UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

WALTER ZOLCSAK NETO NOGARA

CONSTRUÇÃO DE MODELOS PARA PREVISÃO DE RESULTADOS EM PARTIDAS DE TÊNIS UTILIZANDO MACHINE LEARNING

PRESIDENTE PRUDENTE
2023

WALTER ZOLCSAK NETO NOGARA

CONSTRUÇÃO DE MODELOS PARA PREVISÃO DE RESULTADOS EM PARTIDAS DE TÊNIS UTILIZANDO MACHINE LEARNING

Relatório final para Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/Unesp para aproveitamento na disciplina Trabalho de Conclusão de Curso.

Professor Orientador Klaus Schlunzen Júnior

PRESIDENTE PRUDENTE

2023

N774c

Nogara, Walter Zolcsak Neto
Construção de Modelos para Previsão de Resultados Em Partidas de Tênis Utilizando Machine Learning / Walter Zolcsak Neto Nogara. -- Presidente Prudente, 2023
44 p. : tabs.

Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências e Tecnologia, Presidente Prudente
Orientador: Klaus Schlünzen Junior

1. Aprendizado do computador. 2. Modelos lineares. 3. Ciências do esporte. 4. Estatística. 5. Tenis (Jogo). I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências e Tecnologia, Presidente Prudente. Dados fornecidos pelo autor(a).

TERMO DE APROVAÇÃO

Walter Zolcsak Neto Nogara

CONSTRUÇÃO DE MODELOS PARA PREVISÃO DE RESULTADOS EM PARTIDAS DE TENIS UTILIZANDO MACHINE LEARNING

Relatório de Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

Orientador: _____
Prof. Dr. Klaus Schelmzen Junior
Departamento de ESTATÍSTICA

_____
Prof. Dr. SÉRGIO MINORU OIKAWA
Departamento de ESTATÍSTICA

_____
Prof. Dr. MIRIAM RODRIGUES SILVESTRE
Departamento de ESTATÍSTICA

Presidente Prudente, 10 de 02 de 2023.

RESUMO

Estamos na era da revolução tecnológica, um mundo cada vez mais digital e direcionado aos dados, exemplo disso é como as pessoas fazem compras ou pagam contas, na maioria das vezes pelo seu smartphone. Por consequência disso, as empresas vêm manipulando tecnologias avançadas para a obtenção de informações a partir dos seus bancos de dados. Isso possibilita que tomem decisões rápidas e inteligentes, auxiliando no desenvolvimento da empresa e possibilitando também a previsão de certos acontecimentos. A partir disso, surge o interesse em aplicar a "Análise de dados" em diversas áreas, como no setor bancário, alimentício, farmacêutico e até mesmo nos esportes. Na procura do movimento perfeito, de aumentar a velocidade, de resultados positivos, ou seja, maior eficiência nas competições, inúmeras modalidades esportivas vem procurando na tecnologia, com a ajuda da análise de dados, o possível aprimoramento no desempenho esportivo e buscar o melhor dos atletas. Neste trabalho foi abordado a análise de dados aplicada nos esportes, mais especificamente, no tênis. Foram utilizadas técnicas para construção de um modelo para classificar em vitória ou derrota, partidas da ATP (Association of Tennis Professionals). Para isso, nos baseamos em dois algoritmos de classificação, sendo eles, árvore de decisão e regressão logística. Por último, os dois modelos foram comparados e analisados, conclui-se que não é possível assumir que um modelo obteve desempenho melhor que o outro, portanto a solução ideal deveria ser a junção desses modelos, onde cada um ficaria responsável por métricas diferentes.

Palavras-chave: Tênis, análise de dados, previsão, regressão logística, árvore de decisão.

ABSTRACT

We are in the era of the technological revolution, an increasingly digital and data-driven world, an example of which is how people shop or pay bills, most often through their smartphone. As a result, companies have been manipulating advanced technologies to obtain information from their databases. This enables them to make quick and intelligent decisions, helping in the development of the company and also making it possible to predict certain events. From this, there is an interest in applying "Data Analysis" in several areas, such as banking, food, pharmaceuticals and even sports. In search of the perfect movement, to increase speed, of positive results, that is, greater efficiency in competitions, numerous sports modalities have been looking to technology, with the help of data analysis, for the possible improvement in sports performance and to seek the best of athletes. In this work, data analysis applied in sports was approached, more specifically, in tennis. Techniques were used to build a model to classify ATP (Association of Tennis Professionals) matches in victory or defeat. For this, we rely on two classification algorithms, namely, decision tree and logistic regression. Finally, the two models were compared and analyzed, it is concluded that it is not possible to assume that one model performed better than the other, so the ideal solution should be the capacity of these models, where each one would be responsible for different metrics.

Keywords: Tennis, data analysis, forecasting, logistic regression, decision tree.

LISTA DE FIGURAS

Figura 1.1 – Exemplo dos tipos de quadra;

Figura 1.2 – Dimensão da quadra de tênis com medidas;

Figura 2.1 – Processo de criação de um modelo preditivo com um algoritmo de classificação;

Figura 2.2 – Processo de predição da classe para novas instâncias a partir de um modelo preditivo já treinado;

Figura 2.3 – Processo da aplicação de um algoritmo de agrupamento;

Figura 2.4 – Árvore de Decisão para jogar tênis;

Figura 2.5 – Diferença entre classes da AD;

Figura 2.6 – Diferença entre Regressão linear e regressão logística;

Figura 2.7 – Proporção de AVC de acordo com os valores de pressão arterial;

Figura 3.1 – Esquema do pré-processamento de Dados;

Figura 3.2: Exemplo de sumarização dos atributos no R;

Figura 4.1: Fluxograma do processo de validação dos modelos;

Figura 4.2: Lado esquerdo da Árvore;

Figura 4.3: Lado direito da Árvore;

Figura 4.4: Tabela de resíduos RL;

Figura 4.5: Variáveis para estimação;

Figura 4.6: Coeficientes de regressão.

LISTA DE TABELAS E QUADROS

Tabela 1 – Representação da pontuação de cada acerto dentro de um game;

Tabela 2.1 – Exemplo de base de dados para algoritmo de classificação;

Quadro 2.1 – Representação matriz de confusão;

Quadro 2.2: Representação matriz de confusão (Acurácia);

Quadro 2.3: Representação matriz de confusão (Precisão);

Quadro 2.4: Representação matriz de confusão (Recall);

Tabela 4.1: Total de partidas da base de dados;

Tabela 4.2: Matriz de confusão (Árvore de decisão);

Tabela 4.3: Resultados de cada classe (Árvore de decisão);

Tabela 4.4: Resultados das métricas (Árvore de decisão);

Quadro 4.5: Matriz de confusão (Regressão logística);

Quadro 4.6: Resultados de cada classe (Regressão logística);

Quadro 4.7: Resultados das métricas (Regressão logística);

Quadro 4.8: Resultados F1-Score por cada algoritmo;

Quadro 4.9: Comparação dos algoritmos por cada métrica;

LISTA DE SIGLAS

ATP – Association of Tennis Professionals

ML – Machine Learning

AM – Aprendizado de Máquina

RL – Regressão Logística

AD – Árvore de Decisão

SUMÁRIO

1. INTRODUÇÃO 11
2. REFERENCIAL TEÓRICO 14

2.1 ANÁLISE DE DADOS NO MUNDO DO TÊNIS 14

2.2 MACHINE LEARNING 15

2.2.1 Aprendizado Supervisionado 16

2.2.2 Aprendizado Não Supervisionado 17

2.3 ALGORITMOS 18

2.3.1 Árvore de decisão (AD) 18

2.3.2 Regressão Logística (RL) 21

2.4 MÉTRICAS DE AVALIAÇÃO 25

3. PREPARAÇÃO DOS DADOS 29

3.1 CARACTERIZAÇÃO DOS DADOS 29

3.2 PRÉ-PROCESSAMENTO DOS DADOS 31

3.3 LIMPEZA DOS DADOS 32

3.4 CRIAÇÃO DE VARIÁVEIS 32

3.5 ORGANIZAÇÃO DO BANCO DE DADOS 33

3.6 SELEÇÃO DE VARIÁVEIS 33

4. METODOLOGIA 35

4.1 BASE DE DADOS 35

4.2 RESULTADOS 37

4.2.1 Árvore de decisão (AD) 37

4.2.2 Regressão Logística (RL) 39

4.3 COMPARAÇÃO DE DESEMPENHO DOS MODELOS 40

5. CONCLUSÃO 42

REFERÊNCIAS 43

1. INTRODUÇÃO

A origem deste esporte (Tênis) é muito questionável. Acredita-se que os primeiros jogos praticados com bola pelos egípcios, gregos e romanos, tenham sido os predecessores do jogo de tênis. Os primórdios mais próximos do jogo em sua forma mais conhecida, surgiram na França, no século XIII ou XIV através do ¨Le Jeu de Paume¨ (jogo com a palma da mão).

Atualmente, o tênis é praticado em uma quadra de saibro, piso duro ou grama. É disputado por duas equipes adversárias que podem conter um (jogo simples) ou dois integrantes (jogo de duplas), que se utilizam de uma bola de borracha e uma raquete - fabricada com grafite ou fibras de carbono. A duração de cada partida é medida em sets; cada set reúne a contagem de uma segunda pontuação denominada game. Para ganhar o set, o jogador (ou dupla) deverá pontuar no mínimo 6 games com 2 games de diferença em relação ao adversário (WIKIPÉDIA). Caso o jogador chegue a 6 games com menos de 2 games de diferença do adversário, é realizado o game de desempate conhecido como tie-break.

