FERNANDO DAMAS WENZEL

ESTUDOS DE MÉTODOS DE ANÁLISE PREDITIVA EM JOGOS DE FUTEBOL

Monografia apresentada ao Programa de Educação Continuada da Escola Politécnica da Universidade de São Paulo, para obtenção do título de Especialista, pelo Programa de Pós-Graduação em Big Data - Inteligência na Gestão dos Dados.

SÃO PAULO

2019

FERNANDO DAMAS WENZEL

ESTUDOS DE MÉTODOS DE ANÁLISE PREDITIVA EM JOGOS DE FUTEBOL

Monografia apresentada ao Programa de Educação Continuada da Escola Politécnica da Universidade de São Paulo, para obtenção do título de Especialista, pelo Programa de Pós-Graduação em Big Data - Inteligência na Gestão dos Dados.

Área de concentração: Tecnologia da Informação - Big Data

Orientador: Profa. Dr. PedroLuiz Pizzigatti Corrêa

SÃO PAULO

2019

FICHA CATALOGRÁFICA

Wenzel, Fernando
ESTUDOS DE MÉTODOS DE ANÁLISE PREDITIVA EM JOGOS DE FUTEBOL / F. Wenzel -- São Paulo, 2014.
49 p.
Monografia (Especialização em Big Data - Inteligência na Gestão dos Dados) - Escola Politécnica da Universidade de São Paulo. PECE – Programa
de Educação Continuada em Engenharia.
1.Big Data 2.Machine Learning 3.Mineração de dados 4.Data Mining I.Universidade de São Paulo. Escola Politécnica. PECE – Programa de
Educação Continuada em Engenharia II.t.

AGRADECIMENTOS

Agradeço ao orientador deste estudo, de grande valia para meu aprendizado e desenvolvimento pessoal. Ao meu pai, que sempre prezou pelo cultivo da minha sabedoria. Ao meu amigo, Guilherme, que me motivou em seguir no curso. Ao meu irmão, Rodrigo, que me incentivou a entrar neste ramo. Aos meus gestores, que possibilitaram a minha participação no curso.

CURSO BIG DATA: INTELIGENCIA NA GESTÃO DOS DADOS

Coord.: Profa. Dra. Solange N. Alves de Souza
Vice-Coord.: Profa. Dra. Lucia V. Leite Filgueiras

Perspectivas profissionais alcançadas com o curso:

O curso de Big Data foi responsável por me apresentar diversas tecnologias diferentes, sejam aquelas que têm impacto direto no Big Data como também outras que a complementam. Como minha formação não é em TI, o curso também foi sólido em ensinar as bases de todos os processos.

Julgando-me novo no mercado de trabalho, não tenho dúvidas que o curso me abriu e abrirá mais possibilidades no futuro, sendo que hoje sou capaz de discutir a área em um nível alto e podendo me especializar em uma das vertentes que fora apresentada no curso.

RESUMO

A previsibilidade do mercado esportivo é um tema que chama a atenção de diversos públicos distintos: entusiastas, estudantes e até mesmo investidores. O futebol, especificamente, é o esporte com maior número de fãs no mundo. Suas variáveis podem ser separadas em duas diferentes categorias: as pré-jogo e durante o jogo. As variáveis pré-jogo são as possíveis de se analisar antes mesmo do jogo acontecer, como o local em que a partida será realizada, quem será o mandante, qual o histórico de jogos de ambos os times e quem tem, possivelmente, o melhor elenco. Já as durante o jogo são fatores que acontecem no decorrer da partida - uma expulsão ou lesão, fatores que não são capazes de prever, por exemplo. Assim sendo, o trabalho teve como escopo o levantamento de hipóteses seguido da aplicação de modelos estatísticos que utilizaram fontes de dados abertos, tentando predizer não somente o resultado da partida antes dela acontecer, mas, deixando espaço para futuros trabalhos que poderão atuar também no momento em que a partida ocorre. No decorrer do trabalho, são estudadas diversas técnicas já consideradas em trabalhos prévios, com o intuito de encontrar qual modelo do *Machine Learning* é o mais apropriado para a realidade e ambiente do futebol brasileiro, um dos mais competitivos do mundo.

Palavras-chave: *Machine Learning*, mineração de dados, modelo preditivo, futebol.

ABSTRACT

The predictability of the sports market is a topic that catches the attention of many different audiences: enthusiasts, students and even investors. Football, specifically, is the sport with the most fans in the world. Your variables can be selected into two categories: pre-game and during-game. Pre-match variables are possible to analyze even before they happen, such as where a match will be played, who will be the principal, what is the game history of both times and who has possibly the best cast. Already as during the game are factors that occur during the match - an expulsion or injury, factors that are not able to predict, for example. Thus, the work has as its scope the application of theoretical models and model applications that use public and diverse data sources, trying the predator not only the match result before it happens, but also adapting during it. In the course of the work, several techniques already performed by other theorists will be studied to find the *machine learning* model or the most appropriate to the reality and environment of Brazilian football, one of the most competitive in the world.

Keywords: *Machine Learning*, data mining, predictive model, soccer.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Machine Learning (ML) é uma das áreas de inteligência que vem sendo destacada e sendo elencada como uma das mais promissora nos últimos anos. O esporte é um dos setores que mais vem sendo procurada pelos estudantes da técnica. Muito desse interesse deve do alto crescimento do mercado de apostas em jogos de futebol (THABTAH; BUNKER, 2017). Muitas técnicas e indicadores já foram desenvolvidos ao longo dos anos e são frequentemente base de estudos para entusiastas e investidores. Porém, a empregabilidade das técnicas pode variar de acordo com o ambiente aplicado, tendo muitas delas se mostrado ineficientes quando aplicadas em períodos espaçados (P. BUNKER,RORY; THABTAH, FADI, 2013).

Estudos demonstram que a utilização de técnicas de mineração de dados em redes sociais associadas com modelos padrões podem gerar resultados melhores do que quando aplicados individualmente (ADAMIDES; KAMPAKIS, 2014).

O futebol é um jogo complexo, que concilia técnica e estratégia, combinando diversas variáveis que podem ser coletadas e analisadas, podendo auxiliar na previsão de seu resultado. Estas variáveis podem ser trazidas à tona antes da partida iniciar – como histórico de jogos recentes, local da partida e jogadores – e durante – como expulsões e lesões (AALBERS; VAN HAAREN, 2018).

Avanços na área de modelos preditivos e métodos de *Machine Learning* para processamentos massivos de dados possibilitam que a incorporação de variáveis estruturadas, ou não, possam ser consideradas para aplicação prática. Porém, poucos são os trabalhos acadêmicos que se propuseram a aplicar técnicas baseadas no ambiente brasileiro de futebol.

1.1 Motivação

Alguns pontos foram determinantes para a motivação do trabalho em questão. O mercado de esportes é, naturalmente, algo que atrai a atenção de grande parte da

população. Tal motivo também pode explicar a relevância que se tem dado ao mercado de apostas e os diversos materiais expostos sobre o tema, que concilia ganho financeiro com o entendimento do jogo.

Outro motivo concentra-se no possível ganho esportivo. O ganho esportivo é quando o resultado não se estende apenas ao interesse comercial, mas principalmente na contribuição ao entendimento pleno do esporte (A. MARTINS AND A. UFF, 2009). Este ganho pode ser obtido através de um modelo estatístico, já que uma das equipes da disputa pode entender melhor as variáveis que tendem a influenciar mais no resultado final.

Um ponto também relevante trata-se da escassez de artigos científicos no Brasil, quando comparado à quantidade de artigos publicados em países mais bem desenvolvidos (https://www.tecmundo.com.br/mercado/126770-brasil-tem-maior-percentual-publicacoes-cientificas-acesso-aberto.htm). Assim, contribuirá também para estudos nacionais.

1.2 Objetivo

O trabalho tem como objetivo o uso de técnicas de *Machine Learning* (ML) para avaliar o impacto de diversas variáveis resposta, identificadas no capítulo 4, sobre jogos de futebol, sejam estas variáveis avaliadas antes do jogo acontecer ou durante a partida.

Sendo assim, estima-se que este estudo com abordagens de mineração de dados na área futebolística, resulte em informações sobre a probabilidade de uma equipe ganhar, empatar ou perder um determinado jogo que está para ocorrer.

1.3 Justificativa

Há, nos dias atuais, uma grande variedade e volume de dados sobre futebol - estatísticas de jogadores, jogos, clubes, treinamentos, notícias, etc. Porém, não encontrou-se conteúdo que objetiva a estruturação destes dados a fim de tentar predizer qual o resultado da partida.

Um ponto relevante para a distinção deste estudo frente aos demais é a utilização de dados provenientes da franquia mais bem sucedida de jogos que simulam o futebol - FIFA, produzido pela empresa EA Games. Estes dados foram extraídos do site Sofifa.com, sendo consumido como um serviço da web. O site atualiza semanalmente estes dados. No jogo, cada jogador recebe uma pontuação para os vários atributos importantes para a qualidade de um jogador. Isto será abordado nos próximos capítulos do trabalho.

Combinados aos dados individuais de cada jogador que constitui um time, histórico de jogos reais dos campeonatos inglês e espanhol foram extraídos através de uma API disponibilizada pelo site https://football-data-api.com.

Posterior as tratativas de dados, foram separadas as variáveis principais para aplicação de algoritmos a fim de se obter uma porcentagem estatística acerca do mais provável resultado de um jogo.

1.4 Contribuição

O esporte mais popular do mundo é o futebol (DVORAK, JIRI & JUNGE; ASTRID & GRAF-BAUMANN; TONI & PETERSON, LARS, 2004). Logo, o jogo atrai a atenção de fãs, apostadores (MAURICIO MURAD, 2016), treinadores e da mídia em si. Todos estes grupos são amplamente interessados em conhecer o jogo e quais aspectos podem influenciar no resultado de uma partida. Os aspectos podem ser dos mais variados possíveis, desde os mais dedutíveis, como a qualidade dos jogadores de um time, o local da partida, moral recente da equipe, até os que a princípios não tão claros, como a média de escanteios por jogo.

