UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

VÍTOR BALDEZ MACIEL

Predição de resultados de partidas do Campeonato Brasileiro de Futebol

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação

Orientador: Prof[a]. Dr[a]. Renata de Matos Galante

Porto Alegre
2024

AGRADECIMENTOS

Primeiramente eu agradeço imensamente à minha família. Meu pai, Gilberto Luiz Vieira Maciel, minha mãe, Sônia Regina Baldez Maciel, minhas irmãs, Fernanda Baldez Maciel e Isadora Baldez Maciel e à minha namorada, Bruna Marques Rosa e sua família que me apoiaram durante todos esses anos e foram fundamentais para toda a minha formação, me tranquilizando e transmitindo amor tanto nos momentos felizes, quanto nos momentos de tensão dentro da faculdade.

Agradeço também a todos os amigos que fiz antes de iniciar minha jornada na faculdade e aos que fiz durante minha graduação que me foram importantes tanto intelectualmente sempre estudando e fazendo trabalhos juntos nas cadeiras cursadas, quanto sentimentalmente cuja amizade foi e continua sendo muito importante para mim.

Reservo agradecimentos a todos os professores e funcionários de diversas funções que conheci na UFRGS, que me ensinaram muito tecnicamente questões da minha graduação e sempre me trataram com muita cordialidade, pessoas que levarei boas lembranças dessa experiência.

Agradeço também ao professor Anderson Rocha Tavares, que nunca tive aula durante a graduação, mas que se disponibilizou prontamente com o meu pedido de auxílio com algumas questões relacionadas à aprendizado de máquina que tive e foi muito importante para a conclusão deste trabalho.

Por último e não menos importante, agradeço imensamente à professora Renata de Matos Galante, orientadora deste trabalho, que se disponibilizou desde o início na cadeira de dados que fiz com ela a me orientar e se esforçou ao máximo para que este trabalho fosse concluído da melhor maneira possível. Além disso, também sempre me acalmando e ajudando.

Muito obrigado.

RESUMO

A obtenção de lucro em apostas esportivas não é uma tarefa fácil de se executar. Normalmente o que vemos nas casas de apostas são apostadores casuais, que mais perdem dinheiro do que lucram com essa atividade. A proposta deste trabalho é a criação de um modelo de predição de apostas para ser utilizado no Campeonato Brasileiro de Futebol. Primeiramente é mostrada toda a etapa de pré-processamento de dados, que foram retirados do site *Sofascore* e representam uma série de estatísticas dos jogos envolvendo ambos os times envolvidos na partida. Foi aplicado um algoritmo de aprendizado não supervisionado também como parte do pré-processamento para dividir os dados em subconjuntos e verificar se a predição obtém alguma vantagem dessa divisão. Depois foram aplicados 7 algoritmos de aprendizado supervisionado, sendo 6 de classificação e 1 de regressão. Os 2 melhores modelos de classificação foram selecionados a partir de 3 métricas de desempenho para serem comparados entre si e com o modelo de regressão realizando apostas nos jogos disputados no Campeonato Brasileiro de 2023. Foi verificado nos 3 modelos o lucro, ou prejuízo e o acerto proporcionado por eles em uma simulação real de apostas esportivas usando a base de dados pré-processada e testada anteriormente.

Palavras-chave: Aprendizado de máquina. predição.

Prediction of match results Brazilian Football Championship

ABSTRACT

Making a profit in sports betting is not an easy task to perform. Normally what we see in betting houses are casual bettors, who lose more money than they profit from this activity. The purpose of this work is to create a betting prediction model to be used in the Brazilian Football Championship. Firstly, the entire data pre-processing stage is shown, which was taken from the Sofascore website and represents a series of game statistics involving both halves involved in the match. An unsupervised learning algorithm was also applied as part of the pre-processing to divide the data into subsets and check whether the prediction obtains any advantage from this division. Then, 7 supervised learning algorithms were applied, 6 for classification and 1 for regression. The 2 best classification models were selected based on 3 performance metrics to be compared with each other and with the regression model for bets on games played in the 2023 Brazilian Championship. The 3 profit or loss models and the accuracy provided were selected. by them in a real sports betting simulation using a pre-processed and previously tested database.

Keywords: machine learning, prediction.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	*Machine Learning* - Aprendizado de máquina
LR	*Logistic Regression* - Regressão Linear
KM	*K-means* - K Médias

SUMÁRIO

1 INTRODUÇÃO

O engajamento com as apostas esportivas no Brasil tem crescido muito nos últimos anos. Em uma pesquisa feita pelo (GABRIEL; SALDAÑA, 2024), foi registrado que 15% dos brasileiros fazem ou já fizeram apostas online, atingindo majoritariamente o público de homens e jovens. Desses, 30% do público entre 16 e 24 anos já praticou essa atividade (GABRIEL; SALDAÑA, 2024), sendo o dobro da média registrado para todo o país. Fazer apostas esportivas é uma atividade que acaba se tornando bem satisfatória, adicionando mais um atrativo para o público aficionado por esportes.

No Brasil, pela paixão existente pela população ao futebol, faz esse esporte ter uma grande quantidade de apostadores. No entanto, existe um grande problema relacionado a isso. Normalmente, quem faz apostas não tem uma estratégia para acertar os resultados, muitas vezes a aposta fica muito relacionada à questão da sorte, ou o clássico 'sentimento' do resultado que ocorrerá no final da partida apostada e isso acarreta em uma considerável quantidade de pessoas obtendo resultados financeiramente não satisfatórios. O gasto mensal médio dentre os apostadores de aposta atinge o valor de 263 reais (GABRIEL; SALDAÑA, 2024), um valor bem expressivo levando em conta que representa 20% do salário mínimo de 2023. Além disso, 50% dos apostadores afirmam que perderam mais dinheiro do que ganharam (GABRIEL; SALDAÑA, 2024). Levando em conta o prejuízo que uma parte dos apostadores acabam tendo na realização dessa atividade e o fato de ser possível encontrar padrões relacionados às apostas e não depender estritamente da sorte, motivo que levou o governo a não considerar esse um jogo de azar (GRANCHI, 2023), surgiu a ideia deste trabalho relacionada a predição de partidas de futebol do Campeonato Brasileiro.

Existem algumas técnicas para fazer predição e uma parte delas foi utilizada neste trabalho. As técnicas usadas são: aprendizado supervisionado que visam um atributo alvo a partir de seus atributos dependentes e baseiam a predição nesse atributo alvo. O aprendizado supervisionado se divide em classificação onde o atributo alvo é categórico e regressão onde o atributo alvo é contínuo. A outra técnica utilizada foi aprendizado não supervisionado que não tem atributo alvo e visa achar padrões, normalmente dividindo os dados em subgrupos semelhantes.

O objetivo desse trabalho é fazer predições das partidas utilizando algoritmos de Aprendizado de Máquina(AM), selecionar os 2 melhores modelos de classificação e depois testá-los em uma simulação real de apostas junto com o modelo de regressão, verifi-

cando o resultado das partidas do Campeonato Brasileiro de Futebol de 2023 e o possível lucro, ou prejuízo obtido de acordo com as cotações dadas às partidas.

O restante deste trabalho está organizado da seguinte forma: o Capítulo 2 apresenta os conceitos e tecnologias envolvidas para a aplicação dos modelos realizados no trabalho. O Capítulo 3 mostra os trabalhos relacionados a predição de partidas usando AM. O Capítulo 4 explica a metodologia que serviu de base para o desenvolvimento do trabalho. O Capítulo 5 analisa os resultados obtidos a partir dos modelos criados e simulação em casas de apostas. Por fim, o Capítulo 6 revisa o desenvolvimento do trabalho, resultados positivos e negativos, além de possibilidades futuras que podem ser desenvolvidas para expansão das aplicações realizadas.

2 CONCEITOS E TECNOLOGIAS UTILIZADAS

Neste capítulo, são apresentados os conceitos e as tecnologias que foram utilizados para o cálculo de predição de resultados de partidas de futebol.

2.1 Futebol no Brasil

Existem muitas teorias a respeito do esporte que deu origem ao futebol, mas o que mais encontramos relatos como precursor do futebol moderno é o Tsu chu, um esporte que tem como tradução literal 'chuta bola', que era jogado há aproximadamente 3000 a.C. na China e que usava uma bola de couro preenchida com pelos de animais (FIFA, 2020). O futebol como conhecemos hoje foi regulamentado na Inglaterra em 1863 pela Football Association, uma entidade que existe até hoje e que ainda organiza o esporte neste país (MOSCA, 2006). Obviamente, desde os primórdios, muitas regras mudaram, foram criadas e novas tecnologias implementadas para melhorias de qualidade do esporte, como o mais recente e polêmico 'VAR', ou árbitro de vídeo, mas a essência do jogo permanece até hoje como no princípio. No Brasil, o futebol começou a se popularizar com Charles Miller, brasileiro filho de pai escoces e mãe brasileira que após passar um tempo estudando na Inglaterra voltou ao Brasil com bolas e regras sobre o esporte, passou a ensinar seus conhecidos a jogar e organizar partidas amadoras (MILLS, 2005). Para se ter um panorama da força que o futebol teve no Brasil, o remo era um esporte muito praticado antes da chegada do futebol. Prova disso, é podermos ver vários times que foram criados a partir do remo e hoje são clubes com o foco principal no futebol e que ainda possuem referências no escudo e até mesmo no nome relacionadas ao antigo esporte principal, como, por exemplo, três dos quatro maiores clubes do estado do Rio de Janeiro que ainda levam a palavra 'Regatas' no nome (ROCHA, 2008). O futebol brasileiro foi ganhando notoriedade perante o surgimento de grandes jogadores, inicialmente com o dito 'Rei do Futebol' Edson Arantes do Nascimento, o Pelé, que junto de Manoel Francisco dos Santos, o Garrincha, capitaneou tecnicamente uma seleção brasileira que veio a vencer três Copas do Mundo no período de 14 anos(1958-1962) e colocou o país nos holofotes do mundo inteiro. A habilidade individual dos jogadores brasileiros continuou chamando atenção e contribuindo na revelação de inúmeros jogadores com grande técnica que levou o país a conquistar mais duas Copas do Mundo, culminando em um total de cinco, sendo um recorde ainda não batido até os dias atuais (COELHO, 2018). Com uma história tão

longínqua quanto o masculino, o futebol feminino passou por muitas dificuldades para ser reconhecido, até mesmo enfrentando um decreto-lei de 1941 assinado por Getúlio Vargas que proibia a prática de esportes por mulheres no país, que foi revogado apenas em 1983 (WESTIN, 2023). Após essa triste marca, tivemos alguns destaques, o principal deles com a revelação de Marta Vieira da Silva que foi muito importante para o crescimento da modalidade feminina no país e que é considerada por muitos a 'Rainha do futebol'. Felizmente, nos últimos anos tivemos outro grande marco, com a obrigatoriedade imposta pela Confederação Brasileira de Futebol em 2019 de todos os times masculinos da série A do campeonato brasileiro terem também um time feminino e em 2023 o anúncio que a partir de 2027 essa lei será estendida também para todas as quatro divisões do futebol brasileiro (SIMõES, 2023). Por fim, todos os fatores citados contribuíram para a disseminação de uma famosa frase reconhecida por muitos que fala que 'o Brasil é o país do futebol', mesmo não sendo o país de origem desse esporte.

2.2 Apostas Esportivas

Historiadores afirmam que as apostas chegaram ao Brasil ainda no período colonial, com a chegada dos europeus trazendo jogos de cartas, dados entre outros. No século XVIII surgiram as primeiras casas de apostas, que normalmente eram mais utilizadas por pessoas de poder econômico mais elevado. Nos últimos anos o mercado de apostas esportivas vem se consolidando no país, com os mais diversos mercados disponíveis como futebol, vôlei, basquete, cassino e até mesmo jogos online como League of Legends e Counter Strike (IBJR, 2023). No contexto do futebol, esse mercado cresceu de maneira muito significativa nos últimos anos, com uma grande quantidade de times brasileiros sendo patrocinados por casas de apostas. Em 2023, dos vinte clubes da série A do Campeonato Brasileiro, dezenove são patrocinados por alguma empresa desse ramo, sendo doze deles com patrocínio máster. Além de todos os times da série B, várias competições profissionais oficiais de futebol e emissoras de televisão também tem essas empresas como parceiras. Em valores a estimativa é que esse setor investe 3,5 bilhões de reais anuais em patrocínios no Brasil dentre todos os segmentos (MAGATTI, 2023). Apesar de todo esse investimento, ainda existe um entrave em relação às apostas. Ainda que em 2018 foi decretada a Lei nº 13.756 que criou a modalidade lotérica Apostas de Quota Fixa que confere a autorização da atividade das apostas esportivas em território nacional, ainda não existe uma regulamentação para esse novo mercado. Essa desregulamentação causa

insegurança para os consumidores, pois essas empresas, na sua maioria estrangeiras, não tem uma sede e nem são registradas no Brasil, tirando o direito das pessoas de recorrer ao Código de Defesa do Consumidor por qualquer desacordo relacionado a prática das apostas (PÓVOA et al., 2023). Felizmente, em 2023 foi protocolado o projeto de Lei nº 3626 que ainda passa por processos burocráticos, mas que propõe a regulamentação das apostas esportivas e passa por vários pontos que serão benéficos para esse mercado, dando maior segurança aos consumidores (SENADO, 2023) e que tende a consolidar as apostas no país.