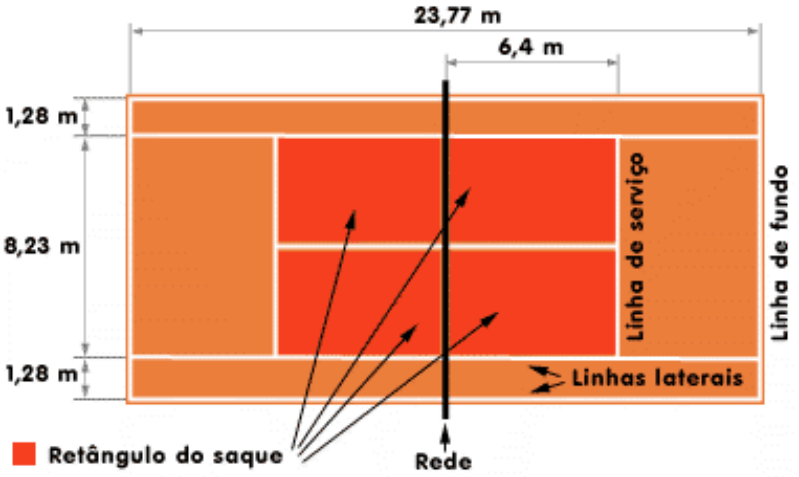
Figura 1.1 - Exemplo dos tipos de quadra.

Fonte: https://teniscuritiba.com.br/tipos-de-quadras-de-tenis/

Para o jogo simples, a quadra de tênis tem dimensões de um retangulo de 23,77 metros de extensão e 8,23 metros de largura. Uma rede de fio torcido divide a quadra ao meio com 1,07 metros de altura nas laterais e 0,91 metros de altura no centro. Posteriormente, cada lado da quadra é subdividido ao meio e a metade mais

próxima a rede também é subdivida ao meio com aproximadamente 4,1 metros de largura. Essa subdivisão serve para determinar as áreas de saque.

Figura 1.2 - Dimensão da quadra de tênis com medidas.



Fonte: http://hiegotenis.blogspot.com/2012/02/saiba-tudo-sobre-as-quadras-de-tenis-i.html

Há duas formas principais de se pontuar em uma partida: winner e erro. O ponto de winner se dá quando o jogador arremessa a bola de forma que ela quique na quadra adversária sem que o oponente consiga rebater para o lado oposto antes do segundo quique, independentemente de onde ele seja. O erro se dá quando o jogador adversário rebate a bola e esta não passa pela rede ou não quica dentro da quadra válida (REGRAS DOS ESPORTES, 2017). Há três maneiras de ganhar um set: (1) o jogador ganhar seis games e ter pelo menos dois games de vantagem sobre o oponente; (2) se a pontuação do set estiver 6x5 deve ser jogado mais um set, se o jogador em vantagem fizer 7x5 ele ganha; (3) porém se o adversário fizer 6x6 deve ser jogado um set chamado tie-break. Para o jogo de tie-break a pontuação varia de acordo com o torneio. Quem ganhar o tie-break será o vencedor.

Tabela 1 - Representação da pontuação de cada acerto dentro de um game.

1° ponto	15
2° ponto	30
3° ponto	40
4° ponto	Game

Fonte: Autor

Em esportes de alta performance, simples mudanças geram grandes consequências, a diferença entre o primeiro e o segundo colocado pode ser tão mínima quanto alguns décimos de segundo, e pequenas especificidades resultam entre estar no pódio ou não. Grande parte dos esportes têm a assistência de diversas áreas da ciência para melhorar o rendimento dos jogadores, desde acessórios tecnológicos, como equipamentos mais leves e resistentes até assistência médica.

A Estatística também passou a ser utilizada nas práticas esportivas para fiscalizar e prever cada ação de um atleta durante uma competição, ou seja, com esta análise a equipe técnica é capaz de convocar os jogadores, elaborar táticas e tomar decisões durante uma partida e classificar os oponentes. Sua relevância é tanta que é aplicada em inúmeras áreas para favorecer os atletas e agradar os telespectadores com uma quantidade enorme de dados, previsões e possibilidades. A Estatística está presente em vários esportes, como por exemplo no beisebol, onde há um caso popular do uso da ferramenta de análise de dados, no filme ‘’O homem que mudou o jogo’’, de 2011.

Além de ser uma grande atração para assistir e torcer, o tênis também possibilita, através de casas de apostas online, chances do espectador se beneficiar sobre os resultados dos jogos. Atualmente vemos algumas opções de apostas, não se limitando apenas a quem vai ganhar. Por isso, a previsão de resultados esportivos tornou-se fundamental para torcedores, apostadores, treinadores, etc., porém, prever os resultados de uma partida de tênis não é uma tarefa fácil, existem vários fatores externos que podem influenciar diretamente no resultado. Sendo assim, para realizar as predições, estudaremos alguns métodos estatísticos de aprendizado de máquina. E como aplicação, partiremos dos dados históricos das partidas, ou seja, pretendemos predizer o resultado de um confronto direto entre dois times utilizando os dados de jogos que já aconteceram.

O objetivo principal deste trabalho foi identificar, através das técnicas de aprendizado de máquina, um modelo preditivo final capaz de prever resultados de confrontos nos torneios de tênis.

2. REFERENCIAL TEÓRICO

2.1 ANÁLISE DE DADOS NO MUNDO DO TÊNIS

Grande parte dos jogadores profissionais estão utilizando a análise de dados, porém poucos dizem sobre. O tênis está passando por uma revolução analítica. Os principais tenistas estão pagando salários exorbitantes para equipes de analistas na confiança de ganhar alguma vantagem sobre o adversário (SAMUELS, 2019).

Este conceito não é novo, a análise de dados tem sido utilizada por times em diversas modalidades por quase duas décadas, em especial quando Billy Beane, dirigente do Oakland Athletics, manuseou estatísticas para construir uma equipe com orçamento limitado. Porém, ao contrário do beisebol ou do futebol, a análise de dados no tênis não se limita apenas em vasculhar uma equipe, mas também em otimizar o desempenho de todos os jogadores.

O crescimento dos dados esportivos tornou essa análise inevitável. No passado, era complicado adquirir tais informações das práticas e dos jogos de um oponente, atualmente, quantidades ilimitadas dessas imagens, vídeos permitem que os analistas identifiquem padrões e prevejam comportamentos para preparar um atleta contra um oponente específico.

Um retrato disso é a “árvore decisao”, análise que revela onde um jogador é mais disposto a acertar a bola em cada posição na quadra. Utilizando tais dados, os treinadores podem criar padrões de treinamentos para explorar as fraquezas dos seus atletas.

Por último, a análise de dados ainda é algo referente a jogadores profissionais, pois certos dados podem oferecer insights sobre o jogo de um adversário além do que o olho nu pode enxergar, permitindo que os jogadores alterem e ajustem seu plano tático de jogo. À medida que a tecnologia evolui, é notável que possa ser usada durante partidas que estão em andamento, como por exemplo na ATP CUP, onde ocorre análises simultâneas e quando há o intervalo, de um game para outro, essas informações são repassadas para o jogador.

2.2 MACHINE LEARNING

Tradicionalmente, grande parcela dos problemas a serem esclarecidos metodicamente necessitam de pessoas com experiência, ou seja, profissionais nos tópicos classificados. O Machine Learning é um termo criado em 1959 por Arthur Samuel**,** engenheiro do MIT que é apontado como o pioneiro da inteligência artificial (IA). Na época, referiu o conceito como "um campo de estudo que dá aos computadores a habilidade de aprender sem terem sido programados para tal".

> “... [machine learning é] uma área de Inteligência Artificial que tem como objetivo desenvolver técnicas computacionais que permitam a predição e o aprendizado de determinados comportamentos ou padrões automaticamente a partir de experiências acumuladas por exemplos e problemas anteriores.” (MITCHEL, p.147, 1997).

A tradução do termo “machine learning” já mostra a sua essência, que, em português, reflete "aprendizado de máquina (AM)". Esse método firma-se no conceito de que programas podem aprender com dados por meio de ferramentas estatísticas típicas, além de aplicar uma diversidade de algoritmos para identificar padrões e tomar decisões com o mínimo de interferência humana.

Em razão das complicações dos problemas a serem solucionados e ao crescente volume de informações obtidas, quando nos referimos aos dados é humanamente impossível realizar a análise sem algum processo computacional. Os algoritmos de aprendizado de máquina procuram resolver o problema da análise, tendo como meta a criação de regras que o solucionem. Utilizam-se de informaçães passadas para aprenderem padrões que podem ser aplicados na predição ao de dados futuros.

No presente temos a utilização de algoritmos de aprendizado de máquina em inúmeras corporações. Como por exemplo, análise de risco na concessão de crédito, diagnóstico de doenças, meteorologia e previsões esportivas. Os algoritmos são subdivididos em dois grandes grupos: supervisionados e não supervisionados.