Prever o resultado de uma partida de futebol é uma tarefa difícil. Atletas, jornalistas, comentadores e treinadores tentam prever há algum tempo, sem um consenso claro (BHATTACHARYYA S, JHA S, THARAKUNNEL K, AND WESTLAND JC, 2011). A tecnologia avançou nos últimos anos como um todo e a forma como coletamos os dados e os mineramos foram aprimorados. Já é possível realizar complexas consultas de dados e armazenar grandes volumes de informação com certa facilidade. Porém, o dado, por si só, não é expressivo se não conseguirmos extrair alguma informação ou conclusão a partir de sua análise. Para que seja possível tratar o dado desde sua origem, surgiu a mineração de dados (J. HUCALJUK AND A. RAKIPOVIC, 2011).

A Mineração de Dados é uma forma de lidar com problemas referentes a qualidade de dados, cujo objetivo é apoiar alguma decisão técnica. As técnicas de mineração de dados são aplicadas em diversas áreas produtivas. Indústrias automobilísticas, por exemplo, são conhecidas por reduzir problemas no processo operacional através de análises de falhas. As companhias de telecomunicação utilizam a mineração de dados para evitar que um cliente troque de operadora (ASUR, S., & HUBERMAN, B. A, 2010).

Portanto, este trabalho é mais uma fonte de aprendizagem, ampliando o conhecimento teórico e técnico destas atividades, essenciais aos tempos modernos.

1.5 Metodologia

Para este estudo, a metodologia utilizada é um processo já consolidado para o tratamento de problemas variantes da mineração de dados, CRISP-DM (Cross-Industry Standard Process for Data Mining).

A metodologia CRISP-DM consiste em trabalhos com amostras limitadas de dados (https://semantix.com.br/blog/como-explorar-e-gerenciar-dados-com-o-crisp-dm/

). Para esta pesquisa, foram consideradas duas estratégias para esta metodologia – *growing window* e *sliding window*.

1.6 Organização do trabalho

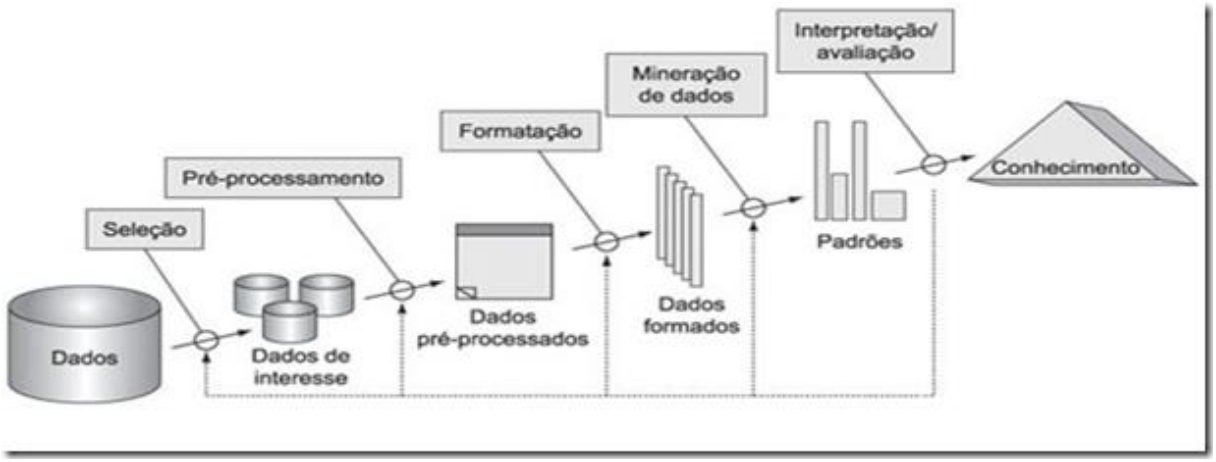
A introdução visa destacar os objetivos e motivação do trabalho. O desenvolvimento teórico formula e respalda a validade técnica do estudo, contendo diversas fontes relevantes para a construção desta pesquisa. A preparação dos dados, parte fundamental da construção da mineração de dados, detalha como foi feita a captura e tratamento do dado, seguindo a metodologia descrita anteriormente. A modelagem tem como fim utilizar os dados já estruturados e tirar dele algum valor. A conclusão abrevia o trabalho e seus resultados.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Mineração de dados

O dado é importante ativo de grande valor de decisão para uma empresa. Assim, para lidar com o processo de extração e manipulação de dados relevantes à uma atividade, alcunhou-se a mineração de dados.

Figura 1 - Processo de mineração de dados



Fonte Figura 1 – Navega, 2002

2.2 Metodologia CRISP-DM

A mineração de dados é um processo lógico e racional utilizado para descobrir informações que de fato são relevantes para um processo específico, onde o volume de dados é muito grande. O objetivo desta técnica é encontrar padrões que antes eram desconhecidos. Uma vez que estes padrões são encontrados, eles podem e devem contribuir para tomadas de decisões multilaterais.

Os passos que devem ser considerados, seguindo a metodologia CRISP-DM (CHAPMAN, P., 2000) são:

- Exploração dos dados;
- Identificação de padrões;
- Aplicação da mineração de dados.

2.2.1 Exploração de dados

Na primeira etapa da exploração de dados, o dado é, primeiramente, limpo. Isso significa que todo o excesso não necessário é retirado a fim de manter uma base de dados sólida e organizado. Essa transformação é posterior à detecção das variáveis chaves para a resolução do problema em questão.

A exploração de dados foi fundamental para que o modelo a ser criado no decorrer do trabalho tivesse resultados relevantes. Sem esta etapa, variáveis importantes seriam desconsideradas da análise e não haveria tempo hábil para descobrir novos padrões.

Aqui, trabalhamos com hipóteses que poderiam ou não ajudar a identificação de padrões de dados. Estes trabalhos serão detalhados no capítulo 4. Após a comprovação ou não destas hipóteses, o próximo passo é a identificação destes padrões.

2.3 Identificação de padrões

Após a fase de exploração de dados ser concluída, o segundo passo é criar um método de identificação de padrões. Identificar e escolher, de maneira clara e objetiva, os padrões que realizem as melhores predições.

Os padrões reconhecidos são, então, aplicados em prol do resultado esperado. Para a aplicação, a seleção de uma técnica descrita na seção anterior será utilizada, posterior a análise de eficácia de cada uma delas.

2.3.1 Técnicas de predição de desempenho

Para avaliar o desempenho de um modelo criado, é necessário a obtenção de um conjunto de dados para treinamento em que o valor da variável Y - os pontos de uma equipe, por exemplo - é conhecido. Também é necessário que os dados do treinamento sejam diferentes dos dados de testes, para que o modelo possa ser validado.

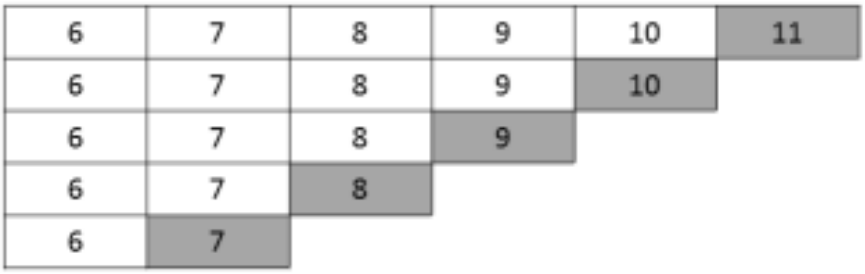
Essa separação dos dados é necessária para que não ocorra o chamado *overfitting*, que é um ajustamento aos dados de treinamento. Esta situação se dá quando o modelo criado depende de um conjunto de dados já conhecido e que, quando aplicado em outros conjuntos, não apresenta resultados satisfatórios. O modelo apenas memorizou os dados de treino. Sendo assim, ele é ótimo para lidar com o atual conjunto de dados, mas ruim quando tenta prever outros dados. Ou seja, tecnicamente, conforme se aumenta a precisão dada por um modelo para um conjunto de dados já conhecido, ele tende a ser menos confiável para novos conjuntos de dados não conhecidos.

Esta separação de dados pode ser feita utilizando várias técnicas conhecidas no mercado, utilizaremos duas em específico: *Growing Window e Sliding Window*

2.3.2 Growing Window

Nesta estratégia, o conjunto de dados de treino sempre vai crescendo conforme as previsões são efetuadas.

Figura 2 - Estratégia Growing Window



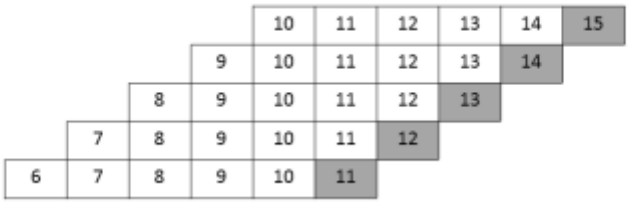
Fonte Figura 2 - Han, 2011

A figura acima mostra os cinco primeiros testes de implementação do modelo criado, em que a 6ª interação é definida como o ponto de partida para as demais. As células sem preenchimento são as de treinamento, enquanto as marcadas de cinza são de teste.

2.3.3 Sliding Window

Nesta técnica, a interação inicial é variada conforme ocorrem. O conjunto de dados de treinamento contém, obrigatoriamente, sempre a mesma quantidade de jogos.

Figura 3 - Estratégia Sliding Window



Fonte Figura 3 - Han, 2011

A figura 3 demonstra cinco interações utilizando a estratégia *Sliding Window*. O conjunto de dados é sempre constituído pelas cinco rodadas anteriores à do teste.

A desvantagem dela é justamente a perda de memória que a envolve. Ao retirar dados antigos, podemos perder também características interessantes à um modelo.

2.4 Aplicações da mineração de dados

A mineração de dados é uma tecnologia moderna e que ainda está em processo de amadurecimento. Apesar disso, há várias empresas que já a utiliza com certa regularidade.