2.3 Odds nas casas de apostas

O que faz as casas de apostas girarem, popularmente chamadas de odds, são as cotações dadas aos eventos disponíveis nas casas de apostas relacionadas a chance que a casa avaliou daquilo ocorrer. No Brasil e na maior parte do mundo elas são dadas por números decimais. Se, por exemplo uma casa de apostas calcular que a probabilidade de um certo time ganhar um jogo é de 80%, então a odd será dada com a fórmula 1/0.8 que dá uma odd de 1,25 e esse valor será retornado para o vencedor da aposta para cada 1 real apostado (MILLER; DAVIDOW, 2019).

2.4 Aprendizado de Máquina

Desde quando os computadores foram inventados, nos perguntamos se eles poderiam ser feitos para aprender (MITCHELL, 1997) e motivado por esse pensamento emergiu o Aprendizado de Máquina(AM). Esse é um campo da ciência da computação, definido por Arthur Samuel, onde um computador faz análises baseadas em dados, aprende e toma decisões sem a interferência humana. Temos três tipos principais de aprendizado: Supervisionado, Não-Supervisionado e por Reforço, os quais os dois primeiros são utilizados ao longo deste trabalho.

2.5 Aprendizado Supervisionado

Nessa abordagem, tentamos prever uma variável dependente a partir de uma lista de variáveis independentes, para isso são utilizados dados rotulados para criar um modelo

e fazer previsões de outros dados ainda sem uma resposta identificada. Existem dois tipos de aprendizado supervisionado: classificação e regressão (FACELI et al., 2021).

2.5.1 Classificação

Representa a predição de uma classe discreta, ou categoria. No livro de Saikat Dutt (DUTT; CHANDRAMOULI, 2018), temos um exemplo para partidas de críquete que podemos fazer um paralelo para o futebol. No problema de prever o resultado em uma partida, o classificador vai atribuir um valor de vitória/derrota para o atributo alvo baseado no valor de outras características como estatísticas dos times e desempenho em partidas anteriores. Uma classificação é considerada correta se, por exemplo, foi previsto pelo modelo que o time iria vencer e ele de fato venceu.

2.5.1.1 K-Nearest-Neighbor

É um algoritmo que classifica as instâncias baseada na similaridade com outras instâncias. Cada instância representa um ponto em um gráfico e suas coordenadas são definidas por todas as variáveis independentes, cada qual representando um eixo do gráfico. A instância com categoria desconhecida é prevista a partir dos K vizinhos mais próximos do seu ponto representado pelo gráfico, assim fazendo uma espécie de votação dos K vizinhos e definindo a classificação da instância. O número K é definido previamente e pode variar a depender dos dados disponíveis.

2.5.1.2 Naive Bayes

Esse algoritmo entra na área dos métodos probabilísticos de classificação. Nesse caso se procura a classe mais provável estatisticamente dada a execução do modelo e não a melhor classe como podemos ver nos modelos anteriores apresentados neste trabalho. Para entender melhor esse modelo, precisamos primeiro apresentar alguns conceitos:

- Probabilidade anterior: é a probabilidade atribuída a um evento antes de qualquer experimento ser realizado sobre ele;
- Probabilidade posterior: é a probabilidade atribuída a um evento após a realização de experimentos com novas informações sobre ele.

No teorema de NB temos que a probabilidade do conjunto A ocorrer dado B é definida

pela probabilidade do evento A ocorrer, multiplicado pela probabilidade do evento B ocorrer dada a ocorrência de A, dividido pela probabilidade anterior de B ocorrer.

2.5.1.3 Árvore de Decisão

Este é um modelo muito simples, pois é de fácil interpretação. Para se entender melhor, temos 3 termos que precisam ser explicados:

- Nó raiz: é primeiro nó da árvore, ele não possui ramos como entradas, só como saídas;
- Nó interno: são todos os nós intermediários da árvore, tendo tanto ramos de entrada como ramos de saída;
- Nó folha: é o último nó de um caminho da árvore, por isso não tem nenhum ramo na saída. Podemos ter múltiplos nós folha, onde cada qual representa a classificação final de um caminho percorrido.

Nesse modelo temos apenas um caminho da raiz até cada nó folha e ele representa um conjunto de classificação do tipo se-então. A construção da árvore funciona na forma *top down* usando a técnica de divisão e conquista recursivamente e existem algoritmos que podem ser usados para definir os melhores atributos para o desempenho final do modelo. É escolhido um atributo como raiz e dividimos nossas instâncias de acordo com os valores dos ramos criados a partir do atributo na raiz, chegando a novos nós que terão essas instâncias já separadas. Posteriormente, o processo é repetido recursivamente para os outros nós da árvore até chegar na folha. Para o nós folha, temos 2 possibilidades:

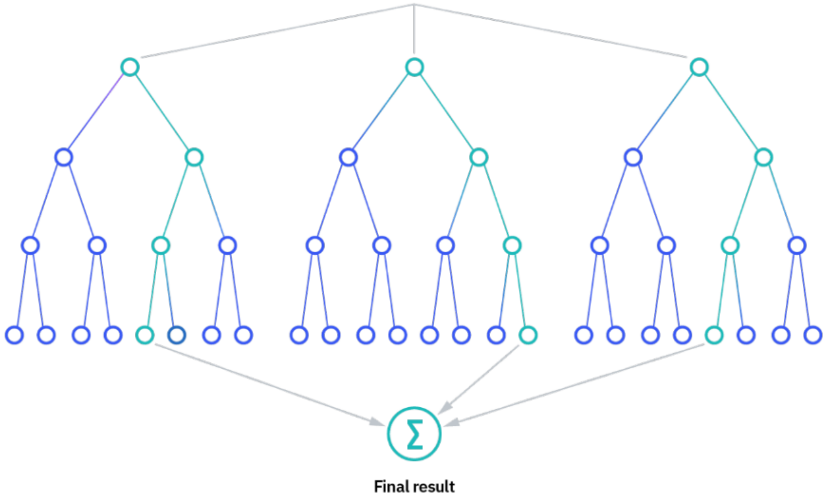
- Chegar em um nó com apenas instâncias da mesma classe, assim criando o nó folha com a classe correspondente a classe dessas instâncias;
- Não termos mais atributos para escolher, ou chegar em um nó vazio, assim é criada a folha é definida a classe a partir da classe predominante nas instâncias do último nó gerado.

2.5.1.4 Floresta Aleatória

O algoritmo de Floresta Aleatória se utiliza da combinação de inúmeras Árvores de Decisão para fazer a predição. Ele pode ser usado tanto para tarefas de classificação, como de regressão, mas neste trabalho utilizamos para classificação. A Floresta Aleatória toma a decisão em uma espécie de votação de cada árvore componente da floresta, a

Figura 2.1 ilustra isso, onde temos 3 árvores que chegam a uma decisão individualmente e fazem uma votação final para concluir a previsão da floresta.

Figura 2.1 – Flosreta Aleatória

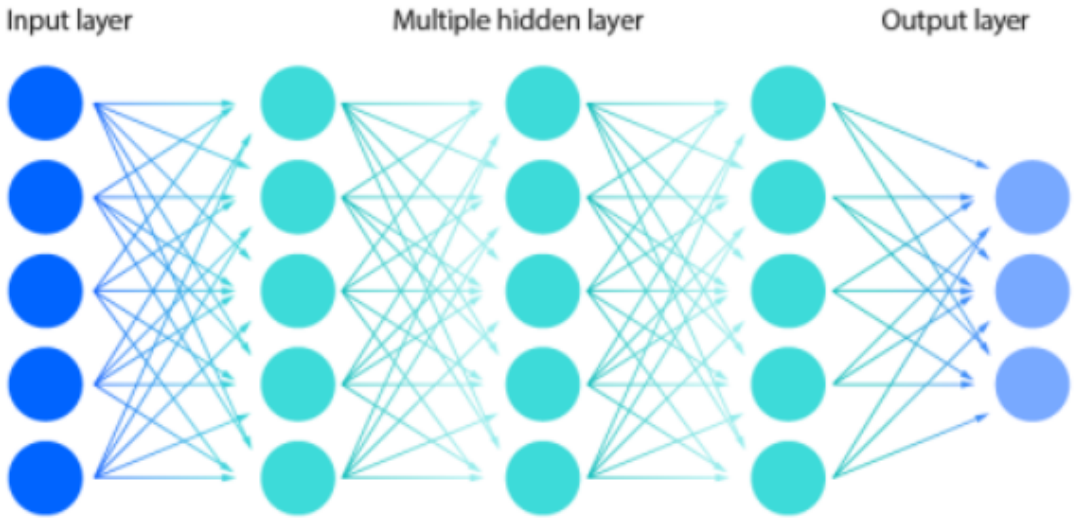


Fonte: IBM Corporation

2.5.1.5 Redes Neurais

Os algoritmos de Redes Neurais são inspirados no cérebro humano, pois tem como objetivo simular a maneira como os neurônios mandam sinais uns para os outros. Os neurônios são compostos por camadas de 1 nós, contendo uma camada na entrada, 1 ou mais camadas ocultas e uma camada de saída. Cada neurônio conecta-se a outro e tem um peso e um limite associado a ele. Se a saída de um nó estiver acima do valor limite, este nó será ativado, caso contrário nenhum dado será transmitido à próxima camada. Na Figura 2.2 temos uma exemplificação, com o nó de entrada(*Input layer*), nós intermediários(*Multiple hidden layer*) e nó de saída(*Output layer*) (IBM, 2023).

Figura 2.2 – Rede Neural



Fonte: IBM Corporation

2.5.2 Regressão

Na regressão, em vez de fazer a predição de resultados categoricamente com classes bem definidas, o atributo alvo é um valor contínuo (FACELI et al., 2021). Voltando ao exemplo do futebol, substituindo a definição binária de vitória e não vitória do time da casa que se tinha na classificação, é possível trazer a probabilidade de vitória para o time mandante medida em porcentagem como atributo alvo.

2.5.2.1 Regressão Logística

Apesar de ter 'regressão' no nome, esse é um algoritmo que é usado para a classificação. Na execução da Regressão Logística é retornado um número contínuo que pode atingir valores de escalas bem diferentes, tornando mais difícil a classificação. Devido a isso, é utilizado um método que leva todos os valores preditos pelo algoritmo para o intervalo [0,1], transformando esses valores em uma variável categórica e fazendo a classificação a partir disso. Como exemplo, podemos dizer que para a classificação valores abaixo de 0,5 assumem 1 categorização e acima de 0,5 assume outra categorização, nesse caso transformando o problema que antes tinha resultados com valores contínuos em resultados categóricos.

2.5.3 *XGBoosting*

Essa é atualmente a forma mais evoluída da aplicação de árvores de decisão para predição de resultados. É um algoritmo que também usa o *essemble*, assim como a floresta aleatória, que se utiliza de modelos distintos para gerar a predição. Ela usa uma técnica de *Gradient Boosting* mais aprimorada nesse algoritmo, que consiste na minimização de erros, executando os modelos sequencialmente e com o objetivo de corrigir o modelo executado anteriormente com a execução do atual. É um algoritmo que também pode ser usado em classificação, mas foi utilizado para a regressão no escopo deste trabalho.

2.6 Aprendizado Não-Supervisionado

Nessa abordagem, não temos um rótulo esperado para cada instância dos nossos dados. Na execução desses algoritmos, é buscado um padrão ou similaridade entre os dados sem qualquer classificação pré estabelecida. Eles podem ser divididos entre algoritmos de transformação e algoritmos de agrupamento. Nos algoritmos de transformação, é gerada uma nova representação para os dados, que podem ser mais facilmente compreendidos por humanos, ou utilizados em outros algoritmos de AM. Nos algoritmos de agrupamento, os dados são particionados em grupos por similaridades (FONTANA, 2020).

2.6.1 K-Means

O *K-Means* é um algoritmo que tem por objetivo fazer agrupamento de instâncias de modo que elas sejam semelhantes entre si. Esses grupos são chamados de agrupamentos e cada um idealmente tem elementos que sejam semelhantes entre si e que se diferem dos outros grupos (PÉREZ-ORTEGA et al., 2019). Inicialmente, definimos k que representa o número de grupos e cada grupo tem um ponto como centróide inicial, sendo definido aleatoriamente. Através de uma fórmula de proximidade, calculamos qual centróide cada instância está mais próxima, então a partir disso são formados os agrupamentos. A partir dos grupos formados, recalculamos cada centróide pelo cálculo da média das instâncias pertencentes àquele cluster e voltamos ao passo de calcular a proximidade das instâncias aos centróides formando novos grupos. Continuamos fazendo essas itera-

ções até que os centróides se estabilizem e não se tenham mais elementos mudando de grupos. Como os centróides iniciais são definidos de maneira aleatória, é interessante executar o algoritmo diversas vezes para se ter a melhor noção de qual centróide gera os melhores grupos que representam bem os dados. Para se definir qual o melhor valor para k normalmente é utilizado o método do cotovelo, que roda o *K-Means* diversas vezes para diferentes quantidades de grupos. Ele normalmente é representado por um gráfico que ao atingir a quantidade ótima de grupos forma uma curva semelhante a um cotovelo no ponto ideal.