2.2.1 Aprendizado Supervisionado

O conceito principal do aprendizado supervisionado é dispor de preditores (dados de entrada ou inputs) para prognosticar uma ou mais respostas (dados de saída ou outputs), que podem ser quantitativas ou qualitativas (categorias, atributos ou fatores). O caso de respostas qualitativas representa a problemas de classificação e aquele de respostas quantitativas, a problemas de regressão (MORETTIN, SINGER, 2019). Nesta classe de aprendizado temos a existência de uma variável de destino (ou dependente), em que a finalidade é fazer predições sobre essa variável sobre de um conjunto de outras variáveis que designamos de explicativas (ou independentes). Demonstramos ao computador qual é cada entrada e automaticamente, há um entendimento de quais características fazem as entradas serem o que são, através da observação de um conjunto de objetos descritos, como se fosse um treinamento.

Referente a classificação, o objetivo é prever classes dentro das possibilidades limitadas existentes, ou seja, o algoritmo irá desenvolver um agrupador com a finalidade de dizer se um objeto pertence ou não a uma determinada categoria, podendo ser de duas classes (variável binária) ou mais.

Como exemplo, pode-se prever uma decisão da realização de um jogo de tênis, dado as condições climáticas. A Tabela 2.1 demonstra uma típica tabela na classificação e exemplifica os atributos descritivos da condição climática de alguns jogos e a decisão final se o jogo foi ou não realizado. A coluna ”Joga” é a classe do problema.

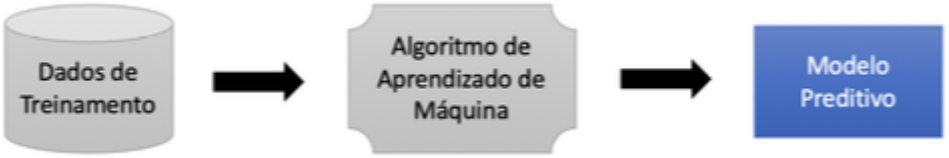
Tabela 2.1 - Exemplo de base de dados para algoritmo de classificação.

Tempo	Temperatura	Umidade	Vento	Joga
Chuvoso	71	91	Sim	Não
Ensolarado	69	70	Não	Sim
Ensolarado	80	90	Sim	Não
Nublado	83	86	Não	Sim
Chuvoso	70	96	Não	Sim
Chuvoso	65	70	Sim	Não
Nublado	64	65	Sim	Sim
Nublado	72	90	Sim	Sim
Ensolarado	75	70	Sim	Sim
Chuvoso	68	80	Não	Sim
Nublado	81	75	Não	Sim
Ensolarado	85	85	Não	Não
Ensolarado	72	95	Não	Não
Chuvoso	75	80	Não	Sim

Fonte: Autor

A Figura 2.1 mostra o processo de aquisição de um modelo preditivo, e a Figura 2.2 mostra o processo para predição de novas instâncias.

Figura 2.1 - Processo de criação de um modelo preditivo com um algoritmo de classificação.



Fonte: Autor

Figura 2.2 - Processo de predição da classe para novas instâncias a partir de um modelo preditivo já treinado.

Fonte: Autor

2.2.2 Aprendizado Não Supervisionado

O aprendizado não supervisionado, também é conhecido como aprendizado descritivo, é atribuído a função de identificação de informações significativas ao conjunto de dados. Para utilização da técnica não é necessário o intermédio humano na gestão do aprendizado, nem o conhecimento prévio sobre classes ou categorias.

Na aprendizagem não supervisionada, não contém variável ou resultado para prever ou estimar (TURKSON, BAAGYERE, WENYA, 2016). O conjunto de dados manipulados não possui nenhum tipo de rótulo, portanto, não há a presença da variável dependente, de forma que a funçao desse tipo de aprendizagem é encontrar similaridades entre os objetos estudados com o intúito de produzir agrupamentos de variáveis, verificar associações ou reduzir a dimensão da base de dados.

Como exemplo, a Figura 2.3 ilustra um processo de utilização de um algoritmo de agrupamento. Para uma base de dados inicial onde corresponde ao modo de consumo de indivíduos, aplicam um algoritmo de agrupamento , em seguida, podem ser observados três clusters de saída. O primeiro cluster configura consumidores com alto gasto em abastecimento, o segundo consumidores com alto gasto em lojas e o terceiro consumidores com alto gasto em comida.

Figura 2.3 - Processo da aplicação de um algoritmo de agrupamento.

Fonte: https://www.techedgegroup.com/pt/blog/como-aplicar-machine-learning-em-financas

2.3 ALGORITMOS

2.3.1 Árvore de decisão (AD)

O Algoritmo "árvores de decisão" prega a tática de dividir-e-conquistar (divide-and-conquer), no qual as árvores são estruturas manuseando apenas algumas propriedades. A árvore de decisão é um dos processos de AM, no qual uma adversidade é decomposta em subproblemas. Recursivamente, a mesma estratégia é aplicada a cada subproblema (GAMA, 2000 apud SILVA et al., 2008).

Uma AD é uma técnica que auxilia na tomada de decisão pelo gráfico no formato de árvore e corresponde às condições e as probabilidades para solucionar problemas, são utilizadas porque demonstram ao analista uma fácil interpretação. O algoritmo utilizado na visualização da árvore pertence ao grupo de AM supervisionado. A figura 2.4 ilustra um exemplo de uma árvore de decisão para jogar tênis.

Figura 2.4 – Árvore de Decisão para jogar tênis.

Fonte: Autor

Da mesma forma que um fluxograma, a AD determina nós (decision nodes) que se conectam entre si por uma hierarquia, existe o nó-raiz (root node), que é o de maior relevância, e os nós-folha (leaf nodes), que são os apurações finais. No cenário de ML, o nó-raiz é um dos atributos da base de dados e o nó-folha é a classe ou o valor que será gerado como resposta.

Na ligação entre nós, temos diretrizes de “se-então”. Ao chegar em um nó A, o algoritmo se questiona perante uma condição, como “se a característica X do registro analisado é maior do que 10?”. Se for maior, então ele vai para um lado da árvore; se for menor, então ele vai para outro. No próximo nó, segue a mesma lógica.

É um algoritmo que mantém o que denominamos de “recursivo” na computação. Isto significa que ele repete o mesmo padrão na medida em que vai entrando em novos níveis de profundidade. É como se uma função chamasse a ela mesma como uma segunda função para uma execução paralela, da qual a primeira função depende para gerar sua resposta.

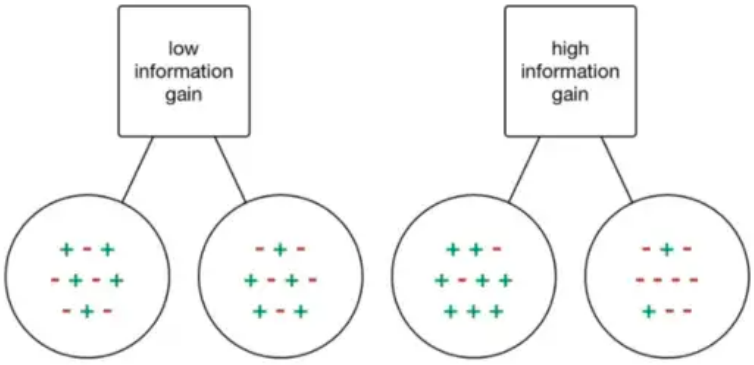
O objetivo da AD é indicar variadas divisões dos dados em subconjuntos, de modo que os subconjuntos ficarão mais uniformes. Um subconjunto dos dados será mais uniforme visto que tenha poucas classes (ou apenas uma) da variável alvo. Elaborar uma árvore de decisão refere-se a descobrir regras sobre as variáveis do modelo que tornam os ramos da árvore mais homogêneos (VIEIRA, 2008).

Para delimitar a base de dados em uma árvore, dependemos das condições. Diante dessas condições, dividimos a base em caminhos com análise dos registros

que satisfazem determinada condição. Se uma condição for “altura > 160” para o lado direito, por exemplo, teremos somente os registros da base que possuem altura maior que 160 cm. Já se a condição for “altura < 160” para a esquerda, teremos somente registros com altura menores que 160 cm para a esquerda.

O ganho de informação é obtido a partir dessa lógica. Ao analisar um atributo, se os registros das bases para cada lado são homogêneos ou próximos disso, temos um alto ganho de informação. Portanto, consideramos que se tenha preferência por uma determinada condição, consecutivamente saibamos exatamente a saída esperada ou estejamos mais próximos de descobrir.

Figura 2.5 – Diferença entre classes da AD.



Fonte: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

O processo de tomada de decisão para o mercado de trabalho, demanda que um profissional conheça o ambiente em que a empresa está inserida e também esteja ciente das suas constantes mudanças. Por deliberação, entende-se a escolha que alguém realiza entre, no mínimo, duas alternativas possíveis, utilizando o meio que julga ser o melhor disponível para atingir um determinado objetivo (CORRAR; THEÓPHILO, 2004 apud SOUTO-MAIOR, 2014).

