UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS SOCIAIS APLICADAS
DEPARTAMENTO DE CIÊNCIAS CONTÁBEIS E ATUARIAIS
CURSO DE CIÊNCIAS ATUARIAIS

JOÃO HENRIQUE MARTINS LIMA

APLICAÇÃO DE MACHINE LEARNING PARA APOSTAS ESPORTIVAS: uso de Regressão Logística, SVM, Árvore de Decisão e Naive Bayes

Recife
2022

JOÃO HENRIQUE MARTINS LIMA

APLICAÇÃO DE MACHINE LEARNING PARA APOSTAS ESPORTIVAS: uso de Regressão Logística, SVM, Árvore de Decisão e Naive Bayes

Trabalho de Conclusão de Curso apresentado à Coordenação do Curso de Ciências Atuariais do Campus Recife da Universidade Federal de Pernambuco, na modalidade de monografia, como requisito parcial para obtenção do grau de bacharel em Ciências Atuariais.

Orientador (a): Prof. Renata Gomes Alcoforado, PhD

Recife

2022

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Lima, João Henrique Martins.
Aplicação de machine learning para apostas esportivas: uso de regressão logística, SVM, árvore de decisão e Naive Bayes / João Henrique Martins Lima. - Recife, 2022.
54 p. : il., tab.

Orientador(a): Dra. Renata Gomes Alcoforado
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Ciências Sociais Aplicadas, Ciências Atuariais, 2022.
Inclui referências, apêndices.

1. Apostas esportivas. 2. Futebol. 3. Machine learning. 4. Previsão. 5. Regressão logística. I. Alcoforado, Dra. Renata Gomes. (Orientação). II. Título.

310 CDD (22.ed.)

JOÃO HENRIQUE MARTINS LIMA

APLICAÇÃO DE MACHINE LEARNING PARA APOSTAS ESPORTIVAS: uso de Regressão Logística, SVM, Árvore de Decisão e Naive Bayes

Trabalho de Conclusão de Curso apresentado à Coordenação do Curso de Ciências Atuariais do Campus Recife da Universidade Federal de Pernambuco, na modalidade de monografia, como requisito parcial para obtenção do grau de bacharel em Ciências Atuariais.

Aprovada em: 20/10/2022

BANCA EXAMINADORA

Documento assinado digitalmente
gov.br
RENATA GOMES ALCOFORADO
Data: 21/10/2022 11:10:48-0300
Verifique em https://verificador.iti.br

Profa. Dra. Renata Gomes Alcoforado (Orientadora)
Universidade Federal de Pernambuco

Documento assinado digitalmente
gov.br
FILIPE COSTA DE SOUZA
Data: 22/10/2022 08:15:31-0300
Verifique em https://verificador.iti.br

Prof. Dr. Filipe Costa de Souza
Universidade Federal de Pernambuco

Documento assinado digitalmente
gov.br
WILTON BERNARDINO DA SILVA
Data: 25/10/2022 17:37:50-0300
Verifique em https://verificador.iti.br

Prof. Dr. Wilton Bernardino da Silva
Universidade Federal de Pernambuco

Dedico este trabalho primeiramente a Deus por ter me dado saúde e força nos momentos mais difíceis e aos meus pais por me ensinar e encorajar a ir atrás dos meus objetivos, independente dos obstáculos que por hora venham a surgir.

AGRADECIMENTOS

Agradeço a Deus por ter me dado força para superar as adversidades que surgiram nesta caminhada e conseguir ir em busca dos meus objetivos.

Aos meus pais, João Martins e Ivaneide Teixeira, por sempre me mostrarem o valor do esforço e comprometimento perante nossos objetivos e por todo cuidado, dedicação e incentivo para que eu não esmorecesse.

Aos meus amigos por todo o encorajamento durante esse processo, principalmente ao Lauro Henrique pela companhia e troca de aprendizado nas longas horas de estudo na biblioteca, além de seus conselhos valiosíssimos que me fizeram perseverar apesar dos momentos de dúvidas. À Patricia Teixeira por acreditar verdadeiramente em mim quando as vezes nem eu mesmo acreditava, por sempre enxergar potencial e torcer por mim. E ao Marcos Aurélio por sempre me incentivar a aprender, a descontrair nas horas difíceis e me motivar a batalhar pelos meus objetivos.

À coordenação do curso e aos professores por serem solícitos no atendimento na resolução de problemáticas que surgiram ao longo do curso e por compartilharem tanto conhecimento nesses anos, e em especial a minha orientadora Renata Alcoforado por me ajudar de forma tão generosa e me incentivar tão pacientemente a persistir com este trabalho, apesar das minhas dificuldades.

Por fim, todos recebem ajuda na vida, diretamente ou indiretamente, e com isso agradeço a todos que desde os mais simples gestos me ajudaram a chegar até este momento.

“Não há nada nobre em ser superior ao seu semelhante. A verdadeira nobreza é ser superior ao seu antigo eu”. (HEMINGWAY).

RESUMO

As práticas esportivas surgiram de atividades visando lazer e entretenimento, com o passar dos anos surgiu também um interesse econômico, e uma das ramificações é a aposta esportiva, cujo primeiro relato remonta aos Jogos Olímpicos da Antiguidade na época da Grécia antiga. Com o desenvolvimento da tecnologia e o advento da internet, as informações, os resultados, e o interesse pelo esporte e pelas apostas aumentaram, sendo possível fazer diferentes tipos de apostas em variados esportes. O futebol, por exemplo, influencia a vida de 3,5 bilhões de pessoas no mundo todo e é preferência nas apostas dos brasileiros. No Brasil entre 2018 e 2022 mesmo com a pandemia da Covid-19 o mercado local de apostas saiu de 2 bilhões de reais em movimentação para 7 bilhões, enquanto o mercado global foi avaliado no ano de 2020 em 59,6 bilhões de dólares, podendo chegar em 127,3 bilhões em 2027. O presente trabalho pretende gerar previsões dos resultados de jogos de futebol a partir de quatro métodos de *machine learning* e cinco ligas nacionais de futebol, com o objetivo de identificar qual a melhor combinação de método e liga que gera maior nível de assertividade nas previsões, além de verificar se com uma série de dados mais longa é possível obter uma melhoria na assertividade das previsões. Para tanto, desenvolvemos códigos, na linguagem de programação *python*, abordando quatro técnicas de *machine learning*: regressão logística, SVM, árvore de decisão e *Naive Bayes*. Utilizamos dados de cinco ligas nacionais de futebol: Brasil, Inglaterra, Itália, Espanha e França, com uma divisão dos dados para as 5 e as 10 últimas temporadas completas para efeito de comparação de resultados. Como resultado, foi possível observar que ao utilizar dados de 5 temporadas a melhor combinação foi aplicar regressão logística na *Premier League* e para 10 temporadas aplicar regressão logística ou SVM na *Serie A*, já que ambos tiveram mesmo desempenho. Quanto à comparação entre dados de 5 *versus* 10 temporadas houve ganhos em 3 das 5 ligas em relação às assertividades das previsões.

Palavras-chave: apostas esportivas; futebol; *machine learning*; previsão; regressão logística.

ABSTRACT

Sports practices emerged from activities aimed at leisure and entertainment, over the years an economic interest has also emerged, and one of the ramifications is sports betting, whose first report dates back to the Olympic Games of Antiquity at the time of ancient Greece. With the development of technology and the advent of the internet, information, results, and interest in sports and betting have increased, making it possible to place different types of bets on various sports. Football, for example, influences the lives of 3.5 billion people around the world and is preferred in betting by Brazilians. In Brazil, between 2018 and 2022, even with the Covid-19 pandemic, the local betting market went from 2 billion reais in movement to 7 billion, while the global market was valued in 2020 at 59.6 billion dollars, being able to reach 127.3 billion in 2027. The present work intends to generate predictions of the results of soccer games from four machine learning methods and five national soccer leagues, in order to identify the best combination of method and league that generates a higher level of assertiveness in forecasts, in addition to verifying whether with a longer data series it is possible to obtain an improvement in the assertiveness of forecasts. For that, we developed codes, in the python programming language, approaching four machine learning techniques: logistic regression, SVM, decision tree and Naive Bayes. We used data from five national football leagues: Brazil, England, Italy, Spain and France, with a breakdown of the data for the last 5 and 10 complete seasons for the purpose of comparing results. As a result, it was possible to observe that when using data from 5 seasons the best combination was to apply logistic regression in the Premier League and for 10 seasons to apply logistic regression or SVM in Serie A, since both had the same performance. As for the comparison between data from 5 versus 10 seasons, there were gains in 3 of the 5 leagues in relation to the assertiveness of the forecasts.

Keywords: sports betting; soccer; machine learning; prediction; logistic regression.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Segundo Lira (2018), o interesse por atividades de lazer e entretenimento é algo constante e que existe desde a origem da humanidade. Por tal motivo surgiram as práticas esportivas, que foram propagadas com o passar das gerações como atividades de entretenimento e também econômicas.

Devido a globalização, a propagação do esporte pelo mundo e sua democratização, além do interesse enquanto atividade econômica, o mercado de apostas vem se expandindo ao longo dos anos (LIRA, 2018). Segundo Soares (2019), com o advento da internet e a sua expansão, as casas de apostas puderam realizar apostas online, facilitando o acesso para os usuários em todo o mundo.

E isso também ocorre devido ao alto nível de investimento realizado pela indústria de apostas esportivas, pois ao verificar o seu potencial de crescimento, investe em publicidade para captar novos clientes, realizando contratos com celebridades, marcas esportivas, atletas, ligas e clubes a nível internacional, movimentando bilhões de dólares por ano, contando com a participação do Brasil como um dos maiores consumidores (SOARES, 2019).

Barbosa & Filho (2020) destacam que no Brasil é comum fazer apostas esportivas online, e embora a gama de esportes ofertados seja diversificada, onde é possível apostar em futebol, basquete, vôlei, e corrida de cavalos, há uma preferência pelo futebol. Segundo Costa (2021), há um interesse mundial pelo futebol, visto que esse esporte influencia diariamente aproximadamente a vida de 3,5 bilhões de pessoas.

De acordo com Globo (2021), os números das apostas esportivas no Brasil são animadores, e com uma regulamentação das apostas no país há uma possiblidade do país se destacar como um dos principais mercados devido a sua grande população e a ligação cultural do povo brasileiro com esportes. No país, o mercado de apostas esportivas movimentou 7 bilhões de reais mesmo durante a pandemia de Covid-19, paralisando boa parte dos jogos. Entre os anos de 2018 e 2022 houve um crescimento na movimentação do setor no Brasil saindo dos 2 bilhões de reais para os 7 bilhões de reais. Ainda segundo Globo (2021), o mercado global de apostas esportivas cresce 11,5% ao ano e em 2020 foi avaliado em 59,6 bilhões de dólares, além da possibilidade de atingir 127,3 bilhões no ano de 2027.

O presente trabalho pretende gerar previsões dos resultados de jogos de futebol a partir de quatro métodos de *machine learning* e cinco ligas nacionais de futebol, com o objetivo de identificar qual a melhor combinação de método e liga que gera maior nível de assertividade nas previsões, além de verificar se com uma série de dados mais longa é possível obter uma melhoria na assertividade das previsões. Pois de acordo com Bertozzo (2019), para o aprendizado de máquina aprender e extrair bons resultados é necessário ter grandes volumes de dados.

Com o crescente acesso às apostas esportivas, principalmente por conta da internet e o alto nível de investimento para atrair novos apostadores, o presente trabalho reforça sua importância ao apresentar a utilização de programação e estatística como ferramentas de auxílio na tomada de decisão em apostas esportivas.

Para atingir os objetivos propostos foi utilizado no estudo quatro técnicas de *machine learning*, dados de cinco ligas nacionais de futebol de acordo com o ranqueamento realizado em 2021 pela IFFHS (Federação Internacional de História e Estatísticas do Futebol), sendo estas a liga do Brasil, da Inglaterra, da Itália, da Espanha e da França. Dados esses que foram divididos em duas partes, a primeira com as últimas cinco temporadas completas, e a segunda com as últimas dez temporadas completas.

A partir desses dados foi desenvolvido um conjunto de códigos na linguagem de programação *python*, em que uma parte dos códigos é utilizada para organizar as bases de dados por liga e corte temporal, e a outra parte dos códigos para fazer o treinamento das técnicas de *machine learning*, e consecutivamente gerar as previsões dos jogos de futebol.

O trabalho está estruturado da seguinte forma: o capítulo 2 contém o referencial teórico; a descrição da metodologia é feita no capítulo 3, no capítulo 4 são apresentados e discutidos os resultados da pesquisa; por fim, no capítulo 5, encontra-se as considerações finais do estudo.

2 REFERENCIAL TEÓRICO

2.1 HISTÓRIA DO FUTEBOL

Segundo Gerhardt (1979), a forma mais antiga de jogar uma bola com os pés, para a qual se tem evidências científicas remonta ao século II e III a.C. na China. Trata-se de um exercício de educação física, o *Tsu'Chu*, presente em um manual militar da época da Dinastia Han. A prática consistia em chutar uma bola de couro cheia de penas e cabelos através de uma abertura, medindo apenas 30 - 40 cm de largura, em uma pequena rede fixada em longas canas de bambu.

Apesar de existirem práticas antigas que se assemelham ao futebol, o esporte regulamentado e que conhecemos hoje, é chamado de futebol moderno. De acordo com Murray & Murray (1998), foi apenas em 1863, na cidade de Londres na Grã Bretanha, que o jogo de futebol foi regulamentado, a partir da criação da *Football Association*.