Muitas dessas companhias estão combinando técnicas de mineração de dados com estatística, reconhecimento de padrões, *Machine Learning*, etc. A mineração de dados é importante, pois auxilia a empresa a conhecer mais sobre seus clientes e, a partir daí, tomar decisões mais eficientes.

2.4.1 Algoritmos e técnicas

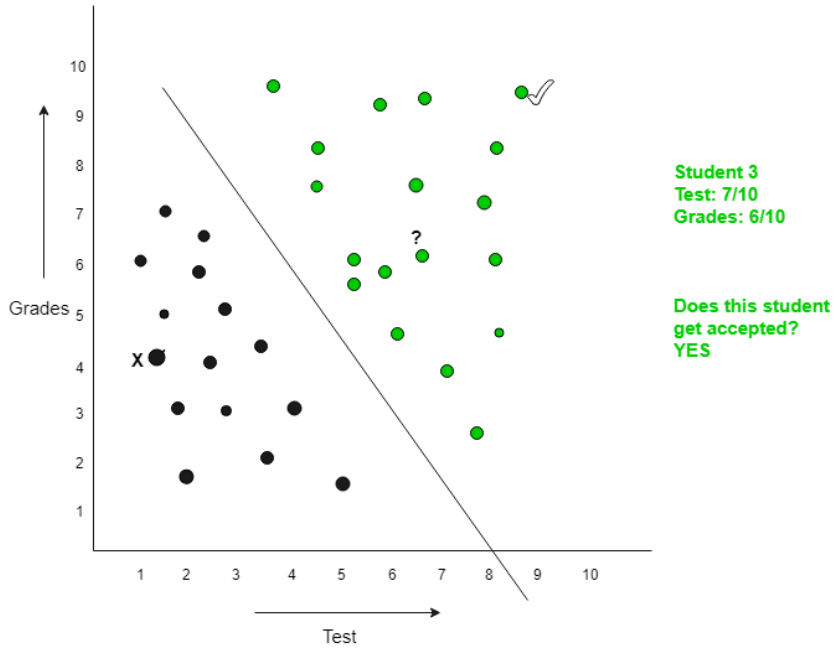
Vários algoritmos e técnicas, como classificação, *clustering*, regressão e árvores de decisão podem ser utilizadas para gerar conhecimentos de uma base de dados. Na área acadêmica, estes algoritmos são explorados vastamente.

Na sequência, serão detalhadas técnicas de modelagem de dados cujo objetivo base é a geração de algoritmos de previsão de dados. Todos eles têm como fonte principal de conhecimento a busca por padrões nos dados, seja qual for sua abordagem.

2.4.1.1 Classificação

A Classificação é o método mais aplicado de mineração de dados. Ela consiste em utilizar uma série de exemplos previamente determinados para classificar um grande volume de dados em poucas séries. É bastante utilizado na detecção de fraudes e análises de risco de crédito.

Figura 4 - Método de classificação

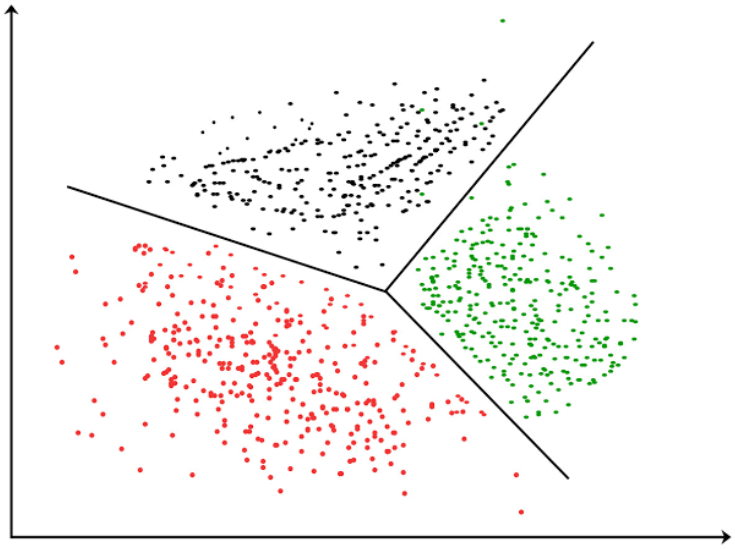


Fonte Figura 4 - Kamber, 2011

2.4.1.2 Clustering

O método *clustering* pode ser definido como uma identificação de classes similares de objetos. Utilizando as técnicas *clustering*, podemos identificar regiões mais densas de dados, onde grupos se formam. O método de classificação também é efetivo para distinguir grupos ou classes de objetos, porém é mais lento em seu processamento.

Figura 5 - Método de clusterização



Fonte Figura 5 - Kamber, 2011

Como pode ser observado na imagem 5, este método se baseia na identificação de grupos nos dados. Identificados os grupos, padrões individuais para cada classe podem vir a ser explorados com mais detalhe em um universo mais reduzido.

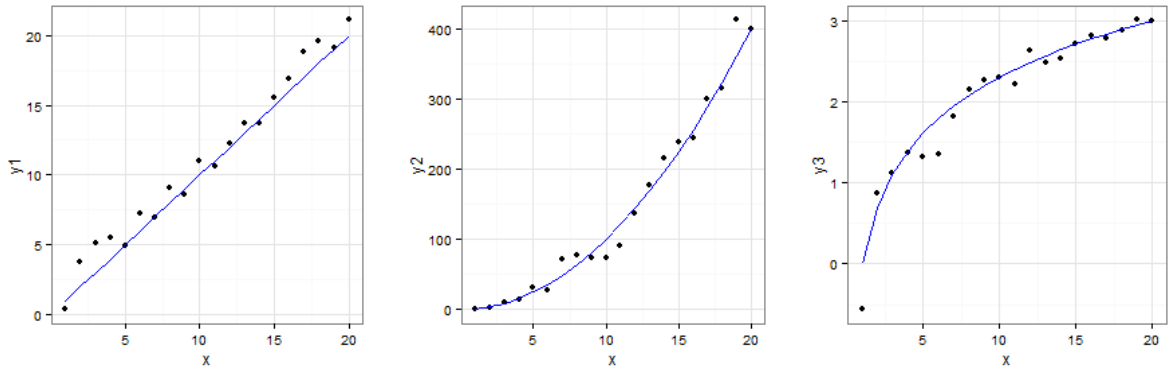
2.4.1.3 Regressões

A regressão pode ser utilizada para conjecturar o relacionamento entre uma ou mais variáveis dependentes ou independentes. Em mineração de dados, as variáveis

independentes são aquelas cujos atributos já são conhecidos e a resposta é o que queremos prever.

A maioria das situações reais não é simples de se prever, por exemplo, volume de vendas e preço de ações, pois dependem de muitas variáveis, e algumas delas desconhecidas. Com isso, técnicas mais complexas podem ser adotadas a fim de solucionar o problema em questão, como regressão logística, árvores de decisão ou redes neurais.

Figura 6 - Métodos de regressão



Fonte Figura 6 - Kampakis, 2014

2.4.1.4 Regras associativas

Associação e correlação são normalmente utilizadas para encontrar frequência em uma base de dados, que auxiliam na tomada de decisões nos negócios. Um exemplo é a análise de comportamentos de compra.

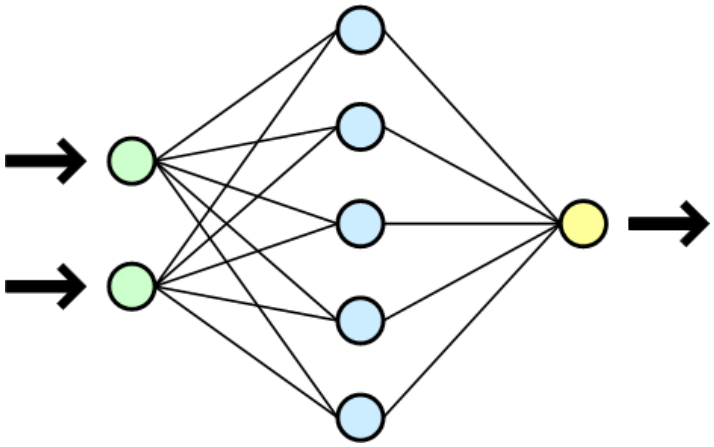
Os algoritmos dentro desta classe geram um indexador menor que 0. Porém, o número de correlação para uma base de dados é, geralmente, muito grande e a quantidade de proporção dessas regras são, muitas das vezes, de pouco valor.

2.4.1.5 Redes neurais

As redes neurais são uma série de informações de entrada e saída conectadas, ao passo que cada conexão possui um peso atribuído a ela. Durante a fase de aprendizagem, essas redes ajustam os pesos para poder prever as classes combinadas que realmente ajudam na construção do modelo

As redes neurais têm a habilidade de trazer resultados claros de dados imprecisos ou complexos, que passariam despercebidos à percepção humana ou até as outras técnicas computacionais.

Figura 7 - Redes neurais



Fonte Figura 6 - Kamber, 2011

.

3 TRABALHOS CORRELATOS

A inteligência artificial e seu promissor início tem dado para a sociedade a habilidade de criar sistemas de predição com grau de assertividade nunca antes visto. O aprendizado de máquinas vem sendo utilizado em basicamente todas as áreas, seja qual seja o objetivo de atuação, dada seu alto grau de efetividade. Uma área onde os sistemas de predição ganharam muita popularidade é no futebol - tentar prever os resultados de uma partida.

