o clima através das estações do ano na data em que o torneio ocorre, todas variáveis que podem influenciar os resultados. Também se vai analisar detalhadamente o perfil dos jogadores, os seus rankings, experiência prévia e a sua idade, comparando-os com os seus oponentes para perceber se isso afeta o desempenho e o número de sets.

No ténis profissional existem jogos entre singulares (1 vs 1) e em pares (2 vs 2). Neste estudo apenas se vai tratar dos jogos entre singulares (1 vs 1).
O primeiro jogador a ganhar quatro jogos num set ganha o set e o primeiro a ganhar três ou cinco sets (dependendo do tipo de torneio: melhor de 3 ou melhor de 5) ganha o jogo. Se o resultado de um jogo estiver empatado 3-3 ou 5-5, é jogado um tie-break. O tie-break é um jogo de sete pontos, em que o primeiro jogador a chegar aos sete pontos, com uma vantagem mínima de dois pontos, ganha o tie-break, o set. (Federação Internacional de Ténis, 2023).

A Associação de Tenistas Profissionais (ATP) é o órgão regulador dos circuitos de ténis profissionais masculinos. Foi constituído em setembro de 1972 por Donald Dell, Cliff Drysdale e Jack Kramer para proteger os interesses dos tenistas profissionais. Drysdale foi o primeiro presidente. A partir de 1990, a associação organizou o circuito ATP: ciclo mundial do tênis masculino, vinculando-o ao nome da organização.

É importante perceber bem a pontuação atribuída a cada competição. Os torneios estão classificados em ATP 250 (que concede 250 pontos ao vencedor), ATP 500 (500 pontos) e ATP Masters 1000 (1000 pontos). Ou seja, o sistema de pontuação e o subsequente ranking dos atletas fica claro na nomenclatura da competição. Além disso, os quatro Grand Slams concedem 2000 pontos para o vencedor. Sendo assim, o jogador de ténis recebe uma pontuação conforme o seu desempenho nos campeonatos citados anteriormente.

Relativamente aos torneios que ocorrem na Índia, o torneio mais conhecido é o Open da Índia (Chennai Open). O Chennai Open foi um torneio ATP 250 realizado anualmente em Chennai, Índia. Era disputado em quadras duras e geralmente acontecia em janeiro. Foi fundado em 1996 e acabou em 2018. Este torneio era visto com os jogadores como uma forma de preparação para o Open da Austrália, que é considerado um Grand Slam e ocorre pouco tempo depois.

Outro torneio bastante conhecido a nível mundial é o Open de Maharashtra, que também é um torneio ATP 250 e é realizado em Pune. Também é disputado em quadras duras. Começou a ser realizado em 1996, inicialmente em Nova Deli, mas foi transferido para Pune em 2018. O torneio é importante na medida em que atrai uma mistura de jogadores experientes e jovens talentos, contribuindo para a promoção do ténis na Índia.

Data Understanding

Data Collection

O dataset principal que foi utilizado foi fornecido pelos docentes e foi retirado do site ATP[1]. A base de dados inicial fornece informações sobre jogos de ténis, torneios e jogadores de inúmeros países, excluindo jogos de pares (2 vs 2) e de singulares (1 vs 1) entre atletas femininas. Sobre os jogos contém dados sobre a ronda do jogo no torneio em si e o resultado do jogo. Sobre os torneios tem o nome do torneio, a localização, a data, o piso e o prémio do torneio. Já sobre os jogadores é fornecido o nome, o local de nascimento, a altura, a mão dominante, o ranking ATP, o oponente daquele determinado jogo e se ganhou o jogo ou não.

O primeiro critério a ser atendido é o filtro geográfico para o país que se pretendia estudar, a Índia. Feito o filtro para o país em questão, o dataset inicial contém 17917 observações e cerca de 14 variáveis, com registros de jogos desde maio de 1973 até março de 2022 realizados na Índia.

Para um melhor entendimento dos dados e futuramente uma melhor análise, retirou-se uma base de dados em formato excel que contém informações mais detalhadas dos jogadores que estão no dataset inicial. Este csv contém informações como o nome, a mão dominante, o seu país de origem, o ranking ATP do jogador e a sua data de nascimento. Com estas novas informações é possível associá-las a novas variáveis no dataset original. Esta última variável (data de nascimento) é importante para ser calculada a idade que vai ser relevante para estudar a variável alvo. Retirou-se ainda informações sobre as estações do ano na Índia de forma a perceber se o clima influencia o que se pretende concluir (número de sets).

A variável alvo, o número de sets, foi deduzida através do placar da partida. É calculada pela contagem do número de dígitos numéricos da variável score formatada. Ainda através do score é possível recolher a informação se o jogo é à melhor de 3 ou à melhor de 5 e se há a uma desistência quer seja por "walk off" que se refere a desistir antes do início da partida, quer seja por "retired", que descreve a desistência durante o jogo devido a lesão ou outro motivo.

[1] https://www.atptour.com/en

Data description

Com a inserção das novas informações retiradas de fora, passou-se de 14 variáveis para 35 variáveis.

- PlayerName, OpponentName
  O Nome do jogador em causa e do seu oponente
  Tipo: Categórico Nominal

- PlayerBorn, OpponentBorn
  O local de nascimento(cidade e país) do jogador e do seu oponente
  Tipo: Categórico Nominal

- PlayerHeight, OpponentHeight
  A altura do jogador e do seu oponente em cm
  Tipo: Numérico contínuo

- PlayerHand, OpponentHand
  A mão usada pelo jogador e pelo seu oponente
  Tipo: Categórico Nominal
  Classes: 'Right-Handed, Two-Handed Backhand', 'Left-Handed, Two-Handed Backhand', 'Right-Handed, One-Handed Backhand', 'Left-Handed, One-Handed Backhand', 'Right-Handed, Unknown Backhand', 'Left-Handed, Unknown Backhand', 'Ambidextrous, Two-Handed Backhand'

- PlayerRank, OpponentRank
  O ranking ATP do jogador e do seu oponente
  Tipo: Numérico discreto

- PlayerBirth, OpponentBirth
  Data de nascimento do jogador e do seu oponente
  Tipo: Categórico Nominal (Date)

- PlayerAge, OpponentAge
  A idade do jogador e do seu oponente
  Tipo: Numérico contínuo

- PlayerMainHand, OpponentMainHand
  A mão dominante do jogador e do seu oponente
  Tipo: Categórico Nominal
  Classes: 'Right-Handed', 'Left-Handed', 'Ambidextrous'

- PlayerBackHand, OpponentBackHand
  A mão menos dominante do jogador e do seu oponente
  Tipo: Categórico Nominal
  Classes:'Two-Handed Backhand', 'One-Handed Backhand', 'Unknown Backhand'

- Tournament
  O nome do torneio em que o jogo se vai realizar
  Tipo: Categórico Nominal
  Classes: ‘Chennai Open’, ‘Maharashtra Open’, etc…

- Bo3_Bo5
  Se o jogo é à melhor de 3 ou de 5
  Tipo: Categórico Ordinal
  Classes: ‘3’ (Melhor de 3), ‘5’ (Melhor de 5)

- Location
  A localização do torneio (cidade e país)
  Tipo: Categórico Nominal
  Classes: ‘Chennai’, ‘Pune’, etc…

- Date
  A data de início e fim do torneio em que o jogo se vai realizar
  Tipo: Categórico Nominal (Date)

- Start_Date
  A data de início do torneio em que o jogo se vai realizar
  Tipo: Categórico Nominal (Date)

- End_Date
  A data de fim do torneio em que o jogo se vai realizar
  Tipo: Categórico Nominal (Date)

- Season
  A estação do ano na data de início do torneio em que o jogo se vai realizar
  Tipo: Categórico Nominal
  Classes: 'Shishira (Inverno)', 'Sharad (Outono)', 'Hemanta (Pré-Inverno)', 'Vasanta (Primavera)', 'Varsha (Monções)', 'Grishma (Verão)'

- Ground
  O tipo de piso do jogo
  Tipo: Categórico Nominal
  Classes: 'Hard', 'Grass', 'Clay', 'Carpet'

- Prize
  O prémio que o vencedor do torneio ganha
  Tipo: Numérico Contínuo

- GameRound
  A ronda do torneio do jogo que se está a disputar
  Tipo: Categórico Ordinal
  Classes: 'Finals', 'Semi-Finals', 'Quarter-Finals', 'Round of 16', 'Round of 32', 'Round Robin', '2nd Round Qualifying', '1st Round Qualifying', '3rd Round Qualifying', 'Round of 64'

- WL
  Se o jogador ganha ou perde o jogo
  Tipo: Categórico Nominal
  Classes: 'W' (Venceu o jogo), 'L' (Perdeu o jogo)

- Score
  O resultado do jogo
  Tipo: Numérico Discreto
  Classes: ‘6-1 6-2’, ‘6-3 4-6 6-3’, etc…

- Withdrawal
  Se o jogador desistiu do jogo e se desistiu de que forma foi
  Tipo: Categórico Nominal
  Classes: 'NA’ (Não desistiu do jogo), ‘W/O’ (desiste antes de começar o jogo), ‘RET’ (jogador desiste durante o jogo)

- Sets
  Número de sets que o jogo durou
  Tipo: Categórico Ordinal
  Classes: ‘2’, ‘3’, ‘4’, ‘5’

- RankDif
  Diferença do ranking ATP entre o jogador e o oponente
  Tipo: Numérico Discreto

- AgeDif
  Diferença de idades entre o jogador e o oponente
  Tipo: Numérico Contínuo

- HeightDif
  Diferença de alturas entre o jogador e o oponente
  Tipo: Numérico Contínuo

Data Preparation

Numa fase inicial da etapa de preparação de dados, avaliou-se os valores em falta de cada variável. Para isso realizou-se um mapa de calor. Neste tipo de mapas, as cores indicam a magnitude do valor associado a cada ponto de dados. Cores mais claras indicam a ausência de dados ou valores mais baixos, enquanto cores mais escuras podem representar valores mais altos ou a presença de dados. Por exemplo, se uma coluna específica tem muitas células claras, isso indica que muitos casos têm informações específicas em falta. Se uma linha tem muitas células escuras, quer dizer que a variável em questão tem muitos dados disponíveis.

