



Curso de Licenciatura em Ciência de Dados

2º ano 2º semestre - 2024/2025

Unidade Curricular:

Projeto Aplicado em Ciência de Dados I

Projeto:

**Previsão do número de sets para a conclusão de
um jogo de ténis Profissional**

Grupo 6:

António Santos, nº 123434;

Bernardo Alho 123431;

Gonçalo Henriques, nº 123422;

José Alberto nº 121959;

Maria Inês nº 123393.

Docente: Sérgio Moro

Lisboa, Abril de 2025

ÍNDICE

1.	INTRODUÇÃO	1
2.	BUSINESS UNDERSTANDING	3
3.	DATA PRE-PROCESSING	4
3.1	Variáveis do <i>Dataset</i> Original	4
3.2	<i>MongoDB</i>	5
3.2.1	<i>Datasets</i> Utilizados	5
3.2.2	<i>Collections Grounds, Rounds e Tournaments</i>	7
3.2.3	<i>Collection ATP2</i>	7
3.2.4	<i>Collection Players</i>	8
3.2.5	Variável <i>Born</i>	8
3.2.6	Variável <i>Height</i>	9
3.2.7	Variável <i>Hand</i>	10
3.2.8	<i>Collection Confrontos</i>	11
3.2.9	<i>Collection Season</i>	12
3.2.10	<i>Collection Game</i>	12
3.3	<i>MySql</i>	14
3.4	Tratamento base em <i>Python</i>	15
3.4.1	<i>Players</i>	17
3.4.2	<i>Games</i>	19
3.4.3	Tratamento Aprofundado	20
4.	DATA UNDERSTANDING	23
4.1	Descrição das variáveis originais	23
4.2	Análises iniciais	24
4.3	Criação e enriquecimento de variáveis	25
4.4	Análise Exploratória	26

5. DATA PREPARATION	28
5.1 Seleção de Variáveis e Motivação baseada no contexto do Ténis.....	29
5.2 <i>SMOTE Family</i>	30
5.3 <i>Dummy</i>	31
5.4 <i>Standard Scaler</i>	31
5.5 Técnicas de Avaliação dos Modelos.....	32
5.5.1 Divisão em treino e teste	32
5.5.2 Validação cruzada (<i>Cross Validation</i>)	32
6. MODELING.....	34
7. EVALUATION	36
8. DEPLOYMENT	40
9. CONCLUSÃO.....	41
10. REFERÊNCIAS BIBLIOGRÁFICAS	42
11. ANEXOS.....	43
Anexo 1	43
Anexo 2	43
Anexo 3	44
Anexo 4	44
Anexo 5	44
Anexo 6	45
Anexo 7	45
Anexo 8	45
Anexo 9	46
Anexo 10	46
Anexo 11	47
Anexo 12	47

Anexo 13	48
Anexo 14	48
Anexo 15	49
Anexo 16	49
Anexo 17	49
Anexo 18	50
Anexo 19	50
Anexo 20	50
Anexo 21	51

1. INTRODUÇÃO

O presente relatório foi desenvolvido no âmbito da unidade curricular “Projeto Aplicado em Ciência de Dados I”, onde o mesmo é referente a um projeto que incidiu na criação de um modelo preditivo capaz de prever o número de sets de um jogo de ténis realizado na Áustria. Com o crescente volume e acessibilidade de dados públicos sobre torneios da *Association of Tennis Professionals (ATP)*, incluindo características físicas dos jogadores, estatísticas de jogo e contextos dos encontros, abrem-se novas oportunidades para desenvolver modelos preditivos acessíveis e úteis para jogadores, treinadores e analistas.

O ténis ocupa um lugar de destaque no mundo do desporto, sendo o quarto desporto mais popular a nível global, com aproximadamente mil milhões de adeptos. A modalidade é regulada por várias organizações, entre elas a *ATP*, que supervisiona os torneios individuais e em duplas masculinos, a *Women's Tennis Association (WTA)*, responsável pelas competições femininas, e a *International Tennis Federation (ITF)*, que atua como entidade reguladora geral, definindo as regras e organizando torneios em ambas as categorias. Cada uma destas entidades desempenha um papel essencial na estruturação e evolução do ténis profissional, desde a definição de regulamentos à realização de eventos internacionais.

Este trabalho enquadra-se, assim, na área da análise preditiva aplicada ao desporto, com foco na criação de um modelo de *machine learning* para prever o número de sets jogados numa partida de ténis de jogadores profissionais, com base em dados históricos de torneios *ATP*. Este tipo de previsão é relevante não só do ponto de vista competitivo e estratégico, mas também, como referido anteriormente, pode ser uma ferramenta útil para treinadores, jogadores e analistas, permitindo uma preparação mais informada e eficaz antes das partidas.

O desenvolvimento do modelo preditivo seguirá, como indicado, a metodologia *CRISP-DM* (*Cross-Industry Standard Process for Data Mining*). É uma abordagem estruturada e amplamente utilizada em projetos de *data science*. O trabalho será conduzido ao longo das seguintes fases, que ao longo do projeto foram sendo iteradas:

1. *Business Understanding*, para clarificar os objetivos e contexto do problema;

2. *Data Understanding*, para explorar e conhecer a origem, estrutura e qualidade dos dados disponíveis;
3. *Data Preparation*, onde serão realizadas as transformações e integrações necessárias;
4. *Modeling*, que envolve a aplicação de algoritmos de *machine learning*;
5. *Evaluation*, para medir a performance dos modelos e selecionar o mais adequado;
6. *Deployment*, onde se discute a aplicação prática dos resultados e o seu impacto.

No que diz respeito às tecnologias e linguagens utilizadas, o projeto adotou uma abordagem que combina várias ferramentas de tratamento e análise de dados. Inicialmente, foi utilizado *MongoDB* para armazenar e organizar dados não estruturados provenientes de fontes públicas, como o site *ATP*. Posteriormente, estes dados foram convertidos e integrados num modelo relacional em *SQL*, permitindo uma estruturação mais adequada para análises estatísticas. A fase de análise estatística e visualização dos resultados dos modelos foi realizada em *Python*, com recurso a bibliotecas especializadas em ciência de dados. Por fim, foi utilizado o *RStudio* para a implementação dos modelos classificativos desenvolvidos com as diferentes técnicas de validação.

O projeto poderá futuramente ser expandido para incluir jogos de *Grand Slam*, bem como torneios femininos, abrangendo assim uma variedade mais ampla de competições e contextos no ténis profissional alargando o alcance e aplicabilidade do modelo desenvolvido.

2. BUSINESS UNDERSTANDING

A análise e previsão do número de sets num jogo de ténis apresenta um valor prático relevante para vários intervenientes no contexto desportivo. Para atletas e treinadores, esta informação pode apoiar a gestão da carga física, a definição de estratégias personalizadas e o planeamento de treinos mais eficazes. As equipas técnicas e analistas beneficiam também de uma base adicional de apoio à decisão, enquanto o mercado de apostas desportivas pode tirar proveito de previsões mais fiáveis para otimizar lucros. Além disso, entidades como canais de televisão podem utilizar essas previsões para melhorar o planeamento da grelha de programação, antecipando a duração dos encontros.

Esta abordagem permite identificar nuances importantes do jogo como o equilíbrio competitivo entre os jogadores ou o tipo de piso que influencia a duração do jogo e consequentemente o número de sets disputados.

Deste feito, o problema em análise, pode assim ser definido da seguinte forma: Como prever, antes de um jogo, o número de sets que é provável serem disputados com base em dados históricos e características dos jogadores e do contexto?

O objetivo central do projeto consiste em construir um modelo preditivo baseado em *Machine Learning* que estime com elevada precisão o número de sets que serão disputados num jogo de ténis, com base num conjunto de variáveis disponíveis relacionadas com os jogadores, como altura, país de naturalidade, e se é destro ou esquerdino, e com o contexto do jogo, como data, tipo de piso, prémio, ronda e número de sets para a tomada de decisão da partida.

Para garantir a utilidade do modelo desenvolvido no contexto real, o mesmo deverá ser suficientemente intuitivo e explicável, permitindo que profissionais de ténis sem formação em ciência de dados o possam utilizar com confiança. Este projeto procura também dar resposta a questões relevantes para o planeamento e a análise no desporto, tais como: existirão padrões em certos tipos de torneios que favoreçam jogos mais longos? Jogadores com estilos semelhantes tendem a ter confrontos mais prolongados? E de que forma se pode ajustar o planeamento dos treinos quando se antecipa um jogo com maior duração?

3. DATA PRE-PROCESSING

Para a concretização deste projeto, foi fornecido um *dataset* no formato *json*, relativo a jogos individuais de 10.361 jogadores masculinos, incluindo os 500 melhores jogadores que jogaram entre 28/03/1973 e 14/02/2022. Trata-se de uma base de dados recolhida de <https://www.atptour.com/en>, fornecida pela *ATP*, cujos dados estão publicamente disponíveis a fins de pesquisa e análise (*open data*), que consiste, como referido, no circuito mundial de ténis profissional masculino de primeira linha.

O processo inicial de pré-processamento exigiu uma compreensão detalhada do significado de cada uma das 15 variáveis iniciais, que foi vista como uma fase de pré entendimento dos dados. Para tal, foram realizadas várias pesquisas, nomeadamente no próprio site do *ATP* e, assim garantir que os ajustes realizados ao *dataset* fossem os mais acertados.

Este *dataset* já havia sido previamente alvo de estudo em um projeto desenvolvido na Unidade Curricular em Armazenamento para Big Data (1º Semestre do decorrente ano letivo), contudo com o objetivo de construir um modelo relacional de modo a garantir assim a máxima integridade e consistência dos dados para a realização dos *Select's* solicitados. Para o presente projeto, o foco não está na construção de um modelo relacional, mas sim na construção de um modelo preditivo, contudo estes passos foram reutilizados e ajustados, a fim de proceder à extração do *dataset* final com a máxima consistência dos dados, realizar o tratamento das variáveis que o exigiram e, por fim, às análises e seleções das variáveis preditoras e à identificação e criação da variável alvo do modelo.

Esta seção tem, assim, como foco a descrição de todos os passos tomados na construção do modelo relacional para a obtenção do *dataset* final em *Sql*. Importante mencionar que todos estes passos podem ser realizados com auxílio do ficheiro “*Step by Step.txt*”, que contém um *roadmap* de como devem ser utilizados os ficheiros que permitiram obter este modelo relacional.

3.1 Variáveis do Dataset Original

Como mencionado foram estudadas previamente as *features* do *dataset* fornecido e assim foi possível retirar as seguintes conclusões descritas. O *dataset* contém inicialmente 1308835 registos, em que cada um diz respeito a um jogo realizado entre 2

jogadores e, 15 variáveis que reúnem informações sobre os jogadores, torneio e o jogo em que se enquadra. Cada linha possui um identificador único (*_id*) e detalhes sobre os jogadores, como os nomes (*PlayerName* e *Oponent*) e, para o respetivo jogador que se encontra em *PlayerName* são descritas as seguintes variáveis: local de nascimento (*Born*), país, cidade ou estado, altura (*Height*), mão predominante e tipo de backhand (*Hand*) e um link identificador do próprio site do *ATP* (*LinkPlayer*). Em relação ao jogador que se encontra em *Oponent* é fornecido o seu *rank* no momento do decorrer do respetivo jogo (*GameRank*).

Os torneios são descritos com variáveis como o nome (*Tournament*), o local (*Location*) que também pode ser país, cidade ou estado, data de início e fim (*Date*), tipo de superfície (*Ground*) e o prémio que é distribuído pelos jogadores consoante a sua performance no torneio em questão (*Prize*).

No contexto dos jogos, são fornecidas informações como a ronda (*GameRound*), (*WL*) que indica se o jogador que se encontra em *PlayerName* saiu vitorioso da partida e a variável (*Score*) que indica o resultado em relação ao jogador referente ao *PlayerName* do jogo ocorrido.

3.2 MongoDB

Dado o formato original dos dados, *Json*, estes passos, para a construção do modelo relacional, foram realizados inicialmente em *MongoDB* dado que se trata de um banco de dados *NoSql* (*Not Only Sql*) baseado em documentos *Json* e permite o processamento eficiente de um grande volume de dados semi-estruturados, o que se alinha às necessidades deste *dataset* fornecido (*atpplayers.json*).

3.2.1 Datasets Utilizados

É importante realçar que os *datasets* utilizados neste primeiro processamento foram de acordo às necessidades surgidas no projeto realizado anteriormente com este dataset original. No entanto, ao longo deste relatório, serão apresentados outras base de dados que também foram necessárias para o desenvolvimento completo do trabalho.

Para o estudo das referidas variáveis *Atp* (secção [3.1](#)) foi feito o *import* do *dataset* original para a *collection* *atp*.

Os ficheiros *csv* a seguir mencionados foram utilizados para solucionar problemas que as variáveis *Born*, *Hand* (na *collection* *Players*) e *Location* (nas *collections* *Season* e

Confrontos) apresentaram. *Datasets* esses que foram obtidos da web e devidamente ajustados, em *Python*, de modo a estarem adequados à tarefa pretendida com o mesmo.

O *csv* designado de “*all_players_singles.csv*”, obtido a partir de um *dataset* disponível no *Kaggle* (Hallmar, 2018), cujo foi importado para a *collection all_players*. Ficheiro este utilizado para o tratamento da variável *Born* (servindo de apoio para a padronização dos dados relativos ao local de nascimento dos jogadores) em 2 formas: atualizando os campos dos jogadores já presentes na base de dados inicial, convertendo os valores para siglas, com o objetivo de reduzir a redundância dos países, cidades e estados e, inserir as siglas dos países dos jogadores que não continham este campo, mas que se encontravam neste *csv*. O ajuste do mesmo incluiu apenas os jogadores individuais, tendo sido excluídas as observações que continham o carácter “_”, por se referirem a duplas.

Ainda no tratamento da *collection Players*, para a variável *Hand* foi utilizado um ficheiro *csv* (*all_players_hand.csv*) obtido a partir do mesmo website do *csv* anterior mencionado, *Kaggle* (Brownlow, 2021), que foi importado para a *collection all_players_hand*. O seu processamento apenas consistiu na filtragem dos jogadores do sexo masculino e na exclusão dos registo em que esta variável se encontrava vazia e, após esta filtragem, criou-se o *csv* referido.

Por fim, para o tratamento nas *collections Confrontos* e *Season* para a variável *Location* foi semelhante ao que foi aplicado à variável *Born* na *collection Players*. A principal diferença consistiu na forma de atribuição das siglas dos correspondentes países, que, neste caso, foi realizada manualmente, sem o auxílio de um ficheiro *csv* e, assim, contribuir para o objetivo de reduzir a redundância dos valores que se referiam ao mesmo país. As siglas atribuídas em *Location* e *Born* foram baseadas em três códigos: ISO-ALPHA-3 (ISO 3166-1 alfa-3), IOC (*International Olympic Committee*) e FIFA (*Fédération Internationale de Football Association*). Para esse fim, foi utilizado um *csv* ao qual se designou *codes_siglas.csv*, obtido a partir do *Kaggle* (Bohnacker, 2022), cujo *import* se realizou para a *collection codes_siglas*. O ajuste consistiu na padronização das siglas de países para códigos de três caracteres, com o objetivo de garantir consistência e uniformidade. Dessa forma, cada país passou a ser representado por uma única sigla, eliminando possíveis ambiguidades. Este processo foi essencial para a criação de um modelo relacional robusto, onde cada país será unicamente associado a uma única sigla.

3.2.2 Collections Grounds, Rounds e Tournaments

A criação destas *collections*, que dizem respeito ao tipo de pavimento em que ocorreu o torneio, à ronda em que foi realizado o jogo e ao nome do torneio em questão, foi baseada na ideia de manter a ideologia do modelo relacional, de modo, que quando algum registo fosse inserido nas tabelas “filhas” tivesse de cumprir com as informações encontradas nestas tabelas e, assim, posteriormente foram exportadas do *MongoDB* para o *MySQL*.