Ao longo dos anos esse esporte se popularizou e conforme os dados mais recentes, em uma pesquisa realizada pela FIFA (Federação Internacional do Futebol) com ajuda de suas federações-membro, em 2006 havia 265 milhões de jogadores de futebol (masculinos e femininos) ao redor do mundo, e se considerar os árbitros e oficiais a quantia chegava a 270 milhões de pessoas envolvidas ativamente com o futebol (FIFA, 2007).

As ligas como conhecemos hoje em dia levaram bastante tempo para serem criadas, mas a primeira liga de futebol surgiu na Inglaterra em 8 de setembro de 1888 (MURRAY & MURRAY, 1998). Com o passar dos anos, o futebol foi se propagando ao redor do mundo e assim novos campeonatos surgiram em outros países, possibilitando uma estruturação melhor do esporte e seu desenvolvimento até chegar nos moldes atuais. A seguir será detalhada a história das 5 principais ligas nacionais de futebol em 2021.

2.1.1 Brasil: Série A

Devido à grande dimensão territorial do Brasil houve muitas dificuldades para se implementar uma competição a nível federal (SARMENTO, 2006). Apesar das dificuldades, houve 3 torneios que posteriormente foram reconhecidos como

campeonatos brasileiros de futebol pela Confederação Brasileira de Futebol (CBF), a Taça Brasil, o Torneio Roberto Gomes Pedrosa e Campeonato Nacional de Clubes. A Taça Brasil é considerada como a primeira edição, em 1959 (BETING, 2016).

Até o momento, a competição nacional brasileira já passou por diferentes torneios e formatos, contou com o sistema mata-mata, em que existem partidas eliminatórias (*playoffs*), um sistema misto, em que existe jogos classificatórios e depois os *playoffs* e atualmente o sistema de pontos corridos, em que 20 equipes se enfrentam e a mais bem classificada leva o título de Campeão Brasileiro da Série A (SANTOS, 2019).

Tipicamente a temporada começa entre o final de abril e começo de maio até dezembro, totalizando 380 jogos por temporada, 38 rodadas com 10 confrontos a cada rodada. Disputado de forma contínua, em turno e returno, 19 jogos de ida e 19 jogos de volta (CBF, 2017). A liga brasileira é reconhecida pelo nível parelho dos times, dificultando o palpite de quem será o campeão da vez. Diferente das ligas europeias, não há 2 clubes com grande diferença para os demais.

Com o passar dos anos e o alto investimento europeu, muitos talentos rumaram aos clubes estrangeiros, entretanto a liga brasileira conta com grandes jogares do futebol mundial.

2.1.2 Inglaterra: *Premier League*

A *Premier League* foi criada em 1992 após os clubes romperem com a *Football League*, visando aumentar as receitas provenientes dos direitos televisivos (ANDRADE, 2021b). De acordo com Fabbri (2021), originalmente eram 22 clubes que compunham a liga. A partir da temporada 1995/1996 passou a ser com 20 clubes.

O formato da liga é de pontos corridos, em um total de 38 rodadas com 10 confrontos em cada rodada, totalizando 380 jogos por temporada, seguindo o calendário europeu, com duração de 10 meses, geralmente com as primeiras partidas em agosto e as últimas em maio do ano seguinte.

O nível de disputa no campeonato inglês é altíssimo, contando inclusive com o *Big Six*, termo que representa os seis grandes times ingleses, que são Arsenal, Chelsea, Liverpool, Manchester City, Manchester United e Tottenham (COLLINSON, 2019). Todos os times de muita tradição e a maioria com uma galeria recheada de

títulos importantes. Dos títulos de maior expressão do *Big Six* estão 11 *UEFA Champions League* e 3 Copas do mundo de clubes da FIFA.

Segundo Alberti & Tassi (2022), o campeonato inglês segue hegemônico e mais atraente que as demais ligas europeias refletindo assim os altos investimentos colocam a liga no topo de gastos nas janelas de transferências de verão desde a temporada 2012/2013. Por se tratar uma liga em que há muito investimento e altas cifras em patrocínios e direitos televisivos, muitos jogadores se interessam em jogar na *Premier League* e não só jogar em grandes clubes, mas também jogar contra os grandes.

2.1.3 Itália: *Serie A*

A mais alta divisão do campeonato italiano de futebol se iniciou da forma que existe atualmente, como uma liga nacional da Itália em 1929 (ANDRADE, 2021a). Durante esses longos anos de existência houve uma variação na quantidade de times participantes, desde 16 times até 21.

Segundo Estadão (2022a), a competição é realizada por pontos corridos e conta atualmente com 20 times, que se enfrentam em turno e returno. Dos times com mais destaque no país estão Milan, Juventus e Internazionale, e todas conquistaram a *UEFA Champions League,* sendo o Milan o segundo maior vencedor do torneio, atrás apenas do Real Madrid, time espanhol. Os 3 grandes clubes italianos somam 12 títulos da *UEFA Champions League* e 2 Copas do mundo de clubes da FIFA.

De acordo com Leal (2014), a competição italiana era o lugar para se estar entre as décadas de 1980 e 1990, pois qualquer clube, mesmo os pequenos, tinham condições de contratar grandes jogadores das principais seleções do mundo, fazendo com que entrassem nas competições europeias já como candidatos a título.

2.1.4 Espanha: *La Liga*

Entre o final de 1928 e o início de 1929 era fundada a mais alta divisão do futebol espanhol, conhecida como *La Liga*. Segundo Müller (2020), a primeira temporada da liga teve a participação de apenas 10 times, Arenas, Athletic Bilbao, Atlético de Madrid, Barcelona, Espanyol, Europa, Racing de Santander, Real Madrid, Real Sociedad e Real Union, onde a disputa se deu através de 2 turnos e com o

sistema de pontos corridos. Trata-se de uma liga muito respeitada mundialmente devido aos seus maiores clubes e o respectivo sucesso na maior competição de clubes da europa, a *UEFA Champions League* (Liga dos Campeões da Europa).

A *La Liga* funciona atualmente com o sistema de pontos corridos, com um total de 20 times e 38 confrontos durante a temporada, disputando entre si no sistema de turno e returno (ESTADÃO, 2022b). Nas ligas europeias a duração é geralmente de agosto (final do verão do hemisfério norte) até maio do ano seguinte (final da primavera do hemisfério norte).

Dois times se destacam na liga, e no cenário europeu de futebol, os protagonistas do clássico conhecido como *El Clásico*, sendo estes o Barcelona e o Real Madrid, também os 2 maiores campeões do torneio. O Barcelona possui entre os principais títulos, 26 títulos da *La Liga*, 5 *UEFA Champions League* e 3 Copas do mundo de clubes da FIFA contra 35 títulos da *La Liga*, 14 *UEFA Champions League* e 4 Copas do mundo de clubes da FIFA do Real Madrid.

2.1.5 França: *Ligue 1*

A liga francesa da primeira divisão foi organizada profissionalmente pela primeira vez na temporada 1932/1933, e desde então mudou várias vezes o número de participantes, variando entre 18 e 20, sendo o último o mais viável e que está em rigor atualmente. Cada time do campeonato faz 38 jogos, sendo 19 jogos de ida e 19 jogos de volta, fazendo referência ao mando de campo, ou seja, 19 jogos como mandante e 19 como visitante (ESTADÃO, 2022c). Entretanto de acordo com TNT Sports (2021), a *Ligue 1* comunicou uma alteração futura na quantidades de times da liga, passando de 20 para 18 times a partir da temporada 2023/2024.

Das ligas europeias vistas anteriormente a *Ligue 1* é uma liga que vem ganhando prestígio apenas nos últimos anos devido ao alto investimento principalmente pelo time Paris Saint-Germain, que protagonizou até então a transferência mais cara da história, ao trazer Neymar Jr. do Barcelona para a equipe de Paris por nada menos que 222 milhões de euros em 2017. Em seguida contratou grandes nomes como Kylian Mbappé, Gianluigi Donnarumma, Keylor Navas, Sergio Ramos, variando desde jovens jogadores com um futuro muito promissor até nomes de peso no mundo do futebol, como um dos maiores jogadores da história do futebol,

Lionel Messi. Com contratações desse nível o interesse pelo clube subiu bastante e a *Ligue 1* não ficou de fora.

Com o aumento da visibilidade novos investidores têm interesse em participar e o exemplo disso é que de acordo com MKTEsportivo (2022), a CVC Capital Partners está negociando para assumir 13% da liga por um montante de € 1.5 bilhão por meio da nova empresa comercial que negociará a venda dos direitos de transmissão da *Ligue 1,* e os clubes da liga aprovam tal acordo.

2.2 APOSTAS ESPORTIVAS

Aposta esportiva é quando uma pessoa aposta seu dinheiro em um evento esportivo, independente de qual seja o esporte. Se o cenário apostado for se tornar realidade o apostador aufere lucro, caso contrário perde o dinheiro apostado (REJANE, 2021).

A história das apostas esportivas é paralela a história do esporte e as primeiras apostas são da época da Grécia antiga, nos Jogos Olímpicos da Antiguidade. Onde os espectadores apostavam enquanto acompanhavam as competições, tais como salto à distância, boxe e lançamento de disco (CHAGAS, 2016).

De acordo com Bayer (2014), as práticas desportivas antes vinculadas essencialmente a elite aos poucos foram se popularizando nas demais camadas sociais, os acadêmicos e a classe operária.

De início, o jornal impresso era o único meio de comunicação disponível acerca dos resultados dos jogos e esportes, mas de acordo com Thompson (2010), a invenção da televisão ampliou a divulgação de tais resultados e também popularizou ainda mais as práticas esportivas, provocando uma crescente nas apostas.

Seguindo o mesmo caminho, a internet possibilitou uma expansão das informações dos esportes, ajudando na globalização do esporte, democratizando o acesso à informação e assim propiciando à indústria esportiva ser uma organização internacional (SALVARO, 2016). Consequentemente a essa estruturação tecnológica foram desenvolvidas as apostas online.

Os sites de apostas esportivas disponibilizam as *odds*, que são as cotações do jogo. Uma *odd* é calculada a partir da probabilidade relacionada a ocorrência do evento em questão. Essas probabilidades são calculadas através de diversos critérios, como classificação do time no campeonato, escalação do time, momento do time no

campeonato, entre outros. Segundo Wheatcroft (2020), para calcular uma *odd* basta inverter a cotação, como visto a seguir.

$$odd_i = \frac{1}{p_i}$$

Em que, odd_i representa a i-ésima cotação e p_i representa a i-ésima probabilidade de ocorrência do evento estudado. O indivíduo ao apostar que determinado evento aconteça, irá multiplicar seu dinheiro apostado pela referida *odd*, e se ele acertar receberá esse valor, que constitui o valor apostado acrescido a um percentual extra, assim obtendo lucro, porém caso o evento não ocorra, o indivíduo perde a aposta e consequentemente o dinheiro apostado.

De acordo com MKTEsportivo (2020), as casas de apostas não utilizam nas cotações somente as probabilidades de ocorrência do evento, há também um acréscimo relacionado ao lucro das casas de apostas. Uma forma de verificar é observar, em jogos de futebol, a soma das probabilidades de vitória do time da casa, empate e vitória do time visitante. Caso a soma das probabilidades exceda 100%, o valor excedente trata-se da margem de lucro das casas de apostas.

Há diversas opções de apostas, podendo ser feitas sobre o resultado de uma partida de futebol, número de gols, número de escanteios, cartões para um determinado jogador, entre outros. É interessante ressaltar que cada esporte tem sua modalidade de aposta, no caso do basquete é possível apostar na quantidade de pontos do jogo, no resultado final, no resultado por quartos, qual jogador mais marcará pontos, entre outros, o que difere do futebol por exemplo.

2.2.1 Legislação

Apresentaremos nessa subseção a legislação de apostas esportivas nos cinco países que estão sendo abordados neste trabalho. Em ordem temos: Brasil, Inglaterra, Itália, Espanha e França.

No Brasil os jogos de apostas estavam banidos desde 1946 de acordo com o Decreto-Lei nº 9.215, até que em 2018 foi promulgada a Lei nº 13.756/18, criando a modalidade de apostas esportivas no território nacional. As únicas apostas permitidas no país se referem as geridas pelo Estado (BARBOSA, 2019).

De acordo com Salvaro (2016), embora não sejam regulamentadas, as apostas online não podem ser consideradas ilegais, visto que os sites ficam hospedados em outros países. Entretanto a fim de captar as apostas e pagar os prêmios dos apostadores, as empresas ligadas aos sites de apostas abrem uma conta corrente no Brasil (SALVARO, 2016). Ainda segundo Salvaro (2016), estimativas indicam que os brasileiros apostam em sites estrangeiros uma quantia na faixa dos 4 bilhões de reais por ano.

Para a Inglaterra e os demais países que compõem o Reino Unido há três leis fundamentais que atuam sobre as apostas esportivas, sendo o *Betting gaming and lotteries act* de 1963, o *Horserace betting and Olympic Lottery act* de 2004 e o *Gambling act* de 2005 (CHAGAS, 2016).

Segundo Olmeda (2011), o Reino Unido foi o primeiro da União Europeia a buscar regular a aposta esportiva de maneira firme e aberta, abrangendo diversas tecnologias como telefone, rádio, televisão, internet e novas tecnologias, contanto que se adeque a Lei de Jogos.

