UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE COMPUTAÇÃO
CURSO DE CIÊNCIA DA COMPUTAÇÃO

GABRIEL AUGUSTO DE DEUS

UTILIZAÇÃO DE APRENDIZADO DE MÁQUINA PARA PREVISÃO DE RESULTADOS DE JOGOS DE FUTEBOL

TRABALHO DE CONCLUSÃO DE CURSO

MEDIANEIRA

2019

GABRIEL AUGUSTO DE DEUS

UTILIZAÇÃO DE APRENDIZADO DE MÁQUINA PARA PREVISÃO DE RESULTADOS DE JOGOS DE FUTEBOL

Trabalho de Conclusão de Curso apresentado ao Departamento Acadêmico de Computação da Universidade Tecnológica Federal do Paraná como requisito parcial para obtenção do título de "Bacharel em Computação".

Orientador: Prof. Dr. Arnaldo Candido Junior

Co-orientador: Prof. Dr. Alan Gavioli

MEDIANEIRA

2019

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Diretoria de Graduação e Educação Profissional
Coordenação do Curso de Ciência da Computação

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

TERMO DE APROVAÇÃO

UTILIZAÇÃO DE APRENDIZADO DE MÁQUINA PARA PREVISÃO DE RESULTADOS DE JOGOS DE FUTEBOL

Por

GABRIEL AUGUSTO DE DEUS

Este Trabalho de Conclusão de Curso foi apresentado às 10:20h do dia 11 de novembro 2019 como requisito parcial para a obtenção do título de Bacharel no Curso de Ciência da Computação, da Universidade Tecnológica Federal do Paraná, Câmpus Medianeira. O candidato foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

Prof. Arnaldo Candido Junior
UTFPR - Câmpus Medianeira

Prof. Pedro Luiz De Paula Filho
UTFPR - Câmpus Medianeira

Prof. Jorge Aikes Junior
UTFPR - Câmpus Medianeira

A folha de aprovação assinada encontra-se na Coordenação do Curso.

RESUMO

De Deus, Gabriel Augusto. UTILIZAÇÃO DE APRENDIZADO DE MÁQUINA PARA PREVISÃO DE RESULTADOS DE JOGOS DE FUTEBOL. 55 f. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade Tecnológica Federal do Paraná. Medianeira, 2019.

Mineração de dados e Aprendizado de Máquina são duas tecnologias crescentes que estão sendo utilizadas em várias segmentos da sociedade, comercial, acadêmico, entre outros. Esse trabalho une essas duas tecnologias a um esporte amplamente popular no Brasil, o futebol. O objetivo dessa pesquisa foi utilizar da Mineração de Dados e Aprendizado de Máquina para fazer a previsão de resultados de jogos e campeonatos de futebol. Para a realização do trabalho foram utilizados os princípios da mineração de dados, que são, obtenção dos dados, pré-processamento, a aplicação com os algoritmos de Aprendizado de Máquina, cujo os escolhidos foram, a Regressão Linear, o Bayesiano Ingênuo, Redes MLP e por último SVMs, esse algoritmos serão responsáveis pelos resultados de regressão e classificação. E por último a avaliação dos resultados, utilizando Teste T sobre Acurácia e Coeficiente de Correlação. Os resultados obtidos foram próximos a 60% no caso de classificadores e um coeficiente de correlação de aproximadamente 0,55 nos regressores. E com a execução dos testes foi possível ver leve vantagem das SVMs e MLPs como classificadores, e um empate técnico geral nos regressores.

Palavras-chave: mineração de dados, esportes, futebol, aprendizado de máquina

ABSTRACT

De Deus, Gabriel Augusto. USE OF MACHINE LEARNING TO PREDICT SOCCER MATCH RESULTS. 55 f. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade Tecnológica Federal do Paraná. Medianeira, 2019.

Data Mining and Machine Learning are two growing technologies that are being used in various segments of society, including commercial and academic application, among others. This work makes combines these two technologies and an widely popular sport in Brazil, soccer (or football). The purpose of this research is to use Data Mining and Machine Learning to forecast the results of soccer games and championships. In order to perform the work, were used the principles of data mining, data acquisition, preprocessing, the application with machine learning algorithmswhich were chosen, Linear Regression, Naive Bayes, MLP Networks and finally SVMs, these algorithms will be responsible for the results of regression and classification. And finally the evaluation of the results, using T-Test over Accuracy and Correlation Coeficient. The results obtained were close to a case of classifiers of 60% and a correlation coefficient of approximately 0.55 in regressors. And by running the tests it was possible to get a slight advantage from SVMs and MLPs as classifiers, and a general technical draw on the regressors.

Keywords: data mining, sports, soccer, machine learning

AGRADECIMENTOS

Agradeço em primeiro lugar a meu pai e minha mãe por toda ajuda e suporte e apoio durante a confecção desse trabalho. Agradeço a meu irmão que durante toda essa jornada me apoiou. Agradeço a minha namorada Ana Caroline que além de muita ajuda me deu muita força para que eu pudesse avançar nesse trabalho.

Agradeço muito ao meu orientador Professor Dr. Arnaldo Candido Júnior por toda ajuda não só nesse trabalho mas também durante toda o período na graduação. Agradeço ao professor Jorge Aikes por todo apoio durante a disciplina de TCC.

Agradeço a todos os colegas de universidade que fiz nessa jornada incrível que vou guardar comigo para todo sempre.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

AM	Aprendizado de Máquina
KDD	Knowledge Discovery in Databases
MLP	Multilayer Perceptron
SVM	*Suport Vector Machines*

SUMÁRIO

1 INTRODUÇÃO

Com o crescimento exponencial dos hardwares no início dos anos 90 até atualmente (2019), a computação passou a superar vários obstáculos, sendo um deles o armazenamento de dados em massa. Nos dias atuais, é comum a existência de grandes bancos de dados, armazenados por vários anos. Com o armazenamento de dados em massa, passou-se a trabalhar com uma nova ciência, a mineração de dados. Segundo Fayyad et al. (1996), "mineração de dados é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados". A mineração de dados já vem sendo utilizada nos esportes há alguns anos, para previsão de vencedores, como no trabalho Graettinger e Graettinger (2014).

Outra área que também tem se destacado na predição de resultados de eventos esportivos é o Aprendizado de Máquina (AM), definido como a capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência (MITCHELL, 1997).

Esta pesquisa trata especificamente da área esportiva do futebol. Segundo a Federação Internacional do Futebol (Fédération Internationale de Football Association – FIFA), o futebol é o esporte mais assistido e mais jogado no mundo, com cerca de 3,5 bilhões de fãs e mais de 260 milhões de profissionais envolvidos, levando em consideração jogadores, técnicos entre outros (FIFA, 2007).

A mineração de dados e o aprendizado de máquina têm se expandido cada vez mais e em várias áreas. Nos esportes, essas duas tecnologias aliadas, têm servido para aumento de desempenho de atletas (MIN et al., 2008), e também para previsão de resultados, mesmo sendo o futebol um esporte com grande número de variáveis externas envolvidas (LOUZADA et al., 2011).

1.1 OBJETIVOS GERAL E ESPECÍFICOS

Esse trabalho tem como objetivo geral utilizar mineração de dados e aprendizado de máquina, com bases de dados (*data sets*) de resultados e estatísticas esportivas para prever o resultado de jogos e times vencedores de campeonatos.

Os objetivos específicos são:

- Selecionar e pré-processar um base de dados e estatísticas de partidas de futebol;
- Treinar algoritmos de aprendizado de máquina para predizer os resultados das partidas;
- Comparar os modelos de aprendizado de máquina desenvolvidos.

1.2 JUSTIFICATIVA

O futebol é uma atividade que movimenta grandes quantias de dinheiro. Segundo a revista britânica Deloitte, o clube espanhol Real Madrid gerou mais de 750 milhões de Euros em receita na temporada de 2018/2019 (Sports Deloitte, 2019). O planejamento dos clubes, o quanto vão gastar na temporada, quantos jogadores contratar e outras decisões são tomadas em torno do desempenho da equipe na temporada, por isso utilizar ferramentas que ajudam a prever o desempenho de um clube na temporada também ajudaria a diretoria a tomar melhores decisões.

Desde de 1930, ano da primeira Copa do Mundo realizada no Uruguai, já existem relatos e imagens de estádios com lotação máxima. Porém, agora existem várias competições de grande porte e de apelo internacional. Essa popularidade do futebol tem grandes efeitos nas cidades e nas regiões que recebem os eventos. Isso gera um aumento no número de turistas, no policiamento, melhora na logística de transporte da cidade entre outros elementos que precisam ser preparados para receber os fãs e turistas.

Um ótimo exemplo em que a previsão de resultados ter ajudado é a final da *UEFA Europa League* 2019 realizada em Baku, capital do Afeganistão. A final foi disputada entre dois times o Arsenal e o Chelsea ambos da capital inglesa, Londres. Baku ter sido a sede do jogo trouxe dois problemas, o primeiro deles é o fato de um jogador do Arsenal ser Armênico, e como o Afeganistão e a Armênia tem relações diplomáticas complicadas o jogador não pode

viajar para a disputa da final. O segundo problema é que não existiam voos diretos de Londres para Baku dificultando muito a logística dos torcedores fazendo com que nem todos os ingressos fossem vendidos e sobraram lugares na final (Folha de SP, 2019).

Outra área de destaque são as apostas esportivas. Segundo Louzada et al. (2011) estima-se que o mercado de apostas esportivas no Brasil faturou entre 1,8 e 4 bilhões por ano sem qualquer controle do governo brasileiro. O uso da tecnologia ajuda tanto às casas de apostas a embasar os valores ofertados, quanto aos apostadores a decidirem a maneira de efetuarem suas apostas.

Os esportes sempre foram populares no mundo todo, existem vários registros históricos que demonstram o interesse da população por esse tipo de atração na Roma Antiga, Grécia Antiga e até alguns registros com Incas e Astecas. Com a profissionalização dos esportes, os dados passaram a ser armazenados, porém no princípio esses dados eram apenas armazenados e apresentados, não utilizados em tomada de decisão. Em um novo cenário, treinadores e olheiros passaram a olhar esses dados de um nova forma, tomando decisões baseadas nos dados disponíveis, como contratar um jogador baseado em suas estatísticas. Porém, apesar de visualizar os dados, essas decisões tomadas por treinador e olheiros ainda eram mais baseadas em intuição do que em métodos científicos. Mais recentemente, em um terceiro cenário, inicia-se a análise desses dados com algoritmos, o que auxilia na tomada de decisão para escolher a melhor jogada, ou ver se um jogador se encaixa nas características de um time (SCHUMAKER, 2010).

1.3 ORGANIZAÇÃO DO DOCUMENTO

Nos próximos capítulos desse documento, o Capítulo 2 - Referencial Teórico, abordará sobre o que será utilizado para execução de trabalho. Em seguida no Capítulo 3 - Materiais e Métodos, será trabalhado como serão obtidos os materiais e os métodos utilizados para se obter conhecimento através desses dados. O Capítulo 4 - Experimentos discutirá o resultado dos experimentos e seus passos. O Capítulo 5 trará as conclusões.

2 REFERENCIAL TEÓRICO

Nessa seção serão descritas as tecnologias que serão utilizadas para o desenvolvimento dessa pesquisa. Uma introdução sobre mineração de dados e aprendizado de máquina, sobre o esporte, em especial o futebol, e os algoritmos que serão utilizados na tentativa de prever resultados de jogos e eventos esportivos.

2.1 ESPORTE

O esporte no contexto histórico é divido em três partes: Esporte Antigo, Esporte Moderno e Esporte Contemporâneo (KRAVCHYCHYN et al., 2012).

Existem duas versões para o surgimento do esporte, uma que cita que surgiu com caráter educacional, outra, já fala que desde os primeiros tempos o esporte surgiu com caráter biológico. Antes do esporte surgir, as atividades físicas eram praticadas com fim de sobrevivência, como caçar e plantar, entre outros. Com a evolução e alguns povos deixando de ser nômades, começaram então a prática de atividades físicas com fim militar, tanto para atacar quanto para se defender. Os historiadores indicam que o primeiro local a começar a utilizar atividades físicas de forma educativa foi Atenas na Grécia. Com isso surgiram os jogos gregos, que mais tarde inspiraram os jogos olímpicos. Os jogos gregos são considerados importantíssimos pois são considerado a concepção do esporte (TUBINO, 1993).

Existem muitos relatos de competições esportivas nas antiguidade como em 1830 a.C uma competição de arremessos na Irlanda, na Ilha de Creta em 1500 a.C competições de boxe, em 1200 a.C o poeta grego Homero relatava os Jogos Fúnebres em seus poemas (DUARTE, 2019).

Os jogos gregos em 776 a.C posteriormente se tornariam o evento hoje mundialmente conhecido como Jogos Olímpicos. Em 500 a.C começaram a se der valores em dracmas de ouro e especiarias para os vencedores das competições. Já na Ásia o surgimento do polo hípico

praticado pelos Persas em 651 d.C e na China o *wan-chin* (esporte muito próximo ao golfe) e o sumô em 754 d.C (DUARTE, 2019).

