UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE ENGENHARIA DE COMPUTAÇÃO

IAGO MARTINS BOUCINHA

Um modelo de previsão de resultados de futebol utilizando Machine Learning

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

Orientador: Prof. Dra. Renata Galante

Porto Alegre
2023

AGRADECIMENTOS

Agradeço primeiramente à minha família, que me motivou e apoiou durante toda esta longa etapa chamada graduação.

Agradeço também a todos os professores que me auxiliaram durante o curso, em especial à professora Renata Galante, orientadora deste trabalho. Todos professores demonstraram um profundo conhecimento dentro das suas áreas de atuação e foram fundamentais para o meu desenvolvimento.

Muito Obrigado.

RESUMO

Nos últimos anos, os eventos esportivos vêm se tornando cada vez mais relevantes dentro da sociedade, cativando mais e mais pessoas. Esse movimento gerou um crescimento exponencial de casas de apostas relacionadas a essa modalidade, cada uma com seu próprio esquema de precificação. Este trabalho se propõe a criar um modelo de precificação utilizando aprendizado de máquina para prever os possíveis resultados e probabilidades de uma partida de futebol. O modelo proposto será avaliado utilizando métricas clássicas da área de aprendizado de máquina e também será comparado com as probabilidades ofertadas nas casas de aposta. A criação de um modelo robusto, capaz de encontrar falhas de precificação dentro dos eventos, possibilita uma grande vantagem competitiva, podendo ser utilizado para ajustes na precificação ou até mesmo para se obter lucro no mercado de apostas esportivas. Os resultados que serão apresentados nos próximos capítulos reforçam a teoria de que existe muito potencial a ser explorado neste segmento, porém não é uma tarefa fácil.

Palavras-chave: Futebol. apostas esportivas. aprendizado de máquina.

A model to predict outcomes of soccer matches using machine learning

ABSTRACT

In recent years, sporting events have become increasingly relevant within society, captivating more and more people. This movement generated an exponential growth of bookmakers related to this modality, each one with its own pricing scheme. This work aims to create a machine learning model to predict the results and probabilities of a football match. The proposed model will be evaluated using classic metrics from the area of machine learning and the odds offered by bookmakers will also be compared. The creation of a robust model, able to find misfits prices, would allow a great competitive advantage, being able to be used for market pricing or even to obtain profit in the sports betting market. The results that will be presented in the next chapter reinforce the theory that there is a lot of potential to be explored in this segment, but is not an easy task.

Keywords: soccer, sports betting, machine learning.

LISTA DE ABREVIATURAS E SIGLAS

SQL	Structured Query Language - Linguagem de consulta estruturada
Odds	Cotações oferecidas para uma aposta
LR	Logistic Regression
KNN	K Neighbors Classifier
DT	Decision Tree
NB	GaussianNB
XGB	XGB Classifier
RF	Random Forest Classifier

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Os esportes estão inseridos em nossa sociedade desde as épocas mais primitivas, Victor Matheson (2021) afirma que eventos como lutas entre gladiadores e corridas de bigas já eram realizadas na Grécia antiga, atraindo um grande número de espectadores. O interesse em analisar e prever tais eventos sempre esteve presente no mundo esportivo e, o acesso à informação facilitado que temos hoje em dia, aumentou consideravelmente o número de pessoas provendo diferentes pontos de vista.

Alguns espectadores utilizam seus conhecimentos para tentar alcançar retornos financeiros através das apostas, algo popular desde a época dos gladiadores. Segundo o relatório da consultoria *Grand View Research*, o mercado de apostas esportivas foi avaliado em quase 70 bilhões de dólares em 2020 e tem previsão de ultrapassar 140 bilhões até o ano de 2028. Apenas uma partida de futebol da Premier League é responsável por movimentar, em média, cerca de 1 milhão[1]. Dito isto, precificar corretamente as probabilidades de um determinado resultado é algo fundamental para manter a saúde financeira das casas de apostas. Os avanços tecnológicos tem proporcionado muitas ferramentas para a realização deste trabalho, disponibilizando desde o acesso aos dados de forma simplificada até modelos de aprendizado complexos como é o caso do *Tensorflow*, uma biblioteca de código aberta utilizada para criação de redes neurais. Atualmente cada casa de aposta aparenta possuir seu próprio modelo de precificação, visto que as cotações dificilmente são iguais entre elas. Esses métodos não são públicos por serem considerados os responsáveis pelo lucro da empresa. Dito isso, sabe-se apenas que os preços se baseiam em dois fatores chave, informações históricas de performance dos times e o volume de aposta em cada um dos resultados.

A análise de uma partida envolve diversos fatores. Alguns são facilmente mensuráveis, enquanto outros são mais subjetivos, demandando um trabalho mais detalhado. Analisar cuidadosamente todos os jogos diários é uma tarefa praticamente impossível de ser feita manualmente, isso tornou as automações muito atraentes nesse segmento. O aprendizado de máquina é uma das técnicas mais poderosas que temos para automatizar processos e decisões de alta complexidade, capacitando as máquinas a tomar decisões tão boas quanto um especialista no assunto.

O presente trabalho contribuirá com a criação de um modelo de probabilidades para futebol utilizando aprendizado de máquina, onde as principais ligas do mundo serão

[1]Fonte:https://bleacherreport.com/articles/2200795-mugs-and-millionaires-inside-the-murky-world-of-professional-football-gambling

analisadas a fim de encontrar possíveis falhas nas probabilidades disponíveis no mercado. O histórico de partida dos campeonatos será usado para criação de novas variáveis, buscando assim melhores resultados nos diferentes tipos de modelos.

O restante do trabalho está organizado da seguinte forma: O capítulo 2 apresenta o referencial teórico que embasou o desenvolvimento deste trabalho, referenciando trabalhos e tecnologias relacionadas. O capítulo 3 descreve, de forma geral, o trabalho e o banco de dados utilizado, detalhando o processo de manipulação e análise dos dados. O capítulo 4 apresenta a avaliação experimental, onde são descritos todos os experimentos realizados e seus respectivos resultados. O capítulo 5 apresenta as conclusões gerais do trabalho e, adicionalmente, melhorias mapeadas para uma possível continuação deste estudo.

2 FUNDAMENTAÇÃO TEÓRICA E TRABALHOS RELACIONADOS

Este capítulo é responsável por apresentar toda base teórica utilizada para o desenvolvimento do trabalho. Também são apresentadas as tecnologias utilizadas durante todas as etapas, desde a criação da base de dados até o desenvolvimento dos modelos. Por fim, são apresentados alguns trabalhos relacionados.

2.1 Futebol:Histórico e Concepções

O futebol é um esporte conhecido e praticado no mundo inteiro, somando diversas teorias sobre o seu surgimento na sociedade. Sua história teve início ainda no período da pré-história, não foi preciso muito tempo para que se consagrasse como o esporte das multidões. Aos finais de semana, o futebol era sinônimo de diversão, mesmo que apresentando alguns modos diferentes de jogar, de acordo com suas localidade (CASTELLANI FILHO, 2019).

Silva (2020) apresenta que o futebol foi introduzido no Brasil em meados do século XIX, por Charles Miller, um estudante paulista que retornou da Inglaterra nos anos 90 e trouxe consigo diversos artigos do futebol, tais como as bolas, uniformes e ainda livros que apresentavam os regulamentos e funcionamentos dos jogos, o que fez com que Miller fosse conhecido como pai do esporte no Brasil. Nessa propositura, o futebol rapidamente se propagou por todo o país, o que o tornou popularizado na sociedade. Inicialmente a adesão ocorreu de forma elitizada, ou seja, ele era restrito à aristocracia da sociedade brasileira, mas com o crescimento urbano do país, logo esse esporte se popularizou para todas as classes e pessoas, inclusive na criação da organização dos clubes de futebol.

Ao longo dos anos, o Brasil se tornou uma grande potência no esporte, fazendo com que a seleção brasileira masculina se tornasse a detentora de 5 (cinco) títulos mundiais, entre os anos de 1958, 1962, 1970, 1994 e 2002. Além disso, diversos clubes brasileiros tornaram-se donos de títulos mundiais, dentre outras vitórias (SILVA, 2020).

O futebol é um esporte bem conhecido e popular no Brasil, sendo considerado um fenômeno em decorrência de sua popularidade, o que vem se passando ao longo das gerações e com influência da mídia, visto que também condiciona a lucratividade desse esporte. Logo essas influências se dissiparam para as escolas, em que o futebol é tem uma grande demanda pelos alunos nas aulas de educação física (CASTELLANI FILHO,

2019).

De acordo com Tubino (2016) o futebol é fenômeno sociocultural, onde a competição é considerada o elemento essencial, contribuindo com a formação e a aproximação dos seres humanos, reforçando o desenvolvimento de valores como a moral, a ética, a solidariedade, a fraternidade e a cooperação.

2.2 Apostas Esportivas

Soares (2019) apresenta que a aposta é como um acordo bilateral, ou seja, entre duas ou mais pessoas de opiniões divergentes em relação a um determinado fato, hipótese ou acontecimento futuro. Devendo aquele que não estiver certo, pagar algo previamente convencionando, ou do contrário, ganhar a quantia ou a coisa que se aposta.

É importante mencionar sobre as incertezas e complexidades envolvendo os jogos de azar. Mesmo que um evento tenha quase 100% (cem porcento) de chance, é impossível afirmar com certeza que ele ocorrerá, sendo assim, passível de apostas. Uma aposta em evento esportivo, geralmente chamada como jogo de azar, muito se assemelha aos próprios investimentos no mercado financeiro e de seus derivados (KELNER, 2016).

Conforme foi mencionado anteriormente, os jogos existem desde os tempos primordiais. De acordo com o estudo levantado por Aquino (2022), os primeiros registros dos jogos de azar, atenua o conhecimento que ocorreu na China, em meados de 2300 a.C. e que foi considerada uma era que as apostas começaram a se destacar, utilizadas principalmente, em questões de conquistas de territórios.