Uma das exigências feitas para o funcionamento de apostas esportivas online é a hospedagem dos sites de apostas em servidores italianos, sites estes que se conectam com suas respectivas empresas para o monitoramento das operações e sua validação. Há também uma lista, que é divulgada mensalmente pela Administração Autônoma dos Monopólios do Estado (AAMS), contendo os eventos autorizados para o funcionamento de apostas esportivas. A AAMS revela ainda mais sua importância sobre o assunto por se tratar do órgão regulador que outorga a licença necessária a todos os operadores que desejem colocar em prática as apostas esportivas online (CHAGAS, 2016). Segundo Kaburakis (2011), na Lei 88/09 é possível constatar a necessidade da obtenção da licença, assim como os desenvolvimentos da política de jogo italiana.

Para a prática das apostas esportivas na Espanha o operador público ou privado necessita da autorização dos órgãos governamentais, da Organização Nacional de Cegos da Espanha (ONCE), dos governos das comunidades autônomas e da empresa pública de Loterias e Apostas do Estado (LAE), que também é responsável pela gestão, exploração e comercialização da loteria esportiva nacional, e é vinculada ao Ministério da Economia e da Fazenda. De acordo com o Decreto Real n. 1710/84 a gestão de apostas de operadores privados é regulamentada pelas

próprias comunidades autônomas ao estabelecerem as diretrizes do setor (CHAGAS, 2016).

Na França a Empresa Reguladora de Jogos Online (ARJEL) é responsável por editar a legislação que define competições, modalidades e formas de apostas permitidas, além de regular a aposta esportiva de acordo com a Lei n. 476/2010, fazendo com que as casas esportivas só atuem sobre eventos nacionais ou internacionais com o seu aval (CHAGAS, 2016).

2.3 *MACHINE LEARNING* EM APOSTAS ESPORTIVAS

Segundo Bunker & Susnjak (2019), a previsão de resultados esportivos não é algo recente na literatura, mas a aplicação de técnicas de *machine learning* (aprendizado de máquina) sim, sendo publicado em 1996 o primeiro estudo nesta temática. Atualmente na literatura de aprendizado de máquina é possível encontrar variados objetivos para sua aplicação, no que diz a respeito as apostas esportivas, o presente tópico aborda algumas de suas utilizações e seus achados.

É possível encontrar materiais sobre aprendizado de máquina em diferentes esportes, como exemplos de esportes individuais há trabalhos sobre natação por Edelmann-Nusser et al. (2002), tênis por Somboonphokkaphan et al. (2009), corridas de cavalos por Davoodi & Khanteymoori (2010) e golfe por Wiseman (2016).

No trabalho de Wilkens (2021), objetivou-se verificar se a utilização do aprendizado de máquina obteria um desempenho melhor do que as probabilidades das casas de apostas, e também observar se com algum dos modelos utilizados seria possível obter retornos positivos consistentes. Para o desenvolvimento deste trabalho foi realizada uma extensa pesquisa na literatura e uma utilização de 5 técnicas de *machine learning*. Como resultado, verificou-se que na média as previsões alcançam a precisão máxima de 70%, independentemente do método utilizado. Além de que as informações contidas no cálculo das probabilidades das casas de apostas impactam bastante no resultado, fazendo com que a adição de novas variáveis cause baixa melhoria na precisão das previsões. Por fim a utilização das previsões em apostas esportivas mostrou alta variabilidade e negativa no longo prazo.

Já sobre esporte coletivo, o trabalho de Bunker & Susnjak (2019), trata-se de uma revisão de literatura sobre a utilização de diferentes técnicas de aprendizado de máquina em diferentes esportes, com o intuito de identificar qual as técnicas mais

utilizadas, quais os esportes mais utilizados nas pesquisas e qual combinação gerava maior grau de assertividade nos resultados dos jogos analisados. Resultando assim em um trabalho robusto, indicando que para cada esporte pode haver uma técnica que se sobressaia em relação as demais devido as particularidades desse esporte. Tornando assim a pesquisa na área necessariamente bem detalhista para que possa gerar o melhor nível de assertividade possível com os dados utilizados.

Com o intuito de prever os resultados dos jogos e identificar os aspectos que mais contribuíam para o resultado das partidas disputadas na Associação Nacional de Basquete (*NBA*), que é a principal liga de basquetebol profissional da América do Norte. Thabtah et al. (2019) utilizou algumas técnicas de aprendizado de máquina, como *Naive Bayes*, rede neural e árvore de decisão em jogos já disputados da *NBA*. A referida pesquisa identificou que rebote defensivo foi considerado o recurso mais impactante no resultado de uma partida na *NBA*, além de destacar que ao selecionar a porcentagem de três pontos, os lances livres realizados e o número total de rebotes, a assertividade do modelo aumentou entre 2% e 4%.

O trabalho desenvolvido por Schlembach et al. (2022), visa prever a quantidade de medalhas olímpicas para cada país através da técnica de aprendizado de máquina chamada de floresta aleatória, utilizando dados socioeconômicos de 206 países entre os anos de 1991 e 2020. As empresas de apostas esportivas oferecem apostas na contagem de medalhas olímpicas e devido ao desempenho da floresta aleatória em dois estágios é possível com uma comparação detalhada dos modelos e uma possível recalibração ser lucrativa, embora os dados utilizados pelas casas esportivas geralmente são diferentes dos usados neste artigo para determinar as *odds*.

No que se refere à aplicação de *machine learning* em trabalhos voltados especificamente para o futebol, Tax & Joustra (2015) elaboraram um trabalho baseado em um sistema de previsão para jogos de futebol da *Eredivisie*, que é a liga nacional holandesa, a partir de dados públicos. Os autores utilizaram dados de 13 temporadas da liga holandesa, técnicas para redução da dimensão dos dados como a Análise de Componentes Principais (*PCA*) e posteriormente classificadores de aprendizado de máquina como *Naive Bayes* e *Multilayer Perceptron*. Ademais os autores adicionaram dados das probabilidades das casas de apostas, as *odds*. Segundo os pesquisadores os resultados podem ser vistos como promissores ao relacionar dados públicos de futebol e as probabilidades das casas de apostas, concebendo sistemas rentáveis.

Knoll & Stübinger (2019) utilizam o aprendizado de máquina para prever os resultados de partidas de futebol com o intuito de obter lucro em apostas esportivas. O retorno do estudo em questão é que a técnica de árvore aleatória apresentou retornos econômicos significativos obtendo em média 5,42% de retorno para cada aposta realizada. É apontado também, que ao apostar em resultados com a maior probabilidade de acontecer, ou seja, que tenham uma *odd* menor, não se tem retorno financeiro. Pois, quando se acerta, há um baixo retorno e há um não retorno quando se perde a aposta.

Hassanniakalager & Newall (2019), ao discutirem sobre o jogo responsável que tanto é falado nas apostas esportivas, por se tratar de um aviso para evitar problemas com vício em aposta, trazem a reflexão de que a indústria das apostas esportivas cita o jogo responsável, mas não dá números a respeito disso, diferentemente do mercado de bebidas que informam o consumidor sobre o percentual de álcool presente na bebida. Para dar números sobre o risco das apostas os autores utilizaram uma regressão logística mista, para quatro tipos de apostas em dados de oito temporadas da *Premier League*. A perda média calculada variou bastante, entre 1,1% a 58,9%. E foi identificado que as *odds*, que são as probabilidades das apostas são importantes para demonstrar o risco associado ao produto, e que essas grandes diferenças no risco são importantes para promover um jogo mais responsável.

Em seu trabalho, Costa (2021) explora diferentes abordagens e técnicas de aprendizagem de máquina, de forma que possa prever os resultados dos jogos antes e durante o jogo, neste último com a precisão sendo atualizada a cada minuto. Para tanto o autor utiliza desde classificadores mais simples até o uso de redes neurais complexas em dados de jogos de nove campeonatos nacionais e explorando outro tipo de aposta além do resultado, que é “ambas as equipes marcam”, justificado pela crescente procura por esse tipo de aposta. A conclusão do trabalho é de que na média a casa de apostas vai ganhar, mas em alguns cenários onde a casa de aposta oferece cotações justas e a aposta é limitada por algum fator há sim um vislumbre de cenário lucrativo para o usuário. Além do mais o autor enfatiza a dificuldade por trás da previsão seja ela pré-jogo ou ao vivo.

Wheatcroft & Sienkiewicz (2021) apontam que o aprendizado de máquina geralmente não se preocupa em fornecer probabilidades de previsão calibradas. O objetivo dos autores é obter probabilidades de previsão ajustadas através de hiperparâmetros e se necessário uma etapa extra de calibração. Para verificar se esse

trabalho representou uma otimização na tomada de decisão, que é se o usuário aposta ou não, foram utilizadas duas estratégias de apostas. Os pesquisadores encontraram algumas evidências que comprovam melhora no desempenho, mas alertam para que o usuário faça a verificação do uso de hiperparâmetros, visto que em alguns casos foi contraproducente.

Os autores Carloni et al. (2021) empregaram técnicas de aprendizado de máquina buscando prever resultados de partidas de futebol, dividindo em duas partes, sendo a primeira a responsável por extrair os dados dos jogos na internet e a segunda parte rodar os algoritmos para verificar o desempenho perante o Retorno Sobre o Investimento (ROI). Para isso foram utilizados dados de 12 países, 49.319 partidas de futebol e seis técnicas de aprendizado de máquina, a regressão logística, *K-Nearest Neighbors* (KNN), *Support Vector Machines* (SVM), *Naive Bayes*, floresta aleatória e rede neural artificial de quatro camadas. Os achados da pesquisa mostram que para 89 novos jogos, em 65 obtiveram êxito, em 13 obtiveram perdas e em 9 não houve aposta ou apostaram e foi cancelada, apostando 10 créditos em cada aposta chagaram no final a um ROI de 26,54%, concluindo com um resultado animador.

3 METODOLOGIA

3.1 BASES DE DADOS

Os dados de futebol utilizados no trabalho estão disponíveis no site (www.football-data.co.uk), que agrega informações de diversos campeonatos ao redor do mundo. Os dados foram coletados agosto de 2022. Dos campeonatos disponíveis foram escolhidas as cinco melhores ligas de futebol do mundo em 2021, que de acordo com o ranqueamento feito pela IFFHS (2022), são as do Brasil, Inglaterra, Itália, Espanha e França, nesta ordem. Inclusive é a primeira vez que o Brasil fica na mais alta posição do *ranking* posição, classificação essa que é feita anualmente desde 1991 (IFFHS, 2022).

Para o trabalho foram utilizados dados das últimas 5 temporadas completas de cada liga e, também, das 10 últimas temporadas completas para fim de comparação dos resultados. Das variáveis presentes nos arquivos do site, foram selecionadas apenas as variáveis que descreviam a partida e variáveis que qualquer usuário pudesse coletar antes das partidas para poder utilizar no código.

No Brasil a quantidade de variáveis disponíveis é bem inferior à das ligas europeias do presente estudo. Ao utilizar o mesmo critério de seleção de variáveis o estudo da liga do Brasil ficou bem mais simples, mas retratando o mesmo roteiro dos demais campeonatos.

Considerando a quantidade de jogos por temporadas, há 1.900 jogos referentes às últimas 5 temporadas para cada liga e 3.800 jogos referentes às últimas 10 temporadas. Em algumas bases foi necessário a exclusão de alguns poucos jogos em virtude da falta de variáveis, o que inviabilizaria a continuidade do código. No caso da *Ligue 1* (França), na temporada 2019/2020 devido a pandemia de Covid-19 o campeonato foi encerrado com 10 rodadas de antecedência e também ficou faltando o jogo de conclusão da 28ª rodada entre o Strasbourg e o Paris Saint-Germain pelo mesmo motivo. Totalizando 279 partidas neste campeonato em relação aos 380 jogos dos demais anos.

3.2 PROCESSAMENTO DOS DADOS

Para cada base de dados montada a partir dos arquivos coletados, foram selecionadas variáveis que descreviam a partida e as variáveis numéricas desejadas para estudo, sendo apenas as variáveis numéricas usadas para cálculo. Houve a alteração dos resultados dos jogos que estavam como uma variável de texto para uma variável numérica, somente para o correto funcionamento do código (D = empate, H = vitória do time da casa, A = vitória do time visitante para 0 = empate, 1 = vitória do time da casa, 2 = vitória do time visitante). O Quadro 1 apresenta as variáveis mencionadas para os códigos dos países europeus.

Quadro 1 – Bases de dados dos campeonatos europeus

Variável	Significado
Div	Divisão da Liga
Date	Data do jogo
HomeTeam	Time da casa
AwayTeam	Time visitante
FTR	Resultado do jogo (0 = empate, 1 = vitória do time da casa, 2 = vitória do time visitante)
B365H	Odd (probablidade) de vitória do time da casa pela Bet365 (casa de apostas)
B365D	Odd (probablidade) de empate pela Bet365 (casa de apostas)
B365A	Odd (probablidade) de vitória do time visitante pela Bet365 (casa de apostas)
BWH	Odd (probablidade) de vitória do time da casa pela Bet&Win (casa de apostas)
BWD	Odd (probablidade) de empate pela Bet&Win (casa de apostas)
BWA	Odd (probablidade) de vitória do time visitante pela Bet&Win (casa de apostas)
IWH	Odd (probablidade) de vitória do time da casa pela Interwetten (casa de apostas)
IWD	Odd (probablidade) de empate pela Interwetten (casa de apostas)
IWA	Odd (probablidade) de vitória do time visitante pela Interwetten (casa de apostas)
PSH	Odd (probablidade) de vitória do time da casa pela Pinnacle (casa de apostas)
PSD	Odd (probablidade) de empate pela Pinnacle (casa de apostas)
PSA	Odd (probablidade) de vitória do time visitante pela Pinnacle (casa de apostas)
VCH	Odd (probablidade) de vitória do time da casa pela VC Bet (casa de apostas)
VCD	Odd (probablidade) de empate pela VC Bet (casa de apostas)
VCA	Odd (probablidade) de vitória do time visitante pela VC Bet (casa de apostas)

WHH	Odd (probablidade) de vitória do time da casa pela William Hill (casa de apostas)
WHD	Odd (probablidade) de empate pela William Hill (casa de apostas)
WHA	Odd (probablidade) de vitória do time visitante pela William Hill (casa de apostas)

Fonte: Autor.