O trabalho de ULMER, B. & FERNANDEZ, 2013, demonstra o trabalho realizado em criar um modelo de predição geral, montado a partir de resultados da *Premier League*, liga de elite do futebol inglês. Utilizando artefatos de engenharia e exploração de dados, como a mineração de dados e análises exploratória, uma série de atributos que determinam os mais importantes fatores para poder prever o resultado de uma partida foi descrita. A partir deste ponto, foi possível criar um sistema efetivo de predição desses jogos. Foi demonstrada em detalhes, também, a grande dependência de um resultado com estes atributos. O melhor modelo criado, utilizando um sistema de *machine learning* é chamado *gradient boosting*. Este sistema consiste em janelas similares a *sliding window* no capítulo 2. Com ele, sagrou-se um desempenho de 0.2156 no ranking de probabilidade de resultados (RPS) para jogos ocorridos da semana 6 até a 38 na primeira divisão do campeonato inglês, agregados em duas temporadas (de 2014-2015 e 2015-2016). Para o mesmo período, as organizações de apostas oficiais (*Bet365* e *Pinnacle Sports*) obtiveram um índice inferior de 0.2012 para o mesmo cenário. Como um índice de RPS menor representa um nível de assertividade maior, o modelo criado não foi capaz de superar a previsão das organizações em questão, mas, apesar disso, é plausível dizer que os resultados foram promissores.

No trabalho de OWRAMIPUR, FARZIN & ESKANDARIAN, PARINAZ & MOZNEB, FAEZEH. (2013) é destacado que a mineração de dados é um processo de encontrar e descobrir padrões, extraindo informações úteis dos dados para futuras análises e desenvolvimento de aplicações. As aplicações da mineração de dados são extremamente amplas, indo de identificação de fraudes financeiras (BHATTACHARYYA, JHA, THARAKUNNEL, & WESTLAND, 2011; ZHOU & KAPOOR, 2011) até análises de informação e comunicações (KAUR, LECHMAN, &

MARSZK, 2017; KAUR & TAO, 2014). Nos esportes, a mineração de dados e análises preditivas têm sido aplicadas em vários esportes, desde o basquete (ŠTRUMBELJ & VRAČAR, 2012; VRAČAR, STRUMBELJ, & KONONENKO, 2016), corridas de cavalo (LESSMANNA, SUNG, & JOHNSON, 2010) e até mesmo no cricket (ASIF & MCHALE, 2016).

Futebol é o esporte mais popular do mundo (DVORAK, JIRI & JUNGE; ASTRID & GRAF-BAUMANN; TONI & PETERSON, LARS, 2004). Uma das mais importantes ligas de futebol é a inglesa, tema mais recorrente de trabalhos correlatos analisados neste trabalho, o que acabou por influenciar a decisão. Chamada de *Premier League*, a liga é a mais assistida no mundo, sendo transmitida para 643 milhões de casas em 212 países, com um potencial de audiência de quase 5 bilhões de pessoas *(https:// https://www.premierleague.com/).* Modelar um sistema preditivo para partidas de futebol não pode ser tratado como mero interesse acadêmico, mas também com um grande valor econômico.

Um artigo da BBC estimou que o valor do mercado de apostas de futebol varia de 700 bilhões a 1 trilhão de dólares (*http://www.bbc. com/sport/football/24354124*). Para termos uma medida de comparação, o PIB brasileiro fechou em 2018 em quase 7 trilhões de reais (https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/23886-pib-cresce-1-1-em-2018-e-fecha-ano-em-r-6-8-trilhoes).

Um grande problema encontrado enquanto modelando os resultados foi a alta natureza competitiva da divisão inglesa e a alta incidência de times julgados como fracos ganharem de times superiores. Um exemplo disso foi o título quase que inexplicável do Leicester City em 2016. A conquista veio contra todas as probabilidades, o que mostra a imprevisibilidade do jogo e a dificuldade em lidar com o problema principal.

A dificuldade mais relevante em prever o resultado de um jogo é que ele se encaixa em um problema de classificação multiclasses, com três resultados possíveis: vitória do time da casa, vitória do time visitante e empate.

Vários estudos com diferentes abordagens já foram divulgados. Há métodos que focam na marcação de gols, enquanto outros baseiam-se na busca dos três resultados possíveis – esta é a abordagem escolhida para o estudo.

Outro trabalho neste âmbito foi feito por JOSEPH, FENTON E NEIL (2016). O principal foco do estudo foi a criação de uma rede bayesiana para prever os resultados da equipe Tottenham Hotspurs, no período de 1995 a 1997. Lá, foi demonstrado que a rede criada pelos autores conseguiu resultados melhores do que outras técnicas de aprendizado de máquinas, como o *K-nearest*, a *naive Bayesian* e a árvore de decisões.

Apesar de o resultado obtido ter tido uma taxa de 40.79%, um alto índice de acerto, a predição do modelo focava em um time em particular e num período de tempo também específico.

OWRAMIPUR, ESKANDARIAN E MOZNEB (2013) também produziram um conteúdo com redes bayesianas. Este estudo teve como objetivo prever os resultados do time espanhol FC Barcelona. Sua pesquisa mostrou-se inovadora, já que possuía características não tão comuns para este tipo de análise, como o clima na hora da partida, estado psicológico dos jogadores e se havia ou não algum jogador lesionado no time. O estudo reportou uma acurácia considerada bem alta, com um índice de 92%, porém, novamente, concentrou-se em apenas um time, em apenas uma temporada e envolvendo apenas 20 jogos.

4 ANÁLISE PRELIMINAR DOS DADOS

4.1 Fonte e extração

Os dados utilizados para a construção da análise foram extraídos pelo site Sofifa.com e football.api.

No caso do Sofifa.com, os dados eram referentes às qualidades dos jogadores de cada time. Já no football.api, os dados representavam todas as partidas de futebol do campeonato inglês nas temporadas 2017-2018, 2018-2019 e 2019-2020. Para ambos os casos, o Python foi a linguagem de programação utilizada para importar os dados. O código final pode ser visto ao final do trabalho, no github (*https://github.com/ferweezer/Monografia-USP)*.

O site sofifa.com separa as informações em uma página para cada jogador. Sendo assim, no Python fora criado uma rotina, uma espécie de ponteiro, que captura página por página do host e, baseado na estrutura html criada, os valores de cada atleta. A programação desta rotina está postada no github (*https://github.com/ferweezer/Monografia-USP)*. Logo, significa que o Python realiza uma rotina para entrar em 18 mil páginas e executar um serviço em cada uma delas. Após a coleta, os dados de cada jogador é compilado em um arquivo “.csv”.

Já no caso do football.api o processo é bem mais simples – basta criar uma conexão REST com a API, seguindo a documentação disponibilizada pela própria dona do conteúdo e os dados são, então, coletados e guardados junto a um arquivo ”.csv”.

4.2 Pré-processamento

Esta etapa é crucial para o bom desenvolvimento da mineração de dados, já que ela tem como função a análise, preparação e transformação de dados.

4.3 Preparação de dados

No processo de mineração de um dado, uma das fases mais relevantes para o trabalho de criação dos algoritmos é a capacidade em lidar com a ausência de dados e/ou dados que estejam incompletos. Para lidar com isso, é recomendável a utilização destas técnicas para identificar e minimizar estes problemas (C, M., BARNES, C., ARCHER, D., HOGG, B., & BRADLEY, P., 2015). Assim, uma eliminação de dados inconsistentes, repetidos ou com algum outro tipo de problema é encontrado.

Para realizar tal análise, foi utilizada uma amostra da base de dados original, sendo, portanto, uma base de menor volume que tende a facilitar sua compreensão. No meio do caminho, alguns problemas destes tipos foram encontrados. Ao carregar dados do site sofifa.com, por exemplo, campos duplicados com o nome do jogador foram identificados. Sendo assim, um deles não agrega nenhum valor à base e fora excluído, consistindo em um típico cenário de informação duplicada. Também no mesmo site, a página acabava por repetir o último jogador da página anterior, duplicando os valores presentes. Este tipo de problema é geralmente de fácil detecção e também fácil tratamento.

No meio da tratativa, a ausência de alguns valores foi detectada. Valores nulos se enquadram na categoria de quadros de inconsistência. Os campos com este problema diziam respeito ao valor e salário semanal do jogador. Este fato não foi de tanta importância, pois, como nenhum outro trabalho estudado utilizou destas variáveis, estes campos não foram considerados com grande significado para o contexto do problema.

Já no caso dos eventos dos jogos reais, a maior dificuldade foi para relacionar essas informações com as do jogo virtual – ou seja, o cruzamento das informações do site sofifa.com com os coletados através da API. Para fazer essa relação, algum campo, tanto do time quanto do jogador, deve estar de acordo com as bases do jogo. Para isso, foi considerado o nome do time, escrito por completo, comum em ambas as fontes. As chaves correlacionaram-se de forma satisfatória, com o único problema de acentuação, o que foi tratado prontamente no código.

4.4 Engenharia de variáveis

Esta etapa designa-se pela criação de novas variáveis a partir das variáveis que já existem dentro do nosso mundo de dados. Estas novas variáveis criadas devem mostrar quais relacionamentos são necessários para o casamento das informações.

Aqui, como o objetivo é a previsão de resultados de uma partida de futebol, foram pontuadas todas as variáveis pelo nível de importância seguindo hipóteses criadas a partir de estudos feitos previamente por especialistas do jogo (DVORAK, JIRI & JUNGE; ASTRID & GRAF-BAUMANN; TONI & PETERSON, LARS, 2004) .

Os dados estão todos disponíveis no github (*https://github.com/ferweezer/Monografia-USP*). Como dito anteriormente, há dados que foram coletados do site sofifa.com, a fim analisarmos as variáveis dependentes desta categoria. Também, foram extraídas informações de cada partida através da API da football.uk do campeonato inglês nas temporadas 17/18, 18/19, 19/20. Abaixo, detalhadas as informações geradas:

Figura 8 - Descrição dos atributos e seus pesos

Atributo	Descrição	Pontuação
Points	Número de pontos conquistados por uma equipe em uma partida.	3
Result	Resultado do jogo	0
Goals	Gol feito pelo time em questão	5
HalfTimeGoals	Gol feito pelo time até o intervalo	3
Corners	Escanteios para o time	2
YellowCards	Número de cartões amarelos distribuidos para o time	1
RedCard	Número de cartões vermelhos distribuidos para o time	2
ShotOnTarget	Finalizações que tomaram a direção do gol, mesmo que tenha sido defendido pelo goleiro	4
ShotOffTarget	Finalização que não rumou ao gol	3
Fouls	Número de faltas pelo time	3
Possession	Percentual do tempo total da partida em que o time teve a posse de bola	4
StadiumName	Nome do estádio do jogo	2
Attendance	Número total de pessoas que compareceram à partida no dia do jogo	2
Reference	Se a equipe era mandante ou não da partida	5
Overall	Média da qualidade do time	4

Fonte Figura 8 - Autor

De início, os dados coletados estavam distribuídos por evento. Ou seja, se uma partida terminou com 3 gols e 5 cartões amarelos, 8 instâncias eram consideradas. Sendo assim, é possível dizer que o conjunto de dados não era o mais adequado, já que o objetivo do estudo é a previsão de o resultado de uma partida de futebol, e não as previsões de quais eventos específicos vão ocorrer – apenas apontar o vencedor é o suficiente.