O contexto de esporte moderno foi introduzido por Thomas Arnold. Ele acreditava no esporte como meio educacional, formado por três pilares, um jogo, uma competição e uma formação. Baseado nesses três pilares ele introduziu o esporte no currículo no colégio em que dirigia na Inglaterra (1828 - 1841). Como Thomas Arnold acreditava no esporte como uma evolução, ele deu a liberdade para que os alunos criassem os próprios formatos e regras, assim evoluiriam o jogo até sua forma mais justa de competição. O sucesso foi tão grande que não demorou para que o esporte rompesse as barreiras do colégio e começasse a se espalhar por toda a Inglaterra. Com o crescimento expansivo dos esportes, logo se viu a necessidade da criação de clubes, ligas e federações para ditar o futuro dos esportes, adaptar regras entre outras responsabilidades (TUBINO, 1993).

O esporte contemporâneo surgiu após pessoas e principalmente intelectuais da época se mostrarem extremamente preocupados com os rumos que os esportes e as competições estavam tomando. O esporte era visto apenas na ótica do rendimento, não havia outro contexto como o social e educacional por exemplo. Com isso um grupo de intelectuais da época entre as décadas de 50 e 80, faziam manifestos para que o esporte fosse enxergado de uma nova ótica. Inspirados nisso esse grupo de intelectuais, entre eles *Philip Noel-Becker*, ganhador do Prêmio Nobel da Paz em 1959, assinaram um manifesto que ficou conhecido como "Manifesto do Desporto". Esse manifesto reconhecia a existência de várias formas de se praticar esportes, como o contexto do homem comum e do educacional. Porém, afirma-se que o esporte passou a ser considerado contemporâneo em 1978 quando a Unesco publicou uma carta com o título "Carta Internacional da Educação Física e Esporte", em que reconhecia que o esporte era um direito de todo o cidadão, com forte caráter social e educacional.

Com o passar do tempo, foi percebida a capacidade comercial do esporte. Era visível já nas primeiras décadas das disputas como as pessoas criavam empatia por times, seleções e jogadores, que mais tarde virariam ídolos (TUBINO, 1993). Um ótimo exemplo disso é a Copa do Mundo de 1930 realizada no Uruguai, a final por exemplo, teve um público de quase 70 mil pessoas (LISI, 2007). Isso aliado com o crescimento dos meios de comunicação de massa, primeiramente o rádio e mais tarde a televisão, passou-se a explorar muito a capacidade comercial do esporte com patrocínios, venda de ingressos, camisas, entre outros (TUBINO, 1993).

Na Figura 1, o gráfico mostra no eixo horizontal o evento esportivo, e no vertical o número de pessoas que assistiram o evento em milhões de pessoas no ano de 2014. A final da Copa Do Mundo FIFA com mais de 700 milhões de espectadores, a final da Eurocopa com

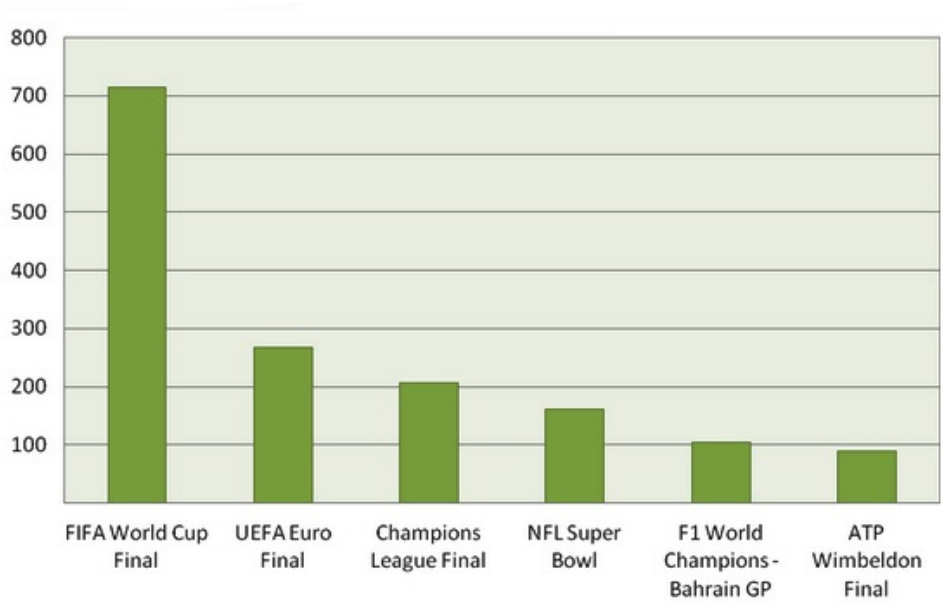

Figura 1 – Audiência dos eventos esportivos em 2014.
Fonte: (REPUCOM, 2014)

mais de 250 milhões e a final da Liga dos Campeões da Europa com mais de 200 milhões. Esses eventos possuem em comum o fato de serem finais de eventos de futebol. Logo, é conclusivo que o futebol é um esporte muito popular e o mais assistido do mundo.

Um ótimo exemplo do alto faturamento dos eventos esportivos é a final da NFL (*National Football Association*), a liga profissional de Futebol Americano, esse evento é conhecido no mundo todo com *Super Bowl*. O segundo do comercial nos intervalos do evento chega a custar 640 mil reais (SUTTO, 2019). A Copa do Mundo FIFA de 2014, realizada no Brasil, teve um faturamento de 4,83 bilhões de dólares, sendo 2,22 bilhões de dólares em despesas. Ou seja um lucro de 2,6 bilhões de dólares, sendo que a venda de direitos televisivos totalizou 2,43 bilhões de dólares (Forbes, 2019).

2.2 FUTEBOL

Os historiadores que estudam e pesquisam sobre o futebol nunca chegaram a um consenso sobre a origem do esporte. Há relatos de um jogo na China onde exércitos que venciam batalhas, tinham como costume chutar a cabeça de seus inimigos e tinham que levar a cabeça

até duas estacas fincadas no chão (SOUZA, 2007). Existem muitos relatos de jogos parecidoS com futebol, como o *Kemari* no Japão, o *Episkiros* na Grécia. Porém, o que obteve destaque foi *gioco del calcio* criado na Itália, mas como qualquer tipo de contato físico era permitido o jogo se tornou muito violento, o que levou jogo a ser proibido pelo Rei Eduardo II. Mas o esporte voltou a ser praticado pela nobreza italiana, agora com regras e conotação menos violenta. O *gioco del calcio* chegou a Inglaterra, lá foram implementadas muitas mudanças e novas regras deixando o jogo mais próximo do que é hoje. Foram introduzidos elementos como as traves e o goleiro e fixado regras como o tempo de jogo e o tamanho do campo. Assim o esporte se aproximava muito do que é conhecido hoje como futebol (SANTOS, 2006). O futebol cresceu muito, e teve um expansão muito rápida para época, e para a organização do esporte, atualização das regras, organização de campeonatos e etc. Em 1904 foi criada Federação Internacional do Futebol e Associação, conhecida mundialmente como FIFA (SANTOS, 2006).

A FIFA[1], é a entidade máxima do futebol. Organiza os principais torneios, em várias categorias, desde a idade livre, até competições com limite de idade, como Copa do Mundo sub-20, sejam elas masculinas ou femininas. Para organização de outros torneios, de clubes e seleções, a FIFA tem suas federações filiadas, são elas:

- AFC, Confederação Asiática de Futebol;
- CAF, Confederação Africana de Futebol;
- CONCACAF, Confederação da América do Norte e Central de Futebol;
- CONMEBOL, Confederação Sul-americana de Futebol;
- UEFA, confederação europeia de futebol;
- OFC, confederação da Oceania de futebol;

Essas confederações ajudam a FIFA na organização de campeonatos locais, como no caso da CONMEBOL que organiza a Libertadores (competição entre os melhores clubes da América do Sul), e a Copa América (competição entre as seleções de futebol da América do Sul).

Além disso, a FIFA ainda tem 211 associações afiliadas, que são as confederações dos países, como a CBF (Confederação Brasileira de Futebol) e a AFA (*Asociación del Fútbol Argentino*).

[1] www.fifa.com

2.3 MINERAÇÃO DE DADOS

A mineração de dados, é um processo que extrai conhecimento e padrões de grandes quantidades de dados (FAYYAD et al., 1996). É uma ciência que tem sido aplicada em diversas áreas, tanto com fins científicos e acadêmicos, como também comercial. Algumas aplicações de destaque utilizam mineração de dados:

- Eleições: a mineração de dados auxiliou o então candidato Barack Obama a vencer a corrida eleitoral americana em 2012 (LAROSE; LAROSE, 2014);
- Cartões de crédito: técnicas de mineração de dados são capaz de identificar anomalias com o objetivo, por exemplo, de detectar fraudes em cartão de crédito;
- Medicina: mineração de dados é utilizada na triagem de pacientes, para facilitar o processo de priorização de pacientes (MACIEL et al., 2015);
- Educação: Baker et al. (2011) apresentam usos da mineração de dados com o objetivo aprimorar a educação brasileira.

A mineração de dados é uma etapa no processo de Descoberta de Conhecimento em Bases de Dados (KDD - Knowledge Discovery in Databases). A mineração de dados é essencial para descoberta do conhecimento. Na Figura 2, é apresentada a demonstração das etapas KDD, o que vai desde a seleção dos dados até a obtenção do conhecimento (FAYYAD et al., 1996).

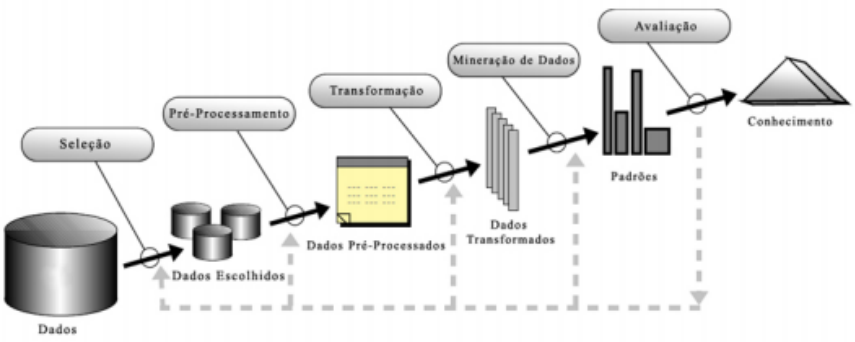

Figura 2 – Etapas da mineração de dados.
Fonte: (FAYYAD et al., 1996)

A primeira etapa é a seleção dos dados, analisando de onde eles serão obtidos, seja de bancos de dados existentes, de planilhas de uma empresa, ou qualquer tipo de armazenamento de dados. A próxima etapa é o pré-processamento, essa etapa visa padronizar os dados e buscar possíveis erros, como campos vazios ou com preenchimento incorreto. Após isso é

feita a transformação dos dados, os dados são colocados em um padrão em que a mineração de dados consiga trabalhar de forma efetiva. Em seguida é feita a mineração em si, em busca de associações, padrões e outros elementos, que por fim são interpretados e avaliados para então chegar ao conhecimento. (FAYYAD et al., 1996).

2.4 APRENDIZADO DE MÁQUINA

Aprendizado de máquina é definido como uma melhora da máquina na realização de uma tarefa com a experiência (MITCHELL, 1997). O aprendizado de máquina pode ocorrer de variadas formas:

- Simbólico: quando são construídas representações simbólicas de um conceito baseado em exemplos e contra exemplos;
- Estatístico: em uma base de dados estatística ocorre a tentativa de através desses dados se aproximar do modelo induzido;
- Baseado em exemplos: dados dois exemplos semelhantes, em que a classe de um deles é conhecida, pode-se usar essa informação para estimar a classe do outro;
- Conexionista: utiliza redes neurais e suas conexões para tomada de decisões.

Existem vários métodos e algoritmos de aprendizado de máquina, porém não é possível afirmar que um algoritmo seja melhor que outro, variados algoritmos se encaixam melhor em variados problemas assim sempre é necessário verificar o problema e executar o teste de vários algoritmos para ver qual é mais adequado (MONARD; BARANAUSKAS, 2003). Esta pesquisa foca em duas categorias do Aprendizado de Máquina: a regressão e classificação. A classificação tem seu resultado baseado em classes, isso é, uma nova entrada de dados e busca-se a qual classe ela pertence. Um exemplo de resultado de classificação é dizer se um atleta pratica basquete, futebol, entre outros, a partir de informações do atleta. Já a regressão é numérica, ela responde perguntas numéricas, do tipo, o valor de um produto, o valor de um jogador, número de pessoas esperadas para um evento, entre outros (MONARD; BARANAUSKAS, 2003).

2.5 ALGORITMOS DE APRENDIZADO DE MÁQUINA

Nesta sessão serão descritos os algoritmos que podem ser utilizados para previsão dos resultados e vencedores de campeonatos.

2.5.1 Regressão Linear

Segundo Camilo e Silva (2009), “as regressões são chamadas de lineares quando a relação entre as variáveis preditoras e a resposta segue um comportamento linear. Neste caso, é possível criar um modelo no qual o valor de y é uma função linear de x”. A regressão linear pode ter múltiplas variáveis. É composta por um vetor de coeficientes de regressão β, um vetor de observação x e um vetor de variáveis dependentes y, e um δ que é o cálculo do erro. (COELHO-BARROS et al., 2008), conforme a Equação 1.