Devido à base de dados do Brasil estar disponibilizada em um modelo diferente das bases de dados dos países europeus a seleção das variáveis presentes na base de dados não foi necessária. No código há a seleção das variáveis de interesse para o andamento do trabalho. O Quadro 2 apresenta as variáveis contidas na base de dados do campeonato brasileiro.

Quadro 2 – Base de dados do campeonato brasileiro

Variável	Significado
Country	País do campeonato
League	Nome da liga
Season	Temporada
Date	Data do jogo
Time	Hora do jogo
Home	Time da casa
Away	Time visitante
HG	Gols marcados pelo time da casa
AG	Gols marcados pelo time visitante
Res	Resultado do jogo (0 = empate, 1 = vitória do time da casa, 2 = vitória do time visitante)
PH	Odd (probablidade) de vitória do time da casa pela Pinnacle (casa de apostas)
PD	Odd (probablidade) de empate pela Pinnacle (casa de apostas)
PA	Odd (probablidade) de vitória do time visitante pela Pinnacle (casa de apostas)
MaxH	Odd (probablidade) máxima de vitória do time da casa
MaxD	Odd (probablidade) máxima de empate
MaxA	Odd (probablidade) máxima de vitória do time visitante
AvgH	Odd (probablidade) média de vitória do time da casa
AvgD	Odd (probablidade) média de empate
AvgA	Odd (probablidade) média de vitória do time visitante

Fonte: Autor.

Utilizamos as informações das casas de apostas que foram encontradas online nas bases de dados supracitadas. Em que as *odds* são as variáveis regressoras utilizadas no estudo. Para o treinamento dos modelos foi definido que aproximadamente 80% dos dados seriam voltados para o treinamento do modelo, 10% para validação do modelo e os últimos 10% para previsão do código e em

seguida a montagem da matriz de confusão para comparação dos resultados de acordo com o previsto nos modelos e o resultado real dos jogos.

O presente trabalho utiliza a linguagem de programação *python* juntamente com o Ambiente de Desenvolvimento Integrado (*IDE*) *Jupyter Notebook* para construção do algoritmo de classificação dos jogos de futebol, devido à grande comunidade de *python* o que permite maior compartilhamento de conhecimentos, auxiliando na construção do código. Segundo Gonçalves (2022), *python* é uma linguagem de programação constantemente empregada para análise de dados, automatização de tarefas, além de construção de sites e softwares.

A confecção dos códigos de programação na linguagem *python* foi separada em duas partes, a primeira visando a criação dos bancos de dados a partir das planilhas fornecidas pelo *site football-data.co.uk* e a segunda parte se refere aos códigos utilizando os métodos abordados no presente trabalho, tanto para as últimas 5 temporadas quanto para as 10 últimas temporadas. É possível visualizar um dos códigos desenvolvidos no trabalho no Apêndice A, juntamente com o *link* onde se encontra o restante dos códigos de programação.

3.3 MODELOS

De acordo com Finkler (2017), o *machine learning* (aprendizado de máquina) pode ser definido como uma área que abrange o desenvolvimento, análise e aplicação de procedimentos para detectar automaticamente padrões em um conjunto de dados.

Para a classificação foram utilizadas quatro técnicas de *machine learning*, sendo estas a regressão logística em sua versão multinomial, o método de SVM, árvore de decisão e o modelo de *Naive Bayes*. Tais modelos foram escolhidos por se tratarem de modelos supervisionados, ou seja, recebem as informações do resultado para assim poderem aprender com os dados apresentados. A seguir estão mais detalhadamente os modelos utilizados.

3.3.1 Regressão Logística

A regressão logística é uma técnica estatística usada no *machine learning*. Produzindo a partir de um conjunto de observações, um modelo para a predição de

valores tomados por uma variável categórica em função de uma ou mais variáveis independentes contínuas e/ou binárias (GONZALEZ, 2018).

De acordo com Masse (2022) em situações reais a variável de resposta pode ser qualitativa como estágio do câncer ou outro tipo de doença. Em casos como esse, o problema é de classificação e para tal a regressão logística se baseia na regressão linear, porém não modela a variável de resposta diretamente, mas sim a probabilidade de que a mesma pertença a uma categoria específica.

Existem 3 tipos básicos de regressão logística:

- Binária: há apenas duas respostas possíveis para a variável categórica;
- Multinomial: há três ou mais respostas possíveis para a variável categórica;
- Ordinal: se assemelha a multinomial por haver três ou mais respostas possíveis para a variável categórica, porém neste caso as variáveis são ordenadas, como exemplo da classificação de estrelas de um hotel.

Para este estudo utilizaremos o modelo de regressão logística multinomial, pois o objetivo é classificar a partir dos dados o resultado das partidas, assim possuindo três respostas possíveis, mas sem necessitar de ordem. Antes de mostrar o modelo de regressão logística multinomial, é interessante observar o funcionamento do modelo binário, que segue a distribuição Bernoulli.

3.3.1.1 Regressão Logística Binária

Com o intuito de determinar a probabilidade de um modelo binário Y_n, e dado um conjunto de preditores X_n, a regressão logística pode ser vista seguindo um modelo de probabilidade Bernoulli. As equações a seguir foram retiradas do artigo de Schein e Ungar (2007), em que temos a probabilidade condicional de Y_n dado x_n como uma função logística binária com parâmetros x_n e w, que é o resultado discreto da classificação.

$$P(Y_n = 1|x_n) \doteq \sigma(w \cdot x_n)$$

Obtemos sua função de verossimilhança por meio do produtório de uma Bernoulli:

$$P(y|x_n, n = 1, \ldots, N) = \prod_n \sigma(w \cdot x_n)^{y_n}(1 - \sigma(w \cdot x_n))^{(1-y_n)}$$

$$= \prod_n \sigma(w \cdot x_n)^{y_n}\, \sigma(-w \cdot x_n)^{(1-y_n)}$$

A função logística utilizada na regressão logística é:

$$\sigma(\theta) = \frac{1}{1 + \exp\,[-\theta]}$$

Por se tratar uma função crescente contínua mapeando qualquer valor real θ no intervalo (0,1) é apropriada para representar a probabilidade de um ensaio de Bernoulli (SCHEIN & UNGAR, 2007).

3.3.1.2 Regressão Logística Multinomial

Segundo Schein & Ungar (2007), o caso multinomial Y_n pode assumir três ou mais resultados discretos ao invés de 0 ou 1, quantidade essa representada por c, portanto a probabilidade é definida pela função logística:

$$P(Y_n = c|x_n) \doteq \pi(c, x_n, w) = \frac{\exp\,(w_c \cdot x_n)}{\sum_{c'} \exp\,(w_{c'} \cdot x_n)}$$

O vetor de parâmetros w da regressão logística binária é dividido para cada categoria gerando um conjunto de vetores w_c. Portanto a função de verossimilhança chega no produtório:

$$P(y|x_n, n = 1, \ldots, N) = \prod_{nc} \pi(c, x_n, w)^{y_{nc}}$$

O caso multinomial é uma generalização do caso binário, e ao utilizarmos os valores de $w_0 = 0$ e $w_1 = w$ é possível retornar à função logística do caso binário.

3.3.1.3 Estimação dos parâmetros

Para a estimação dos parâmetros é utilizado o método de estimação da máxima verossimilhança. Schein & Ungar (2007), apresentam os parâmetros da seguinte forma:

$$\mathcal{L}(\sqrt{n}(\hat{w} - w)) \rightarrow N(0, F^{-1}(w))$$

$$\hat{w}_n = \bar{w}_n + O_p(\frac{1}{m^{1/2}})$$

Em que, $\mathcal{L}$ é a distribuição que o seu argumento segue, $F(w)$ é a matriz de informação de Fischer do modelo, $\hat{w}_n$ e $\bar{w}_n$ são a estimativa baseada em uma amostra e a estimativa esperada de w, respectivamente, enquanto O_p se trata de uma taxa de convergência de probabilidade.

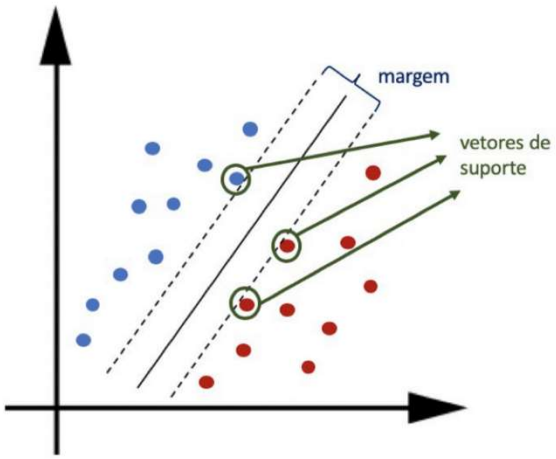
3.3.2 *Support Vector Machines*

Segundo Santos (2002) os fundamentos do SVM têm origem na Teoria de Aprendizagem Estatística, teoria esta que foi trabalhada inicialmente pelo pesquisador russo Vladmir Vapnik e seus colaboradores. Na segunda edição de livro *The Nature of Statistical Learning Theory* é dito que no lançamento de sua primeira edição os métodos de aprendizado SVM eram totalmente novos e que nesta segunda edição havia três novos capítulos exclusivamente sobre o SVM (VAPNIK, 2000). De acordo com Noble (2006), SVM é um algoritmo computacional que aprende a rotular objetos. Exemplos de aplicações citados por Lorena & Carvalho (2007), vão desde a categorização de textos, análise de imagens e utilização em Bioinformática.

Para produzir o resultado desejado o modelo utiliza funções kernel para transformar os dados de treino originais para uma dimensão maior, o objetivo é que nesta nova dimensão seja possível separar os dados linearmente de forma ótima através de hiperplanos (ESCOVEDO & KOSHIYAMA, 2020).

O modelo utiliza vetores de suporte e margens para encontrar esses hiperplanos que melhor separem os dados, para assim conseguir chegar na classificação dos dados. Na Figura 1 é possível ver um exemplo mais básico disso.

Figura 1 – Funcionamento do SVM



Fonte: Escovedo e Koshiyama (2020).

Este exemplo retrata a classificação em dois grupos, em que a reta está fazendo a separação do conjunto de dados utilizando-se das retas pontilhadas. A margem é a distância entre essas as duas retas pontilhadas, e que ao serem movimentadas determinam os vetores auxiliares, que são os primeiros pontos do conjunto de dados a serem interceptados pelas retas pontilhadas.

3.3.3 Árvore de Decisão

A primeira vez que as árvores de decisão foram apresentadas foi em 1975 por J. Ross Quinlan em seu livro *Machine Learning*, e em 1983 ele criou o primeiro algoritmo capaz de criar as árvores de decisão, que foi nomeado de *Iterative Dichotomiser 3* (*ID3*), devido ao seu trabalho o autor ficou conhecido por ser o pai das árvores de decisão (NAPOLEÃO, 2018).

Segundo Escovedo & Koshiyama (2020), há duas formas de utilizar as árvores, para classificação ou para regressão, em que se chamadas de árvores de classificação e árvores de regressão, respectivamente.

A árvore de decisão pode ser utilizada no dia a dia para várias situações que envolvam tomadas de decisão, como a escolha de um roteiro de viagem, onde as ramificações mostrariam as paradas durante o percurso (NAPOLEÃO, 2018).

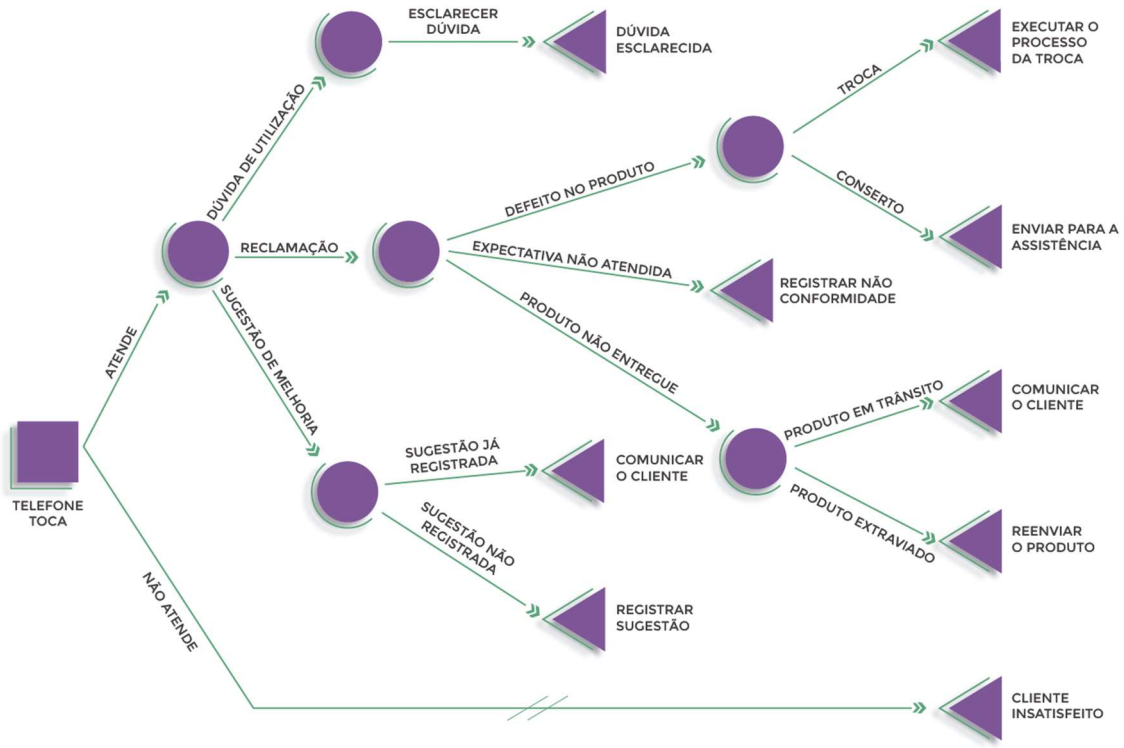
Esse método de aprendizado de máquina usa vários algoritmos para decidir dividir um nó em dois ou mais nós. Quanto maior for a quantidade de nós criados/divididos maior deve ser a homogeneidade dos nós resultantes, que são os nós finais. Para isso a árvore de decisão divide o conjunto de treinamento formando

subconjuntos mais homogêneos em relação a variável dependente e, repete o processo até que consiga um conjunto bem homogêneo, para chegar à atribuição de um único valor, ou seja a classificação (MEIRA et al., 2008).