2.7 Escolha de algoritmos

Todos os algoritmos selecionados para esse trabalho foram utilizados porque são muito usuais no desenvolvimento de trabalhos de AM. Além disso, existem algoritmos de diferentes poderes de processamento e o objetivo foi verificar como eles se comportaram e verificar quais se saíram melhor com os dados propostos. Como exemplo, temos dois algoritmos baseados em Árvores de Decisão usados para a classificação e um para a regressão e todos tiveram resultados finais distintos.

2.8 Dados Qualitativos vs Dados Quantitativos

Todos os dados que são usados em modelos de AM são estruturados, podendo representar objetos físicos, ou noções mais abstratas (FACELI et al., 2021). Normalmente representados por tabelas, cada linha representa uma instância que é uma unidade da entidade a qual estamos observando e as colunas representam características, chamadas de atributos, dessa entidade que diferem em cada instância apresentada. Temos um atributo dito o atributo alvo que a depender dos dados pode estar ou não na nossa tabela e é estimado a partir da análise dos demais atributos existentes. Quando temos o atributo alvo nossos dados são ditos rotulados, quando não temos são ditos não rotulados e existem modelos de AM que são específicos para cada tipo (GUPTA, 2023). Em relação ao formato dos dados podemos ter dois tipos (MARTINS, 2011):

- Dados quantitativos: definidos por qualidades da instância, como sexo e cor do cabelo, que podem ser representados por palavras ou números e que não podem ser usados em operações aritméticas, são apenas categóricos;

- Dados qualitativos: definem características mensuráveis das instâncias, como altura e peso, representados por números discretos ou contínuos, mas que podem ser usados em operações aritméticas ao utilizar os algoritmos

2.9 Normalização

Na análise de dados podemos ter dados dos mais diversos formatos. No caso desse trabalho, os dados são compostos por números contínuos. Temos uma série de variáveis com seus valores tendo escalas bem distintas, que pode ser representada desde números decimais a números que chegam na casa das centenas. Isso pode causar uma discrepância na predição, pois atributos com ordem de grandeza mais elevadas vão ter peso maior que atributos de números mais baixos no momento da aplicação dos algoritmos. Por esse motivo, nesses casos, é necessário a aplicação da normalização, onde tratamos todos os dados para o uma mesma escala. Por exemplo, nos intervalos entre 0 e 1 para números positivos e -1 e 1 no caso de existirem números negativos nos dados (HAN; PEI; KAMBER, 2011), assim erradicando o problema de ter atributos com pesos muito distintos e prejudicando o resultado final dos modelos. A normalização utilizada nos modelos desse trabalho é a min-max (JUNIOR, 2020), conforme representada na Equação 2.1:

$$X_{norm} = X_i - X_{min}/X_{max} - X_{min} \tag{2.1}$$

2.10 Validação Cruzada

Para uma boa avaliação de um modelo de AM, precisamos remover o *overfitting*. *overfiting* é um problema que ocorre quando treinamos nossos modelos e o desempenho fica muito atrelado a amostra de dados usados e acaba não tendo um desempenho efetivo para outros casos fora da amostra (WEBB; SAMMUT, 2010). Para resolver isso, nesse trabalho, utilizaremos o método de validação cruzada k-fold estratificado que vai auxiliar os modelos a ter um desempenho mais generalizado para diferentes conjuntos de dados. Nessa abordagem, após termos todos os dados pré-processados e os algoritmos definidos para fazer a predição, vamos dividir nossos dados em k partições e cada uma delas é dividida em conjuntos de treinamento e teste. As partições são utilizados nos algoritmos e temos o retorno do desempenho calculado a partir de certas métricas. Após a execução

com todas as k partições , calcula-se a média para verificar o desempenho final do modelo para o algoritmo específico utilizado. Com a estratificação, mantemos a proporcionalidade da classificação dos nossos dados intactos para cada partição. Isto significa que, se nós temos nos nossos dados 80% das instâncias pertencentes a classe x e 20% pertencentes a classe y, o algoritmo idealmente vai tentar manter essa mesma divisão nas partições geradas (FACELI et al., 2021).

2.11 Métricas de Desempenho

Para a avaliação de desempenho de cada modelo são necessárias algumas métricas que são calculadas a partir da validação e demonstradas nesta seção. Em AM temos um teorema dito No Free Lunch Theorem (WOLPERT; MACREADY, 1997), que diz que não existe um algoritmo de predição que tenha um desempenho satisfatório para todos os problemas existentes, então, precisamos avaliar a qualidade do modelo e decidir qual o melhor algoritmo para o problema de interesse. Para visualizar a performance de um modelo, utilizaremos a matriz de confusão. No caso deste trabalho, temos uma classificação binária e, por simplificação uma classe é dada como positiva e a outra como negativa. A Figura 2.3 apresenta uma exemplificação da matriz, onde:

- Verdadeiros Positivos(VP): número de instâncias preditas como positivas e que estão corretamente classificadas;
- Verdadeiros Negativos(VN): número de instâncias preditas como negativas e que estão corretamente classificadas;
- Falsos Positivos(FP): número de instâncias preditas como positivas e que estão incorretamente classificadas;
- Falsos Negativos(FN): número de instâncias preditas como negativas e que estão incorretamente classificadas.

Figura 2.3 – Matriz de Confusão

		Classe predita	
		+	-
Classe verdadeira	+	VP	FN
	-	FP	VN

Fonte: Katti Facelli (2021)

A partir da matriz de confusão, podemos calcular as métricas para avaliação do desempenho dos algoritmos de predição. As métricas (FACELI et al., 2021) utilizadas são as seguintes:

- Acurácia: calcula a porcentagem de acerto total do algoritmo, Fórmula 2.2. Faz a soma de todas as instâncias positivas e negativas classificadas corretamente e divide pelo número total de instâncias.

$$accuracy = \frac{VP + VN}{n} \tag{2.2}$$

- Precisão: calcula a porcentagem de acerto das instâncias classificadas como positivas, Fórmula 2.3. Faz a divisão das instâncias classificadas corretamente como positivas por todas as instâncias classificadas como positivas.

$$precision = \frac{VP}{VP + FP} \tag{2.3}$$

- *Negative Predictive Value*(NPV): calcula a porcentagem de acerto das instâncias classificadas como negativas, Fórmula 2.4 faz a divisão das instâncias classificadas corretamente como negativas por todas as instâncias classificadas como negativas.

$$nvp = \frac{VN}{VN + FN} \tag{2.4}$$

- Sensibilidade: calcula a taxa de acertos na classe positiva. A Fórmula 2.5 faz a divisão das instâncias classificadas corretamente como positivas pelas instâncias classificadas como corretamente como positivas mais instâncias classificadas como negativas, mas que na verdade são positivas.

$$recall = \frac{VP}{VP + FN} \tag{2.5}$$

- Medida F1: faz a união da precisão e da sensibilidade a fim de trazer um número único que calcule a qualidade geral do nosso modelo. A Fórmula 2.6 faz a multiplicação da precisão pela sensibilidade e pelo número 2, dividindo pela soma da precisão, mais a sensibilidade.

$$f1 = \frac{2 * precision * recall}{precision + recall} \tag{2.6}$$

Essas métricas foram escolhidas, pois ao longo do trabalho veremos que não existe

um foco principal sobre nenhuma das classes, apenas o objetivo de acertar o máximo de instâncias possíveis. Foi definido que nos cálculos, a possibilidade de vencer é designada como classe positiva e a possibilidade de não vencer(empatar ou perder) como classe negativa. Dessa maneira, as métricas são utilizadas visando algoritmos que maximizem o desempenho de acerto das duas classes. A sensibilidade não foi utilizada como comparação apenas por achar que a utilização da precisão encaixou melhor com o entendimento do trabalho, mas ainda assim o F1 foi colocado nas análises por fazer uma espécie de união entre sensibilidade e precisão e basicamente calculando as duas métricas em apenas uma. A F1 foi utilizada para as duas classes definidas no trabalho.

2.12 Tecnologias Utilizadas

Nesta seção, são apresentadas as tecnologias utilizadas neste trabalho que foram utilizadas para fazer a aquisição e pré-processamento dos dados e posterior implementação dos algoritmos de AM para fazer a análise de predição.

2.12.1 *Google Colab*

O *Google Colab* (GOOGLE, 2023), ou *Colaboratory*, é um ambiente interativo que permite desenvolvimento e execução de código em *Python*, além de utilização de forma colaborativa de outros formatos como texto e imagens em um só documento. Ele é um serviço baseado no *Jupter Notebook*, que diferentemente do *Colaboratory* só pode ser usado localmente na máquina. Ele proporciona muitas vantagens em relação aos ambientes de desenvolvimento habituais, são elas: a gratuidade e facilidade de acesso ao serviço precisando apenas de uma conta ativa no *Google*, a possibilidade de desenvolver de forma totalmente online e com salvamentos automáticos, a dispensabilidade de uma máquina robusta, pois o serviço oferece acesso gratuito aos recursos computacionais do Google pela nuvem, o acesso simultâneo de vários usuários a um documento e o compartilhamento dos documentos que pode ser feito de maneira rápida através de links. Além disso, a ferramenta é muito poderosa e serve perfeitamente para as atividades de AM e Ciência de Dados que são o foco deste trabalho.

2.12.2 *Google Sheets*

O *Google Sheets* (SHEETS, 2024) é uma ferramenta para criação de planilhas eletrônicas. Resumidamente essas planilhas são tabelas compostas por linhas e colunas e dentro de cada célula é possível conter textos, números ou fórmulas matemáticas. Os dados colocados dentro da planilhas podem ser classificados, filtrados, colocados dentro de outras tabelas e usados para criação de gráficos. Toda essa manipulação pode ser feita visando uma melhor visualização dos dados e posterior análise. Muito semelhante ao *Excel*, o *Google Sheets* pode ser utilizado de maneira online, gratuita. Essa ferramenta será utilizada no trabalho para o salvamento dos dados adquiridos, a facilitação de alguns pré-processamentos dos dados e melhor integração com o *Google Colab*.

2.12.3 *Python*

Python (FOUNDATION, 2023b) é uma linguagem de programação de alto nível, interpretada, interativa, orientada a objeto e de código aberto criada em 1989 por Guido van Rossum, um desenvolvedor que por não encontrar nenhuma outra linguagem com as especificidades que ele necessitava acabou por implementar uma nova que resolvesse seu problema. É uma linguagem de propósito geral, com uma sintaxe simples, inúmeras bibliotecas e um grande suporte da comunidade. Atrelado a isso o *Python* oferece uma vasta quantidade de recursos relacionados a AM que auxiliam no processo de análise de dados, tema que será abordado neste trabalho. As bibliotecas são códigos já escritos por outros programadores que facilitam e agilizam o desenvolvimento de um código novo. O *Python* já tem uma biblioteca padrão bem grande que oferece inúmeros recursos para programação (FOUNDATION, 2023a). Um exemplo de uma biblioteca padrão que é utilizada de forma significativa nesse trabalho é a *sqlite3-python* que disponibiliza um banco de dados que não necessita de um servidor separado e permite o acesso ao banco de uma forma facilitada, permitindo manipulação de dados e consultas (FOUNDATION, 2023c). Existem duas bibliotecas que não são padrão do *Python* e que merecem uma maior atenção, pois serão usadas ao longo deste trabalho: o *scikit-learn* (SCIKIT-LEARN, 2023) é uma biblioteca com foco em AM que apresenta inúmeros algoritmos de aprendizado supervisionado e não-supervisionado, além de um suporte para o pré-processamento de dados e outras ferramentas relacionadas aos algoritmos. O *pandas* (ALMEIDA, 2023) é uma biblioteca utilizada para manipulação de dados relacionais e pode ser utilizada para

uma série de processos como visualização de dados, consulta em banco de dados e suporte para atividades de AM.

2.13 Considerações Finais

Neste capítulo foram apresentados os conceitos e as ferramentas que foram usados no desenvolvimento deste trabalho para a execução e análise dos algoritmos utilizados para predição de resultados de partidas de futebol. Foram usados o *Google Colab* visando a facilitação do desenvolvimento e salvamento do código criado, sem precisar fazer downloads localmente para o computador; a linguagem *Python* que apresenta bibliotecas que impactam positivamente se comparadas a outras linguagens de programação nas atividades propostas que envolvem os assuntos de AM e Ciência de Dados; e a ferramenta e *Google Sheets* que foi utilizada para algumas atividades de pré-processamento de dados.

3 TRABALHOS RELACIONADOS

Neste capítulo, são apresentados alguns trabalhos relacionados a predição de resultados de partidas de futebol. O capítulo é finalizado com uma análise comparativa entre os trabalhos pesquisados.