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \delta \tag{1}$$

2.5.2 Bayesiano Ingênuo

O teorema de Bayes que inspirou o *Naive Bayes* foi criado por Thomas Bayes (1701 - 1761). A Equação 2 mostra a fórmula do teorema de Bayes, que tem as seguintes características.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2}$$

- P(A |B): é a probabilidade do evento A ocorrer dado que B ocorreu;
- p(B |A): é a probabilidade do evento B ocorrer dado A ocorreu;
- p(A): é a probabilidade do evento A ocorrer;
- p(B): é a probabilidade do evento B ocorrer;

Mesmo sendo um algoritmo simples, o Bayesiano Ingênuo tem se mostrado

extremamente efetivo em diversas tarefas, como classificação de texto (GAIGEROV et al., 1982) e diagnóstico médico (VIAENE et al., 2004). Uma característica desse algoritmo é, como seu próprio nome diz, *Naive*, em tradução literal ingênuo. Ele desconsidera a relação de dependência entre os atributos (MITCHELL, 1997).

Um problema de classificação contém a classe a ser comparada Y, e um vetor X com instâncias desse problema de classificação. Essas instâncias todas tem seus respectivos valores. Para o cálculo desse problema é utilizada a função *argmax*, essa função tem como objetivo não retornar a maior probabilidade, mas sim a qual classe pertence a maior probabilidade, como na Equação 3.

$$f = argmax P_{yj} \in Y(y_j|a_1, a_2 \wedge ... \wedge a_n) \tag{3}$$

Agora a equação acima será aplicada no teorema de bayes, ficando como na Equação 4.

$$f = argmax_{yi} \frac{P(a_1, a_2 ... a_n)P(y_j)}{P(a_1, a_2 ... a_n)} \tag{4}$$

Dado a Equação 4, pode-se verificar que o seu denominador é sempre constante para determinada instância, assim pode ser removido da notação como na Equação 5.

$$f = argmax P_{yj}(a_1 ... a_n|v_j)P(y_j) \tag{5}$$

Após isso pode-se aplicar a hipótese ingênua (Equação 6). A hipótese ingênua é a probabilidade de dois eventos independentes acontecerem.

$$P(A|B) = P(A)P(B) \tag{6}$$

Agora aplica-se a hipótese ingênua ao bayesiano ingênuo, ficando como na Equação 7.

$$f = argmax P_{yj} P(v_1|y_j) ... P(v_n|y_j)P(y_j) \tag{7}$$

Após analisar está equação é visível que pode ser colocado um produtório para simplificar a notação, como na Equação 8 (MITCHELL, 1997).

$$f = argmax \prod P(y_j)P(x_i|y_j) \tag{8}$$

Enquanto a regressão linear tem como resposta um score de aproximação a alguma classe, o Bayesiano Ingênuo permite dizer a classe a qual a entrada de dados pertence, se ganhou, perdeu ou empatou.

2.5.3 Neurônio Artificial e Perceptron

O neurônio artificial é baseado no neurônio biológico, que recebe sinais de entrada através das sinapses, ocorrendo entrada e saída e de íons. Na sequência um potencial de membrana dá a resposta sobre o sinal de entrada, e se irá ocorrer um sinal de saída. Sempre que esse potencial de membrana for superior a um determinado valor são liberadores neurotransmissores (ZUBEN, 2013). O peso representa o processo biológico que ocorre nas sinapses. O neurônio artificial é composto por peso, junção somadora e função de ativação.

Na Figura 3, um exemplo de neurônio artificial. O vetor $[x_1, x_2, ..., x_m]$ como entrada, em seguida o vetor $[w_k1, w_k2, ..., w_km]$ que corresponde ao peso das conexões, depois b_k que corresponde a junção somadora já ponderada por seus pesos, por último a função de ativação, e y_k como saída.

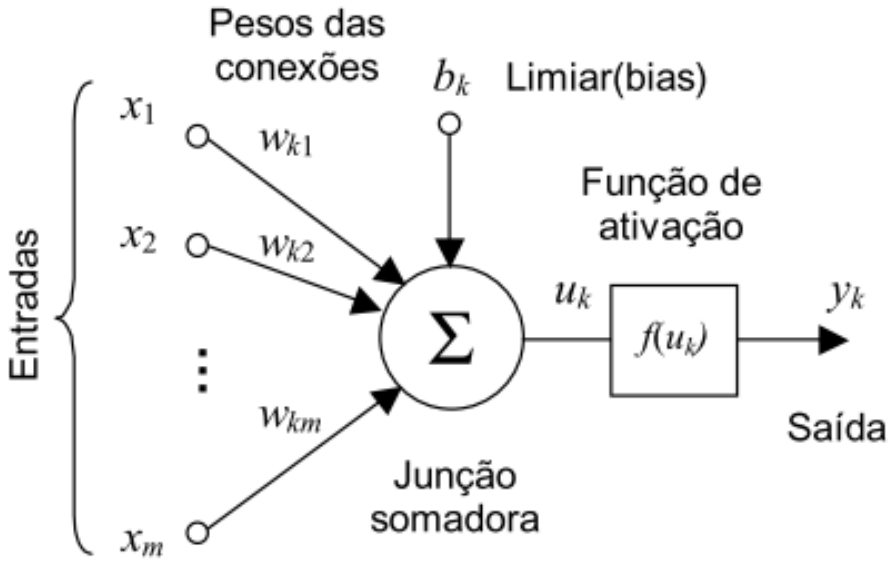

Figura 3 – Representação de um neurônio artificial.

Fonte: (ZUBEN, 2013)

Perceptron é uma rede neural de uma única camada. O Perceptron simples é composto por uma camada de entrada que recebe informações do exterior, uma última camada traz o resultado de acordo com o peso. O Perceptron simples só é capaz de resolver problemas linearmente separáveis de classes binárias. A principio, o Perceptron tem saída aleatória, mas, com o ajuste de pesos feito no treinamento, ele é treinado para fornecer saídas de acordo com os dados recebidos (BRAGA et al., 2007).

A função de ativação mais utilizada no Perceptron é o função passo utilizada como

classificador binário, quando é necessário dizer sim ou não como resposta de uma classe. O Perceptron só resolve problemas linearmente separáveis.

$$f(t) = u(t-a) \tag{9}$$

Na Equação 9, onde t é o potencial ativação de um neurônio ou não, e o a representa o bias do perceptron (MESQUITA; PIMENTA, 2012). A principal característica dessa função é quando a é maior do que zero, sua resposta é 1. E quando o o valor de a é negativo, sua resposta é nula ou seja 0, como é visível no gráfico da função na Figura 4 (RADTKE, 2008).

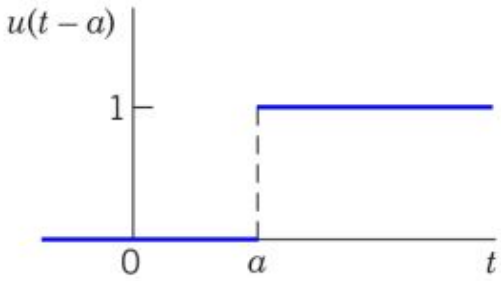

Figura 4 – Gráfico da função passo

.

Fonte: (RADTKE, 2008)

2.5.4 Rede Neural Multilayer Perceptron

Na Seção 2.5 foi afirmado que Perceptrons só resolvem problemas linearmente separáveis, para resolver problemas de funções matemáticas mais complexas, surgiram as redes MultiLayer Perceptron (MLP). Nas redes MLP, os neurônios são agrupados em camadas. Essa rede é composta por uma camada de entrada, zero ou mais camadas ocultas, e a camada de saída. Uma MLP com uma camada oculta pode aproximar de qualquer função contínua, com duas camadas ocultadas é possível a aproximação de qualquer função (BRAGA et al., 2007). As redes MLP podem ser do tipo *feedfoward* (em tradução literal sempre em frente) ou recorrentes. Em redes *feedforward*, o sinal nunca para de uma camada posterior a uma camada anterior, ao

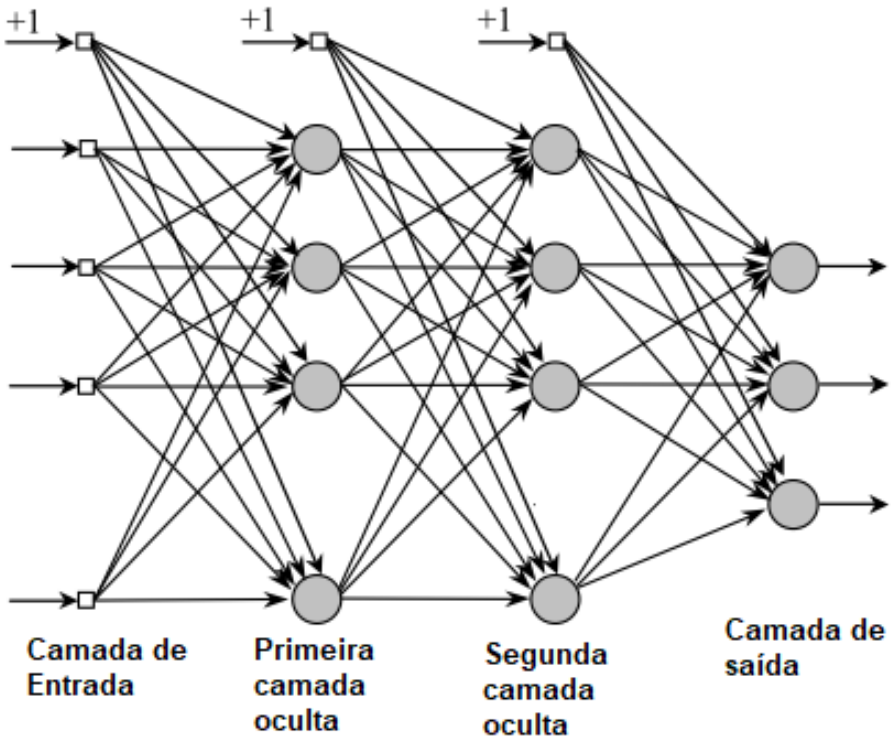

Figura 5 – Representação de uma MLP *feedforward*

.

Fonte: (ZUBEN, 2013)

contrário da rede recorrente.

Na Figura 5, a rede MLP segue o padrão *feedforward*, é possível ver a camada de entrada que recebe valores do conjunto de treinamento e em seguida as saídas dos neurônios sempre indo para uma camada adiante, nunca retornando ou indo para um neurônio de camada de mesmo nível (ZUBEN, 2013). Outro tipo de rede MLP muito utilizada são as redes recorrentes. Segundo Zuben (2013) "são as denominadas redes recorrentes, pois elas possuem, pelo menos, um laço realimentando da saída de neurônios para outros neurônios da rede".

Um item importante das redes MLP é a função de ativação, que foi introduzida nas redes MLP para resolver o problema do ajuste do peso e do *bias*, que antes das funções de ativação, um pequeno ajuste nos pesos e no *bias* de redes MLP poderiam causar mudanças drásticas na saída de rede. Com as funções de ativação, tem-se apenas uma pequena mudança no peso e no *bias* da rede. As funções de ativação têm como objetivo tomar a decisão se um neurônio deve ser ativado ou não. Elas usam o sinal de entrada e o aplica em uma função não linear. A saída dessa função serve como entrada para a próxima camada de neurônios. Existem vários tipos de funções de ativação, que devem ser utilizadas em diferentes situações (MITCHELL, 1997).

Em redes MLP as mais comuns são (ZUBEN, 2013):

- Linear: utilizada para fazer regressão;
- Sigmóide Logística: essa função é muito utilizável, já que não é linear e diferenciável, isso significa que quando uma MLP utiliza esse tipo de função sua saída também não será linear;
- Tahn: muito semelhante a função Sigmóide, sua principal diferença é que varia entre -1 e 1.

Com a evolução dos estudos das redes MLP, foi necessário o uso de funções mais modernas e complexas para redes mais profundas.

- ReLU: é uma função não linear, seu maior defeito é que seu gradiente é sempre 0 quando x é menor do que 0;
- Leaky ReLU: uma versão otimizada da ReLU, resolve o problema para quando x for menor que 0, que na ReLU seu gradiente sempre é 0. A Leaky ReLU resolve esse problema;
- Softmax: é um tipo de função sigmóide, que ajuda a lidar com problemas de classificação. A saída de neurônio em uma camada sigmoide segue sendo em 0 e 1, e seus valores podem ser interpretados como probabilidades.

2.5.5 Backpropagation

Para o treinamento das redes MLP *feedforward* o algoritmo mais utilizado têm sido o *Backpropagation*. O algoritmo é baseado em duas fases, primeiro a fase *forward*, onde ocorrem as entradas e os dados passam pela rede, depois a fase *backward*, onde é calculado o erro, por isso o nome *Backpropagation* (MITCHELL, 1997).

A fase *forward* do algoritmo tem o seguinte funcionamento: a entrada, chamada de sinal, é enviada da camada entrada para cada camada da rede. O sinal é propagado até que chegue na camada de saída. A última camada tem seu resultado que é comparado com as saídas desejadas.

A fase *backward* do algoritmo tem o seguinte funcionamento: a partir da última camada até a entrada, baseado no gradiente são reajustados os pesos da atual camada para reduzir seus erros (BRAGA et al., 2007).

```
Inicializar pesos e parâmetros
Repetir até o erro ser mínimo ou até a realização de um dado número de ciclos:

    Para cada padrão de treinamento X
        Definir a saída da rede através da fase forward
        Compara saídas produzidas com as saídas desejadas
        Atualizar os pesos dos nodos através da fase backward
    FimPara

FimRepetir
Fim
```

Figura 6 – Pseudo-código Backpropagation.