De acordo com Chauhan (2022a), o fluxo de funcionamento da árvore de decisão se baseia em prever um rótulo de classe para um registro, começando da raiz da árvore, e em seguida fazer a comparação dos valores do atributo raiz com o atributo do registro. A partir dessa comparação passamos para o ramo correspondente a esse valor e saltamos para o próximo nó da árvore.

Na Figura 2 é possível ver um exemplo mais simples do funcionamento da árvore de decisão no setor de atendimento ao cliente.

Figura 2 – Árvore de decisão no setor de atendimento ao cliente



Fonte: Napoleão (2018).

Neste exemplo, quando um funcionário do setor de atendimento ao cliente notar o telefone tocando o mesmo precisará tomar uma decisão inicial, que é representada pelo quadrado e chamada de nó raiz. Se optar por atender o telefone irá chegar em um círculo, representando uma probabilidade, chamado de nó folha. Para cada ramificação, a representação é feita por uma seta. Já o triângulo representa um resultado final. É possível então visualizar a quantidade de opções relacionadas à

decisão inicial de atender ou não o telefone, conforme a estrutura de uma árvore de decisão.

3.3.4 *Naive Bayes*

O modelo possui esse nome por ser baseado no Teorema de Bayes, já que ambos tratam do cálculo de probabilidades condicionais (SACRAMENTO, 2021). Segundo Escovedo & Koshiyama (2020), o Teorema de Bayes determina a probabilidade de um evento com base em um conhecimento anterior podendo estar relacionado a este evento. Um exemplo clássico de utilização desse modelo é o classificador de *spam*. Ele analisa e-mails e busca classificar se o mesmo é um spam ou não, baseando-se em sua estrutura e informações. Segundo Chauhan (2022b), o algoritmo de *Naive Bayes* tem tido sucesso em diversas situações, ainda mais quando se trata de problemas de processamento de linguagem natural.

De acordo com Hrouda-Rasmussen (2021), na implementação de *Naive Bayes* na forma Gaussiana, é assumido que as densidades condicionais de classe são normalmente distribuídas para um conjunto de dados de treinamento de N variáveis de entrada x com variáveis-alvo correspondentes t como é possível observar a seguir:

$$P(x \mid t = c, \mu_c, \sum_c) = N\ (x \mid \mu_c, \sum_c)$$

Sendo c a classe, μ_c o vetor médio específico da classe e Σ_c a matriz de covariância do vetor específico. Ao aplicar o Teorema de Bayes conseguimos calcular a classe posterior:

$$\overbrace{P(t = c|x, \mu_c, \sum_c)}^{classe\ posterior} = \frac{\overbrace{P(t = c|x, \mu_c, \sum_c)}^{densidade\ condicional\ da\ classe} \cdot \overbrace{P(t = c)}^{classe\ anterior}}{\sum_{k=1}^{K} P(x \mid t = k, \mu_k, \sum_c)\ P(t = k)}$$

E em seguida classificar a classe de x:

$$\hat{h}(x) = \underset{c}{\operatorname{argmax}}\, P(t = c|x, \mu_c, \sum_c)$$

3.4 MÉTRICAS UTILIZADAS

As métricas utilizadas neste trabalho para a avaliação das previsões são a acurácia e *F1-Score*. Ambas as medidas são comumente utilizadas pela literatura de *machine learning*. Segundo Lima et al. (2021), a acurácia representa a proporção de classificações corretas do total de classificações realizadas, indicando o desempenho do modelo. Tal métrica é dada por:

$$Acurácia = \frac{TP + TN}{TP + FP + TN + FN}$$

Em que, TP é o número de verdadeiros positivos, TN o número de verdadeiros negativos, FP o número de falsos positivos e FN o número de falsos negativos. Ainda de acordo com Lima et al. (2021), *F1-Score* é uma média harmônica entre precisão e sensibilidade, que é dada por:

$$F1 - Score = \frac{2 \cdot Precisão \cdot Sensibilidade}{Precisão + Sensibilidade}$$

$$Precisão = \frac{TP}{TP + FP} \qquad Sensibilidade = \frac{TP}{TP + FN}$$

4 RESULTADOS

Abordando os métodos previamente discutidos serão apresentados a seguir os valores das acurácias por métodos e ligas nacionais de futebol. De acordo com o ranqueamento feito pela IFFHS (2022), Federação Internacional de História e Estatísticas do Futebol, as 5 melhores ligas de futebol do mundo no ano de 2021 são a do Brasil, Inglaterra, Itália, Espanha e França, nesta ordem.

4.1 UTILIZANDO 5 TEMPORADAS

Com o intuito de obter o resultado de um jogo de futebol para cada uma das 5 ligas abordadas neste trabalho, foram utilizados dados das últimas 5 temporadas, desde a temporada 2017 até a 2021 no caso do Brasil e da temporada 2017/2018 até a 2021/2022 para as ligas europeias. Para tanto foram criados códigos voltados para a utilização dos quatro métodos estatísticos supracitados. É possível visualizar os resultados encontrados na Tabela 1.

Tabela 1 – Resultados das ligas com 5 temporadas

Liga	***Método***							
	Regressão Logística		**SVM**		**Árvore de Decisão**		***Naive Bayes***	
	Ac. (%)	***F1* (%)**	**Ac. (%)**	***F1* (%)**	**Ac. (%)**	***F1* (%)**	**Ac. (%)**	***F1* (%)**
Série A	**44,21**	44,21	43,16	43,16	42,63	41,67	43,68	43,68
Premier League	**56,32**	56,32	56,32	56,32	43,68	42,29	47,89	47,89
Serie A	55,26	55,26	**55,79**	55,79	44,74	42,54	47,89	47,89
La Liga	**47,89**	47,89	47,89	47,89	44,74	43,38	38,42	38,42
Ligue 1	**50,56**	50,56	50,00	50,00	37,78	36,85	38,89	38,89

Fonte: Autor.

A tabela 1 indica os valores de acurácia e *F1-Score* encontrados ao gerar as previsões dos jogos. Sendo possível observar que na liga inglesa houve uma assertividade maior nas previsões para todos os métodos em relação a liga brasileira e também para a liga francesa.

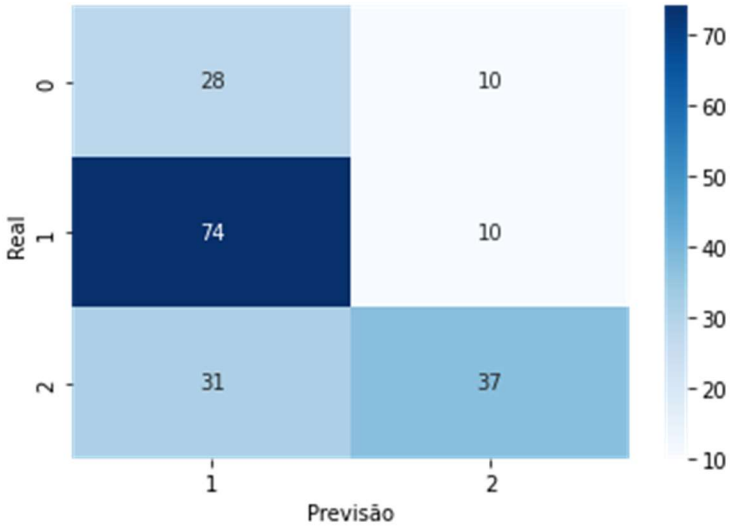
Os resultados em destaque se referem à melhor acurácia por liga e método, em que nos dois casos onde houve empate entre os métodos de regressão logística e SVM foi considerado o de regressão logística para contabilização dos métodos que

se sobressaíram na geração das previsões, com o intuito de manter uma padronização nos métodos.

Segundo Santos (2020), uma outra forma de avaliar o desempenho dos classificadores presente na literatura é a estimação da curva Característica de Operação do Receptor (ROC), porém a mesma é utilizada em classificadores binários. Porém não foram utilizados classificadores binários no presente estudo.

Como a melhor acurácia foi obtida ao utilizar o método de regressão logística na Premier League. É possível observar na Figura 3 os resultados da previsão do código em confronto com os resultados reais, de acordo com os últimos 10% dos dados. Também fica visível que para este conjunto de dados o método não chegou a prever resultados de empate.

Figura 3 – Matriz de confusão com 5 temporadas da *Premier League*



Fonte: Autor.

Na matriz de confusão o eixo x representa a previsão gerada pelo algoritmo, enquanto no eixo y está o resultado real da partida de futebol. Ao observarmos o par $(1,2)$ notamos que houve 31 partidas em que o algoritmo gerou a previsão como vitória do time da casa, mas o resultado correto da partida foi a vitória do time visitante. A tonalidade da cor azul apenas está relacionada à quantidade de partidas por combinação entre o eixo x e y.

Uma exemplificação é observar como na combinação $(2,1)$ a cor está bem clara por só haver 10 partidas nesta situação, enquanto no par $(1,1)$ a cor fica escura por apresentar muitas partidas nesta situação, em que o algoritmo classificou o resultado

desses jogos como vitória do time da casa e o resultado real destas 74 partidas foi de vitória do time da casa, o que significa que a previsão está correta.

Na Tabela 2 encontra-se a classificação dos métodos utilizados de acordo com a maior acurácia em cada liga nacional. Nota-se que o método que mais obteve expressão para as últimas 5 temporadas foi o de regressão logística, obtendo uma acurácia em 4 das 5 ligas, embora em 2 delas houve um empate com o método de SVM.

Tabela 2 – Classificação dos métodos com 5 temporadas

Método	**Nº de vezes com melhor desempenho**	**Ligas onde obteve o melhor desempenho**
Regressão Logística	4	Série A *Premier League* *La Liga* *Ligue 1*
SVM	1	*Serie A*
Árvore de Decisão	0	-
Naive Bayes	0	-

Fonte: Autor.

4.2 UTILIZANDO 10 TEMPORADAS

Ao gerar as previsões com dados das 10 temporadas é possível observar na Tabela 3 o cenário geral das acurácias e *F1-Score* por método e liga nacional de futebol. Nota-se que a assertividade da liga brasileira se aproximou bastante da liga inglesa e francesa, para os métodos de regressão logística, SVM e árvore de decisão, destoando das demais apenas em relação ao método de *Naive Bayes*. Porém para essas 3 ligas o desempenho ficou abaixo dos 50%, sendo somente as ligas inglesa e italiana que conseguiram obter uma assertividade acima dos 50%.

Tabela 3 – Resultados das ligas com 10 temporadas

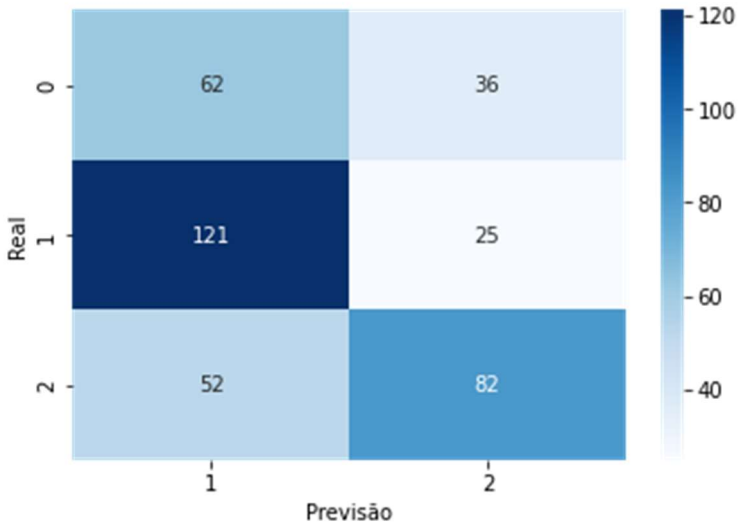
Liga	***Método***							
	Regressão Logística		**SVM**		**Árvore de Decisão**		***Naive Bayes***	
	Ac. (%)	***F1* (%)**	**Ac. (%)**	***F1* (%)**	**Ac. (%)**	***F1* (%)**	**Ac. (%)**	***F1* (%)**
Série A	48,42	48,42	48,68	48,68	40,00	37,05	**49,47**	49,47
Premier League	**49,74**	49,74	49,47	49,47	40,26	37,26	38,16	38,16
Serie A	**56,61**	56,61	56,61	56,61	46,56	44,62	50,26	50,26
La Liga	**53,56**	53,56	52,51	52,51	41,95	40,56	41,16	41,16
Ligue 1	**48,64**	48,64	48,64	48,64	41,85	40,38	40,22	40,22

Fonte: Autor.