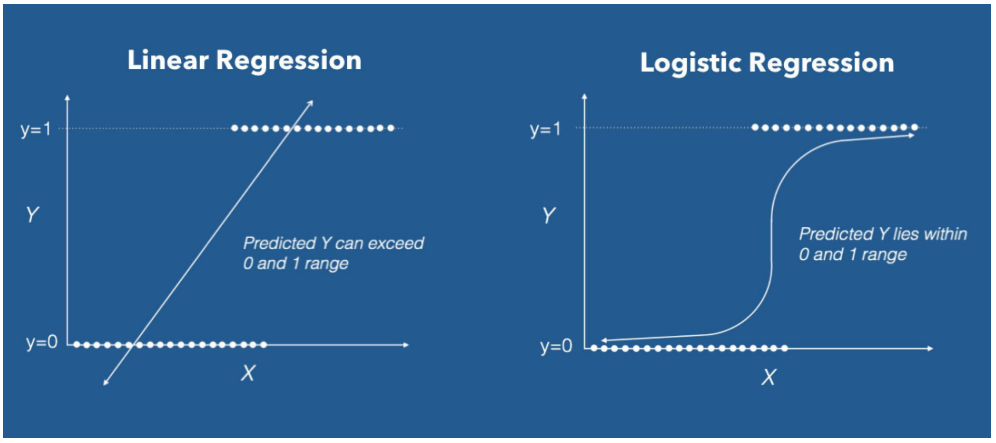
Além disso, toda árvore possui um nó chamado raiz, que possui o maior nível hierárquico (o ponto de partida) e ligações para outros elementos, denominados filhos. Esses filhos podem possuir seus próprios filhos que por sua vez também possuem os seus. O nó que não possui filho é conhecido como nó folha ou terminal. Portanto, uma árvore de decisão nada mais é que uma árvore que armazena regras em seus nós, e as nossas folhas representam a decisão a ser tomada.

2.3.2 Regressão Logística (RL)

É uma técnica utilizada para ocasiões em que a variável dependente é de natureza dicotômica ou binária, ou seja, os resultados da análise ficam contidos no intervalo de zero a um. Quanto às independentes, tanto podem ser categóricas ou não. A regressão logística é um recurso que nos permite estimar a probabilidade associada à ocorrência de determinado evento em face de um conjunto de variáveis explanatórias, portanto, procura estimar a probabilidade da variável dependente assumir um determinado valor em função dos conhecimentos de outras variáveis.

É relevante falarmos sobre regressão linear e sobre as suas diferenças com a regressão logística (RL). No modelo de regressão linear, a variável dependente é considerada contínua, enquanto na regressão logística é categórica. Ela pertence a uma família de modelos chamada modelos lineares generalizados (GLM) utilizando a mesma fórmula básica da regressão linear, mas, em vez do contínuo, está regredindo para a probabilidade de um resultado categórico. No uso, é manuseada para classificação binária ou multi-classe (onde é chamado de regressão logística multinômial).

Figura 2.6 – Diferença entre Regressão linear e regressão logística.



Fonte: https://datapeaker.com/pt/big--data/regresion-logistica-guia-para-principiantes-de-regresion-logistica

Por exemplo, A RL é utilizada em áreas como as seguintes: Em medicina, permite por exemplo determinar os factores que caracterizam um grupo de indivíduos doentes em relação a indivíduos sãos; No domínio dos seguros, permite encontrar frações da clientela que sejam sensíveis a determinada política securitária em

relação a um dado risco particular; Em instituições financeiras, pode detectar os grupos de risco para a subscrição de um crédito; Em econometria, permite explicar uma variável discreta, como por exemplo as intenções de voto em atos eleitorais.

Através da RL é possível avaliar os fatores que influenciam a ocorrência de um determinado evento, buscando estimar a probabilidade de a variável dependente assumir um determinado valor em função dos valores conhecidos de outras variáveis, e os resultados da análise ficam contidos no intervalo de zero a um.

Ela utiliza a função de ligação "logit", o que possibilita a interpretação dos resultados em função da Razão de Chances (Odds Ratio). Algumas vantagens da regressão logística são a facilidade de lidar com variáveis independentes categóricas, fornecimentos de resultados em termos de probabilidade e alto grau de confiabilidade.

$$ln(odds) \Rightarrow ln(\frac{p}{1-p});$$

$$Logit^{1} g(x) = \frac{1}{1+e^{-g(x)}}$$

Seja y_i a variável resposta binária e $x_i = (1, x_{i1},..., x_{ip})^T$ o vetor de p covariáveis do i-ésimo indivíduo $(i = 1,..., n)$. Assumimos que $y_i \sim Bernoulli\,(p_i)$, em que p_i é a probabilidade de ocorrência de um determinado evento e desejamos calculá-la a partir das covariáveis preditoras presentes em x_i, em que $g(x_i)$ é o preditor linear e $\beta_0, \beta_1,..., \beta_p$ representam parâmetros do modelo:

$$g(x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}$$

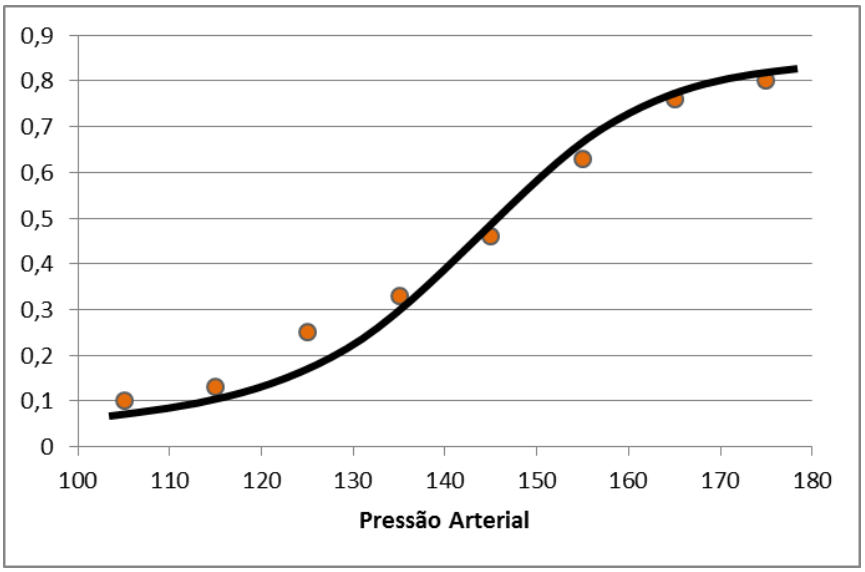
$$p = \frac{1}{1+e^{-g(x_i)}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_p x_{ip})}} = \frac{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_p x_{ip}}{1+e^{(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_p x_{ip})}}$$

Obs: Nota-se, que a fórmula do preditor linear $g(x_i)$, é na verdade, uma regressão Linear.

Considerando uma certa combinação de coeficientes $(\beta_0, \beta_1,..., \beta_p)$ e variando

os valores de X, observa-se que a curva logística tem um comportamento probabilístico no formato da letra S, o que é uma característica da regressão logística (HOSMER; LEMESHOW; STURDIVANT, 2013)

Figura 2.7 – Proporção de AVC de acordo com os valores de pressão arterial.



Fonte:https://ebmacademy.wordpress.com/2015/08/17/o-fantasma-da-regressao-logistica/

Os coeficientes $(\beta_0, \beta_1, ..., \beta_p)$ são estimados a partir do conjunto dados, pelo método da máxima verossimilhança, em que encontra uma combinação de coeficientes que maximiza a probabilidade da amostra ter sido observada:

$$L(\beta) = \prod_{i=1}^{n} (x_i)^{y_i} [1 - (x_i)]^{1-y_i}$$

Aplicando-se o logaritmo natural em ambos os lados da equação, obtemos a função log-verossimilhança:

$$L(\beta) = ln[L(\beta)] = \sum_{i=1}^{n} [y_i ln(x_i) + (1 - y_i) ln(1 - (x_i))]$$

O valor β que maximiza $ln[L(\beta)]$ é obtido após derivar $l(\beta)$ em relação aos parâmetros $(\beta_0, \beta_1, ..., \beta_p)$:

$$\frac{\partial\, ln\,[L(\beta)]}{\partial\, \beta_0} = \sum_{i=1}^{n} [y_i - (x_i)]\ ;$$

$$\frac{\partial\, ln\,[L(\beta)]}{\partial\, \beta_1} = \sum_{i=1}^{n} x_i [y_i - (x_i)];$$

$$\frac{\partial\, ln\,[L(\beta)]}{\partial\, \beta_p} = \sum_{i=1}^{n} x_{ip} [y_i - (x_i)]$$

Após a estimação dos coeficientes, há a importância em averiguar o nível de significância das variáveis do modelo. Isso inclui a formulação e teste de uma hipótese estatística para determinar se as variáveis independentes do modelo são significativamente relacionadas com a variável dependente. Para isto, há testes para avaliar o modelo logístico. Os testes mais utilizados são os testes da Razão da Verossimilhança, o teste de Wald e Pseudo R2 de Cox e Snell.

2.4 MÉTRICAS DE AVALIAÇÃO

Para elaborar estudos de machine learning, é crucial a aplicação de métricas adequadas de validação para solucionar erros. Tem consequência na apuração da qualidade dos modelos, portanto, indicam a informação do desempenho desses modelos. Logo, se a escolha for mal realizada, terá dificuldade em avaliar se o modelo de fato está sendo eficiente.

Posteriormente ao cálculo de probabilidade, há a necessidade em associar um valor estimado para y, denotado por ŷ, que pode estar correspondente ou não, com o valor real. A tabela abaixo representa a chamada matriz de confusão:

Quadro 2.1 - Representação matriz de confusão.