3.1 Um modelo de previsão de resultados de futebol utilizando Machine Learning

No trabalho de conclusão de curso da faculdade de Ciência da Computação da UFRGS (BOUCINHA, 2023) é apresentado um modelo de predição de resultados para predição de partidas de futebol. No trabalho foi proposta a análise das temporadas desde o ano de 2018 até 2022 das principais ligas do mundo do futebol, retirando os dados do site Sportmonks, tratando e aplicando em uma série de algoritmos de AM com o intuito de comparação para visualizar qual teria o melhor desempenho. Com os dados, foram criadas variáveis relacionadas à gols feitos e sofridos, além do retrospecto de vitórias e derrotas. Essas variáveis foram atribuídas igualmente ao retrospecto do time mandante e do time visitante; e divididas em 3 grupos: histórico das últimas 10 partidas dentro do campeonato atual, histórico em todos os jogos do campeonato vigente e histórico dos jogos das últimas 3 temporadas. Foram usados algoritmos de AM supervisionados com 2 rótulos: vitória do time mandante, ou não vitória do time mandante que representa os resultados tanto de empate, quanto de vitória do time visitante. Foram definidos os melhores algoritmos e é feita uma comparação usando os modelos entre as probabilidades apresentadas nas casas de apostas calculando a possibilidade de lucro. Foi avaliado que os modelos poderiam gerar um lucro para apostas em partidas que não tem um favorito, onde as probabilidades de vitória ficavam entre 45% e 55%.

3.2 Utilizando Aprendizado de Máquina para predição de resultados da NBA

No trabalho de conclusão de curso da faculdade de Engenharia de Produção e Transportes da PUCRS (MELO, 2021) é uma proposta de um modelo de previsão para partidas da National Basketball Association(NBA). Nesse caso, o autor criou um banco de dados a partir da API oficial da NBA, onde é possível encontrar inúmeras estatísticas relacionadas às partidas. Após esse levantamento, esses dados foram aplicados em algoritmos

de AM para posterior avaliação de qual modelo teve um melhor desempenho na predição. No basquete, caso uma partida termine empatada no último quarto, o desempate é processado através de períodos prorrogatórios de 5 minutos até termos um vencedor. Por isso, neste trabalho foram usados dados rotulados com dois classificadores: vitória do time da casa ou vitória do time visitante. Um ponto interessante a se destacar nesse trabalho foi a criação de uma variável baseada no 'Elo Rating', um cálculo criado por Arpad Elo para classificar o nível de jogadores de xadrez baseado no desempenho deles em partidas anteriores que foi adaptado para o contexto do basquete. Neste trabalho, foi possível verificar o retorno positivo de alguns times, mas não foi possível achar um padrão relacionado ao posicionamento final do time de basquete no campeonato e o retorno.

3.3 Análise de Variáveis em Partidas de Futebol: Previsão de Resultados com Naïve Bayes e Poisson

Neste trabalho desenvolvido para o Programa de Pós-graduação em Sistemas e Processos Industriais na UNISC (SEHNEM et al., 2021) e apresentado no XVIII Encontro Nacional de Inteligência Artificial e Computacional em 2021 é possível ver uma abordagem onde o banco de dados foi preenchido manualmente com estatísticas das partidas retirando os dados de sites esportivos e criando uma variável autoral que calcula o desempenho das equipes em partidas anteriores. Após a criação da base de dados, foram feitos modelos com 14 conjuntos diferentes de variáveis aplicando o algoritmo de Naive Bayes e outros 3 conjuntos de variáveis aplicando o cálculo de Poisson, com posterior comparação dos modelos de predição para verificar qual deles tem o melhor desempenho. Foi possível verificar que o Naive Bayes teve um desempenho superior ao cálculo de Poison e um lucro previsto em mais de um conjunto de variáveis testadas. O desempenho máximo de acertos ao final do trabalho foi de 53%, frisando que nesse caso existem 3 possibilidades de resultados, naturalmente tornando mais difícil a previsão.

3.4 Comparação dos Trabalhos

Em todos os trabalhos observados podemos ver os mesmos objetivos: criar o melhor modelo preditivo para resultados de partidas de algum esporte, visando aplicação dos modelos em casa de apostas esportivas e obter o melhor lucro. No trabalho (MELO,

2021), a diferença se dá pelo esporte utilizado, apesar dessa diferença, é possível notar que foram usadas estatísticas das partidas da NBA para poder criar modelos de previsão. Um dado interessante que é utilizado no nosso trabalho é o ELO, que foi adaptado para ser usado para o futebol e criar outra variável aplicada nos algoritmos usados. No trabalho (SEHNEM et al., 2021), é possível ver a aplicação de vários conjuntos no algoritmos de Naive Bayes para achar o melhor modelo. Em nosso trabalho é utilizado um algoritmo para filtrar quais as melhores variáveis para serem utilizadas e aplicação de uma série de algoritmos de AM para encontrar o melhor modelo a ser adotado. No trabalho (BOUCINHA, 2023), temos também a aplicação de uma série de algoritmos de AM para analisar qual é o melhor modelo. Em nosso trabalho, são utilizadas as variáveis utilizadas em (BOUCINHA, 2023), além de adição de outras variáveis relacionadas às estatísticas das partidas como chutes ao gol, defesas do goleiro, passes certos etc. Com o objetivo de melhorar o desempenho, nós fazemos a seleção de variáveis, além da aplicação de um algoritmo de aprendizado não supervisionado antes da aplicação dos algoritmos de aprendizado supervisionado para comparação.

A Tabela 3.1 apresenta uma comparação entre o trabalho proposto e os trabalhos relacionados citados anteriormente com o objetivo de esclarecer a diferença entre eles.

Tabela 3.1 – Tabela comparativa com trabalhos relacionados

	Predição Futebol	**TR 1**	**TR 2**	**TR 3**
Esporte	Futebol	Futebol	Basquete	Futebol
Campeonatos	Brasileirão	Mundo	NBA	Brasil
Classificação	Sim	Sim	Sim	Sim
Regressão	Sim	Não	Não	Não
Não-Supervisionado	Sim	Não	Não	Não

4 PROPOSTA E METODOLOGIA

Neste capítulo é apresentada a metodologia que foi utilizada como inspiração para o desenvolvimento do trabalho, uma breve explicação das atividades normalmente realizadas em cada etapa e descrição das atividades realizadas nesse trabalho por etapa.

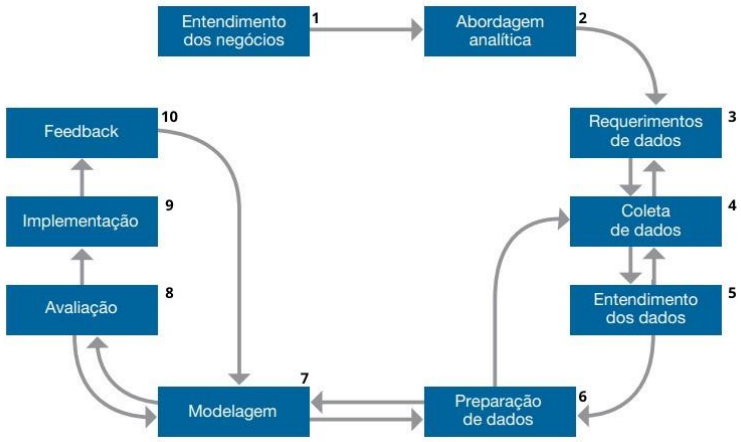
4.1 Visão Geral da Proposta

O objetivo deste trabalho é a criação de um modelo de predição de resultados de partidas de futebol usando AM e a metodologia de Ciência de Dados. São levantados dados estatísticos relacionados aos dois times envolvidos em cada partida do Campeonato Brasileiro de 2018 até 2022 e um comparativo aplicando algoritmos de AM tanto de aprendizado supervisionado quanto não-supervisionado para avaliar qual, ou quais apresentam os melhores resultados. Posteriormente, os melhores modelos avaliados são aplicados para as partidas jogadas no Campeonato Brasileiro de 2023, simulando uma aposta baseado nos resultados levantados pelos modelos e conceitos relacionados a apostas, verificando o lucro/perda total que seria obtido ao fim do campeonato se fosse aplicado no mundo real.

4.2 Metodologia

A metodologia de Ciência de Dados utilizada nesse trabalho é baseada no (ROLLINS, 2015) que apresenta 10 estágios que vão desde o entendimento do negócio até a obtenção de feedback após a finalização de todo o processo, conforme ilustrado na Figura 4.1. Essa metodologia visa ser satisfatória independentemente do escopo do projeto, seja relacionado ao tamanho do banco de dados utilizado, ou as tecnologias e abordagens envolvidas. A seguir são descritas as atividades realizadas nas 10 etapas do processo:

Figura 4.1 – Fluxo Metodologia de Base para Ciência de Dados



Fonte: IBM Corporation

- **Etapa 1: Entendimento do negócio**: Etapa inicial de todo projeto, inicia-se com o apontamento do problema, passa pela definição de objetivos e requisitos da solução que por fim culminarão na resolução do problema. Essa é uma etapa muito importante, pois a base de todo o projeto será definida aqui, impactando em todas as outras fases. Neste trabalho, o problema foi definido a partir das apostas esportivas, que apesar de não serem consideradas um jogo de azar, é necessária uma estratégia para conseguir lucro. Baseada nisso, a proposta foi idealizada para criar um modelo de AM para predição de partida de futebol do Campeonato Brasileiro com o intuito de aplicar em casas de apostas.
- **Etapa 2: Abordagem analítica**: Após a definição do problema, é o momento de definir a abordagem para resolução desse problema. Aqui são identificadas as técnicas que mais fazem sentido para se obter o melhor resultado. Neste trabalho, a ideia é criar o melhor modelo de AM para resolver o problema. É explorada uma série de algoritmos de aprendizado supervisionado e não supervisionado, técnicas de classificação e de regressão. Por fim, é realizada uma análise dos melhores modelos e, então, selecionados para aplicação em uma simulação real de apostas.
- **Etapa 3: Requisitos de dados**: A partir da abordagem analítica escolhida, é necessário escolher os dados com uma formatação condizente para que possam ser aplicados a essa abordagem. Neste trabalho, são desenvolvidos algoritmos que percorrerem cada linha de um banco de dados e criam variáveis para serem aplicadas aos algoritmos de AM. Nesse sentido, temos arquivos nos formatos .csv com dados representando números que são colocados dentro de um banco de dados e manipulados criando as variáveis para depois serem usados no trabalho.

- **Etapa 4: Coleta de dados**: É onde os dados necessários para a análise proposta no projeto são reunidos. A depender do escopo, as informações podem ser mais complexas de adquirir, apresentando restrições de acesso e precisando de investimentos financeiros para captação. Pode-se precisar adicionar mais dados conforme o desenvolvimento do projeto, ao revisar os requisitos, com isso impactando no resultado final. Neste trabalho, foram levantados dados de estatísticas de partidas do Campeonato Brasileiro de Futebol do ano de 2018 até 2022. Os dados foram retirados do *Sofascore*[1], um site gratuito que possui uma grande quantidade de informações referentes a partidas de futebol.
- **Etapa 5: Entendimento dos dados**: Nesta etapa, é dada uma atenção maior para os dados adquiridos, avaliando e tendo percepções iniciais sobre eles. Em certos casos, os dados levantados são muito complexos e precisam de algumas técnicas para auxiliar no entendimento. Neste trabalho, temos dados brutos de estatísticas sobre as partidas como número de chutes no gol, gols marcados e passes certos, além de alguns dados em porcentagem que representam posse de bola, porcentagem de acerto no passe, entre outros.
- **Etapa 6: Preparação dos dados**: Esse é um processo bem importante de Ciência de Dados que pode tomar até 80% do tempo de um projeto (SERVICES, 2023). Abrange todos os passos para construção de um conjunto de dados para ser utilizado em um modelo de AM. Nessa etapa, dados são aglutinados quando provêm de fontes diferentes, limpos, formatados e são realizadas todas as ações que forem necessárias para deixá-los prontos para serem aplicados nas próximas etapas. Em nosso trabalho, os dados vieram de apenas uma fonte, porém algumas ações foram tomadas para ajustes, como: limpar informações desnecessárias, completar campos que vieram em branco, modificar campos para ficarem no formato desejado e adicionar informações faltantes em algumas instâncias. Depois disso, foram criadas as variáveis que são utilizadas nos modelos, colocadas em uma nova tabela e normalizadas para não se ter nenhum impacto relacionado à diferença de escala entre elas.
- **Etapa 7: Modelagem**: Após a preparação do conjunto de dados, inicia-se a construção do modelo a partir da abordagem analítica definida. Neste trabalho, exploramos vários algoritmos de AM para analisar qual o melhor modelo para predição, foram usados 7 algoritmos. Os algoritmos de aprendizado supervisionado de classi-

[1]Disponível em: https://www.sofascore.com/en-us/

ficação *K-Nearest Neighbor*, *Naive Bayes*, *Decision Tree*, *Random Forest*, *Logistic Regression* e *Neutral Network*. O algoritmo de aprendizado supervisionado de regressão *XGBoosting*. O algoritmo de aprendizado não-supervisionado *K-Means*.

- **Etapa 8: Avaliação**: É o momento de avaliar o modelo construído, usando ferramentas para mensurar a qualidade do modelo e eficácia para a resolução do problema. Neste trabalho, foi utilizada a validação cruzada, que cria vários conjuntos diferentes de dados divididos entre teste e treino e executa o algoritmo de predição em cada conjunto tirando a média para calcular o desempenho final de cada modelo. Como métricas de desempenho para comparação dos modelos foram utilizadas acurácia, precisão e NPV.
- **Etapa 9: Implementação**: Quando o melhor modelo é aprovado, ele é implementado em um ambiente de produção, ou em um ambiente de teste similar para poder aperfeiçoar até ser possível colocá-lo em produção. No caso deste trabalho, não temos uma implementação em produção, apenas é feita uma simulação de apostas com as partidas do Campeonato Brasileiro de 2023 para avaliarmos o possível ganho/prejuízo do modelo no mundo real de apostas esportivas de futebol.
- **Etapa 10: Feedback**: Após a implementação e aplicação desse modelo em um ambiente mais robusto, podemos ter o *feedback* para refiná-lo ainda mais, podendo aplicar automatizações, ou outras features que não tinham sido notadas antes desse teste. No caso deste trabalho a porcentagem de acerto e de lucro/prejuízo obtido nas casas de apostas pela simulação nos mostra se é possível aplicar este algoritmo de predição no mundo real.