Embora o método de *Naive Bayes* tenha se destacado na Série A, o mesmo não ocorreu na *Premier League*, *La Liga* e na *Ligue 1*, visto que o método obteve o pior desempenho perante os demais. No caso da liga italiana o método só não obteve novamente o pior desempenho devido ao baixo desempenho do método de árvore de decisão.

Em virtude do melhor desempenho do método de regressão logística na *Serie A* italiana. É possível observar na Figura 4 os resultados da previsão do código em confronto com os resultados reais, de acordo com os últimos 10% dos dados. Também fica visível que para este conjunto de dados o método não chegou a prever resultados de empate.

Figura 4 – Matriz de confusão com 10 temporadas da *Serie A*



Fonte: Autor.

Observa-se que os dois casos que mais vezes se repetiram foram casos em que as previsões estavam corretas, sendo 121 partidas em que o algoritmo previu que o resultado da partida seria a vitória do time da casa e que de fato aconteceu, como nas 82 partidas que a previsão se referia à vitória do time visitante, como de fato ocorreu. O terceiro caso de maior repetição foi quando o algoritmo previu que a partida terminaria com a vitória do time da casa, mas o resultado real foi empate.

A seguir a Tabela 4 mostra a classificação dos métodos utilizados de acordo com a maior acurácia em cada liga nacional. Onde a regressão logística se destacou mais ao obter melhor assertividade com os dados das últimas 10 temporadas em 4 das 5 ligas, embora que na *Serie A* e na *Ligue 1* houve empate com o método de SVM. O outro método que obteve destaque, porém em apenas 1 liga, foi o de *Naive Bayes* com dados da liga italiana.

Tabela 4 – Classificação dos métodos com 10 temporadas

Método	**Nº de vezes com melhor desempenho**	**Ligas onde obteve o melhor desempenho**
Regressão Logística	4	*Premier League* *Serie A* *La Liga* *Ligue 1*
SVM	0	-
Árvore de Decisão	0	-
Naive Bayes	1	Série A

Fonte: Autor.

4.3 5 TEMPORADAS *VERSUS* 10 TEMPORADAS

A expectativa ao analisar um maior número de informações é de que o treinamento do código obtenha um melhor resultado, ou seja, que a assertividade dos métodos aumente, porém o que foi observado é que não houve esse resultado de forma unânime, como é apresentado na Tabela 5. Isto pode se dar pelo fato de que uma informação de 10 anos atrás para uma liga que tenha recebido uma recente injeção monetária não traga "boas informações" para a previsão.

Tabela 5 – Resultados das ligas por número de temporadas

Liga	Acurácia das últimas 5 temporadas (%)	Acurácia das últimas 10 temporadas (%)
Brasil - Série A	44,21	**49,47**
Inglaterra – *Premier League*	**56,32**	49,74
Itália – *Serie A*	55,79	**56,61**
Espanha – *La Liga*	47,89	**53,56**
França – *Ligue 1*	**50,56**	48,64

Fonte: Autor.

Na liga brasileira, italiana e espanhola o maior número de temporadas no treinamento dos métodos proporcionou a melhoria na assertividade das previsões geradas, porém para a liga inglesa e francesa o efeito foi oposto.

De todos os resultados obtidos o método de regressão logística foi o que mais vezes esteve presente como o método que proporcionou melhor acurácia como é possível observar na Tabela 6.

Tabela 6 – Classificação dos métodos por acurácia nas ligas

Método	Nº de vezes com melhor desempenho	Ligas onde obteve o melhor desempenho	
		Últimas 5 temporadas	Últimas 10 temporadas
Regressão Logística	8	Série A *Premier League* *La Liga* *Ligue 1*	*Premier League* *Serie A* *La Liga* *Ligue 1*
SVM	1	*Serie A*	-
Árvore de Decisão	0	-	-
Naive Bayes	1	-	Série A

Fonte: Autor.

A quantidade de ligas em que a regressão logística esteve presente como melhor método para previsão dos resultados das partidas não se alterou em relação a 5 e 10 temporadas, mas houve uma alteração nas ligas. Em que a *Serie A* possuía melhor desempenho com o método de SVM com dados de 5 temporadas, mas para 10 temporadas o melhor método é o de regressão logística. Em relação a Série A

brasileira houve uma mudança de método passando de regressão logística para *Naive Bayes*.

Apenas o método de árvore de decisão que não conseguiu representar melhor assertividade em qualquer uma das 5 ligas utilizadas no trabalho, tanto para 5 temporadas como para 10.

5 CONSIDERAÇÕES FINAIS

O presente trabalho gerou previsões dos resultados de jogos de futebol a partir de quatro métodos de *machine learning* e cinco ligas nacionais de futebol, com o objetivo de identificar qual a melhor combinação de método e liga que geraria maior nível de assertividade nas previsões, além de verificar se com uma série de dados mais longa seria possível obter uma melhoria na assertividade das previsões.

Ao gerar as previsões para os 4 métodos escolhidos para as últimas 5 temporadas completas de 5 ligas nacionais de futebol e também para 10 temporadas, em que foram utilizados dados de 18.841 partidas de futebol, foi possível chegar à conclusão de que a combinação que representou maior assertividade nas previsões para 5 temporadas foi ao utilizar regressão logística na *Premier League* obtendo uma assertividade de 56,32%, já na combinação para 10 temporadas foi a utilização de regressão logística ou SVM na *Serie A*, visto que ambos os métodos representaram uma acurácia de 56,61%.

Dos 4 métodos utilizados, o que mais se destacou foi o de regressão logística ao figurar 8 vezes com o maior nível de assertividade nas previsões do total de 10 testes, enquanto o algoritmo de árvore de decisão obteve o pior desempenho, pois não chegou a representar a melhor assertividade em nenhum dos casos avaliados.

Ao comparar dados dos melhores desempenhos nas previsões por liga com o conjunto de dados de 5 temporadas para 10 temporadas seria possível imaginar que para uma maior quantidade de informações disponíveis o nível de assertividade da previsão pudesse melhorar, porém não foi uma unanimidade, já que das 5 ligas, em 3 casos (Série A, *Serie A* e *La Liga*) houve uma melhoria na previsão dos resultados dos jogos enquanto na *Premier League* e *Ligue 1* foi o oposto.

Ao utilizar dados das 10 temporadas ao invés de 5 temporadas para o aprendizado de máquina e a previsão de resultados, o resultado mais positivo é notado na Série A que obteve um ganho de 11,90% na previsão, seguido da *La Liga* com 11,84%. O incremento na *Série A* foi mínimo, com 1,47%, já para a *Ligue 1* obteve uma perda de 3,80%, enquanto a *Premier League* teve o pior resultado, perdendo 11,68%. É interessante notar que as duas ligas que apresentaram redução na assertividade das previsões ao utilizar dados mais antigos são a *Premier League* e a *Ligue 1*, ligas que passaram tiveram grandes injeções de dinheiro nas temporadas

mais recentes. Diante do exposto é possível visualizar que os objetivos propostos neste trabalho foram atendidos.

Como dificuldades enfrentadas durante a pesquisa é importante citar a seleção dos dados, pois embora o futebol seja um esporte amplamente conhecido e haja diversas estatísticas foi difícil para encontrar bases de dados com grandes quantidades de variáveis e de livre acesso, além de deter as mesmas variáveis para diferentes ligas, visto que há níveis de exposição diferentes para as ligas. Dificuldades estas que possivelmente estão refletidas nas acurácias encontradas no trabalho, que foram baixas, próximas de 60%, e que se sanadas podem gerar bons resultados.

Por fim, como sugestão para trabalhos futuros, sugere-se a utilização de hiperparâmetros com o intuito de aprimorar a assertividade dos modelos, pois nas técnicas de *machine learning* os parâmetros dos modelos são ajustados diretamente pelo processo de aprendizado, enquanto nos hiperparâmetros as variáveis são definidas antes do processo de treinamento. Outra sugestão é trabalhar com outras variáveis além das probabilidades de casas de apostas, como o número de faltas, cartões e dados financeiros, visando diversificar ainda mais as informações das partidas de futebol e melhorar os treinamentos, buscando uma previsão mais assertiva. Diante dos resultados observados na comparação entre 5 e 10 temporadas para a *Premier League* e para a *Ligue 1*, outra sugestão é utilizar dados mais recentes como por exemplo as 3 últimas temporadas, em virtude do investimento feito nas ligas em temporadas mais recentes.

REFERÊNCIAS

Alberti, C., & Tassi, T. (2022). *Inglês começa após R$ 6 bi gastos na janela e como base da seleção de Tite*. UOL. https://www.uol.com.br/esporte/futebol/ultimas-noticias/2022/08/04/premier-league-dominio-mercado-da-bola-e-selecao-brasileira.htm

Andrade, G. (2021a). *Campeonato Italiano : campeões , artilheiros e estatísticas*. Esportelândia. https://www.esportelandia.com.br/futebol/campeonato-italiano/

Andrade, G. (2021b). *Premier League : Maiores Campeões , Artilheiros e Recordes*. Esportelândia. https://www.esportelandia.com.br/futebol/premier-league/

Barbosa, A. (2019). *Apostas esportivas online é legal ou ilegeal no Brasil?* Futebol Latino. https://futebolatino.lance.com.br/apostas-esportivas-online-e-legal-ou-ilegal-no-brasil/#

Barbosa, F. A. B., & Filho, T. N. (2020). Responsabilidade civil das casas de apostas esportivas. In *Ensaios de Responsabilidade Civil* (pp. 24–36). Editora Fi. https://www.editorafi.org/43responsabilidade

Bayer, R. S. (2014). *A AUTONOMIA DAS ORGANIZAÇÕES INTERNACIONAIS ESPORTIVAS* [Universidade Federal de Santa Catarina]. https://repositorio.ufsc.br/xmlui/handle/123456789/129270

Bertozzo, R. J. (2019). *Aplicação de machine learning em dataset de consultas médicas do sus* [Universidade Federal de Santa Catarina]. https://repositorio.ufsc.br/bitstream/handle/123456789/202663/TCC.pdf?sequence=1&isAllowed=y

Beting, M. (2016). *Taça Brasil é Brasileirão? Robertão é Brasileirão? Entenda. Ou não.* UOL. https://maurobeting.blogosfera.uol.com.br/2016/05/31/taca-brasil-e-brasileirao-robertao-e-brasileirao-entenda-ou-nao/

Bunker, R., & Susnjak, T. (2019). *The Application of Machine Learning Techniques for Predicting Match Results in Team Sport: A Review*. https://arxiv.org/abs/1912.11762

Carloni, L., Angelis, A. De, Sansonetti, G., & Micarelli, A. (2021). A Machine Learning Approach to Football Match Result Prediction. In *Communications in Computer and Information Science - HCI International 2021 - Posters* (Vol. 1420, pp. 473–480). Springer. https://doi.org/https://doi.org/10.1007/978-3-030-78642-7_63

Regulamento Específico da Competição Campeonato Brasileiro da Série A 2017, 15 (2017). https://www.cbf.com.br/futebol-brasileiro/noticias/campeonato-brasileiro-serie-a/brasileirao-regulamento-especifico-e-plano-geral

CHAGAS, J. M. (2016). A (IM)POSSIBILIDADE DE REGULAMENTAÇÃO DAS APOSTAS ESPORTIVAS NO ORDENAMENTO JURÍDICO BRASILEIRO [Universidade Federal de Santa Catarina]. In *Universidade Federal de Santa Catarina*. https://repositorio.ufsc.br/handle/123456789/166160

Chauhan, N. S. (2022a). *Decision Tree Algorithm, Explained*. KDnuggets. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

Chauhan, N. S. (2022b). *Naïve Bayes Algorithm: Everything You Need to Know*. KDnuggets. https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html

Collinson, C. (2019). *New Premier League football season : Can the ' Big Six ' monopoly be broken ?* BBC. https://www.bbc.com/sport/football/49165976

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

http://dspace.sti.ufcg.edu.br:8080/xmlui/bitstream/handle/riufcg/20618/ÍGOR BARBOSA DA COSTA - TESE %28PPGCC%29 2021.pdf?sequence=3&isAllowed=y

Davoodi, E., & Khanteymoori, A. R. (2010). Horse racing prediction using Artificial Neural Networks. *RECENT ADVANCES in NEURAL NETWORKS, FUZZY SYSTEMS & EVOLUTIONARY COMPUTING*, *Junho*, 155–160. https://www.researchgate.net/publication/228847950_Horse_racing_prediction_using_artificial_neural_networks

Edelmann-Nusser, J., Hohmann, A., & Henneberg, B. (2002). Modeling and prediction of competitive performance in swimming upon neural networks. *European Journal of Sport Science*, *2*(2), 1–10. https://doi.org/10.1080/17461390200072201

Escovedo, T., & Koshiyama, A. S. (2020). *Introdução a Data Science - Algoritmos de Machine Learning e métodos de análise*. Casa do Código. https://www.casadocodigo.com.br/products/livro-data-science

Estadão. (2022a). Campeonato Italiano: quando começa, onde assistir, tabela, times e regulamento da Serie A 2022/23. *Estadão*. https://esportes.estadao.com.br/blogs/bate-pronto/campeonato-italiano-quando-comeca-onde-assistir-tabela-times-e-regulamento-da-serie-a-2022-23/