	ŷ = 0	**ŷ = 1**
y = 0	VN	FP
y = 1	FN	VP

Onde:

VN - Verdadeiro Negativo: são as observações que o modelo previu como negativas e realmente eram negativas.

FP - Falso Positivo: são as observações que o modelo previu como positivas, mas na realidade eram negativas.

FN - Falso Negativo: são as observações que o modelo previu como negativas mas na realidade eram positivas.

VP - Verdadeiro Positivo: são as observações que o modelo classifica como positiva e realmente eram positivas.

A partir desses quatro valores, podemos definir três conceitos para avaliar se a predição foi eficiente:

- A **acurácia** mede a proporção de predições corretas, independentemente de serem verdadeiros positivos ou negativos, ou seja, a métrica de acurácia nos diz quantos de nossos exemplos foram de fato classificados corretamente, independente da classe. Ela é definida pela fórmula abaixo e representada na sua matriz de confusão a seguir:

$$ACURÁCIA = \frac{VP+VN}{VN+FP+FN+VP}$$

Quadro 2.2: Representação matriz de confusão (Acurácia).

	ŷ = 0	**ŷ = 1**
y = 0	VN	FP
y = 1	FN	VP

- A **precisão** mede a proporção de verdadeiros positivos, a capacidade do modelo classificar como positivo dado que ele é de fato positivo, ou seja, representa a capacidade de um modelo em prever a classe positiva. Logo,

demonstra maior ênfase para os erros por falso negativo, é representada pela fórmula a seguir e também pela fórmula:

$$PRECISÃO = \frac{VP}{VP+FN}$$

Quadro 2.3: Representação matriz de confusão (Precisão).

	ŷ = 0	**ŷ = 1**
y = 0	VN	FP
y = 1	FN	VP

- A **revocação** é a proporção dos verdadeiros positivos entre todas as observações que são positivas no seu conjunto de dados, ou seja, é utilizada para indicar a relação entre as previsões classificadas corretamente como positivas e o total de observações classificadas como positivas. É possível observar que a precisão dá um destaque maior para os erros por falso positivo, é representada pela fórmula a seguir:

$$RECALL = \frac{VP}{VP+FP}$$

Quadro 2.4: Representação matriz de confusão (Recall).

	ŷ = 0	**ŷ = 1**
y = 0	VN	FP
y = 1	FN	VP

A métrica **Score F1** é usada a partir de classes desbalanceadas, ou seja, quando uma base de dados possui muitos exemplos de uma classe e poucos da outra classe. É dada pela média harmônica entre a revocação e a precisão:

$$Score\ F1\ = \frac{2 \times Precisão \times Recall}{Precisão + Recall}$$

Portanto, um modelo que apresenta um F1 Score alto é um modelo qualificado, em relação aos acertos das predições (alta precisão) e em recuperar as observações da classe de interesse (alta revocação). Grande parte dos problemas, o F1 Score é uma métrica melhor que a acurácia, especialmente em casos onde os erros possuem impactos diferentes no modelo.

Métricas de avaliação em um modelo de classificação desempenham um peso essencial para o sucesso do estudo. Compreender o que cada mensuração representa e quais serão suas restrições, facilitam no desenvolvimento. Recomenda-se levar em conta o objetivo do modelo no mundo real, o peso de cada tipo de erro e o quão interpretável deve ser a métrica, entre outros fatores.

3. PREPARAÇÃO DOS DADOS

3.1 CARACTERIZAÇÃO DOS DADOS

Os banco de dados utilizados para a realização do estudo foram disponibilizados pelo site Kaggle. Este conjunto de dados contém registros de todas as partidas referentes a dezessete temporadas do circuito ATP, da temporada 2000 até a temporada 2017, totalizando 53119 partidas.

Os dados disponibilizados sobre as partidas são:

- torney_id - Número de identificação do torneio;
- tourney_name - Nome do torneio;
- surface - Tipo de quadra (Saibro; Rápida; Grama);
- draw_size - Quantidade de jogadores no card principal;
- tourney_level - Nivel do torneio (Master 1000; Atp 1000; Atp 500; Atp 250);
- tourney_date - Data no torneio;
- match_num - Número da partida;
- winner_id - Número de identificação do vencedor;
- winner_seed - Cabeça de chave do vencedor;
- winner_entry - Se entrou no card principal ou qualificatórias;
- winner_name - Nome do vencedor;
- winner_hand - Mão dominante do vencedor (Direita; Esquerda);
- winner_ht - Altura do vencedor (cm);
- winner_ioc - Nacionalidade do vencedor;
- winner_age - Idade do vencedor;
- winner_rank - Ranking do vencedor;
- winner_rank_points - Quantidade de pontos adquiridos pelo vencedor;
- loser_id - Número de identificação do perdedor;
- loser_seed - Cabeça de chave do perdedor;
- loser_name - Nome do perdedor;
- loser_hand - Mão dominante do perdedor;
- loser_ht - Altura do perdedor;
- loser_ioc - Nacionalidade do perdedor;
- loser_age - Idade do perdedor;

- loser_rank - Ranking do perdedor;
- loser_rank_points - Quantidade de pontos adquiridos pelo perdedor;
- score - Resultado da partida;
- best_of - Quantidade de Sets;
- round - Numero da rodada;
- minutes - Tempo da partida;
- w_ace - Quantidade de ace do vencedor;
- w_df - Quantidade de duplas faltas do vencedor;
- w_svpt - Quantidade de pontos jogados no serviço do vencedor;
- w_1stIn - Quantidade de pontos dentro com o primeiro saque do vencedor;
- w_1stWon - Quantidade de pontos vencidos com o primeiro saque do vencedor;
- w_2ndWon - Quantidade de pontos vencidos com o segundo saque pelo vencedor;
- w_SvGms - Quantidade de games servido pelo vencedor;
- w_bpSaved - Quantidade de pontos de quebra salva pelo vencedor;
- w_bpFaced - Quantidade de pontos de quebra pelo vencedor;
- l_ace - Quantidade de ace do perdedor;
- l_df - Quantidade de duplas faltas do perdedor;
- l_svpt - Quantidade de pontos jogados no serviço do perdedor;
- l_1stIn - Quantidade de pontos dentro com o primeiro saque do perdedor;
- l_1stWon - Quantidade de pontos vencidos com o primeiro saque do perdedor;
- l_2andWon - Quantidade de pontos vencidos com o segundo saque pelo perdedor;
- l_SvGms - Quantidade de games servido pelo perdedor;
- l_bpSaved - Quantidade de pontos de quebra salva pelo perdedor;
- l_bpFaced - Quantidade de pontos de quebra pelo perdedor.

3.2 PRÉ-PROCESSAMENTO DOS DADOS

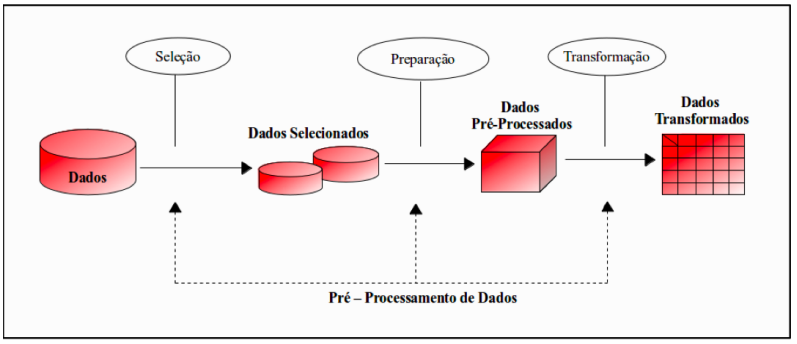
Esta é uma etapa fundamental, se o conjunto de dados não está organizado, não iremos produzir resultados ótimos, resultando em análises e modelagens não válidas. Diante disso, um pré-processamento confiável é essencial para certificar a qualidade e confiabilidade.

O estágio de pré-processamento dos dados consiste em uma série de ações que abrangem preparação, organização e estruturação dos dados para que os mesmos estejam nos padrões permitidos em cada algoritmo de AM. Refere-se de uma etapa fundamental que precede a realização de análises e predições (CAMILO, SILVA, 2009).

Apesar de que o entendimento dos dados consiste em uma etapa separada do pré-processamento de dados, é de grande importância entender os dados para realizar o pré-processamento. Através das análises, podemos identificar possíveis problemas nos dados, verificar como os dados estão distribuídos e realizar uma melhor transformação.

Portanto, separamos os dados em dois grupos, a base de treinamento e a base de testes. Os dados de treinamento, que representam cerca de 80% do total da base de dados, serão explicitados anteriormente ao algoritmo de ML para construção do modelo. Os dados de teste, que representam cerca de 20% do total da base de dados, serão explicitados ao modelo após a sua construção, permitindo que sua real eficiência seja apurada.

Figura 3.1 - Esquema do pré-processamento de Dados.



Fonte:https://cepein.femanet.com.br/BDigital/arqTccs/1011330122.pdf

3.3 LIMPEZA DOS DADOS

Sobre a base de dados, pode abranger inúmeras informações irrelevantes, duplicadas ou ausentes (missing values). Trabalhar com isso pode prover inconsistência nos resultados , logo a limpeza da base é crucial, abrangendo toda a manipulação, identificação e exclusão de outliers, elaboração de dados faltantes e a diminuição da inconsistência da sua base de dados. Portanto, feita a manipulação, a base está pronta para a próxima etapa.