No entanto, alguns arqueólogos já davam indícios sobre os objetos antigos dos jogos de azar em sociedades na Índia, Egito e na Grécia, destacando as pinturas encontradas em tumbas no Egito, onde os jogos de azar eram explorados dentro das limitações do período. Foram encontrados tabuleiros esculpidos em mármores e as mesas de pedras que serviam como base para a diversão de homens ociosos (AQUINO, 2022).

Ao longo dos anos, os jogos e apostas foram sendo aprimorados, passando inclusive por proibições em alguns locais, em que as leis eram brandas e ineficazes, pois as multas eram de baixos valores, acarretando apenas o título perante a sociedade que era contra as apostas, marginalizando a atuação das casas de apostas e dos jogos de azar.

De acordo com as pesquisas realizadas por Lira (2018) as apostas esportivas começaram a se destacar e rapidamente se propagar com a expansão da internet, em que as empresas deste setor começaram a surgir em praticamente todos os países. Fazendo com

que empresas também desenvolvessem técnicas próprias para obter suas licenças e atrair apostadores do mundo inteiro.

As apostas são condizentes com os jogos de azar, que possuem uma variedade de conceitos que podem ser utilizados de forma pejorativa. O mais conhecido da sociedade é a contravenção penal, apresentando consigo a diplomacia legal que acentua a sorte como fator fundamental para vitória, porém são diferentes, as apostas esportivas não são eventos que dependem exclusivamente de sorte. Devido as inúmeros formas de se analisar um determinado jogo, é possível utilizar-se de métodos e estudos como base para se obter uma maior previsibilidade daquele evento, reduzindo consideravelmente o fator sorte (DE LIRA, 2018).

Aoki (2017) menciona uma dificuldade muito grande na predição de resultados esportivos com precisão. Segundo ele, essa dificuldade existe porque nem sempre o time mais qualificado vence o duelo, uma vez que existe o fator sorte envolvido na equação, o que implica em uma limitação na acurácia dos modelos.

Pereira (2017) relata que, mesmo com as incertezas apresentadas anteriormente, os apostadores de longo prazo não contam com a sorte em suas análises, pois ela é responsável apenas pelas variações no curto prazo. O número de amostras é algo extremamente importante para avaliar se um apostador é ou não lucrativo, da mesma forma que em outros investimentos de renda variável, a consistência é a chave do sucesso.

2.3 Engenharia de Atributos

A engenharia de atributos é a etapa mais importante da preparação de dados para o aprendizado de máquina. Funções de transformação, como aritmética e agregações, são utilizadas para que novos atributos sejam criados. O produto final deste trabalho é a conversão de informações brutas em uma base tratada de acordo com o problema em questão. (Nargesia, 2017)

Segundo Khurana (2016), a engenharia de atributos é uma etapa predominante humana e demorada, porém é fundamental para o fluxo de trabalho. Sendo responsável por melhorar o desempenho da modelagem preditiva em um conjunto de dados, transformando as informações disponíveis em novos dados otimizados para o aprendizado.

2.4 Aprendizado de Máquina

O termo *Aprendizado de Máquina* surgiu em meados de 1959, criado por Arthur Samuel, um engenheiro do Instituto de Tecnologia de Massachussetts (MIT), definindo como uma área de estudo que atribui aos computadores a habilidade de aprender automaticamente com experiência. Existem técnicas que auxiliam neste processo sem terem sido programados com regras fixas que resultam em uma determina saída. Elas se baseiam em um conjunto de algoritmos capazes de identificar padrões e, em seguida, utilizar esses padrões encontrados para prever situações futuras. (Mitchell,1997)

O mundo vem se transformando constantemente, contando com uma série de dados, fotos, músicas, textos, vídeos, dentre outros meios que estão armazenados em diferentes dispositivos. Estamos entrando em uma era onde encontrar os dados não é mais um problema. O desafio atual é analisar esse conjunto gigante de dados extraindo informações úteis que auxiliem nos processo de decisão.

O processo de aprendizado pode ser conduzido por diferentes caminhos, entender o problema é um passo fundamental para escolher quais métodos devem ser aplicados. Segundo Langaroudi (2019), os métodos mais importantes de aprendizado de máquina são: Aprendizado Supervisionado (Classificação e Regressão), Aprendizado Semi-Supervisionado (Cadeia de Markov e Programação Dinâmica) e Aprendizado Não Supervisionado (Clusterização). Alguns exemplos e comparações serão apresentadas ao longo deste estudo.

2.4.1 Máquina de Vetores de Suporte

Máquina de vetores de suporte é um algoritmo criado com o intuito de identificar o melhor modo de separação entre classes. Vapnik (1995) descreve esse algoritmo como uma utilização de planos em conjunto com vetores aplicados na base de teste para separar as classes linearmente, criando hiperplanos que podem ser utilizados para classificação de novas entradas. Os hiperplanos são criados no ponto médio entre duas classes, dessa forma o risco de *overfitting* é minimizado e uma simetria entre os planos e os pontos de suporte é assegurada.

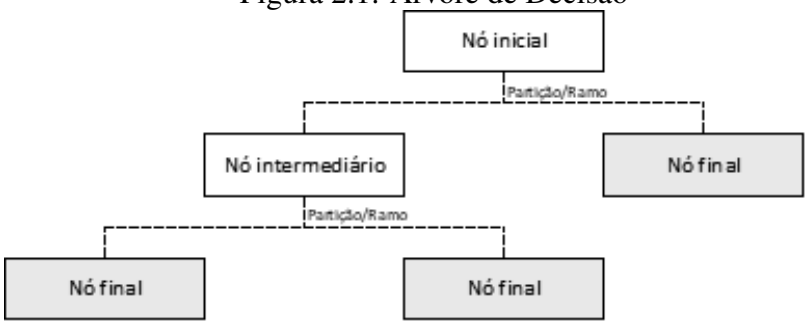
2.4.2 Árvore de Decisão

Casali (2021) descreve os algoritmos baseados em árvore de decisão como os mais populares e acessíveis, principalmente por se assemelharem a um fluxograma de tomada de decisão, permitindo uma compreensão facilitada frente a outros modelos de aprendizado de máquina.

A árvore de decisão é um algoritmo de aprendizado supervisionado que pode ser usado tanto para classificação quanto para regressão (Russell and Norvig, 2009). Isto é, ele é capaz de prever categorias e também valores numéricos. Neste modelo, o fluxo de uma decisão se inicia no nó-raiz e, logo abaixo, se divide em dois possíveis resultados, que se ramificam em outras possibilidades formando uma árvore completa. Os nós folhas são os últimos nós, eles que possuem a informação final, representando o resultado da decisão.

A entropia é um conceito muito importante na criação desse tipo de modelo, visto que ela representa um modo de quantificar as impurezas de cada ramo. Como o objetivo final é identificar um resultado de forma consistente, é necessário reduzir as impurezas o máximo possível, por isso a entropia é o parâmetro mais utilizado durante a fase de criação da árvore.

Figura 2.1: Árvore de Decisão



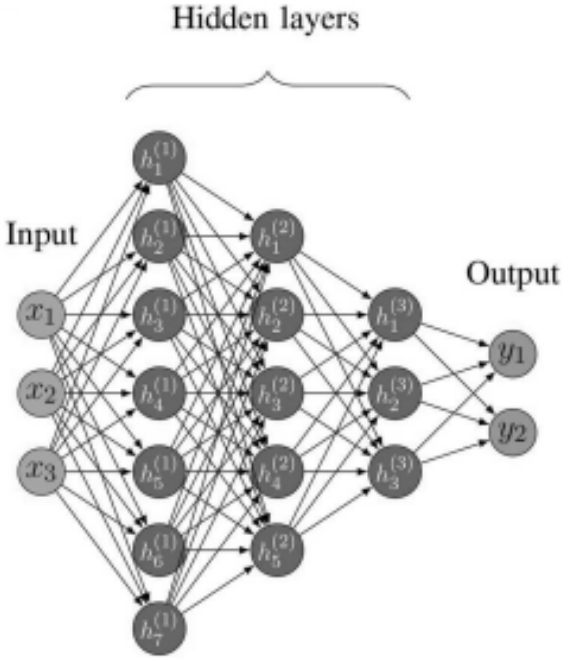
Fonte: (Taconeli, 2008)

2.4.3 Redes Neurais Artificiais

As *Redes Neurais Artificiais* foram desenvolvidas com a intenção de imitar o sistema de aprendizagem biológico, ou seja, modelos de como o cérebro humano aprende. Elas são compostas por uma série de neurônios interconectados em diferentes camadas, tornando-as capazes de resolver sistemas extremamente complexos (Goodfellow, 2016).

Cada neurônio possui uma conexão ponderada com os neurônios da camada anterior e posterior, fazendo com que a informação dos valores de entrada variem ao longo da rede até atingir seu valor final na camada de saída. Uma rede é composta por, no mínimo, uma camada de entrada e uma camada de saída, podendo conter camadas ocultas entre elas quando necessário. O treinamento da rede consiste em processar um conjunto de dados de treinamento, transmitindo as informações pela rede e refinando os pesos de cada ligação.

Figura 2.2: Esquema ilustrando o funcionamento de uma Rede Neural Artificial



Fonte: (Riguzz, Fabrizio (2020)

2.4.4 Aprendizado supervisionado *versus* aprendizado não supervisionado

O aprendizado supervisionado é aquele em que são fornecidos ao algoritmo da aprendizagem um conjunto de exemplos que são destinados ao treinamento, nos quais os rótulos de classe são conhecidos. Seu objetivo é a construção de um classificador que consiga definir os padrões da classe. O modelo supervisionado pode ser utilizado para identificar classes discretas, problemas de classificação, e para rótulos contínuos, onde o tipo do problema é identificado como uma regressão.