As *collections* *Grounds* e *Rounds* não exigiram o seu processamento em *MongoDB*, apenas foram criadas para manter a organização dos dados, facilitando a futura migração para o modelo relacional. O único tratamento efetuado foi nos registos da *collection* *Tournaments*, que foram ajustados posteriormente em *MySQL*.

A criação destas *collections* provém da que foi importada inicialmente, *atp*, onde nesta foram identificados 6 torneios (em uma data específica) que não continham a variável “*Ground*” preenchida, apresentando este valor como “”, pelo que foi atualizado para “*Unknown*”. Irá ser mencionado mais à frente neste relatório, na criação da *collection* *Season*, o procedimento que foi tomado no tratamento destes registos.

3.2.3 Collection ATP2

A criação desta *collection* foi motivada pela necessidade do tratamento de algumas variáveis importantes como, *Date*, *Born*, *Location* e *Prize*, pelo que no processo da criação da mesma foram tomados em consideração os seguintes critérios.

Primeiramente, considerou-se que as datas podiam ser analisadas de forma separada para o mesmo torneio, ou seja, criou-se a variável *Start*, que se refere à data em que o mesmo se iniciou e *End*, à data final do mesmo.

No que diz respeito às variáveis referentes a localizações, *Born* e *Location*, foram tratadas como referido na secção [3.2.1](#). Contudo, antes de se considerar apenas a sigla para a respetiva localização (do país), esta variável foi atualizada de modo a conter apenas o nome após a última vírgula e espaço deste campo.

Por fim, no que toca ao valor monetário do torneio, a variável *Prize*, verificou-se que apenas existiam 2 tipos de moedas, os dólares (\$) e os dólares Australianos (A\$) na nossa base de dados. Verificou-se, em alguns casos, que para o mesmo torneio foram atribuídos dois valores monetários diferentes, em que apenas diferia no primeiro caractere

(?), indicando um possível lapso no momento da criação desta base de dados. Foi, assim, verificado para estes campos que continham estes caracteres que diziam respeito à moeda dólar (\$), pelo que foram atualizados para esta respetiva moeda.

3.2.4 Collection Players

Na *collection* *atp2* verificou-se que existiam alguns jogos em que o campo *Oponent* estava vazio e o *Score* não. Assim, foram feitas pesquisas no site do *ATP* e *ITF* e conseguiu-se encontrar os jogadores em falta, tendo agora em consideração mais jogadores na criação desta *collection Players* proveniente da *collection atp2*.

A criação desta *collection* iniciou-se com a inserção dos jogadores com a variável *LinkPlayer* que diz respeito aos jogadores que se encontravam pelo menos uma vez no campo *PlayerName*, inserção esta que resultou em 9960 jogadores. Após isto, verificou-se que existiam 7 jogadores com o mesmo nome, mas como estes continham a variável do identificador único foi possível de os distinguir, não resultando em possíveis uniões de jogadores diferentes.

Por fim, inseriu-se os jogadores que apenas se encontravam no campo *Oponent* e assim obteve-se todos os jogadores da base de dados. Esta inserção foi feita através da variável *Oponent*, podendo vir a resultar no agrupamento do mesmo jogador, pois poderiam ter o mesmo nome. Contudo em *Oponent* isso não se verifica, pois, o número de jogadores com o mesmo nome manteve-se igual (7) na inserção destes jogadores. Deste modo obteve-se um número total de jogadores distintos igual a 22696.

3.2.5 Variável Born

Esta variável é bastante importante para o problema em questão, pelo que, verificou-se que a grande maioria dos jogadores não continha o campo desta variável. Foram feitas algumas análises/pesquisas e chegou-se à conclusão que a maneira mais eficiente de tratar destes campos vazios (imputação de dados) foi arranjar o *csv* referido na secção [3.2.1](#), importado para a *collection all_players*, que contém vários jogadores de ténis, onde cada registo contém o país de nascimento do respetivo jogador, em sigla.

Após realizado o *import* deste *csv*, em primeiro lugar, foram tratados os campos dos jogadores na *collection Players* que se encontravam nesta *collection* e continham o mesmo preenchido com a finalidade de reduzir a redundância dos “países”, ou seja, jogadores que tenham nascido no mesmo país possuírem apenas uma designação para este. Destes jogadores que já possuíam esta variável preenchida, nem todos os valores

diziam respeito a um país em concreto, ou seja, existia casos em que se referiam a cidades, a estados e outros a países. Verificou-se que destes jogadores, a grande maioria estava presente nesta *collection*, e, portanto, foi alterada esta variável para reduzir a referida redundância. Não se verificou a atualização para todos estes jogadores dado que uma parte dos mesmos já continha a sigla correspondente à que se encontrava na *collection*. Em relação aos que possuíam esta variável e não se encontravam nesta *collection*, tiveram de ser atualizados manualmente os valores dos mesmos, de modo que a ideia de reduzir a redundância fosse alcançada.

Posteriormente procedeu-se ao tratamento do preenchimento dos campos vazios desta variável em análise, que inicialmente se referiam à grande maioria dos jogadores. Importante voltar a referir que todos os valores destas atualizações encontram-se no script *atpplayers.json*. O código que nos permitiu a inserção destes campos vazios foi crucial neste passo, contudo, como era de esperar, ainda nos deparamos com uma quantidade razoável de jogadores que não continham esta variável e também não constavam nesta *collection*. Para estes foi feita a atribuição do valor “*GHOST-FLAG*” que é o tradicionalmente utilizado para indicar que uma determinada localização não é conhecida.

Verificou-se que existia um país, nesta *collection* (*all_players*) que foi introduzido com o nome do país ao invés da sigla a que lhe diz respeito, país este que corresponde à Alemanha (*Germany*), pelo que foi feita a alteração do campo dos jogadores com este valor para “*DEU*”.

Por fim, foram feitos alguns ajustes às siglas que foram atribuídas aos jogadores após estas atualizações, nomeadamente através da collection *codes_siglas*, para que estas estivessem de acordo com o *csv* que se irá utilizar em *MySQL* (*All_Country_code.csv*) para fazer a correspondência da sigla ao país, ou seja, para torná-las homogéneas. Isto acontece, pois como mencionado, as siglas foram atribuídas de acordo com 3 códigos (ISO-Alpha-3, IOC e FIFA) ou até outros que já não se encontram em uso atualmente. Desta feita, podemos reduzir de forma significante o número de jogadores sem país na base de dados.

3.2.6 Variável Height

Esta variável apesar de se puder demonstrar relevante nas fases seguintes de análise em *Python* não foi tratada nesta fase no sentido da imputação de dados para os jogadores que não a continham. Contudo pudemos tirar algumas conclusões iniciais e

tomadas de decisão em relação à mesma, nomeadamente como ilustrado no [anexo 1](#), é de refletir que a atribuição das alturas dos jogadores encontra-se em diferentes medidas de comprimento como, o metro e o *foot* (habitualmente utilizado nos Estados Unidos). Também se pode inferir que existem bastantes jogadores com a altura igual a 0, pelo que, de modo a uniformizar os valores foi também colocado o valor 0 aos jogadores que apenas se encontravam em *oponent*.

3.2.7 Variável Hand

Esta variável foi bastante importante no que toca ao objetivo deste projeto, prever o número de *sets*, dado que dependente do tipo de *ForeHand* e *Backhand* que os jogadores apresentam o jogo poderá ser resolvido em menos ou mais sets. Esta análise relacional irá ser reforçada mais à frente no presente relatório. Dado este critério, seria essencial puder analisar em separado o tipo de *Hand* e *Backhand* que um determinado jogador apresente. Assim, o primeiro tratamento desta variável consiste na separação desta variável em 2 variáveis, “*Hand*” e “*BackHand*”.

Verificou-se que cerca de metade dos jogadores não apresentavam a variável *Hand* preenchida e, como esta foi crucial para o objetivo em análise do presente projeto, utilizou-se a *collection all_players_hand* proveniente do *csv* mencionado na secção [3.2.1](#) com vários nomes de jogadores e com a respetiva “*Hand*” em sigla, ou seja, “R” para “*Right-Handed*”, “L” para “*Left-Handed*” e “U” para “*Unknown*”. Como esperado esta quantidade de jogadores sem esta variável, no sentido literal, coincide com o número de jogadores que se encontram apenas do lado do *oponent*.

Observou-se que a maior parte dos valores desta variável, ou continha valores nulos (*null*) ou vazio. Assim no final do tratamento desta variável uniu-se estes 2 valores para “U”, correspondendo a “*Unknown*” (juntamente com os “U” que surgiram dos *updates* feitos de seguida).

Verificou-se, assim, que existia uma grande maioria dos jogadores sem esta variável em *Players* que se encontravam em *all_players_hand*. Surpreendentemente a grande maioria desta correspondência a esta *collection* também continham a sigla “U” para esta variável, ou seja, também não era conhecida o tipo de *Hand* deste jogador. Este fator indica que se não é possível obter as *Hands* para estes jogadores nesta base de dados também não será em outras.

Ainda assim, foi feita a atualização destes jogadores em correspondência com esta *collection* e, de seguida, todos os campos que continham “*null*” e “” passaram para “U”, de modo a reduzir a redundância dos mesmos.

Posteriormente, atualizou-se esta variável de modo a associar os nomes correspondentes às suas “siglas”, como mencionado anteriormente. Deste feito, o número que elucidava a quantidade de jogadores com “*Hand*” antes do updates, conseguiu-se aumentar o número de jogadores com esta variável preenchida e que tivesse significado.

Por fim, atualizou-se a variável *Backhand* seguindo a mesma ideologia que o tratamento de “*Hand*”, no caso de “*Two-Handed Backhand*” passou a “T”, “*One-Handed Backhand*” passou a “O” e “*Unknown Backhand*” passou a “U”, juntamente com os “U” que foram adicionados no momento da criação desta variável.

3.2.8 Collection Confrontos

Na *collection* *atp2* foram observados registo duplicados, pelo que no momento da criação desta *collection* Confrontos, descartou-se estes registo. Isto foi verificado, pois na função *aggregate*, agrupou-se os registo por um conjunto de variáveis que refletem um registo único, tal como *GameRound*, *PlayerName*, *Oponent*, *Tournament*, *Start*, *End*, entre outras, reduzindo o número de registo inicial para 1306005 confrontos.

Na criação desta *collection* foi alterado o nome da variável *GameRank* para *OponentRank*, pelo motivo mencionado na secção [3.1](#).

Uma outra análise, permitiu que fossem encontrados registo em que a variável *Oponent* continha o valor “*bye*”. A grande maioria deste acontecimento acontece quando os oponentes designados como “*bye*”, que por definição segundo *TennisCompanion* é quando ocorre o avanço automático de um jogador de uma ronda para a próxima sem ter de competir com um adversário pois o número de jogadores não permite que todos os jogadores participem em uma partida das primeiras rondas. O “*WL*” não realçou a importância do seu tratamento, assumimos que estes valores foram atribuídos corretamente. No caso do valor “” nesta variável *WL* existiam registo que continham esta variável vazia e, verificou-se que, como era esperado, este valor diz respeito à grande maioria dos oponentes que se encontravam como “*bye*”. Para os restantes registo que não diziam respeito a *bye*, foram feitas análises sobre os mesmos e chegou-se à conclusão de que estes jogos à exceção de 1 dizem respeito a torneios que não foram concluídos, pelo que na criação de *collection* resultante desta, “*Game*”, criou-se uma variável

denominada de *Released* com valores (0,1), em que 0 significa que não foram realizados os jogos e 1 que foram. O único registo em que realmente aconteceu o jogo foi atualizado manualmente, logo o valor correspondente à variável criada *Released* será 1.

O tratamento da variável *Location* consistiu no que foi referenciado na secção [3.2.1](#), quando se referiu que a principal diferença entre o tratamento em *Born* na *collection Players* surgiu na forma da atribuição das siglas aos correspondentes países, que, neste caso, foi realizada de forma manual, sem o auxílio de um ficheiro csv. Contudo, de modo a reduzir a referida redundância foi utilizada a mesma *collection* que se usou para uniformizar as siglas na *collection Players*, como mencionado, designada de *codes_siglas*. Realçar que esta atribuição das siglas foi sempre tendo em consideração o código da *ISO-ALPHA-3*.

3.2.9 Collection Season

A motivação para a criação da mesma surgiu de para o mesmo torneio ter um único registo para este, ou seja, esta *collection* diz respeito às variáveis que caracterizam um torneio, sendo estas *Tournament, Start, End, Location, Prize* e *Ground*.

Relativamente aos valores que haviam sido atualizados para *Unknowns*, conforme descrito na secção [3.2.2](#), procedeu-se a uma pesquisa na *web*, tendo-se verificado que apenas foi possível identificar o tipo de pavimento em 3 dos 6 torneios em questão. Especificamente, no torneio “M15 Opava” o piso foi identificado como *Carpet*; no “M15 Cancun” foi atualizado para *Hard*; e no “M15 Antalya” o piso passou a ser *Clay*.

No que diz respeito à variável comum da localização (*Location*) com a *collection Confrontos*, o tratamento foi o mesmo que foi descrito na secção anterior.

Verificou-se ainda que existem 3 datas, em que o torneio começa no dia a seguir a acabar. Contudo, não se viu relevância no tratamento da mesma dado a utilização da mesma na análise em *Python*.

3.2.10 Collection Game

A criação desta *collection* surge com o intuito de remover os duplicados, não no sentido literal, mas no que diz respeito aos regtos que para o mesmo jogo aparecem 2 vezes na base de dados. Estes regtos representam o mesmo jogo, mas diferem apenas na forma como os dados são apresentados, consoante o jogador surge na variável *PlayerName* ou em *Opponent*. Para consolidar essa informação, foram criadas novas

variáveis que permitem reunir, em um único registo, os dados que anteriormente estavam divididos em dois. Esta abordagem visa, sobretudo, possibilitar a obtenção do *rank* de ambos os jogadores para um mesmo jogo.

Nesta *collection* foram feitas algumas alterações nomeadamente a variável *PlayerName* passou a *Player1*, *Oponent* para *Player2* e criou-se a variável *RankPlayer1* que obviamente diz respeito ao *rank* do *Player1* e *RankPlayer2* ao *rank* do *Player2*. Na inserção dos registos da mesma foi atribuído o nome do vencedor à variável *Winner* criada. Importante realçar que a variável *Score* é referente ao resultado visto do lado do *Player1* como já mencionado.

Primeiramente foram inseridos os 20 registos dos jogos não concluídos em torneios referidos anteriormente, contendo a variável *Released* igual a 0. De seguida, adicionou-se registos sem possibilidade de serem registos “duplicados”, que correspondem aos confrontos em que os jogadores só apareciam em *PlayerName* ou em *Oponent*, incluindo os “*bye*”, dado que neste comando, foram inseridos os registos dos confrontos em que o *WL* encontra-se como “W”. Na seguinte fase, destes últimos registos inseridos, foi criada uma variável “mapa”, *gameDuplicatedMap*, que concatena numa *string* as variáveis que identificam um jogo único nomeadamente *Player1*, *Player2*, *Tournament*, *Start*, *End* e *GameRound*.

Por último, no processo de verificação dos registos *WL* = “L”, ou seja, com o intuito de analisar a possibilidade de se tratarem de registos duplicados ou não, criou-se 2 variáveis, *bulkOpsInsert* e *bulkOpsUpdate*, que no caso de se tratar de um registo duplicado, encontrando-se na variável *gameDuplicatedMap* mas com os *Players* trocados, adiciona, assim, à variável *bulkOpsUpdate* a operação de atualizar o registo duplicado no sentido de inserir no registo já encontrado em *Game* o *Rank* do *Player1*. Dado o espaço computacional e por consequente o tempo que levaria a atualizar todos os registos que são duplicados, correspondendo a quase metade da base de dados, apenas foram registadas as atualizações necessárias para o país em análise neste projeto, que diz respeito à Áustria. Contudo, ainda assim, como estas atualizações são passíveis de demorarem algum tempo a correr, desenvolveu-se uma função que permitisse realizá-las em *batches* de 10 operações. Por outro lado, caso o registo não se encontre na variável “mapa” é adicionado à variável *bulkOpsInsert*, a operação de inserir o registo em questão.