Para uma análise de dados mais eficiente, foi excluída as temporadas de 2000 até 2010, pois no circuito da ATP há uma grande rotatividade, ou seja, os jogadores que hoje pertencem ao top 100 (Os 100 melhores rankeados), daqui 5 anos, provavelmente serão outros 100, poucos conseguem se manter por um longo período entre os melhores. Além dessas, foi excluída também a temporada de 2017, pois as informações apresentaram inconsistência, visto que continha apenas metade dos jogos da temporada regular.

3.4 CRIAÇÃO DE VARIÁVEIS

A criação de novas variáveis estabelece-se como o processo de produzir variáveis a partir da sua base de dados inicial, exemplo: utilizando a variável "l_1stWon" (Quantidade de pontos dentro com o primeiro saque do vencedor) e dividindo pela variável "w_1stIn" (Quantidade de pontos vencidos com o primeiro saque do vencedor) temos um novo atributo, "w_1serve", que representa uma taxa de eficiência do primeiro saque. Variáveis novas, caracteriza uma ótima performance do modelo, logo, o conhecimento na área auxilia na construção.

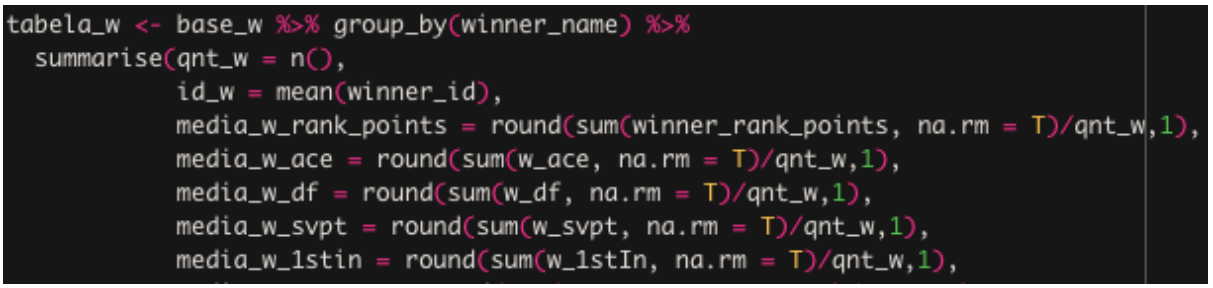
3.5 ORGANIZAÇÃO DO BANCO DE DADOS

A Base de dados concedida não é apropriada para solucionar o problema, tendo em vista a necessidade de partir dos dados históricos para prever o resultado, isto significa, informações de partidas anteriores para predição da atual.

Para solucionar isto, foi realizada uma organização, onde foram produzidas novas variáveis para o conjunto, que podem ser divididas em dois grupos de informações. O primeiro grupo contém características do vencedor (Winner) e o segundo, contém características do perdedor (Loser).

Foram construídas diversas variáveis, tais como a médias dos seguintes atributos: saque, duplas faltas, pontos com primeiro saque, pontos com segundo saque e pontos de quebra de saque. Após a construção desses dois grupos, é necessário unificar todas as informações em apenas uma base de dados.

Figura 3.2 - Exemplo de sumarização dos atributos no R.



```
tabela_w <- base_w %>% group_by(winner_name) %>%
  summarise(qnt_w = n(),
            id_w = mean(winner_id),
            media_w_rank_points = round(sum(winner_rank_points, na.rm = T)/qnt_w,1),
            media_w_ace = round(sum(w_ace, na.rm = T)/qnt_w,1),
            media_w_df = round(sum(w_df, na.rm = T)/qnt_w,1),
            media_w_svpt = round(sum(w_svpt, na.rm = T)/qnt_w,1),
            media_w_1stin = round(sum(w_1stIn, na.rm = T)/qnt_w,1),
```

Fonte: Autor

3.6 SELEÇÃO DE VARIÁVEIS

Posteriormente a solidificação da base de dados, contendo as principais variáveis, é necessário averiguar se há variáveis correlacionados e quais são as variáveis de maior relevância, pois o objetivo da elaboração de um modelo ótimo é ter o mesmo resultado com o menor número de variáveis possíveis. Portanto, foram retiradas as variáveis correlacionadas da interação. Porém houve variáveis, que apesar de ser correlacionadas, tem grande importância no resultado final, por exemplo, a variável "quantidade de ace" é correlacionada com a “quantidade de pontos vencido com primeiro saque”, pois ambas, indicam o ponto vencedor com primeiro saque, porém ao retirar alguma delas diminui-se a quantidade de informações adquiridas dos jogadores, induzindo o modelo a atingir menos sucesso, com um acerto reduzido.

4. METODOLOGIA

4.1 BASE DE DADOS

A base de dados representa todas as partidas realizadas nas temporadas 2010 a 2016 do circuito da ATP “Association of Tennis Professionals”. Cada linha dessa base representa uma partida, onde contém informações sobre o jogo em geral, como por exemplo: Data; Tipo de quadra; Tempo de duração da partida. Contém também informações sobre o jogador vencedor e sobre o jogador perdedor, como por exemplo: Mão de dominância; Idade; País de origem. A base de dados completa contém todas as variáveis, totalizando 48 variáveis e 20872 partidas.

Para solucionar o problema, será definida como vitória ou derrota, a classificação do ranking, ou seja, em um confronto entre jogador A e jogador B, se o jogador A, melhor ranqueado, ganhar do jogador B, será contado como vitória, caso contrário, será considerado como derrota. Exemplo: Roger Federer está na 1ª colocação do ranking e irá jogar com Nick Kyrgios, que mantém a 15ª posição do ranking, se Roger Federer ganhar, será considerado vitória, se Roger Federer perder, contará como derrota. Sendo assim, para compreender a variável resposta, segue a distribuição das classes a serem previstas:

Tabela 4.1: Total de partidas da base de dados.

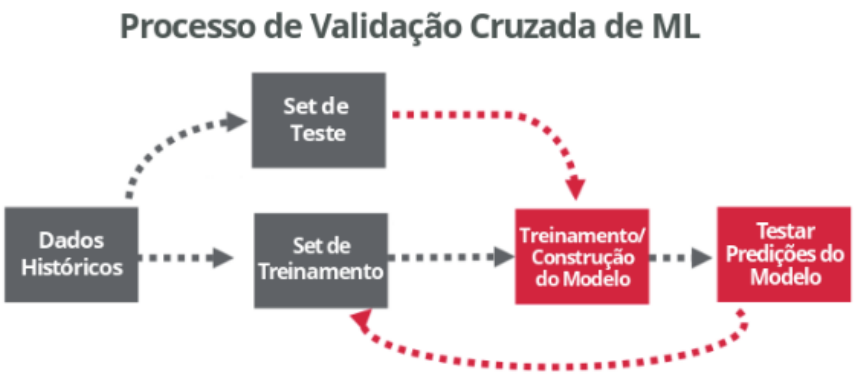
	DERROTA	**VITÓRIA**	**TOTAL**
PARTIDAS	6933	13939	20872
FREQUÊNCIA (%)	33,22%	66,78%	100%

Poderíamos utilizar o conjunto inteiro dos dados para a construção de um modelo preditivo, mas dessa forma não saberíamos o real desempenho deste modelo. Logo, para conseguir avaliar sua eficiência, é fundamental realizar testes, utilizando dados diferentes dos que foram apresentados em sua origem.

Nesta etapa, foi necessário separar os dados existentes em dois grupos, a base de treinamento e a base de testes. Os dados de treino serão introduzidos aos

algoritmos para a construção dos modelos, estes dados significam aproximadamente 80% do total. Para os dados de teste, serão introduzidos ao modelo após a sua criação, para que sua eficácia seja verificada, estes dados significam aproximadamente 20% do total.

Figura 4.1: Fluxograma do processo de validação dos modelos.



Fonte: https://www.direction.com.br/audit

Como o estudo é fundamentado no uso de informações de partidas passadas para prever o resultado futuro, é inviável testar os modelos que os dados de treino são dados após aos dados de teste, logo, o fator tempo é relevante nessa separação e não poderá haver partidas de treino com datas superiores as partidas de teste. Portanto, o circuito da ATP de 2010 até 2014 equivale aos dados de treino, com 14973 partidas e o circuito de 2015 até 2016 os dados de testes, com 5899 partidas.

4.2 RESULTADOS

Para avaliar o real desempenho dos modelos de classificação, foram empregadas as métricas citadas na seção 2.4, em que, serão calculadas a partir da matriz de confusão.

O próximo passo é combinar os resultados das métricas, por cada classe em um único valor, para verificar o real desempenho de cada classificador. Existem diversas maneiras de realizar isso, porém foram escolhidas somente duas dessas

medidas para a análise. A primeira é trazer uma média aritmética das pontuações por classe, chamada de pontuação macro. Ao utilizar a pontuação macro, são atribuídos pesos iguais para cada classe, portanto, para os casos de classes desbalanceadas, não é recomendado. Logo, a segunda maneira foi utilizar a média ponderada das pontuações por classe, em que os pesos de cada uma delas são atribuídos pelo tamanho amostral de cada classe.

