MINERAÇÃO DE DADOS EM APOSTAS ESPORTIVAS

Matheus Percário Bruder, Matheus Rosisca Padovani, Kevin Barrios, Cristiano José Furlan

RESUMO

A pesquisa foi executada a fim de encontrar regras de associação que deem suporte para apostadores de jogos de futebol apostarem em diversas partidas durante uma temporada. Para alcançar o objetivo, foi utilizada uma base de dados que continha informações estatísticas pertencentes a jogos da Premier League, isto é, a liga inglesa de futebol. A base de dados foi submetida ao software Weka, desenvolvido na Universidade de Waikato na cidade de Hamilton na Nova Zelândia, o qual foi responsável pela mineração de dados, ou seja, fornece as regras de associação que darão suporte aos apostadores.

INTRODUÇÃO

a) O que é mineração de dados?

O processo de minerar é um dos trabalhos mais antigos existentes no planeta Terra. Esse processo consiste basicamente encontrar algo precioso, em meio a uma imensidão de lixo (coisas comuns, não preciosas). O exemplo clássico de mineração é o garimpo, em que, centenas de pessoas despendem horas peneirando pedras a fim de encontrar pequenas quantidades de metais preciosos. Portanto, percebe-se que a mineração é um processo árduo, pois, deve-se empregar bastante trabalho e dedicação para encontrar aquilo que realmente importa.

A mineração de dados, ideologicamente, é realizada da mesma forma que a mineração de pedras preciosas. Contudo, minerar dados é a maneira mais eficaz para extrair o conhecimento e as informações a partir de grandes volumes de dados, descobrindo relações intrínsecas, padrões e gerando regras para predizer dados [1][2]. Então, ao final da mineração de dados, as informações relevantes e o conhecimento adquirido serão utilizados com o intuito de auxiliar a tomada de decisão [1]. A mineração de dados traz inúmeros benefícios consigo, porém, para alcançar o objetivo, isto é, extrair conhecimento e informações, pertinentes, em meio a massa de dados, é preciso conhecer algumas técnicas.

Para obter um bom resultado ao minerar dados, é preciso que os mesmos estejam dispostos em um único arquivo e distribuídos em objetos (conhecidos também como registros) e atributos, que são, respectivamente, linhas e colunas. Vale ressaltar que o arquivo em que estão os dados é chamado de base de dados.

b) Tarefas de mineração de dados

Após entender a maneira com que os dados devem ser distribuídos, é necessário conhecer os principais tipos de tarefas da mineração de dados. Cada tarefa possui um método diferente para realizar a mineração [2]. Uma boa analogia seria um cenário em que cada tarefa da mineração de dados seria correspondente a uma variante do método de lavra subterrânea da mineração de pedras preciosas, pois, cada variante possui um método diferente, contudo, todos os métodos devem alcançar o mesmo objetivo, a extração dos metais preciosos. As três principais tarefas da mineração de dados, são:

i) Tarefa de Agrupamento

A tarefa de agrupamento envolve algoritmos que buscam similaridades entre os registros ou objetos, então, o objetivo é fazer com que os registros similares estejam dispostos em um mesmo grupo (*cluster*) [2]. A similaridade em questão é dada por meio da distância entre os objetos. Tal distância é determinada pelo algoritmo, como por exemplo, o algoritmo de agrupamento K-MEANS utiliza a distância euclidiana para determinar o nível de similaridade.

Essa é uma tarefa que não necessita que os registros sejam rotulados (tarefa não supervisionada). Contudo, a desvantagem existente é que encontrar o melhor agrupamento para um conjunto de objetos não é simples quando há um grande número de objetos e grupos.

ii) Tarefa de Classificação

A tarefa de classificação possui como principal objetivo gerar um modelo capaz de predizer uma saída para novos registros, ou seja, registros cuja classe ou valor de saída são desconhecidos [2]. A diferença entre a tarefa de agrupamento para a tarefa de classificação é dada pelo fato de que ao classificar deve haver um atributo a mais, o qual é denominado como classe ou rótulos do objeto (tarefa supervisionada). Então, a partir dos rótulos, os algoritmos de classificação encontrarão regras para elaborar o modelo preditivo.

Para gerar o modelo preditivo é necessário dividir a base de dados em duas partes. A primeira parte será o treinamento, que engloba todo o processo de elaboração do modelo preditivo. A segunda parte da base de dados será o teste, isto é, a avaliação do modelo preditivo gerado durante o treinamento.

iii) Tarefa de Associação

A tarefa de associação difere das tarefas de mineração de dados anteriores, agrupamento e classificação, pois a base de dados utilizada é transacional, ou seja, cada um dos registros é chamado de transação.

É uma das tarefas mais conhecidas da mineração de dados, seu principal objetivo é encontrar regras de associação com base nos itens mais frequentes dentre os registros, isto é, a tarefa de associação permite obter resultados do tipo: "*SE compra leite e pão TAMBÉM compra manteiga*" [2]. Então, com base nas regras de associação encontradas são extraídas informações importantes, que serão responsáveis pelo auxílio à tomada de decisão. Contudo, o apoio a decisão proveniente das regras de associação não é dado de qualquer maneira, é preciso que um analista avalie individualmente cada uma das regras geradas, para então dizer se aquela regra é relevante ou não.

Para realizar essa análise é preciso conhecer dois conceitos chaves, o suporte e a confiança. O suporte retrata a quantidade de vezes que a regra aparece dentre a base de dados, logo, é uma medida quantitativa. Em contrapartida, a confiança retrata a taxa de vezes para qual aquela regra realmente foi verdadeira, logo, é uma medida qualitativa. A confiança é calculada pela razão entre o suporte da regra e o suporte do antecedente da regra.