Antes da realização em lote de cada uma destas variáveis de inserção e atualização verificou-se o tamanho das mesmas e, concluiu-se que existem 9400 registos duplicados

para os jogos realizados na Áustria, correspondendo ao número de duplicados da variável *bulkOpsUpdate*. Concluiu-se assim que dos 1306005, apenas 701228 são registos únicos.

3.3 MySql

Esta etapa foi crucial para o desenvolvimento do projeto, pois permitiu-nos construir um modelo relacional a partir dos dados previamente armazenados e processados em *MongoDB*, como descrito todo o processo anterior. Ao estruturar e dividir esses dados em tabelas, conseguimos atingir o mais alto nível de integridade e consistência, constituindo elementos fundamentais para a fiabilidade do nosso sistema.

Esta secção tem como foco assim, a descrição para a obtenção final deste modelo relacional após o processamento em *MongoDB*, a fim de ser possível extrair o *dataset* final para análise com todas as variáveis exigidas no enunciado. Importante voltar a frisar que todo o processo até o momento da obtenção do modelo relacional final criado encontra-se detalhado, de como devem ser utilizados todos os ficheiros para a obtenção do mesmo, no *road map* do ficheiro “*Step by Step.txt*”

Feitos os *exports* para *csv's* das *collections* *Tournaments*, *Players*, *Season*, *Game*, *Grounds* e *Rounds* a partir do Mongo, procedeu-se à criação de todas as tabelas necessárias à obtenção do referido modelo. Estas são: *backhand*, *country*, *game*, *ground*, *hand*, *player*, *round*, *season* e *tournaments*. Adicionalmente, foram criadas tabelas temporárias, nomeadamente, *tempgame*, *tempseason* e *temptournaments*, para facilitar o processo de importação e modelação dos dados.

No momento da criação destas tabelas foram adicionados manualmente os valores das tabelas referentes ao tipo de *Hand* e *BackHand*, já os dados das restantes tabelas foram introduzidos através dos *csv's* criados a partir do Mongo com a exceção do que diz respeito à atribuição da sigla ao respetivo país, que originou a tabela *country*. Csv este designado de “*All_Country_code.csv*”, que como já mencionado na secção [3.2.5](#) surge da necessidade de atribuir a sigla no código ISO-ALPHA-3 ao respetivo país, pelo que para este ficheiro foi necessário gerar a partir do “*codes_sigla.csv*”. As únicas tabelas que não necessitaram de uma tabela temporária sendo logo obtidas a partir dos dados provenientes dos *csv's* gerados foram as referentes a *Round*, *Ground* e *Player* que corresponde à respetiva ronda em que o jogo se decorreu no torneio, o tipo de pavimento do mesmo e, às características pessoais dos respetivos jogadores, respetivamente.

As tabelas temporárias referidas foram criadas de modo a ser possível serem geradas as finais, dado que existiam alguns problemas com a integridade dos dados nos respetivos *csv*'s. Problemas estes como registos duplicados, dado que no *Sql* as letras maiúsculas e minúsculas são vistas da mesma maneira, ao contrário do *Mongo*, a formatação das datas em que ocorreram os torneios e, por fim, a necessidade de obter os *ids* na tabela *Game* dos respetivos jogadores e da determinada temporada (*Season*) na qual ocorreu o torneio. Assim, após estes ajustes e feita a imputação dos dados para as tabelas definitivas destas mencionadas, foi possível obter assim, uma primeira versão do modelo relacional. É possível visualizar este modelo e, as respetivas chaves primárias e estrangeiras, a partir do [anexo 2](#) que contém já futuras tabelas que foram implementadas após as análises feitas em *Python*.

Estas análises feitas, posteriormente, em *Pyhton*, referenciadas mais adiante neste relatório, exigiram a criação das tabelas e variáveis antes ainda não mencionadas que se encontram no modelo relacional do anexo referido. Dizem respeito às tabelas *continent* e *continente-country* com a intuito de se introduzir a variável “continente” do nascimento do correspondente jogador em consideração, visto que a variável *Born* continha bastantes classes. Para além disto, como se pode verificar no respetivo anexo, existe uma variável referente à data de nascimento dos jogadores, que foi utilizada para criar a variável da idade dos jogadores no momento do jogo. Por fim, foi feita a imputação das idades dos jogadores encontrados no *csv* “*atp_players.csv*”. Este relato reflete as iterações realizadas entre as etapas do *CRISP-DM*.

3.4 Tratamento base em *Python*

Pela seleção sugerida pelo professor e por pesquisas realizadas sobre o tema de previsões para cenários relacionados a resultados dos jogos de ténis, foram selecionados do *SQL* um total de 19 variáveis para serem analisadas e seguidas de uma segunda seleção para modelação. Entre estas variáveis, 12 são relacionadas aos jogadores intervenientes da partida e 7 relacionadas a partida em si. Esta seleção foi feita com o objetivo de gerarmos um ficheiro *csv* com os dados que consideramos úteis para realizarmos o estudo solicitado e assim garantir que conseguíssemos obter os dados de forma mais acelerada isto porque a base de dados contém muitos registos o que faz com que os resultados dos *select*'s acabem sendo mais demorados.

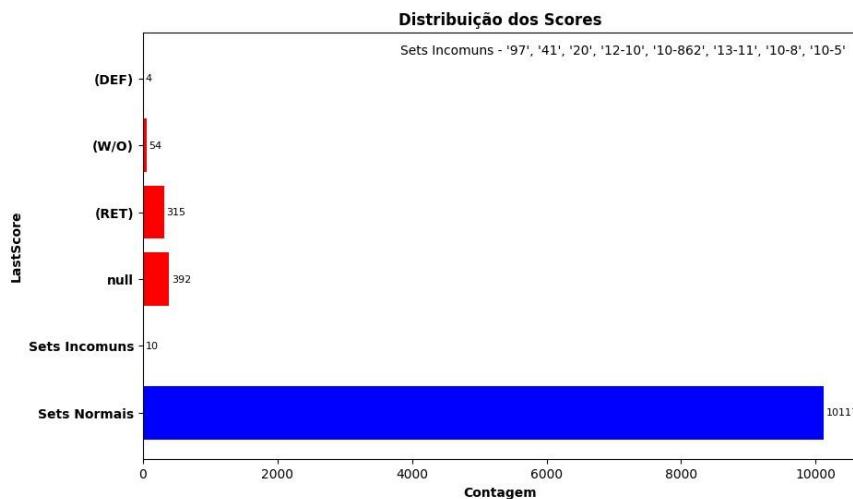
Para tal foi utilizada a linguagem de programação *Python* de modo a ser feita uma *query* com os dados necessários, transformar num *dataset* e de seguida converter para o tipo de ficheiro desejado.

De forma direta os dados obtidos não continham a variável alvo em estudo, pelo que foi decidido tratar da mesma antes de se definir qualquer outro tipo de avanço no que diz respeito aos dados ou de se obter o *csv*. Com os mesmos já estruturados em um *DataFrame* do *pandas* avançamos para o tratamento da variável alvo.

O processo de compreensão do negócio permitiu-nos identificar a variável pela qual podia ser extraída a variável alvo, que diz respeito à variável *Score*, esta que mostra os resultados obtidos ao fim de cada set. Nesta, cada set encontrava-se separado por um espaço, pelo que bastou-nos transformar cada um em elementos de uma lista e contar quantos elementos tinha a lista.

Com estes tratamentos realizados foi possível gerar o ficheiro *csv* e assim prosseguir para as etapas seguintes.

Inicialmente, criou-se a variável alvo definida “*Sets*”, criada a partir da variável “*Score*”. O processo consistiu em dividir a *string* de scores por espaços e contar o número de elementos resultantes, que corresponde ao número de sets jogados. Este processo exigiu a análise e exclusão de regístos inválidos ou ambíguos, que não fariam sentido para os casos em que os jogos que apresentavam “*DEF*”, “*RET*” ou “*W/O*”, no entanto, criou-se também a variável “*LastScore*” com o objetivo de estudar estes casos referidos. Esta variável identifica partidas com problemas no *score*. Os jogos marcados com *sets* inválidos foram verificados em detalhe e excluídos do *dataset* para assegurar consistência na variável alvo. Também se considerou a introdução do *tie-break* e o impacto nos dados. No gráfico seguinte é possível verificar estes jogos discutidos e, os que realçaram a sua importância de análise foram os que se encontram a vermelho no gráfico.



Destes, 4 jogos de “*DEF*” (desqualificação) foram removidos, 315 jogos de “*RET*” (desistência), dos quais 4 foram discutidos, mas também removidos, casos com ‘*W/O*’ (*walkover*) em que foram excluídos 54 jogos por ausência total de jogo, casos com ‘*Score*’ incorreto ou incoerente (ex: ‘10-862’, ‘20’) tratados manualmente e/ou removidos, como foi o caso dos registo com *score* vazio e o respetivo valor da variável *Sets* ser 0. No entanto nos registo que continham sets incomuns decidiu-se, então, remove-los.

Além disto, para completar o nosso *dataset*, foram utilizados ficheiros auxiliares, da *web*, em *csv* para enriquecer os dados, como “*all_players_hand.csv*” que desta vez serviu para a criação da variável da data de nascimento, a qual foi criada de modo a refletir a idade dos jogadores num determinado jogo, mas, mais uma vez, como esperado, não foi possível obter esta data para todos os jogadores.

Nas variáveis “*player1Born*” e “*player2Born*”, foram identificados valores chegados da base de dados em *SQL* com “\r” no final de cada campo e, como não era útil tê-los nesse estado decidiu-se remover.

3.4.1 Players

As variáveis relacionadas aos jogadores fornecem um retrato valioso sobre os jogadores de cada partida, permitindo, assim, facilitar na previsão do número de *sets* por jogo.

A identificação dos nomes dos jogadores é bastante completa e diversificada, o que permite uma análise aprofundada do desempenho individual ao longo do tempo. A ausência de valores nulos nessas colunas é um ponto positivo, pois evita a necessidade de tratamentos prévios e assegura que os dados estão bem estruturados nesse aspecto.

Entretanto, variáveis como a altura apresentaram desafios importantes. Apesar de a maioria dos valores parecer coerente, há um número bastante considerável de alturas registadas como zero (1832 observações), o que é evidentemente inválido. Esse tipo de inconsistência pode afetar os modelos preditivos a serem criados ou estatísticas descritivas que dependam dessa informação. Além disso, é possível verificar, como já referido, que existem diferentes unidades de medidas (como metros e *foot*), o que exige uma padronização cuidadosa para evitar distorções.

As variáveis que descrevem a lateralidade dos jogadores (como a mão dominante e o tipo de *backhand*) são úteis para entender estilos de jogo e padrões táticos. No entanto, a presença de valores registrados como “*Unknown*” em grande volume para ambas as variáveis *hand* e *backhand* compromete a sua utilidade. Em especial, quando se sabe que essas informações deveriam estar sempre disponíveis, já que se referem a características públicas e estáveis dos atletas.

Quanto à nacionalidade, há tanto potencial como desafios. Saber de onde os atletas vêm é interessante, especialmente num contexto em que todos os jogos ocorreram no mesmo país (Áustria), podendo influenciar nos desempenhos. No entanto, há um desequilíbrio significativo na representação de certos países, o que faz sentido dado que todos os campeonatos foram realizados num país atribuído, existindo assim sub-representação de países, com apenas 1 observação para alguns deles.

Já no caso das idades, a maioria dos jogadores encontra-se entre os 20 anos, embora a distribuição geral seja plausível, há registo com idades negativas, o que sugere erros na codificação das datas de nascimento ou dos eventos, uma vez que esta foi criada para cada jogo a diferença entre a data do jogo com a data de nascimento. Além disso, ao contrário de outras variáveis, os valores nulos não foram codificados como “*Unknown*”, tendo sido totalizados 600 valores nulos em cada uma das idades dos jogadores.

Por fim, a classificação (*rank*) dos jogadores traz consigo tanto valor analítico quanto complexidade. É uma variável essencial para compreender a diferença de nível entre os adversários, mas a ocorrência frequente de valores zerados levanta dúvidas sobre a sua fiabilidade em certos registo. Como se trata de uma métrica normalmente bem definida e amplamente divulgada, é importante verificar se o valor zero corresponde à ausência de *rank* ou a um erro de registo, dado que a maior parte dos jogos que o *dataset* apresenta são jogos de primeiras fases dos torneios não tendo sido atribuídos *ranks* aos jogadores, os *ranks* vão alterando à medida que as partidas dos torneios vão ocorrendo.

3.4.2 Games

As variáveis associadas aos jogos oferecem uma base sólida para compreender o contexto em que os mesmos ocorreram, sendo assim a base para encontrar padrões ou possibilidades para prever jogos mais disputados.

A variável que indica a fase do torneio (*GameRound*) apresenta uma diversidade adequada, havendo sempre, como referido anteriormente, mais rondas de fases iniciais de torneios e menos rondas de fases finais, como o “normal” dado que se trata de registo sobre jogos com eliminatórias à medida que os jogadores avançam para outras fases. Assim, este comportamento da variável reflete as diferentes etapas de uma competição eliminatória. Embora algumas fases estejam pouco representadas, isso é esperado, uma vez que há naturalmente menos jogos nas fases finais, sendo coerente com a estrutura típica dos torneios.

A variável que identifica o torneio (*tournament*) mostra uma ampla variedade de competições, o que pode enriquecer bastante a análise, permitindo verificar se o local ou o torneio específico influencia o desempenho dos jogadores. No entanto, a sub-representação de alguns torneios pode limitar análises comparativas mais profundas nesses casos. Mesmo assim, o valor informativo continua elevado, especialmente para os torneios com mais registo.

As datas de início e fim dos jogos (*Start & End*) fornecem uma boa cobertura temporal, permitindo estudar a evolução dos torneios ao longo do tempo ou até detetar padrões sazonais. A diversidade de datas sugere um histórico robusto, o que é valioso para análises temporais.

A variável que indica o tipo de piso (*Ground*) é particularmente relevante, já que o desempenho dos jogadores pode variar muito consoante o tipo de superfície. A predominância dos jogos em *Clay* é esperada, tendo em conta o contexto do conjunto de dados. A presença de diferentes tipos de piso, mesmo que em menor quantidade, permite explorar variações táticas e de performance dos jogadores associadas a cada superfície.

Por fim, a variável relacionada ao valor dos prémios monetários (*Prize*) tem potencial para ser uma métrica interessante de prestígio dos torneios ou de motivação dos jogadores. No entanto, não se verifica existência de diferentes unidades monetárias utilizadas, apenas o dólar, pois só temos em estudo um único país.

3.4.3 Tratamento Aprofundado

Com o tratamento inicial realizado em cada variável, decidiu-se transformar os dados de modo que seja possível fazer análises gráficas que sejam necessárias, seleções das variáveis admissíveis para os estudos a serem realizados e possível remoção de registo que não vão de acordo com a ideia da estrutura dos dados atuais.

Em *Python*, foi identificado que existem alguns jogadores em que a respetiva altura seria 0, deste modo, considerou-se colocar nessas alturas o respetivo valor da média de todas as alturas obtidas até ao momento, no entanto, voltou-se ao *SQL* para realizar a imputação das alturas dos jogadores que se encontravam no *atp_players.csv*, contudo após essa imputação ainda se verificou um número considerável de valores nulos.

Quanto as variáveis relativas à mão predominantes e à utilização de uma ou duas mãos no lado oposto ao da mesma de cada jogador, decidiu-se eliminar os registo com *Ambidextrous*, para a variável *Hand*. No entanto, para as duas variáveis, a expressão nos registo “*Handed*” foi eliminada e estudou-se qual mão seria a mais provável de ser a predominante de cada jogador com o campo da variável *Hand* preenchido com “*Unknown*”.

Tendo em conta o elevado número de países de nascimento, criou-se uma variável relativa aos continentes, de modo a reduzir a dimensionalidade dos dados, colocando em conta que com esta decisão estamos propensos a perder informações. Assim, estudou-se como estará a distribuição por continentes, sendo o mais esperado a Europa com predominância, devido ao torneio ser realizado num país europeu e a maior parte dos países de nascimento serem de jogadores austríacos e a influência de cada continente para o número de sets. Dado esta maioria absoluta, percebe-se que esta classe está sobre representada, causando problemas em análises futuras.