Primeiramente, optou-se por utilizar os atributos que influenciam de maneira mais clara o resultado de uma partida. Por exemplo, os gols. Porém, já que o objetivo do trabalho visa utilizar outras variáveis, consideramo-las para ver o resultado. Para isso, pontuamos baseado em análises feitas por BABOOTA, R., KAUR, H (2018) E C. PEACE, E. OKECHUKWU (2015), os atributos que faziam mais diferença em um jogo. As demais variáveis foram excluídas do nosso conjunto de dados. Com base nestes atributos, os dados foram agregados para que as análises posteriores pudessem ser feitas.

Além dos campos identificados na tabela acima, estes dados tratados possuem os campos como a identificação da competição, a temporada, o código do jogo e a identificação das equipes.

Este conjunto de dados formado, então, serviu como apoio para a construção de novas variáveis nesta primeira. Pela conexão da API, é possível ter dados de praticamente todos os campeonatos possíveis, porém, para que a base fosse mais enxuta e, consequentemente, mais fácil de entender quais fatores abrangem o jogo no geral. O campeonato inglês foi o escolhido.

4.5 Análise exploratória

As análises exploratórias buscam responder as hipóteses que ajudam na identificação de padrões. Com isso, podemos comprovar que uma consumpção de jogo é verdadeira e utilizar nos modelos preditivos finais.

As hipóteses baseiam-se em entender os seguintes pontos:

- Se uma equipe que joga como mandante do jogo faz mais gols do que quando joga de visitante.
- Se a forma recente do time impacta em sua pontuação
- Se há alguma relação entre a porcentagem de posse de bola em um jogo com a pontuação da equipe.
- Se uma equipe que joga como mandante do jogo faz mais pontos do que quando joga de visitante.
- Se a qualidade da equipe, baseada nos dados do site sofifa.com, acompanha a pontuação final de uma competição.

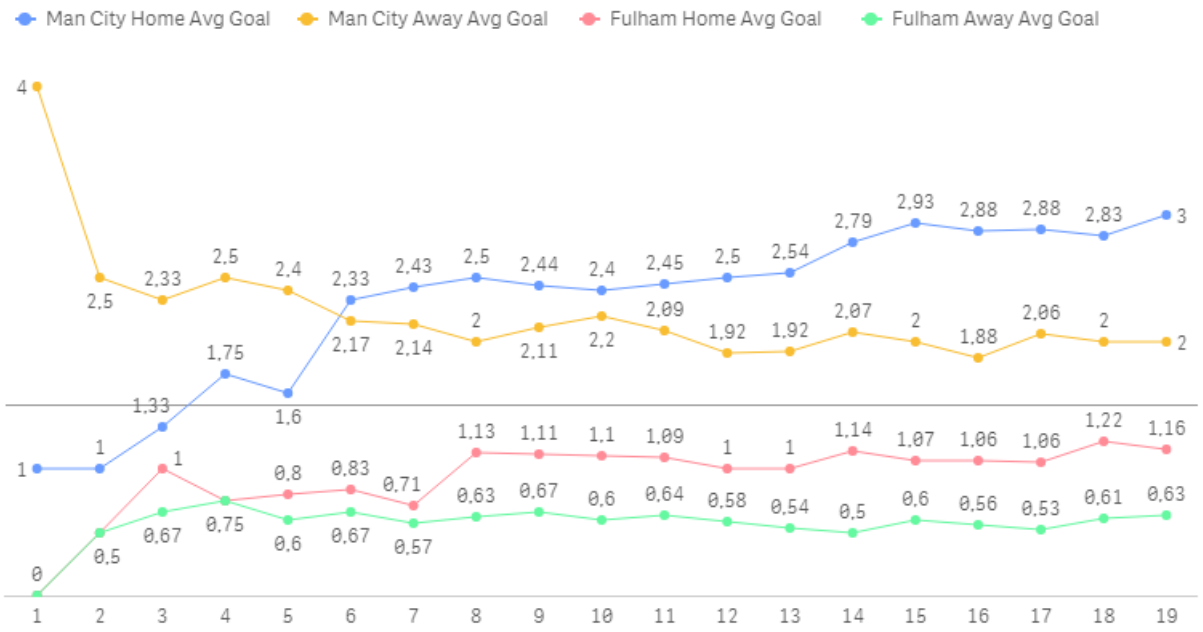
Para comprovar as hipóteses, comparações entre equipes que se posicionaram entre os primeiros e últimos colocados da *Premier League* foram realizadas. Nem sempre os mesmos times foram escolhidos para as comparações, pois o objetivo é tirar a possível casualidade da situação.

4.5.1 Gols marcados e mando de campo

Os dados então foram coletados e tratados a fim de fazer uma análise exploratória que se uniria para nos dar visualizações necessárias para o trabalho. Foi tomado, então, sempre como base de comparação dois times, um que se posicionou bem na tabela de classificação e um que se posicionou mal. Isso faz com que se destaque bem quais as variáveis são relevantes para o nosso modelo.

Para o nosso primeiro exemplo, avaliamos o que chamamos de fator casa. O fator casa é o nome dado à relevância do mando de campo ao resultado da partida. Naturalmente, uma hipótese é a de que o time que joga em seu campo tem mais chances de ganhar a partida. Logo, é preciso validar que isto tem incidência no número de gols de uma equipe.

Figura 9 - Comparação de gols médios entre equipe bem e mal qualificada



Fonte Figura 9 - Autor

Na imagem 9, comparamos o primeiro colocado da temporada 18-19, o Manchester City, com o décimo oitavo, o Fulham. Traçamos duas linhas para cada equipe, uma contendo a média do número de gols marcados pela equipe quando mandante, e outra quando visitante.

O Manchester City quando mandante tem a legenda em azul e o nome *Man City Home Avg Goal*. Quando ele é visitante, ela é laranja e com a descrição *Man City Away Avg Goal*. Já o Fulham, possui a legenda em vermelho quando joga em casa e verde quando fora.

Obviamente, o Fulham teve uma média de gols marcados menor do que a do Manchester City ao longo da competição. Porém, a quantidade de gols marcados quando mandante foi quase o dobro. Com o Manchester City, a diferença demorou um pouco a mais a aparecer, mas veio da mesma forma.

Assim, é possível concluir que um time que joga em casa tem mais chances de marcar gols em casa do que fora.

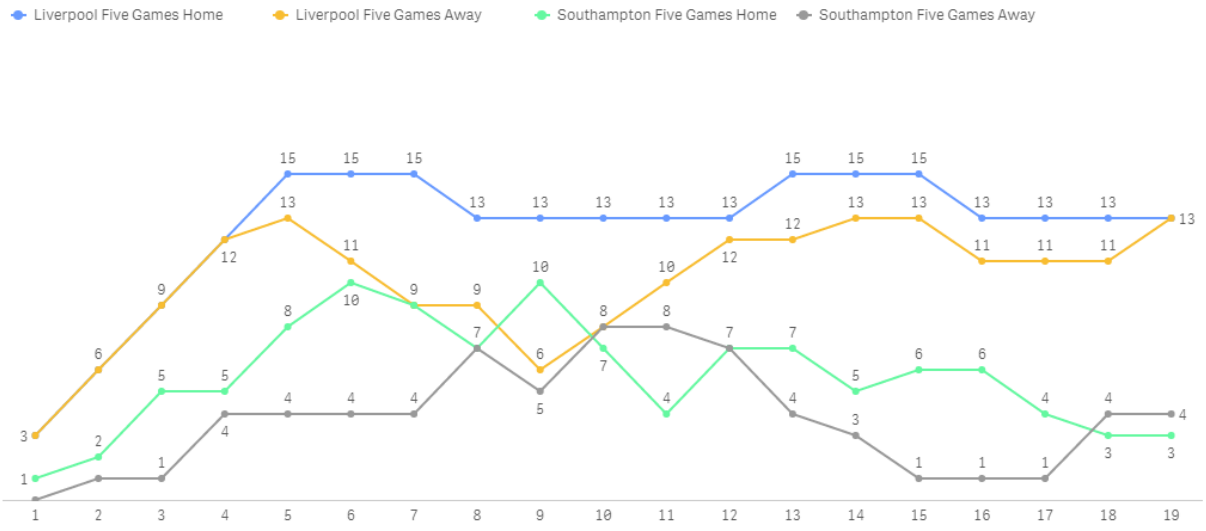
4.5.2 Últimos cinco jogos e mando de campo

Outra variável que é julgada importante é a forma recente de um clube. A forma recente do clube é um valor sobre desempenho que a equipe teve nos últimos 5 jogos.

Um clube que tem ganhado os jogos recentes atinge uma moral mais alta do que um que perde, e isso influencia em como a equipe pratica o futebol. Uma equipe, porém, pode se sentir confiante jogando em casa e menos confiante jogando fora de casa.

A importância desta análise é ver se há uma evolução gradativa do estado da equipe com os resultados futuros – ou seja, uma equipe que perde um jogo tende a seguir a inércia.

Figura 10 - Comparação dos últimos cinco jogos entre equipe bem e mal qualificada



Fonte Figura 10 - Autor

No gráfico representado pela imagem 10, há quatro linhas – duas para uma equipe bem qualificada na temporada de 18-19 e outras duas para uma mal qualificada. As linhas representam o número de pontos conquistados nos últimos 5 jogos da equipe. Cada equipe tem duas linhas, conforme à legenda da foto, uma para jogos em que é mandante e outra para quando é visitante. Cada jogo possui três resultados

possíveis : vitória, o que resulta em 3 pontos ; Empate, que vale 1 ponto ; e derrota, não pontuando.