Conforme a própria palavra aduz, nesse método conta com um supervisor, em tese, a informação prévia acerca dos dados inerentes em que a máquina poderá avaliar sua eficácia e seus aprendizado, além disso, esses algoritmos são subdivididos entre clas-

sificadores e regressores. No caso da primeira é observado o conjunto dos dados como a atribuição de cada uma das classes ou categorias. Enquanto que a regressão, para cada observação ou valor de entrada, existe um valor numérico de saída, com o intuito da máquina de aprender a estimativa de determinado valor inerente a ideia de saída para cada conjunto de valores de entrada, como no caso dos resultados dos jogos de futebol (ALMEIDA; CARVALHO; MENINO, 2020).

O aprendizado não supervisionado é utilizado quando não existe um rótulo para os registros históricos, ou seja, não temos as saídas desejadas. Assim, o algoritmo busca encontrar padrões e correlações entre os registros, agrupando-os em clusters por similaridade.

Diante ao que vem sendo especificado sobre os tipos de aprendizado, este trabalho será baseado no aprendizado supervisionado.

2.4.4.1 Modelo

Para cada tipo de algoritmo existente existe uma forma distinta, na qual condiz com um conjunto de dados em treinamento que serão compreendidos e utilizados. Através do algoritmo, no qual tendem a atualizar o modelo de acordo com o comportamento dos ouros dois componentes, cabe mencionar a importância e a eficácia do modelo que é determinada de acordo com o componente de avaliação do algoritmo (ALMEIDA; CARVALHO; MENINO, 2020).

2.4.4.2 Avaliação

Nesse caso, cada algoritmo que possui uma função de avaliação diferenciada, variando o método com a estruturação dos dados e dos problemas eminentes, assim como a função que propaga a qualidade na aprendizagem do algoritmo, e a comparação com as saídas que são estimadas com as saídas reais no conjunto dos dados de treinamento, em ressalva que deverá ser amenizada ou maximizada, no que depende de cada algoritmo em questão. Dessa forma, o processo de minimização ou de maximização é compreendida como função da avaliação que é efetivado como componente de aprimoramento desse algoritmo.

2.4.4.3 Otimização

Consiste na forma pela qual a função decorre da avaliação que será aprimorada e varia de acordo com o algoritmo. Além disso, os diferentes métodos para o aprimoramento existem conforme suas variedades e dos problemas, assim como a eficácia do algoritmo da aprendizagem da máquina que depende da escolha como o bom algoritmo de otimização, atualizando os parâmetros do modelo.

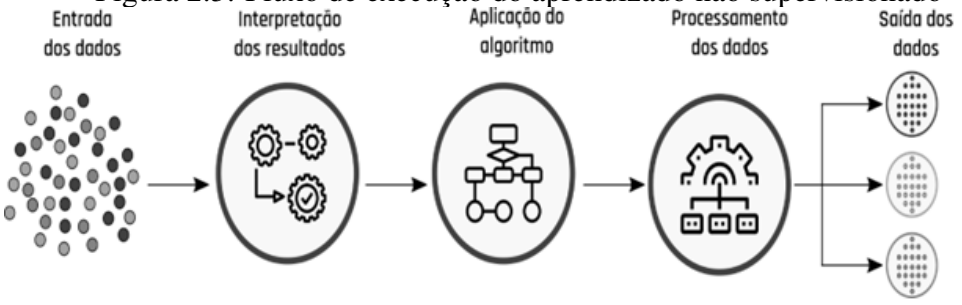
2.4.4.4 Principais Diferenças

Percebe-se que no aprendizado supervisionado ocorre um conjunto de exemplos de acordo com a programação de cada um que está intrinsecamente relacionados com sua classe, construindo uma série de classificações para contribuir com os novos exemplos, além das regressões que também já foram devidamente discorridas.

Diante ao que vem sendo discorrido, em relação ao não supervisionado, são fornecidos ao indutor um conjunto de dados para que possam ser analisados e agrupados, o que permite a formação dos agrupamentos, que são definidos por meio de uma análise de forma a identificar as contextualizações dos agrupamentos e como cada um deles se relacionada com o problema a ser analisado (ALMEIDA; CARVALHO; MENINO, 2020).

Após esse processo de definição dos agrupamentos, é necessário fazer uma análise para a definição do sentido de cada um deles no contexto da sua aplicação, pois devem fazer a inferência a partir dos pressupostos inerentes aos dados que não foram rotulados, classificados ou categorizados previamente.

Figura 2.3: Fluxo de execução do aprendizado não supervisionado



Fonte: Almeida, Carvalho, Menino (2020)

Conforme pode se perceber com o detalhamento da Figura 2.3 que as atividades do agrupamento e da amenização do dimensionamento que estão entre as principais atribuições executadas pelos algoritmos da aprendizagem não supervisionada, além de ser amplamente aplicada na identificação das irregularidades dos dados.

O aprendizado não supervisionado pode ser aplicado para a detecção das transações fraudulentas. Além disso, as anomalias podem ser bem complexas de serem identificadas, principalmente quando estão dentro de uma grande quantidade de informações, o que pode afetar significativamente o treinamento de um modelo.

2.4.5 Indicadores de desempenho

Os indicadores são utilizados como uma ferramenta para o acompanhamento dos objetivo em todos os segmentos. Isso torna possível manter os dados como protagonistas no processo da tomada de decisão.

Referindo-se ao aprendizado de máquina, existem diversos indicadores de desempenho que poderão ser utilizados com o intuito de estimar e comparar os resultados em diferentes classificações, além da comparação da saída predita com a esperada. Cada um desses indicadores é compreendido de formas distintas, sendo necessária uma avaliação em conjunto dos dados para que se tenha um entendimento completo do desempenho.

Os modelos são construídos com base em dados históricos, porém, para a avaliação do desempenho, é necessário a aplicação em um conjunto de dados que não são previamente conhecidos, sendo necessária a divisão do conjunto de dados em duas vertentes, que é o treinamento (dados históricos utilizados para a construção do modelo) e o teste (usado para a avaliação do desempenho real do modelo segundo a métrica avaliativa).

Os modelos desenvolvidos neste estudo serão avaliados com base nas métricas tradicionais referentes ao aprendizado de máquina, porém, o principal indicador de desempenho a ser considerado é retorno total que seria atingido utilizando o modelo como fonte para as apostas.

2.5 Apostas Esportivas e a necessidade de regulamentação no Brasil

Nesse ponto, é traçado que cabe não aos Estados em pontuar a regulamentação ou não dos jogos esportivos, mas a União, esse fato é fundamentado na Constituição Federal de 1988 e que não especifica questão de apostas esportivas, mas determina a quem compete essa legislação em termos de sistematização de consórcios e sorteios, mais precisamente no artigo 22.

Dessa forma, o Supremo Tribunal Federal consolidou termos atualizados sobre as

apostas, em seu artigo 33, da Constituição Federal Brasileira, impedindo que os Estados e Municípios legislem sobre o sistema de consórcios e dos sorteios. Acentuando sobre o entendimento de bingos e loterias, porém, se atendo aos métodos de custeio da seguridade social, e a normativa constitucional na premissa de conceder a possibilidade dos recursos provenientes de sorteio a serem realizados com benefícios a sociedade.

As pessoas ou empresas que realizam apostas, aceitando e intermediando os prognósticos de um evento futuro, que são chamadas como casas de aposta, são denominadas como *bookmakers*, recebendo um percentual de todas as apostas que são realizadas, como forma de remuneração, sendo o responsável pela apresentação de todas as informações eminentes as apostas no mercado, e no final do evento pagar a remuneração da aposta ao vencedor.

A qualificação quando se fala sobre jogo de azar não tem ligação com as apostas esportivas, pois conforme já mencionado, elas não possuem vínculo apenas de azar ou sorte, mas de um fator conhecido, a maestria de que não cabe exclusivamente a sorte, o que não condiz com a situação em eminência (SALVARO, 2019).

O que foi apontado no parágrafo anterior, condiz com a definição de um acordo entre as partes, tendo aquele que não acerta, ou perde, mas que depende exatamente do que se trata, pagando a outra parte, diante ao que foi acordado. Desta forma, as apostas esportivas não são condicionadas com os jogos de azar (tipificado na Lei de Contravenções Penais, no art. 50, § 3º).

É dentro dessa perspectiva que se abrem parâmetros sobre a legislação brasileira, o direito que veda as práticas de jogos de azar, além das Contravenções Penais que vedam por completo a exploração e o estabelecimento dos jogos de azar em locais públicos, mediante o pagamento ou não e nesta mesma norma, se atendo a alguns requisitos, o que não confere as apostas esportivas.

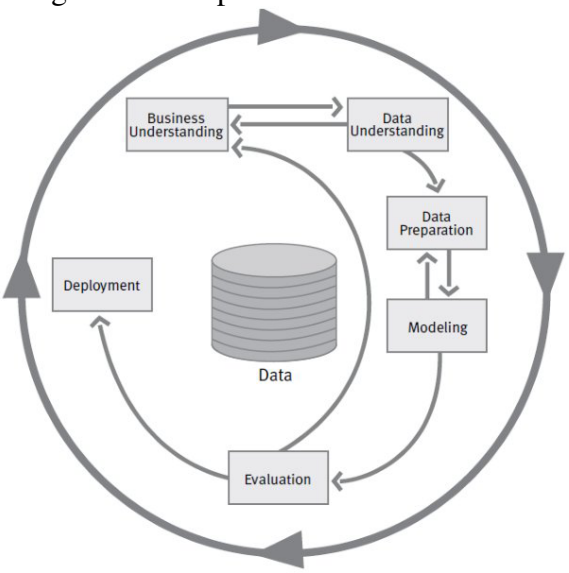
Dessa forma, a prática das apostas esportivas que foge dos requisitos, então, a punibilidade pelo fato de não poder ser alcançada com as sanções previstas, comprovando a análise no impositivo legal, compreendendo então, a possibilidade de elucidar dúvidas legais sobre esse assunto.