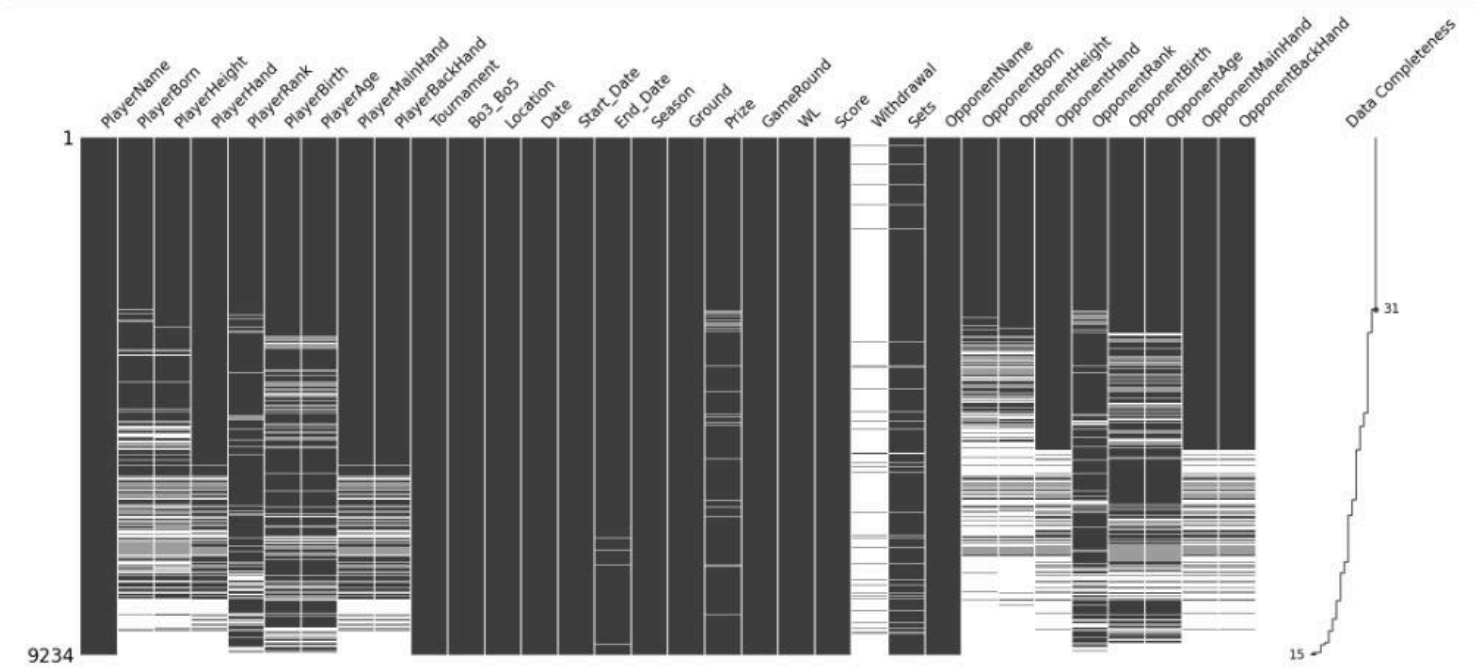

Fig. 1 – Valores Nulos das Variáveis do Dataset

Através da análise do gráfico, consegue-se observar que quase todas as variáveis relacionadas com os jogadores têm dados em falta e apenas as variáveis *PlayerName e OpponentName* não têm valores em faltas, o que é esperado, já que se trata de uma informação principal.

Em relação às variáveis relacionadas com os jogos e torneios, observa-se, maioritariamente, a presença de poucos valores nulos. Apenas nas variáveis *End_Date, Prize* e *Sets* é que se observa a existência de alguns valores nulos mas não significantes para o objetivo pretendido.

A variável Withdrawal tem inúmeros dados em falta e até muito poucos dados disponíveis, mas isso deve-se ao facto desta variável indicar se o jogador desistiu do jogo e, se desistiu, de que forma foi. Quando o jogador não desiste, esta variável assume o valor de NA e daí a justificação para quase todos os valores desta variável serem nulos.

Posteriormente, foram analisadas as variáveis da altura, do ranking e a da idade dos jogadores, bem como as suas diferenças. São as variáveis numéricas que, provavelmente, têm maior potencial para influenciar a variável alvo.
Inicialmente foram analisados os rankings dos jogadores (*PlayerRank*).

```
count    9233.000000
mean      662.389689
std       507.032436
min         2.000000
25%       279.000000
50%       556.000000
75%       927.000000
max      3000.000000
```

Fig 2 - Estatísticas de PlayerRank

Os valores em falta na variável rank foram colocados com o rank de 3000, visto que o rank máximo é de 2246. Decidiu-se fazer isso porque se não há registo do ranking do jogador, muito provavelmente este ou não é minimamente conhecido ou tem um ranking muito baixo, que acaba por nem ter registo no ATP.
A média de 662.39 com um desvio padrão de 507.03 indica que existe uma variedade considerável de rankings entre os jogadores. Um desvio padrão alto sugere que os valores estão mais espalhados em relação à média.
Os valores dos quartis fornecem informações sobre a distribuição dos dados. Por exemplo, o terceiro quartil (75%) de 927 sugere que a maioria dos jogadores tem um ranking relativamente alto, enquanto o primeiro quartil (25%) de 279 indica que ainda há uma parcela significativa mas substancialmente mais pequena de jogadores com rankings melhores.

O valor máximo real (2246) em relação à média sugere a presença de outliers. Para provar isto, realizou-se um boxplot, que facilita a perceção da existência de outliers nesta variável.

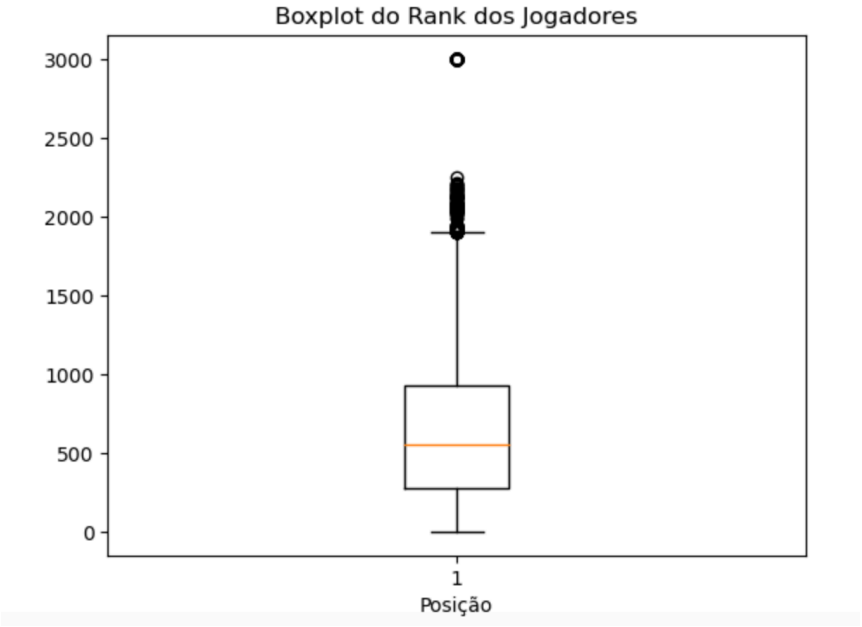

Fig.3 - Boxplot do Rank dos Jogadores

Pela observação do boxplot, consegue-se concluir que há um número significativo de outliers moderados acima do bigode superior, o que sugere a presença de muitos jogadores com o ranking extremamente alto. A existência do outlier extremo de 3000 não interessa para o estudo em si, pois esse valor substitui todos os valores nulos presentes nesta variável, como já foi falado anteriormente.

De seguida foram analisados os ranks dos oponentes (*OponnentRank*).

```
count    9233.000000
mean      881.972490
std       748.723383
min         2.000000
25%       334.000000
50%       699.000000
75%      1181.000000
max      3000.000000
```

Fig 4. - Estatísticas de OponnentRank

A média dos rankings dos oponentes é de 881.97, sugerindo que, em média, os jogadores enfrentam oponentes classificados em torno de 882 no ranking mundial. Isso indica que os oponentes geralmente não estão entre os melhores do mundo, uma vez que rankings mais baixos são considerados melhores (por exemplo, um ranking de 1 é melhor do que um ranking de 1000).

O desvio padrão de 748.72 indica uma grande variação nos rankings dos oponentes, mostrando que os jogadores enfrentam tanto adversários de alto ranking (mais próximos do ranking 1) quanto de baixo ranking. O melhor ranking enfrentado é 2, representando um oponente de altíssimo nível, no topo do ranking mundial. Por outro lado, o pior ranking enfrentado é 3000, representando os valores nulos desta variável.

O primeiro quartil (25%) é 334, o que significa que 25% dos oponentes têm um ranking melhor que 334. Estes são considerados bons jogadores, dentro dos 334 melhores do mundo. A mediana é 699, indicando que metade dos oponentes está classificada entre os 699 melhores jogadores, o que sugere que os jogadores frequentemente enfrentam adversários na metade superior do ranking global. O terceiro quartil (75%) é 1181, mostrando que a maioria dos oponentes está razoavelmente bem classificada, com apenas 25% tendo um ranking pior que 1181.

A diferença entre a média e a mediana sugere que a distribuição dos rankings é assimétrica, com uma cauda longa à direita. Isso significa que há um número significativo de oponentes com rankings relativamente piores (mais altos), o que puxa a média para cima. O alto desvio padrão reflete a grande variabilidade na qualidade dos oponentes, indicando que os jogadores enfrentam uma mistura de adversários, desde altamente qualificados até menos qualificados. A amplitude interquartil de 847 confirma a diversidade nos níveis de competição, com uma grande faixa de rankings entre os oponentes mais frequentes.