Na imagem 10, nota-se que o Liverpool, segundo colocado geral, teve uma campanha regular dentro de casa a temporada toda. Já para quando foi visitante, houve uma oscilação no meio da tabela. É possível também demonstrar que, na maioria dos casos, a equipe quando a equipe cai, tende a permanecer no movimento e, quando sobe, também.

O Southampton, que acabou por ser rebaixado na temporada regular, teve um desempenho instável. Porém, o mesmo ritmo é verificável – suas quedas de desempenho vieram acompanhadas de outras quedas também sequenciadas.

Em ambas as situações as equipes possuiram campanhas, na média, melhor quando jogando como mandante do que fora.

4.5.3 Pontuação e qualidade do time

Aqui, a intenção é verificar se há correlação entre a qualidade do time, baseada no jogo de simulação FIFA, produzido pela EA Sports, com a classificação e resultados do time.

Para isso, não poderíamos tomar as estatísticas de todos os jogadores do time, pois lá haveriam muitos jogadores que são reservas e participam pouco dos jogos. Assim, foi feito, anteriormente, uma triagem a fim de trazer apenas os 15 jogadores com as melhores pontuações médias de qualidade.

Desta forma, garantimos que os melhores jogadores são os considerados para análise. Assim, uma análise de pontuação da equipe contra a qualidade média do time foi efetuada para ver a correlação.

Abaixo, a tabela com os nomes das equipes e a qualidade dele, baseada no jogo FIFA.

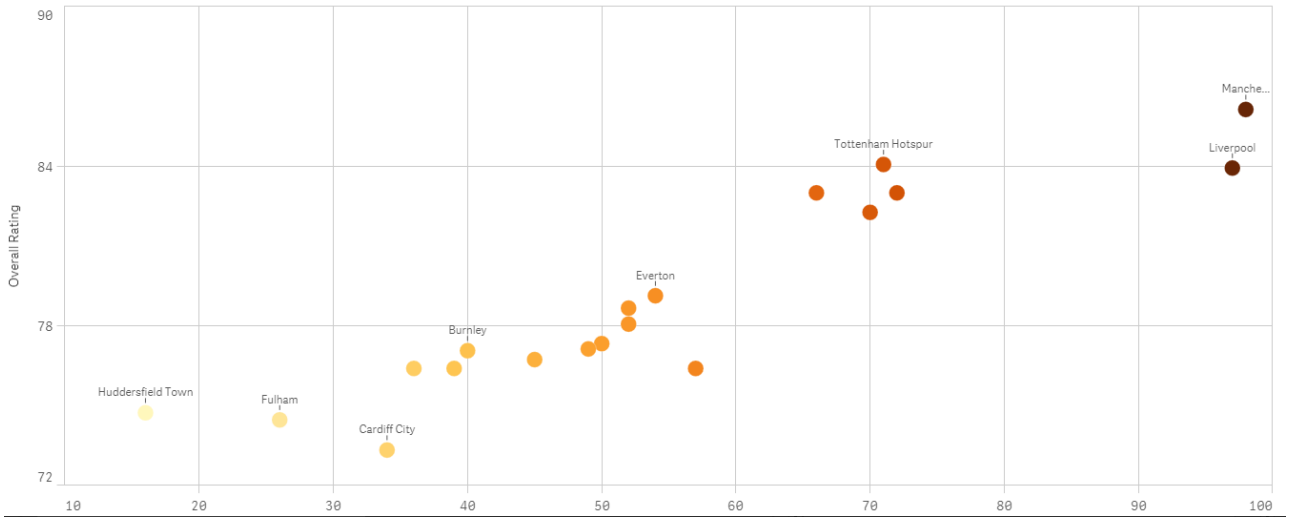
Tabela 1 - Relação entre pontuação e qualidade do elenco

Equipe	Pontos	Qualidade do time
Manchester City	98	86,1
Liverpool	97	83,9
Chelsea	72	83,0
Tottenham Hotspur	71	84,1
Arsenal	70	82,3
Manchester United	66	83,0
Wolverhampton Wanderers	57	76,4
Everton	54	79,1
West Ham United	52	78,7
Leicester City	52	78,1
Watford	50	77,3
Crystal Palace	49	77,1
Newcastle United	45	76,7
Burnley	40	77,1
Southampton	39	76,4
Brighton & Hove Albion	36	76,4
Cardiff City	34	73,3
Fulham	26	74,5
Huddersfield Town	16	74,7

Fonte Tabela 1 - Autor

Fazendo então uma correlação de matrizes entre os pontos conquistados e a qualidade do elenco, temos um índice de 0,92. Considerando que, quanto mais próximo de 1, maior a correlação, é fato que os dois pontos estão correlacionados, tendo, na maioria das vezes, o time com maior qualificação também um número maior de pontos.

Figura 11 - Gráfico de dispersão associando qualidade da equipe com pontuação



Fonte Figura 11 - Autor

No gráfico representado pela imagem 11, cada bolha dentro do quadro é uma equipe. O eixo Y é a qualidade média da equipe com base nos dados do Sofifa.com. Já o eixo X é a quantidade de pontos nesta edição da *Premier League*, de 2018/2019.Quanto mais escura é a cor da bolha, maior sua pontuação e qualidade.

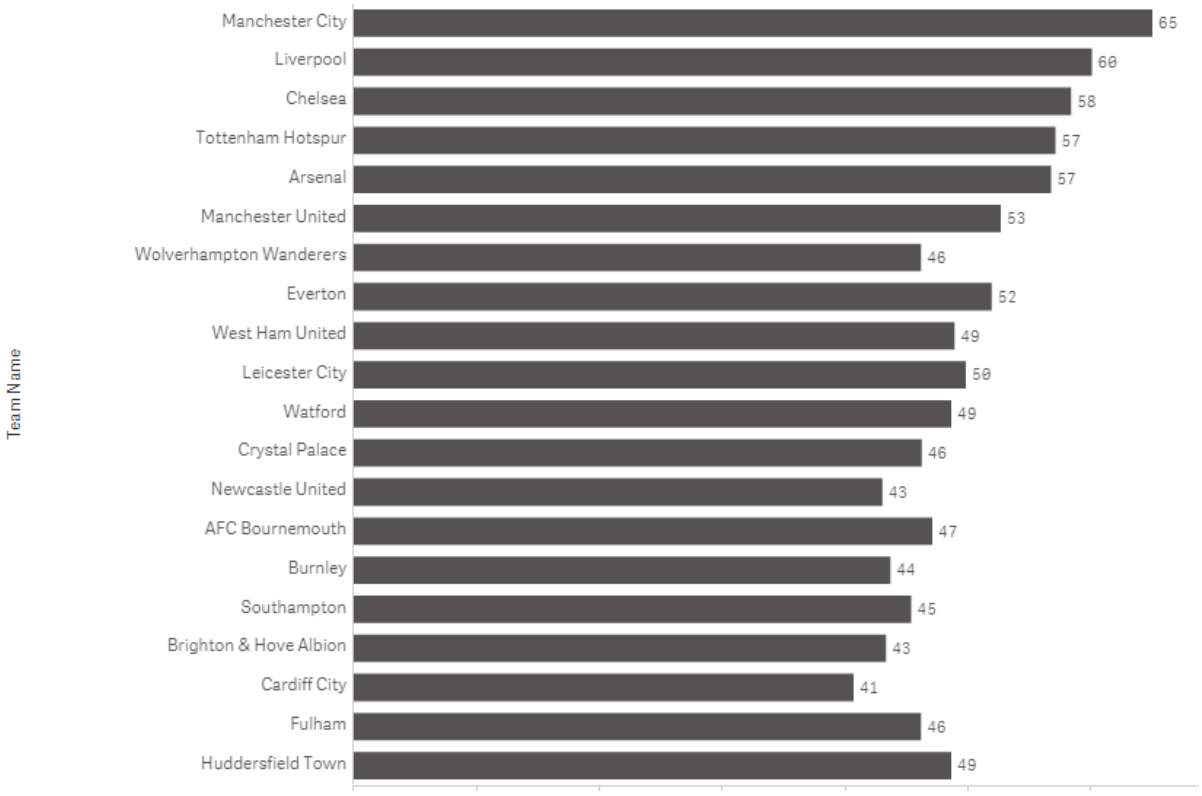
Verificamos que há, sim, uma linha progressiva clara que delimita que quanto melhor a equipe, mais bem posicionada ela ficou na temporada.

4.5.4 Pontuação e posse de bola média

Uma variável importante para o jogo também é em relação à posse de bola que uma equipe possui ou tende a possuir dentro da partida. É razoável pensar que uma equipe que fica mais tempo com a bola tende a criar mais chances de gol e, portanto, ganhar mais jogos.

Esta análise associa a posição final das equipes da primeira divisão inglesa com a média de posse bola.

Figura 12 - Gráfico associando posse de bola com posição na liga



Fonte Figura 12 – Autor

O gráfico acima está ordenado pela posição da equipe, ou seja, o campeão da temporada de 18-19, Manchester City, é a primeira equipe mostrada no gráfico, enquanto o último, Huddersfield Town, o último. Cada barra demonstra a média de posse de bola que a equipe em questão teve no campeonato.

Analisando o gráfico, é verifica que os seis primeiros colocados também foram as seis equipes que possuiram mais a bola durante o campeonato, seguindo, inclusive, uma ordem decrescente. Já na segunda metade do gráfico observa-se que nem sempre a equipe que possuiu mais a bola obteve uma posição mais alta.