Inicialmente foram verificadas idades negativas, como referido anteriormente, com isto verificou-se a quantidade de valores inferiores a 0 e como são poucos registo, 3, no total, decidiu-se eliminar essas observações. Além disto, os registo em que tanto a idade como a altura dos jogadores não são conhecidas vão acabar por ser eliminados quando criado o *dataset* relativo aos registo sem alturas nulas de pelo menos um dos jogadores.

Relativamente à variável *rank* de cada jogador, ao analisar os dados identificou-se que existem múltiplos registo onde o valor do *rank* está igual a zero para um ou ambos

os jogadores. Dado que o *rank* é uma variável essencial para avaliar o nível de desempenho e o equilíbrio entre os adversários, a presença de valores nulos ou inválidos (como o zero) compromete a fiabilidade desta informação. Inicialmente, foi avaliado o impacto de diferentes abordagens, como a remoção de registo com *rank* nulo ou a imputação da média. Optou-se, então, por eliminar todos os jogos em que ambos os jogadores apresentavam *rank* igual a zero. Esta decisão foi tomada para garantir que, em todos os confrontos considerados, pelo menos um dos jogadores possua um *rank* válido. Adicionalmente, começou-se a avaliar a qualidade de outras variáveis, como as alturas dos jogadores, com o intuito de eventualmente aplicar filtros combinados, garantindo ainda mais consistência nos dados finais.

Na variável relativa aos torneios foram encontrados torneios que não são originários da Áustria, mas competições entre a Áustria e alguns dos diferentes países e que tiveram uma das suas rondas realizadas na Áustria, como este torneio apresentava poucos registo foi tomada a decisão de os apagar. Além do mais, dos torneios restantes verificaram-se que 2 deles não demonstravam pertencer ao país em estudo, uma vez que os nomes eram “*Italy F25*” e “*Italy F12*”, que correspondem à Itália e, ainda se encontrou um torneio com o respetivo nome escrito de forma incorreta, em que o “*M25 Vagau*”, mais tarde, se alterou para “*M25 Vogau*”.

Através das datas de início e fim de cada torneio foi criada uma variável “*Days*”, que corresponde à duração em dias de cada campeonato os valores dessa variável correspondiam a 0, 5, 6, 13 e 20, como é possível verificar, existiam torneio realizados em 0 dias o que não faz sentido, tendo sido pesquisado que a duração dos torneios correspondia a mais 1 dia do que a diferença das datas, sendo assim, os valores 1, 6, 7, 14 e 21. Durante a análise desta variável, observou-se que o valor correspondente a 6, mais tarde alterado para 7, se ilustrava sobre representado na amostra, havendo uma elevada quantidade deste valor na mesma.

A partir da data criou-se uma variável relativa às estações do ano (*Season*), que foi estudada e se verificou que a estação do ano correspondente ao verão encontrava-se com valores demasiado elevados comparativamente a outras estações.

A variável *prize* foi transformada em valor numérico uma vez que apresentava o símbolo do dólar inserido na mesma, tendo este sido removido. Dos registo restantes do tratamento das outras variáveis encontraram-se alguns torneios sem o respetivo prémio. Para a resolução deste problema foram discutidas soluções, como substituir pela média

ou excluí-los de modo a não ser necessário excluir mais valores dado que os que serão removidos irão reduzir muito os nossos dados, existindo a possibilidade de já se estar em risco de ter falta de dados.

Após a análise da variável *Sets*, verificou-se que existiam jogos com 1, 2, 3, 4 e 5 sets, os jogos com 1 foram eliminados, uma vez que não foram acabados. De modo a facilitar a exploração criou-se a variável *BestOf*, esta indicou-nos que a quantidade de jogos realizados à melhor de 5 sets e à melhor de 3. O processo de criação desta variável passou por se entender que os jogos com 2 sets vão para à melhor de 3 e os jogos com 4 e 5 vão para a melhor de 5. Desta maneira, constatou-se o ponto mais fulcral, que foi nos 3 sets, dado que ter-se-ia de fazer a separação para as quais os jogos que eram supostos ir a um máximo de 5 ou de 3 sets. Assim, os jogos que apresentavam 3 vitórias para um determinado jogador foram classificados como à melhor de 5 e o caso contrário foi definido que seriam jogos à melhor de 3. Por fim, após a criação de *BestOf*, observou-se que da amostra total, os jogos que seriam à melhor de 5 eram menos de 1%, tendo sido acordado prever se os jogos iriam até aos 2 ou 3 sets.

4. DATA UNDERSTANDING

4.1 Descrição das variáveis originais

A presente fase de *Data Understanding* teve como objetivo compreender a fundo a estrutura e qualidade dos dados fornecidos, os quais dizem respeito a jogos de disputados entre 28 de março de 1973 e 14 de fevereiro de 2022. O conjunto original, contendo 10.894 registos, foi recolhido a partir do site oficial da organização, disponibilizado em formato *JSON* e posteriormente convertido para um modelo relacional em *SQL* após tratamento em *MongoDB*.

Cada linha representa um jogo entre dois jogadores, com variáveis relativas às características dos atletas, ao torneio e ao resultado da partida. As variáveis originais incluem, nomes dos jogadores, alturas, mão dominante, tipo de *backhand*, país de nascimento, *rank*, tipo de torneio, superfície do campo, ronda do torneio, prémio monetário, resultado da partida (*Score*), bem como as datas de início e fim do torneio (*start* e *end*).

Nome da variável	Tipo de dados	Dados estatísticos
<i>player1Name</i>	Categórica nominal	1596 nomes únicos
<i>player2Name</i>	Categórica nominal	2554 nomes únicos
<i>player1Height</i>	Numérica contínua	Min: 0, Max: 510; Média: ~ 152 cm
<i>player2Height</i>	Numérica contínua	Min: 0, Max: 510; Média: ~ 126 cm
<i>player1Hand</i>	Categórica nominal	Right: 7494; Left: 1489; Unknown: 1134
<i>player2Hand</i>	Categórica nominal	Right: 7185; Left: 1219; Unknown: 1712
<i>player1Backhand</i>	Categórica nominal	Two: 3645; One: 1586; Unknown: 4886
<i>player2Backhand</i>	Categórica nominal	Two: 3012; One: 1175; Unknown: 5930
<i>player1Born</i>	Categórica nominal	81 países únicos
<i>player2Born</i>	Categórica nominal	92 países únicos

<i>player1Rank</i>	Numérica Ordinal	Min: 0, Max: 2233; Média: ~ 308
<i>player2Rank</i>	Numérica Ordinal	Min: 0, Max: 2233; Média: ~ 417
<i>GameRound</i>	Categórica nominal	10 classes; mais frequente: Round of 32
<i>tournament</i>	Categórica nominal	78 torneios distintos
<i>start</i>	Data	350 datas únicas
<i>end</i>	Data	350 datas únicas
<i>ground</i>	Categórica nominal	Clay: 7200; Hard: 1583; Carpet: 1244; Grass: 90
<i>prize</i>	Numérica contínua	68 valores distintos
<i>Score</i>	Texto estruturado	1003 valores distintos

4.2 Análises iniciais

Durante a análise inicial, foram detetados diversos problemas de qualidade dos dados. Em particular, 392 registos apresentavam valores nulos na variável *Score*, o que impossibilita a extração do número de sets. Para além disso, foram encontrados 315 jogos com o valor "*RET*" (*Retired*), indicando desistência durante o jogo; 54 jogos marcados como "*W/O*" (*Walkover*), nos quais a partida não chegou sequer a iniciar-se; e 4 casos de "*DEF*" (*Default*), indicando desqualificação de um jogador. Adicionalmente, identificaram-se valores anómicos na variável *Score*, como "10-862", "20" ou "41", incompatíveis com a realidade e que exigiram a sua remoção.

Outros problemas foram identificados, nomeadamente alturas com valor zero, que não são viáveis e indicam ausência de dados. Estas alturas foram mais tarde imputadas com a média ou removidas, consoante o caso. Verificaram-se ainda *rank* nulos ou com valores bastante elevados e fora do intervalo de maior concentração dos dados (até 2233), o que exigiu tratamento adicional. A variável relativa ao torneio apresentou elevada cardinalidade, com 78 categorias distintas, o que motivou, numa fase posterior, a decisão de restringir a análise a torneios realizados na Áustria.

4.3 Criação e enriquecimento de variáveis

De modo a permitir uma compreensão mais aprofundada dos dados e identificar relações relevantes com a variável alvo, foram criadas diversas variáveis auxiliares. A variável *Sets*, criada a partir do conteúdo da coluna *Score*, representa o número total de sets jogados num jogo e constitui a variável alvo deste projeto. A partir da mesma variável *Score* foi ainda extraída a variável *BestOf*, que indica se o jogo foi disputado à melhor de três ou de cinco sets. Foi também criada a variável *LastScore*, com o objetivo de identificar rapidamente casos com anomalias, como resultados incompletos, desistências ou situações invulgares mencionadas anteriormente.

Entre outras variáveis criadas internamente, destaca-se a variável *days*, que representa a duração de cada torneio em dias, sendo calculada com base nas datas de início e fim (*start* e *end*). Com base na data de início foi também derivada a variável *Season*, permitindo categorizar os jogos por estação do ano (Primavera, Verão, Outono e Inverno).

Por outro lado, houve variáveis cuja criação exigiu recorrer novamente ao *SQL* e à integração de fontes externas. A criação da variável *Continent* para ambos os jogadores exigiu a conversão do local de nascimento em códigos continentais, com o objetivo de reduzir a elevada cardinalidade presente na variável *Born*. Da mesma forma, foram revistas e completadas as variáveis *Hand* e *Backhand* com base em ficheiros adicionais, de modo a colmatar falhas ou inconsistências nos registos. Foi também através do *SQL* que se obteve a data de nascimento dos jogadores, necessária para o cálculo da variável *Age*. Além disso, procedeu-se à imputação parcial de alturas que se encontravam nulas, tendo algumas sido preenchidas com valores recolhidos via *SQL* e outras mantidas como nulas pela impossibilidade de encontrar os valores reais das mesmas.

Adicionalmente, foram criadas variáveis para representar diferenças absolutas entre os dois jogadores ao nível da altura, idade e rank (*HeightDifference*, *AgeDifference*, *RankDifference*). Estas variáveis foram fundamentais para permitir análises centradas no confronto entre os dois jogadores em cada partida, em vez de tratar separadamente *player1* e *player2*.

Entre as ideias criativas discutidas no grupo, foi também proposta a criação de uma variável relacionada com o peso dos jogadores, a partir da qual seria possível futuramente calcular o IMC e, assim, criar indicadores comparativos relacionados com a

robustez física e o seu possível impacto na performance dos atletas. No entanto, como esta variável não é constante ao longo do tempo, ou seja, a modificação do peso não é constante, podendo haver alterações semanais, mensais ou até mesmo anuais, não seria possível ter a mesma atenção que a idade, isto é, para cada jogo pode haver uma diferença de peso, seja devido à massa muscular, ou até mesmo ao aumento de gordura. Assim, tendo em conta este fator de inconsistência decidiu-se não utilizar estas variáveis (Peso e IMC).

4.4 Análise Exploratória

Após a criação das variáveis, os dados foram divididos em dois conjuntos: um contendo valores nulos na variável *Height* (*withNull.csv*), utilizado para efeitos comparativos e de exploração, e um segundo conjunto limpo (*withoutNull.csv*), com todos os registo prontos para a fase de análise e modelação. A utilização desta versão limpa permitiu avançar para análises mais avançadas, assegurando que os resultados obtidos se baseavam apenas em observações completas.

Com base neste conjunto de dados limpo, foi realizada uma análise exploratória aprofundada com o objetivo de investigar o comportamento das variáveis explicativas em relação à variável alvo *Sets*. Foram realizadas análises univariadas e bivariadas, bem como análises visuais e estatísticas complementares. Tendo em conta que apenas 0,7% dos jogos realizados eram à melhor de cinco *sets* ([Anexo 3](#)), foi decidido restringir a análise aos jogos à melhor de três *sets*, de forma a garantir maior homogeneidade nos dados. Entre estes, observou-se que 66,6% dos jogos terminaram com dois *sets* e 33,4% com três *sets* ([Anexo 4](#)), sendo estas as distribuições mais representativas.

A maioria dos jogos foi disputada em pisos de terra batida (*Clay*), com esta superfície a demonstrar predominância como se observa no gráfico das distribuições da variável *ground* ([Anexo 5](#)). A variável *Season* revelou que os jogos ocorreram maioritariamente no verão ([Anexo 6](#)), refletindo o calendário de torneios da ATP.

Através da análise das variáveis diferenciais (*RankDifference*, *AgeDifference*, *HeightDifference*), foi possível observar que jogos com maior desequilíbrio entre jogadores tendem a terminar em menos sets ([Anexos 7, 8 e 9](#)), reforçando a ideia de que maior competitividade está associada a maior duração. Para além dessa tendência geral, os *boxplots* revelaram ainda a presença de *outliers* significativos, sobretudo em *RankDifference*, indicando que existem jogos com discrepâncias muito acentuadas entre

os adversários, como por exemplo confrontos entre jogadores do *top 10* e outros com *rank* superiores a 1000. Embora representem uma minoria, estes casos extremos devem ser tidos em conta, pois podem influenciar o desempenho dos modelos e justificar abordagens específicas no tratamento dos dados.

Foram também analisados as combinações entre as mãos dominantes e os tipos de *backhand* dos jogadores, tendo sido criadas novas variáveis combinadas (player1 + player2) com o intuito de capturar traços técnicos que pudessem influenciar a duração dos jogos, partindo da hipótese de que o estilo de jogo poderia ter impacto no número de sets. No entanto, essa relação não se confirmou, uma vez que as distribuições foram semelhantes tanto para jogos com dois como com três sets (Anexos [10](#) e [11](#)).

Por outro lado, foi ainda analisada a frequência dos confrontos entre combinações continentais dos jogadores. Verificou-se que os confrontos entre jogadores europeus (*Europe - Europe*) foram largamente predominantes ([Anexo 12](#)), o que indica um desequilíbrio na representatividade geográfica da base de dados.

No que respeita à relação estatística entre variáveis, foram conduzidas análises de correlação com o coeficiente *V de Cramér* (para variáveis categóricas) e coeficiente *ETA* (para variáveis numéricas). Contudo, não se observou qualquer correlação significativa entre as variáveis explicativas analisadas e a variável alvo *Sets* ([Anexo 13](#) e [14](#)), o que reforça a complexidade do problema preditivo e a potencial necessidade de explorar novas fontes de informação ou variáveis transformadas.

Esta etapa permitiu garantir uma visão clara e estruturada da qualidade, abrangência e limitações dos dados, bem como enriquecer o conjunto original com variáveis relevantes para a análise. As observações obtidas orientaram decisões fundamentais para a preparação dos dados e definiram a base analítica sobre a qual será construído o modelo preditivo.

5. DATA PREPARATION

Inicialmente, foi criada uma cópia do *dataset*, de modo a fazer alterações provisórias e caso se verificasse que seria adequado então refazê-las no *dataset* original. Este tratamento foi realizado no final do *NoteBook* “ExploratoryDataAnalisisNoNulls.ipynb”

Assim, para além das restantes colunas vindas da base de dados processada, foram adicionadas outras de modo a facilitar as fases seguintes, tendo sido a coluna *Hands* que corresponde às mãos predominantes dos jogadores de cada jogo, *Backhands*, se agarra na raquete com 1 ou 2 mãos no lado da mão contrária à predominante, *Rank* e *Ages*, ambos se referem à diferença dos valores da classe respetiva classe de cada jogador num determinado confronto.

Ao serem analisadas as variáveis referidas acima, estas apresentaram alguns problemas. Nas variáveis *player1Age* e *player2Age* foram identificados 514valores nulos, tendo sido acordada a remoção destes, ficando com um total de 4692 registos.