Para identificar possíveis outliers nesta variável procedeu-se à realização de um box plot.

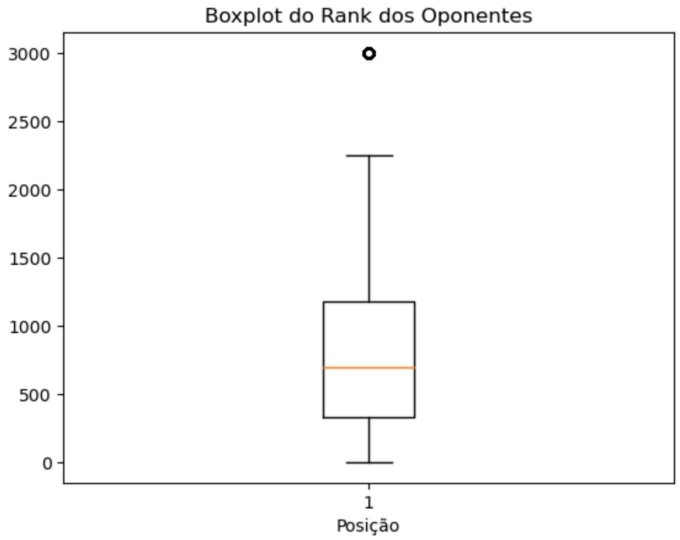

Fig. 5 - Boxplot do Rank dos Oponentes

Analisando o boxplot intitulado de “Boxplot do Rank dos Oponentes”, pode-se observar que a maioria dos ranks dos oponentes está concentrada num intervalo mais estreito, representado pela caixa, que indica o intervalo interquartil (IQR). Relativamente à possível existência de outliers, excluindo o valor 3000 que não interessa para o estudo em si devido a ser a substituição dos valores nulos da variável, não parece haver outliers através da observação do box plot.

De seguida, procedeu-se à análise da diferença de rankings (*RankDif*).

```
count    9233.000000
mean      477.312683
std       579.541944
min         0.000000
25%        92.000000
50%       267.000000
75%       629.000000
max      2971.000000
```

Fig. 6 - Estatísticas de RankDif

A média da diferença de ranking é 477.31, indicando que, em média, há uma diferença significativa nos rankings dos jogadores e respetivos oponentes. Isto sugere que muitos jogos não ocorrem entre jogadores com ranking próximo.

O desvio padrão é alto, 579.54, mostrando uma grande variação nas diferenças de ranking. Isso significa que algumas partidas são entre jogadores de rankings muito próximos, enquanto outras envolvem jogadores com rankings bastante distantes.

O valor mínimo é 0, indicando que em pelo menos uma partida, os jogadores tinham o mesmo ranking, representando uma disputa entre adversários teoricamente de igual habilidade. O primeiro quartil é 92, o que significa que 25% das diferenças de ranking são menores que 92, indicando que um quarto das partidas ocorre entre jogadores cujas habilidades, conforme o ranking, são relativamente próximas.

A mediana é 267, mostrando que metade das partidas tem uma diferença de ranking menor que 267. Isso ainda indica uma diferença considerável, mas menor do que a média. O terceiro quartil é 629, indicando que 75% das partidas têm uma diferença de ranking menor que 629. A maioria dos jogos ocorre entre jogadores com diferenças de ranking significativas, mas não extremas.

A diferença entre a média (477,31) e a mediana (267) sugere uma distribuição assimétrica, possivelmente com uma cauda longa à direita. Isso indica que existem algumas partidas com diferenças de ranking extremamente grandes que puxam a média para cima. O alto desvio padrão reflete a grande variabilidade nas diferenças de ranking, indicando que a competitividade dos jogos varia amplamente.

A amplitude interquartil (IQR) é de 537 (629 - 92), mostrando uma grande faixa de diferenças de ranking na maioria dos jogos.

Para identificar possíveis outliers, procedeu-se à realização de um boxplot.

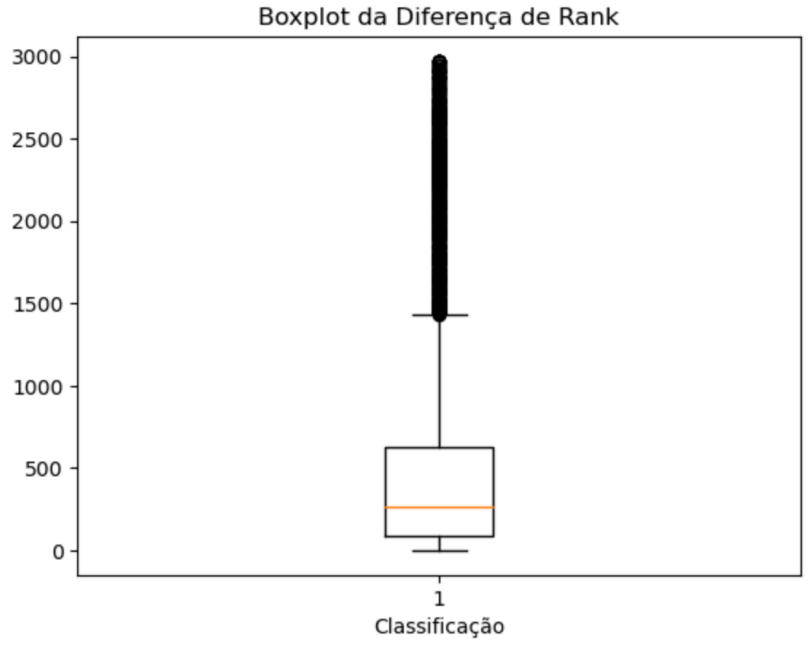

Fig. 7 - Boxplot da Diferença de Rank

Com base no boxplot, há indicação de outliers na distribuição das diferenças de rank entre jogadores e oponentes em partidas de ténis. Através da observação do boxplot, há uma quantidade significativa de outliers acima do bigode superior, indicando algumas diferenças de rankings de jogadores e oponentes significativamente superiores à maioria. Pode-se ver que, sensivelmente, a partir da diferença de ranking entre jogadores e oponentes de 1500 já é considerado outlier. Com uma diferença de 1500 entre rankings, é bastante notória a melhor capacidade física e mental do jogador com melhor ranking.

Seguidamente, passou-se para o estudo das idades.

No estudo desta variável devido à existência de diversos valores nulos (1311) procedeu-se a alteração desses valores pela média de idades em prol de um resultado melhor. Tentou-se encontrar externamente a idade dos jogadores em falta mas sem sucesso.
De seguida foram analisadas as idades dos jogadores (*PlayerAge*).

count	9233.000000
mean	22.681772
std	3.315820
min	14.000000
25%	20.000000
50%	22.681772
75%	24.000000
max	52.000000

Fig. 8 - Estatísticas de PlayerAge

A média das idades dos jogadores é de 22.68, com um desvio padrão de 3.32 indicando que as idades dos mesmos variam nesse intervalo. Porém, o jogador mais velho e o mais jovem apresentam idades de 52 e 14 anos, respetivamente.

O primeiro quartil (25%) é 20, o que significa que 25% dos jogadores têm uma idade igual ou inferior a 20. A mediana é 22.68, indicando que metade dos jogadores apresenta uma idade igual ou inferior a 22. O terceiro quartil (75%) é 25, mostrando que três quartos dos jogadores apresentam uma idade igual ou inferior a 25. Ainda assim com a mediana, igual à média de 22.68, reforça a simetria da distribuição das idades, indicando que a maioria dos jogadores tem uma idade em torno dessa média.

Para identificar possíveis outliers nesta variável procedeu-se à realização de um box plot.

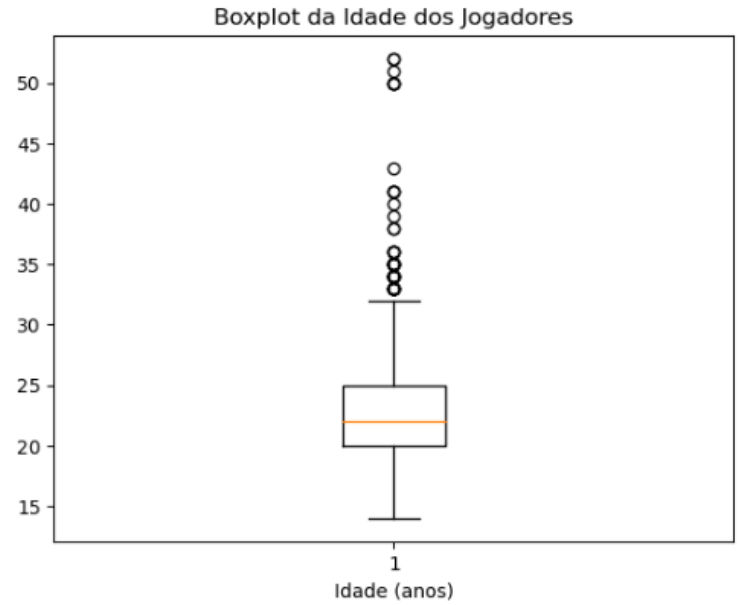

Fig. 9 – Boxplot das Idades dos Jogadores

Através da análise do boxplot, consegue-se observar que não há outliers abaixo dos bigodes, mas há 8 outliers acima do bigode superior, indicando alguns jogadores com idades significativamente superiores à maioria. Pode-se ver que, sensivelmente, a partir dos 33 anos de idade já é considerado outlier. Isso deve-se ao fato de que, nessas idades, muitos jogadores acabam por se retirar do ténis profissional ou diminuem o ritmo devido às exigências físicas.