Existe uma premissa, comprovada na maioria das vezes, que diz que uma boa regra deve ter um baixo suporte e uma alta confiança. Em outras palavras a premissa diz que a regra será boa quando pouco aparecer (baixo suporte), porém, quando aparece é assertiva (alta confiança).

c) Descrição do problema

Nos últimos anos, com o avanço da tecnologia, as redes de comunicações cresceram exponencialmente, principalmente a *internet*, portanto, uma consequência desse crescimento é que a moeda se torna cada vez menos física. Atualmente, praticamente toda transação monetária é feita, em segundos, por meio de um *smartphone*. Logo, a virtualização da moeda é responsável por fornecer a sensação de que todos, e não apenas grandes empresários, podem ser capitalistas, isto é, fazer com que o capital também trabalhe a fim de proporcionar juros.

Além disso, o avanço da *internet* também gerou uma grande quantidade de possibilidades de compra, pois, é possível comprar um produto que está do outro lado do mundo sem problema algum e, além disso, com a *internet* surgiram as compras de produtos virtuais, como por exemplo, vantagens para um determinado jogo. Ainda no campo dos produtos não físicos, pode-se dizer que as apostas também são um ótimo exemplo. Embora os ambientes de jogos de azar sejam ilegais no Brasil, não há nenhuma lei que proíba a prática de apostas ou jogos de azar na *internet*, isso porque quando as leis de proibição foram criadas a internet sequer existia [3][4]. Ao juntar a virtualização da moeda com o entretenimento gerado pelas apostas, bem como, pelo futebol, certamente, a ideia de realizar apostas esportivas surgiria em algum momento.

A partir desse cenário, é possível desenvolver essa pesquisa, a qual tem como objetivo encontrar regras de associação que deem suporte para apostadores de jogos de futebol apostarem em diversas partidas durante uma temporada. Esse suporte será dado por meio da mineração dos dados estatísticos referentes aos jogos de futebol da liga inglesa. As apostas podem ser realizadas em diversos *sites* presentes na *internet*, contudo, no Brasil os mais populares são, *Bet365, Betway Esportes e Betfair*, pois, são *sites* que possuem alta credibilidade, como também, possuem a maior praticidade e facilidade para realizar uma aposta. Por fim, para realizar a mineração de dados, a pesquisa utilizará a tarefa de associação.

A tarefa de associação foi escolhida, porque para encontrar regras que dêem suporte aos apostadores, é necessário encontrar semelhanças, padrões ou repetições. Por exemplo, uma pessoa deseja realizar uma aposta, afirmando que um determinado time que está jogando fora de casa sofrerá mais de cinco escanteios, então, por meio da tarefa de associação e através do suporte e da confiança, é possível indicar se essa é uma regra boa ou não. Portanto, a tarefa de associação será responsável por dar apoio à decisão do apostador. Por fim, nesta pesquisa o algoritmo usado para realizar a tarefa de associação foi o APRIORI.

MATERIAL E MÉTODO

Os dados utilizados para o desenvolvimento da pesquisa são públicos, isto é, estão disponíveis no site *Datahub* [5], o qual fornece uma grande quantidade de base de dados (*Big Data*) relativo a inúmeras áreas. Para realização dessa pesquisa foi utilizado um banco de dados referente às últimas nove temporadas da Premier League, ou seja, são dados que retratam as estatísticas dos jogos referentes aos últimos nove anos da liga inglesa de futebol.

Após a coleta dos dados, deve-se submetê-los ao software Weka 3.8, a fim de encontrar regras de associação que deem para tais apostadores. O software Weka foi desenvolvido por pesquisadores da universidade de Waikato na cidade de Hamilton na Nova Zelândia. O Weka tem como objectivo agregar algoritmos provenientes de diferentes abordagens e paradigmas na sub-área da inteligência artificial dedicada ao estudo de aprendizagem de máquina e procede à análise computacional e estatística dos dados fornecidos recorrendo a técnicas de mineração de dados tentando, indutivamente, a partir dos padrões encontrados gerar hipóteses para soluções e no extremos inclusive teorias sobre os dados em questão [6].

Para o processamento dos dados foi utilizado uma máquina com as seguintes características: Sistema operacional Windows 10 Single Language, processador Intel Core i7-8770K e 16Gb de memória RAM.

A base de dados estava fragmentada em nove arquivos que continham, respectivamente, as informações estatísticas dos nove anos de jogos da Premier League. Cada um dos arquivos possuía 381 registros e 22 atributos.

Field Name	Type (Format)	Description
Date	date (%Y-%m-%d)	Match Date (dd/mm/yy)
HomeTeam	string (default)	Home Team
AwayTeam	string (default)	Away Team
FTHG	integer (default)	Full Time Home Team Goals
FTAG	integer (default)	Full Time Away Team Goals
FTR	string (default)	Full Time Result (H=Home Win, D=Draw, A=Away Win)
HTHG	integer (default)	Half Time Home Team Goals
HTAG	integer (default)	Half Time Away Team Goals
HTR	string (default)	Half Time Result (H=Home Win, D=Draw, A=Away Win)
Referee	string (default)	Match Referee
HS	integer (default)	Home Team Shots
AS	integer (default)	Away Team Shots
HST	integer (default)	Home Team Shots on Target
AST	integer (default)	Away Team Shots on Target
HF	integer (default)	Home Team Fouls Committed
AF	integer (default)	Away Team Fouls Committed
HC	integer (default)	Home Team Corners
AC	integer (default)	Away Team Corners
HY	integer (default)	Home Team Yellow Cards
AY	integer (default)	Away Team Yellow Cards
HR	integer (default)	Home Team Red Cards
AR	integer (default)	Away Team Red Cards

Figura 1 - Atributos de cada uma das bases de dados fornecidas pelo site Datahub.io

a) Pré-processamento

Devido a fragmentação da base de dados, o primeiro passo foi realizar a integração dos dados, isto é, os nove arquivos foram concatenados em apenas um, pois para realizar a mineração de dados é preciso que todas as informações estejam em um único local. Após a integração a base de dados possuía os mesmos 22 atributos, porém, 3429 registros.