4.3 Considerações Finais

Neste capítulo, foram apresentados a proposta do trabalho e a metodologia que será utilizada para fazer uma análise de predição de resultados de partidas de futebol. Os processos feitos em cada etapa da metodologia são especificados na avaliação experimental.

5 AVALIAÇÃO EXPERIMENTAL

Neste capítulo, são descritos os experimentos realizados na tentativa de predição do resultado de partidas do Campeonato Brasileiro de Futebol. Primeiro, descrevemos a implementação da etapa de pré-processamento de dados e então apresentamos e comparamos os resultados dos modelos criados. Em seguida, apresentamos os melhores modelos avaliados e aplicamos nas partidas do Brasileirão 2023, cujas cotações foram registradas, para calcular o possível lucro, ou prejuízo que teríamos aplicando os modelos em um cenário real de casas de apostas. As seções a seguir são baseadas na metodologia proposta e as atividades práticas executadas dentro de cada etapa.

5.1 Coleta de Dados

Os dados utilizados para a realização do trabalho foram retirados do *Sofascore*, uma das maiores ferramentas de estatísticas do planeta, que tem informações de times e campeonatos dos principais esportes disputados no mundo. Esse é um site gratuito e disponibiliza fácil acesso a seus métodos. Foi feita uma coleta de dados com a linguagem Python por meio de 3 endpoints do site, dos quais foram retiradas no total de 67 informações das partidas dos ultimos 5 anos do Campeonato Brasileiro de futebol. Os dados são relacionados a diversas informações e estatísticas dos times envolvidos nas partidas. Na Tabela 5.1, é possível visualizar uma exemplificação das informações levantadas, que foram transformadas em um arquivo no formato .csv para facilitar a compreensão.

Figura 5.1 – Tabela com estatísticas do Campeonato Brasileiro

	Date	Year	Principal	Visitor	Winner	Score_principal	Score_visitor	Ball_possession_home	Ball_possession_away	Total_shots_home	Total_shots_away	Shots_on_target_home
2	2018-04-14	2018	Cruzeiro	Gremio	3	0	1	0.40	0.60	12	6	2
3	2018-04-14	2018	Vitoria	Flamengo	2	2	2	0.66	0.34	21	10	6
4	2018-04-15	2018	Vasco	Atletico Mineiro	1	2	1	0.66	0.34	19	11	7
5	2018-04-15	2018	Corinthians	Fluminense	1	2	1	0.64	0.36	11	13	2
6	2018-04-15	2018	Internacional	Bahia	1	2	0	0.45	0.55	13	12	4
7	2018-04-15	2018	Santos	Ceara	1	2	0	0.54	0.46	22	10	3
8	2018-04-15	2018	Athletico	Chapecoense	1	5	1	0.69	0.31	18	14	8
9	2018-04-15	2018	America Mineiro	Sport Recife	1	3	0	0.39	0.61	14	15	4
10	2018-04-17	2018	Botafogo	Palmeiras	2	1	1	0.46	0.54	13	7	3
11	2018-04-17	2018	Sao Paulo	Parana Clube	1	1	0	0.51	0.49	10	7	5
12	2018-04-21	2018	Flamengo	America Mineiro	1	2	0	0.49	0.51	14	18	5
13	2018-04-21	2018	Bahia	Santos	1	1	0	0.53	0.47	12	8	5
14	2018-04-22	2018	Fluminense	Cruzeiro	1	1	0	0.32	0.68	3	16	1
15	2018-04-22	2018	Chapecoense	Vasco	2	1	1	0.36	0.64	10	17	2
16	2018-04-22	2018	Ceara	Sao Paulo	2	0	0	0.51	0.49	14	11	3
17	2018-04-22	2018	Atletico Mineiro	Vitoria	1	2	1	0.60	0.40	12	11	5
18	2018-04-22	2018	Parana Clube	Corinthians	3	0	4	0.48	0.52	17	9	6
19	2018-04-22	2018	Gremio	Athletico	2	0	0	0.50	0.50	22	7	6
20	2018-04-22	2018	Palmeiras	Internacional	1	1	0	0.45	0.55	13	7	2
21	2018-04-24	2018	Sport Recife	Botafogo	2	1	1	0.50	0.50	20	10	11
22	2018-04-28	2018	Botafogo	Gremio	1	2	1	0.55	0.45	19	9	6

Fonte: Elaboração Própria

5.2 Preparação dos dados

A etapa de Preparação dos dados foi a que mais levou tempo na execução do trabalho. Primeiramente, os dados trazidos do Sofascore foram tratados dentro do *Google Sheets*, realizando uma série de atividades de pré-processamento para a criação das variáveis que serão descritas a seguir:

- *Redução de campos*: a raspagem inicial feita no Sofascore trouxe um total de 73 campos relacionados às partidas, porém 6 campos foram removidos por só estarem presentes em parte das instâncias trazidas e não ter sido possível encontrar dados precisos fora da ferramenta de estatísticas utilizada para poder completar esses campos faltantes.
- *Remoção de partida*: Dentre todas as partidas dos Campeonatos Brasileiros de 2018 a 2022, foi necessário remover uma partida que envolvia os times Palmeiras e CSA pelo certame de 2019. Essa decisão foi tomada porque essa partida veio com seus campos em branco, ou com suas estatísticas zeradas e não foram encontradas informações precisas para completar todos os campos. Foi realizada a conferência e esse jogo foi de fato disputado, porém, por algum motivo, os dados não estavam disponíveis no site. Além disso, por ser só um jogo com problema dentre os 1800 obtidos, o impacto de remover é mais satisfatório do que manter essa partida com dados totalmente inconsistentes.
- *Formatação dados de posse*: os dados de posse de bola vieram no formato de porcentagem com o simbolo % ao lado, por isso foi necessário editar esses dados e colocá-los em números decimais.
- *Formatação dados de acerto*: existem colunas que calculam acerto de passes, bolas longas, cruzamentos e dribles feitos. Nesses casos, vieram um conjunto de dados com os índices de acerto divididos pelas quantidade totais e a porcentagem de acerto com o símbolo %. Como exemplo, temos nestes campos um dado no formato 10/25 (40%). Realizando a formatação foram removidas as informações iniciais e formatadas as informações que geram porcentagem para números decimais.
- *Preenchimento de campos vazios*: algumas colunas trazidas tinham campos em branco, mas que após conferência foi constatado que esses campos vieram dessa maneira porque nas partidas relacionadas não foram constatadas ações referentes a esses campos para o time em questão. Por esse motivo, os dados em branco foram

preenchidos com 0.

- *Adição de novo campo*: foi adicionado um campo novo que não havia no *Sofascore*. Denominado 'Year'. Ele foi utilizado para colocar o ano do campeonato correspondente ao que a partida foi disputada. Isso se justifica, pois tivemos o campeonato de 2020, por exemplo, onde tivemos vários jogos disputados em 2021 por conta da pandemia de Covid-19, e esse campo serviu de auxílio para o algoritmo entender mais facilmente por qual campeonato a partida foi disputada.
- *Alteração de campos*: os dados correspondentes ao resultado das partidas trouxeram números categóricos onde 1 corresponde a vitória do mandante, 2 a empate e 3 a vitória do visitante. Como no trabalho proposto teremos um atributo alvo binário, os campos que correspondiam a vitória do visitante também foram editados para 2, unindo os dois outros resultados possíveis além da vitória do mandante.

5.2.1 Transformação em Variáveis

Após tratar todos os dados trazidos do *Sofascore* com o *Google Sheets*, dentro do *Google Colab* foram criadas variáveis com *Python* para serem usadas nos algoritmos de predição. No total, foram criadas 121 variáveis dependentes divididas em 4 conjuntos principais e mais uma variável chamada Elo Rating. Como atributo alvo, temos a variável categórica para a classificação que é apresentada nos dados como 1(para vitória do time mandante) e 2(para empate ou vitória do time visitante). Para o atributo alvo da regressão, temos uma variável contínua que representa a cotação para o time da casa vencer a partida. As variáveis dependentes criadas representam tanto o time mandante, quanto o time visitante e serão descritas a seguir.

5.2.1.1 Desempenho recente

Variáveis para calcular o desempenho das últimas 10 partidas dos times no campeonato, Figura 5.4 Nesse caso, são incluídos jogos de campeonatos anteriores, caso ainda não tenham sido disputados 10 jogos no campeonato atual. Elas foram elaboradas baseadas em gols feitos e sofridos, além de contagem de vitórias e derrotas nas últimas partidas do campeonato. Todas foram desenvolvidas somando os campos correspondentes ao time envolvido nas últimas 10 partidas e tirando a média no final.

Tabela 5.1 – Tabela desempenho últimos 10 jogos

Mandante	**Visitante**
Ultimas10Mandante_GolsMarcados	Ultimas10Visitante_GolsMarcados
Ultimas10Mandante_GolsSofridos	Ultimas10Visitante_GolsSofridos
Ultimas10Mandante_Vitorias	Ultimas10Visitante_Vitorias
Ultimas10Mandante_Derrotas	Ultimas10Visitante_Derrotas
Ultimas10Mandante_Jogos1GolMarcado	Ultimas10Visitante_Jogos1GolMarcado
Ultimas10Mandante_JogosMaisde1GolMarcado	Ultimas10Visitante_JogosMaisde1GolMarcado
Ultimas10Mandante_Jogos1GolSofrido	Ultimas10Visitante_Jogos1GolSofrido
Ultimas10Mandante_JogosMaisDe1GolSofrido	Ultimas10Visitante_JogosMaisDe1GolSofrido
Ultimas10Mandante_VitoriasMaisDe2Gols	Ultimas10Visitante_VitoriasMaisDe2Gols
Ultimas10Mandante_DerrotasMaisDe2Gols	Ultimas10Visitante_DerrotasMaisDe2Gols

5.2.1.2 Desempenho campeonato atual

Essa variáveis fazem o mesmo cálculo que é feito para o desempenho recente dos times, mas abrange todas as partidas disputadas pelos times no campeonato vigente que a partida em questão está sendo disputada, Figura 5.2.

Tabela 5.2 – Tabela desempenho da temporada do time

Mandante	**Visitante**
TemporadaMandante_GolsMarcados	TemporadaVisitante_GolsMarcados
TemporadaMandante_GolsSofridos	TemporadaVisitante_GolsSofridos
TemporadaMandante_Vitorias	TemporadaVisitante_Vitorias
TemporadaMandante_Derrotas	TemporadaVisitante_Derrotas
TemporadaMandante_Jogos1GolMarcado	TemporadaVisitante_Jogos1GolMarcado
TemporadaMandante_JogosMaisde1GolMarcado	TemporadaVisitante_JogosMaisde1GolMarcado
TemporadaMandante_Jogos1GolSofrido	TemporadaVisitante_Jogos1GolSofrido
TemporadaMandante_JogosMaisDe1GolSofrido	TemporadaVisitante_JogosMaisDe1GolSofrido
TemporadaMandante_VitoriasMaisDe2Gols	TemporadaVisitante_VitoriasMaisDe2Gols
TemporadaMandante_DerrotasMaisDe2Gols	TemporadaVisitante_DerrotasMaisDe2Gols

5.2.1.3 Desempenho últimos 3 campeonatos

Seguindo a mesma ideia dos dois outros conjuntos já apresentados, as variáveis mostradas na Tabela 5.3 visam calcular o retrospecto a longo prazo do time, fazendo a média do desempenho nos 3 campeonatos mais recentes, incluindo os jogos disputados no certame atual.

Tabela 5.3 – Tabela desempenho das últimas 3 temporadas do time

Mandante	Visitante
Ultimas3TemporadaMandante_GolsMarcados	Ultimas3TemporadaVisitante_GolsMarcados
Ultimas3TemporadaMandante_GolsSofridos	Ultimas3TemporadaVisitante_GolsSofridos
Ultimas3TemporadaMandante_Vitorias	Ultimas3TemporadaVisitante_Vitorias
Ultimas3TemporadaMandante_Derrotas	Ultimas3TemporadaVisitante_Derrotas
Ultimas3TemporadaMandante_Jogos1GolMarcado	Ultimas3TemporadaVisitante_Jogos1GolMarcado
Ultimas3TemporadaMandante_JogosMaisde1GolMarcado	Ultimas3TemporadaVisitante_JogosMaisde1GolMarcado
Ultimas3TemporadaMandante_Jogos1GolSofrido	Ultimas3TemporadaVisitante_Jogos1GolSofrido
Ultimas3TemporadaMandante_JogosMaisDe1GolSofrido	Ultimas3TemporadaVisitante_JogosMaisDe1GolSofrido
Ultimas3TemporadaMandante_VitoriasMaisDe2Gols	Ultimas3TemporadaVisitante_VitoriasMaisDe2Gols
Ultimas3TemporadaMandante_DerrotasMaisDe2Gols	Ultimas3TemporadaVisitante_DerrotasMaisDe2Gols

5.2.1.4 Desempenho estatístico recente

Com essas variáveis, temos o cálculo da média aritmética das últimas 10 partidas do time relacionado às estatísticas dos jogos vistas na Tabela 5.4 como chutes a gol, bolas na trave, desarmes, passes certos etc. Com essas variáveis não é calculado o histórico a longo prazo como feito anteriormente, pois nesse caso o desempenho estatístico está muito mais relacionado ao momento do time na temporada em questão de desempenho. Como podemos ter uma variação de estilo de jogo durante a temporada, principalmente no futebol brasileiro, isso impacta muito mais no desempenho estatístico do time do que no desempenho geral como gols, vitórias e derrotas, aspectos analisados nos outros conjuntos de variáveis, que é acredito ser possível encontrar um padrão mais facilmente em cada time.