De seguida foram analisadas as idades dos oponentes (*OpponentAge*).

```
count    9233.000000
mean       23.670415
std         3.383476
min        14.000000
25%        22.000000
50%        23.670415
75%        25.000000
max        54.000000
```

Fig. 10 – Estatísticas de Opponentage

A idade média dos oponentes é aproximadamente 23,67 anos, com um desvio padrão de, aproximadamente, 3,38 anos. O oponente mais novo apresenta uma idade de 14 anos e o mais velho de 54 anos.

O primeiro quartil (25%) é 22, o que significa que 25% dos jogadores têm uma idade igual ou inferior a 22. A mediana é 23.67, indicando que metade dos jogadores apresenta uma idade igual ou inferior a 23.67. O terceiro quartil (75%) é 25, mostrando que três quartos dos jogadores apresentam uma idade igual ou inferior a 25. Ainda assim, com a mediana igual à média de 23,67, reforça-se a simetria da distribuição das idades, indicando que a maioria dos jogadores tem uma idade em torno dessa média.

Para identificar possíveis outliers nesta variável procedeu-se à realização de um box plot.

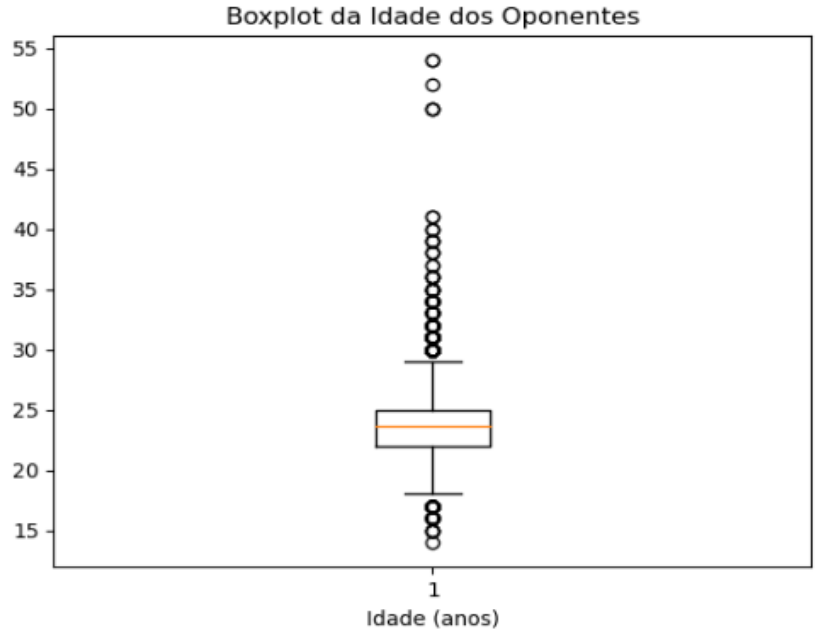

Fig. 11 – Boxplot das Idades dos Oponentes

Através da análise do boxplot, consegue-se observar que há outliers abaixo dos bigodes e há também acima do bigode superior, indicando alguns oponentes com idades significativamente superiores e inferiores à maioria. Pode-se ver que, sensivelmente, a partir dos 30 anos de idade já é considerado outliers. Os outliers abaixo dos bigodes são considerados outliers em idades de 18 para baixo, sensivelmente.

De seguida, foram analisadas as diferenças de idades dos jogadores com os oponentes (*AgeDif*).

count	9233.000000
mean	3.482385
std	2.926897
min	0.000000
25%	1.000000
50%	3.000000
75%	5.000000
max	36.000000

Fig. 12 – Estatísticas de AgeDif

A diferença média de idades entre os jogadores e os seus oponentes, é aproximadamente de 3.48 anos, com um desvio padrão de 2.93 anos. A menor diferença de idade é 0 anos e a maior é de 36 anos, o que nos indica que existem pelo menos um jogador e um oponente com a mesma idade. O primeiro quartil (25%) é 1, o que significa que 25% dos jogadores têm uma diferença de idades menor ou igual a 1 ano. A mediana é 3, indicando que metade dos jogadores apresenta uma diferença de idades igual ou inferior a 3 anos. O terceiro quartil (75%) é 5, mostrando que três quartos dos jogadores apresentam uma idade igual ou inferior a 5 anos.

Para identificar possíveis outliers nesta variável procedeu-se à realização de um box plot.

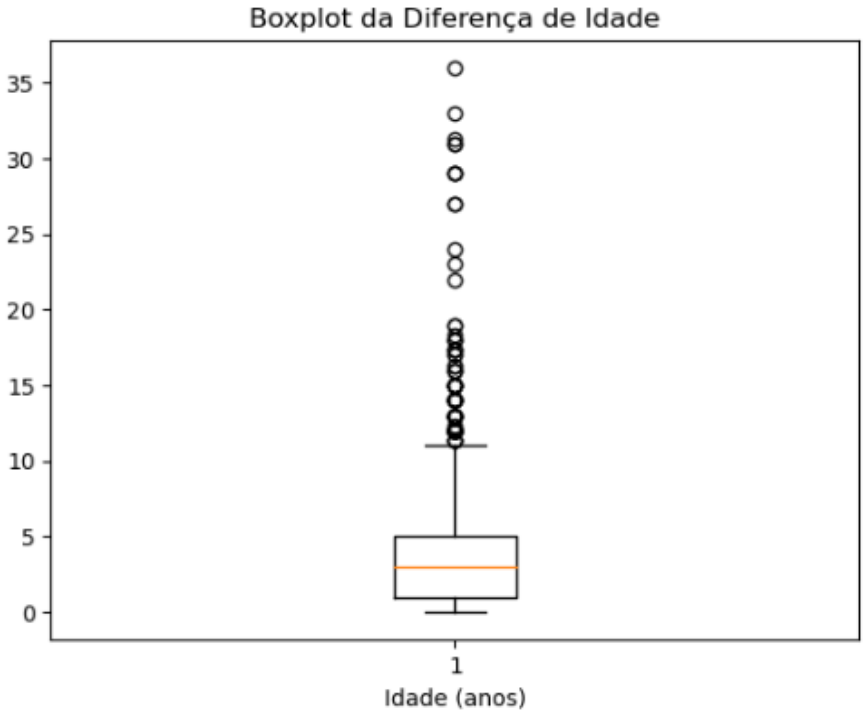

Fig. 13 – Boxplot das Diferenças de Idades

Através da análise do boxplot, consegue-se observar que não há outliers abaixo dos bigodes mas há uma quantidade significativa acima do bigode superior, indicando algumas diferenças de idades de jogadores e oponentes significativamente superiores à maioria. Pode-se ver que a partir da diferença de idade de 12 anos entre jogadores e oponentes já é considerado outliers. Uma diferença de idades de 12 anos pode ser significativa e influenciar o estilo de jogo e a estratégia.

Passando para a variável altura, tinha algumas anomalias nos seus dados e muitos dados em falta. Tinha jogadores e oponentes com altura igual a 0 e igual a 15, (variável em cm) que iam prejudicar a análise e futuramente os modelos. Tinha cerca de 2431 valores nulos, o que é um número bastante alto e não pode ser ignorado.
Primeiramente, substitui-se os valores dos jogadores e oponentes com altura igual a 0 e igual a 15 por NA, para ficarem dados sem anomalias e apenas com valores nulos. Depois para a análise desta variável não ficar comprometida, substitui-se os valores nulos pela média de altura, quer dos jogadores quer dos oponentes.
Após feito isto, procedeu-se à análise da variável da altura dos jogadores (*PlayerHeight*).

```
count    9233.000000
mean      182.750221
std         5.829769
min       165.000000
25%       180.000000
50%       182.750221
75%       185.000000
max       211.000000
```

Fig. 14 – Estatísticas de PlayerHeight

A média das alturas é de 182.75 cm, sugerindo que a altura ideal para jogadores de ténis tende a ser superior à média da população geral. Isso deve-se às vantagens que a altura oferece, como maior alcance para jogadas difíceis, fatores que podem tornar os jogadores mais competitivos.
O desvio padrão de 5.83 cm indica uma variação não muito alta nas alturas dos jogadores. Embora a altura ofereça vantagens, a habilidade técnica e a agilidade também são cruciais no ténis, permitindo que os jogadores que têm diferentes alturas possam competir em alto nível, daí o desvio padrão não ser muito alto.

O primeiro quartil, com 25% dos jogadores tendo altura de 180 cm ou menos, sugere que uma altura ligeiramente acima da média pode ser benéfica, mas não é um requisito absoluto para o sucesso. A mediana, igual à média de 182.75 cm, reforça a simetria da distribuição das alturas, indicando que a maioria dos jogadores tem altura em torno dessa média. O terceiro quartil, com 75% dos jogadores tendo altura de até 185 cm, sugere que ser mais alto que a média pode oferecer vantagens competitivas, mas não é determinante, já que 25% dos jogadores têm uma altura superior a essa.

A altura máxima registrada é de 211 cm, significativamente maior que a média. Jogadores extremamente altos podem tirar vantagem do seu alcance e poder de serviço, mas também podem enfrentar desafios de mobilidade e resistência. A altura mínima registrada é de 165 cm, bem abaixo da média, o que mostra que jogadores mais baixos podem compensar a falta de altura com outras habilidades, como agilidade e velocidade.

Para identificar possíveis outliers, procedeu-se à realização de um box plot.