A partir disso foi necessário descobrir quais os times que haviam jogado em todas as temporadas, pois haveria uma inconsistência caso todos os times estivessem sido submetidos a mineração de dados, uma vez que, alguns times jogaram muitas vezes e outros poucas. Isso certamente causaria um enviesamento da mineração e então os resultados obtidos não iriam condizer com real problema.

TIMES	2009-10	2010-11	2011-12	2012-13	2013-14	2014-15	2015-16	2016-17	2017-18	SOMA
Arsenal	1	1	1	1	1	1	1	1	1	9
Chelsea	1	1	1	1	1	1	1	1	1	9
Everton	1	1	1	1	1	1	1	1	1	9
Liverpool	1	1	1	1	1	1	1	1	1	9
Manchester City	1	1	1	1	1	1	1	1	1	9
Manchester United	1	1	1	1	1	1	1	1	1	9
Stoke City	1	1	1	1	1	1	1	1	1	9
Tottenham Hotspur	1	1	1	1	1	1	1	1	1	9
Sunderland	1	1	1	1	1	1	1	1	0	8
West Ham United	1	1	0	1	1	1	1	1	1	8
Aston Villa	1	1	1	1	1	1	1	0	0	7
Newcastle United	0	1	1	1	1	1	1	0	1	7
Swansea City	0	0	1	1	1	1	1	1	1	7
Southampton	0	0	0	1	1	1	1	1	1	6
Fulham	1	1	1	1	1	0	0	0	0	5
West Bromwich Albion	0	1	1	1	1	1	0	0	0	5
Crystal Palace	0	0	0	0	1	1	1	1	1	5
Burnley	1	0	0	0	0	1	0	1	1	4
Hull City	1	0	0	0	1	1	0	1	0	4
Wigan Athletic	1	1	1	1	0	0	0	0	0	4
Norwich City	0	0	1	1	1	0	1	0	0	4
Leicester City	0	0	0	0	0	1	1	1	1	4
Blackburn Rovers	1	1	1	0	0	0	0	0	0	3
Bolton Wanderers	1	1	1	0	0	0	0	0	0	3
Wolverhampton	1	1	1	0	0	0	0	0	0	3
Queens Park Rangers	0	0	1	1	0	1	0	0	0	3
Bournemouth	0	0	0	0	0	0	1	1	1	3
Watford	0	0	0	0	0	0	1	1	1	3
West Bromwich	0	0	0	0	0	0	1	1	1	3
Birmingham	1	0	0	0	0	0	0	0	0	1
Portsmouth	1	0	0	0	0	0	0	0	0	1
Birmingham	0	1	0	0	0	0	0	0	0	1
Blackpool	0	1	0	0	0	0	0	0	0	1
Reading	0	0	0	1	0	0	0	0	0	1
Cardiff City	0	0	0	0	1	0	0	0	0	1
Middlesbrough	0	0	0	0	0	0	0	1	0	1
Brighton & Hove Albion	0	0	0	0	0	0	0	0	1	1
Huddersfield Town	0	0	0	0	0	0	0	0	1	1

Figura 2 - Seleção dos times que participaram de todas as nove temporadas da Premier League

A Figura 2 revela que os únicos times que participaram de todas as temporadas analisadas da Premier League foram: Arsenal, Chelsea, Everton, Liverpool, Manchester City, Manchester United,

Stoke City e Tottenham. Logo, a mineração de dados foi realizada baseada nos dados estatísticos referentes aos jogos dos times citados.

De acordo com a Figura 1, que representa os atributos da base de dados original, é perceptível que cada um dos registros é referente a um jogo da temporada. Contudo, além da integração, também foi necessário transformar a base de dados para que cada um dos registros fosse referente a um jogo de um determinado time (como mostrado na Figura 4). Assim sendo, ao invés de existirem atributos referentes ao time de fora e o time de casa, o que passará a diferenciar o local da partida é o domínio do atributo.

Além do mais, foi preciso transformar os domínios dos atributos, os quais deixaram de ser atributos numéricos e passaram a representar um intervalo. Por exemplo, determinado registro tinha como valor do atributo 'escanteio' o número 7, então após a transformação o número 6 agora faz parte do intervalo '6 - 10', pois como os valores eram muito discrepantes, a transformação em intervalo ajudou a encontrar regras de associação mais relevantes.

Nome do campo	Tipo (Formato)	Descrição
Time	Cadeia de caracteres (string)	Nome do time
Local	Cadeia de caracteres (string)	Local do jogo (casa ou fora)
Gols_total	Cadeia de caracteres (string)	Total de gols feitos pelo time
Resultado_total	Cadeia de caracteres (string)	Resultado do time
Gols_parcial	Cadeia de caracteres (string)	Gols marcados pelo time na primeira metade
Resultado_parcial	Cadeia de caracteres (string)	Resultado do time na primeira metade
Juiz	Cadeia de caracteres (string)	Nome do arbitro
Chutes	Intervalo	Intervalo referente ao total de chutes
Chutes_aoGol	Intervalo	Intervalo referente ao total de chutes ao gol
Faltas	Intervalo	Intervalo referente ao número de faltas
Escanteios	Intervalo	Intervalo referente ao número de ecanteios
Cartões amarelos	Cadeia de caracteres (string)	Quantidade de cartões amarelos
Cartões vermelhos	Cadeia de caracteres (string)	Quantidade de cartões vermelhos

Figura 3 - Atributos da nova base de dados após a transformação

A Figura 3 retrata os atributos que serão utilizados para realizar a mineração de dados referente a essa pesquisa. Todos os atributos são relativos aos dados estatísticos sobre um determinado time em uma partida de futebol.