Fonte: (BRAGA et al., 2007)

A Figura 6 demonstra um pseudo-código do funcionamento do *Backpropagation*.

A principal vantagem do *backpropagation* é a minimização de erros, pois o cálculo do erro da fase *backward* faz o ajuste nos pesos da última camada até primeira, fazendo assim com que, no conjunto de treinamento, os erros vão se minimizando (MITCHELL, 1997). Um exemplo dado por Mitchell (1997) utiliza como função de ativação a função Sigmóide e como custo a função Erro Quadrático Médio.

Na Equação 10 o w_i representa o peso que deve ser recalculado, x_i é o dado de entrada, e o δ é erro calculado para o neurônio associado ao peso em questão do erro. A Equação 11 refere-se ao ajuste do bias, muito semelhante ao ajuste do peso.

$$w_i j = w_i j - \eta x_i \delta_j \tag{10}$$

$$b_j = b_j - \eta \delta_j \tag{11}$$

O cálculo do erro δ é feito de duas maneiras diferentes. Uma para última camada e outra para camadas ocultas (intermediárias). A Equação 12 refere-se a última camada. A saída obtida é representando por y_j e a saída desejada por d_j. Nas camadas intermediárias, o principio do cálculo é o mesmo, a principal diferença como é visível na Equação 13 é que é feito um somatório dos pesos e dos erros δ que encontrados na camada seguinte.

$$\delta_j = y_j(1 - y_j)(y_j - \delta_j) \tag{12}$$

$$\delta_j = y_j(1 - y_j)\sum w_j k \delta_k \tag{13}$$

2.5.6 Máquina de Vetores de Suporte - SVM

As Máquinas de Vetores de Suporte (SVM) é uma técnica para reconhecimento de padrões baseada em aprendizado supervisionado. O conceito é que a SVM recebe dados que sejam separados por um hiperespaço em dois grupos e quando houver uma nova entrada, o algoritmo seja capaz de dizer em qual das classes a nova entrada se encaixa. Entre o hiperplano e as instâncias mais próximas de cada uma das classes tem-se o que é chamado de margem, e a primeira entrada é aquela que está mais próxima da margem, logo é o dado que menos diverge da outra classe (CORTES; VAPNIK, 1995).

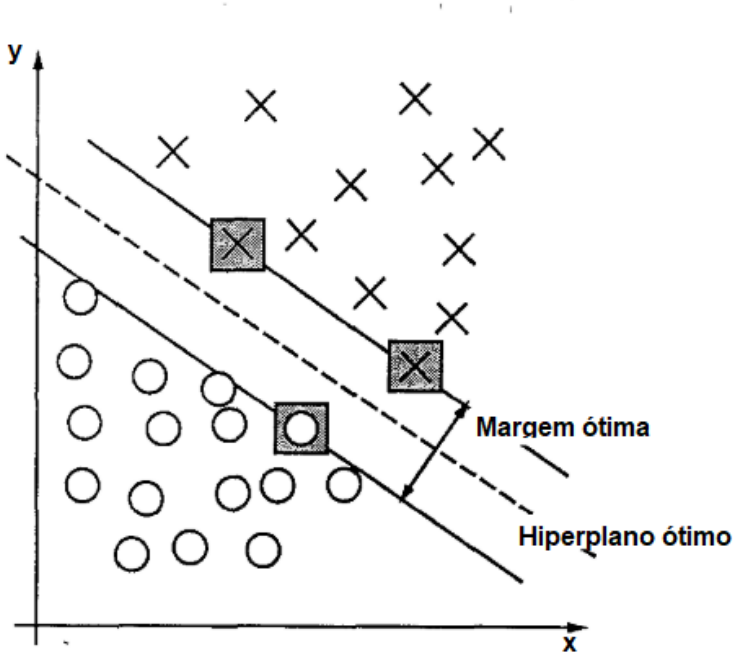

Figura 7 – Exemplo de SVM.
Fonte: (CORTES; VAPNIK, 1995)

Na Figura 7, tem-se um exemplo de SVM, no meio, a linha trajada representa o hiperplano ótimo, e as linhas contínuas representam as margens ótimas. E em destaque com quadrados em volta são as três entradas mais próximas ao hiperplano, e que dão origem a margem.

Uma função muito utilizada para as SVMs é a função Kernel, as funções Kernel são conhecidas como funções que independente da ordem dos parâmetros o resultado não é alterado. Essas funções podem ser utilizadas para mapear problemas não linearmente separável em um problema linearmente separável em um plano maior por exemplo $\mathbb{R}^2 -> \mathbb{R}^3$. (HOFMANN,

2006).

O *kernel trick*, em tradução literal truque do kernel é baseada em sistemas de equação, e usa produtos escalares para aumentar o número de dimensões. Atualmente as funções kernel mais utilizadas são a Polinomial, Gaussiana e Sigmoide. O motivo é porque elas tendem a aumentar o número de dimensões consideravelmente (LINDSETMO et al., 2008).

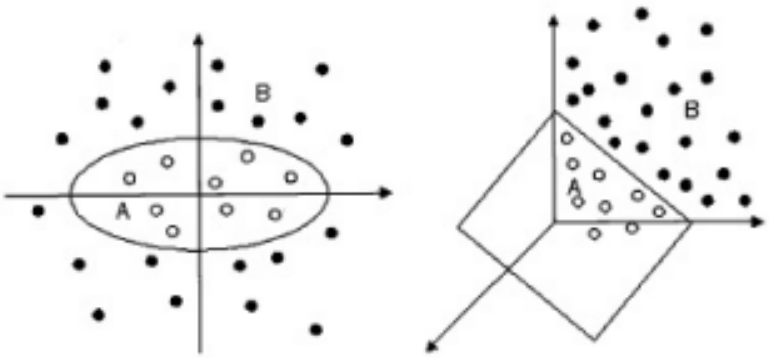

Figura 8 – Mapeamento de dados não linearmente separáveis no $\mathbb{R}^2 -> \mathbb{R}^3$ para o $\mathbb{R}^2 -> \mathbb{R}^3$.
Fonte: (HOFMANN, 2006)

Na Figura 8, tem-se o exemplo de uma função que no R^2 não é linearmente separável e ao passar ela para o R^3 ela se torna separável por hiperplano.

2.6 TESTE DE HIPÓTESES

O teste de hipóteses é utilizado para que seja feita medição e comparação entre dados estatísticos(MILAN, 2011). As hipóteses são divididas em hipótese nula H_0 e hipóteses alternativas H_n. A hipótese nula é normalmente expressada por uma igualdade enquanto a hipótese alternativa pode ser por diferente, menor ou maior (LOPES, 2003).

Os testes podem ocorrer de três formas: (a) bicaudal ou bilateral, quando as hipóteses são diferentes; (b) unilateral direito, quando o valor da hipótese alternativa é maior do que a nula; e (c) unilateral esquerdo, quando o valor da hipótese alternativa é menor do que a nula, como apresentado na Figura 9.

Os testes de hipótese podem conter dois tipos de erros que são conhecidos como erros

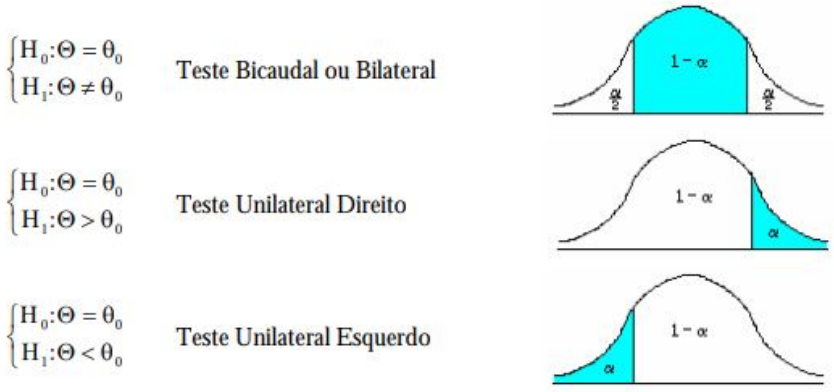

Figura 9 – Tipos de testes de hipótese.

Fonte: (LOPES, 2003)

Tipo I e Tipo II. O erro Tipo I é quando H_0 é rejeitado e a hipótese é verdadeira. E o Tipo II quando H_1 é rejeitado e a hipótese é verdadeira (LOPES, 2003). Um exemplo prático seria, a 80 por cento de chance de chover a noite, a H_0 é vai que chover a noite, e H_1 é que não vai chover a noite. O erro Tipo I é rejeitar H_0 ir sem guarda-chuva e se molhar, enquanto o erro Tipo II é rejeitar H_1 levar guarda-chuva, não chover e não utilizá-lo (FARIAS; LAURENCEL, 2000).

O Teste-t é um teste muito utilizado para comparar o desempenho de algoritmos em um ou mais conjuntos de dados específicos. Ele compara os algoritmos dois a dois, indicando qual dos algoritmos foi superior e qual a confiança dessa afirmação. É utilizado o mesmo padrão para o Teste-t a principal diferença é o desvio padrão que é removido, e é iniciado o trabalho com a estatística T com distribuição de t de *Student* com n-1 graus de liberdade (MILAN, 2011).

Segundo Vieira (1997) o Teste t, tem o seguinte funcionamento. Primeiro são escolhidas as hipóteses H_0 e H_1, depois os níveis de significância desejados. Após isso é necessário medir as diferenças d entre todas as observações x_i feitas, como apresentado na equação 14.

$$d = x_2 - x_1 \tag{14}$$

Depois disso é necessário calcular as médias das diferenças $\overline{d}$, como na equação 15, em que n representa o total de diferenças.

$$\overline{d} = \frac{\sum d}{n} \tag{15}$$

Após isso, é calculada a variância s^2 dessas diferenças da amostra, como na equação 16.

Tabela 1 – Tabela Teste-T

Graus de liberdade	10%	5%	1%
1	6,31	12,71	63,66
2	2,92	4,30	9,92
3	2,35	3,18	5,84
4	2,13	2,78	4,60
5	2,02	2,57	4,03
6	1,94	2,45	3,71
7	1,90	2,36	3,50
8	1,86	2,31	3,36
9	1,83	2,26	3,25

Fonte: Adaptado de Vieira (1997)

$$s^2 = \frac{\sum(d-\overline{d})^2}{n-1} \tag{16}$$

Por fim, é calculado o valor de t, como na Equação 17.

$$t = \frac{\overline{d}}{\sqrt{\frac{s^2}{n}}} \tag{17}$$

Dado o valor de t absoluto, ele deve ser comparado com o valor crítico na Tabela t (Tabela 1 para até 9 graus de liberdade) , no nível de significância estabelecido com os mesmos graus de liberdade. Se o valor crítico for maior igual ao valor de t, significa que essa hipótese deve ser rejeitada.

2.7 TRABALHOS RELACIONADOS

O trabalho de Schmidt (2017), foram utilizados os algoritmos *Random Forest*, Máquinas de Vetores de Suporte e Redes MLP para prever resultados de ligas de futebol. Seu melhor resultado é de 58,77% utilizando a base de dados do campeonato inglês das temporadas 2016-2017. Porém, é utilizada uma abordagem diferenciada usando dados do jogo de entretenimento e simulador de futebol, FIFA, produzido pela *EA Sports*, no seus *datasets* são inseridos atributos do jogo FIFA, os *ratings* médios do ataque, da defesa e do meio campo dos times. *Ratings* sãs as taxas que medem a habilidade de um jogador, por exemplo se um jogador com o *rating* de 89 de drible, for tentar driblar um jogador com 72 de defesa, o jogador com

rating maior tem grandes chances de conseguir o drible. Esses dados podem vir a ser utilizados, porém a *EA Sports* não revela como eles são calculados, se utilizado algum método científico como um algoritmo ou uma fórmula matemática. O que a *EA Sports* já assumiu que o sucesso internacional do jogador contribui no *rating* do jogador.

Já no trabalho de Carpita et al. (2015), melhores resultados são obtidos, chegando a 84% no melhor dos testes. Porém, esse trabalho tem uma abordagem diferente, o trabalho não busca prever resultados das partidas, mas sim eventos que ocorrem dentro da partida que levam o time a vencer. Por exemplo, ele utiliza em seus *datasets* dados que já ocorreram dentro da partida, como posse de bola, número de chutes no alvo, número de cruzamentos e etc. Com esses dados ele tenta prever qual time vencerá a partida. Já no presente trabalho, não são utilizados dados de dentro da partida que tenta prever, mas apenas dados passados de partidas passadas dos times que vão se enfrentar.

No trabalho de Joseph et al. (2006) ele utiliza Bayesiano Ingênuo, Redes Bayesianas, KNN e árvore de decisão para tentar prever resultados de um único time o *Tottenham Hotspur* da Inglaterra. O *dataset* montado para o desenvolvimento da pesquisa, foram os jogos do *Tottenham* das temporadas de 1995/1996 e 1996/1997. E os atributos escolhidos foram, a presença dos melhores jogadores do *Tottenham* na partida (verdadeiro ou falso), a qualidade da equipe adversário (alta, média ou baixa), e o local se o *Tottenham* joga em casa ou fora. O melhor resultado ficou com as Redes Bayesianas que acertou 59,21%.