3 PROPOSTA E METODOLOGIA DO TRABALHO

Este trabalho consiste na criação completa de um modelo de precificação para apostas de futebol, incluindo a solução como um todo, desde a extração e manipulação dos dados até a criação do modelo de aprendizado de máquina. O desenvolvimento do trabalho teve como inspiração a metodologia *CRISP DM (Cross Industry Standard Process for Data Mining)*, metodologia esta que teve sua primeira versão apresentada no *Jounal of Data Warehousing* (2000). Ela auxilia no processo de entendimento dos dados, transformando-os em informações que suportam uma tomada de decisão.

A metodologia *CRISP DM* define o ciclo de vida do projeto, dividindo-o nas seguintes etapas: Entendimento do problema, Compreensão dos dados, Preparação dos dados, Modelagem e Avaliação.

Figura 3.1: Etapas da modelo CRISP DM



Fonte: Jounal of Data Warehousing (2000)

3.1 Base de Dados

A base de dados utilizada para análise contempla os últimos 5 anos das principais ligas mundiais, somando um total de mais de 100 mil jogos. Todos os dados das partidas foram extraídos da plataforma Sportmonks[1] utilizando as APIs de consulta disponíveis, sendo complementados com as cotações da Bet365 [2], umas das maiores casas de aposta

[1]https://www.sportmonks.com/
[2]https://www.bet365.com/

do mundo. A empresa sportmonks disponibiliza todas as estatísticas dos jogos encerrados, contudo, não existe nenhum dado referente a performance geral dos times. A criação de novas variáveis para mensuração de performance histórica é algo extremamente necessário para o sucesso do projeto. Para auxiliar nesse processo, uma base de dados SQL foi criada para armazenar todos esses dados, facilitando as manipulações e suportando os testes de hipóteses, afim de modelar novas variáveis customizadas para potencializar o poder de aprendizagem. As informações de uma das partidas estão listadas na tabela abaixo para exemplificar os dados.

Tabela 3.1: Estatísticas da Partida Argentina x Venezuela pela Copa do Mundo 2022

Estatística	*Argentina*	*Venezuela*
Gols	3	0
Chutes no Alvo	6	1
Chutes Fora do Alvo	10	3
Ataques Perigosos	57	20
Ataques	152	69
Posse de Bola	73	27

Fonte: Sportmonks

3.1.1 Viés Existente

Identificar a existência de um viés na base de dados é um passo fundamental para construção e, posteriormente, avaliação de um modelo de aprendizado de máquina. Um algoritmo que identifica a vitória do time da casa pode parecer excelente ao apresentar uma acurácia de 65%, contudo, não seria muito útil caso a distribuição inicial da base apresente 64% de vitórias do time da casa.

A análise inicial, baseada apenas na organização e visualização dos dados, apresentou um pequeno viés referente a vitória do time da casa, algo já esperado nesse esporte, visto que o time mandante conta com o apoio da torcida e tende a buscar o resultado.

O histograma da Figura 3.2 apresenta um percentual de vitórias quase 30% maior para o time da casa em relação ao time visitante, enquanto a Figura 3.3 mostra a diferença entre os placares, ou seja, placar do mandante - placar do visitante.

Figura 3.2: Distribuição dos resultados em razão do vencedor do confronto

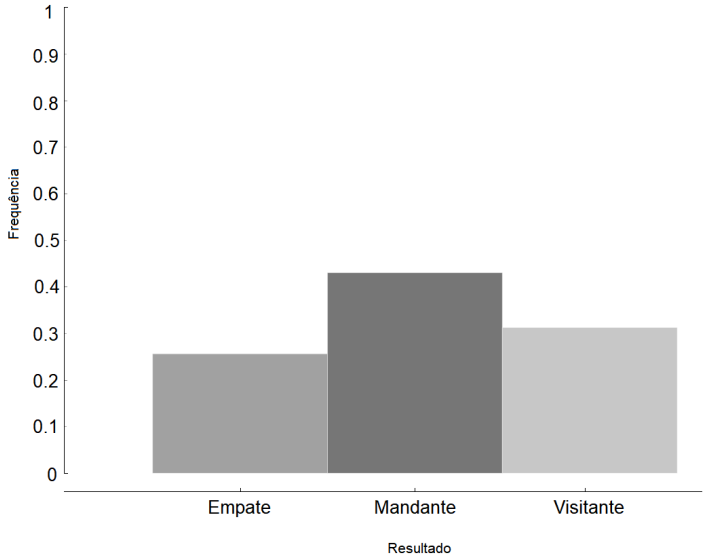
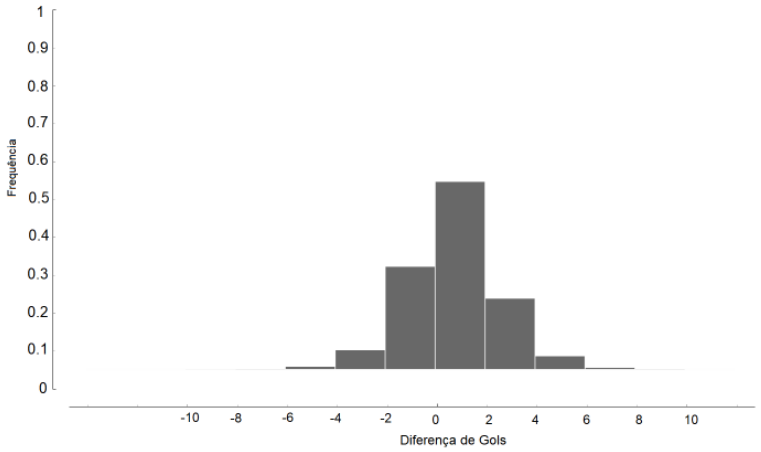


Figura 3.3: Diferença de Gols



3.2 Precificação dos resultados

A cotação de uma aposta determina a quantia de dinheiro paga por uma aposta unitária vencedora, consequentemente, quanto menor o risco, menor o retorno. As cotações estão diretamente relacionadas as probabilidades de ocorrência do resultado em questão, podendo ser derivada através da fórmula “probabilidade = 1/cotação´´. Lisi e Zanella (2017) apresentaram uma análise comparativa entre probabilidade e risco em seu artigo. Eles afirmam que quando dois times possuem chances iguais de vitória, a cotação justa seria de 2.0, retornando o dobro do valor investido em caso de vitória. No entanto, manter a precificação totalmente justa não é vantajoso para as casas de aposta. As cotações esperadas para o evento descrito anteriormente se aproximam de 1.85 no mercado atual, garantido uma boa margem de segurança para a casa e dificultando ainda mais a

lucratividade dos jogadores no longo prazo.

A grande popularidade deste mercado vem obrigando as empresas a atender cada vez mais ligas e modalidades, tornando esse um dos mais importantes diferenciais competitivos. Esse grande volume de jogos torna muito difícil o trabalho de precificação em tempo real de todas as partidas, fazendo com que boas oportunidades de investimento apareçam, mesmo que por curtos períodos de tempo. Alguns artigos focam exclusivamente nas variações existentes dentro de uma partida em relação aos acontecimentos do evento.

O trabalho de Ricado Gil (2006) avaliou a eficiência do mercado dentro da Copa do Mundo de 2002, uma das competições mais populares do mundo. Foi constatado que existe uma brusca variação após um gol ser marcado, porém existe uma tendência de correção dessa variação até 15 minutos depois do acontecimento. Uma segunda consideração importante foi apresentada neste mesmo trabalho e reavaliada por Robert Simmons (2008) que afirma que os valores oferecidos pela vitória de times populares tende a ser mais baixos do que o valor considerado justo. Isso acontece porque a base de fãs do time é muito grande e, consequentemente, existem muitas apostas a seu favor. Dito isso, podemos considerar que a estratégia de apostar em um time favorito antes do jogo começar, dificilmente será lucrativa no longo prazo.

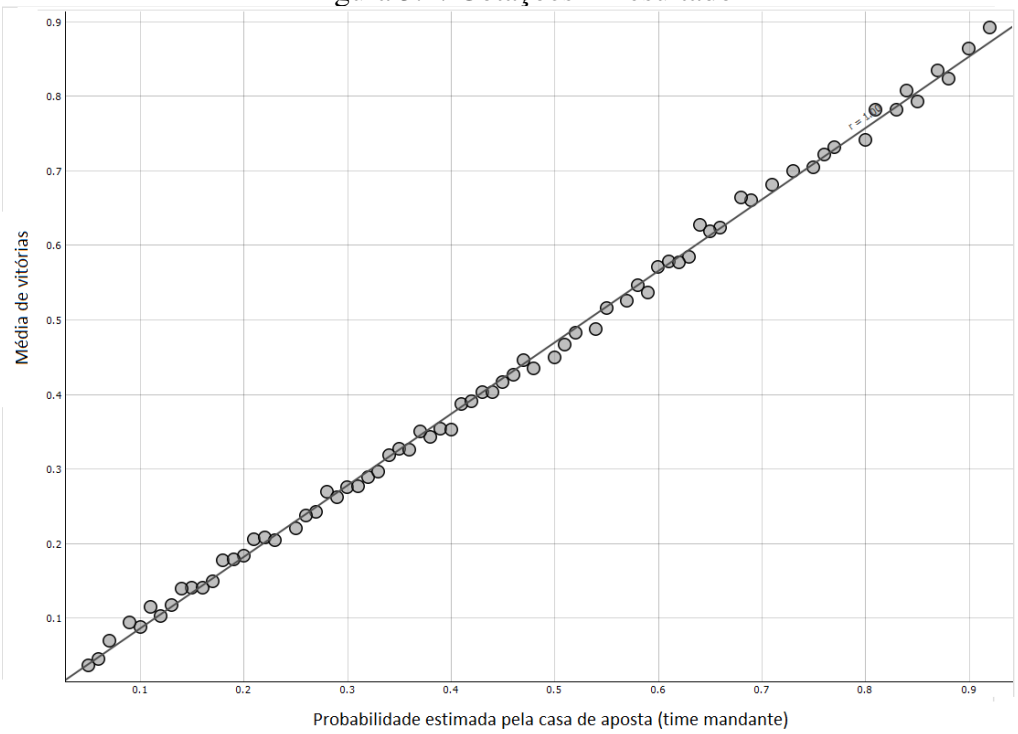
3.3 Eficiência das Casas de Aposta

Não é novidade para ninguém que todas as casas de aposta possuem vantagens frente a seus jogadores, isso ocorre por dois motivos básicos. O primeiro deles é a aplicação de uma margem de segurança em suas cotações, fazendo com que as probabilidades estejam sempre a seu favor. Mesmo com uma boa margem, é necessário que a avaliação das probabilidades seja feita ao menos de forma aproximada, e é aqui que vamos ao segundo ponto. Não é raro encontrarmos cotações completamente desajustadas na abertura dos mercados, porém as casas possuem uma informação adicional, elas conseguem mensurar exatamente quanto está entrado de dinheiro em cada um dos resultados e, com essa informação, reajustar suas próprias cotações mitigando o erro inicial.