Por exemplo, um jogo entre Liverpool (jogando em casa) e Stoke City (jogando fora), faz referência a dois registros da base de dados, em que, um dos registros relata as estatísticas do time de casa e o outro do time de fora. Logo, se o Liverpool ganhar o jogo, o domínio do atributo '*Resultado_total*' será '*Vitoria_H*', consequentemente, o valor domínio do mesmo atributo para o Stoke City será '*Derrota_A*'. O mesmo vale para o atributo '*Resultado_parcial*' e, por fim, o valor do atributo '*Juiz*' será o mesmo em ambos registros.

b) Base de dados após o pré-processamento

Após o pré-processamento, isto é, com a integração e a todas as transformações dos dados, a base de dados está pronta para ser submetida ao software Weka. Essa base agora possui 2736 registros e 12 atributos.

TIME	GOLS_TOTAL	RESULTADO_TOTAL	GOLS_PARCIAL	RESULTADO_PARCIAL	JUIZ	CHUTES	CHUTES_AOGOL	FALTAS	ESCANTEIOS	AMARELOS	VERMELHOS
Arsenal_H	QUATRO_GOLS_H	Vitoria_H	DOIS_GOLS	Empatando	M Dean_H	25-30_HS	6-10_HST	6-10_HF	6-10_HC	ZERO_HY	ZERO_HR
Arsenal_H	TRES_GOLS_H	Vitoria_H	DOIS_GOLS	Ganhando_H	A Taylor_H	10-15_HS	6-10_HST	10-15_HF	6-10_HC	ZERO_HY	ZERO_HR
Arsenal_H	DOIS_GOLS_H	Vitoria_H	UM_GOL	Ganhando_H	R Madley_H	15-20_HS	6-10_HST	6-10_HF	6-10_HC	UM_HY	ZERO_HR
Arsenal_H	DOIS_GOLS_H	Vitoria_H	UM_GOL	Ganhando_H	K Friend_H	20-25_HS	10-15_HST	6-10_HF	6-10_HC	ZERO_HY	ZERO_HR
Arsenal_H	DOIS_GOLS_H	Vitoria_H	ZERO_GOLS	Perdendo_H	L Mason_H	15-20_HS	0-5_HST	6-10_HF	0-5_HC	ZERO_HY	ZERO_HR
Arsenal_H	DOIS_GOLS_H	Vitoria_H	DOIS_GOLS	Ganhando_H	M Dean_H	10-15_HS	0-5_HST	10-15_HF	6-10_HC	QUATRO_HY	ZERO_HR
Arsenal_H	CINCO_GOLS_H	Vitoria_H	UM_GOL	Ganhando_H	G Scott_H	20-25_HS	6-10_HST	6-10_HF	6-10_HC	ZERO_HY	ZERO_HR
Arsenal_H	UM_GOL_H	Derrota_H	ZERO_GOLS	Perdendo_H	A Marriner_H	30-35_HS	10-15_HST	10-15_HF	10-15_HC	TRES_HY	ZERO_HR
Arsenal_H	UM_GOL_H	Vitoria_H	UM_GOL	Ganhando_H	S Attwell_H	20-25_HS	0-5_HST	10-15_HF	6-10_HC	DOIS_HY	ZERO_HR
Arsenal_H	TRES_GOLS_H	Empate	ZERO_GOLS	Perdendo_H	M Atkinson_H	10-15_HS	0-5_HST	6-10_HF	6-10_HC	UM_HY	ZERO_HR
Arsenal_H	DOIS_GOLS_H	Empate	ZERO_GOLS	Empatando	A Taylor_H	10-15_HS	6-10_HST	10-15_HF	6-10_HC	TRES_HY	ZERO_HR
Arsenal_H	QUATRO_GOLS_H	Vitoria_H	QUATRO_GOLS	Ganhando_H	C Kavanagh_H	15-20_HS	6-10_HST	6-10_HF	0-5_HC	ZERO_HY	ZERO_HR
Arsenal_H	CINCO_GOLS_H	Vitoria_H	QUATRO_GOLS	Ganhando_H	N Swarbrick_H	10-15_HS	6-10_HST	6-10_HF	0-5_HC	DOIS_HY	ZERO_HR