3 MATERIAIS E MÉTODOS

Neste capítulo serão descritos os materiais e métodos utilizados para o desenvolvimento deste trabalho. O método utilizado é o *Knowledge Discovery in Databases* representado pela sigla KDD, que é a descoberta de conhecimento em base de dados. Com isso foi seguida a metodologia de Fayyad et al. (1996): Identificação e coleta dos dados, pré-processamento, mineração de dados e obtenção de conhecimento.

3.1 MATERIAIS

Nesse seção serão descritos os materiais utilizados para a realização desse trabalho, como foi construída e a origem da base de dados, o ambiente e software utilizados para realização do trabalho.

3.1.1 Obtenção dos dados

Dados de partidas de futebol estão amplamente distribuídos na Internet. A plataforma *Kaggle*[1], traz uma quantidade muito grande de base de dados. Nessa plataforma existem muitas base de dados esportivas e muitos delas sobre futebol. Uma dessas base de dados chamada *European Soccer Database* contém dados de mais de 25 mil partidas das 11 principais ligas europeias segundo a UEFA que são as ligas:

1. *Belgium Jupiler League*: liga da Bélgica;
2. *Premier League*: liga da Inglaterra (pode conter times do Reino Unido);

[1] https://www.kaggle.com/

3. *Ligue 1*: liga da França;
4. *Italy Serie A*: liga da Itália;
5. *Bundesliga*: liga da Alemanha;
6. *Eredivisie*: liga da Holanda;
7. *Poland Ekstraklasa*: liga da Polônia;
8. *Portugal Liga ZON Sagres*: liga de Portugal;
9. *Scotaland Premier League*: liga da Escócia;
10. *La Liga BBVA*: liga da Espanha;
11. *Switzerland Super League*: liga da Suiça;

A base de dados contém mais de 10 mil jogadores das temporadas de 2008 a 2016. Essa base de dados vem em um formato de banco de dados relacional, com 7 tabelas *Country* (País), *League* (Liga), *Team* (Time), *Match* (Partida), *Player* (Jogador), *Player Attributes* (Atributos do jogador), *Team Attributes* (Atributos do time). A Figura 10 mostra como é o relacionamento desse banco de dados.

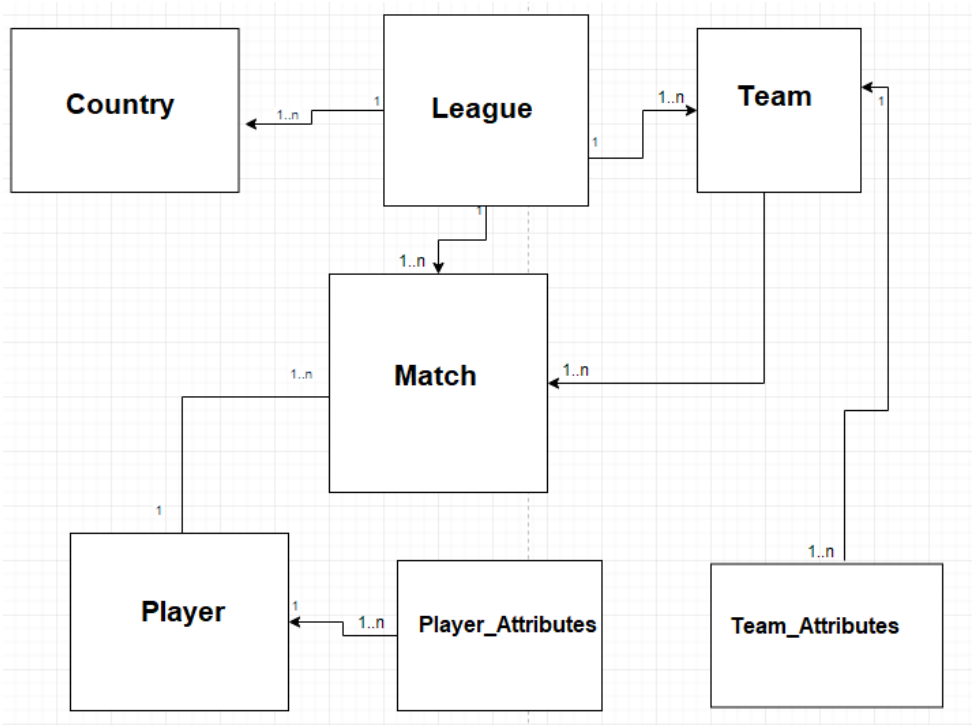

Figura 10 – Relacionamento banco de dados *European Soccer Database*.

Fonte: Autoria própria

O que chama atenção nessa base de dados é a enorme quantidade de informação dentro dela. A tabela *Match*, que tem informações sobre a partida, contém muitas informações como: horário, local, resultado, posse de bola, cartões, escanteios, chutes no gol, chutes fora do gol, jogadores e até posicionamento. O banco de dados utilizado por essa base de dados é o SQLite.

Por isso para utilizar esse banco foi utilizada a ferramenta portátil *SQLite Studio*[2] para consultas, edição e o que mais for necessário com os dados. Além do fato que a página oficial dessa base de dados na plataforma *Kaggle* indica o uso dessa ferramenta.

```
1 Select m.home_team_api_id, m.home_team_goal, m.away_team_api,
     m_away_team_goal, m.possession
2 from match m, team t
3 where m.away_team_api_id = t.team_api_id
4 and m_league_id = 21518
5 and m.away_team_api_id = 8634
6 and m.date = '2009-05-02 00:00:00';
```

Código 3.1 – Consulta SQL

O código SQL 3.1 faz uma consulta para a partida Real Madrid x FC Barcelona na temporada 2008-2009 no dia 2 de maio de 2009 utilizando o *SQLite Studio*, em que o Barcelona venceu por 6-2 no estádio do Real Madrid. O resultado desta consulta traz alguns dados da partida:

1. *home_team_api*: 8633, identificador do time mandante no caso Real Madri;
2. *home_team_goal*: 2, quantidade de gols que o time mandante fez;
3. *away_team_api*: 8634, identificador do time visitante no caso o FC Barcelona;
4. *away_team_goal*: 6, quantidade de gols que o time visitante fez;

Outra base de dados interessante é o da Copa do Mundo. Essa base de dados contém dados de todos as Copas do Mundo com exceção da última edição realizada na França, desde a primeira edição realizada no Uruguai em 1930 até a de 2014 realizada no Brasil.

Esses dados estão separados em três planilhas, no formato csv. Uma das planilhas contém os dados das partidas. Por qual fase da Copa do Mundo a partida era válida, local, resultado, entre outras informações. Outra dos jogadores, que tem as informações se o jogador era titular, a posição, se marcou um gol, número da camisa e seu treinador. A terceira planilha contém os resultados das respectivas Copas do Mundo, vencedores, vice-campeões, terceiros e quartos colocados.

[2] https://sqlitestudio.pl/index.rvt

3.1.2 Ambiente

O Ambiente utilizado para execução dos algoritmos e padronização dos dados é um computador com um processador Ryzen 7 1700 3.8 GHz, 16 GB de memória RAM, 220GBs de SSD e uma placa de vídeo AMD RX 480 de 8 GBs. Utilizado o sistema operacional Microsoft Windows 10 Ultimate, o software LibreOffice Calc[3], o SQLite Studio 3.0 portátil e o Weka[4]. Por mais que esse seja um computador pessoal, é um computador considerado potente, já que ele tem um processador recém lançado e com clock consideravelmente alto, uma placa de vídeo potente e um quantidade vasta de memória RAM.

3.1.3 Weka

Desenvolvido pela *University of Waikato*, da Nova Zelândia. O Weka (*Waikato Environment for Knowledge Analysis*) é um software livre desenvolvido na linguagem Java que contém várias ferramentas e algoritmos que auxiliam na modelagem e análise de dados. O Weka faz várias tarefas de mineração de dados como, pré-processamento, agrupamento, classificação entre outros.

A interface gráfica do Weka pode ser acessada no item Explorer do menu, dentro dessa tela temos várias opções de mineração de dados, separadas em abas na parte superior da tela.

- *Preprocess*: Essa tela é a responsável pelo pré-processamento dos dados na ferramenta. É possível normalizar dados, deletar linhas ou colunas e preparar os dados segundo o estudo realizado;
- *Classify*: Neste local é possível encontrar os algoritmos de classificação e regressão, o objetivo é classificar dados e verificar se os algoritmos estão acertando de acordo com a predição desejada;
- *Cluster*: Esse painel é focado nos algoritmos de agrupamento;
- *Associate*: Focado nos algoritmos de associação, essa tela é focada em criar associação entre dados;
- *Select Attributes*: Seu objetivo é selecionar atributos que podem ser mais precisos para previsões em conjuntos de dados;

[3]https://pt-br.libreoffice.org/
[4]https://www.cs.waikato.ac.nz/ml/weka/

- *Visualize*: Essa tela da ferramenta mostra gráficos, matriz de dispersão entre outros de forma gráfica, ou seja visual.

3.1.4 SQLite Studio

No site *Kaggle* onde é encontrado a base de dados *European Soccer Database* na página dessa base de dados é indicado a utilização do SQLite Studio como ferramenta de visualização dos dados dessa base de dados. Além da indicação pelos criadores dessa base de dados, o SQLite Studio é uma ferramenta extremamente leve e simples de usar. O SQLite Studio faz o uso da linguagem de Bancos de Dados relacionais, SQL, para consultas, edição ou até deletar os dados.

3.2 MÉTODOS

Nesta sessão serão descritos os métodos utilizados para realização do trabalho. O pré-processamento de dados, que precisam ser deixados no padrão para a realização de treinamentos, comparações, retreinamentos (se necessário) e testes de acurácia, precisão entre outros.

3.2.1 Pré-Processamento

Essa etapa visa padronizar os dados para a descoberta do conhecimento, Excluir dados desnecessários, e corrigir possíveis erros.

No caso da base de dados *European Soccer Database*, na tabela *Match* existem muitas colunas com dados que não serão relevantes para essa pesquisa, por tanto podem ser excluídos. Já a base de dados das Copas do Mundo no formato csv também tem muitos dados desnecessários que não serão relevantes para a pesquisa, por exemplo os árbitros das partidas,

ou o número da camisa do jogador.

Verificando as duas base de dados e que existem muitos dados que serão irrelevantes para a pesquisa, a Tabela 2 contém uma versão otimizada da base de dados da Copa do Mundo em formato de *Data Sets*, e o tratamento necessário para os dados, que fiquem de forma otimizada para melhor execução dos algoritmos.

O *dataset* da Copa do Mundo conforme Tabela 2, é formado por todos os confrontos das copas de 1930 realizada no Uruguai até 2014 realizada no Brasil com um total de 836 jogos. Nos testes pre-liminares foi detectado que existiam vários dados que não seriam úteis para essa pesquisa, eles então foram retirados.

Em busca de melhores resultados, foram adicionados vários atributos para o *dataset*, baseado no desempenho do time nos últimos jogos, e também seu histórico contra o adversário atual. Para isso foram criados atributos em formato de taxas que variam de 0 a 1, ou seja, taxa de vitória, empate e derrota do time nos últimos x jogos, onde x assume vários valores e permite a criação de diversas taxas. Por exemplo, se $x = 4$ e nos últimos quatro jogos do Brasil houveram três vitórias e um empate, sua taxa de vitória será de 0.8, sua taxa de empate 0.2 e sua taxa de derrota 0.0. Os valores de x definidos para criação das taxas são dos últimos 1, 2, 3, 5, 10 e todos os jogos anteriores.

Outros atributos extraídos são referentes ao histórico do confronto atual, analisando confrontos passados com as mesmas equipes e extraindo taxas de vitória, empate e derrota para as duas equipes da partida a ser analisada. Por exemplo, no confronto Brasil vs Argentina, se $x = 3$ e, nos últimos 3 jogos, o Brasil venceu 2 vezes a Argentina, a taxa de vitória do Brasil é de 66,66%, quanto a Argentina tem uma taxa de vitória de 33,33%. Para o confronto direto, os valores de x utilizando foram 1, 2, 3, 5 e 10 e todos jogos anteriores.

Os testes iniciais apontaram para seis versões do *dataset*. O *dataset* **Copa do Mundo - V1** com os dados originais o time mandante visitante e o resultado numérico ou categórico. A versão **Copa do Mundo - V2** é o *dataset* com todas as taxas calculadas. Para executar testes em diversas situações, foram realizados testes com diferentes versões do *dataset*. Criando várias versões com as taxas calculadas diferentes. Esses atributos foram retirados da versão **Copa do Mundo - V3**. Da mesma maneira, foram removidos os atributos dos últimos cinco confrontos para recuperar apenas confrontos recentes entre as equipes em Copas do Mundo, essa versão é representado como **Copa do Mundo - V4**. A quinta versão (**Copa do Mundo - V5**), contém apenas as taxas para $x = 1$ e $x = 2$. Por fim, a versão **Copa do Mundo - V6**, contém apenas as informações do último jogo. A Tabela 2 é a versão do **Copa do Mundo - V5**, com mais taxas calculados aumenta-se o número de atributos do *dataset*.