Tabela 5.4 – Tabela desempenho estatístico últimos 10 jogos

Mandante	Visitante
Ultimas10Mandante_Posse_De_Bola	Ultimas10Visitante_Posse_De_Bola
Ultimas10Mandante_Chutes	Ultimas10Visitante_Chutes
Ultimas10Mandante_Chutes_No_Gol	Ultimas10Visitante_Chutes_No_Gol
Ultimas10Mandante_Chutes_Fora	Ultimas10Visitante_Chutes_Fora
Ultimas10Mandante_Chutes_Bloqueados	Ultimas10Visitante_Chutes_Bloqueados
Ultimas10Mandante_Escanteios	Ultimas10Visitante_Escanteios
Ultimas10Mandante_Impedimentos	Ultimas10Visitante_Impedimentos
Ultimas10Mandante_Faltas	Ultimas10Visitante_Faltas
Ultimas10Mandante_Cartoes_Amarelos	Ultimas10Visitante_Cartoes_Amarelos
Ultimas10Mandante_Cartoes_Vermelhos	Ultimas10Visitante_Cartoes_Vermelhos
Ultimas10Mandante_Chutes_Livres	Ultimas10Visitante_Chutes_Livres
Ultimas10Mandante_Laterais	Ultimas10Visitante_Laterais
Ultimas10Mandante_Tiros_De_Metas	Ultimas10Visitante_Tiros_De_Meta
Ultimas10Mandante_Grandes_Chances	Ultimas10Visitante_Grandes_Chances
Ultimas10Mandante_Grandes_Chances_Perdidas	Ultimas10Visitante_Grandes_Chances_Perdidas
Ultimas10Mandante_Chutes_Na_Traves	Ultimas10Visitante_Chutes_Na_Trave
Ultimas10Mandante_Defesas_Goleiro	Ultimas10Visitante_Defesas_Goleiro
Ultimas10Mandante_Passes	Ultimas10Visitante_Passes
Ultimas10Mandante_Acerto_De_Passes	Ultimas10Visitante_Acerto_De_Passes
Ultimas10Mandante_Acerto_Bolas_Longas	Ultimas10Visitante_Acerto_Bolas_Longas
Ultimas10Mandante_Acerto_Cruzamentos	Ultimas10Visitante_Acerto_Cruzamentos
Ultimas10Mandante_Acerto_Dribles	Ultimas10Visitante_Acerto_Dribles
Ultimas10Mandante_Perdas_Posse_De_Bola	Ultimas10Visitante_Perdas_Posse_De_Bola
Ultimas10Mandante_Duelos_Ganhos	Ultimas10Visitante_Duelos_Ganhos
Ultimas10Mandante_Duelos_Aereos_Ganhoss	Ultimas10Visitante_Duelos_Aereos_Ganhos
Ultimas10Mandante_Desarmes	Ultimas10Visitante_Desarmes
Ultimas10Mandante_Interceptacoes	Ultimas10Visitante_Interceptacoes
Ultimas10Mandante_Cortes	Ultimas10Visitante_Cortes
Ultimas10Mandante_Finalizacoes_Dentro_Da_Area	Ultimas10Visitante_Finalizacoes_Dentro_Da_Area
Ultimas10Mandante_Finalizacoes_Fora_Da_Area	Ultimas10Visitante_Finalizacoes_Fora_Da_Area

5.2.1.5 Elo Rating

O método Rating Elo foi criado por Arpad Elo para calcular a força de jogadores de xadrez (ELO, 2008) e é utilizado até hoje pela Federação Internacional de Xadrez para raquear jogadores profissionais. Esse método foi disseminado e adaptado para outros esportes, como basquete e futebol. Neste trabalho, foi criada uma variável baseada no Elo Rating da *World Football Elo Ratings*[1] que desenvolveu essa fórmula para o contexto do futebol (RATINGS, 2024). Foi utilizada uma pontuação base de 1500 para todos os times e calculado o rating a partir das partidas iniciais do Brasileirão de 2018, sendo atualizado para todos os times em cada partida até o fim do Brasileirão 2022.

Tabela 5.5 – Tabela Elo Rating

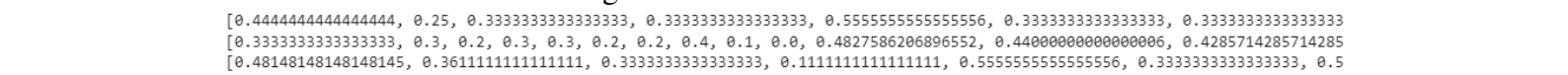
Mandante	Visitante
Elo_Rating_Mandante	Elo_Rating_Visitante

[1]Disponível em: https://eloratings.net/

5.2.1.6 Normalização

Como analisado anteriormente, os dados levantados neste trabalho tem ordens de grandeza diferentes. Como exemplo, temos a quantidade de passes dada por um time que em alguns jogos chegou a quase 900 por determinado time. Em comparação, temos valores bem menores, na casa dos decimais, que representam a porcentagem de posse de bola nas partidas. Para uma variável não ter um peso muito maior que outra na execução dos algoritmos e causar erros de predição, todos os dados da tabela gerada foram normalizados com o método *MinMaxScaler* que converte todos os valores dos dados para o intervalo [0,1]. Na Tabela 5.2 é possível visualizar como os valores ficaram após normalização.

Figura 5.2 – Tabela Normalizada



Fonte: Elaboração Própria

5.2.1.7 Seleção de Variáveis

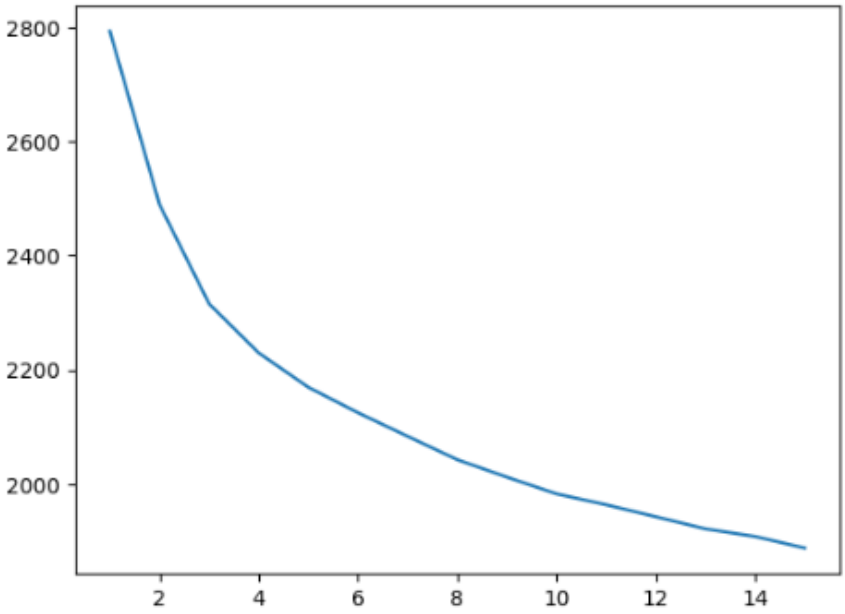
A seleção de variáveis dentro do processo de pré-processamento dos dados auxilia o modelo a não lidar com atributos irrelevantes na sua execução. Ao executar *SelectKBest* que seleciona as K melhores variáveis para o modelo, fiz o teste com 40, 60, 80, 100 e 120 variáveis. Apesar de não ter se obtido uma grande diferença referente às métricas, os melhores resultados foram obtidos com 100 variáveis.

5.2.1.8 Clusterização

O método não-supervisionado proposto no trabalho foi o *K-Means*, um algoritmo que faz a divisão dos dados do domínio em subconjuntos chamados de clusters. O objetivo dessa abordagem foi usar esse algoritmo como pré-processamento apenas para encontrar conjuntos parecidos de dados que possam acarretar em um melhor desempenho na execução dos algoritmos supervisionados, mas sem aplicar uma análise relacionada a quais semelhanças existem entre os dados de cada cluster. Foi usado o Método do Cotovelo, que executa o *K-Means* para várias quantidades diferentes de agrupamentos e define o melhor número para ser aplicado ao conjunto total dos dados disponível. Como é possível ver na Figura 5.3, foram testados até 15 grupos e o número ótimo definido para os dados propostos foi 3. Vale ressaltar, que apesar de ter sido aplicado tanto nos modelos de classificação quanto de regressão com atributos alvo diferentes, ainda assim o melhor

número de grupos se manteve em 3 para ambos.

Figura 5.3 – Método do Cotovelo - Elbow Method



Fonte: Elaboração Própria

5.3 Modelagem e Avaliação

Depois de realizar o pré-processamento dos dados, os modelos de predição foram criados. Foram utilizados 7 algoritmos de aprendizado supervisionado, sendo 6 de classificação e 1 de regressão. Em todos os modelos, foi utilizada a validação cruzada com os dados divididos em 5 subconjuntos na tentativa de remover o *overfiting*. Nos algoritmos de classificação foram executados e avaliados através das métricas de acurácia, precisão e NPV, foi feita uma comparação entre todos os modelos na Tabela 5.6 e os 2 melhores avaliados a partir das métricas definidas foram selecionados para serem visualizados mais a fundo e utilizados na simulação real de casa de apostas. Como nesses algoritmos foi realizada o agrupamento com *K-Means*, os dados da Tabela 5.6 foram preenchidos calculando a média aritmética simples das métricas de cada execução do conjunto total de dados e subconjuntos utilizados para facilitar o entendimento das comparações. Após a seleção dos melhores modelos, foi aprofundada a análise mostrando os resultados de cada cluster separadamente.

Tabela 5.6 – Comparação Modelos

Algoritmos	Acurácia	Precisão	NPV	F1 1	F1 2
K Nearest Neighbor	0.6275	0.5109	0.6354	0.4589	0.5408
Naive Bayes	0.5958	0.5528	0.5871	0.5554	0.5510
Decision Tree	0.5748	0.5341	0.5548	0.5089	0.5426
Random Forest	**0.6275**	**0.5708**	**0.6153**	**0.4808**	**0.5810**
Logistic Regression	**0.6271**	**0.5903**	**0.6269**	**0.4894**	**0.5450**
Neutral Network	0.6090	0.5049	0.4657	0.4710	0.5158

Para selecionar os melhores modelos foi dada prioridade para a acurácia, pois como o objetivo é maximizar o acerto dentro do modelo sem dar prioridade para uma das categorias, essa métrica vai sanar esse objetivo. Como dito anteriormente, para facilitar o entendimento dos resultados, a vitória do mandante foi definida como classe positiva e o empate/vitória do visitante foi definido como classe negativa. A precisão e o NPV foram utilizados pois, um outro objetivo do trabalho, além de fazer uma aposta que gere lucro, é não realizar uma aposta que gere prejuízo. Por esse motivo, como essas duas métricas calculam os acertos baseados nos seus falsos positivos e falsos negativos, respectivamente, quanto mais próximo de 1 for o resultado, mais temos os verdadeiros positivos e verdadeiros negativos se sobressaindo sobre os falsos positivos e falsos negativos. Um modelo que maximize essas métricas é um modelo que vai evitar que se faça apostas em 'falsos resultados'. Além disso o F1 foi utilizado para ter um panorama melhor dos resultados de precisão e sensibilidade das duas classe. A maioria dos algoritmos tiveram resultados muito semelhantes, com ligeiras variações entre as métricas, mas ainda assim é possível verificar os que foram melhores em um panorama geral. Fazendo uma análise por eliminatória, os modelos de *Decision Tree* e *Naive Bayes* foram deixados de lado por terem uma baixa acurácia comparados aos outros. Por fim os modelos de *Logistic Regression* e *Random Forest* foram escolhidos por terem na média melhor precisão, NPV e F1 para as duas classes em relação aos outros 2. Na regressão foi definido o algoritmo *XGBoosting*, mas não foram feitas comparações porque esse foi o único algoritmo de regressão utilizado. Não foi realizada uma seleção prévia, só comparamos esse algoritmo com os outros dois modelos de classificação selecionados no teste real nas casas de apostas.