Além disso, na variável *Backhands* existia uma elevada quantidade de valores desconhecidos, “*Unknown*”, para resolver este problema os registos nesta variável que apresentam “*One-Unknown*” foram atualizados para “*One*” por ser a classe modal para as observações com “*One*”, já os registos que continham apenas “*Unknown*”, foram alterados para “*Two*”, uma vez que esta é a classe modal geral, ou seja, a categoria mais comum no conjunto de dados para a variável *Backhands*, introduzindo, assim, o menor viés possível na distribuição.

A variável *Hands*, continha uma subrepresentação da classe “*Left*”, tendo sido testado a união da respetiva classe com a “*Right*”, intitulando, esse vínculo de “*Same*”, contudo, no contexto real é muito mais habitual existirem jogadores destros do que canhotos o que leva a esta subrepresentação, não só neste desporto, mas em todos, desta forma, foi decidido manter ambas as classes separadas.

Para diminuir os *outliers* na variável *Rank*, aplicou-se o método baseado no Intervalo Interquartil (IQR), o *Turkey Method* uma técnica estatística para deteção de valores extremos. O IQR é calculado como a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1), ou seja, $IQR = Q3 - Q1$. A partir deste intervalo, define-se um limite superior para deteção de *outliers* como $Q3 + 1.5 * IQR$. Com este cálculo, todos os *outliers* passaram para 434 para esta variável.

Da mesma forma que o feito para a variável *Rank* baseado no intervalo interquartil (IQR), foram identificados os valores extremos na variável *Ages*. O limite superior resultou num valor de corte de 12 anos. Todos os registos com valores superiores a este limite foram considerados *outliers*. Deste feito, optou-se pelos *outliers* serem substituídos pelo próprio valor do limite superior (12).

5.1 Seleção de Variáveis e Motivação baseada no contexto do Ténis

A variável “Sets” foi selecionada como variável alvo por refletir a duração e competitividade de um jogo de ténis. Assim, para conseguir prever o número de sets realizados num determinado jogo, as seguintes variáveis foram selecionadas com base na análise estatística e conhecimento do domínio do ténis:

- *Hand*, uma vez que numa visão geral é autoexplicativo porque é que os tipos de hand vão influenciar no número de sets, mas para além disso confrontos entre jogadores com diferentes tipos de hand vão ter um jogo mais tático e competitivo potencialmente indo para 3 sets

- *Backhand*, já que o *backhand* a duas mãos oferece mais estabilidade e controlo permitindo trocas de bola (rallies) mais longas favorecendo jogos equilibrados e com muitos pontos disputados aumentando a chance de o jogo ir para 3 sets. Além disso, o backhand a uma mão apesar de ser mais elegante é uma técnica muita mais complicada de executar existindo maior margem de erro, o que pode desequilibrar o jogo rapidamente e levar a vitórias mais rápidas (2 sets). E, por fim, quando um jogo ocorre em que a mão predominante dos jogadores é a oposta um do outro, à um choque de estilos em que um apresenta uma técnica mais arriscada e outro opta pela defensiva, cria um confronto mais equilibrado (3 sets) visto que existe um ajuste constante de ambos os jogadores.

- *Rank*, devido ao *boxplot* da relação entre esta variável e a variável alvo, jogos com diferenças de rank mais acentuadas, ou seja, tendem a ser jogos mais curtos e mais rápidos a serem decididos indo para 2 sets.

- *Ground* que para as diferentes classes, terra batida (*clay*) apresenta um piso mais lento e um ressalto alto o que corresponde a pontos longos e trocas demoradas tendencialmente sendo jogos resolvidos em 3 sets, relva (*grass*) apresenta um piso mais rápido com baixo ressalto indicando pontos curtos favorecendo jogadores com bom serviço correspondendo a jogos decididos em 2 sets, *hard* apresenta um piso equilibrado o que acaba por depender mais da capacidade dos jogadores para decidir o jogo. Já o

carpet é um piso muito rápido, com ressalto baixo e pouca fricção, favorecendo ainda mais o jogo ofensivo e encurtando os pontos, com tendência para jogos curtos.

- *Gameround*, uma vez que nas rondas iniciais, como os jogos são sorteados, são tendencialmente decididas em 2 sets visto que são confrontos entre jogadores com grandes diferenças de rank, já jogos em fases finais tendem a ser mais demorados indo para 3 sets.

- *Prize*, logicamente quando um confronto apresenta um prémio superior, vai ser um jogo mais competitivo logo mais longo do que um jogo com um prémio inferior.

- *Age*, já que esta variável tem uma influência direta na capacidade física, recuperação muscular, e resistência cardiorrespiratória dos jogadores.

Assim, com os dados tratados e as variáveis selecionadas, o *DataFrame* foi convertido para *CSV*, as variáveis categóricas foram convertidas em *dummies*, variáveis contínuas normalizadas e o *dataset* foi dividido em conjunto de treino e teste. Além disso, foi aplicada a técnica *SMOTE* para balanceamento da variável alvo, a validação cruzada (*Cross Validation*) foi utilizada para garantir a robustez do modelo e o *Standard Scaler* transforma os dados para que tenham média zero e variância unitária, facilitando o desempenho de algoritmos de machine learning.

5.2 SMOTE Family

Já fora referido anteriormente que a variável alvo de estudo se encontra ligeiramente desbalanceada, o custo de classificar erroneamente um exemplo anômalo (jogos terminados com 3 sets) como normal (jogos terminados com 2 sets) é geralmente maior do que o oposto, para garantir a qualidade do modelo é importante que se tenha uma porção similar de ambas as classes, para tal foi utilizada uma técnica de *Over-sampling* chamada *SMOTE* (*Synthetic Minority Over-sampling Technique*) esta que cria exemplos sintéticos ao longo dos segmentos de linha entre vizinhos mais próximos da classe minoritária. Ao contrário das técnicas de *over-sampling* simples que levam a regiões de decisão muito específicas, resultando em *overfitting*, o *SMOTE* resulta em regiões de decisão mais gerais, melhorando a generalização do classificador.

O *SMOTE* cria os exemplos sintéticos com base nos vizinhos mais próximos (para este trabalho foram utilizados os 5 vizinhos mais próximos), faz uso de medidas de distâncias para a sua aplicação, o que cria algumas implicações sobre os dados a serem

utilizados na sua aplicação. Os dados precisam ser todos numéricos e estar numa mesma escala para que as distâncias sejam reais. As variáveis selecionadas para o modelo não cumprem de todo com os requisitos impostos pelo *SMOTE* pelo que foram necessárias aplicar algumas técnicas para que os dados se adequassem a esta abordagem e estes serão abordados nos próximos pontos.

5.3 *Dummy*

Nos dados selecionados havia 4 variáveis que não se encontravam no formato numérico, o que, como já referido no ponto anterior, seria um problema para a técnica de balanceamento escolhida dado que ela utiliza medidas de distâncias. De modo a ultrapassar este problema foram criadas variáveis *dummy*.

Variáveis *dummy* são variáveis binárias criadas para representar uma variável com duas ou mais categorias. Para cada uma das variáveis foram criadas $n - 1$ variáveis *dummy*, sendo n o número de classes que uma variável categórica tem, esta aplicação é padrão de modo a evitar a existência de multicolinearidade dos dados. A classe excluída é refletida quando todas as outras classes não se verificam.

5.4 *Standard Scaler*

Tratadas das variáveis categóricas ainda existiam as variáveis numéricas que se encontravam em escalas totalmente diferentes, uma referindo a moedas, outras referindo a diferenças de idades e alturas e com o tratamento anterior, surgiram também variáveis numéricas binárias, entre 0 e 1, e elas também precisavam fazer parte das contas.

A escala das novas variáveis *dummy* acabaram forçando a escolha da técnica de estandardização da escala das variáveis selecionadas para a modelação. Existem várias técnicas de estandardização das escalas, as mais conhecidas são a *z-score* (eliminar a média e dividir pelo desvio padrão) e a normalização zero-um (que coloca os dados numa escala de 0 a 1), por esta pequena explicação já deve ser fácil acertar qual foi a técnica utilizada. Foi utilizada a normalização zero-um, sobre as variáveis numéricas não *dummy* e assim todas as variáveis passaram a estar na mesma escala.

Nestas condições os dados já se encontravam prontos para as fases a seguir.

5.5 Técnicas de Avaliação dos Modelos

De modo a se garantir segurança na escolha do melhor modelo a ser utilizado foram utilizadas duas técnicas de avaliação de modelos bastante populares, divisão em treino e teste e a validação cruzada, no final foram feitas comparações dos desempenhos de cada uma delas para a decisão. Mas, no que consiste cada uma delas e como foram implementadas?

5.5.1 Divisão em treino e teste

É o método mais simples de ser implementado, é feita a divisão do conjunto de dados em duas partes, onde um conjunto serve para treinar o modelo (chamado conjunto de treino) e o outro conjunto serve para testar o modelo (chamado conjunto de teste), esta técnica procura avaliar como o modelo construído lida com o conjunto utilizado para construção e como ele poderá lidar com um conjunto de dados totalmente novo. De modo a garantir que os modelos capturem o máximo de informação possível dos dados estas divisões são feitas geralmente 70%-80% para o conjunto de treinamento e 20%-30% para o conjunto de teste.

Para o presente estudo foi feita a divisão 70/30, visto que esta obteve melhores resultados face a outra divisão, também popular.

5.5.2 Validação cruzada (*Cross Validation*)

Esta técnica, mais complexa do que a anterior, tem todo um conjunto de implementações, a utilizada neste trabalho foi a *K-fold*, que consiste em dividir os dados em K subconjuntos, treinar o modelo K vezes, cada vez usando um subconjunto diferente como conjunto de teste e o restante como treinamento. Em geral, ela acaba sendo uma técnica que fornece uma estimativa mais confiável do desempenho do modelo dado que utiliza o conjunto de completo para o treinamento.

Para a avaliação final do modelo treinado sobre esta técnica é considerado o desempenho médio dos subconjuntos utilizados para teste ao longo dos treinamentos do modelo. A implementação dele baseia-se mais na escolha do K , que para este trabalho foi escolhido *10-folds*, para esta escolha não foram feitas experiências, apenas utilizou-se o padrão.

A técnica de balanceamento mencionada anteriormente foi sempre utilizada sobre os conjuntos utilizados para o treinamento dos modelos e nunca para o teste, isto porque só se espera obter representatividade de ambas as classes no treinamento dos dados.

6. MODELING

Com os dados devidamente preparados e a variável alvo definida na fase anterior, avançou-se finalmente para a fase de modelação preditiva. O objetivo desta fase consistiu em desenvolver modelos de machine learning capazes de prever, com elevada precisão e rigor, o número total de sets disputados num jogo de ténis profissional, considerando exclusivamente jogos disputados no formato *best of 3*. Neste contexto, a variável alvo é binária e assume dois valores possíveis: 2 – jogos resolvidos em dois sets; ou 3 – jogos resolvidos em 3 sets. Trata-se, portanto, de um problema no qual se pretende antecipar, com base nas características dos jogadores e do contexto do jogo, se a partida será mais curta (2 sets) ou mais equilibrada e demorada (3 sets).

Para garantir a robustez e generalização dos modelos, como já foi referido numa fase anterior, os dados foram divididos em conjunto de treino (70%) e teste (30%) de forma estratificada, assegurando a manutenção proporcional da distribuição das classes da variável alvo.

Foram testados quatro algoritmos de classificação binária, selecionados com base na sua robustez, capacidade preditiva e interpretabilidade. Foram eles: *Decision Tree*, *Random Forest*, *XGBoost (Extreme Gradient Boosting)* e *SVM (Support Vector Machine)*.

A *Decision Tree* é um dos modelos mais intuitivos e interpretáveis da aprendizagem automática. Baseia-se na criação de regras sequenciais do tipo se-então que dividem recursivamente o espaço de decisão com base nos valores dos preditores. Cada nó representa uma decisão sobre um atributo, e cada ramo define uma saída possível. A árvore termina em folhas, que correspondem às previsões finais. Este modelo é particularmente útil quando se pretende compreender a lógica da decisão por detrás das previsões, já que permite visualizar os caminhos tomados com base em combinações de atributos. A escolha de incluir este modelo deve-se à sua capacidade de servir como base comparativa e à sua utilidade pedagógica para perceber como diferentes atributos influenciam o resultado.

O *Random Forest* é uma extensão do modelo de árvore, baseada no conceito de *ensemble learning*. Em vez de criar uma única árvore de decisão, constrói-se uma floresta composta por múltiplas árvores independentes, treinadas sobre subconjuntos aleatórios dos dados e dos preditores. A previsão final é obtida por agregação (tipicamente por

votação). Esta abordagem permite reduzir a variância das previsões e aumentar a robustez face a *outliers* e ruído nos dados. Num contexto como o nosso — com variáveis fracas e desbalanceamento —, o *Random Forest* pode captar relações não lineares entre preditores e evitar sobre ajuste, ao contrário de uma árvore isolada. Foi incluído por ser um modelo robusto, amplamente validado em problemas de classificação complexos.

O *XGBoost* é uma implementação otimizada do algoritmo *Gradient Boosting*, um método que constrói o modelo de forma sequencial, acumulando pequenas correções feitas por árvores fracas, aos erros cometidos pelas anteriores. Cada nova árvore é treinada para corrigir os erros residuais do modelo anterior, com foco nos exemplos mal classificados. Esta abordagem mostrou-se particularmente eficaz para os dados do nosso projeto, uma vez que mesmo preditores fracos podem, em conjunto, gerar sinais preditivos relevantes, desde que bem combinados. O *XGBoost* destaca-se por incorporar mecanismos internos de regularização, como a penalização de árvores muito profundas, o que reduz significativamente o risco de *overfitting* — algo crucial num dataset com fraca correlação entre variáveis.

O *Support Vector Machine* trabalha a partir do princípio de construir um hiperplano de separação ótimo num espaço multidimensional gerado pelos preditores. Cada jogo é representado como um ponto nesse espaço, e o objetivo do modelo é encontrar a fronteira que melhor separa os jogos com 2 e 3 sets, maximizando a margem entre os pontos mais próximos de cada classe — os chamados vetores de suporte. Contudo, num problema como o nosso, em que os preditores não têm elevada capacidade discriminativa e a separação linear no espaço original não é evidente, o SVM recorre à utilização de *kernels*. Em especial, é uma excelente escolha quando os preditores não têm uma estrutura relacional forte com a variável alvo, como neste caso.

Esta fase permitiu testar e comparar diferentes abordagens algorítmicas, consolidando uma base robusta para compreender o comportamento preditivo dos modelos e sustentar, com rigor, a análise dos seus resultados.

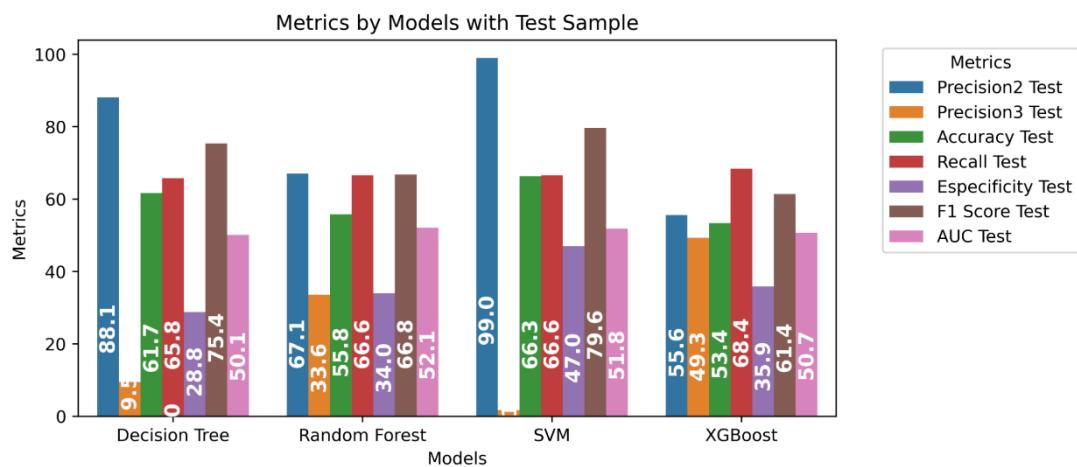
7. EVALUATION

Concluída a fase de modelação, procedeu-se à avaliação dos modelos treinados com o objetivo de comparar e interpretar o desempenho dos modelos desenvolvidos na tarefa de prever, com base em características do jogo e dos jogadores, se uma partida de ténis no formato *best of 3* será decidida em dois ou três sets. Esta análise foi conduzida em duas abordagens: uma divisão simples em treino e teste; e uma validação cruzada estratificada com 10 *folds*, de forma a permitir avaliar a generalização dos modelos em diferentes contextos.