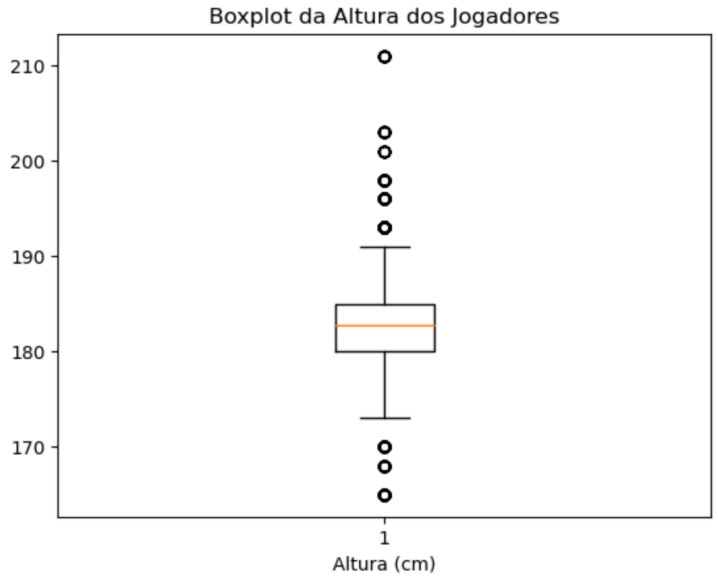

Fig. 15 – Boxplot das Alturas dos Jogadores

Através da observação do boxplot, consegue-se concluir a existência de bastantes outliers. Estes são jogadores cuja altura está significativamente acima ou abaixo da média. No caso deste box plot, temos vários outliers na parte superior, indicandojogadores muito mais altos do que a maioria e alguns na parte inferior, indicando jogadores muito mais baixos do que a maioria.

Após esta análise, procedeu-se à análise da altura dos oponentes (OponnentHeight).

```
 count    9233.000000
mean      182.603666
std         4.974902
min       165.000000
25%       182.603666
50%       182.603666
75%       183.000000
max       211.000000
```

Fig. 16 – Estatísticas de OponnentHeight

A média de altura dos oponentes é de 182.60 centímetros. Isso sugere que, em geral, os jogadores enfrentam adversários com alturas próximas a esse valor. A média e a mediana sendo iguais (182.60 cm) sugerem que a distribuição das alturas é simétrica. O desvio padrão de 4.97 cm é pequeno em comparação com a média, indicando que a maioria dos jogadores tem alturas próximas à média de 182.60 cm. Isso sugere que não há muita variabilidade nas alturas dos oponentes, o que pode indicar uma certa homogeneidade no grupo de jogadores.

Os quartis mostram pouca variação entre o 1º quartil (182.60 cm) e o 3º quartil (183 cm). Isto indica que a maioria dos jogadores está concentrada em um intervalo estreito de alturas, com 50% dos dados situados entre 182.60 cm e 183 cm. Esta concentração num intervalo pequeno reforça a ideia de pouca variabilidade nas alturas.

A diferença entre a altura mínima (165 cm) e a máxima (211 cm) é significativa, sugerindo a presença de alguns jogadores que são consideravelmente mais baixos ou mais altos que a média. Esses valores extremos, ou outliers, podem influenciar as dinâmicas de jogo, mas não parecem afetar a simetria geral da distribuição.

Para avaliar a existência de outliers, realizou-se um boxplot da altura dos oponentes.

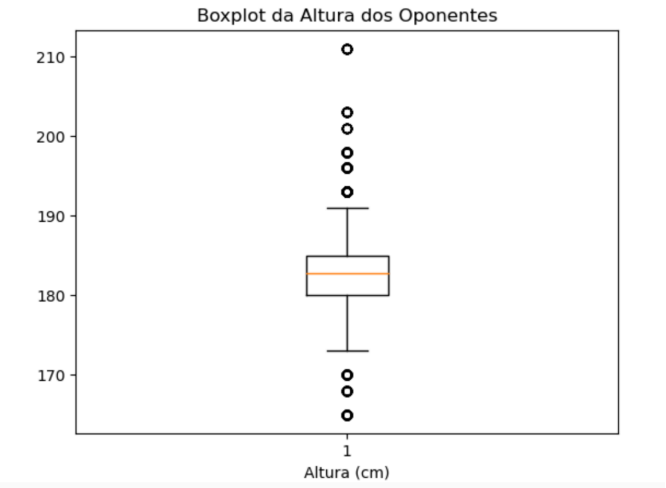

Fig. 17 – Boxplot das Alturas dos Oponentes

Através da análise do boxplot, consegue-se observar pontos acima do bigode superior e abaixo do bigode inferior que representam os oponentes com alturas extremamente altas e extremamente baixas para a realidade dos nossos dados.

Os jogadores mais altos têm vantagens como alcance e potência, mas podem ter dificuldade na movimentação dentro do jogo. Por outro lado, os jogadores mais baixos podem destacar-se em agilidade e velocidade, mas podem enfrentar dificuldades no alcance e na potência. Portanto, estratégias específicas devem ser desenvolvidas para enfrentar oponentes com alturas extremas.

Após estas conclusões, procedeu-se à análise da diferença de alturas (HeightDif).

```
count    9233.000000
mean        5.371574
std         5.275982
min         0.000000
25%         0.249779
50%         4.603666
75%         8.000000
max        36.000000
```

Fig. 18 – Estatísticas de HeightDif

A distribuição das diferenças de altura apresenta algumas características interessantes. A média (5.37 cm) e a mediana (4.60 cm) são diferentes, sugerindo uma distribuição ligeiramente assimétrica, possivelmente com uma cauda longa à direita, devido à presença de valores extremos (diferença máxima de 36.00 cm).

O desvio padrão de 5.28 cm, bastante grande em relação à média, indica uma considerável variabilidade nas diferenças de altura entre jogadores e oponentes.

Isto significa que, enquanto algumas partidas têm jogadores de alturas muito próximas, há um grande número de partidas onde os jogadores e oponentes apresentam diferenças significativas.

O primeiro quartil (0.25 cm) indica que 25% das diferenças de altura são muito pequenas, quase insignificantes. O terceiro quartil (8.00 cm) mostra que 75% das diferenças de altura são menores que 8.00 cm, indicando que grandes diferenças de altura são menos comuns, mas ainda significativas.

A diferença máxima de 36.00 cm é bastante significativa, sugerindo partidas onde um jogador é consideravelmente mais alto que o oponente. Estes valores extremos podem ter um impacto importante na dinâmica do jogo. A diferença mínima de 0 cm e um primeiro quartil de 0.25 cm indicam que uma proporção significativa das partidas envolve jogadores de alturas muito similares, sugerindo que algumas partidas são bem balanceadas em termos de altura dos jogadores.

Acerca desta variável, ainda tem de se avaliar a existência de outliers.

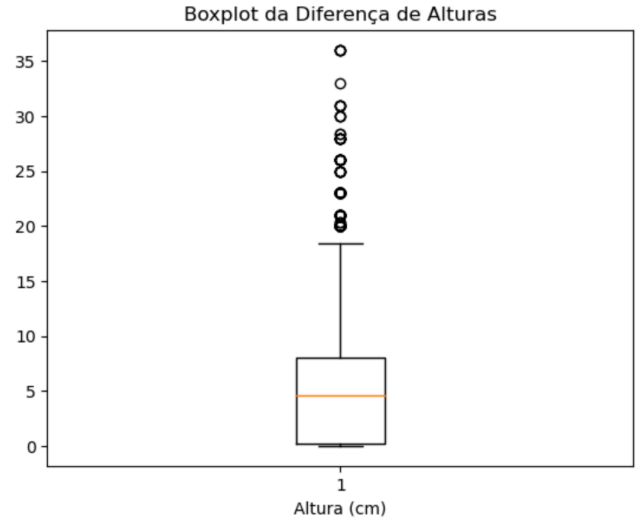

Fig. 19 – Boxplot da Diferença de Alturas

Através da observação do boxplot, consegue-se observar pontos acima do bigode superior que representam outliers em que a diferença de alturas entre jogadores e oponentes é extremamente grande. Pode-se ver que, sensivelmente, a partir da diferença de 20 cm de jogadores e oponentes já é considerado outliers. Uma diferença de altura de 20 cm pode ser significativa e influenciar o estilo de jogo e a estratégia.

Após as análises a estas variáveis, tratou-se melhor a variável inicial “*PlayerHand*” e “*OpponentHand*”. Dividiu-se os valores de ambas as colunas em duas partes e atribuiu-se a duas novas colunas. Sendo assim, os jogadores e os oponentes ficaram com a informação das suas mãos melhor organizada, dividida em *mainhand* (mão dominante) e *backhand* (mão não dominante). Além disso, também removeu-se os espaços em branco ao redor das novas colunas.

Criou-se uma codificação para facilitar o estudo das variáveis, em que 'R' era igual a 'Right-Handed', 'L' a 'Left-Handed', 'U' aos valores nulos e 'A' a 'Ambidextrous'.

Por fim, os NAs destas variáveis foram substituídos por "Unknown MainHand" e "Unknown BackHand", de forma a não ter valores nulos, com o intuito de facilitar o estudo das correlações e a construção do modelo.Após a limpeza de dados foram criadas variáveis dummies. Transformou-se todas as variáveis categóricas em n-1 dummies, em que n representa o número total de categorias diferentes da variável em questão. Com variáveis dummies, é possível analisar o impacto de cada categoria individualmente nos modelos preditivos, onde cada dummy pode ter o seu próprio coeficiente, revelando a influência de cada categoria na variável dependente.

Seguidamente, calculou-se as correlações das variáveis com a variável alvo (nº de sets) e os respetivos p-values. O método escolhido para o estudo das correlações foi o ponto bisseral. Este método apenas pode ser utilizado porque há variáveis binárias, que surgiram após fazer-se as dummies.