Estadão. (2022b). LaLiga: quando começa, onde assistir, tabela, times e regulamento do Campeonato Espanhol 2022/23. *Estadão*. https://esportes.estadao.com.br/blogs/bate-pronto/laliga-quando-comeca-onde-assistir-tabela-times-e-regulamento-do-campeonato-espanhol-2022-23/

Estadão. (2022c). Ligue 1: quando começa, onde assistir, tabela, times e regulamento do Campeonato Francês 2022/23. *Estadão*. https://esportes.estadao.com.br/blogs/bate-pronto/ligue-1-quando-comeca-onde-assistir-tabela-times-e-regulamento-do-campeonato-frances-2022-23/?utm_source=headtopics&utm_medium=news&utm_campaign=2022-08-02

Fabbri, B. (2021). *Conheça a história da Premier League , o Campeonato Inglês*. Jornal DCI. https://www.dci.com.br/esporte/futebol/premier-league/conheca-a-historia-da-premier-league-o-campeonato-ingles/113132/

FIFA. (2007). FIFA Big Count 2006: 270 million people active in football. *FIFA Communications Division, Information Services*, *31*, 1–12. https://digitalhub.fifa.com/m/55621f9fdc8ea7b4/original/mzid0qmguixkcmruvema-pdf.pdf

Finkler, A. C. (2017). APRENDIZAGEM DE MÁQUINA APLICADA À PREVISÃO DOS MOVIMENTOS DO IBOVESPA [Universidade Federal do Paraná]. In *Universidade Federal do Paraná*. https://acervodigital.ufpr.br/bitstream/handle/1884/49395/R - D - ALINE CRISTIANE FINKLER.pdf?sequence=1&isAllowed=y

Gerhardt, W. (1979). *The Colorful History of a Fascinating Game: More than 200 Years of Football*. FIFA News. http://web.archive.org/web/20040803075659/http://fifa.com/en/history/history/0,1283,1,00.html

Globo. (2021). *O mercado de apostas esportivas*. Globo. https://gente.globo.com/o-mercado-de-apostas-esportivas/

Gonçalves, G. A. (2022). COMPARAÇÃO DE MODELOS DE MACHINE LEARNING PARA PREVISÃO DE PREÇO DE FECHAMENTO DE UMA AÇÃO DO SETOR BANCÁRIO LISTADA NA B3 [Universidade Federal de Uberlândia]. In *Universidade Federal de Uberlândia*. https://repositorio.ufu.br/handle/123456789/35505

Gonzalez, L. de A. (2018). *Regressão Logística e suas Aplicações* [Universidade Federal do Maranhão]. https://monografias.ufma.br/jspui/bitstream/123456789/3572/1/LEANDRO-GONZALEZ.pdf

Hassanniakalager, A., & Newall, P. W. S. (2019). A machine learning perspective on responsible gambling. *Behavioural Public Policy*, *6*(2), 237–260. https://doi.org/10.1017/bpp.2019.9

Hrouda-Rasmussen, S. (2021). *(Gaussian) Naive Bayes*. Towards Data Science. https://towardsdatascience.com/gaussian-naive-bayes-4d2895d139a

IFFHS. (2022). *THE STRONGEST NATIONAL LEAGUE IN THE WORLD 2021 by IFFHS*. https://www.iffhs.com/posts/1607

Kaburakis, A. (2011). European Union Law, Gambling, and Sport Betting. European Court of Justice Jurisprudence, Member States Case Law, and Policy. In *Sports Betting: Law and Policy* (pp. 27–97). https://doi.org/10.1007/978-90-6704-799-9_4

Knoll, J., & Stübinger, J. (2019). Machine-Learning-Based Statistical Arbitrage Football Betting. *KI - Kunstliche Intelligenz*, *34*(1), 69–80. https://doi.org/10.1007/s13218-019-00610-4

Leal, U. (2014). *Serie A Italiana anos 1980 e 1990 : O melhor campeonato nacional da história*. Trivela. https://trivela.com.br/italia/serie-a/serie-a-italiana-anos-1980-e-1990-o-melhor-campeonato-nacional-da-historia/

Lima, J. P. de O., Almeida, C. D. F. de, Filho, L. C. S. de A., & Oliveira, R. C. de. (2021). Sistema em Nuvem para Identificação de Células Sanguíneas Infectadas pelo protozoário da Malária utilizando Redes Neurais Convolucionais. *Procedings Do XV Simpósio Brasileiro de Automação Inteligente*, 1668–1673. https://doi.org/10.20906/sbai.v1i1.2791

LIRA, P. E. M. DE. (2018). *OS DESAFIOS PARA A REGULAMENTAÇÃO DAS APOSTAS ESPORTIVAS FRENTE AO SISTEMA JURÍDICO BRASILEIRO* [Universidade Federal de Campina Grande]. http://dspace.sti.ufcg.edu.br:8080/jspui/handle/riufcg/15330

Lorena, A. C., & Carvalho, A. C. P. L. F. de. (2007). Uma Introdução às Support Vector Machines. *Revista de Informática Teórica e Aplicada*, 25. https://doi.org/https://doi.org/10.22456/2175-2745.5690

Masse, N. R. (2022). *Scholarship at UWindsor* (Issue August) [Universidade de Windsor]. https://scholar.uwindsor.ca/cgi/viewcontent.cgi?article=1257&context=major-papers

Meira, C. A. A., Rodrigues, L. H. A., & Moraes, S. A. (2008). Análise da epidemia da ferrugem do cafeeiro com árvore de decisão. *Tropical Plant Pathology*, *33*(2), 114–124. https://doi.org/10.1590/S1982-56762008000200005

MKTEsportivo. (2020). *Como as casas de apostas determinam as odds*. MKTEsportivo. https://www.mktesportivo.com/2020/09/como-as-casas-de-apostas-determinam-as-odds/

MKTEsportivo. (2022). *Clubes aprovam investimento de € 1.5 bilhão da CVC na Ligue 1*. MKTEsportivo. https://www.mktesportivo.com/2022/03/clubes-aprovam-investimento-de-e-1-5-bilhao-da-cvc-na-ligue-1/

Müller. (2020). *História de LaLiga : a criação e seu primeiro campeão Regulamento pronto As correntes A criação oficial de LaLiga*. Minha Torcida. https://www.minhatorcida.com.br/geral/8522-historia-de-laliga-a-criacao-e-seu-primeiro-campeao

Murray, B., & Murray, W. J. (1998). *The world's game: a history of soccer*. University of Illinois Press. https://books.google.com.br/books?hl=pt-BR&lr=&id=CRkIAcCB_0EC&oi=fnd&pg=PR11&dq=soccer+history&ots=YeqyBzKeFA&sig=2_wIXHJvwa20tJKs7yd4AucnroY#v=onepage&q=soccer history&f=false

Napoleão, B. M. (2018). *Árvores Decisórias*. Ferramentas Da Qualidade. https://ferramentasdaqualidade.org/arvores-decisorias/

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565

Olmeda, A. P. (2011). El juego on line. In *News.Ge* (1st ed.). Aranzadi. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwiMiemBt8r6AhX-O7kGHQOyCN8QFnoECBcQAQ&url=https%3A%2F%2Fwww.grupocodere.com%2Fsala-de-prensa%2Fdocumentacion%2Fdocumentacion-sector%2Fel-juego-online%2F&usg=AOvVaw35UncgKXgJsJxbF3SRm

Rejane, S. (2021). *Apostas esportivas no Brasil : Entenda como funciona*. JUS.com.br. https://jus.com.br/artigos/92170/apostas-esportivas-no-brasil-entenda-como-funciona

Sacramento, G. (2021). *Naive Bayes: como funciona esse algoritmo de classificação*. Somostera. https://blog.somostera.com/data-science/naive-bayes

SALVARO, R. D. F. (2016). *PERSPECTIVAS DE TRIBUTAÇÃO COM A LEGALIZAÇÃO DAS APOSTAS ESPORTIVAS NO BRASIL* [Universidade do Extremo Sul Catarinense]. http://repositorio.unesc.net/handle/1/7442

Santos, E. M. dos. (2002). *Teoria e Aplicação de Support Vector Machines à Aprendizagem e Reconhecimento de Objetos Baseado na Aparência* [Universidade Federal da Paraíba]. http://docs.computacao.ufcg.edu.br/posgraduacao/dissertacoes/2002/Dissertacao_EulandaMirandadosSantos.pdf

Santos, J. M. A. dos. (2019). *Previsões de Resultados em Partidas do Campeonato Brasileiro de Futebol* [FGV]. https://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/27672/joao_marcos_amorim_dos_santos.pdf

Santos, G. C. (2020). *Algoritmos De Machine Learning Para Previsão De Ações Da B3* [Universidade Federal de Uberlândia]. https://repositorio.ufu.br/bitstream/123456789/29897/7/AlgoritmosMachineLearning.pdf

Sarmento, C. E. (2006). *A regra do jogo : uma história institucional da CBF*. https://bibliotecadigital.fgv.br/dspace/handle/10438/6703

Schein, A. I., & Ungar, L. H. (2007). Active learning for logistic regression: An evaluation. In *Machine Learning* (Vol. 68, Issue 3). https://doi.org/10.1007/s10994-007-5019-5

Schlembach, C., Schmidt, S. L., Schreyer, D., & Wunderlich, L. (2022). Forecasting the Olympic medal distribution – A socioeconomic machine learning model. *Technological Forecasting and Social Change*, *175*(November 2021), 121314. https://doi.org/10.1016/j.techfore.2021.121314

Soares, I. de C. (2019). *REGULAÇÃO E TRIBUTAÇÃO DE APOSTAS ESPORTIVAS NO BRASIL: LEI 13.756/18 E A COMPATIBILIDADE COM O ORDENAMENTO JURÍDICO BRASILEIRO* [Universidade Federal da Paraíba]. https://repositorio.ufpb.br/jspui/handle/123456789/16211

Somboonphokkaphan, A., Phimoltares, S., & Lursinsap, C. (2009). Tennis Winner

Prediction based on Time-Series History with Neural Modeling. *International Multi-Conference of Engineers and Computer Scientists*, *I*, 127–132. https://www.semanticscholar.org/paper/Tennis-Winner-Prediction-based-on-Time-Series-with-Somboonphokkaphan-Phimoltares/fd3dc6d565e6a0e90d3df5b016df214dca39d9ba

Tax, N., & Joustra, Y. (2015). Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach. *Transactions On Knowledge And Data Engineering*, *10*(10), 15. https://doi.org/10.13140/RG.2.1.1383.4729

Thabtah, F., Zhang, L., & Abdelhamid, N. (2019). NBA Game Result Prediction Using Feature Analysis and Machine Learning. *Annals of Data Science*, *6*(1), 103–116. https://doi.org/10.1007/s40745-018-00189-x

Thompson, W. N. (2010). The International Encyclopedia of Gambling. In *Reference & User Services Quarterly* (1st ed., Vol. 50, Issue 1). ABC-CLIO.

TNT Sports. (2021). *Ligue 1 define redução de clubes na primeira divisão francesa a partir da temporada 2023/24*. TNT Sports. https://tntsports.com.br/melhorfuteboldomundo/Ligue-1-define-reducao-de-clubes-na-primeira-divisao-francesa-a-partir-da-temporada-202324-20211013-0022.html

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*. https://link.springer.com/book/10.1007/978-1-4757-3264-1

Wheatcroft, E. (2020). Profiting from overreaction in soccer betting odds. *Journal of Quantitative Analysis in Sports*, *16*(3), 193–209. https://doi.org/10.1515/jqas-2019-0009

Wheatcroft, E., & Sienkiewicz, E. (2021). Calibration and hyperparameter tuning in football forecasting with Machine Learning. *Mathsport 2021*, *June*, 1–6. https://www.researchgate.net/publication/352846677_Calibration_and_hyperparameter_tuning_in_football_forecasting_with_Machine_Learning

Wilkens, S. (2021). Sports prediction and betting models in the machine learning age: The case of tennis. *Journal of Sports Analytics*, *7*(2), 99–117. https://doi.org/10.3233/jsa-200463

Wiseman, O. (2016). *Using Machine Learning to Predict the Winning Score of Professional Golf Events on the PGA Tour* [National College Of Ireland]. http://trap.ncirl.ie/2493/

APÊNDICE A – CÓDIGOS EM *PYTHON*

Os demais códigos estão disponibilizados no meu repositório do GitHub através do link (https://github.com/johnhenry291/TCC-Joao_Henrique).