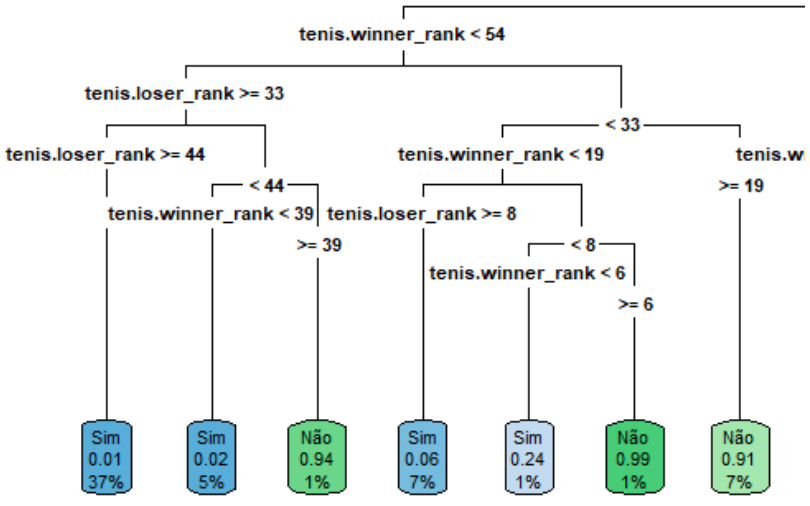
Foram escolhidos 2 algoritmos de classificação, descritos na seção 2.3, para previsão dos resultados. Os algoritmos utilizados serão especificados:

4.2.1 Árvore de decisão (AD)

Para construção do modelo de árvore de decisão foi utilizado o pacote “rpart” do software R. Recordando, uma AD é um modelo que particiona recursivamente um conjunto de dados em subconjuntos cada vez mais “puros”, em relação a uma variável de resposta binária.

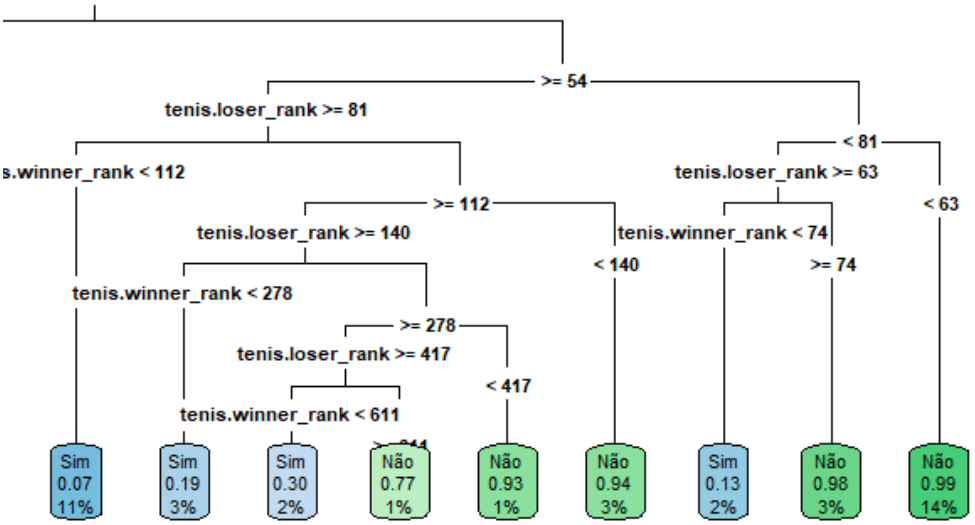
A AD foi treinada com 32 variáveis selecionadas e o seu real desempenho foi checado pela base de teste, logo, foram obtidos os seguintes resultados:

Figura 4.2 - Lado esquerdo da Árvore



Fonte: Autor

Figura 4.3: Lado direito da Árvore



Fonte: Autor

Tabela 4.2: Matriz de confusão (Árvore de decisão).

	ŷ = 0	**ŷ = 1**	**TOTAL REAL**
y = 0	1069	877	**1946**
y = 1	1848	2105	**3953**
TOTAL PREVISTO	**2917**	**2982**	**5899**

Na tabela 4.2 observa-se a representação da matriz de confusão da AD. Os valores na diagonal principal representam a quantidade de acertos do modelo. As linhas representam os valores reais e as colunas os valores previstos.

Tabela 4.3: Resultados de cada classe (Árvore de decisão).

Classe	**Precisão**	**Revocação**	**F1-Score**
Derrota	54,9%	36,6%	43,9%
Vitória	53,2%	70,5%	60,7%

A precisão do modelo está relacionada ao quanto o classificador acerca dos resultados que previu como positivo, a revocação é a proporção do quanto o modelo previu como positivo do total de positivos da base, e por fim, o f1 score que é dado pela média harmônica entre essas duas métricas citadas acima.

Tabela 4.4: Resultados das métricas (Árvore de decisão).

Classificador	Acurácia	F1 Ponderado	F1 Macro	Precisão Ponderada	Precisão Macro	Revocação Ponderada	Revocação Macro
AD	53,6%	55,6%	52,3%	53,1%	54%	60,3%	53,5%

O desempenho do modelo em relação a taxa de acerto é dado pela acurácia de 53.8%. Portanto, a tabela 4.4 indica que o modelo tem mais facilidade em prever vitórias do derrotas, dado que é a classe majoritária da base de dados, próximo de 66,7%. Para derrotas, mostra que a predição do modelo está ruim, indicado pelo F1-Score baixo, próximo de 43,9%.

4.2.2 Regressão Logística (RL)

Um dos modelos lineares generalizados mais utilizados é o modelo de regressão logística binária, onde a variável resposta do modelo tem distribuição de Bernoulli (ou Binomial) e a função de ligação é a função logística. Esse modelo, vai associar a variável dependente à variável independente e recebe a denominação de Modelo de Regressão Linear Simples (MRLS).

Para construção do modelo de regressão logística foi utilizada a função “glm” do software R e os resultados são mostrados a seguir:

Figura 4.4 - Tabela de resíduos RL

```
Deviance Residuals:
      Min         1Q     Median         3Q        Max
-3.565e-04  -2.000e-08   2.000e-08   2.000e-08   3.661e-04
```

Fonte: Autor

Figura 4.5 - Variáveis para estimação

```
Coefficients:
                Estimate Std. Error
(Intercept)     2.250e+02  5.346e+04
surfaceGrass   -7.848e-01  1.301e+03
surfaceHard    -2.722e-01  6.871e+02
tourney_levelD -4.383e+01  2.522e+05
tourney_levelF  3.748e-01  1.518e+03
tourney_levelG  3.537e+01  2.829e+05
tourney_levelM  2.314e-01  6.924e+02
winner_rank    -4.639e+02  7.014e+03
loser_rank      4.644e+02  7.027e+03
best_of        -1.349e+01  1.107e+05
minutes         2.603e-01  8.972e+02
w_ace          -2.827e-02  4.504e+02
w_df           -8.151e-02  3.588e+02
w_svpt         -5.925e-01  3.013e+03
w_1stIn         3.113e-01  1.777e+03
w_1stWon        7.727e-01  2.733e+03
w_2ndWon        7.085e-01  1.463e+03
w_SvGms        -1.326e+00  2.499e+03
w_bpSaved      -7.321e-01  1.774e+03
w_bpFaced       8.067e-01  2.161e+03
l_ace           4.522e-02  4.185e+02
l_df           -1.499e-01  3.696e+02
l_svpt         -5.312e-01  2.681e+03
l_1stIn        -3.731e-01  1.583e+03
l_1stWon        1.509e+00  2.860e+03
```

Fonte: Autor

Figura 4.6 - Coeficientes de regressão

```
Coefficients:
   (Intercept)    surfaceGrass     surfaceHard  tourney_levelD  tourney_levelF
     225.03542        -0.78482        -0.27218       -43.83080         0.37480
tourney_levelG  tourney_levelM     winner_rank     loser_rank         best_of
      35.37491         0.23143       -463.94675       464.43092       -13.49336
       minutes           w_ace            w_df          w_svpt         w_1stIn
       0.26032        -0.02827        -0.08151        -0.59248         0.31125
      w_1stWon        w_2ndWon         w_SvGms       w_bpSaved       w_bpFaced
       0.77272         0.70853        -1.32604        -0.73208         0.80673
         l_ace            l_df          l_svpt         l_1stIn        l_1stWon
       0.04522        -0.14986        -0.53117        -0.37309         1.50872
      l_2ndWon         l_SvGms       l_bpSaved       l_bpFaced
       0.36422        -1.17887        -1.02429         1.20568
```

Fonte: Autor

Quadro 4.5 - Matriz de confusão (Regressão logística).

	ŷ = 0	**ŷ = 1**	**TOTAL REAL**
y = 0	1099	847	**1946**
y = 1	1874	2079	**3953**
TOTAL PREVISTO	**2973**	**2926**	**5899**

Observa-se a representação da matriz de confusão da RG. Os valores na diagonal principal representam a quantidade de acertos do modelo. As linhas representam os valores reais e as colunas os valores previstos.

Quadro 4.6 - Resultados de cada classe (Regressão logística).