Correlations where P-Value: 0.05 with Sets

	Variable	Correlation	P-value
4	RankDif	-0.104335	1.700203e-21
2	OpponentRank	-0.095978	2.030625e-18
11	GameRound_Round Robin	0.070059	1.723029e-10
14	WL_W	-0.059673	5.453177e-08
18	OpponentMainHand_Unknown MainHand	-0.058133	1.187920e-07
17	OpponentMainHand_Right-Handed	0.052950	1.417594e-06
12	GameRound_Round of 32	-0.052275	1.926219e-06
9	Ground_Grass	0.052048	2.133674e-06
3	OpponentAge	0.043946	6.290046e-05
0	PlayerRank	-0.042977	9.103814e-05
13	GameRound_Semi-Finals	0.034659	1.602924e-03
8	Season_Varsha (Monções)	-0.034387	1.744185e-03
16	OpponentHand_Right-Handed, Unknown Backhand	0.033668	2.175510e-03
1	OpponentHeight	0.033649	2.188307e-03
7	PlayerMainHand_Unknown MainHand	-0.033438	2.333452e-03
6	PlayerMainHand_Right-Handed	0.028835	8.667405e-03
15	OpponentHand_Right-Handed, Two-Handed Backhand	0.027722	1.161804e-02
5	PlayerHand_Right-Handed, Unknown Backhand	0.027019	1.391078e-02
19	OpponentBackHand_Two-Handed Backhand	0.026981	1.404726e-02
10	GameRound_Quarter-Finals	0.026043	1.775554e-02
20	OpponentBackHand_Unknown Backhand	-0.023983	2.902739e-02

Fig. 20 – Correlações das Variáveis com a Variável Alvo de Forma Decrescente em Módulo

As variáveis com p-value menor que 0.05 rejeitam a hipótese nula, isto é, são consideradas estatisticamente significativas para a variável alvo. O que se pretende através deste estudo é ver qual a variável mais correlacionada (em módulo) com a variável alvo e o p-value menor que 0.05. Como se pode observar pelo output, a variável mais correlacionada é o RankDif (diferença de ranking de jogadores e oponentes), ou seja, a diferença de rankings dos jogadores e oponentes é o que mais influencia o número de sets.

Após esta análise, estudou-se a correlação do RankDif (variável mais correlacionada com a variável alvo) em módulo com as outras variáveis.

Correlations where P-Value: 0.05 with RankDif

	Variable	Correlation	P-value
20	Ground_Hard	-0.022220	4.311730e-02
17	Season_Vasanta (Primavera)	0.023313	3.383195e-02
27	OpponentHand_Left-Handed, Two-Handed Backhand	-0.024973	2.301270e-02
12	PlayerBackHand_Unknown Backhand	0.025735	1.914987e-02
19	Ground_Grass	-0.026087	1.756547e-02
18	Ground_Clay	0.027479	1.237004e-02
28	OpponentHand_Left-Handed, Unknown Backhand	-0.028604	9.217943e-03
21	GameRound_3rd Round Qualifying	-0.032163	3.411049e-03
16	Season_Varsha (Monções)	0.039453	3.278942e-04
32	OpponentMainHand_Left-Handed	-0.040775	2.051046e-04
14	Season_Sharad (Outono)	0.046949	1.908285e-05
9	PlayerHand_Right-Handed, Unknown Backhand	-0.048148	1.161668e-05
4	OpponentHeight	-0.049521	6.487282e-06
8	PlayerHand_Right-Handed, One-Handed Backhand	-0.051567	2.647946e-06
13	Season_Hemanta (Pré-Inverno)	0.052927	1.432403e-06
2	PlayerAge	-0.056011	3.365401e-07
29	OpponentHand_Right-Handed, One-Handed Backhand	-0.070982	9.896642e-11
7	HeightDif	-0.075550	5.742546e-12
10	PlayerMainHand_Right-Handed	-0.081067	1.463824e-13
22	GameRound_Finals	-0.086292	3.592002e-15
30	OpponentHand_Right-Handed, Two-Handed Backhand	-0.089359	3.662836e-16
11	PlayerMainHand_Unknown MainHand	0.093189	1.896276e-17
0	PlayerHeight	-0.093463	1.526514e-17
35	OpponentBackHand_Two-Handed Backhand	-0.095964	2.054756e-18
3	Sets	-0.104335	1.700203e-21
23	GameRound_Quarter-Finals	-0.109345	1.831031e-23
25	GameRound_Semi-Finals	-0.110271	7.737722e-24
24	GameRound_Round of 32	0.122500	4.493211e-29
36	OpponentBackHand_Unknown Backhand	0.129352	2.985219e-32
6	OpponentAge	-0.140363	9.970233e-38
31	OpponentHand_Right-Handed, Unknown Backhand	-0.141828	1.720357e-38
26	WL_W	0.160851	3.800077e-49
15	Season_Shishira (Inverno)	-0.191065	5.797950e-69
33	OpponentMainHand_Right-Handed	-0.244819	2.347864e-113
1	PlayerRank	0.256947	4.643887e-125
34	OpponentMainHand_Unknown MainHand	0.277044	7.011876e-146
5	OpponentRank	0.767402	0.000000e+00

Fig. 21 – Correlações das Variáveis com Rankdif de Forma Crescente em Módulo

O estudo destas correlações tem como principal objetivo perceber a existência de multicolinearidade entre as variáveis e RankDif. A multicolinearidade é identificada através da análise de correlações entre variáveis independentes. Correlações altas indicam multicolinearidade entre variáveis. O que se pretende concluir com isto sãoas variáveis que têm um menor valor de correlação, em módulo, para ser usadas nos modelos.

O output mostra por ordem crescente as variáveis mais correlacionadas com a varíavel RankDif em módulo. Como se pode observar pelo output, as cinco variáveis menos correlacionadas são o Ground_Hard, o Season_Vasanta (Primavera), Opponent_Left-Handed, Two-Handed BackHand, PlayerBackHand_Unknown BackHand e o Ground_Grass.

Os modelos testados na próxima fase incluirão a variável mais correlacionada com a variável alvo (RankDif) e aquelas menos correlacionadas com RankDif.

Modeling

Nesta fase de modelação foram testados diversos modelos. Foi usada a variável alvo (número de sets), a variável mais correlacionada com a variável alvo (RankDif) e as variáveis que estão menos correlacionadas com a RankDif (Ground_Hard, o Season_Vasanta (Primavera), Opponent_Left-Handed, Two-Handed BackHand, PlayerBackHand_Unknown BackHand e Ground_Grass). Todos os modelos foram divididos em conjuntos de treino e teste, com proporções de 70% e 30%, respectivamente.

Para testar todos os modelos, foram filtrados os dados apenas para se estudar os jogos à melhor de 3. Decidiu-se fazer isto porque os dados relativos a jogos à melhor de 5 eram muito poucos e os jogos à melhor de 3 representavam mais de 70% dos dados.
Foram retirados todos os valores nulos e também os jogos onde houvesse desistência, seja antes ou durante o jogo. Tomou-se esta decisão porque não era pertinente prever o número de sets com base em dados dessa natureza.

Num jogo de ténis à melhor de 3, excluindo desistências, os únicos resultados possíveis são o jogo ter 2 sets ou ter 3 sets. Inicialmente, os dados estavam desbalanceados. O desbalanceamento ocorre quando o número de amostras da classe minoritária (sets 3, 1880 dados) é significativamente menor do que o número de amostras da classe maioritária (sets 2, 4748 dados). Esse desequilíbrio pode causar problemas nos modelos, podendo torná-los enviesados em favor da classe maioritária, resultando num mau desempenho na previsão ou classificação pela classe minoritária.

Os modelos iniciais testados foram feitos com os dados desbalanceados. Os resultados destes modelos previam todos os casos para a classe modal (sets 2) e, por isso, foram descartados.

Para resolver o problema do desbalanceamento de classes, primeiramente realizou-se o over-sampling e procedeu-se a uma ferramenta chamada SMOTE.

O SMOTE (Synthetic Minority Over-sampling Technique) resolve esse problema, gerando novas amostras sintéticas para a classe minoritária, em vez de simplesmente duplicar as existentes. O processo funciona da seguinte maneira:

O primeiro passo é identificar as amostras da classe minoritária no conjunto de dados (sets 3).
O segundo passo é, para cada amostra minoritária identificada, o SMOTE seleciona alguns dos seus vizinhos mais próximos com base em uma métrica de distância, como a distância Euclidiana. Isso é feito para garantir que as novas amostras sintéticas sejam semelhantes às amostras minoritárias originais, mas com alguma variação.
O terceiro passo é que as novas amostras sintéticas sejam geradas por interpolação entre a amostra minoritária selecionada e seus vizinhos mais próximos. Especificamente, um dos vizinhos mais próximos é escolhido aleatoriamente, e uma nova amostra é criada ao longo da linha que conecta a amostra minoritária original e o vizinho selecionado. A posição da nova amostra é determinada por uma interpolação linear que coloca a nova amostra num ponto intermediário entre as duas amostras.
No quarto e último passo as novas amostras sintéticas geradas são adicionadas ao conjunto de dados, aumentando o número de amostras da classe minoritária e ajudando a balancear as classes, ou seja, as duas classes ficam com o mesmo número de dados.
Portanto, com esta ferramenta, a classe sets 3 passou a ter 4748 dados, o mesmo número que a classe sets 2.

Foram testados 2 modelos com esta ferramenta, uma regressão logística e uma árvore de decisão.

Numa outra forma de testar os resultados, implementou-se uma estratégia adicional: o undersampling, que consiste em reduzir o número de amostras da classe maioritária para equilibrar a proporção de amostras entre as classes.
O primeiro método utilizado foi o Near Miss. O Near Miss adota uma abordagem mais sistemática e baseada em distância para selecionar as amostras a serem removidas.

Inicialmente, o Near Miss identifica as amostras das classes maioritária e minoritária no conjunto de dados. A técnica calcula a distância entre cada amostra da classe maioritária e as amostras da classe minoritária, utilizando uma métrica de distância, como a distância Euclidiana. Depois seleciona as amostras da classe maioritária cuja média das distâncias aos k vizinhos mais próximos da classe minoritária é a menor. Isto é, mantém as amostras maioritárias que estão, em média, mais próximas das amostras minoritárias. Por fim, o novo conjunto de dados é formado combinando todas as amostras da classe minoritária com as amostras selecionadas da classe maioritária, resultando num conjunto de dados mais equilibrado.
Portanto, com esta ferramenta, a classe sets 2 passou a ter 1880 dados, o mesmo número que a classe sets 3.