5.3.1 Escolha dos Hiperparâmetros

Para escolher os hiperparâmetros foi utilizada a biblioteca hyperopt do Python, onde são escolhidos os melhores valores a partir de um conjunto pré-definido utilizando

uma métrica de forma comparativa. Nesse caso, a métrica utilizada para a classificação foi a acurácia e para a regressão foi o erro quadrático médio que calcula a diferença entre o valor predito e o real. Quanto mais próximo de zero for esse valor, melhor é o algoritmo. O conjunto de valores definidos antes da execução do algoritmos são mostrados nas Figuras 5.4, 5.5 e 5.6, além dos hiperparâmetros escolhidos por cada cluster e pelo conjunto com todos os elementos são mostrados nas figuras 5.7, 5.8 e 5.9.

Figura 5.4 – Conjunto teste hiperparâmetros Regressão Logística

```
"solver": hp.choice("solver", ["lbfgs", "liblinear", "newton-cg", "newton-cholesky", "sag", "saga"]),
"max_iter": hp.choice("max_iter", [100, 200, 300, 400, 500, 600, 700, 800, 900, 1000]),
"tol": hp.uniform("tol", 0.0001, 0.01),
"intercept_scaling": hp.choice("intercept_scaling", [1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20]),
"C" : hp.uniform("C", 0, 20),
```

Figura 5.5 – Conjunto teste hiperparâmetros Floresta Aleatória

```
"criterion": hp.choice("criterion", ["gini", "entropy", "log_loss"]),
"max_depth": hp.choice("max_depth", [10, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000]),
"max_features": hp.choice("max_features", ["sqrt", "log2", None]),
"min_samples_leaf": hp.choice("min_samples_leaf", [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]),
"min_samples_split": hp.choice("min_samples_split", [2, 3, 4, 5, 6, 7, 8, 9, 10, 11]),
"n_estimators": hp.choice("n_estimators", [10, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000]),
```

Figura 5.6 – Conjunto teste hiperparâmetros *XGBoosting*

```
"learning_rate": hp.choice("learning_rate", ["constant", "invscaling", "adaptive"]),
"learning_rate_init": hp.uniform("learning_rate_init", 0.001, 0.1),
"max_iter": hp.choice("max_iter", [100, 200, 300, 400, 500, 600, 700, 800, 900, 1000]),
"momentum": hp.uniform("momentum", 0, 1),
"solver": hp.choice("solver", ["lbfgs", "sgd", "adam"]),
"tol": hp.uniform("tol", 0.0001, 0.01),
```

Figura 5.7 – Hiperparâmetros utilizados Regressão Logística

```
Todos:     LogisticRegression(C = 0.02837672781180345, intercept_scaling = 13, max_iter = 500, solver = "sag", tol = 0.007066654416181151)
Cluster 1: LogisticRegression(C = 0.25662478309802694, intercept_scaling = 16, max_iter = 400, solver = "saga", tol = 0.00549547593035501)
Cluster 2: LogisticRegression(C = 0.41536254133900175, intercept_scaling = 3, max_iter = 800, solver = "newton-cg", tol = 0.0070070767274921925)
Cluster 3: LogisticRegression(C = 0.5520056996405969, intercept_scaling = 5, max_iter = 800, solver = "liblinear", tol = 0.002646831232970823)
```

Figura 5.8 – Hiperparâmetros utilizados Floresta Aleatória

```
Todos:     RandomForestClassifier(criterion = "entropy", max_depth = 200, max_features = "log2", min_samples_leaf = 9, min_samples_split = 5, n_estimators = 300)
Cluster 1: RandomForestClassifier(criterion = "entropy", max_depth = 500, max_features = "log2", min_samples_leaf = 8, min_samples_split = 6, n_estimators = 500)
Cluster 2: RandomForestClassifier(criterion = "gini", max_depth = 50, max_features = "sqrt", min_samples_leaf = 11, min_samples_split = 2, n_estimators = 600)
Cluster 3: RandomForestClassifier(criterion = "entropy", max_depth = 600, max_features = "log2", min_samples_leaf = 10, min_samples_split = 10, n_estimators = 600)
```

Figura 5.9 – Hiperparâmetros utilizados *XGBoosting*

```
Todos:     XGBRegressor(colsample_bytree = 0.6644918116199707, gamma = 7.386382266259656, max_depth = 1200, min_child_weight = 3.0, n_estimators = 700, reg_alpha = 178.0, reg_lambda = 0.31027966220258874)
Cluster 1: XGBRegressor(colsample_bytree = 0.8603153047041996, gamma = 1.630283981572731, max_depth = 1000, min_child_weight = 3.0, n_estimators = 500, reg_alpha = 131.0, reg_lambda = 0.90293917494035)
Cluster 2: XGBRegressor(colsample_bytree = 0.7400940735133774, gamma = 4.242558922823842, max_depth = 10, min_child_weight = 0.0, n_estimators = 50, reg_alpha = 84.0, reg_lambda = 0.4620139292601921)
Cluster 3: XGBRegressor(colsample_bytree = 0.9992066601170332, gamma = 7.198826927503546, max_depth = 300, min_child_weight = 9.0, n_estimators = 50, reg_alpha = 126.0, reg_lambda = 0.3845636021613974)
```

5.3.2 Analise dos modelos escolhidos

Para fazer uma análise mais aprofundada dos algoritmos, foram criadas duas tabelas comparativas dos modelos selecionados, tabelas 5.7 e 5.8. Nos dois modelos, é

possível notar que separar os dados em subconjuntos gerou melhora na acurácia para os clusters 1 e principalmente o 3, porém esse teve uma queda de desempenho na precisão. Apesar disso, como a acurácia é o nosso principal ponto de análise, uma possível utilização do cluster 3 para fazer as apostas pode gerar melhor resultado que apostas usando todos os dados em conjunto, para ambos algoritmos. Em compensação, verificando o cluster 2, ele cai de desempenho em todas as métricas em comparação ao conjunto inteiro para os dois modelos,

Tabela 5.7 – Comparação Linear Regression

Partições	**Acurácia**	**Precisão**	**NPV**	**F1 1**	**F1 2**
Completo	0.6244	0.6271	0.6230	0.5564	0.6697
Cluster 1	**0.6371**	0.6416	0.6135	0.7655	0.1053
Cluster 2	0.5871	0.5696	0.5987	0.4975	0.6202
Cluster 3	**0.6598**	0.5227	0.6727	0.1384	0.7849

Tabela 5.8 – Comparação Random Forest

Partições	**Acurácia**	**Precisão**	**NPV**	**F1 1**	**F1 2**
Completo	0.6138	0.6041	0.6222	0.5626	0.6548
Cluster 1	**0.6350**	0.6373	0.5625	0.7706	0.2171
Cluster 2	0.6051	0.5919	0.6131	0.5312	0.6584
Cluster 3	**0.6560**	0.45	0.6634	0.0588	0.7935

5.4 Implementação

Nessa etapa, foram utilizados os 2 melhores modelos de classificação selecionados e o modelo de regressão em uma simulação real do mundo de apostas. Essa simulação foi realizada fazendo predição das partidas do Campeonato Brasileiro de 2023. Foi criada uma nova tabela de dados contendo as odds para 2 mercados de apostas:

- Resultado da partida: nesse mercado são dadas 3 alternativas de resultado (vitória mandante, empate e vitória visitante) e suas odds para cada possibilidade e ganha a aposta que acertar qual resultado vai ocorrer.
- Dupla chance: nesse mercado também 3 alternativas de resultado, mas com uma chance dupla em cada uma delas. As alternativas são a vitória do mandante ou empate, vitória do visitante ou empate e vitória de um dos dois times envolvidos na partida. Por dar duas chances dentro de uma única aposta acabam normalmente com as odds menores e gerando menos lucro em comparação a outros mercados em caso de acerto.

Para a classificação, foi realizada a junção desses dois mercados, escolhida uma alternativa de cada um deles e transformada em uma das categorias de classificação definidas neste trabalho. Por fim, temos como categoria 1 a vitória do mandante (mercado 'Resultado da partida') e como categoria 2 a vitória do visitante ou empate (mercado 'Dupla chance'. As categorias ficaram dessa forma, pois na etapa de análise do trabalho foi definida que uma abordagem de predição binária para o modelo seria mais eficaz do que uma com múltiplas categorias. Para a regressão, temos um atributo alvo contínuo, então a análise foi mais focada na vitória do time mandante, trazendo as odds relacionadas a essa possibilidade do mercado 'Resultado da partida'.

5.4.1 Porcentagem de Equilíbrio

Para se dar bem no mundo das apostas, é importante também entender ao menos um pouco de probabilidade. A porcentagem de equilíbrio é a quantidade de vezes que você deve ganhar uma aposta para não ganhar nem perder dinheiro ao longo do tempo. Por exemplo, a probabilidade de girar um dado e dar um determinado número tem 1 chance em 5, isso representa uma porcentagem de equilíbrio de 16,7%. Se a casa de apostas achar que a probabilidade de ganhar é 1 para 4(20%), então você já está perdendo porque a casa está calculando uma chance maior do que a que você realmente tem. Se a casa achar que a chance é 1 para 7(12,5%) então você vai ganhar mais vezes essa aposta do que a casa prevê. Trazendo para o futebol, o algoritmo de Regressão Logística calculou que a chance de acertar uma partida é de 65%, então as apostas neste trabalho são baseadas nas odds que preveem uma probabilidade menor do que os 65% que o modelo criado nesse caso prevê, pois teoricamente, ao longo do tempo pelo menos prejuízo não teremos com as apostas.

5.4.2 Simulação

Neste ponto, foi realizada a simulação dos algoritmos selecionados em um cenário real de casas de apostas. Foi criada uma tabela com as odds da *Bet365*, uma das maiores casas de apostas do mundo, realizadas as execuções dos modelos e comparadas com as odds. Na regressão foi calculada a odd para o time mandante ganhar a partida a partir das odds já conhecidas dos jogos dos campeonatos anteriores, então se a odd prevista for

menor que a da casa de apostas, a aposta será realizada. Na classificação, se os modelos previam um resultado e a probabilidade do time ganhar representado pelas odds tem uma porcentagem de equilíbrio menor do que a acurácia prevista nos algoritmos, então essa aposta será realizada. Nas Tabelas 5.11, 5.10 e 5.9, é possível visualizar alguns dados relacionados às apostas feitas e o lucro/prejuízo que obtivemos no final, sempre baseando as apostas nas partidas disputadas no Campeonato Brasileiro de Futebol de 2023. Todos os resultados relacionados ao valor retornado e o lucro/prejuízo se baseiam em uma aposta realizada com o valor de R$ 1,00.

Tabela 5.9 – Análise XGBoosting

Métricas	Clusterizado			Não Clusterizado
	Cluster 1	Cluster 2	Cluster 3	
Apostas Feitas	40	55	57	194
Acertos	15	29	19	85
Valor Retornado	31.21	50.23	52.50	172.88
Porcentagem Acerto	37.5%	52.72%	33.33%	43.81%
Lucro/Prejuízo	-8.79	-4.77	-4.50	-21.11

Iniciando a análise com o *XGBoosting*, é possível visualizar que trazendo esse algoritmo para o mundo real obtivemos um prejuízo em todos os conjuntos de dados utilizados. Além disso, tivemos uma porcentagem de acerto muito baixa para 3 dos 4 conjuntos testados, e mesmo o conjunto que ultrapassou 50% de acerto ainda assim acabou dando prejuízo. Acredito que o possível motivo do prejuízo seja seu atributo alvo. No caso de prever a vitória de um time, mapear um atributo alvo que faça sentido para os modelos de regressão é uma tarefa difícil, pois até mesmo focando no objetivo das casas de apostas no mercado estudado, é possível perceber que ela se aproxima muito mais de uma tarefa classificatória.

Tabela 5.10 – Análise Floresta Aleatória

Métricas	Clusterizado			Não Clusterizado
	Cluster 1	Cluster 2	Cluster 3	
Apostas Feitas	75	32	81	159
Acertos	39	12	38	85
Valor Retornado	70.97	23.10	72.08	158.73
Porcentagem Acerto	52%	37.5%	46.91%	53.45%
Lucro/Prejuízo	-4.02	-8.89	-8.91	-0.26

No modelo criado com Floresta Aleatória, nas execuções mostradas também foi possível notar um prejuízo em todos os conjuntos testados. No caso do conjunto sem

nenhum agrupamento, algumas execuções deram um certo lucro, porém não acredito que esse modelo seja confiável o bastante para ser aplicado no mundo real, pois a inconstância de causar prejuízo em um momento e lucro em outro não convence para ser utilizado em mais testes. O acerto para todos os conjuntos também ficaram bem abaixo da acurácia prevista inicialmente com os dados de 2018 a 2022, dando indícios de um possível *overfiting* relacionados a execução inicial sem os dados do Brasileirão 2023, mesmo com todos os processamentos feito para tentar retirar essa possibilidade.