A análise do desempenho dos modelos foi realizada com base num conjunto abrangente de métricas, tais como: *accuracy*, que representa a proporção de previsões corretas; *precision*, referente à capacidade do modelo de prever corretamente uma classe; *recall*, é a capacidade de identificar corretamente todos os exemplos de uma classe; *specificity*, capacidade de identificar corretamente os negativos; *F1 Score*, trata-se da média entre o *precision* e o *recall*; *AUC*, mede a capacidade do modelo de distinguir corretamente as classes; e por último, o Índice de *Huberty*, que mede o quanto o modelo melhora face à escolha da classe mais frequente.

Na abordagem treino/teste, os resultados revelaram diferenças significativas entre os modelos.

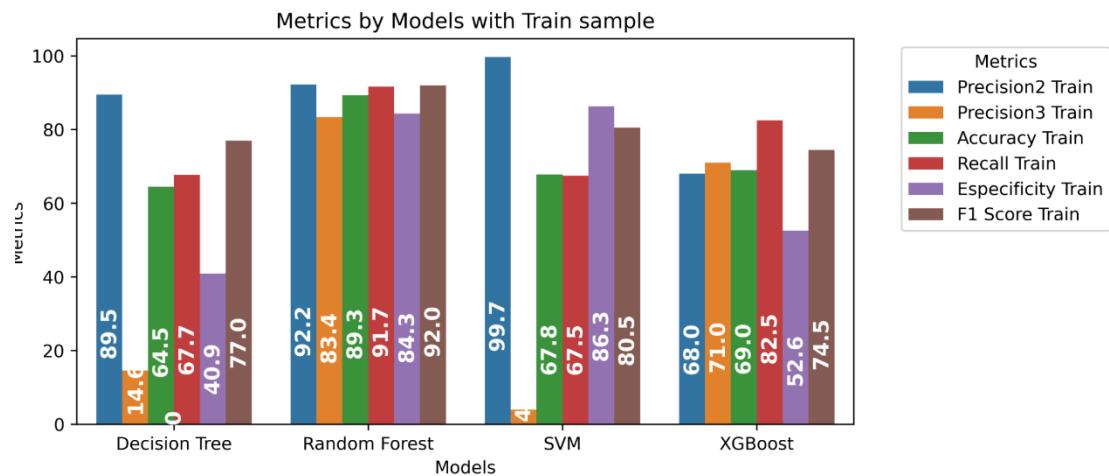
Conjunto Teste



No conjunto de teste, a comparação entre os quatro modelos revelou padrões distintos em termos de desempenho e equilíbrio entre classes. O modelo *SVM* destacou-se por apresentar o maior valor de *accuracy* (66.3%) e o valor mais elevado de *F1 score* (79.6%), o que sugere uma elevada capacidade preditiva global. No entanto, uma análise

mais detalhada das métricas por classe revelou um desempenho extremamente assimétrico: a *precision* para a classe 2 (jogos decididos em dois sets) foi de 99.0%, enquanto, por sua vez, a *precision* para a classe 3 (jogos decididos em três sets) foi apenas de 1.7%. Este comportamento mostra que, apesar da elevada performance global, o modelo estava altamente enviesado para a classe maioritária, falhando sistematicamente na identificação da classe minoritária. Este padrão foi também visível, ainda que em menor grau, noutras modelos. A *Decision Tree* embora tenha atingido uma *accuracy* razoável (61,7%) e um *F1 Score* de 75,4%, demonstrou um comportamento fortemente enviesado, com uma *precision* muito baixa na classe 3 (9,5%) e baixa *specificity* (28,8%), indicando dificuldades em reconhecer corretamente jogos dessa classe. O modelo *XGBoost* apresentou a *accuracy* mais baixa entre os quatro (53,4%), mas destacou-se por ser o mais equilibrado entre as classes. Este modelo foi o único a mostrar uma verdadeira capacidade de identificar eficazmente jogos com três sets, o que é particularmente relevante num contexto com desbalanceamento da variável alvo.

Conjunto Treino



No conjunto de treino, os modelos apresentaram, em geral, desempenhos mais elevados do que no conjunto de teste, como seria de esperar, dada a familiaridade com os dados.

O *Random Forest* foi o modelo com melhor desempenho global em treino, com uma *accuracy* de 89,3% e um *F1 Score* de 92,0%, revelando-se altamente eficaz na classificação dos exemplos já vistos. Também se destacou com valores muito elevados de precisão para ambas as classes e um *recall* de 91,7%. A *specificity* (84,3%) reforça esta conclusão, sugerindo que o modelo distinguiu bem as duas classes. No entanto, a discrepância em relação ao desempenho no conjunto de teste sugere possível sobreajuste

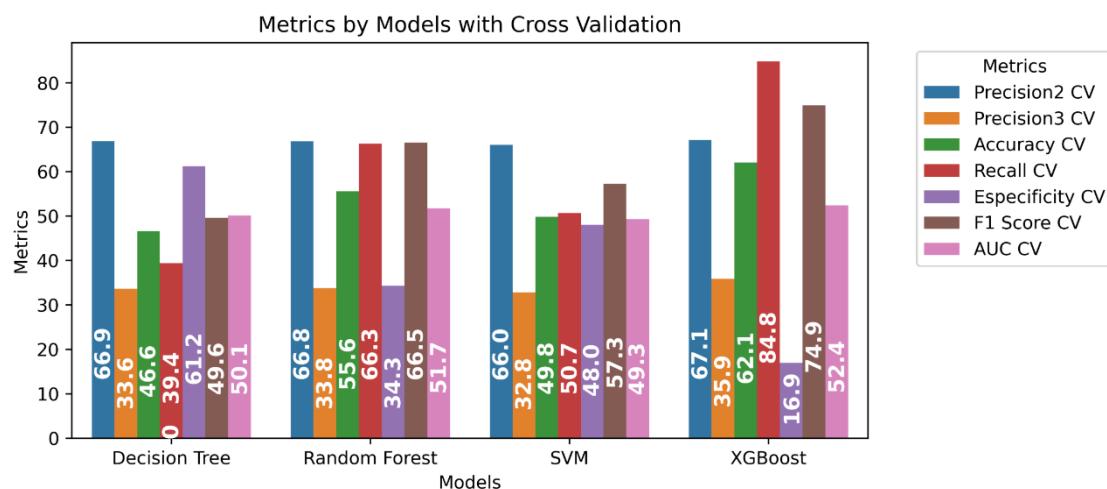
(*overfitting*) — ou seja, o modelo adapta-se demasiado aos dados que viu, mas perde capacidade de generalização para novos dados (anexos [15,16,17,18,19,20](#)).

O modelo *SVM*, por sua vez, embora apresente a maior *precision* na classe 2 (99,7%) e uma *specificity* de 86,3%, repete o padrão observado em teste com um desempenho fraco na classe 3 (apenas 4,0% de precisão). Isto indica que o modelo continua fortemente enviesado para a classe majoritária.

De um modo geral, os resultados em treino reforçam que os modelos mais complexos (*Random Forest* e *XGBoost*) conseguem adaptar-se melhor aos dados, com alta performance, embora o primeiro mostre sinais claros de sobreajuste.

Já na abordagem do *Cross Validation*, os resultados foram, em geral, ligeiramente inferiores, como esperado, mas mais estáveis.

Cross Validation



Globalmente, o modelo *XGBoost* destacou-se pela maior *accuracy* média (62,1%) e pelo valor mais elevado de *F1 Score* (74,9%), além de obter também o melhor *recall* (84,8%). Além disso, foi o único modelo a ultrapassar os 50% de *precision* para ambas as classes. A sua *precision* para a classe 3 (35,9%) foi a melhor do grupo, o que confirma a sua eficácia na identificação da classe minoritária. No entanto, esta performance foi acompanhada de uma *specificity* extremamente baixa (16,9%), sinalizando que o modelo comprometeu fortemente a identificação correta da classe 2.

A *Decision Tree* foi, novamente, o modelo com desempenho mais fraco, apresentando *accuracy* de 46,6%, *recall* de apenas 39,4% e um *F1 Score* de 49,6%.

Em termos de *AUC*, os modelos apresentaram valores entre 49,3% (*SVM*) e 52,4% (*XGBoost*), todos relativamente próximos, o que sugere que nenhum modelo possui uma capacidade de discriminação excepcional neste problema específico.

Foi ainda analisado também o Índice de *Huberty*, com base no gráfico que compara os resultados obtidos em teste e em *cross validation*, que revelou valores negativos para todos os modelos ([Anexo 21](#)). Este comportamento indica um desempenho inferior ao que seria esperado por pura aleatoriedade. Ainda assim, a proximidade relativa entre os resultados de teste e de *cross validation* indica alguma estabilidade nos desempenhos.

Em conclusão, embora o *XGBoost* apresente os melhores resultados absolutos, o *SVM* revelou-se o modelo mais equilibrado e estável entre todas as métricas e abordagens. Esta consistência indica elevada capacidade de generalização e menor risco de sobreajuste, tornando-o, do ponto de vista técnico, a solução mais confiável para prever o número de sets em jogos de ténis profissionais.

8. DEPLOYMENT

Ao longo da escrita deste relatório esta fase já foi subliminarmente evidenciada, porém éposta agora aqui em destaque. Já se tornou numa prática popular a realização de apostas desportivas online com o surgimento de plataformas para tal credenciadas mundialmente. Estas plataformas começaram a se popularizar com o surgimento das apostas direcionadas para o universo do futebol e à medida que as pessoas se foram envolvendo passou-se a trabalhar em alastrar para os restantes desportos como o basquetebol, andebol, ténis e entre outros desportos.

Porém, em alguns desportos as apostas ainda não são tão generalistas, por exemplo, no futebol, fora a equipa vencedora, apostam-se também o número de cantos, cartões amarelos ou vermelhos, número de golos marcados por um jogador específico, praticamente a maior parte dos detalhes por trás de uma partida de futebol são alvos de apostas. No ténis ainda não é tão genérico e se tem trabalhado para tal, este projeto visa contribuir nesta evolução.

As apostas, de modo a retornarem lucros para quem decide arriscar o seu palpite, precisam ser calculadas *odds* que refletem o quanto o apostador deverá ganhar por cada unidade da sua moeda investida no seu palpite e estas *odds* são calculadas com base nas probabilidades de ocorrência de qualquer fenómeno, o modelo desenvolvido neste trabalho, não retorna simplesmente o número exato de sets a serem jogados, ele retorna também a probabilidade desse número ser real, pelo que este modelo está habilitado a ser utilizado pelas casas de apostas para a implementação de uma nova aposta que é o número de sets a serem jogados, ou algo que para o contexto dos jogos que contém o *dataset* utilizado para a modelagem (jogos à melhor de 3) uma aposta poderia ser ambos os jogadores ganham um set, isto que se reflete no número de sets, de realçar que já existem algumas plataformas de apostas com esta aposta.

Porém, este modelo não precisa servir só para as plataformas de apostas como também para as equipas técnicas dos jogadores de ténis e com base nestas probabilidades e nas previsões em si definir estratégias de jogos de modo a dar vantagem aos seus jogadores como dito antes na fase introdutória.

9. CONCLUSÃO

Com este projeto foi-nos permitido aprofundar diversas competências nas diferentes fases da metodologia *CRISP-DM*, promovendo o contacto com diversas tecnologias e linguagens de programação, nomeadamente *MongoDB*, *MySQL* e *Python*, estas foram essenciais para tratar, estruturar e analisar este conjunto de dados real. A divisão das etapas por diferentes softwares revelou-se vantajosa, o que nos favorece no sentido de estarmos mais aptos às diferentes linguagens de programação.

Num contexto real este projeto poderia não ser viável dadas os problemas relatados ao longo do relatório e, apesar do esforço investido na preparação e enriquecimento dos dados, enfrentaram-se limitações significativas no que diz respeito à qualidade e correlação entre as variáveis preditoras e a variável alvo (número de *sets*). Esta falta de correlação refletiu-se nos resultados obtidos, com métricas de performance medíocres, aproximando os modelos desenvolvidos a modelos baseados na classe modal (como indicado pela métrica do índice de *huberty*).

Ainda assim, a aplicação rigorosa de técnicas de validação cruzada e balanceamento de classes (*SMOTE*) permitiu-nos selecionar um modelo com métricas mais equilibradas, demonstrando que, mesmo em contextos desafiantes, é possível obter soluções preditivas minimamente aceitáveis. Embora a viabilidade de aplicação prática deste modelo num cenário real seja condicionada pelas limitações identificadas, a sua implementação, conforme proposta, continua a ser realizável.

Deste modo, este projeto revelou-se uma experiência valiosa para consolidar os conhecimentos adquiridos ao longo do curso, expondo-nos a um problema real, mostrando que não se irá lidar com variáveis preditoras com boas correlações com a variável alvo, estando, assim, preparados para lidar com situações semelhantes no futuro profissional.

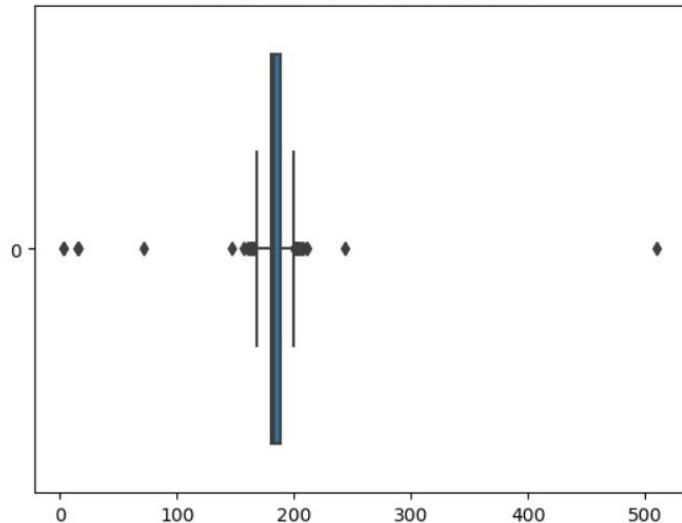
10. REFERÊNCIAS BIBLIOGRÁFICAS

- Brownlow, T. (2021). *WTA/ATP Tennis Data*. Kaggle. Acessado a 1 de abril de 2025, de [WTA/ATP Tennis Data](https://www.kaggle.com/datasets/wbrownlow/wta-atp-tennis-data)
- Bohnacker. (2022). *Country Longitude Latitude*. Kaggle. Acessado a 3 de Abril de 2025, de <https://www.kaggle.com/datasets/bohnacker/country-longitude-latitude>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: syntetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 321-357.
- Filipcic, A., Zecic, M., Reid, M., & Crespo, M. (26 de dezembro de 2015). Differences in performance indicators of elite tennis players in the period 1991–2010. Obtido em 6 de abril de 2025, de https://www.researchgate.net/publication/295705032_Differences_in_performance_indicators_of_elite_tennis_players_in_the_period_1991-2010
- HALLMARK, E. (2018). *A large Tennis dataset for ATP and ITF betting*. Kaggle. Acessado a 3 de Abril de 2025, de <https://www.kaggle.com/datasets/ehallmar/a-large-tennis-dataset-for-atp-and-itf-betting>
- Hegarty, T., & Whelan, K. (2025). Forecasting soccer matches with betting odds: A tale of two markets. *International Journal of Forecasting*, 41(2), 803-820. doi:<https://doi.org/10.1016/j.ijforecast.2024.06.013>
- Rui, B., Hing, K. C., Haoyuan, L., & Yew, J. T. (abril de 2024). Forecasting Tennis Player Matches Based on Machine Learning. *Proceedings of 2024 International Conference on Machine Learning and Intelligent Computing*". Obtido em 6 de abril de 2025, de <https://proceedings.mlr.press/v245/rui24b.html>
- Serrano, A. D. (2020). *PREDICTING TENNIS MATCH OUTCOME A MACHINE LEARNING*. Obtido em 6 de abril de 2025, de <https://lib.ugent.be/>
- ÜNAL, T. (2023). *PREDICTING TENNIS MATCH OUTCOME: A MACHINE LEARNING APPROACH USING THE SRP-CRISP-DM FRAMEWORK*. Obtido em 6 de abril de 2025, de <https://open.metu.edu.tr/handle/11511/106394>