Foram testados 2 modelos com esta ferramenta, uma regressão logística e uma árvore de decisão.

O segundo método utilizado de forma a realizar undersampling, foi o Tomek Links. Esta técnica para além de reduzir o número de amostras da classe maioritária, também ajuda a limpar o conjunto de dados, removendo exemplos ruidosos ou ambíguos, podendo reduzir o número de amostras da classe minoritária.

Os Tomek Links são pares de amostras de classes opostas (uma da classe maioritária e outra da classe minoritária) que são os vizinhos mais próximos um do outro. Formalmente, um par de amostras (x_i, x_{ji}) forma um Tomek Link se x_i e x_j são os vizinhos mais próximos um do outro e se x_i pertence a uma classe e x_j pertence à outra classe.

Explicando como funciona esta técnica, inicialmente, para cada amostra no conjunto de dados, identifica-se o seu vizinho mais próximo utilizando uma métrica de distância, como a distância Euclidiana. Seguidamente, verifica-se quais pares de amostras formam Tomek Links. Um par (x_i, x_{ji}) é um Tomek Link se x_i e x_{ji} são vizinhos mais próximos um do outro e pertencem a classes opostas. Por fim, uma vez identificados os Tomek Links, pode-se remover uma ou ambas as amostras do par. Normalmente, remove-se a amostra da classe maioritária para ajudar a balancear o conjunto de dados. No entanto, em alguns casos, pode-se optar por remover ambas as amostras para limpar o conjunto de dados de possíveis ruídos ou ambiguidades.

Foram testados 2 modelos com esta ferramenta, uma regressão logística e uma árvore de decisão.

Após a testagem dos 6 modelos, fez-se validação cruzada. A validação cruzada permite estimar o desempenho do modelo em dados não vistos e ajuda a identificar se o modelo está a sofrer de overfitting (quando um modelo se ajusta excessivamente aos dados de treino) ou underfitting (quando o modelo é muito simples para capturar os padrões subjacentes nos dados de treino).

Posteriormente, realizou-se uma curva roc que compara o desempenho dos modelos realizados. A Curva ROC é um gráfico que plota a taxa de verdadeiros positivos (True Positive Rate - TPR) contra a taxa de falsos positivos (False Positive Rate - FPR) para diferentes pontos de corte.

Em seguida, é apresentada a curva roc.

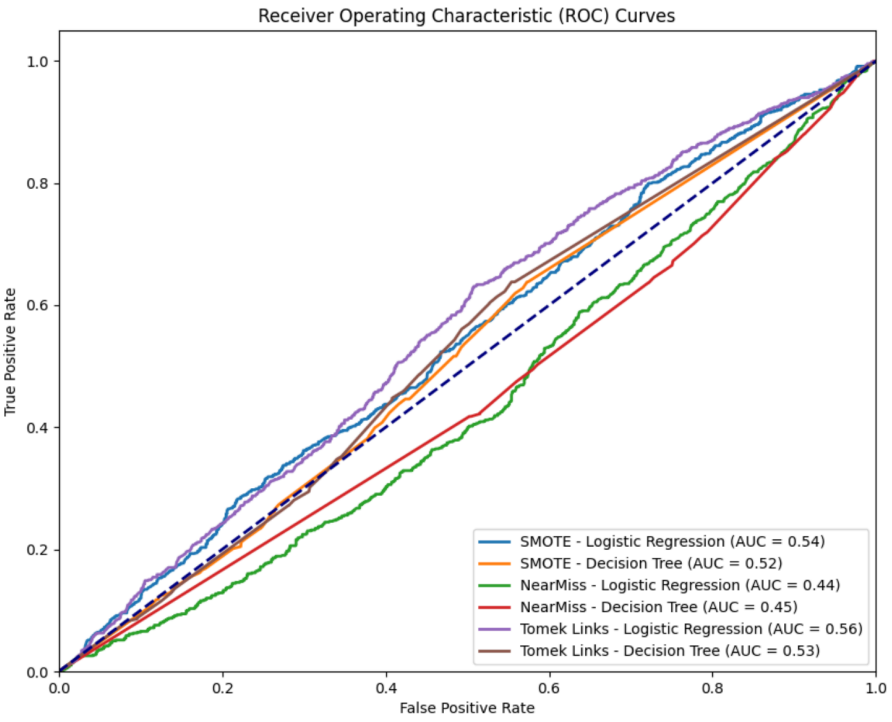

Fig. 22 – Curva Roc

Analisando a curva ROC, revela-se que todos os modelos possuem AUCs relativamente próximos, o que sugere que nenhum modelo se destaca substancialmente em termos de desempenho. O AUC fornece uma medida agregada da performance do modelo em todos os limiares possíveis, com valores entre 0 e 1.

A curva ROC da Regressão Logística e Árvore de Decisão com SMOTE mostram um AUC de 0.54 e de 0.52, respetivamente, ligeiramente acima da linha de classificação aleatória (AUC = 0.5). Isso indica uma capacidade discriminativa marginalmente melhor que a aleatória. No entanto, as performances gerais não são

robustas, o que sugere que, embora o SMOTE ajude a lidar com o desbalanceamento das classes, os resultados para estes modelos específicos ainda são limitados.

A Regressão Logística e Árvore de Decisão com Near Miss apresentam uma AUC de 0.44 e de 0.45, respetivamente, abaixo da linha de aleatoriedade. Estes desempenhos sugerem que a técnica do Near Miss pode não ser adequada para estes modelos específicos, resultando numa pior capacidade discriminativa do que um classificador aleatório.

A Regressão Logística com Tomek Links mostra um AUC de 0.56, o valor mais alto entre os modelos analisados. Embora a AUC seja ligeiramente superior, a matriz de confusão revela que o modelo classificou todas as amostras como negativas, resultando em recall de 0 e F1-Score de 0. Isso destaca uma grave limitação, pois o modelo falha em identificar qualquer amostra positiva, o que é inaceitável em muitos cenários.

A Árvore de Decisão com Tomek Links apresenta uma AUC de 0.53, sugerindo uma capacidade discriminativa ligeiramente melhor que a aleatoriedade. Apesar disso, a performance global em termos de identificação de amostras positivas é fraca, com um recall de 0.33, indicando que o modelo consegue identificar algumas amostras positivas, mas ainda com muitas limitações.

Estes resultados indicam que, enquanto as técnicas de balanceamento podem oferecer algumas melhorias, é crucial considerar a aplicação e ajustar os modelos de acordo. É importante avaliar as curvas ROCs com as matrizes de confusão e com os resultados das métricas de ambos os modelos.

Evaluation

De seguida são mostrados os resultados dos modelos realizados usando o método do SMOTE.

Resultados Modelos SMOTE

Modelos	Accuracy	Cross-Validation	Precision	Recall	F1-Score	Support	AUC
Regressão Logística	0.56	0.5542 ± 0.1146	0.57 0.56	0.56 0.56	0.57 0.56	1206 1177	0.54
Árvore de Decisão	0.65	0.5968 ± 0.0191	0.64 0.66	0.71 0.59	0.67 0.63	1206 1177	0.52

Matriz de Confusão SMOTE - Regressão Logística

679	527
516	661

Matriz de Confusão SMOTE - Árvore de Decisão

851	355
477	700

Analisando os resultados, a Árvore de Decisão apresentou um desempenho superior à Regressão Logística na maioria das métricas avaliadas. A accuracy da Árvore de Decisão foi de 0.65, enquanto a Regressão Logística obteve uma accuracy de 0.56. Este resultado sugere que a Árvore de Decisão tem uma maior capacidade de classificar corretamente as amostras nos seus respetivos grupos.

A precisão da Árvore de Decisão foi de 0.64, enquanto a da Regressão Logística foi de 0.57. Este resultado indica que a Árvore de Decisão é mais eficaz em prever corretamente as amostras positivas. Adicionalmente, o recall da Árvore de Decisão foi significativamente mais alto (0.66) em comparação com a Regressão Logística (0.56), evidenciando uma melhor capacidade de identificar todas as amostras

positivas. Este equilíbrio entre precisão e recall reflete-se no F1-Score, onde a Árvore de Decisão obteve 0.65, comparado com 0.56 da Regressão Logística.

A AUC da Regressão Logística foi de 0.54, ligeiramente superior à AUC da Árvore de Decisão, que foi de 0.52. No entanto, ambos os valores estão muito próximos do valor de 0.5, o que indica que nenhum dos modelos tem uma separação significativa entre as classes. Este resultado sugere que, embora a Regressão Logística tenha um desempenho ligeiramente melhor em termos de AUC, a diferença não é suficiente para contrabalançar o desempenho inferior nas outras métricas.

A matriz de confusão fornece informações detalhadas sobre os tipos de erros cometidos por cada modelo. A Árvore de Decisão teve 851 verdadeiros positivos e 700 verdadeiros negativos, enquanto a Regressão Logística teve 679 verdadeiros positivos e 661 verdadeiros negativos. Além disso, a Árvore de Decisão cometeu menos erros, com 355 falsos positivos e 477 falsos negativos, comparados com 527 falsos positivos e 516 falsos negativos da Regressão Logística. Estes números mostram que a Árvore de Decisão não só é mais precisa na previsão de positivos, como também é mais eficaz na minimização de falsos negativos, o que é crucial em muitos cenários aplicados.

A accuracy da validação cruzada foi ligeiramente inferior para ambos os modelos, com a Árvore de Decisão a obter 0.5968 e a Regressão Logística 0.5542. Esta diferença sugere um ligeiro sobreajuste dos modelos aos dados de treino, mas a Árvore de Decisão ainda mantém uma vantagem significativa sobre a Regressão Logística.

A Árvore de Decisão demonstrou ser o modelo mais eficaz para este conjunto de dados balanceado com SMOTE, superando a Regressão Logística em quase todas as métricas. Apenas no valor da AUC é que a Árvore de Decisão é ligeiramente melhor. No entanto, a baixa AUC para ambos os modelos sugere que há espaço para melhorias.
De seguida são mostrados os resultados dos modelos realizados usando o método do Near Miss.