Para garantir a correta avaliação do trabalho, precisamos garantir que a nossa base esteja aderente à situação real do mercado, onde temos probabilidades ofertadas muito próximas das probabilidades concretizadas. O estudo apresentado na Figura 3.4 valida essa afirmação utilizando as cotações de vitória do time da casa, porém também exibe alguns pequenos pontos de desalinhamento que podem ser explorados caso o motivo seja

identificado. Embora a precificação apresentada esteja próxima da perfeição, grande parte das observações não estão sobre a reta, indicando que existe uma margem de erro nas cotações oferecidas.

Figura 3.4: Cotações x Resultado



3.4 Mercados

Os mercados representam as diferentes formas de se apostar em uma partida, isso significa que existem inúmeras maneiras de se expor dentro de um mesmo evento, é possível apostar em quem será o vencedor do confronto e até mesmo no número de cartões amarelos da partida. Todos esses mercados possuem particularidades bem especificas que precisam ser levadas em conta durante a avaliação do risco e retorno ao efetuar uma aposta. Este trabalho tem como objetivo mensurar as probabilidades de cada equipe vencer a partida, dito isso, a avaliação final será feita com base nos mercados referentes ao resultado final da partida. Cada mercado é descrito nas subseções a seguir.

3.4.1 Resultado Final

O mercado de Resultado Final, também conhecido como mercado de probabilidades, é o mais conhecido no meio das apostas esportivas. Nele é possível realizar 3 tipos de apostas, vitória do mandante, empate ou vitória do time visitante. Este mercado apresenta grandes variações em caso de gols, permitindo que os apostadores encerrem suas posições assegurando seu lucro ou minimizando o prejuízo, mesmo antes do jogo ser encerrado.

3.4.2 Dupla Chance

O mercado dupla chance agrupa dois resultados em uma mesma aposta, a vitória de um dos times e o empate. Você vence se acertar o time vencedor ou se a partida terminar empatada, o que o torna um mercado extremamente conservador. Esse conservadorismo acaba tendo um preço, ele apresenta cotações muito mais baixas que o mercado de resultado final, exigindo uma taxa de acerto mais elevada para garantir lucro a longo prazo. Esse tipo de aposta costuma ser explorada contra o favorito, desse modo é possível encontrar cotações melhores.

3.4.3 Empate Anula Aposta

O mercado empate anula aposta é utilizado para eliminar o empate da equação, facilitando consideravelmente a análise da partida. O futebol é um esporte em que o empate muitas vezes assombra os times favoritos que, mesmo pressionando o jogo inteiro, não conseguem vencer a partida. Nesse cenário, anular sua aposta em caso de empate parece ser algo muito vantajoso para os apostadores, contudo, nenhuma casa visa facilitar a vida dos seus jogares. Ele precisa ser utilizado com extrema cautela pois também apresenta cotações mais baixas, tornando-se praticamente inviável em jogos onde existe um claro favorito.

3.4.4 Comparação de Mercados

Oferecer diversas opções de mercados se tornou um grande diferencial para as casas de apostas, atraindo diversos clientes casuais por proporcionar diversão através de

apostas inusitadas. Esse grande leque de opções acaba gerando algumas dúvidas no momento de realizar uma entrada, ser conservador pode te levar a uma variância menor e, ao mesmo tempo, reduzir drasticamente seus lucros. A tabela 3.2 apresenta as cotações oferecidas para um jogo da copa do mundo. Uma aposta de cem reais na vitória da Argentina retornaria um lucro de dezoito reais, enquanto uma aposta na vitória da Argentina e no empate retornaria apenas cinco reais. Essa diferença nas cotações precisa ser levada em consideração durante a analise, visto que resultados inesperados são comuns no futebol.

Tabela 3.2: Comparação de *Odds*

Mercado	*Argentina*	*Venezuela*
Resultado Final	1.18	12
Dupla Chance	1.05	4.5
Empate Anula Aposta	1.05	11

Fonte: Bet365

4 AVALIAÇÃO EXPERIMENTAL

O objetivo deste capítulo é apresentar os experimentos realizados e quais métricas foram utilizadas para comparar os seus resultados. Os modelos serão treinados, inicialmente, sem passar por um processo de otimização. Essa comparação tem como objetivo identificar o melhor modelo para este problema especifico, levando-o para uma segunda fase onde será otimizado e validado.

Alguns modelos de classificação trabalham apenas com resultados binários, por isso os testes serão baseados apenas na previsão de vitória do time mandante. Mesmo representando apenas um grupo dos possíveis resultados, uma boa performance nesse grupo é suficiente para validar o estudo, podendo inclusive ser replicado para os demais resultados futuramente.

4.1 Criação e Preparação das Variáveis

A criação de novas variáveis para mensurar a performance dos times é um passo fundamental para o sucesso deste estudo. Novas informações precisam ser calculadas para complementar nosso entendimento sobre as equipes que estão se enfrentando, assim será possível avaliar o duelo.

4.1.1 Variáveis disponíveis na Fonte

As fontes de consulta nos disponibilizam o histórico completo das principais ligas mundiais, incluindo as principais estatísticas de uma partida de futebol. Todas as estatísticas foram carregadas para uma base de dados destinada ao estudo, contemplando uma tupla por partida.

Figura 4.1: Base Inicial

	MatchId	HomeTeamScore	AwayTeamScore	HomeTeamShotsOnTarget	HomeTeamShotsOffTarget	HomeTeamDangerousAttacks	HomeTeamAttacks	AwayTeamShotsOnTarget	AwayTeamShotsOffTarget	AwayTeamDangerousAttacks	AwayTeamAttacks
1	12011751	1	2	5	11	41	94	4	4	32	82
2	17883125	2	0	7	8	67	109	5	5	46	110
3	12015790	0	2	2	14	57	103	2	6	26	75
4	18677018	0	0	1	4	67	115	3	7	80	119
5	18677049	1	2	7	10	75	131	6	9	70	136
6	18677077	2	1	9	9	98	143	4	4	59	130
7	18677094	0	2	4	5	72	123	3	4	73	96
8	18677118	0	0	4	7	74	144	2	3	31	119
9	18677008	1	1	2	2	55	99	4	12	65	90
10	18677022	2	0	3	9	80	146	0	19	110	152
11	18677038	1	2	1	14	88	137	2	12	99	136
12	18677055	0	1	2	5	55	103	8	15	73	124
13	18677072	3	2	5	6	64	141	10	10	98	131
14	18677087	3	0	3	7	21	133	0	7	17	137
15	18677098	1	1	4	1	82	91	2	5	70	72

4.1.2 Partidas Recentes

O trabalho de criação das novas variáveis teve inicio com a mensuração da performance recente de ambas as equipes. Essa avaliação tem o objetivo de avaliar a situação atual das equipes, levando em conta os últimos dez jogos de cada equipe em qualquer campeonato, sendo replicadas para ambas equipes e em dois grupos distintos, um considerando o mesmo mando de campo e outro geral. A Tabela 4.1 apresenta as medidas utilizadas como base para mensurar a performance histórica das equipes, enquanto as Tabela 4.2 e 4.3 representam as variáveis posteriormente criadas.

Tabela 4.1: Medidas Criadas

Variável	*Descrição*
ProGoals	Média de gols feitos
AgainsGoals	Média de gols sofridos
Win	Média de vitórias
Lose	Média de derrotas
ProOver05	Média de jogos com ao menos um gol feito
ProOver15	Média de jogos com mais de um gol feito
AgainstOver05	Média de jogos com ao menos um gol sofrido
AgainstOver15	Média de jogos com mais de um gol sofrido
WinByTwoOrMoreGoals	Média de vitórias com dois ou mais gols de diferença
WinByTwoOrMoreGoals	Média de derrotas com dois ou mais gols de diferença

Tabela 4.2: Performance nos últimos dez jogos

Medida Base	*Variável Criada*
ProGoals	Last10HomeTeamProGoals
AgainsGoals	Last10HomeTeamAgainstGoals
Win	Last10HomeTeamWin
Lose	Last10HomeTeamLose
ProOver05	Last10HomeTeamProOver05
ProOver15	Last10HomeTeamProOver15
AgainstOver05	Last10HomeTeamAgainstOver05
AgainstOver15	Last10HomeTeamAgainstOver15
WinByTwoOrMoreGoals	Last10HomeTeamWinByTwoOrMoreGoals
WinByTwoOrMoreGoals	Last10HomeTeamLoseByTwoOrMoreGoals
ProGoals	Last10AwayTeamProGoals
AgainsGoals	Last10AwayTeamAgainstGoals
Win	Last10AwayTeamWin
Lose	Last10AwayTeamLose
ProOver05	Last10AwayTeamProOver05
ProOver15	Last10AwayTeamProOver15
AgainstOver05	Last10AwayTeamAgainstOver05
AgainstOver15	Last10AwayTeamAgainstOver15
WinByTwoOrMoreGoals	Last10AwayTeamWinByTwoOrMoreGoals
WinByTwoOrMoreGoals	Last10AwayTeamLoseByTwoOrMoreGoals