Arsenal_H	ZERO_GOLS_H	Derrota_H	ZERO_GOLS	Perdendo_H	A Marriner_H	6-10_HS	0-5_HST	10-15_HF	6-10_HC	UM_HY	ZERO_HR
Arsenal_H	TRES_GOLS_H	Vitoria_H	UM_GOL	Ganhando_H	M Atkinson_H	10-15_HS	6-10_HST	10-15_HF	0-5_HC	DOIS_HY	ZERO_HR
Arsenal_H	TRES_GOLS_H	Vitoria_H	ZERO_GOLS	Empatando	C Pawson_H	20-25_HS	10-15_HST	6-10_HF	6-10_HC	UM_HY	ZERO_HR
Arsenal_H	TRES_GOLS_H	Vitoria_H	DOIS_GOLS	Ganhando_H	A Marriner_H	10-15_HS	6-10_HST	10-15_HF	6-10_HC	DOIS_HY	UM_HR
Arsenal_H	QUATRO_GOLS_H	Vitoria_H	ZERO_GOLS	Empatando	L Mason_H	15-20_HS	6-10_HST	10-15_HF	6-10_HC	TRES_HY	ZERO_HR
Arsenal_H	CINCO_GOLS_H	Vitoria_H	DOIS_GOLS	Ganhando_H	A Marriner_H	15-20_HS	6-10_HST	6-10_HF	0-5_HC	ZERO_HY	ZERO_HR
Chelsea_H	DOIS_GOLS_H	Derrota_H	ZERO_GOLS	Perdendo_H	C Pawson_H	15-20_HS	6-10_HST	15-20_HF	6-10_HC	TRES_HY	DOIS_HR
Chelsea_H	DOIS_GOLS_H	Vitoria_H	DOIS_GOLS	Ganhando_H	J Moss_H	15-20_HS	6-10_HST	10-15_HF	6-10_HC	DOIS_HY	ZERO_HR
Chelsea_H	ZERO_GOLS_H	Empate	ZERO_GOLS	Empatando	M Oliver_H	10-15_HS	0-5_HST	10-15_HF	0-5_HC	UM_HY	UM_HR
Chelsea_H	ZERO_GOLS_H	Derrota_H	ZERO_GOLS	Empatando	M Atkinson_H	0-5_HS	0-5_HST	6-10_HF	0-5_HC	ZERO_HY	ZERO_HR
Chelsea_H	QUATRO_GOLS_H	Vitoria_H	UM_GOL	Empatando	J Moss_H	10-15_HS	6-10_HST	10-15_HF	6-10_HC	DOIS_HY	ZERO_HR
Chelsea_H	UM_GOL_H	Vitoria_H	ZERO_GOLS	Empatando	A Taylor_H	15-20_HS	6-10_HST	15-20_HF	0-5_HC	UM_HY	ZERO_HR
Chelsea_H	UM_GOL_H	Vitoria_H	ZERO_GOLS	Empatando	N Swarbrick_H	20-25_HS	6-10_HST	6-10_HF	10-15_HC	UM_HY	ZERO_HR
Chelsea_H	TRES_GOLS_H	Vitoria_H	DOIS_GOLS	Ganhando_H	K Friend_H	20-25_HS	6-10_HST	6-10_HF	10-15_HC	ZERO_HY	ZERO_HR
Chelsea_H	UM_GOL_H	Vitoria_H	UM_GOL	Ganhando_H	R East_H	20-25_HS	6-10_HST	6-10_HF	6-10_HC	UM_HY	ZERO_HR
Chelsea_H	DOIS_GOLS_H	Vitoria_H	ZERO_GOLS	Empatando	M Dean_H	20-25_HS	6-10_HST	6-10_HF	10-15_HC	ZERO_HY	ZERO_HR
Chelsea_H	CINCO_GOLS_H	Vitoria_H	TRES_GOLS	Ganhando_H	K Friend_H	20-25_HS	10-15_HST	0-5_HF	6-10_HC	UM_HY	ZERO_HR
Chelsea_H	ZERO_GOLS_H	Empate	ZERO_GOLS	Empatando	M Jones_H	15-20_HS	6-10_HST	10-15_HF	6-10_HC	DOIS_HY	ZERO_HR
Chelsea_H	ZERO_GOLS_H	Derrota_H	ZERO_GOLS	Empatando	L Probert_H	20-25_HS	0-5_HST	0-5_HF	6-10_HC	UM_HY	ZERO_HR
Chelsea_H	TRES_GOLS_H	Vitoria_H	UM_GOL	Ganhando_H	L Mason_H	15-20_HS	6-10_HST	10-15_HF	6-10_HC	ZERO_HY	ZERO_HR
Chelsea_H	DOIS_GOLS_H	Vitoria_H	DOIS_GOLS	Ganhando_H	A Taylor_H	25-30_HS	10-15_HST	6-10_HF	10-15_HC	UM_HY	ZERO_HR