Resultados Modelos Near Miss

Modelos	Accuracy	Cross-Validation	Precision	Recall	F1-Score	Support	AUC
Regressão Logística	0.64	0.4196 ± 0.0383	0.60 0.72	0.79 0.50	0.68 0.59	454 478	0.44
Árvore de Decisão	0.56	0.4305 ± 0.0241	0.53 0.62	0.77 0.35	0.63 0.45	454 478	0.45

Matriz de Confusão Near Miss - Regressão Logística

360	94
239	239

Matriz de Confusão Near Miss - Árvore de Decisão

349	105
309	169

Analisando os resultados, a Regressão Logística apresentou um desempenho superior à Árvore de Decisão na maioria das métricas avaliadas. A accuracy da Regressão Logística foi de 0.64, enquanto a Árvore de Decisão obteve uma accuracy de 0.56. Este resultado sugere que a Regressão Logística tem uma maior capacidade de classificar corretamente as amostras nos seus respetivos grupos.

A precisão da Regressão Logística foi de 0.60, comparada com 0.53 da Árvore de Decisão. Este resultado indica que a Regressão Logística é mais eficaz em prever corretamente as amostras positivas. Adicionalmente, o recall da Regressão Logística foi significativamente mais alto (0.72) em comparação com a Árvore de Decisão (0.62), evidenciando uma melhor capacidade de identificar todas as amostras positivas. Este equilíbrio entre precisão e recall reflete-se no F1-Score, onde a Regressão Logística obteve 0.68, comparado com 0.63 da Árvore de Decisão.

A AUC da Árvore de Decisão foi de 0.45, ligeiramente superior à AUC da Regressão Logística, que foi de 0.44. No entanto, ambos os valores estão muito próximos do valor de 0.5, o que indica que nenhum dos modelos tem uma separação significativa

entre as classes. Este resultado sugere que, embora a Árvore de Decisão tenha um desempenho ligeiramente melhor em termos de AUC, a diferença não é suficiente para contrabalançar o desempenho inferior nas outras métricas.

A Regressão Logística teve 360 verdadeiros positivos e 239 verdadeiros negativos, enquanto a Árvore de Decisão teve 349 verdadeiros positivos e 169 verdadeiros negativos. Além disso, a Regressão Logística cometeu menos erros, com 94 falsos positivos e 239 falsos negativos, comparados com 105 falsos positivos e 309 falsos negativos da Árvore de Decisão. Estes números mostram que a Regressão Logística não só é mais precisa na previsão de positivos, como também é mais eficaz na minimização de falsos negativos.

A accuracy de validação cruzada foi baixa para ambos os modelos, com a Árvore de Decisão a obter 0.4305 e a Regressão Logística 0.4196. Esta diferença sugere um ligeiro sobreajuste dos modelos aos dados de treino, mas a Regressão Logística ainda mantém uma vantagem significativa sobre a Árvore de Decisão.

A Regressão Logística demonstrou ser o modelo mais eficaz para este conjunto de dados balanceado com Near Miss, superando a Árvore de Decisão em termos de accuracy, precisão, recall e F1-Score. Apesar da ligeira vantagem da Árvore de Decisão em AUC, a Regressão Logística apresentou uma performance globalmente melhor, sendo mais eficiente na detecção de amostras positivas e na minimização de erros. No entanto, a baixa AUC para ambos os modelos sugere que há espaço para melhorias adicionais.

Resultados Modelos Tomek Links

Modelos	Accuracy	Cross-Validation	Precision	Recall	F1-Score	Support	AUC
Regressão Logística	0.70	0.7188 ± 0.0002	0.70 0	1 0	0.82 0	1153 493	0.56
Árvore de Decisão	0.66	0.6455 ± 0.0107	0.70 0.33	0.88 0.13	0.78 0.19	1153 493	0.53

Matriz de Confusão Tomek Links - Regressão Logística

1153	0
493	0

Matriz de Confusão Tomek Links - Árvore de Decisão

1018	135
428	65

Analisando os resultados, a Regressão Logística obteve uma accuracy de 0.70, superior à accuracy da Árvore de Decisão, que foi de 0.66. Esta diferença sugere que a Regressão Logística possui uma capacidade ligeiramente maior de classificar corretamente as amostras.

A precisão da Regressão Logística foi de 0.70, igual à da Árvore de Decisão. No entanto, o recall da Regressão Logística foi de 0, enquanto a Árvore de Decisão obteve um recall de 0.33. Este resultado indica que a Regressão Logística não conseguiu identificar nenhuma das amostras positivas (classes minoritárias), ao contrário da Árvore de Decisão, que teve algum sucesso nessa tarefa. A ausência de recall na Regressão Logística resulta num F1-Score de 0, comparado com 0.13 da Árvore de Decisão.

A AUC da Regressão Logística foi de 0.56, ligeiramente superior à AUC da Árvore de Decisão, que foi de 0.53. Esses valores indicam que ambos os modelos têm uma capacidade limitada de discriminar entre as classes, embora a Regressão Logística apresente uma ligeira vantagem.

A matriz de confusão da Regressão Logística mostrou 1153 verdadeiros negativos e 493 falsos negativos, sem identificar nenhum verdadeiro positivo ou falso positivo. Isto sugere que a Regressão Logística classificou todas as amostras como negativas. Por outro lado, a matriz de confusão da Árvore de Decisão indicou 1018 verdadeiros negativos, 135 falsos negativos, 428 falsos positivos e 65 verdadeiros positivos. Embora a Árvore de Decisão tenha identificado alguns positivos, ainda há uma quantidade significativa de falsos negativos e falsos positivos, o que prejudica sua eficácia geral.

A accuracy de validação cruzada foi mais alta para a Regressão Logística (0.7188) em comparação com a Árvore de Decisão (0.6455). Este resultado sugere que a Regressão Logística pode ter um desempenho mais consistente em novos conjuntos de dados, apesar das limitações observadas na matriz de confusão.

Em resumo, a Regressão Logística apresentou uma accuracy ligeiramente superior, mas falhou em identificar qualquer amostra positiva, o que é evidenciado por um recall e F1-Score de 0. Por outro lado, a Árvore de Decisão, embora com uma accuracy ligeiramente inferior, mostrou-se capaz de identificar algumas amostras positivas, refletindo-se em métricas de recall e F1-Score superiores. Ambos os modelos têm uma AUC relativamente baixa, indicando uma capacidade limitada de discriminação entre classes.

Estes resultados sugerem que, apesar da Regressão Logística apresentar uma accuracy ligeiramente superior, a Árvore de Decisão oferece um desempenho mais equilibrado em termos de identificação de classes minoritárias. A Regressão Logística não é recomendável devido à sua falha em reconhecer a classe minoritária. A Árvore de Decisão, embora não ideal, mostrou-se mais equilibrada.

Deployment

Nenhum dos modelos demonstra um equilíbrio ótimo de alta accuracy, alta sensibilidade e alta especificidade. O desafio aqui parece estar no desenvolvimento de um modelo que possa efetivamente distinguir entre as classes no conjunto de dados sem ser tendencioso para uma classe ou comprometer a precisão geral.

Com a ferramenta SMOTE ambas as técnicas (Regressão Logística e Árvore de Decisão) apresentaram uma ligeira melhoria na capacidade discriminativa em comparação com a aleatoriedade, mas ainda assim não robustas o suficiente. A Árvore de Decisão com SMOTE apresenta a melhor combinação de accuracy (0.65) e F1-Score (0.65), mas a AUC de 0.52 indica uma capacidade discriminativa marginal. Os modelos com Near Miss apresentaram AUCs abaixo da aleatoriedade e valores da accuracy muito baixos, indicando uma eficácia limitada ou até prejudicial para este conjunto de dados específico. A Regressão Logística com Tomek Links teve a maior AUC (0.56) e accuracy (0.70), mas falhou em prever qualquer amostra positiva. A Árvore de Decisão com Tomek Links teve uma AUC razoável de 0.53 e um bom accuracy de 0.66, mas uma baixa capacidade de identificar amostras positivas e um recall muito baixo.

No geral, estas limitações são um indicativo de um problema subjacente com os dados ou pode sugerir que o problema de prever o número de séries em uma partida de ténis é particularmente difícil de modelar com precisão.

No futuro, possíveis melhorias e possíveis soluções são considerar outras técnicas de balanceamento, como ADASYN ou combinar técnicas (por exemplo SMOTE+Tomek Links). Outro procedimento poderia ser ajustar hiperparâmetros dos modelos de classificação ou experimentar com outros algoritmos, como Random Forest ou Gradient Boosting, bem como realizar uma validação cruzada mais extensa e análise em diferentes subsets de dados para garantir que os resultados sejam consistentes e generalizáveis.

References

- ATP Tour (n.d.). History. Retrieved October 12, 2023, from https://www.atptour.com/en/.
- Bryman, A., & Cramer, D. (2001). Quantitative Data Analysis with SPSS Release 10 for Windows. Philadelphia: Routledge.
- Candila, V. and Palazzo, L. , (2020) , Neural Networks and Betting Strategies for Tennis, Risks, 8: , 68.
- Cornman, A. , Spellman, G. and Wright, D. , (2017) , Machine Learning for Professional Tennis Match Prediction and Betting, Working Paper, Stanford University, December.
- De Araujo Fernandes, M. , (2017) , Using Soft Computing Techniques for Prediction of Winners in Tennis Matches, Machine Learning Research, 2: (3), 86–98.
- Sipko, M. , (2015) , Machine Learning for the Prediction of Professional Tennis Matches, Master's Thesis, Imperial College London, June.
- Somboonphokkaphan, A. , Phimoltares, S. and Lursinsap, C. , (2009) , Tennis Winner Prediction Based on Time-Series History with Neural Modeling, Proceedings of the International MultiConference of Engineers and Computer Scientists, Hong Kong