Tabela 4.3: Performance nos últimos dez jogos com o mesmo mando de campo

Medida Base	*Variável Criada*
ProGoals	Last10HomeInHomeTeamProGoals
AgainsGoals	Last10HomeInHomeTeamAgainstGoals
Win	Last10HomeInHomeTeamWin
Lose	Last10HomeInHomeTeamLose
ProOver05	Last10HomeInHomeTeamProOver05
ProOver15	Last10HomeInHomeTeamProOver15
AgainstOver05	Last10HomeInHomeTeamAgainstOver05
AgainstOver15	Last10HomeInHomeTeamAgainstOver15
WinByTwoOrMoreGoals	Last10HomeInHomeTeamWinByTwoOrMoreGoals
WinByTwoOrMoreGoals	Last10HomeInHomeTeamLoseByTwoOrMoreGoals
ProGoals	Last10AwayInAwayTeamProGoals
AgainsGoals	Last10AwayInAwayTeamAgainstGoals
Win	Last10AwayInAwayTeamWin
Lose	Last10AwayInAwayTeamLose
ProOver05	Last10AwayInAwayTeamProOver05
ProOver15	Last10AwayInAwayTeamProOver15
AgainstOver05	Last10AwayInAwayTeamAgainstOver05
AgainstOver15	Last10AwayInAwayTeamAgainstOver15
WinByTwoOrMoreGoals	Last10AwayInAwayTeamWinByTwoOrMoreGoals
WinByTwoOrMoreGoals	Last10AwayInAwayTeamLoseByTwoOrMoreGoals

O futebol é um esporte onde a tradição da equipe conta muito dentro de campo, estando diretamente relacionada aos fatores subjetivos de pressão e experiencia. O termo popular utilizado para representar esse fator é o famoso “peso da camisa´´. Para mensurar essa característica, mesmo que de forma imprecisa, avaliamos a performance da equipe nas últimas 3 temporadas, replicando os mesmo critérios anteriormente utilizados.

Tabela 4.4: Performance nas últimas três temporadas

Medida Base	*Variável Criada*
ProGoals	Last3SeasonsHomeTeamProGoals
AgainsGoals	Last3SeasonsHomeTeamAgainstGoals
Win	Last3SeasonsHomeTeamWin
Lose	Last3SeasonsHomeTeamLose
ProOver05	Last3SeasonsHomeTeamProOver05
ProOver15	Last3SeasonsHomeTeamProOver15
AgainstOver05	Last3SeasonsHomeTeamAgainstOver05
AgainstOver15	Last3SeasonsHomeTeamAgainstOver15
WinByTwoOrMoreGoals	Last3SeasonsHomeTeamWinByTwoOrMoreGoals
WinByTwoOrMoreGoals	Last3SeasonsHomeTeamLoseByTwoOrMoreGoals
ProGoals	Last3SeasonsAwayTeamProGoals
AgainsGoals	Last3SeasonsAwayTeamAgainstGoals
Win	Last3SeasonsAwayTeamWin
Lose	Last3SeasonsAwayTeamLose
ProOver05	Last3SeasonsAwayTeamProOver05
ProOver15	Last3SeasonsAwayTeamProOver15
AgainstOver05	Last3SeasonsAwayTeamAgainstOver05
AgainstOver15	Last3SeasonsAwayTeamAgainstOver15
WinByTwoOrMoreGoals	Last3SeasonsAwayTeamWinByTwoOrMoreGoals
WinByTwoOrMoreGoals	Last3SeasonsAwayTeamLoseByTwoOrMoreGoals

4.1.3 Temporada atual

Além do desempenho recente dos times, também é importante avaliar sua performance geral dentro do campeonato em questão.

Tabela 4.5: Performance dentro do campeonato

Medida Base	*Variável Criada*
ProGoals	SeasonHomeTeamProGoals
AgainsGoals	SeasonHomeTeamAgainstGoals
Win	SeasonHomeTeamWin
Lose	SeasonHomeTeamLose
ProOver05	SeasonHomeTeamProOver05
ProOver15	SeasonHomeTeamProOver15
AgainstOver05	SeasonHomeTeamAgainstOver05
AgainstOver15	SeasonHomeTeamAgainstOver15
WinByTwoOrMoreGoals	SeasonHomeTeamWinByTwoOrMoreGoals
WinByTwoOrMoreGoals	SeasonHomeTeamLoseByTwoOrMoreGoals
ProGoals	SeasonAwayTeamProGoals
AgainsGoals	SeasonAwayTeamAgainstGoals
Win	SeasonAwayTeamWin
Lose	SeasonAwayTeamLose
ProOver05	SeasonAwayTeamProOver05
ProOver15	SeasonAwayTeamProOver15
AgainstOver05	SeasonAwayTeamAgainstOver05
AgainstOver15	SeasonAwayTeamAgainstOver15
WinByTwoOrMoreGoals	SeasonAwayTeamWinByTwoOrMoreGoals
WinByTwoOrMoreGoals	SeasonAwayTeamLoseByTwoOrMoreGoals

4.1.4 Viés da Liga

Existem campeonatos em que o mando de campo apresenta um impacto muito acima da média no resultado final, por isso foram criadas duas variáveis que auxiliam no aprendizado dessa tendência. A primeira delas representa a média geral de vitória dos mandantes nas últimas três temporadas do campeonato, enquanto a outra representa a média dos visitantes.

Tabela 4.6: Variáveis representativas do viés nas últimas três temporadas da competição

Variável
Last3SeasonsLeagueHomeWin
Last3SeasonsLeagueAwayWin

Tabela 4.7: Exemplo que clarifica as grandes variações

Country Name	*League Name*	*Home Win Percentage*
Italy	Coppa Italia	0.66
Italy	Primavera Cup	0.64
Vietnam	Vietnamese Cup	0.57
World	CONCACAF Champions League	0.57
Bolivia	Liga de Futbol Prof	0.56
Australia	Ffa Cup	0.32
Jordan	Shield Cup	0.30
Germany	DFB Pokal	0.30
Poland	Polish Cup	0.29
World	International Tournament (Cyprus) Women	0.27

4.2 Avaliação Preliminar dos Modelos

Essa seção tem como objetivo a identificação dos dois melhores modelos para este problema, ambos devem ser otimizados e testados de forma mais detalhada afim de validar este estudo. Para garantir a qualidade dos testes finais, dividimos a nossa base em duas partes, a primeira com 80% das amostras é utilizada para o treinamento dos modelos, enquanto o restante é nossa amostra para a validação final. A comparação de modelos fez uso apenas da base de treino, utilizando o processo de validação cruzada.

A avaliação dos modelos foi realizada com base nas métricas tradicionais do aprendizado de máquina.

1) *ROC* - Avalia a taxa de verdadeiro positivo e falso positivo.

2) *Acurácia* - Avaliação geral referente a quantos registros foram classificados corretamente, tanto positivos quanto negativos.

3) *Precisão* - Avalia a taxa de classificação correta dos registros positivos frente a todos registros previstos com saída positiva.

4) *Revocação* - Avalia a taxa de classificação correta dos registros positivos frente a todos registros originalmente positivos

5) *F1* - Representa a média harmônica entre precisão e revocação.

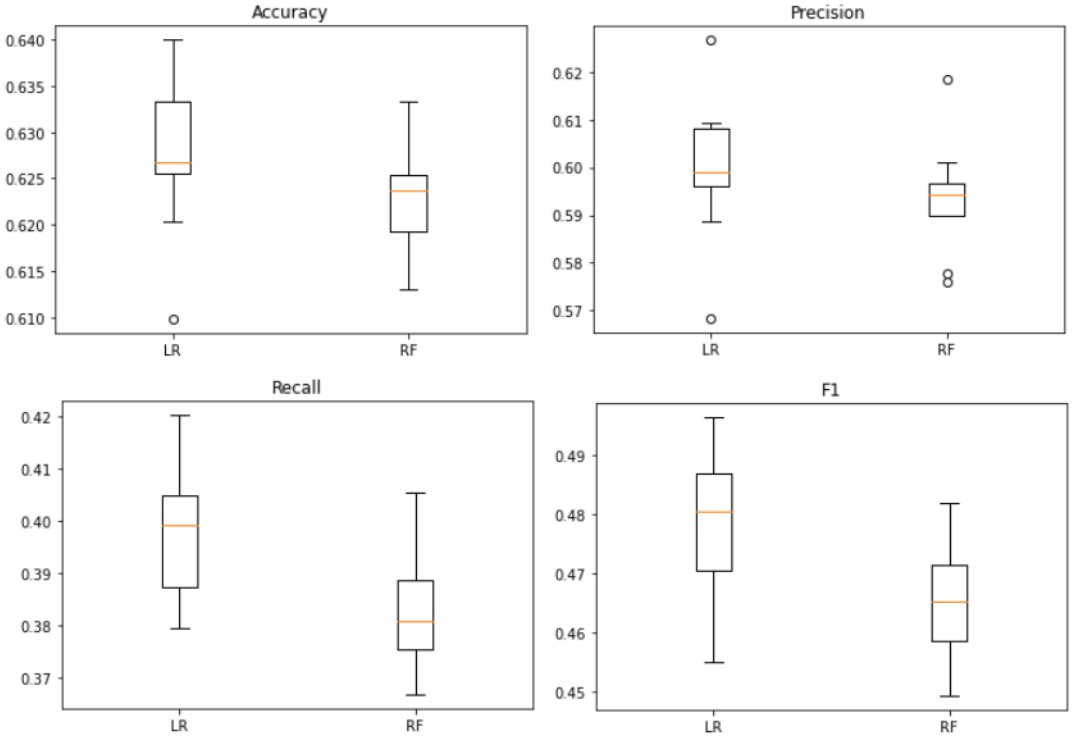
Tabela 4.8: Metricas de Validação

Modelo	*Accuracy*	*Precision*	*Recall*	*f1*	*ROC*
Logistic Regression	0.627	0.600	0.398	0.478	0.658
K Neighbors	0.554	0.477	0.410	0.441	0.551
Decision Tree	0.548	0.474	0.480	0.477	0.540
Gaussian NB	0.608	0.541	0.581	0.560	0.650
XGB	0.610	0.561	0.423	0.482	0.633
Random Forest	0.622	0.592	0.393	0.466	0.648
Neural Network	0.624	0.590	0.410	0.492	0.600

A Tabela 4.2 apresenta os resultados obtidos durante a primeira bateria de testes, ela evidencia pequenos desvios na performance dos modelos, porém, nenhum deles foi melhor em todos os critérios. As principais métricas utilizadas para descartar os que não serão utilizados foram ROC, acurácia e f1. O objetivo é trabalhar com um modelo poderoso em suas predições e, ao mesmo tempo, estável.