Figura 4 - Parte da base de dados final, após o pré-processamento

RESULTADOS E DISCUSSÃO

A mineração de dados, utilizando a tarefa de associação e a base de dados estatísticos da Premier League, obteve resultados positivos. Foram encontradas diversas regras que proporcionam um bom suporte para aqueles que desejam realizar apostas em partidas de futebol.

A pesquisa utilizou uma base de dados relativamente pequena, isto é, apenas dados estatísticos referente a um único campeonato, a Premier League. Logo, as regras de associação que foram geradas, em sua maioria, são relativas aos times ingleses. Embora as regras sejam limitadas, a pesquisa proporcionou um resultado benéfico, portanto, é possível ampliar ainda mais os horizontes para buscar generalizações das regras, isso por meio de outras pesquisas que envolvam a mineração dos dados estatísticos sobre futebol, abrangendo diversos campeonatos ao redor do mundo.

Portanto, visando obter boas regras, foi preciso configurar a confiança e o suporte mínimo/máximo no software Weka. O programa disponibiliza uma janela (Figura X) que permite configurar as características de processamento do algoritmo APRIORI.

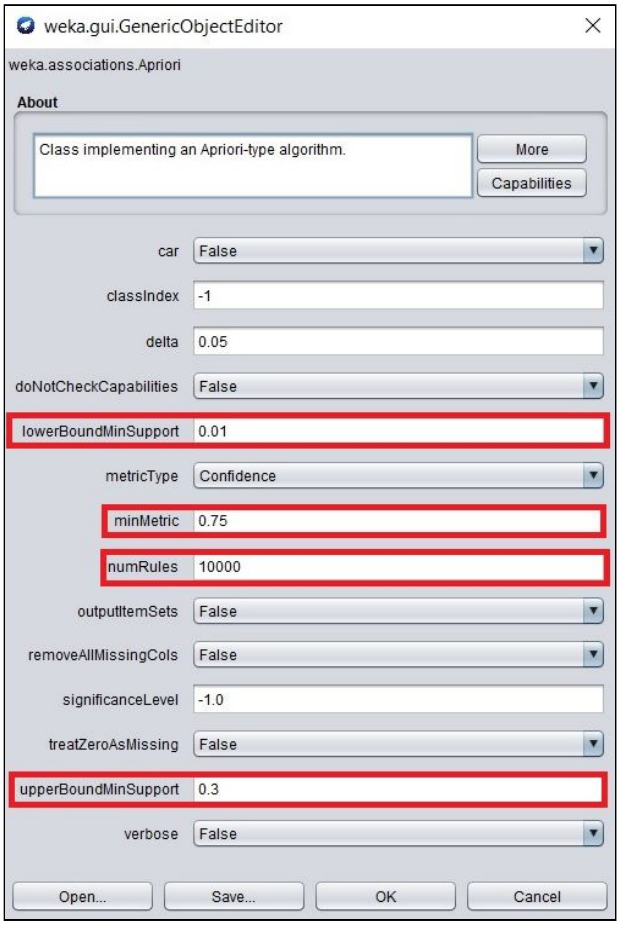

Figura 5 – Janela de configuração do algoritmo APRIORI do software Weka 3.8.

A Figura 5 revela os melhores parâmetros encontrados para essa base de dados e é o resultado de muitos testes. As únicas opções que foram alteradas possuem marcações vermelha na imagem. Os parâmetros que foram alterados e seus respectivos valores são: '*lowerBoundMinSupport*' é o suporte mínimo com valor de 0,01 ou 1%, '*minMetric*' é a confiança com valor de 0.75 ou 75%, '*numRules*' é o número de regras com o valor 10.000 e '*upperBoundMinSupport*' é o suporte máximo com valor de 0.3 ou 30%.

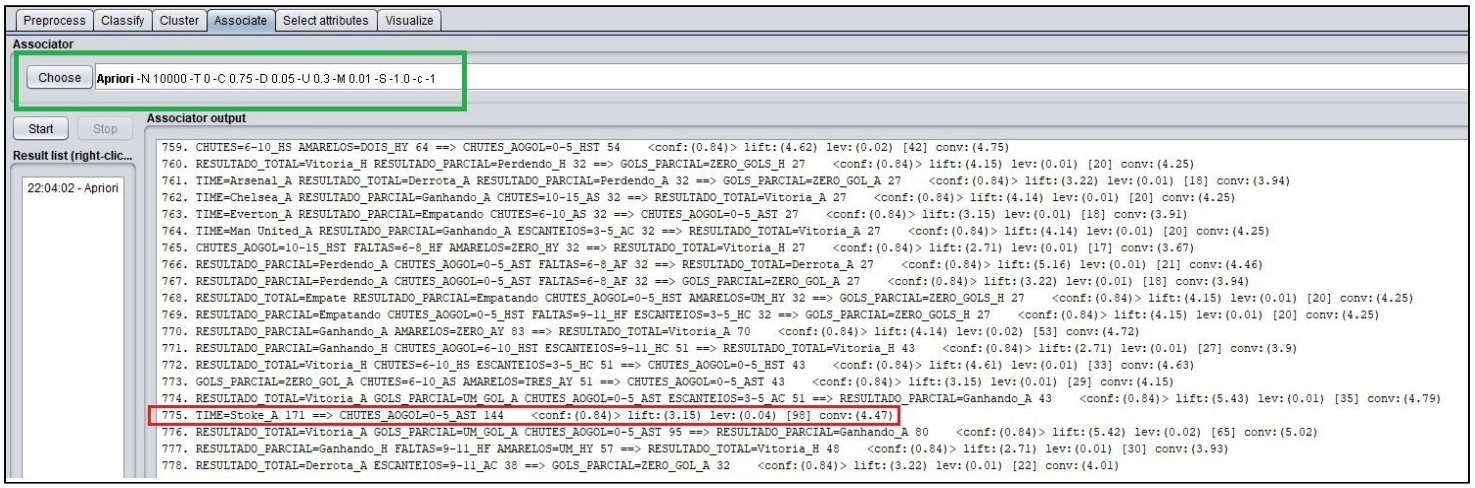

Figura 6 - Captura de tela de uma fração do resultado da mineração de dados, utilizando a tarefa de associação e o algoritmo APRIORI, fornecido pelo software Weka 3.8

A Figura 6 mostra uma fração dos resultados provenientes do software Weka 3.8, isto é, uma parte das regras de associação que foram encontradas utilizando as configurações citadas na Figura 5 e a base de dados sobre as estatísticas da liga inglesa de futebol, a Premier League. Contudo, tais regras ainda não foram analisadas por um profissional da área, logo, grande parte não é relevante para apoiar a decisão de um apostador.