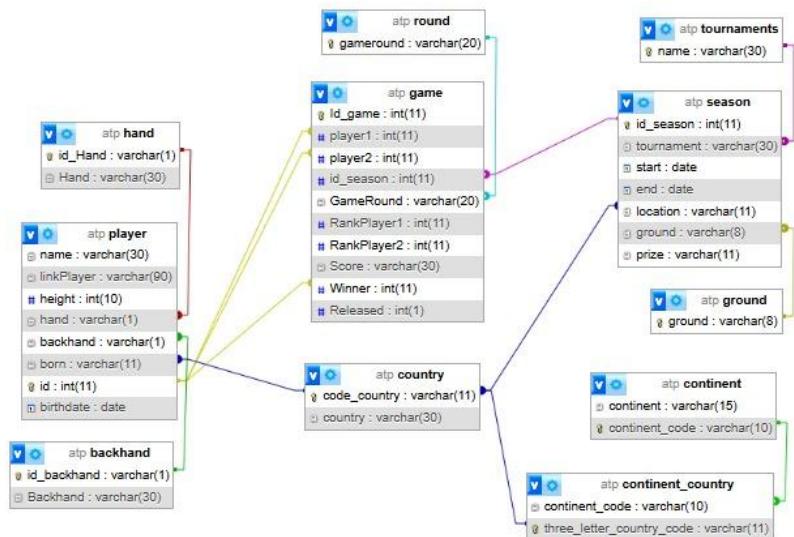
11. ANEXOS

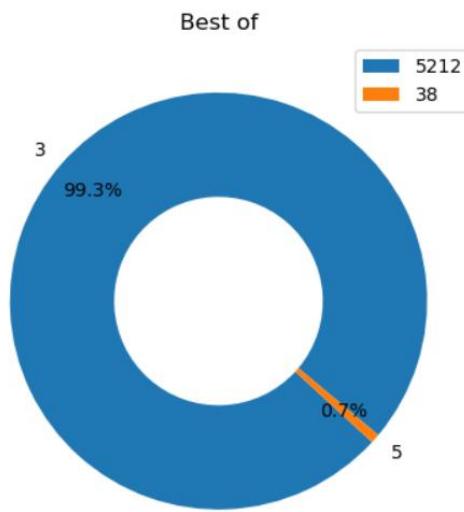
Anexo 1

Histograma altura dos jogadores

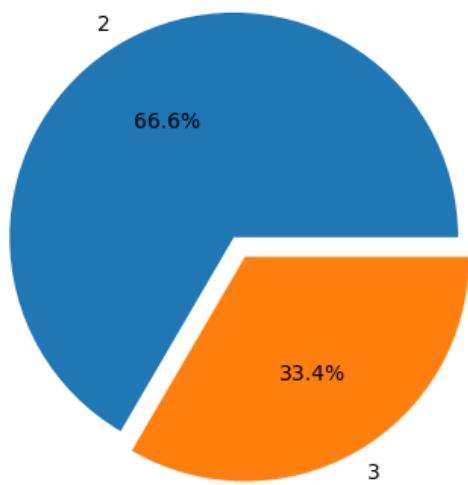
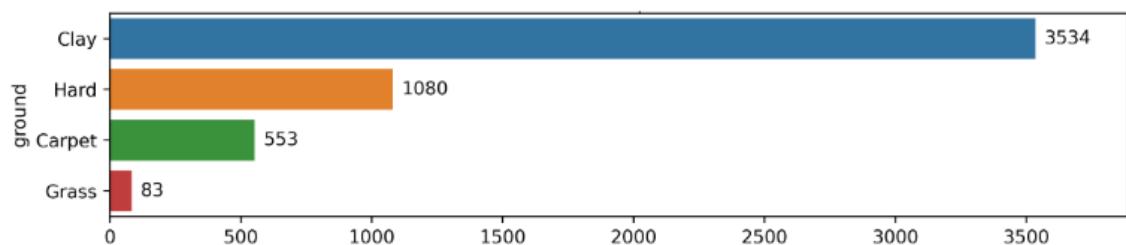


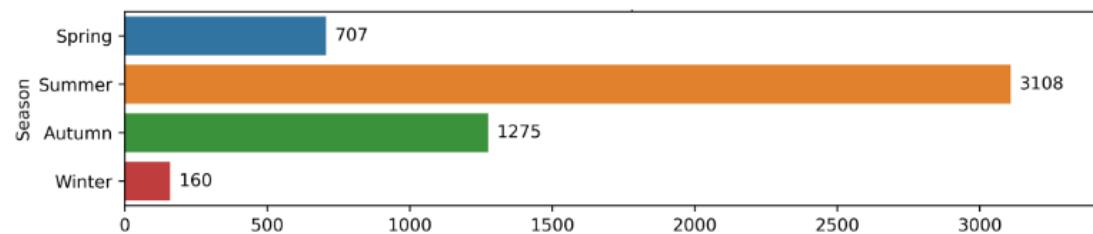
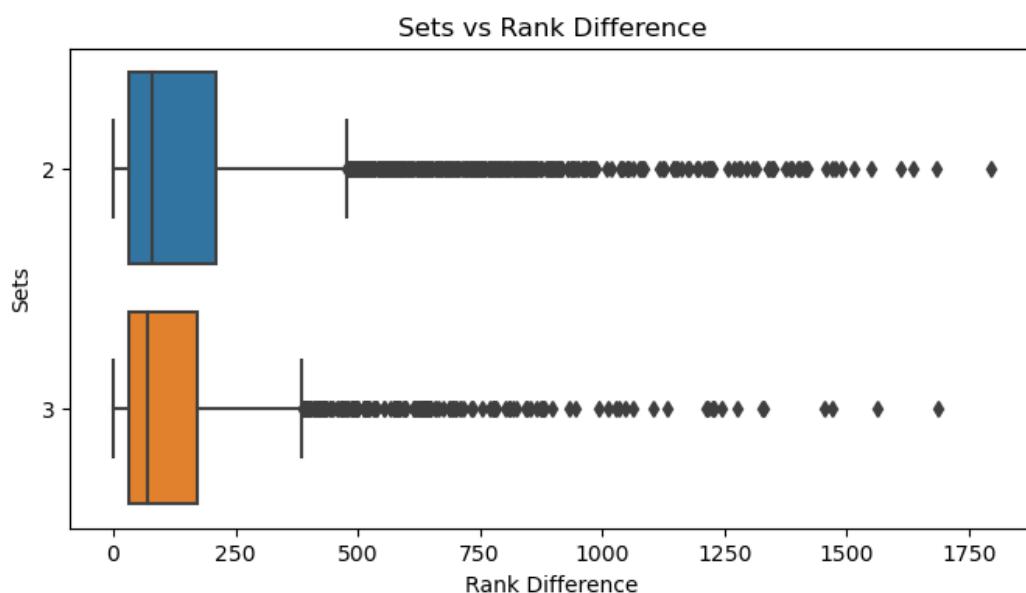
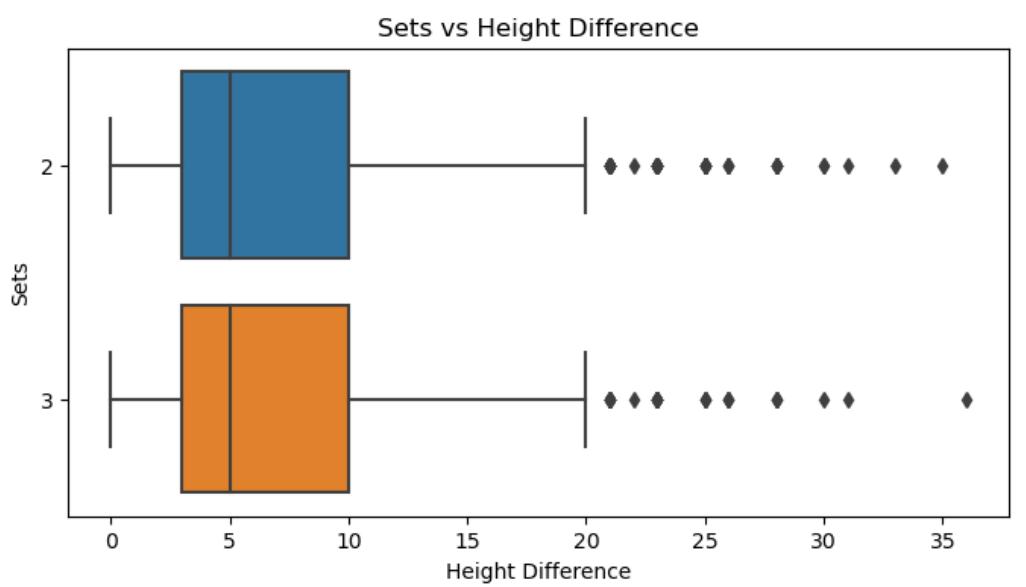
Anexo 2

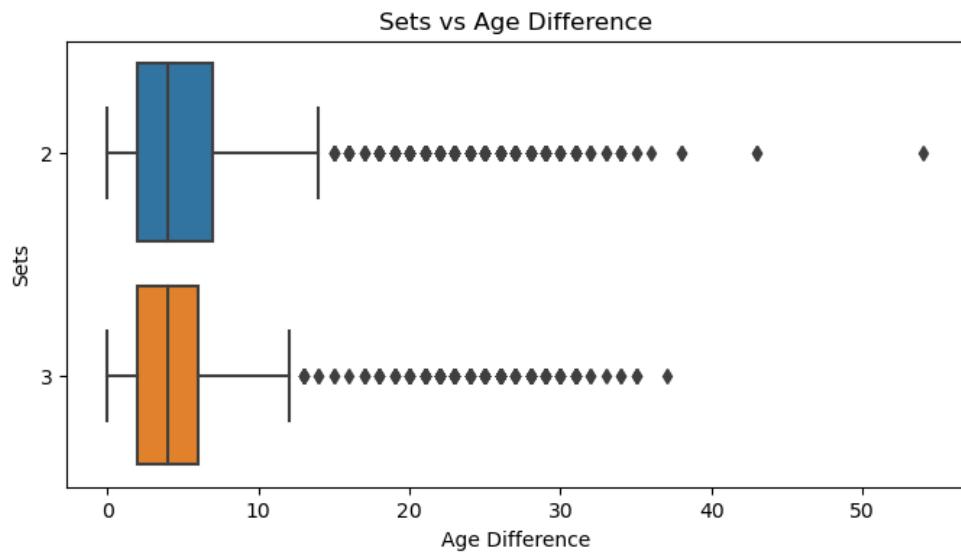
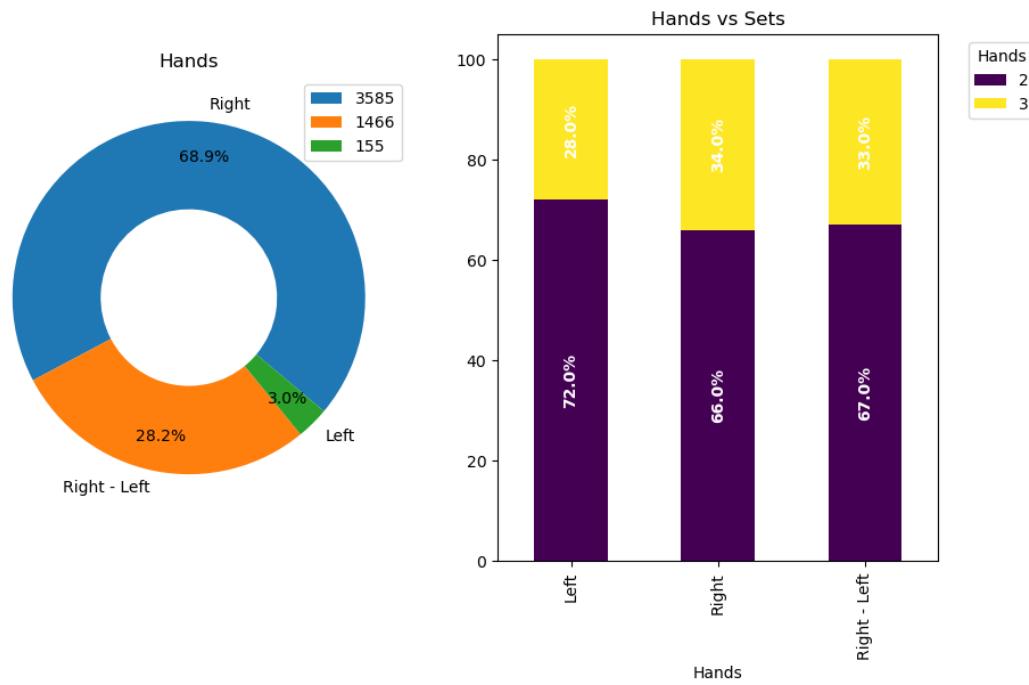