O primeiro modelo selecionado foi a Regressão Logística, um modelo estatístico relativamente simples, o segundo, por sua vez, foi o algoritmo de Floresta Aleatória, um modelo mais complexo que combina múltiplas árvores de decisão para chegar ao seu resultado final.

Figura 4.2: Validação cruzada - Média e Desvio



4.3 Hiperparâmetros

Os hiperparâmetros são os atributos que controlam o aprendizado do modelo, seus ajustes são essenciais para tornar o modelo preparado para enfrentar casos reais. Os parâmetros são os responsáveis por garantir que o modelo tenha um nível de generalização satisfatório, evitando *overfitting* e *underfitting*.

Este estudo empregou a técnica de Grid Search aliado ao K-fold, dois métodos utilizados para testar diferentes combinações e encontrar as melhores especificações.

O Grid Search é responsável por testar todas as combinações de parâmetros possíveis através de um conjunto previamente definido. Enquanto isso, o K-Fold é utilizado para dividir a amostra de dados e possibilitar diferentes testes, assegurando assim a confiabilidade do resultado final.

Figura 4.3: Conjunto de parâmetros utilizados RF

```
max_depth = [1, 10, 50, 100, 200, 300, 400]
min_samples_split = [1, 2, 5, 10, 15, 20, 30]
min_samples_leaf = [1, 2, 3, 4, 8, 12]
bootstrap = [True, False]
criterion=['gini', 'entropy']
random_grid = {'n_estimators': n_estimators,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf,
               'bootstrap': bootstrap,
               'criterion': criterion}
```

Figura 4.4: Conjunto de parâmetros utilizados LR

```
model = LogisticRegression()
solvers = ['newton-cg', 'lbfgs', 'liblinear']
c_values = [100, 10, 1.0, 0.1, 0.01]
grid = dict(solver=solvers,C=c_values)
cv = RepeatedStratifiedKFold(n_splits=5, n_repeats=3, random_state=1)
grid_search = GridSearchCV(estimator=model, param_grid=grid, n_jobs=-1, cv=cv, scoring='accuracy',error_score=0)
grid_result = grid_search.fit(X_train[features], y_train)
```

Figura 4.5: Hiperparâmetros selecionados RF

```
model = RandomForestClassifier(n_estimators= 300,
 min_samples_split= 10,
 min_samples_leaf= 3,
 max_depth= 12,
 criterion= 'entropy',
 bootstrap= True)
```

Figura 4.6: Hiperparâmetros selecionados LR

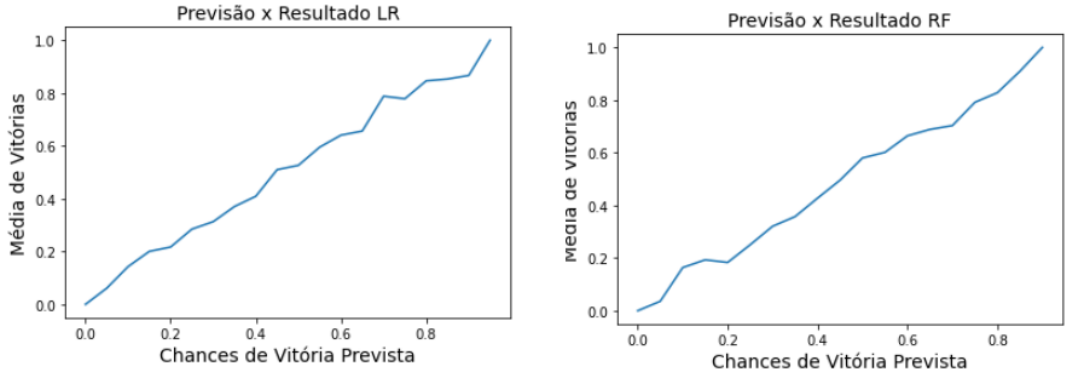
```
model = LogisticRegression(C= 100, solver= 'newton-cg')
```

4.4 Validação

Nesta etapa do estudo abordaremos os testes finais utilizando a nossa base de validação, que permaneceu intacta até o momento. O lucro a longo prazo está diretamente ligado à capacidade de precificação correta dos eventos, pois diferente do que muitas pessoas pensam, não basta apenas saber quem vai ganhar, é preciso avaliar as probabilidades de vitória e compará-las as cotações oferecidas. A figura 4.13 compara a probabilidade de vitória do mandante prevista com a taxa real de vitória, evidenciando a qualidade das previsões.

Conforme apresentado nas seções, podemos converter as cotações oferecidas em probabilidades, isso nos permite uma comparação direta com o resultado dos modelos.

Figura 4.7: Avaliação Geral dos Modelos



Assumindo a premissa de que o *output* do modelo é mais acurado que o da casa de aposta, teremos lucro a longo prazo ao realizar apostas em situações onde a probabilidade prevista for maior que a probabilidade ofertada. Se um determinado evento tem 55% de chances de ocorrer e a casa no oferece uma cotação de 2.0, significa que ela precificou considerando uma probabilidade de 50%, possibilitando um lucro aproximado de 10 reais para cada 100 apostas de 1 real.

Afim de mensurar esses desvios de precificação, uma nova métrica pode ser criada, sendo esta composta pela diferença entre a probabilidade apresentada pelo modelo destes estudo e a probabilidade utilizada pela casa de aposta. Pode-se dizer que essa nova variável representa o retorno esperado dessa aposta no longo prazo, podendo este ser positivo ou negativo.

Figura 4.8: Distribuição das previsões em razão do resultado esperado

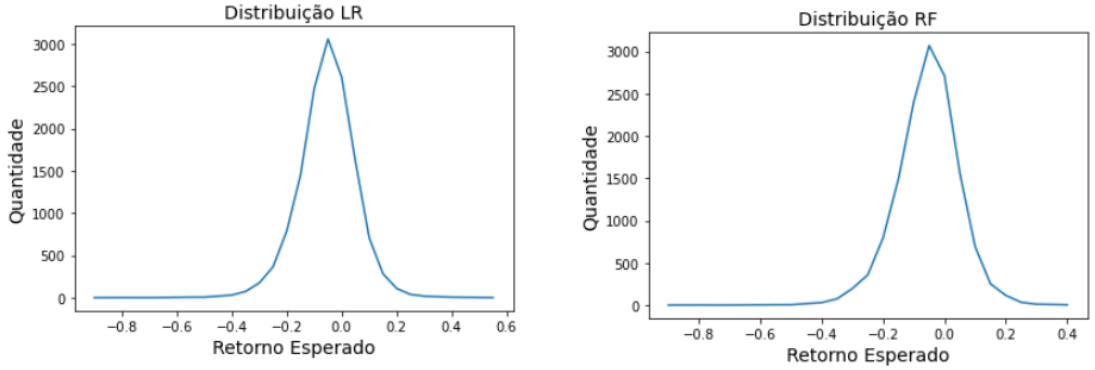
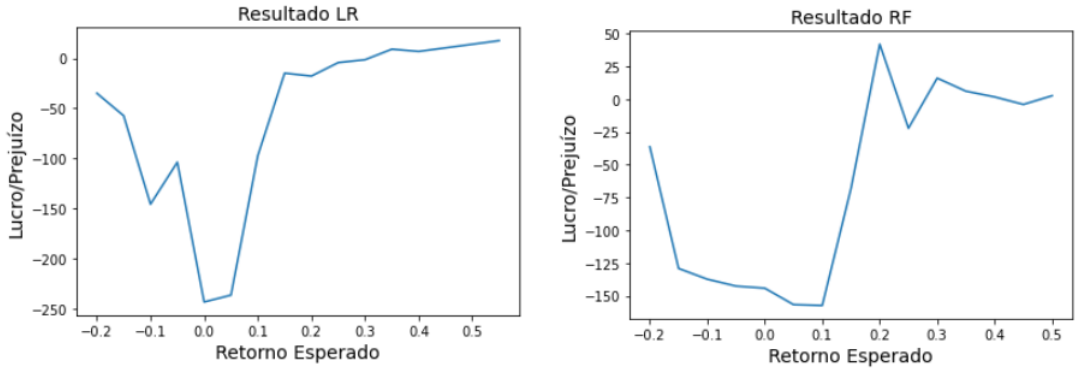
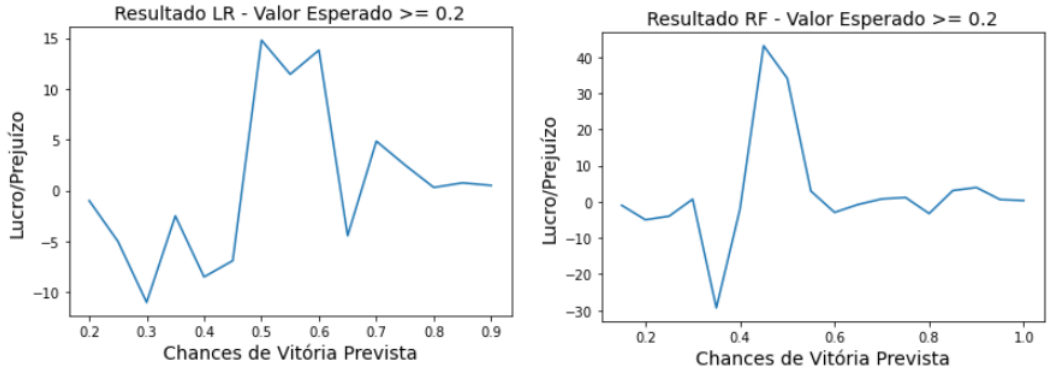


Figura 4.9: Lucro/Prejuízo em razão do resultado esperado



Um retorno positivo, mesmo que pequeno, já é esperado para valores maiores que zero, porém, só é possível observar um potencial lucro quando nossa variável de valor esperado passa a ter um valor mínimo de 0.2. Isso mostra o quanto as precificações ofertadas nas casas de apostas estão ajustadas a realidade.