ID	Regras de Associação	Confiança	Suporte	Suporte Mín.	Suporte Máx.	Confiança Mín.
1	TIME=Man United_H RESULTADO_PARCIAL=Ganhando_H 86 ==> RESULTADO_TOTAL=Vitoria_H 82 **<conf:(0.95)>**	95%	3%	2%	3%	90%
2	TIME=Man United_A CHUTES=6-10_AS ESCANTEIOS=3-5_AC 31 ==> CHUTES_AOGOL=0-5_AST 29 **<conf:(0.94)>**	94%	1%	1%	20%	90%
3	TIME=Man City_H ESCANTEIOS=3-5_HC 40 ==> RESULTADO_TOTAL=Vitoria_H 37 **<conf:(0.93)>**	93%	1%	1%	20%	90%
4	TIME=Liverpool_H GOLS_PARCIAL=DOIS_GOLS_H 33 ==> RESULTADO_TOTAL=Vitoria_H 30 **<conf:(0.91)>**	91%	1%	1%	30%	90%
5	TIME=Stoke_A 171 ==> CHUTES_AOGOL=0-5_AST 144 **<conf:(0.84)>**	84%	5%	3%	10%	80%
6	TIME=Man United_H GOLS_PARCIAL=UM_GOL_H 64 ==> RESULTADO_TOTAL=Vitoria_H 55 **<conf:(0.86)>**	86%	2%	2%	5%	85%
7	GOLS_PARCIAL=UM_GOL_H CHUTES=20-25_HS AMARELOS=ZERO_HY 32 ==> RESULTADO_TOTAL=Vitoria_H 29 **<conf:(0.91)>**	91%	1%	1%	30%	90%
8	CHUTES=20-25_HS ESCANTEIOS=3-5_HC 30 ==> RESULTADO_TOTAL=Vitoria_H 27 **<conf:(0.9)>**	90%	1%	1%	30%	90%
9	RESULTADO_PARCIAL=Ganhando_H ESCANTEIOS=6-8_HC AMARELOS=ZERO_HY 50 ==> RESULTADO_TOTAL=Vitoria_H 50 **<conf:(1)>**	100%	2%	1%	10%	95%
10	RESULTADO_PARCIAL=Ganhando_A JUIZ=J Moss_A 28 ==> RESULTADO_TOTAL=Vitoria_A 28 **<conf:(1)>**	100%	1%	1%	10%	95%
11	JUIZ=N Swarbrick_H 47 ==> VERMELHOS=ZERO_HR 47 **<conf:(1)>**	100%	2%	1%	10%	95%

Figura 7 – Regras de associação encontradas na base de dados relativa às informações estatísticas da Premier League

A Figura 7 retrata as principais regras de associação encontradas na base de dados. Contudo, esse foi o resultado obtido diretamente pelo software Weka 3.8, o qual não é intuitivo, portanto, foi realizada uma tradução para as saídas do software com o intuito de obter uma linguagem mais formal, que expresse melhor o sentido das regras.

ID	Regras de Associação	Confiança	Suporte	Suporte Min.	Suporte Máx.	Confiança Min.
1	*SE Manchester United joga em casa e, tem resultado positivo na primeira metade **TAMBÉM** tem a vitória como resultado final.*	95%	3%	2%	3%	90%
2	*SE Manchester United joga fora e, tem de 6 a 10 chutes e, tem entre 3 e 5 escanteios **TAMBÉM** chutará entre 0 a 5 vezes ao gol.*	94%	1%	1%	20%	90%
3	*SE Manchester City joga em casa e, tem entre 3 a 5 escanteios **TAMBÉM** tem a vitória como resultado final.*	93%	1%	1%	20%	90%
4	*SE Liverpool joga em casa e, marca dois gols na primeira metade **TAMBÉM** tem a vitória como resultado final.*	91%	1%	1%	30%	90%
5	*SE Stoke City joga fora **IMPLICA** que tem de 0 a 5 chutes ao gol.*	84%	5%	3%	10%	80%
6	SE Manchester United joga em casa e, tem um gol na primeira metade **IMPLICA** que terá a vitória como resultado final.	86%	2%	2%	5%	85%
7	*SE um time joga em casa e, chuta entre 20 a 25 vezes e, não toma nenhum cartão amarelo **TAMBÉM** tem a vitória como resultado final.*	91%	1%	1%	30%	90%
8	*SE um time joga em casa e, chuta entre 20 e 25 vezes e, tem entre 3 a 5 escanteios **TAMBÉM** tem a vitória como resultado.*	90%	1%	1%	30%	90%
9	*SE um time joga em casa e, tem resultado positivo na primeira metade e, não toma nenhum cartão amarelo **TAMBÉM** tem a vitória como resultado final.*	100%	2%	1%	10%	95%
10	*SE um time joga fora e, tem resultado positivo na primeira metade e, o árbitro da partida é J Moss **TAMBÉM** tem a vitória como resultado final.*	100%	1%	1%	10%	95%
11	*SE um time joga em casa e, o árbitro da partido é N Swarbrick **IMPLICA** que o time da casa não tomará nenhum cartão vermelho.*	100%	2%	1%	10%	95%

Figura 8 – Regras de associação, em uma linguagem mais simples, encontradas na base de dados relativa às informações estatísticas da Premier League

Por meio da Figura 8, que retrata a tradução das regras de associação, percebe-se o real sentido das regras encontradas. Logo, após a tradução é possível utilizá-las para realizar uma aposta em um determinado *site*, como por exemplo, o Bet365 ou o BetFair.

As regras contidas na Figura 8 são consideradas boas, pois, possuem um baixo suporte, ou seja, aparecem poucas vezes na base de dados e, possuem alta confiança. Desse modo, para as regras de associação que têm baixo suporte e alta confiança, é plausível afirmar que quando o antecedente da regra está presente, certamente o consequente também estará. Por exemplo, de acordo com a primeira regra, se o time Manchester United está jogando em seu estádio e tem um resultado positivo na primeira metade do jogo, certamente ele também terá um resultado positivo ao final do jogo. Portanto, é coerente apostar na vitória do time Manchester United sempre que o mesmo está ganhando ao final do primeiro tempo e jogando em seu estádio, Old Trafford.