Para todas as versões desses *datasets* foi utilizado o resultado numérico (diferença de

Tabela 2 – Tabela *DataSet* Copa do Mundo

Home Team Name	Nominal
Away Team Name	Nominal
Taxa_vitorias_n1	Numérico
Taxa_empate_n1	Numérico
Taxa_derrota_n1	Numérico
Taxa_vitorias_n2	Numérico
Taxa_empate_n2	Numérico
Taxa_derrota_n2	Numérico
Taxa_vitorias_n3	Numérico
Taxa_empate_n3	Numérico
Taxa_derrota_n3	Numérico
Taxa_vitorias_all	Numérico
Taxa_empate_all	Numérico
Taxa_derrota_all	Numérico
Taxa_vitorias_visitante_n1	Numérico
Taxa_empate_visitante_n1	Numérico
Taxa_derrota_visitante_n1	Numérico
Taxa_vitorias_visitante_n2	Numérico
Taxa_empate_visitante_n2	Numérico
Taxa_derrota_visitante_n2	Numérico
Taxa_vitorias_visitante_n3	Numérico
Taxa_empate_visitante_n3	Numérico
Taxa_derrota_visitante_n3	Numérico
Taxa_vitorias_visitante_all	Numérico
Taxa_empate_visitante_all	Numérico
Taxa_derrota_visitante_all	Numérico
Taxa_vitoria_mandante_sobre_visitante_n1	Numérico
Taxa_empate_mandante_sobre_visitante_n1	Numérico
Taxa_derrota_mandante_sobre_visitante_n1	Numérico
Taxa_vitoria_mandante_sobre_visitante_n2	Numérico
Taxa_empate_mandante_sobre_visitante_n2	Numérico
Taxa_derrota_mandante_sobre_visitante_n2	Numérico
Taxa_vitoria_mandante_sobre_visitante_n3	Numérico
Taxa_empate_mandante_sobre_visitante_n3	Numérico
Taxa_derrota_mandante_sobre_visitante_n3	Numérico
Taxa_vitoria_mandante_sobre_visitante_all	Numérico
Taxa_empate_mandante_sobre_visitante_all	Numérico
Taxa_derrota_mandante_sobre_visitante_all	Numérico
Resultado	Numérico (gols mandante – gols visitante)
ResultadoDesc	Nominal (valor a ser predito)

Fonte: Autoria Própria

gols) para os regressores, e a classe (vitória, empate e derrota) para os classificadores.

Para execução do trabalho o *dataset European Soccer Database* foi dividido em vários *datasets* de cinco ligas nacionais. Liga Inglesa, Alemã, Francesa, Italiana e Espanhola com jogos de 2008 á 2016. Esses *datasets* tem muitas semelhanças, o Inglês, Francês, Italiano e Espanhol são campeonatos compostos por 20 times e contém 3040 partidas, com exceção do Italiano que tem 3016, já que algumas partidas foram retiradas devido a escândalos de arbitragem. O campeonato Alemão, diferente dos outros é composto por 18 times, por tanto tem menos jogos. Os *datasets* foram mantidos da mesma forma que os da Copa do Mundo, considerando apenas os atributos relevantes do *dataset* para a tarefa. Além disso, foram adicionados três novos atributos nesses *datasets*, obtidos de uma casa de apostas, referentes às taxas de aposta para vitória, empate e derrota.

Assim como no *dataset* da Copa do Mundo para este também foram criado taxas de vitórias totais e vitórias em confrontos diretos entre as mesmas equipes. Para as ligas foram criadas as seguintes taxas: 1, 2, 3, 5, 10, 15, 19 e todos os jogos. E para confronto direto, 1, 2, 3, 5, 10 e todos os jogos. Para o Campeonato Alemão, considerando que ele é menor, foram criadas as taxas, 1, 2, 3, 5, 10, 13, 17 e todos jogos, confronto direto, 1, 2, 3, 5, 10 e todos jogos. Essas taxas foram escolhidas baseado na quantidade de jogos dos times por campeonatos, por exemplo, em todos os campeonatos um turno tem 19 jogos, enquanto no alemão tem 17 jogos. E as outras taxas escolhidas com o intuito de medir a fase do time a longo, médio e curto prazo.

Os *datasets* para os testes criados foram, da versão 1 **(V1)** até a 8 **(V8)**. Na V1, o *dataset* é mais simples, apenas com o mandante, visitante e o resultado. O V2 tem todas as taxas calculadas, o V3 não usa as taxas longas, retirando-se as taxas de todos os tempos, e dos jogos 19, 15. O V4 são deixadas apenas as taxas médias e curtas, V5 foi retirada a taxa de 5 jogos, V6 a taxa de 3 jogos, V7 a de 2 jogos e V8 apenas o número da apostas, dessa forma os *datasets* das ligas fica muito semelhante com o da Tabela 2.

Todos os *datasets* foram chamados pelo nome do campeonato, seguido da versão, por exemplo, Campeonato Inglês - V1, Campeonato Espanhol - V5.

Tabela 3 – Exemplo do atributo time mandante antes da transformação one-hot

Time mandante	Time visitante
Barcelona	Valencia
Real Madrid	Villarreal
Atlético de Madrid	Levante

Fonte: Autoria Própria

Lembrando ainda que para a execução dos algoritmos o Weka, faz quando necessário, a transformação one-hot, como é visível nas Tabelas 3 e 4 os dados são transformados em

Tabela 4 – Exemplo do atributo Liga após a transformação one-hot

Time	Time Mandante Barcelona	Time Mandante Real Madrid	Time Mandante Atl de Madrid
Barcelona	1	0	0
Real Madrid	0	1	0
Atlético de Madrid	0	0	1

Fonte: Autoria Própria

atributos no caso o time mandante, se aquele for o time mandante, ele toma o valor como 1, se não 0, por exemplo o Barcelona como mandante recebe no atributo Time mandante Barcelona recebe, Time mandante Barcelona = 1, Time mandante Real Madrid = 0 e Time mandante Atlético de Madrid = 0, logo os algoritmos irão concluir que o Barcelona é o time mandante naquele jogo.

3.2.2 Treinamento e comparação

Para o treinamento desse trabalho será utilizada a ferramenta já citada na seção de materiais, o Weka.

Tabela 5 – Algoritmos de aprendizado de máquina e seu retorno

Algoritmo	**Estratégia de Aprendizado Supervisionado**
Regressão Linear	Regressão
Bayesiano Ingênuo	Classificação
Redes MLP	Classificação/Regressão
SVM (SMO, SMOReg)	Classificação/Regressão

Fonte: Autoria Própria

Os algoritmos de aprendizado de máquina a serem utilizados para treinamento e comparação na ferramenta são citados na Seção 2.5, a Tabela 5 mostra quais são os algoritmos e seu retorno é uma regressão ou uma classificação. Esses algoritmos retornam uma classe ou um valor numérico de acordo com a tarefa de classificação ou regressão.

Os parâmetros utilizados para rodar esse algoritmos já foram predefinidos. Para execução da **Regressão Linear** e do **Bayesiano Ingênuo** serão utilizados os parâmetros padrões do algoritmo dados no Software Weka.

Para as **Redes MLP** serão utilizadas taxas de aprendizado com decaimento

exponencial de 0.5, 0,1, 0,05, 0,01, 0,005, 0,001, 0,0005, 0,0001. E o número de camadas ocultas utilizadas e de neurônios por camadas oculta e serão de (5), (10), (10, 10), (50, 50). Para as **SVM** serão utilizados o *Normalized Poly Kernel*, *Puk*, *RBFKernel* e *PolyKernel*.

Serão utilizados um parâmetro de cada vez congelando os outros, essa estratégia é conhecida como descida coordenada de encosta.

Os cálculos e métodos utilizados para comparação e testes, serão o Teste T e Acurácia. O Teste T, é um teste de hipótese para comparar dois a dois, os algoritmos que executarão melhor.

A **Acurácia** é a soma dos acertos, com as entradas dadas como falsos positivos e falsos negativos. Isso pode ser visto na Equação 18 adaptadade de Góes e Carvalho (2018), onde a soma dos Verdadeiros Positivos (vp) com os Verdadeiros Negativos (vn) é dividido pela soma dos Verdadeiros Positivos (vp) com os Verdadeiros Negativos (vn) ainda com, os Falsos Positivos (fp) e os Falsos Negativos (fn) (MONICO et al., 2009).

$$acuracia = \frac{vp + vn}{vp + vn + fp + fn} \tag{18}$$

O coeficiente de correlação de Pearson, criado por Karl Pearson é um número no intervalo de -1 e 1, o sinal indica a direção da correlação, caso o valor da correlação seja 0, quer dizer que não existe correlação entre as variáveis (BRITTO et al., 2009). A correlação quando igual a 1, quer dizer que as variáveis tem uma correlação perfeita, quando igual a 0, não tem correlação, e quando igual a -1, uma correlação negativa perfeita (LIRA, 2004).

A Figura 11 traz como ocorrerá o andamento de busca por resultados nesse trabalho, o primeiro passo é o pré-processamento, deixar os dados da melhor forma possível, para que em seguida se realize o treinamento, o treinamento será feito com os algoritmos citados na Tabela 5, em seguida os resultados serão comparados com o esperado, para isso serão utilizados o cálculo da acurácia citado na Equação 18 e o coeficiente de correlação. Caso os resultados não fiquem dentro do esperado deverá ser feito o retreinamento, até que os resultados fiquem próximos do esperado para que seja possível retirar conhecimento da base de dados.

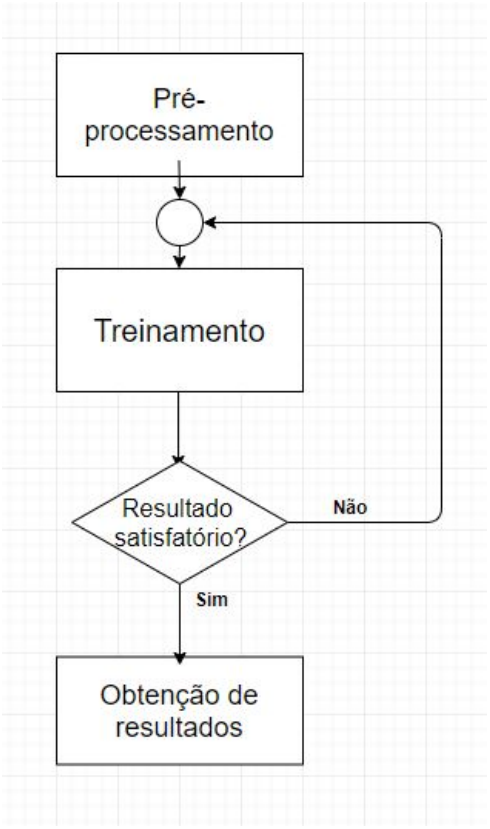

Figura 11 – Fluxograma comparação e treinamento dos dados

Fonte: Autoria própria

4 EXPERIMENTOS

Nesse capítulo serão discutidos os experimentos realizados com os 46 *datasets* e seus respectivos resultados. Conforme apresentado Capítulo 3, os algoritmos a serem executados no trabalho são a Regressão Linear, Bayesiano Ingênuo, Redes MLP, e SVM. Nesse capítulo será descrita a execução dos algoritmos, resultados e como os resultados mudaram de acordo com as mudanças realizadas nos *datasets*.

4.1 REGRESSÃO LINEAR

A Regressão Linear foi testada com todos os *datasets* e os coeficiente de correlação de Pearson podem ser vistos na Tabela 6. No caso da Copa do Mundo, o *dataset* que obteve o melhor resultado foi o V2 aquele que contém todas as taxas calculadas. Já para os campeonatos nacionais, a que gerou melhores resultados na maioria dos experimentos foi o V8 que é o *dataset* com poucos dados, apenas os números das casas de apostas, isso demonstra que os números da casa de apostas são bons indicadores das chances reais de vitória dos times. A exceção é o Campeonato Francês que obteve os piores resultados e que seu melhor *dataset* foi o V1 que tem apenas o nome dos times e o resultado, o motivo possivelmente do desempenho ruim do campeonato francês, é que dentro do período estudado, o campeonato passou a ter investimento milionários dos árabes fazendo assim com que times que não tinham bom desempenho fizesse grandes contratações e passassem a conquistar mais vitórias. O destaque fica para o Copa do Mundo V2 que obteve o melhor resultado 0,5442, onde o coeficiente de correlação mostra que tanto o histórico, quanto a fase atual do time ajudaram na Regressão Linear.

Na Figura 12, é mostrado, em forma gráfica, o resultado do *dataset* de melhor desemepenho o Copa do Mundo - V2. O eixo X é o resultado (diferença de gols) e o eixo Y é o esperado. Com isso, é perceptível que grandes goleadas não são muito esperadas dificultam a análise do algoritmo, a goleada mais esperada é no valor de 4,8. Um exemplo é o 7 a 1

Tabela 6 – Resultado - Regressão Linear

Dataset	Copa do Mundo	Inglês	Alemão	Espanhol	Francês	Italiano
V1	0,5125	0,4253	0,3818	0,5069	**0,3199**	0,3663
V2	**0,5442**	0,4355	0,3879	0,5226	0,2942	0,3829
V3	0,5394	0,4412	0,3919	0,5253	0,2894	0,3876
V4	0,5389	0,4424	0,3967	0,5271	0,2875	0,3866
V5	0,528	0,4451	0,3986	0,5276	0,2894	0,3898
V6	0,5305	0,4458	0,4005	0,53	0,2898	0,3908
V7	-	0,4464	0,402	0,5275	0,2868	0,3927
V8	-	**0,4484**	**0,405**	**0,5311**	0,2847	**0,3982**

Fonte: Autoria própria

da Alemanha sobre o Brasil na Copa do Mundo de 2014, circulado em vermelho, o algoritmo previu por muito pouco a vitória brasileira (praticamente um empate) e o resultado do jogo foi por larga vantagem em favor dos alemães. Na imagem também, é visível que existem $x's$ de diferentes tamanhos, esses $x's$ são as instâncias. Quanto menor o x mas próximo de acertar o resultado o algoritmo ficou, quanto maior mais longe do acerto ele ficou.