Figura 4.10: Lucro/Prejuízo em razão da chance de vitória prevista



A figura 4.16 apresenta uma visão mais direcionada da análise, onde foram selecionadas apenas as partidas em que o valor esperado foi maior ou igual a 0.2. Este conjunto de jogos foi distribuído de acordo com a taxa de vitória prevista com o intuito de identificar possíveis oportunidades.

Evidencia-se que a precificação da casa de aposta, de maneira geral, é mais precisa que o modelo criado neste estudo, porém, ainda foi possível encontrar alguns desajustes nos jogos em que não temos um claro favorito, apresentando um lucro considerável na faixa de jogos entre 0.45 e 0.55.

Os jogos que apresentam um claro favorito geralmente não apresentam oportunidades de valor, isso se da pelo fato de que todos já imaginam quem será o vencedor do

confronto, fazendo com que as apostas sejam centralizadas neste time e, consequentemente, derrubando bastante as suas cotações.

5 CONCLUSÃO

Este estudo baseia-se na utilização de um modelo de aprendizado de máquina para identificar possíveis erros nas precificações oferecidas dentro das casas de apostas. Os testes foram feitos baseados em partidas de futebol, utilizando especificamente o mercado de vencedor do confronto.

As análises iniciais comprovaram uma alta eficiência das casas de apostas ao precificar suas cotações, tornando a identificação de oportunidades algo praticamente impossível. Essa alta taxa de acerto está diretamente relacionada ao modelo utilizado pelas empresas, onde os preços são ofertados, inicialmente, com base em estatísticas passadas, porém sofrem ajustes de acordo com as apostas de seus clientes afim de balancear o mercado.

Embora o estudo tenha apresentado alguns cenários onde existem cotações desajustadas, os modelos aqui apresentados ainda não são bons o suficiente para encontrar falhas no processo atual de precificação. Existem outras variáveis que podem ser incluídas na base para avaliar melhor a situação do time em um confronto especifico, como por exemplo a motivação do time, desfalques, dentre outros. Sabemos que esses fatores impactam diretamente a performance dentro de campo, porém são mais difíceis de se obter.

As apostas esportivas são relativamente novas no Brasil, isso faz co que ainda existam diversas oportunidades no mercado, contudo, encontra-las é uma tarefa extremamente complexa. Falando sobre o aprendizado de máquina, apenas variáveis simples sobre a performance histórica não são suficientes, é necessário ter mais informações sobre as equipes para prever com precisão o resultado dos confrontos.

Este trabalho teve uma base muito forte na etapa de engenharia de atributos, criando diversas novas variáveis para mensurar a qualidade dos times. Uma possível continuidade do trabalho seria buscar mais fontes informativas sobre os times, principalmente no que diz respeito aos acontecimentos próximos a data da partida.

REFERÊNCIAS

AQUINO, Samuel Rodrigues Maia.**Jogos de azar: uma análise de legalidade das apostas esportivas à luz do ordenamento jurídico brasileiro.** Universidade Federal de Campina Grande. Unidade Acadêmico de Direito. Ciências jurídicas e sociais. UFCG. Paraíba, 2022.

ALMEIDA, Adriano; CARVALHO, Felipe; MENINO, Felipe. **Introdução ao Machine Learning.** Revista Grupo DataAt. INPE - Instituto Nacional de Pesquisas Espaciais. Brasil, 2020.

AOKI, Raquel; ASSUNÇÃO, Renato. **Luck is Hard to Beat: The Difficulty of Sports Prediction** Department of Computer Science UFMG, 2017

BUNKER, R.; SUSNJAK, T.**The Application of Machine Learning Techniques for Predicting Match Results in Team Sport: A Review.** https://arxiv.org/abs/1912.11762, 2019.

CASTELLANI FILHO, L. **Educação Física no Brasil: a história que não se conta.** Campinas, 2019.

CHAGAS, Jonathan Machado. **A (im)possibilidade de regulamentação das apostas esportivas no ordenamento jurídico brasileiro.** 2016. 88 f. Monografia (Graduação) - Curso de Direito, Centro de Ciências Jurídicas, Universidade Federal de Santa Catarina, Florianópolis, 2016.

COSTA, Í. B. da. **Modelagem e Predição de Resultados de Futebol Antes e Durante as Partidas Usando Aprendizagem de Máquina** [Universidade Federal de Campina Grande], 2021.

DE LIRA, Pedro Enrick Moraes. **Os desafios para a regulamentação das apostas esportivas frente ao sistema jurídico brasileiro.** Universidade Federal de Campina Grande. Centro de Ciências Jurídicas e Sociais. Paraíba, 2018.

DIETTRICH, Débora. **A importância da regulamentação da aposta esportiva no Brasil.** 2022. Disponível em: < https://www.jb.com.br/jogos-online/2022/04/1037046-a-importancia-da-regulamentacao-da-aposta-esportiva-no-brasil.html> Acessado em: 02 de fev. 2023.

FLÁVIA, Ana.**Futebol: regras, fundamentos e história.** Cola da Web. 2022. Disponível em: <https://www.coladaweb.com/educacao-fisica/futebol> Acessado em: 29 de jan. 2023.

GAGLIANO, Pablo Stolze ; PAMPLONA , Rodolfo.**Manual de Direito Civil.** 2.

ed. São Paulo: Saraiva, v. único, 2018.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning** , 2015.

GIL, R.G.R.; Levitt, S.D.**Testing the efficiency of markets in the 2002 World Cup.** J. Predict. Mark, 2012.

GOULARD, Gabriel Ratto.**Redes neurais para análise de performance de futebol.** Relatório submetido a Universidade Federal de Santa Catarina. Florianópolis, 2019.

Grand View Reserch.**Sports Betting Market Size, Share and Trends Report,** 2020

KELNER, Gregório Ferrer.**SPORT BETTING: um mercado muito além da aposta.** Universidade Federal do Rio de Janeiro. UFRJ. Instituto de Economia. Rio de Janeiro, 2016.

LEAL, J. C.**Futebol: Arte e Ofício.** 2. ed. Rio de Janeiro: Sprint, 2017.

LIMA, João Henrique Martins.**Aplicação de machine learning para as apostas esportivas.** Universidade Federal de Pernambuco. Centro de Ciências sociais aplicadas. Recife, 2022.

LISI, F.; ZANELLA, G.**Tennis betting:can statistics beat bookmakers?** University of Padua, Department of Statistical Sciences, Via Battisti, 241 - 35121 Padua, Italy, 2017.

MATHESON, V.**An Overview of the Economics of Sports Gambling and an Introduction to the Symposium.** Eastern Econ J 47, 1–8 (2021).

MITCHELL, T. M.**Machine Learning** Mcgraw-hill, 1997

MOTTA, Bichara.**Considerações jurídicas sobre a regulamentação das apostas esportivas no Brasil.** Artigo Online. Artigos e Publicações Jurídicas. Considerações Jurídicas. Brasil, 2021.

OJHA, T. R.**Analysis of hey performance indicators in software development.** Thesis (Master's degree in Information Technology) - Faculty of Computing and Electrical Engineering, Tampere University of Technology, Finland 2014.

PEREIRA, Wagner.**Timemania: salvação para os clubes do futebol brasileiro?.** Monografia (graduação) – Universidade Federal de Santa Catarina, Centro de Ciências Jurídicas, Curso de Graduação em Direito, Florianópolis, 2017.

SALVADOR, Paulo Cesar do Nascimento.**Metodologia de ensino de futsal e futebol.** UNIASSELVI. 2010. p. Il. Biblioteca Dante Alighieri. 2016.

SALVARO, Richard de Freitas.**Perspectivas de tributação com a legalização das apostas esportivas no Brasil.** UNESC. Curso de Ciências Contábeis. Criciúma,

2019.

SANTIAGO, Octavio.**Prevendo resultado de partidas de futebol com Machine Learning.** Artigo Online. Medium. Brasil, 2021.

SCHNEIDER, Cristian Felipe.**Machine learning aplicado na previsão de resultados de partida de futebol.** Universidade Federal do Rio Grande do Sul. Escola de Engenharia. Porto Alegre, 2018.

SCHMIDT, H.**Uso de técnicas de aprendizado de máquina no auxílio em previsão de resultados de partidas de futebol.** Santa Cruz do Sul, Universidade de Santa Cruz do Sul, Curso de Ciência da Computação, Trabalho de Conclusão de Curso: [s.n.], 2017

SOARES, Igor de Camargo.**Regulação e tributação de apostas esportivas no Brasil: Lei 13.756 de 2018 e a compatibilidade com o ordenamento jurídico brasileiro.** UFPB. Centro de Ciências Jurídicas - CCJ. João Pessoa, 2019.

TUBINO, M.J.G.**Dicionário enciclopédico Tubino do esporte.** Rio de Janeiro: SENAC, 2016.

VAPNIK, V. **Support-Vector Networks** ATT Bell Labs., Holmdel, NJ 07733, USA, 1995.

KHURANA, U. **Cognito: Automated Feature Engineering for Supervised Learning** IEEE 16th International Conference on Data Mining Workshops (ICDMW), 2016.