As primeiras seis regras geradas fazem referência estritamente aos times que disputam a Premier League, logo, utilizando esse conjunto de regras é possível apostar nesses determinados times. A regra de associação que melhor representa uma boa oportunidade de aposta, dentre o conjunto de regras que fazem referência estritamente aos times, é: "*SE Stoke City joga fora IMPLICA que tem de 0 a 5 chutes ao gol*". Essa regra possui uma confiança de 84% e um suporte de 5%, portanto, quando há um jogo do Stoke City, como visitante, deve-se apostar que haverá menos de 5 chutes por parte do visitante. Isso acontece, pois, provavelmente o Stoke City é um time que joga na defensiva, consequentemente, é um time que chuta ao gol poucas vezes quando visitante.

Contudo, as demais regras de associação, exceto a última (Regra 11), possuem alto grau de generalização, isto é, podem ser utilizadas para apostar em qualquer time que joga a liga inglesa de futebol. O modelo de regras consideradas gerais, são: "*SE um time joga em casa e, chuta entre 20 e 25 vezes e, tem entre 3 a 5 escanteios TAMBÉM tem a vitória como resultado*", ou seja, não há determinação de um time específico. Contudo, a única condição para utilizar esse conjunto de regras é que o apostador precisa saber se o time está jogando em sua casa ou como visitante, o que é uma tarefa simples de verificação.

Por fim, a última regra foi a que gerou maior dificuldade durante análise, essa regra diz que: "*SE um time joga em casa e, o árbitro da partida é N Swarbrick IMPLICA que o time da casa não tomará nenhum cartão vermelho*". Portanto, após algum tempo tentando explicar essa regra, foi percebido que esse é um árbitro que dificilmente marca uma falta, ou seja, é um árbitro que prefere deixar o jogo correr, assim sendo, essa é uma boa regra para apostar que não haverá nenhum cartão vermelho quando o árbitro da partida é o N Swarbrick.

CONCLUSÃO

Essa pesquisa foi fundamentada na mineração de dados estatísticos da liga inglesa de futebol, com o intuito de localizar regras, padrões ou repetições escondidos em meio ao grande volume de dados. Para encontrar tais regras foi utilizado o software Weka 3.8, que permite analisar os dados usando a tarefa de associação e o algoritmo APRIORI.

A Figura 8, gerada após a análise, retrata as mais importantes regras de associação encontradas. Como são regras com um baixo suporte e alta confiança, ao utilizá-las em um site de apostas esportivas, o apostador possui grandes chances de obter um resultado positivo, isto é, conseguir um grande retorno financeiro.

Embora as regras encontradas tenham um suporte baixíssimo, de aproximadamente 1 a 5%, ainda são regras essenciais, pois, a análise foi realizada com base nos times que jogam o campeonato. Ao todo existem 2736 registros na base de dados, contudo, apenas 342 registros são referente a um time, em que metade representam aos jogos daquele time como visitante e outra metade jogando em casa. Logo, na Figura 7, a regra de número 5 revela que em 171 jogos em que o Stoke City jogou como visitante, em 144 jogos o time chutou menos de 6 vezes. Essa é uma regra com 84% de confiança e 5% de suporte, portanto, é bastante considerável, dado que, apenas 342 registros são referentes ao time Stoke City.

Dessa maneira, conclui-se que a mineração de dados, utilizando a tarefa de associação, pode ser muito útil para apostadores, tanto para apostas esportivas, como para apostas em geral, visto que os resultados são padrões e repetições. Além disso, o objetivo da pesquisa foi alcançado, pois, foram geradas inúmeras regras de associação, por meio do software Weka 3.8, que possibilitam um ótimo apoio a decisão do apostador. Assim sendo, cabe a um analista, no caso o próprio apostador, filtrar as regras de associação que realmente são relevantes para aposta e, então, começar usá-las em suas respectivas apostas.

REFERÊNCIAS

[1] DREYER GALVÃO, N.; de FÁTIMA MARIN, H. (2009). **Técnica de mineração de dados: uma revisão da literatura.** Disponível em: <http://www.redalyc.org/pdf/3070/307023846014.pdf>. Acesso em: 03 de jun. 2019.

[2] CAMILO, C. O.; SILVA, J. C. (2009). **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas.** Disponível em: <http://www.portal.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf>. Acesso em: 26 mai. de 2019.

[3] PESSOA, T.; MATURANO G. **Aprendendo apostas esportivas: É permitido fazer apostas esportivas no Brasil?**. Disponível em: <https://www.queroapostar.com/wp-content/uploads/2016/11/Aprendendo-Apostas-Esportivas.pdf>. Acesso em: 04 de jun. 2019.

[4] ALVES, A.; ANDRADE, M.; TAVARES, R. **Clube da Aposta - É proibido apostar no Brasil?**. Disponível em: <https://www.clubedaposta.com/download/Apostas-no-Futebol-eBook-gratis-para-voce-aprender-a-apostar.pdf>. Acesso em: 05 de jun. 2019.

[5] ENGLISH Premier League (football). DataHub, 2019. **Sports data - Premier League.** Disponível em: <https://datahub.io/sports-data/english-premier-league>. Acesso em: 01 de mai. 2019.

[6] FRANK, E.; HALL M. A.; WITTEN I. H. **The Weka Workbench - "Data Mining: Practical Machine Learning Tools and Techniques"**. Disponível em: <https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf>. Acesso em: 06 de jun, 2019.

OUTROS

- Link para o Dataset