Classe	**Precisão**	**Revocação**	**F1-Score**
Derrota	56,4%	36,9%	44,6%
Vitória	52,5%	71%	60,4%

A precisão do modelo está relacionada ao quanto o classificador acerca dos resultados que previu como positivo, a revocação é a proporção do quanto o modelo previu como positivo do total de positivos da base, e por fim, o f1 score que é dado pela média harmônica entre essas duas métricas citadas acima.

Quadro 4.7 - Resultados das métricas (Regressão logística).

Classificador	**Acurácia**	**F1 Ponderado**	**F1 Macro**	**Precisão Ponderada**	**Precisão Macro**	**Revocação Ponderada**	**Revocação Macro**
RL	53,9%	54,3%	52,5%	53,6%	54,4%	60,7%	53,9%

O desempenho do modelo em relação a taxa de acerto é dado pela acurácia de 53.9%. Portanto, indica que o modelo tem mais facilidade em prever vitórias, dado que é a classe majoritária da base de dados, próximo de 66,7%. Para derrotas, mostra que a predição do modelo está ruim, indicado pelo F1-Score baixo, próximo de 44,6%.

4.3 COMPARAÇÃO DE DESEMPENHO DOS MODELOS

Será analisado e comparado o real desempenho dos classificadores, baseando-se nas métricas apresentadas na seção 2.5. Dizer que um modelo é melhor que o outro, se comparado com os resultados obtidos pelas métricas, não é a melhor alternativa. Visto que essas diferenças verificadas, muitas vezes, podem não ser significativas.

O quadro a seguir, traz informações dos resultados da métrica F1 de todos os algoritmos de classificação utilizados, separando-os por classes (vitória e derrota). A métrica F1 score é a mais representativa para esse caso pois contempla informações das métricas de precisão e revocação.

Tabela 4.8 - Resultados F1-Score por cada algoritmo.

Classificador	**Derrota**	**Vitória**
F1 - Árvore de decisão	43,9%	60,7%
F1 - Regressão logística	44,6%	60,4%

Analisando os resultados da tabela 16, notou-se que os modelos obtiveram dificuldades na predição de derrotas, aproximadamente 44%, isso dá pela alta rotatividade de jogadores, isto significa que os modelo tendem a seguir uma média. Percebe-se também, que os modelos não tiveram dificuldade de prever vitórias, ambos com aproximadamente 60%. Além disso, observou-se que o primeiro modelo, de árvore de decisão, obteve um rendimento menor, mesmo que pouco, do que segundo modelo.

Agora, com o intuito de fazer uma análise de desempenho geral dos modelos, o quadro a seguir, representa a taxa de acerto de cada um dos algoritmos utilizados junto com as médias ponderadas das métricas comentadas no quadro.

Tabela 4.9 - Comparação dos algoritmos por cada métrica.

Classificador	**Acurácia**	**F1 Ponderado**	**Precisão Ponderada**	**Revocação Ponderada**
AD	53,6%	55,6%	53,1%	60,3%
RL	53,9%	54,3%	53,6%	60,7%

Na tabela 4.9, podemos notar que o primeiro modelo, AD, possui o pior desempenho visando somente a acurácia como métrica de avaliação, mas, ao utilizar a métrica F1 ponderado, mostra que esse mesmo modelo possui o melhor desempenho. Portanto, notamos a grande importância da verificação e da escolha certa das métricas de avaliação para o sucesso do projeto de classificação.

Verificando as métricas separadamente, observa-se que o modelo que obteve melhor desempenho em relação a taxa de acerto foi o de árvore de decisão, com 60,7%. O algoritmo de classificação que obteve a maior precisão ponderada e revocação ponderada foi de regressão logística, respectivamente, 53,6% e 60,7%.

5. CONCLUSÃO

Analisando o resultado dos classificadores, obtiveram desempenho regulare, visto que em média, houve uma acurácia de 53%, ou seja, a cada 100 partidas, 53 foram previstas com resultado certo, também observou-se que o algoritmo de Regressão Logística apresentou melhor desempenho, com (53.6%) e o classificador que apresentou o maior eficiencia em relação ao F1 ponderado foi, Árvore de Decisão com (55.6%). Por isso, não é possível assumir que um modelo obteve desempenho melhor que o outro, ambos tiveram suas particularidades. Logo, a solução ideal deveria ser a junção desses modelos, onde cada um ficaria responsável por métricas diferentes. Tal como, o algoritmo de regressão logística para prever vitórias e não vitórias, pois obteve melhor acurácia, e o algoritmo de árvore de decisão para prever derrotas e não derrotas, pois obteve o melhor F1 score

Este trabalho, é parte de um modelo sujeito a muitas irregularidades, tendo em vista que é inviável quantificar todos os fatores de influência nos jogos, como por exemplo, sua condição psicológica, contusões, o ambiente, o condicionamento físico, falta de ritmo de jogo após ausência do circuito e utilização de novos equipamentos.

Para aprimorar o estudo, devem ser tomadas medidas que possam trazer melhoria na qualidade de predição, como adicionar novos atributos sobre as partidas e refazer a construção das features. A performance dos tenistas altera-se em relação aos diversos tipos de quadra, inserir uma variável que explique isso poderá trazer ótimos resultados. A premiação em dinheiro obtida representa com um peso maior as vitórias mais importantes, visto que os principais torneios oferecem premiações mais generosas e maiores valores monetários são concedidos em vitórias nas fases mais avançadas dos torneios.

Outro ponto interessante, seria utilizar um modelo de oponentes comuns, com jogadores de épocas distintas, para poder comparar usando as métricas vindas do modelo descrito acima, ou seja, simular um jogo que seria impossível de acontecer.

REFERÊNCIAS

ASSOCIATION of Tennis Professionals Matches: ATP tournament results from 2000 to 2017. *In*: **Kaggle**: Your Machine Learning and Data Science Community. Disponível em: https://www.kaggle.com/datasets/gmadevs/atp-matches-dataset. Acesso em: 2 nov. 2022.

CAMILO, Cássio Oliveira. **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas.** Orientador: João Carlos da Silva. 2009. 29 p. Relatório Técnico (Mestrado em Ciência da Computação) - Instituto de Informática, Universidade Federal de Goiás, Goiás, 2009. Disponível em: https://ww2.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf. Acesso em: 5 jan. 2023.

CORRAR, Luiz J.; THEÓPHILO, Carlos Renato. **Pesquisa operacional para decisão em contabilidade e administração:** Contabilometria. São Paulo: Atlas, 2004.

HOSMER, David W.; LEMESHOW, Stanley; STURDIVANT, Rodney X. Applied **Logistic Regression:** Third Edition. 3ª ed. John Wiley & Sons, Inc., 2013. 510 p. Disponível em: https://onlinelibrary.wiley.com/doi/book/10.1002/9781118548387. Acesso em: 5 jan. 2023.

MITCHEL, Tom M. **Machine Learning.** [*S. l.*: *s. n.*], 1997. 432 p. Disponível em: https://www.cin.ufpe.br/~cavmj/Machine%20-%20Learning%20-%20Tom%20Mitchell.pdf. Acesso em: 5 jan. 2023.

MORETIN, Pedro A.; SINGER, Júlio M. **Introdução à Ciência de Dados: Fundamentos e Aplicações.** 2019. 236 p. Trabalho de Conclusão de Curso (Bacharelado em Estatística) - Universidade de São Paulo, São Paulo, 2019. Disponível em:

https://curso-r.github.io/main-regressao-linear/referencias/Ci%C3%AAncia%20de%20Dados.%20Fundamentos%20e%20Aplica%C3%A7%C3%B5es.%20Vers%C3%A3o%20parcial%20preliminar.%20maio%20Pedro%20A.%20Morettin%20Julio%20M.%20Singer.pdf. Acesso em: 5 jan. 2023.

REGRAS do Tênis. //Regras dos Esportes, 2017. Disponível em: http://www.regrasdosesportes.com/regras-do-tenis/. Acesso em: 5 jan. 2023.

SAMUELS, Mark. **Big data is serving top tennis players a match-winning advantage.** *In*: ZDNET/ Business. 20 nov. 2019. Disponível em: https://www.zdnet.com/article/big-data-is-serving-top-tennis-players-a-match-winning-advantage/. Acesso em: 5 jan. 2023.

SILVA, Wesley Vieira da *et al*. AVALIAÇÃO DA ESCOLHA DE UM FORNECEDOR SOB CONDIÇÃO DE RISCOS A PARTIR DO MÉTODO DE ÁRVORE DE DECISÃO. **Revista de Gestão USP**, São Paulo, ano 2008, v. 15, n. 3, p. 77-94, jul-set 2008.

SOUTO-MAIOR, Cesar Duarte. **Pesquisa Operacional.** 3ª ed. Universidade Federal de Santa Catarina, 2014. 94 p. Disponível em: http://arquivos.eadadm.ufsc.br/somente-leitura/EaDADM/UAB_2014_2/Modulo_4/Pesquisa_Operacional/Pesquisa%20operacional%203ed.pdf. Acesso em: 4 jan. 2023.

TURKSON, Regina Esi; BAAGYERE, Edward Yeallakuor; WENYA, Gideon Evans. A machine learning approach for predicting bank credit worthiness. **2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR)**, Lodz, Poland, Set 2016. Disponível em: https://www.researchgate.net/publication/309151303_A_machine_learning_approach_for_predicting_bank_credit_worthiness. Acesso em: 5 jan. 2023.