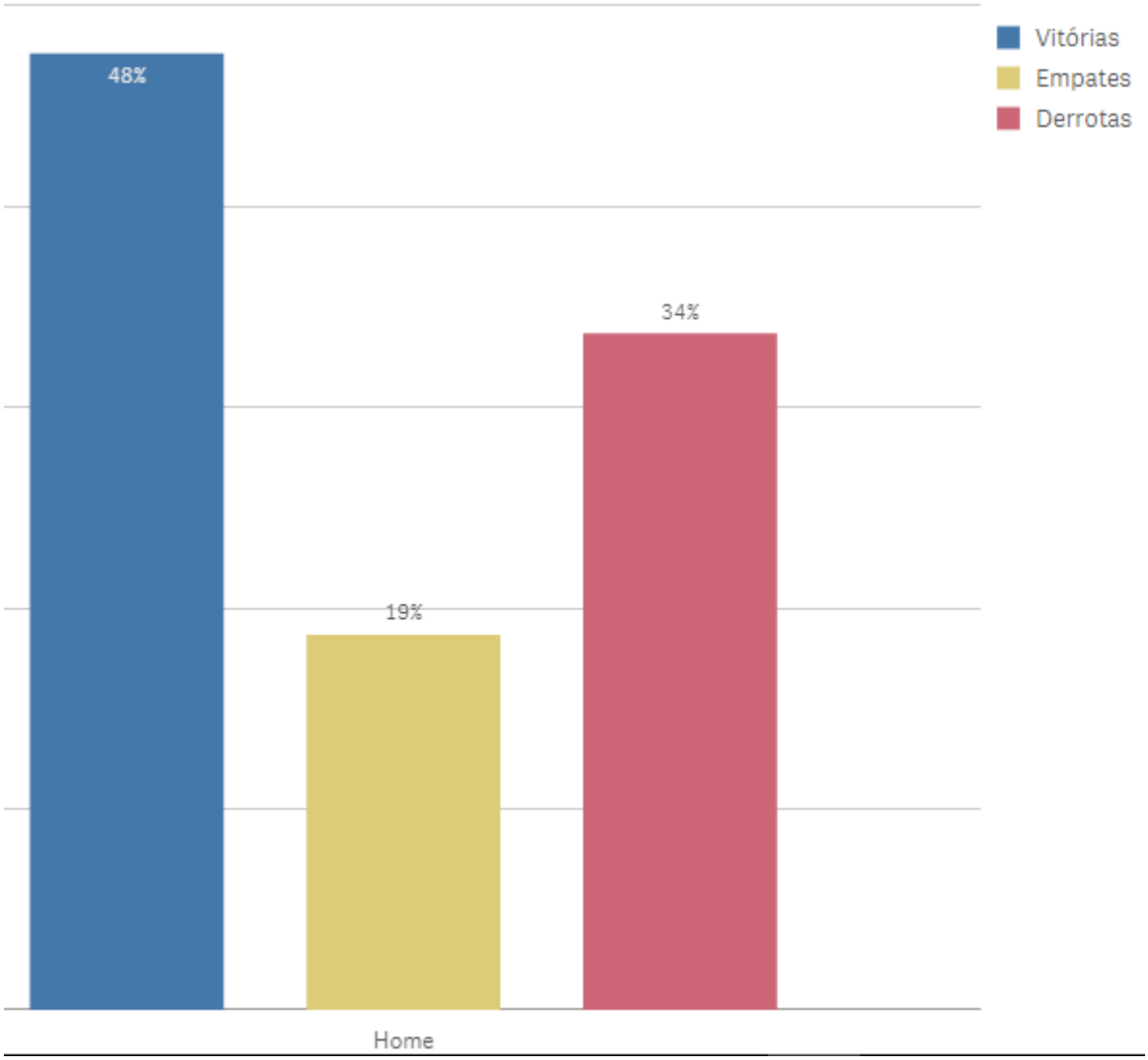
Ainda que a correlação entre posse de bola e resultado não seja tão forte quanto outras variáveis, ela deve ser levada em consideração nas análises feitas.

4.5.5 Mando de campo

O mando de campo representa em qual local a equipe jogará. Por especialistas (AALBERS, BART; VAN HAAREN, JAN, 2018), esse sempre foi considerado como um fator altamente determinante no jogo. Uma equipe que joga no seu estádio, está habituada aos processos, ao ambiente e tem a torcida ao seu favor.

Portanto, a análise a seguir trata relação entre o mando de campo com o número de vitórias e derrotas.

Figura 13 - Vitórias, empates e derrotas para times mandantes



Fonte Figura 13 - Autor

No gráfico acima, vemos os números gerais para todas as equipes nos jogos como mandantes. Há uma grande diferença de vitórias e derrotas para a equipe mandante.

Porém, é necessário ampliar a análise com base de equipe à equipe.

Figura 14 - Análise de desempenho de mandantes e visitantes por posição

TeamName	Reference / Valores					
	Away			Home		
	Vitórias	Empates	Derrotas	Vitórias	Empates	Derrotas
Manchester City	74%	11%	16%	95%	0%	5%
Liverpool	68%	26%	5%	89%	11%	0%
Chelsea	47%	16%	37%	63%	32%	5%
Tottenham Hotspur	58%	0%	42%	63%	11%	26%
Arsenal	37%	21%	42%	74%	16%	11%
Manchester United	47%	16%	37%	53%	32%	16%
Wolverhampton Wanderers	32%	26%	42%	53%	21%	26%
Everton	26%	26%	47%	53%	21%	26%
West Ham United	32%	16%	53%	47%	21%	32%
Leicester City	37%	21%	42%	42%	16%	42%
Watford	32%	26%	42%	42%	16%	42%
Crystal Palace	47%	11%	42%	26%	26%	47%
Newcastle United	21%	42%	37%	42%	5%	53%
AFC Bournemouth	26%	5%	68%	42%	26%	32%
Burnley	21%	26%	53%	37%	11%	53%
Southampton	21%	21%	58%	26%	42%	32%
Brighton & Hove Albion	16%	21%	63%	32%	26%	42%
Cardiff City	21%	11%	68%	32%	11%	58%
Fulham	5%	11%	84%	32%	16%	53%
Huddersfield Town	5%	21%	74%	11%	16%	74%

Fonte Figura 14 – Autor

A tabela acima nos mostra a porcentagem de vitórias, empates e derrotas de cada time participante da primeira divisão inglesa. A ordem da tabela está disposta com relação a classificação delas na liga de futebol.

É possível, a partir daí, então, dar grande valor ao mando de campo. Apesar de que as equipes que foram mal classificados na liga possuem, de fato, um índice baixo de vitórias mesmo quando têm o mando de campo, é visível a diferença quando comparamos com o índice de fora de casa. Se tomarmos o Fulham como exemplo, vemos que a equipe conquistou 5% de vitórias quando jogaram fora de casa. Já quando mandou no jogo, a taxa é de 32%.

De todas as 20 equipes da divisão, apenas uma possuiu um comportamento diferente. O Crystal Palace obteu 47% de vitórias quando jogou fora de casa e apenas 26% quando jogou no seu estádio. Este é um fator relevante, pois nos mostra que, apesar de ser claro a vantagem que uma equipe possui quando é mandante do jogo, há exceções.

Outro ponto interessante é notar que a utilização apenas desta variável de mando de campo não é eficiente. O ideal seria uma conciliação com outros fatores, como a força e qualidade da equipe.

5 ANÁLISE DOS DADOS PROPOSTA PARA OS JOGOS DE FUTEBOL

O capítulo atual será dividido em três seções, conforme descrito no índice. Metodologia, resultados e discussão geral.

5.1 Aplicação da metodologia

Nesta etapa, foram utilizados 4 conjuntos de dados que se diferem entre si apenas nas variáveis que foram utilizados.

Os conjuntos de dados são constituídos com a combinação dos atributos descritos e detalhados no capítulo 4. Também já dito nesta mesma seção, esta etapa deu-se utilizando a primeira divisão inglesa na temporada de 18-19. Nesta competição, foram disputadas 240 partidas de futebol distribuídas em 30 rodadas.

Os jogos transmitem informações inerentes ao jogo, as equipes que participaram dos jogos e o confronto por si. Todas as informações utilizadas para os testes foram tomadas como base os jogos anteriores a eles. Logo, as cinco primeiras rodadas foram excluídas da análise, pois pouco se sabia para a análise. Assim, dos 240 jogos, 40 foram excluídos. Outra exclusão foi feita para aqueles times que não participaram no ano interior da liga. O motivo também foi a falta de informação relevante. Uma equipe pode ter tido resultados ótimos nas divisões inferiores; isso, porém, não garante que ela terá resultados bons na primeira divisão. Com isso, retirou-se outros 48 jogos do conjunto de dados, passando de 200 para 152 jogos.

5.2 Algoritmos

Os modelos escolhidos para a previsão dos jogos de futebol foram descritos na seção 2.5 deste trabalho.

5.2.1 Correlação das variáveis

Uma análise de regressão logística foi, então, gerada. Na tabela 2, os resultados expostos, considerando apenas as variáveis mencionadas e detalhadas aqui neste trabalho

Primeiramente, foram tomados todos os conjuntos de dados possíveis e foi realizada uma análise básica de correlação entre pontos da equipe e qual a correlação entre as variáveis. Utilizamos, para isso, as rodadas 5, 10, 15, 20, 25, 30 e 35.

Os resultados estão expostos na tabela abaixo.

Tabela 2 - Relação entre as variáveis respostas

Rodada	Últimos 5 jogos	Gols acumulado	Posse de bola média acumulada	Qualidade média (sofifa.com)
5	1,00	0,90	0,78	0,77
10	0,91	0,91	0,77	0,87
15	0,74	0,93	0,84	0,91
20	0,74	0,94	0,83	0,91
25	0,72	0,96	0,84	0,94
30	0,74	0,96	0,82	0,94
35	0,75	0,97	0,85	0,94

Fonte Tabela 2 - Autor

Vemos, portanto, que existe forte correlação entre os atributos quando individualizados para com os pontos.

A pontuação dos últimos 5 jogos mostrou uma correlação média de 0,75, já que, nas primeiras rodadas, ela representa uma quantidade muito alta dos jogos. A posse de bola média, gols acumulados e a qualidade média da equipe foram fatores que se demonstraram crescentes na análise, ou seja, conforme o campeonato avança, mais segurança na correlação.