In [1]:
```
# TCC: APLICAÇÃO DE MACHINE LEARNING PARA APOSTAS ESPORTIVAS: uso de Regressão Logística,
#      SVM, Árvore de Decisão e Naive Bayes
#
# Aluno: João Henrique
# Orientadora: Renata Alcoforado
```

In [2]:
```
# Importando as bibliotecas

import pandas as pd
import numpy as np
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score , f1_score, precision_score, recall_score
from sklearn.feature_selection import SelectKBest
from sklearn.model_selection import  GridSearchCV
from sklearn.preprocessing import StandardScaler, MinMaxScaler
from IPython.display import display
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import scale
%matplotlib inline
```

In [3]:
```
# O arquivo utilizado foi uma compilação com dados baixados do site www.football-data.co.uk, que disponibiliza.
# ... informações de diversos campeonatos ao redor do mundo
# Por meio do Excel foram retirados os jogos em que estavam faltando dados de quaisquer variáveis de interesse.
# ...para o estudo


# Legenda das variáveis:

# Div = Divisão da Liga
# Date = Data do jogo
# HomeTeam = Time da casa
# AwayTeam = Time visitante
# FTR = Resultado do jogo (0 = empate, 1 = vitória do time da casa, 2 = vitória do time visitante)
# B365H = Odd (probablidade) de vitória do time da casa pela Bet365 (casa de apostas)
# B365D = Odd (probablidade) de empate pela Bet365 (casa de apostas)
# B365A = Odd (probablidade) de vitória do time visitante pela Bet365 (casa de apostas)
# BWH = Odd (probablidade) de vitória do time da casa pela Bet&Win (casa de apostas)
# BWD = Odd (probablidade) de empate pela Bet&Win (casa de apostas)
# BWA = Odd (probablidade) de vitória do time visitante pela Bet&Win (casa de apostas)
# IWH = Odd (probablidade) de vitória do time da casa pela Interwetten (casa de apostas)
# IWD = Odd (probablidade) de empate pela Interwetten (casa de apostas)
# IWA = Odd (probablidade) de vitória do time visitante pela Interwetten (casa de apostas)
# PSH = Odd (probablidade) de vitória do time da casa pela Pinnacle (casa de apostas)
# PSD = Odd (probablidade) de empate pela Pinnacle (casa de apostas)
# PSA = Odd (probablidade) de vitória do time visitante pela Pinnacle (casa de apostas)
# VCH = Odd (probablidade) de vitória do time da casa pela VC Bet (casa de apostas)
# VCD = Odd (probablidade) de empate pela VC Bet (casa de apostas)
# VCA = Odd (probablidade) de vitória do time visitante pela VC Bet (casa de apostas)
# WHH = Odd (probablidade) de vitória do time da casa pela William Hill (casa de apostas)
# WHD = Odd (probablidade) de empate pela William Hill (casa de apostas)
# WHA = Odd (probablidade) de vitória do time visitante pela William Hill (casa de apostas)

# Lendo o arquivo ENG5anos.csv
data = pd.read_csv('ENG5anos.csv', delimiter = ',')

# Visualizando a base de dados:
display(data.head())
```

	Div	Date	HomeTeam	AwayTeam	FTR	B365H	B365D	B365A	BWH	BWD	...	IWA	PSH	PSD	PSA	VCH	VCD	VCA	WHI
0	E0	11/08/2017	Arsenal	Leicester	H	1.53	4.5	6.50	1.50	4.60	...	6.50	1.53	4.55	6.85	1.53	4.50	6.50	1.5
1	E0	12/08/2017	Brighton	Man City	A	11.00	5.5	1.33	11.00	5.25	...	1.35	10.95	5.55	1.34	10.00	5.50	1.33	10.0
2	E0	12/08/2017	Chelsea	Burnley	A	1.25	6.5	15.00	1.22	6.50	...	13.50	1.26	6.30	15.25	1.25	6.25	15.00	1.2
3	E0	12/08/2017	Crystal Palace	Huddersfield	A	1.83	3.6	5.00	1.80	3.50	...	4.30	1.83	3.58	5.11	1.83	3.60	5.00	1.8
4	E0	12/08/2017	Everton	Stoke	H	1.70	3.8	5.75	1.70	3.60	...	5.00	1.70	3.83	5.81	1.70	3.80	5.75	1.7

5 rows × 23 columns

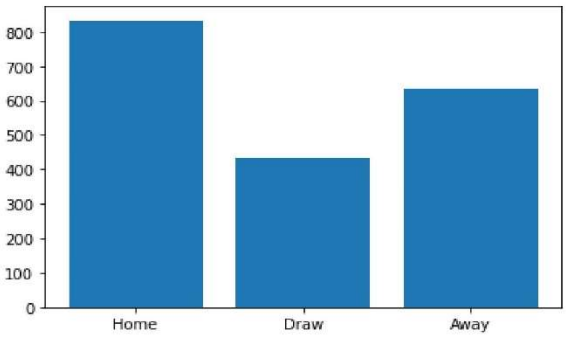
In [4]:
```
# Criando a coluna com o identificador do jogo (Game_id)

data = pd.DataFrame(data)
data = data.reset_index()
data = data.rename(columns = {"index": "Game_id"})
data["Game_id"] = data.index + 1


# Alterando os resultados de string (D = empate, H = vitória do time da casa, A = vitória do time visitante),
# para int (0 = empate, 1 = vitória do time da casa, 2 = vitória do time visitante)
for i in range(0,len(data)):
    if data["FTR"][i] == "D":
        data.at[i,"FTR"] = "0"
    if data["FTR"][i] == "H":
        data.at[i,"FTR"] = "1"
    if data["FTR"][i] == "A":
        data.at[i,"FTR"] = "2"

data["FTR"] = data["FTR"].astype(str).astype(int)

# Verificando as 5 primeiras linhas do arquivo após os primeiros ajustes
display(data.head())
```

	Game_id	Div	Date	HomeTeam	AwayTeam	FTR	B365H	B365D	B365A	BWH	...	IWA	PSH	PSD	PSA	VCH	VCD	VCA
0	1	E0	11/08/2017	Arsenal	Leicester	1	1.53	4.5	6.50	1.50	...	6.50	1.53	4.55	6.85	1.53	4.50	6.50
1	2	E0	12/08/2017	Brighton	Man City	2	11.00	5.5	1.33	11.00	...	1.35	10.95	5.55	1.34	10.00	5.50	1.33
2	3	E0	12/08/2017	Chelsea	Burnley	2	1.25	6.5	15.00	1.22	...	13.50	1.26	6.30	15.25	1.25	6.25	15.00
3	4	E0	12/08/2017	Crystal Palace	Huddersfield	2	1.83	3.6	5.00	1.80	...	4.30	1.83	3.58	5.11	1.83	3.60	5.00
4	5	E0	12/08/2017	Everton	Stoke	1	1.70	3.8	5.75	1.70	...	5.00	1.70	3.83	5.81	1.70	3.80	5.75

5 rows × 24 columns

In [5]:
```
# Explorando a base de dados de dos campeonatos de 2017/2018 a 2021/2022 (últimas 5 temporadas completas)

matches = data.shape[0]
features = data.shape[1]

home_win = len(data[data.FTR==1])
away_win = len(data[data.FTR==2])
draw = len(data[data.FTR==0])
val = [home_win, draw, away_win]

win_rate = (float(home_win)/(matches)) *100

print ('Total de jogos: ', matches)
print ('Total de colunas: ', features)
print ('Total de jogos ganhos em casa: ', home_win)
print ('Total de jogos ganhos pelo visitante: ', away_win)
print ('Total de jogos empatados: ', draw)
print ('Percentual de jogos ganhos em casa: {:.2f}%'.format( win_rate ))

x = np.arange(3)
plt.bar(x, val)
plt.xticks(x, ('Home', 'Draw', 'Away'))
plt.show()
```

```
Total de jogos:  1900
Total de colunas:  24
Total de jogos ganhos em casa:  833
Total de jogos ganhos pelo visitante:  634
Total de jogos empatados:  433
Percentual de jogos ganhos em casa: 43.84%
```



In [6]:
```
# Separando as features e as labels

features = data[['B365H', 'B365D', 'B365A', 'BWH', 'BWD', 'BWA', 'IWH', 'IWD', 'IWA', 'PSH', 'PSD',
                 'PSA', 'VCH', 'VCD', 'VCA', 'WHH', 'WHD', 'WHA']]
labels = data['FTR']

print('Features')
print(features.head())
print('====================')
print('Labels')
print(labels.head())
```

```
Features
   B365H  B365D  B365A    BWH   BWD    BWA   IWH  IWD    IWA    PSH   PSD  \
0   1.53    4.5   6.50   1.50  4.60   6.75  1.47  4.5   6.50   1.53  4.55
1  11.00    5.5   1.33  11.00  5.25   1.30  8.00  5.3   1.35  10.95  5.55
2   1.25    6.5  15.00   1.22  6.50  12.50  1.22  6.2  13.50   1.26  6.30
3   1.83    3.6   5.00   1.80  3.50   4.75  1.85  3.5   4.30   1.83  3.58
4   1.70    3.8   5.75   1.70  3.60   5.50  1.70  3.7   5.00   1.70  3.83

     PSA    VCH   VCD    VCA    WHH  WHD    WHA
0   6.85   1.53  4.50   6.50   1.53  4.2   6.00
1   1.34  10.00  5.50   1.33  10.00  4.8   1.33
2  15.25   1.25  6.25  15.00   1.25  5.5  13.00
3   5.11   1.83  3.60   5.00   1.80  3.3   5.00
4   5.81   1.70  3.80   5.75   1.70  3.5   5.50
====================
Labels
0    1
1    2
2    2
3    2
4    1
Name: FTR, dtype: int32
```

In [7]:
```
# Normalizando os dados de entrada (features)

scaler = MinMaxScaler().fit(features)
features_scale = scaler.transform(features)

print('Features: ',features_scale.shape)
print(features_scale)
```

```
Features:  (1900, 18)
[[0.02142206 0.11971831 0.13490471 ... 0.02406015 0.10743802 0.1087344 ]
 [0.45305378 0.19014085 0.0052658  ... 0.44862155 0.15702479 0.00467914]
 [0.00865998 0.26056338 0.34804413 ... 0.01002506 0.21487603 0.26470588]
 ...
 [0.00364631 0.40140845 0.34804413 ... 0.00501253 0.33884298 0.39839572]
 [0.00455789 0.36619718 0.39819458 ... 0.00501253 0.33884298 0.39839572]
 [0.38468551 0.26056338 0.00325978 ... 0.44862155 0.25619835 0.00289661]]
```

In [8]:
```
# Separação de treino e teste, feita manualmente para manter a ordem cronológica, uma vez que...
# ... temos informação temporal.
# Treino [:1520]     representa 80% dos dados
# Teste [1520:1710]      representa 10% dos dados
# Previsão [1710:1900]     representa 10% dos dados


X_train = features_scale[:1520]
X_test = features_scale[1520:1710]
y_train = labels[:1520]
y_test = labels[1520:1710]

print(len(X_train), len(y_train))
print(len(X_test), len(y_test))
```

```
1520 1520
190 190
```

In [9]:
```
# Rodando o modelo de Regressão Logística em sua versão multinomial

clf_LR = LogisticRegression(multi_class = 'multinomial', max_iter = 2000)
clf_LR.fit(X_train, y_train)
pred = clf_LR.predict(X_test)

lr_acc = accuracy_score(y_test, pred)
f1 = f1_score(y_test, pred, average = 'micro')

print('Regressão Logística')
print('')
print('Acurácia: {:.2f}%'.format(lr_acc*100))
print('F1 Score: {:.2f}%'.format(f1*100))
```

```
Regressão Logística

Acurácia: 56.32%
F1 Score: 56.32%
```

In [10]:
```
# Rodando o modelo de Support Vector Machine (SVM)

clf_SVM = SVC()
clf_SVM.fit(X_train, y_train)
pred = clf_SVM.predict(X_test)

svm_acc = accuracy_score(y_test, pred)
f1 = f1_score(y_test, pred, average = 'micro')

print('SVM')
print('')
print('Acurácia: {:.2f}%'.format(svm_acc*100))
print('F1 Score: {:.2f}%'.format(f1*100))
```

```
SVM

Acurácia: 56.32%
F1 Score: 56.32%
```

In [11]:
```
# Rodando o modelo de Árvore de Decisão

clf_DT = DecisionTreeClassifier(random_state=42)
clf_DT.fit(X_train, y_train)
pred = clf_DT.predict(X_test)

dt_acc = accuracy_score(y_test, pred)
f1 = f1_score(y_test, pred, average = 'macro')

print('Árvore de Decisão')
print('')
print('Acurácia: {:.2f}%'.format(dt_acc*100))
print('F1 Score: {:.2f}%'.format(f1*100))
```

```
Árvore de Decisão

Acurácia: 43.68%
F1 Score: 42.29%
```

In [12]:
```
# Rodando o modelo de Naive Bayes

clf_NB = GaussianNB()
clf_NB.fit(X_train, y_train)
pred = clf_NB.predict(X_test)

nb_acc = accuracy_score(y_test, pred)
f1 = f1_score(y_test, pred, average = 'micro')

print('Naive Bayes')
print('')
print('Acurácia Naive Bayes: {:.2f}%'.format(nb_acc*100))
print('F1 Score: {:.2f}%'.format(f1*100))
```

```
Naive Bayes

Acurácia Naive Bayes: 47.89%
F1 Score: 47.89%
```

In [13]:
```
#Executando a previsão pelo método de maior acurácia (Regressão Logística)

previsao = features_scale[1710:]

game_id_full = data['Game_id']
game_id_prev = game_id_full[1710:]

res_full = data['FTR']
res_prev = res_full[1710:]

pred = clf_LR.predict(previsao)

df = pd.DataFrame({'Real': res_prev, 'Previsão': pred, 'Game_id': game_id_prev})

print(df)
```

```
      Real  Previsão  Game_id
1710     2         1     1711
1711     1         1     1712
1712     1         1     1713
1713     0         1     1714
1714     1         1     1715
...    ...       ...      ...
1895     1         2     1896
1896     1         1     1897
1897     1         1     1898
1898     1         1     1899
1899     2         2     1900

[190 rows x 3 columns]
```

In [14]:
```
#confusion Matrix

df = pd.DataFrame(df,columns = ['Real','Previsão'])

cf_matrix = pd.crosstab(df['Real'], df['Previsão'], rownames = ['Real'] , colnames = ['Previsão'])

sns.heatmap(cf_matrix, annot = True, cmap = 'Blues', fmt = 'g')
```

Out[14]:
```
<AxesSubplot:xlabel='Previsão', ylabel='Real'>
```