Tabela 5.11 – Análise Regressão Logística

Métricas	Clusterizado			Não Clusterizado
	Cluster 1	Cluster 2	Cluster 3	
Apostas Feitas	76	32	81	173
Acertos	36	19	48	91
Valor Retornado	65.95	40.56	93.25	171.53
Porcentagem Acerto	47.36%	59.37%	59.25%	52.60%
Lucro/Prejuízo	-10.05	+8.56	+12.25	-1.46

No modelo com Regressão Logística foi possível visualizar resultados mais animadores. Apesar de ter se obtido prejuízo no Cluster 1 e no conjunto total de dados e um acerto relativamente menor do que a acurácia calculada anteriormente, os grupos 2 e 3 apresentaram dados mais interessantes para possível aplicação, além de todos os conjuntos terem um acerto médio maior que os outros modelos testados, mesmo nos que causam prejuízo. O cluster 2 apresentou uma taxa de acertos até melhor que os testes iniciais feitos, que anteriormente eram de 58% e agora aumentou ligeiramente para 59%. O cluster 3 apresentou uma taxa relativamente menor em relação aos testes iniciais, mas ainda assim proporcionou lucro e foi o conjunto que apresentou o melhor lucro dentre todos os testes. Acredito que os modelos ainda carecem de mais testes para ter sua eficácia comprovada, mas os resultados apresentados pela Regressão Logística para os 2 clusters que geraram lucro indicam um possível bom modelo para ser aplicado em casas de apostas.

5.5 Considerações Finais

Neste capítulo, foram apresentados todos os experimentos realizados durante o desenvolvimento deste trabalho e a análise dos seus resultados. Primeiro, foi apresentada toda a etapa de pré-processamento, desde a captação dos dados, até a aplicação dos algoritmos de normalização e *K-Means* para posterior aplicação nos modelos de predição.

Depois foram mostrados os resultados iniciais de cada algoritmo e realizada uma comparação entre todos e selecionados os melhores para serem aplicados em um cenário real de casas de apostas para verificar o possível lucro, ou prejuízo que eles proporcionaram.

6 CONCLUSÃO

Neste trabalho, foi explorada a aplicação de vários algoritmos de AM com o objetivo de fazer a predição de partidas do Campeonato Brasiliero de Futebol e aplicar em casas de apostas esportivas. Ao final fizemos um comparativo dos 2 melhores modelos de classificação e 1 modelo de regressão para ver o possível lucro/prejuízo que os modelos proporcionaram no campeonato de 2023. Foram utilizadas como base de dados uma série de estatísticas e delas criadas variáveis para tentar fazer a predição dos resultados. Para o modelo de regressão acredito que o atributo alvo escolhido acabou gerando um resultado aquém do esperado, que foi uma tentativa de usar as boas previsões das casas de apostas ao favor dos modelos e inicialmente parecia uma boa escolha. Porém, achar um atributo alvo contínuo que faça sentido para a regressão e que depois auxilie a realização de uma aposta que não tenha alternativas também contínuas é bem difícil e não consegui pensar em outra alternativa a não ser tentar utilizar as odds para isso.

Pensando no mercado de apostas que foi utilizado, tratando-se de um mercado com apostas categóricas, acredito que por esse motivos tivemos um resultado melhor em um dos algoritmos de classificação utilizados. Os dados do campeonato de 2023 testados acabaram revelando um *overfitting* em relação aos dados treinados anteriormente, isso pode ser um pouco minimizado imaginando que aos fazer apostas, basicamente faremos isso de forma intervalada, então a cada rodada ganharemos 10 novas instâncias para serem colocados nos dados e auxiliarem em uma melhor predição, podendo causar também um lucro maior do que o visto neste trabalho e diminuir o *overfitting*. Além disso, a predição de partidas de futebol continua sendo uma tarefa muito difícil, existe um fator relacionado a sorte que em vários momentos acaba se sobressaindo sobre dados estatísticos (AOKI; ASSUNCAO; MELO, 2017). Os dados utilizados, apesar de mostrarem uma série de qualidades sobre o jogo, ainda não mostram muitos fatores que são preponderantes para o desenrolar de uma partida, como qualidade do time que está em campo, podendo existir desfalques que mudam totalmente o resultado de uma partida, o foco do time no campeonato, convocações, o estilo de jogo e até fatores clínicos que influenciam no resultado final.

Como trabalhos futuros, acreditamos que adicionar mais dados estatísticos da partida não causariam uma diferença muito maior do que os resultados apresentados neste trabalho, mas fatores externos a isso, que precisam ser pensados e explorados mais a fundo, poderiam ser testados e teriam potencial de causar um crescimento positivo nos

resultados. Além do dito 'imponderável do futebol', é um fator relacionado a sorte que realmente não é possível de ser previsto. Por fim, acreditamos que uma boa expansão para este trabalho é utilizar os mesmo dados mas explorar outros mercados com resultados menos restritos e focando um pouco mais em modelos de regressão. Por exemplo, imaginamos que um mercado como o de 'número de escanteios', que fornece odds para uma aposta que prevê mais ou menos escanteios que um dado número seria um boa alternativa para ser previsto com modelos de regressão. Existem vários outros mercados mais amplos e que também podem ser explorados, como 'número de chutes no gol' e 'número de laterais' que funcionam como os escanteios e podem trazer um bom resultado com a utilização da regressão.

REFERÊNCIAS

ALMEIDA, M. **Pandas Python: o que é, para que serve e como instalar**. 2023. Disponível em: <https://www.alura.com.br/artigos/pandas-o-que-e-para-que-serve-como-instalar>.

AOKI, R. Y.; ASSUNCAO, R. M.; MELO, P. O. V. de. **Luck is Hard to Beat: The Difficulty of Sports Prediction**. 2017. Disponível em: <https://arxiv.org/abs/1706.02447>.

BOUCINHA, I. M. Um modelo de previsão de resultados de futebol utilizando machine learning. 2023. Disponível em: <https://lume.ufrgs.br/handle/10183/261788>.

COELHO, P. V. **Escola brasileira de futebol**. [S.l.]: Objetiva, 2018. ISBN 9788547000578.

DUTT, S.; CHANDRAMOULI, S. **Machine Learning**. [S.l.]: PEARSON INDIA, 2018. ISBN 9353066697.

ELO, A. E. **The Rating of Chess Players, Past and Present**. [S.l.]: Ishi Press, 2008. ISBN 9780923891275.

FACELI, K. et al. **Inteligência Artificial: uma abordagem de Aprendizado de Máquina (2a edição)**. [S.l.: s.n.], 2021. ISBN 9788521637493.

FIFA, F. I. d. F. A. The history of beach soccer. 2020. Disponível em: <https://www.fifa.com/tournaments/mens/beachsoccerworldcup/russia2021/news/the-history-of-beach-soccer>.

FONTANA, É. Introdução aos algoritmos de aprendizagem supervisionada. 2020. Disponível em: <https://fontana.paginas.ufsc.br/files/2018/03/apostila_ML.pdf>.

FOUNDATION, P. S. **A Biblioteca Padrão do Python**. 2023. Disponível em: <https://docs.python.org/pt-br/3/library/>.

FOUNDATION, P. S. **Python Geral**. 2023. Disponível em: <https://docs.python.org/pt-br/dev/faq/general.html#what-is-python>.

FOUNDATION, P. S. **sqlite3 — Interface DB-API 2.0 para bancos de dados SQLite**. 2023. Disponível em: <https://docs.python.org/pt-br/3/library/sqlite3.html>.

GABRIEL, J.; SALDAÑA, P. **Apostas esportivas atraem jovens e chegam a 15Disponível em: <https://www1.folha.uol.com.br/esporte/2024/01/apostas-atraem-jovens-e-chegam-a-15-da-populacao-que-diz-gastar-r-263-por-mes-mostra-datafolh shtml>.**

GOOGLE. **Olá, este é o Colaboratory**. 2023. Disponível em: <https://colab.research.google.com/notebooks/welcome.ipynb?hl=pt-BR#scrollTo=lSrWNr3MuFUS>.

GRANCHI, G. **Por que jogos de azar são proibidos e sites de apostas são permitidos no Brasil?** 2023. Disponível em: <https://www.bbc.com/portuguese/articles/ce7g64gx1r9o#:~:text=Mudan%C3%A7as%20recentes%20na%20legisla%C3%A7%C3%A3o%20brasileira,tratava%20todos%20de%20maneira%20igual.>

GUPTA, M. Ml | introduction to data in machine learning. **https://www.geeksforgeeks.org/ml-introduction-data-machine-learning/**, 2023.

HAN, J.; PEI, J.; KAMBER, M. **Data Mining: Concepts and Techniques**. 3. ed. [S.l.]: Morgan Kaufmann, 2011. ISBN 9780123814807.

IBJR, I. B. d. J. R. **História das apostas no Brasil**. 2023. Disponível em: <https://ibjr.org/informe-se/historia-apostas-brasil/>.

IBM. O que são redes neurais? 2023. Disponível em: <https://www.ibm.com/br-pt/topics/neural-networks>.

JUNIOR, E. C. V. **Normalização de variáveis**. 2020. Disponível em: <https://geam.paginas.ufsc.br/files/2020/02/feature-scaling.pdf>.

MAGATTI, R. **Como os sites de apostas se tornaram o maior financiador do futebol brasileiro**. 2023. Disponível em: <https://www.estadao.com.br/esportes/futebol/como-os-sites-de-apostas-se-tornaram-o-maior-financiador-do-futebol-brasileiro/>.

MARTINS, C. **Manual de Análise de Dados Quantitativos com Recurso ao IBM SPSS Saber decidir, fazer, interpretar e redigir**. [S.l.]: Psiquilibrios, 2011. ISBN 9789898333087.

MELO, C. R. G. Utilizando aprendizado de máquina para predição de resultados da nba. 2021. Disponível em: <https://www.maxwell.vrac.puc-rio.br/57527/57527.PDF>.

MILLER, E.; DAVIDOW, M. **The Logic Of Sports Betting**. [S.l.]: Ishi Press, 2019. ISBN 9780923891275.

MILLS, J. **Charles Miller: O pai do futebol brasileiro**. [S.l.]: Panda Books, 2005. ISBN 9788587537997.

MITCHELL, T. M. **Machine Learning**. USA: McGraw-Hill Science/Engineering/Math, 1997. ISBN 0070428077.

MOSCA, H. M. B. Fatores institucionais e organizacionais que afetam a profissionalização da gestão do departamento de futebol dos clubes. 2006. Disponível em: <https://www.maxwell.vrac.puc-rio.br/colecao.php?strSecao=resultado&nrSeq=9440@1>.

PÉREZ-ORTEGA, J. et al. The k-means algorithm evolution. 2019. Disponível em: <https://www.semanticscholar.org/reader/da5b2e345b0e1113a46e31b53f433a1db4791131>.

PÓVOA, L. et al. **O Mercado de Apostas Esportivas On-line: impactos, desafios para a definição de regras de funcionamento e limites**. [S.l.], 2023. Disponível em: <https://www12.senado.leg.br/publicacoes/estudos-legislativos/tipos-de-estudos/textos-para-discussao/td315>.

RATINGS, W. F. E. **World Football Elo Ratings**. 2024. Disponível em: <https://www.eloratings.net/about>.

ROCHA, E. C. da. O aspecto social da iconografia do futebol e estudo de caso das agremiações desportivas cariocas. 2008. Disponível em: <https://www.maxwell.vrac.puc-rio.br/colecao.php?strSecao=resultado&nrSeq=12380@1>.

ROLLINS, J. B. **Metodologia de Base para Ciência de Dados**. 2015. Disponível em: <https://www.ibm.com/downloads/cas/B1WQ0GM2>.

SCIKIT-LEARN, D. **Começando**. 2023. Disponível em: <https://scikit-learn.org/stable/getting_started.html>.

SEHNEM, R. et al. Análise de variáveis em partidas de futebol: Previsão de resultados com naïve bayes e poisson. 2021. Disponível em: <https://sol.sbc.org.br/index.php/eniac/article/view/18237>.

SENADO, A. **Regulamentação de apostas esportivas será analisada pelo Senado**. [S.l.], 2023. Disponível em: <https://www12.senado.leg.br/noticias/materias/2023/09/15/regulamentacao-de-apostas-esportivas-sera-analisada-pelo-senado>.

SERVICES, A. W. **O que é preparação de dados?** 2023. Disponível em: <https://aws.amazon.com/pt/what-is/data-preparation/>.

SHEETS, G. **Tome decisões baseadas em dados com o Google Sheets**. 2024. Disponível em: <https://www.google.com/sheets/about/>.

SIMõES, A. **Maioria dos brasileiros não sabia da obrigatoriedade de clubes da Série A terem times femininos**. 2023. Disponível em: <https://www.cnnbrasil.com.br/esportes/maioria-dos-torcedores-brasileiros-nao-sabia-da-obrigatoriedade-de-clubes-da-serie-a-terem-times-femin#:~:text=Desde%202019%2C%20todos%20os%20clubes,Campeonato%20Brasileiro%20ter%C3%A3o%20essa%20obrigatoriedade.>

WEBB, G. I.; SAMMUT, C. **Encyclopedia of Machine Learning**. [S.l.]: Springer, 2010. ISBN 9780387345581.

WESTIN, R. **Futebol feminino já foi proibido no Brasil, e CPI pediu legalização**. [S.l.], 2023. Disponível em: <https://www12.senado.leg.br/noticias/especiais/arquivo-s/futebol-feminino-ja-foi-proibido-no-brasil-e-cpi-pediu-legalizacao#:~:text=Por%20mais%20de%2040%20anos,as%20condi%C3%A7%C3%B5es%20de%20sua%20natureza%E2%80%9D.>

WOLPERT, D. H.; MACREADY, W. G. No free lunch theorems for optimization. **https://www.cs.ubc.ca/ hutter/earg/papers07/00585893.pdf**, 1997.