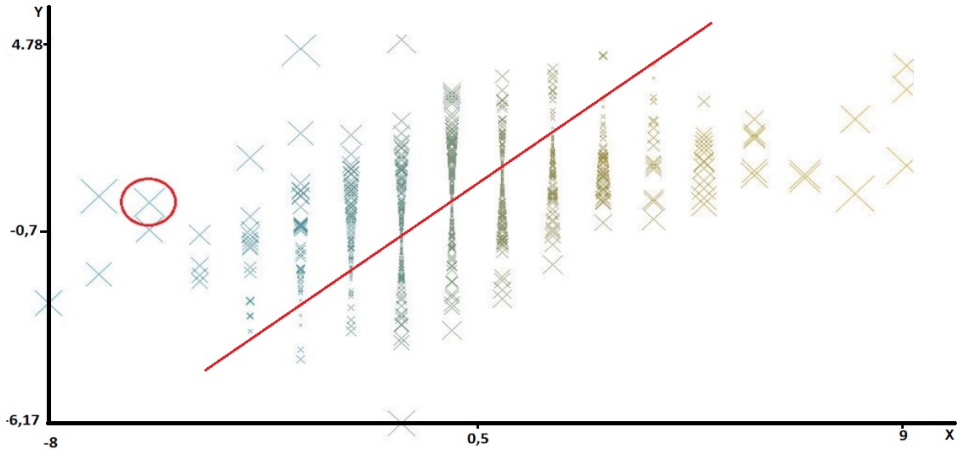

Figura 12 – Resultado experimento 1 - Regressão Linear

Fonte: Autoria própria

Tabela 7 – Resultado - Bayesiano Ingênuo

Dataset	**Copa do Mundo**	**Inglês**	**Alemão**	**Espanhol**	**Francês**	**Italiano**
V1	**60,0478%**	**51,9408%**	**49,1013%**	**53,6356%**	47,3355%	**51,6578%**
V2	49,6411%	48,2797%	44,4444%	46,0376%	44,1188%	46,9496%
V3	49,1627%	46,5789%	43,4641%	46,0376%	45,2632%	46,7838%
V4	51,1962%	46,7434%	43,8317%	46,1601%	45,1645%	47,3475%
V5	53,9474%	47,2039%	43,7908%	45,9150%	46,6776%	48,0438%
V6	58,6124%	47,9605%	43,6683%	45,7108%	46,9079%	49,0053%
V7	-	48,4211%	43,3007%	45,5065%	**47,8289%**	50,6963%
V8	-	48,2895%	42,8922%	44,4036%	46,8750%	50,2653%

Fonte: Autoria própria

4.2 BAYESIANO INGÊNUO

O Bayesiano Ingênuo foi executado em 46 *datasets*. As acurácias obtidas na execução do algoritmo nos respectivos *datasets* são mostradas na Tabela 7. Como o Bayesiano Ingênuo é um classificador, o resultado obtido no *dataset* é o categórico, com os dados empate, vitória e derrota.

Os resultados observado variaram entre 42 e 58%. Na maioria dos *datasets*, uma diferença de 1% na na acurácia corresponde a aproximadamente 30 jogos. Observa-se que existem algum desbalanceamento entre as classes, onde empate ocorre com menos frequência.

O resultado é mostrado na Tabela 7. O Bayesiano Ingênuo se comportou melhor com o *dataset* apenas com o time mandante, visitante e o resultado como classe. Isso demonstra que ele está verificando que os times com mais tradição costumam vencer times com menos tradição. A exceção novamente é o Campeonato Francês que teve um resultado melhor com o histórico único de uma partida. Os destaques ficam para os ***Datasets*** **Copa do Mundo - V1 e Campeonato Espanhol - V1**, o possível motivo desses resultados serem melhores é que o algoritmo detectou que seleções com tradição no futebol costumam vencer as com menos tradição, e no Campeonato Espanhol existem dois times com tradição muito acima da média e que, com frequência, conquistam o campeonato que são o Barcelona e Real Madrid, isso pode ter ajudado no desempenho do algoritmo acima da média sem as taxas calculadas.

Tabela 8 – Resultado - SMOReg

Dataset	**Copa do Mundo**	**Inglês**	**Alemão**	**Espanhol**	**Francês**	**Italiano**
V1	0,4713	0,4161	0,3612	0,4507	0,2499	0,3204
V2	0,495	**0,4355**	0,363	0,5049	0,2799	0,3656
V3	0,5028	0,4256	0,3719	0,5076	0,2749	0,369
V4	0,5133	0,4277	0,3705	0,51	0,281	0,3698
V5	0,5121	0,4344	0,3757	0,5141	0,2807	0,3716
V6	**0,5183**	0,4351	0,3812	0,5168	**0,2846**	0,3763
V7	-	0,4327	**0,3813**	0,5125	0,2833	0,375
V8	-	0,4338	0,3743	**0,5134**	0,2792	**0,379**

Fonte: Autoria própria

4.3 SVM

Para a execução dos experimentos com SVM, foram executados testes nos *datasets* de 2 formas, primeiro na forma de regressão, ou seja buscando prever o resultado, e depois na forma de classe tentando buscar a classe correta, vitória, empate ou derrota. Para SVM foram testados os seguintes *kernels* como parâmetros no ***Dataset* Copa do Mundo - V2**, o *NormalizedPolyKernel* obteve 60,4067% de acerto, o *Puk* com 52,512% de acerto, o RBFKernel com **54,5455% de acerto**, e o *PolyKernel* **61,2440%**. Por tanto o *Kernel* escolhido para rodar todos os testes foi o padrão do Weka 3.8.3.

4.3.1 SMOReg

O SMOReg é um algoritmo regressão, os coeficientes de correlação obtidos ao executar o algoritmo nos respectivos *datasets* podem ser vistos na Tabela 8.

Independente do *dataset* os resultados são muito próximos, isso mostra que o SMOReg é muito flexível dentro dos atributos escolhidos. Porém quando observados os resultados é compreendido que, assim como na Regressão Linear, os *datasets* V8 tem bons resultados ficando em segundo ou terceiro, reafirmando que os atributos das apostas são bons atributos para os regressores.

Tabela 9 – Resultado - SMO

Dataset	**Copa do Mundo**	**Inglês**	**Alemão**	**Espanhol**	**Francês**	**Italiano**
V1	58,9713%	50,4605%	49,6324%	53,2271%	46,8750%	49,7016%
V2	**61,2440%**	51,0855%	49,7958%	53,2271%	46,9737%	51,5584%
V3	59,0909%	51,0526%	49,7141%	**53,3497%**	47,4013%	51,5915%
V4	59,2105%	50,9539%	**49,9592%**	53,0637%	47,2697%	51,3263%
V5	60,0478%	51,0197%	49,6732%	53,1046%	**47,8618%**	**51,6247%**
V6	60,6459%	**51,6579%**	49,5507%	53,2271%	48,0592%	51,3926%
V7	-	50,8224%	49,4281%	53,2680%	47,7303%	51,2268%
V8	-	50,7566%	49,7141%	53,2680%	47,3684%	50,9284%

Fonte: Autoria própria

4.3.2 SMO

O SMO roda como um algoritmo de classificação. A acurácia obtida ao executar o algoritmo nos respectivos *datasets* é mostrada na Tabela 9.

A Tabela 9 mostra os resultados para o algoritmo SMO na forma de classificação. O SMO da Copa do Mundo respondeu bem com todos os atributos ficando acima do 61% de acerto. Nos campeonatos nacionais, assim como no SMOReg, os resultados foram de diferentes *datasets*. Alguns se comportaram melhor com mais atributos sobre o histórico de longo prazo do time, outros com histórico mais recente. Isso prova que, para o SMO, também os atributos são flexíveis. Os resultados também foram muito próximos um do outro quando observado o Campeonato Inglês e Espanhol do melhor resultado para o pior a diferença fica na faixa de 1%.

4.4 MLP

As redes MLP, assim como a SVM, podem executar tanto como um regressor, quanto como um classificador. O teste foi executado nos dois casos e, para o classificador, a acurácia é mostrada na Tabela 11 e, para os regressores, o coeficiente de correlação é mostrado na Tabela 12. Assim como visto no Seção 3.2, os parâmetros foram testados em uma fase preliminar e redes com duas camadas foram aquelas que obtiveram melhores resultados foram. As topologias mais promissores consistiram de 5 neurônios na primeira camada escondida e 3 na segunda camada escondida. A tabela também apresenta redes com 10 e 50 neurônios por camada escondida. A taxa de aprendizado foi de 0,001. O resultado dos testes é mostrado na

Tabela 10 – Resultados testes de parâmetros para Redes MLP

Camadas e nós	Taxa de aprendizado	Resultado
Padrão Weka		49,6411%
(5, 3)	0,5	55,2632%
(5, 3)	0,1	51,555%
(5, 3)	0,05	54,067%
(5, 3)	0,01	54,287%
(5, 3)	0,005	55,8622%
(5, 3)	0,001	**62,4402%**
(5, 3)	0,0005	57,6555%
(10, 10)	0,001	38,2775%
(50, 50)	0,001	38,8756%

Fonte: Autoria própria

Tabela 11 – Resultado - MLP

Dataset	**Copa**	**Inglês**	**Alemão**	**Espanhol**	**Francês**	**Italiano**
V1	59,9282%	51,3158%	49,6324%	52,9820%	**47,9605%**	50,8952%
V2	**62,4402%**	50,79%	49,2239%	53,3088%	44,8684%	51,8899%
V3	62,201%	50,6579%	49,4281%	53,6765%	43,8816%	52,0557%
V4	61,3636%	51,1184%	49,0196%	53,1454%	44,2434%	51,3594%
V5	61,8421%	50,9211%	49,7141%	54,0033%	44,9013%	51,9894%
V6	62,204%	51,2500%	49,8775%	**54,3301%**	45,1974%	52,3873%
V7	-	51,2829%	**50,0000%**	53,8807%	45,0000%	52,7851%
V8	-	**51,9408%**	49,7549%	53,9240%	45,8882%	**52,8846%**

Fonte: Autoria própria

Tabela 10.

No caso do resultado como classificador o *dataset* V2 da Copa do Mundo obteve a melhor acurácia, ficando acima dos 62% de acerto. Para os campeonatos nacionais novamente houve uma mescla de resultados, porém todos com taxas curtas (atributos do passado recente), sem taxas ou apenas com os valores das apostas. Assim como no caso do SMO, novamente houve pouca diferença, por exemplo, o Campeonato Alemão, se observado do melhor resultado para o pior a diferença é menor de que 1%.

Para as MLPs em forma de regressores o melhor resultado de todos ficou novamente com o Copa do Mundo com todas as taxas que obteve um coeficiente de correlação de 0,5637. Nos campeonatos nacionais se comportou como na maioria das regressões, os melhores resultados foram taxas curtas e o as versões do *dataset* V8. Novamente a exceção é o Francês que obteve seu melhor resultado apenas com o time mandante visitante e seu resultado.

Tabela 12 – Resultado - MLPReg

Dataset	**Copa**	**Inglês**	**Alemão**	**Espanhol**	**Francês**	**Italiano**
V1	0,5227	0,423	0,3778	0,505	**0,317**	0,3552
V2	**0,5637**	0,4241	0,3871	0,519	0,2878	0,3724
V3	0,5511	0,4273	0,3923	0,5204	0,2859	0,3737
V4	0,5514	0,4308	0,3943	0,5218	0,2845	0,3742
V5	0,5431	0,4337	0,3975	0,5235	0,2858	0,377
V6	0,5453	0,4375	0,4006	0,5258	0,2873	0,3799
V7	-	0,4386	**0,4023**	0,5238	0,2862	0,3823
V8	-	**0,4408**	0,4013	**0,5266**	0,2855	**0,3848**

Fonte: Autoria própria

4.5 TESTE T

Como mostrado na Seção 2.6, o teste de hipótese escolhido para o trabalho foi o Teste T que é bicaudal, que compara os algoritmos dois a dois. O Weka permite usar a versão corrigida do Teste T, adaptada para o uso de validação cruzada. Nessa seção será mostrado os resultados desses experimentos, e qual algoritmo venceu em cada uma das comparações. Para execução do Teste T no Weka, foi escolhido um nível de significância de 5% logo, uma confiança de 95% e na forma bicudal como é padrão do *Software* Weka.

A comparação foi feita três a três, ou seja, os três classificadores são comparados entre si usando acurácia e os três regressores usando correlação. Para o teste, foram usados todos os *datasets* na V6 para os campeonatos nacionais e V4 para a Copa do Mundo, esses são os *datasets* com as taxas calculados até $x = 3$, esses datasets escolhido para a execução do teste não é totalmente independente ou seja ele passa por uma seleção de atributos. Feitas essas comparações, o Weka apresentou os resultados que são mostrados nas Tabelas 13 e 14.