Anexo 3**Anexo 4**

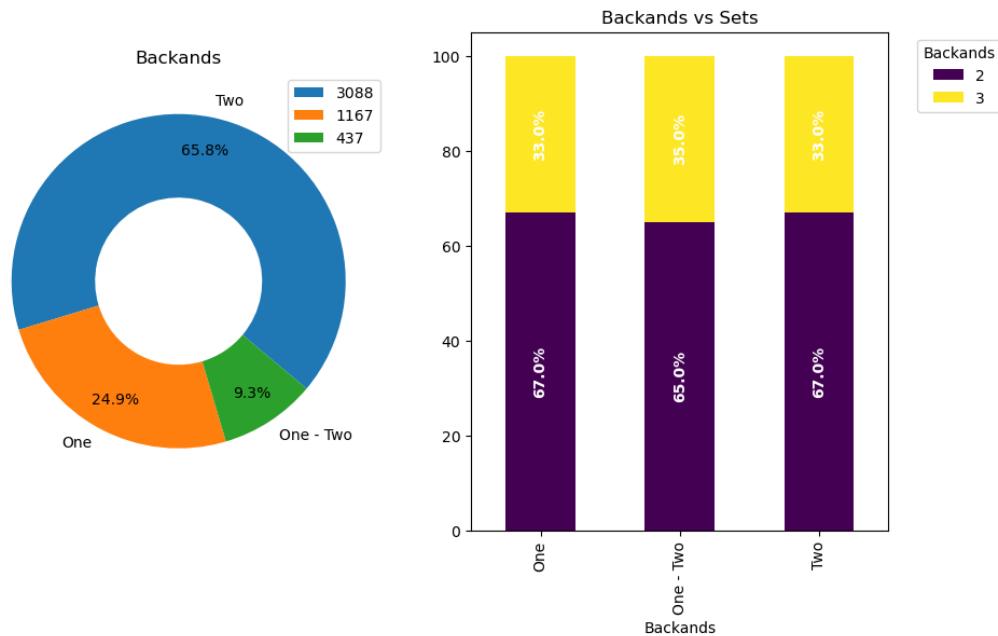
Distribuição da variável Sets

**Anexo 5**

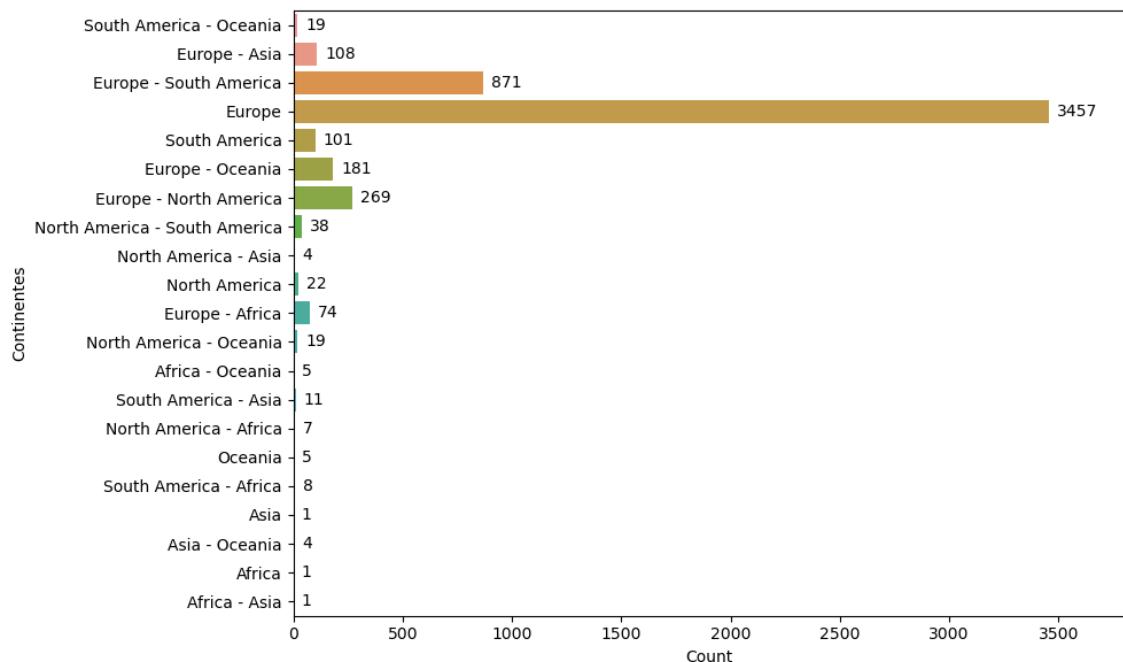
Anexo 6**Anexo 7****Anexo 8**

Anexo 9**Anexo 10**

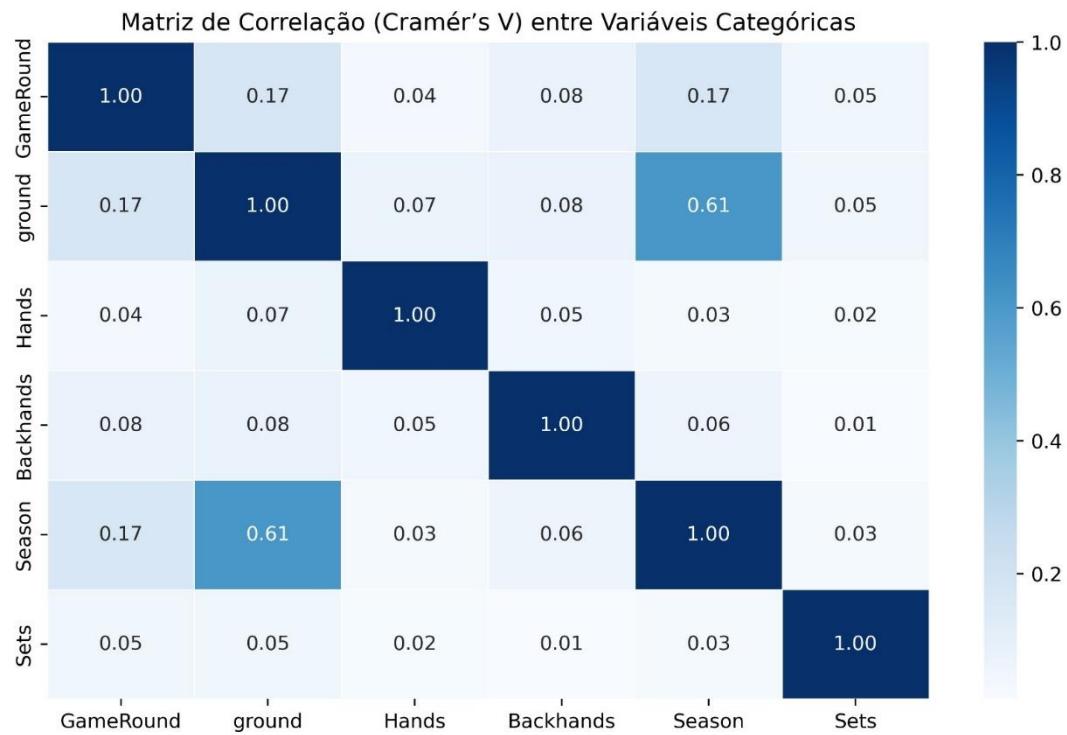
Anexo 11



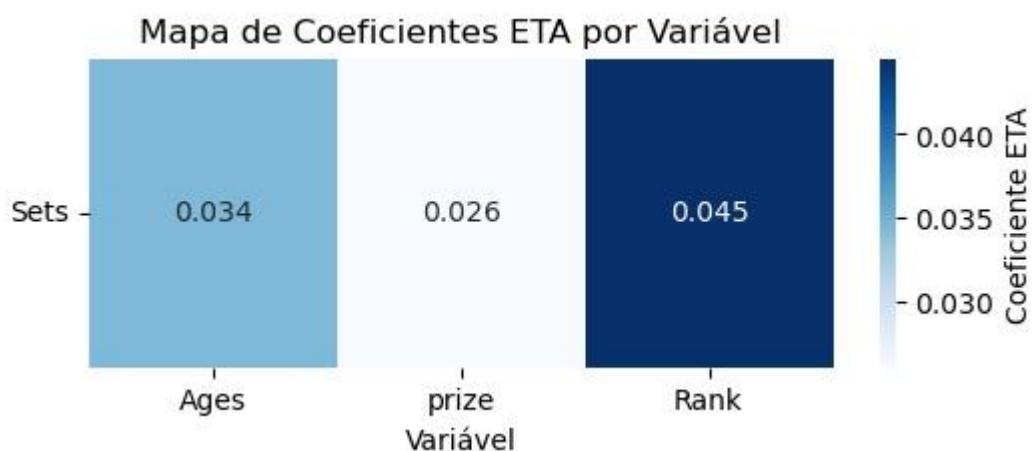
Anexo 12



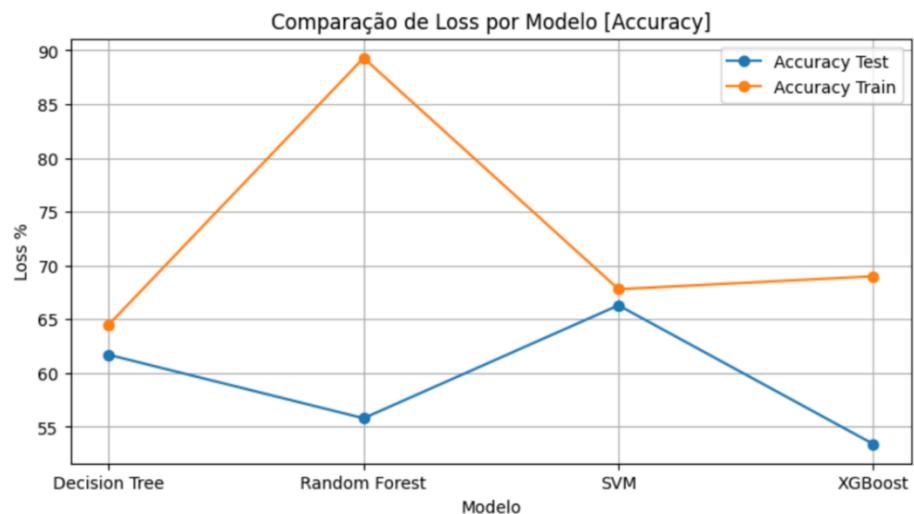
Anexo 13



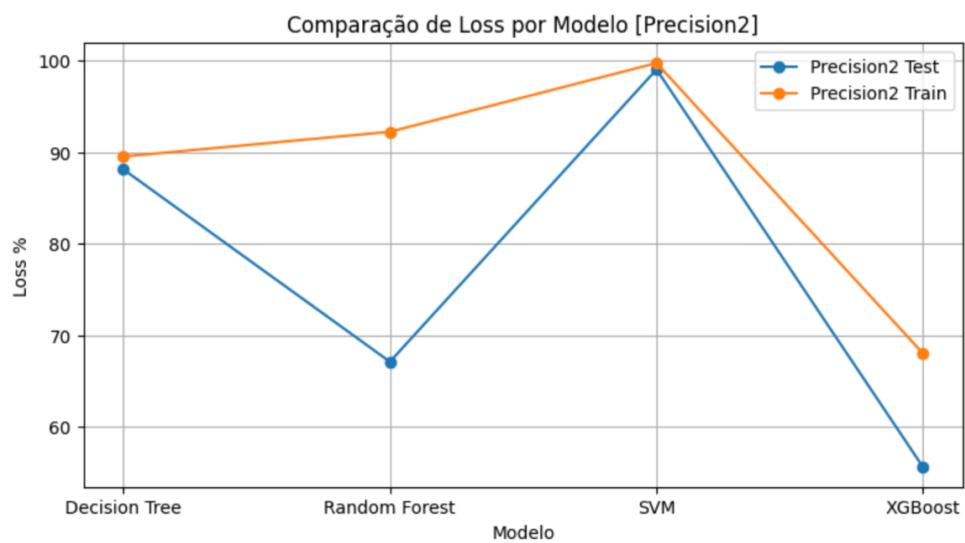
Anexo 14



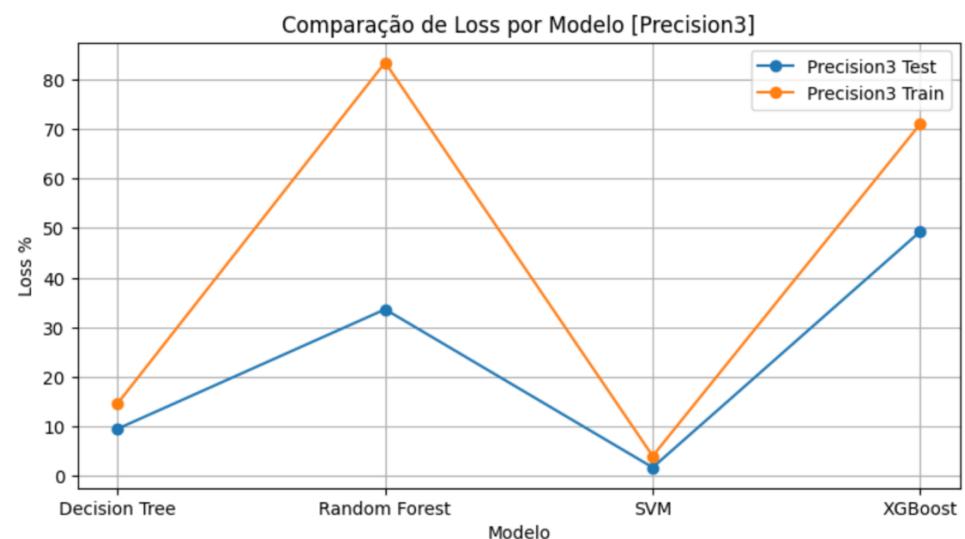
Anexo 15

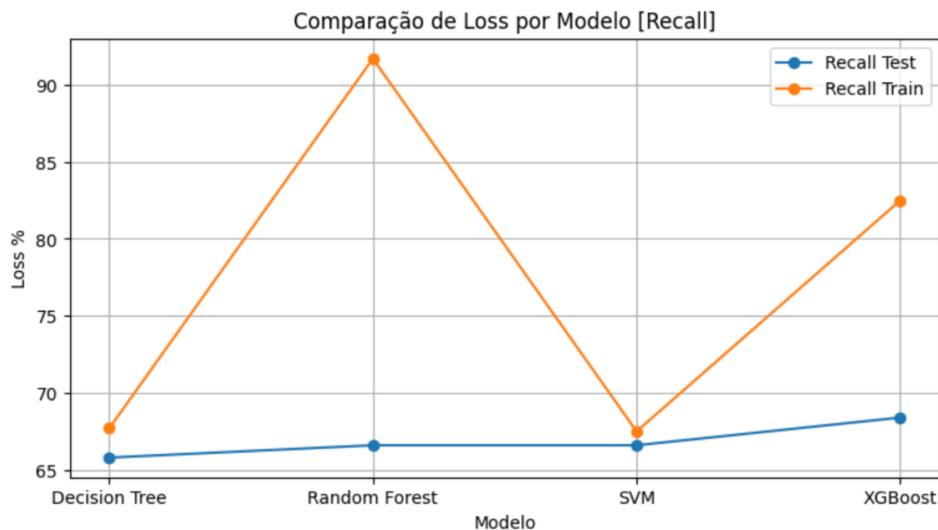
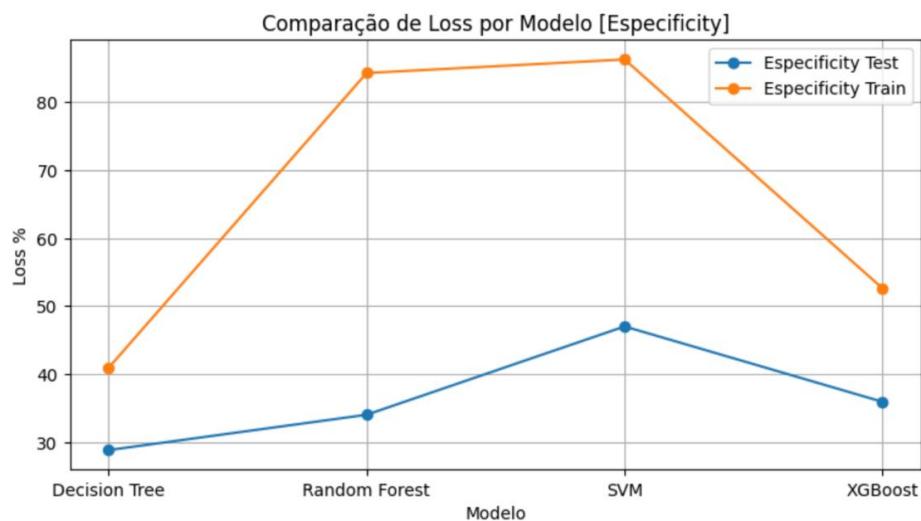
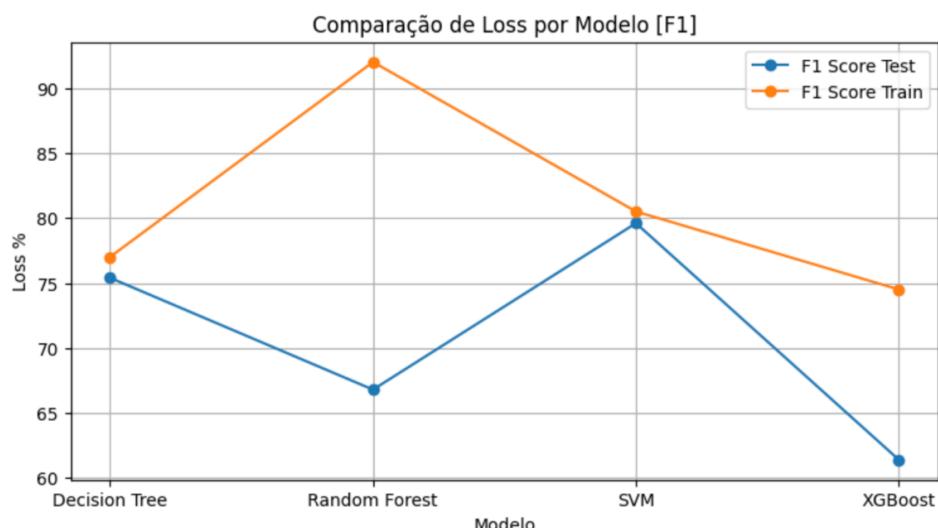


Anexo 16



Anexo 17



Anexo 18**Anexo 19****Anexo 20**

Anexo 21