5.2.2 Regressão logística

Para ter-se acesso às estatísticas de regressão logística, utilizamos a ferramenta Python como apoio. Com a regressão, uma análise foi feita juntando todos os atributos descritos na seção 4.2.1, com a inclusão da variável de mando de campo, já que, aqui, a análise fora feita rodada à rodada.

Tabela 3 - Resumo dos resultados obtidos

RESUMO DOS RESULTADOS

Estatística de regressão	
R múltiplo	0,985039324
R-Quadrado	0,97030247
R-quadrado ajustado	0,97005326
Erro padrão	3,533763154
Observações	722

Fonte tabela 3 - Autor

Um R múltiplo, Quadrado ou quadrado ajustado alto significa em um alto índice de adaptabilidade do modelo às amostras.

Tabela 4 - Resumo dos resultados de regressão

	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*
Interseção	- 18,80	4,58	- 4,11	0,00
Últimos cinco jogos	0,30	0,05	6,45	0,00
Mando de campo	- 0,05	0,26	- 0,18	0,85
Gols acumulados	0,99	0,01	120,93	-
Posse de bola	- 13,35	3,73	- 3,58	0,00
Qualidade da equipe	0,30	0,08	3,95	0,00

Fonte tabela 4 - Autor

Finalmente, realizando os testes com regressão linear, chegamos aos resultados das tabelas acima. Vemos um R quadrado muito próximo de 1 – o máximo possível.

O que demonstra que as variáveis, quando analisadas em conjunto, são, de fato, fortes.

5.2.3 Aplicação ao modelo de dados

Nesta primeira interação, foram construídos quatro conjuntos de dados diferentes que já foram descritos neste trabalho. Estes conjuntos foram testados com 5 algoritmos para, enfim, atingir a melhor relação entre os conjuntos e os algoritmos.

Os algoritmos utilizados foram escolhidos por base na facilidade de uso. Algumas, como as redes neurais, embora efetivas, são de complexa aplicação e fogem do conhecimento teórico básico da área.

Na primeira análise, foi realizada uma equivalência da efetividade quando aplicadas juntamente às estratégias de evolução dos dados *Growing Window* e *Sliding Window*. Foram, então, comparados estes 5 algoritmos utilizando separadamente as duas estratégias.

Tabela 5 - Resultados obtidos com *Growing* e *Sliding Window*

	K-Means	Clustering (Prim)	Regras associativas (Apriori)	Regressão logística	Random Forest
Growing Window	0.45	0.433	0.423	0.455	0.532
Sliding Window	0.42	0.421	0.461	0.452	0.510

Fonte tabela 5 - Autor

Na tabela acima, é possível observar o desempenho de cada um dos 5 algoritmos em cada uma das estratégias *Growing Window* e *Sliding Window*. Realizando esta análise, nota-se que a estratégia *Growing Window* tem melhor desempenho em 4 algoritmos (*K-means*, *Prim*, Regressão logística e *Random Forest*) e o *Sliding Window* em apenas um (*Apriori*).

Uma vez realizada a análise, decidiu-se, então, a utilização por completo da estratégia *Growing Window* no estudo. Ainda analisando a tabela, vemos que o algoritmo *Random* Forest obteve o desempenho mais aceitável. Feita esta análise, foi decidido que a estratégia *Growing Window* mostrou-se mais eficaz neste estudo.

6 CONCLUSÃO

Este estudo teve como objetivo analisar abordagens de Data Mining na tentativa de prever resultados de jogos de futebol reais utilizando como principal variável. O CRISP-DM foi a metodologia de Data Mining adotada. Foram feitas, em linhas gerais, duas interações de todas as fases da metodologia. A abordagem CRISP-DM fez com que todas as etapas do projeto fossem bem estruturadas e planejadas, tendo, consequentemente, um melhor entendimento do problema a ser resolvido.

As fontes de dados utilizados foram duas – uma para capturar dados de jogos reais e outra para os dados do simulador de futebol, FIFA. Para o primeiro, a API aberta do site Football.api foi usada, enquanto para o segundo foi feito um *webscrapping* do site Sofifa.com. Estes dados são mantidos atualizados por terceiros em uma periodicidade semanal.

Logo assim que os dados foram coletados foi feito, primeiramente, uma limpeza nos dados para que, então, fosse possível obtermos as variáveis principais para a criação dos algoritmos.

As variáveis escolhidas foram tomadas com base em estudos do futebol, todos detalhados no trabalho. Então, para confirmarmos as hipóteses, um estudo de correlação fora feito, onde conclui-se que as variáveis eram decentes para as análises.

Com as variáveis já separadas, foi possível, então, construir os algoritmos desejados. O modelo com melhor taxa de acerto foi o *Random Forest*, obtendo um resulto de 59% dos testes. Se comparado com outras taxas já estudadas, o resultado não é positivo, porém, não deixa de ser interessante.

O trabalho contou com a inovação de utilizar dados de um simulador de jogo, o FIFA, que é jogado por milhares de pessoas ao redor do mundo e conta com profissionais especializados para qualificar jogadores ao redor do mundo. Desta forma, garantimos que estamos utilizando outros serviços confiáveis como base de análises e conhecimento.

6.1 Contribuições do trabalho

O trabalho contribui para os estudos na área de Data Mining, além de deixar espaço para evoluções futurasa. Há possibilidade de melhorar os resultados aqui obtidos.

O futebol é o esporte mais popular do mundo e atrai grande interesse da sociedade. Este estudo, além de explorar técnicas de mineração de dado e análise de dados, também tem como objetivo e pode ser utilizado para entender outros aspectos que interferem numa partida de futebol.

6.2 Trabalhos futuros

A realização deste estudo e a análise dos resultados faz com que seja possível a evolução dos modelos criados e desenvolvidos. O modelo e todos os dados foram disponibilizados no site github.com (*https://github.com/ferweezer/Monografia-USP*). Aqui, foram abordados problemas de previsão de resultados de partidas de futebol considerando as três hipóteses possíveis – vitórias, empates e derrotas.

O principal ponto de inovação que pode ser abordado posteriormente é o fato de termos utilizado dados provenientes de um jogo de video game para qualificar os jogadores de uma equipe. Estudos nesse âmbito podem ser aprofundados, utilizando não somente uma média do conjunto dos jogadores, mas como os atributos individualizando, mesclados e agrupados de diferentes maneiras a fim de encontrar um ponto de convergência mais aderente ao entendimento do jogo.

REFERÊNCIAS BIBLIOGRÁFICA

P. Bunker,Rory; Thabtah, Fadi. A machine learning framework for sport result prediction. **Applied Computing and Informatics,** v.1, p. 1 - 7. 2017

Kampakis, Stylianos; Adamides, Andreas. Using Twitter to predict football outcomes. **The Telegraph.** 2014

Ulmer, B.; Fernandez, M. Predicting Soccer Match Results in the English Premier League. **Ph.D. dissertation**, 2013.

Joseph, A.; E. Fenton, N; Neil, M. Predicting football results using Bayesian nets and other machine learning techniques. **Knowledge-Based Systems**, v. 19, pp. 544–553, 2006.

Aalbers, Bart; Van Haaren, Jan. Distinguishing between roles of football players in play-by-play match event data. **arXiv**:1809.05173. 2018

C, M., Barnes, C., Archer, D., Hogg, B., & Bradley, P. (2015). Evolution of match performance parameters for various playing positions in the English Premier League. Human Movement Science, 39, 1-11

Mauricio Murad, Football and Society in Brazil, Konrad-Adenauer-Stiftung e.V. International Reports, Berlin, Aug. 25, 2006.

Baboota, R., Kaur, H., Predictive analysis and modelling football results using machine learning approach for English Premier League. **International Journal of Forecasting** (2018)

Bhattacharyya S, Jha S, Tharakunnel K, and Westland JC (2011) Data mining for credit card fraud: A comparative study. **Decision Support Systems** 50, 602-13

Owramipur, Farzin & Eskandarian, Parinaz & Mozneb, Faezeh. (2013). Football Result Prediction with Bayesian Network in Spanish League-Barcelona Team. International Journal of Computer Theory and Engineering. 812-815. 10.7763/IJCTE.2013.V5.802.

Asur, S., & Huberman, B. a. (2010). Predicting the Future with Social Media. Computers and Society; Physics and Society. doi:10.1016/j.apenergy.2013.03.027

Rue, H., & Salvesen, O. (2000). Prediction and retrospective analysis of soccer matches in a league. Journal of the Royal Statistical Society: Series D, 49(3), 399–418.

J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann Publishers, 2011.

Peace and E. Okechukwu, “**An Improved Prediction System for Football a Match Result**,” vol. 04, no. 12, pp. 12–20, 2014.

A. Martins and A. Uff, “**SIMULAÇÕES DE RESULTADO PARA O CAMPEONATO BRASILEIRO DE 2008 COM BASE EM MODELOS LOGITO**,” 2009.

J. Hucaljuk and A. Rakipovic, “Predicting football scores using machine learning techniques,” 2011 Proceedings of the 34th International Convention MIPRO, vol. 48, pp. 1623– 1627, 2011

P. Chapman, Julian Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, “CRISP-DM 1.0: Step-by-step data mining guide,” Tech. Rep., 2000.[19] A. Tsakonas and G. Dounias, “Soft computing-based result prediction of football games,” The First International Conference on Inductive Modelling ICIM’2002, vol. 3, no. May, pp. 15–21, 2002.

S. Navega, “Princípios Essenciais do Data Mining,” Anais de Infoimagem, Cenadem, 2002.

CHAPMAN, P. CRISP-DM 1.0: Step-By-Step Data Mining Guide. [S.I.]: 2000. Disponível em: . Acesso em: 28 jan. 2018.

Dvorak, Jiri & Junge, Astrid & Graf-Baumann, Toni & Peterson, Lars. (2004). Football is the most popular sport worldwide. The American journal of sports medicine. 32. 3S-4S. 10.1177/0363546503262283.