Os resultados são descritos da seguinte forma, melhor desempenho quando o Weka considera que houve melhora considerável, empate técnico definodo como o experimento no qual não é possível afirmar se um dos algoritmos é melhor de forma estatisticamente significante e pior desempenho quando o outro algoritmo teve melhor desempenho.

Na Tabela 13 é mostrado que a MLP e a SMO se destacaram, ambas ganharam do Bayesiano Ingênuo sendo melhor em cinco *datasets* e empate técnico em um. Já o comparando MLP com a SMO, houve um empate técnico em cinco *datasets* e o SMO foi melhor em um, no geral um empate técnico. No empate técnico, a hipótese nula do Teste-T não foi rejeitada, ou seja, não é possível dizer se um algoritmo se saiu melhor ou pior que o outro.

A Tabela 13 apresenta a comparação entre regressores. Nesse caso, o desempenho foi

Tabela 13 – Resultados classificadores - Comparações

Referência	Adversário	Melhor desempenho	Empate	Pior de desempenho
MLP	Bayesiano Ingênuo	5	1	0
MLP	SMO	0	5	1
Bayesiano Ingênuo	SMO	0	1	5

Fonte: Autoria própria

Tabela 14 – Resultados classificadores - Regressores

Referência	Adversário	Melhor desempenho	Empate	Pior de desempenho
MLP	Regressão Linear	0	6	0
MLP	SMOReg	0	6	0
Regressão Linear	SMOReg	0	3	3

Fonte: Autoria própria

mais equilibrado, quando comparados MLP com Regressão Linear houve um empate em todos os *datasets*, apenas quando comparado Regressão Linear com SMO, o SMOReg se sai melhor por três vezes.

Como já visto, o Teste T não pôde dizer qual foi o melhor de todos os algoritmos, já que ele compara algoritmos dois a dois. Por isso não é possível concluir que o SMOReg foi melhor dos regressores e que MLP e SMO foram o melhor dos classificadores. Contudo, comparando resultados dois a dois, é possível verificar que SMO tende a ser melhor que Regressão Linear e que MLP e SMO tendem a ter melhor desempenho comparados ao Bayesiano Ingênuo.

5 CONCLUSÃO

O objetivo dessa pesquisa era comparar algoritmos de aprendizado de máquina para o cenário esportivo, mais especificamente o futebol. Foram analisados regressores e classificadores e depois foi executado um teste de hipótese, especificamente o Teste T para comparar esses algoritmos. A pesquisa e a execução dos experimentos demonstraram que os classificadores foram todos acima do classificador aleatório 33,33% para vitória, empate ou derrota. O destaque ficou com as *Redes Multi Layer Perceptron* e as Máquinas de Vetores Suporte, que em alguns casos ficaram acima dos 60% de acerto como é mostrado nas Tabelas 9 e 11 e também foram os algoritmos que tiveram maior destaque no Teste T.

Já as regressões tiveram resultados muito próximos uma da outra, os melhores resultados tiveram um coeficiente de correlação, próximo a casa dos 0,55, como é mostrado nas Tabelas 6 e 12. O Teste T também mostra essa proximidade entre os algoritmos com uma leve vantagem da SMOReg sobre a Regressão Linear e SMO e MLP tecnicamente empatados.

Os trabalhos que utilizam essas técnicas para previsão de resultado de jogos de futebol, no geral utilizam *datasets* diferentes e sempre com algumas modificações para tentar melhorar o resultado de suas respectivas pesquisas. Alguns trabalham com média de gols, com o público, com o histórico daquele contra aquele time, em alguns casos prever os resultados de um único time, em outros de toda uma liga. Porém, quando comparado o resultado dessa pesquisa com outros é perceptível que o resultado de 63% é razoavelmente bom e mostra uma pequena melhora.

A pesquisa demonstra que com a evolução da Inteligência Artificial com mais dados que hoje são cada vez mais guardados em relação a todos os esportes e não apenas o Futebol, é possível que num futuro próximo com esses dados possam ser feitas previsões de forma mais eficiente.

5.1 TRABALHOS FUTUROS

Em trabalhos futuros em relação a previsão, é recomendado que se crie atributos que possam medir quanto o elenco do time é bom e quanto contém jogadores com capacidade de desequilibrar o jogo, para que assim melhores previsões possam ser feitas.

Para criação desses atributos, é recomendado que se crie fórmulas que contém quantas esses jogadores decidem jogos e os clubes que estão, por exemplo quantos gols da vitória um jogador fez, ou gols de título, ou por exemplo um zagueiro quantos desarmes fez, ou quantas finalizações bloqueadas. Nos últimos anos esses dados tem sido guardados, então o ideal é buscar um meio de armazenar todos e coloca-los em *datasets* para rodar os mesmos.

Uma boa técnica e que pode ser utilizada no futuro é, utilizar dados das últimas temporadas e prever em uma próxima. Por exemplo, utilizar dados das temporadas 2011-2012, 2012-2013 e 2013-2014 e prever da temporada 2014-2015.

Também utilizar algoritmos de fora do Weka, como o *libsvm* que para alguns casos é um SVM melhor que o Weka. E também investigar melhor a seleção de atributos, ver atributos relevantes para determinados algoritmos e testar para que sejam filtrados e usados os melhores atributos possíveis.

REFERÊNCIAS

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineraçao de dados educacionais: Oportunidades para o brasil. **Brazilian Journal of Computers in Education**, v. 19, n. 02, p. 03, 2011.

BRAGA, A. de P.; FERREIRA, A. C. P. de L.; LUDERMIR, T. B. **Redes neurais artificiais: teoria e aplicações**. Rio de Janeiro: LTC Editora Rio de Janeiro, Brazil, 2007.

BRITTO, D.; FILHO, F.; ALEXANDRE, J. Desvendando os Mistérios do Coeficiente de Correlação de Pearson (r) . **Universidade Federal de Pernambuco (UFPE)**, v. 18, p. 115–146, 2009.

CAMILO, C. O.; SILVA, J. C. d. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. **Universidade Federal de Goiás (UFC)**, p. 1–29, 2009.

CARPITA, M. et al. Discovering the drivers of football match outcomes with data mining. **Quality Technology & Quantitative Management**, Taylor & Francis, v. 12, n. 4, p. 561–577, 2015.

COELHO-BARROS, E. A. et al. Métodos de estimação em regressão linear múltipla: aplicação a dados clínicos. **Revista Colombiana de Estadística**, Universidad Nacional de Colombia, v. 31, n. 1, p. 111–129, 2008.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine learning**, Springer, v. 20, n. 3, p. 273–297, 1995.

DUARTE, O. **História dos esportes**. [S.l.]: Senac, 2019.

FARIAS, A. M. L. d.; LAURENCEL, L. Fundamentos de estatística aplicada: Módulo I: Estatística descritiva. **Rio de Janeiro/RJ: Universidade Federal Fluminense**, 2000.

FAYYAD, U. M. et al. **Advances in knowledge discovery and data mining**. Massachusetts Ave, Cambridge, MA: AAAI press Menlo Park, 1996.

FIFA. FIFA: Big Count 2006 - Comparison 2006 – 2000. **FIFA Communications Division, Information Services**, p. 1–12, 2007.

Folha de SP. Com final na Ásia , Liga Europa não lota estádio em Baku. **Folha de São Paulo**, p. 1–5, 2019. Disponível em: <https://www1.folha.uol.com.br/esporte/2019/05/final-em-baku-nao-anima-torcedores-e-estadio-nao-fica-lotado.shtml>. Acesso em: 21 nov. 2019.

Forbes. Valores bilionários da FIFA na Copa do Mundo. **Forbes**, p. 1–11, 2019.

GAIGEROV, B.; ELKINA, L.; PUSHKIN, S. Metrological characteristics of a group of hydrogen clocks. **Measurement Techniques**, Springer, v. 25, n. 1, p. 23–25, 1982.

GÓES, G. S.; CARVALHO, A. X. Y. d. Curso: Microeconometria. **IDP - Instituto Brasiliense de Direito Público**, 2018.

GRAETTINGER, D.; GRAETTINGER, T. Using data mining to predict the winter olympics medal counts in sochi. **Discovery corp.[Electronic resource].–2013.–**, v. 20, 2014. Disponível em: <http://www.discoverycorpsinc.com/storage/article-files-2013–15/2014/2020Winter/20Olympics/20Medal/20Count/20Prediction>. Acesso em: 21 nov. 2019.

HOFMANN, M. Support vector machines-kernels and the kernel trick. **Houptseminar report**, v. 26, 2006.

JOSEPH, A.; FENTON, N. E.; NEIL, M. Predicting football results using bayesian nets and other machine learning techniques. **Knowledge-Based Systems**, Elsevier, v. 19, n. 7, p. 544–553, 2006.

KRAVCHYCHYN, C. et al. Estudos brasileiros sobre o esporte: ênfase no esporte-educação. **Movimento**, Escola de Educação Física, v. 18, n. 2, 2012.

LAROSE, D. T.; LAROSE, C. D. **Discovering knowledge in data: an introduction to data mining**. Hoboken New Jersey: John Wiley & Sons, 2014.

LINDSETMO, R.-O.; JOH, Y.-G.; DELANEY, C. P. Surgical treatment for rectal cancer: an international perspective on what the medical gastroenterologist needs to know. **World journal of gastroenterology: WJG**, Baishideng Publishing Group Inc, v. 14, n. 21, p. 3281, 2008.

LIRA, S. A. Análise de correlação: abordagem teórica e de construção dos coeficientes com aplicações. **Setores de Ciências Exatas e de**, p. 209, 2004.

LISI, C. A. **A history of the World Cup: 1930-2006**. Londres, UK: Scarecrow Press, 2007.

LOPES, L. F. D. **Apostila de Estatística**. Santa Maria - RS: DE–UFSM, 2003.

LOUZADA, F. et al. Poly-bagging predictors for classification modelling for credit scoring. **Expert Systems with Applications**, Elsevier, v. 38, n. 10, p. 12717–12720, 2011.

MACIEL, T. V. et al. Mineração de dados em triagem de risco de saúde. **Revista Brasileira de Computação Aplicada**, v. 7, n. 2, p. 26–40, 2015.

MESQUITA, M. E. R. V.; PIMENTA, M. A. Perceptrons morfológico de camada única. **Unicamp**, 2012.

MILAN, L. A. Estatística aplicada. **Universidade Federal de São Carlos**, 2011.

MIN, B. et al. A compound framework for sports results prediction: A football case study. **Knowledge-Based Systems**, Elsevier, v. 21, n. 7, p. 551–562, 2008.

MITCHELL, T. M. **Machine learning**. New York: McGraw hill, 1997.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. **Sistemas inteligentes-Fundamentos e aplicações**, v. 1, n. 1, p. 32, 2003.

MONICO, J. F. G. et al. Acurácia e precisão: revendo os conceitos de forma acurada. **Boletim de Ciências Geodésicas**, Universidade Federal do Paraná, v. 15, n. 3, p. 469–483, 2009.

RADTKE, J. J. Função Degrau. **UTFPR**, 2008. Disponível em: <http://paginapessoal.utfpr.edu.br/jonas/disciplinas/calculo-diferencial-e-integral-4/03_funcao_degrau.pdf/view>. Acesso em: 21 nov. 2019.

REPUCOM. **Global Television Audiences for Sports Events**. 2014. Disponível em: <http://www.iboperepucom.com/br/>. Acesso em: 21 nov. 2019.

SANTOS, R. D. Futebol e sua História: Possibilidade de Efetivação da Proposta critica superadora. **Universidade Estadual de Santa Catarina - UNESC**, v. 1, n. 8, p. 1 a 11, 2006.

SCHMIDT, H. L. Uso de técnicas de aprendizado de máquina no auxílio em previsão de resultados de partidas de futebol. **Universidade de Santa Cruz do Sul**, 2017.

SCHUMAKER, R. P. **Sports Data Mining_Book**. New Rochelle, New York: Springer, 2010.

SOUZA, S. de. Disciplina : Aprofundamento em Futebol Disciplina : Aprofundamento em Futebol. **Universidade Salgado de Oliveira**, p. 1–3, 2007.

Sports Deloitte. Bullseye Football Money League. **Sports Deloitte**, n. January, 2019. Disponível em: <https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/sports-business-group/deloitte-uk-deloitte-football-money-league-2019.pdf>. Acesso em: 21 nov. 2019.

SUTTO, G. No Super Bowl , propagandas podem custar R $ 640 mil por segundo. **Info Money**, p. 13–14, 2019. Disponível em: <https://www.infomoney.com.br/noticia/imprimir/7903547>. Acesso em: 21 nov. 2019.

TUBINO, M. **O que é Esporte - Tubino**. Tatuaté, São Paulo: Brasiliense, 1993.

VIAENE, S.; DERRIG, R. A.; DEDENE, G. A case study of applying boosting naive bayes to claim fraud diagnosis. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 16, n. 5, p. 612–620, 2004.

VIEIRA, S. **Introdução à bioestatística**. Rio de Janeiro, RJ: Elsevier Brasil, 1997.

ZUBEN, I.-P. F. J. V. Rede mlp: Perceptron de múltiplas camadas. **DCA/FEEC/Unicamp**, 2013.