Análise de Variáveis em Partidas de Futebol: Previsão de Resultados com Naïve Bayes e Poisson

Rodrigo Sehnem[1], Rejane Frozza[1,2], Daniela Duarte da Silva Bagatini[1], Daniela Saccol Peranconi[1]

Universidade de Santa Cruz do Sul (UNISC)

[1]Departamento de Engenharias, Arquitetura e Computação

[2]Programa de Pós-graduação em Sistemas e Processos Industriais

Avenida Independência, 2293 – Santa Cruz do Sul, RS, 96815-900

rodrigo.sehnem@gmail.com, [frozza, bagatini, danielap]@unisc.br

Abstract. *The objective of this work is to analyze the set of variables that can have more influence on the prediction of the result of a soccer match, using techniques such as probability calculations and prediction algorithms, with the intention of obtaining profits in betting. The techniques used for the development were Bayesian networks, with the Naïve Bayes algorithm, and probability networks, based on the Poisson calculus. The data used for training were from the Brazilian Championship (between 2010 and 2017), considering data from the years 2018 and 2019 for tests. The main results achieved were 53% correctness of the result of a match and the main variables involved were attacking strength and defense strength.*

Resumo. *O objetivo desse trabalho é analisar conjuntos de variáveis que podem ter mais influência na previsão do resultado de uma partida de futebol, utilizando técnicas como cálculos de probabilidade e algoritmos de previsão, com a intenção de obter lucros em apostas. As técnicas utilizadas para o desenvolvimento foram redes bayesianas, com o algoritmo Naïve Bayes, e a de probabilidade, baseada no cálculo de Poisson. Os dados utilizados para treinamento foram do Campeonato Brasileiro, entre 2010 a 2017, sendo considerados os dados dos anos de 2018 e 2019 para testes. Os principais resultados atingidos foram de 53% de acerto do resultado de uma partida e as principais variáveis envolvidas foram força de ataque e força de defesa.*

1. Introdução

O futebol é um esporte popular no mundo todo e, por isso, desperta interesse de uma grande área: o ramo de apostas. Existem apostas para os diversos fatores envolvendo uma partida de futebol, como escanteios, chutes a gol, faltas, cartões amarelos e vermelhos, assim como, obviamente, o resultado: vitória, derrota ou empate. Os autores em Carpita et al. (2015) citam que a previsão de resultados de futebol e o pensamento estatístico nos esportes é documentado desde os primeiros trabalhos da década de 1970. Porém, como citado em Rahman et al. (2018), ganhou notoriedade nos últimos anos e atrai vários tipos de fãs, como analistas com experiência, gerentes de clubes e outros.

Esse ramo movimenta bilhões de Reais por ano e tem cada vez mais pessoas procurando entender como realizar apostas para ter lucro a médio e longo prazo [Keogh e Rose, 2013]. Este lucro é difícil de alcançar, uma vez que o futebol possui inúmeras

variáveis e muitas delas inesperadas. É um esporte coletivo, mas que depende da atuação individual dos jogadores. Por exemplo, não é possível saber se determinado jogador será expulso da partida ou se irá se lesionar; não é possível prever se o árbitro tomará as decisões corretas, entre outros imprevistos.

Existem diversas maneiras de tentar prever o resultado de um jogo de futebol, como estatística, uso de técnicas de aprendizado de máquina (*machine learning*) e mineração de dados (*data mining*), especialista em jogos de futebol e algoritmos que usam previsão a partir de uma base de conhecimento sobre o domínio. *Machine learning* é uma área de investigação da Computação bem reconhecida, que envolve o estudo e desenvolvimento de algoritmos para extrair conhecimento automaticamente a partir de experiências passadas, sem intervenção humana [Rezende, 2003]. A mineração de dados aplica algoritmos específicos para extrair padrões a partir de um conjunto de dados, a fim de descobrir conhecimento útil em banco de dados [Fayyad, Piatetsky e Smith, 1996].

Uma abordagem bastante comum para modelar a previsão dos resultados de uma partida de futebol baseia-se na distribuição dupla de Poisson [Carpita et al., 2015]. Já, a programação genética mostrou-se superior aos modelos fuzzy e às redes neurais na previsão do futebol; assim como a combinação de técnicas de otimização genética e neural pode levar a previsões de partidas de futebol "aceitáveis"; e outros dados mostram que incorporar análise de especialistas humanos com dados em um modelo de rede bayesiana pode levar a previsões mais precisas em comparação com uma série de outras técnicas de aprendizado [Constantinou e Fenton, 2017]. Um fato importante, citado por Esme e Kiran (2018), é que o número de variáveis incluídas no cálculo afeta a precisão da previsão.

Neste contexto, o problema de pesquisa foi definido como: “É possível melhorar a previsão de resultados de partidas de futebol com o uso de técnicas inteligentes, baseadas em aprendizado e conhecimento útil, a partir da análise de um conjunto de variáveis?”.

Assim, o objetivo foi sugerir e validar, dentre um conjunto de variáveis, as que possuem relevância para obter um melhor resultado e, consequentemente, um possível lucro em apostas. Para obter essa resposta, foram realizados testes com um algoritmo de previsão (Naïve Bayes) e o modelo para calcular *odds* (probabilidade), chamado de Poisson. Após, os resultados gerados (previsões das partidas) foram testados com as *odds* pagas por um site de apostas, a fim de verificar se haveria lucro ou prejuízo.

O algoritmo de Naïve Bayes pertence à família dos classificadores probabilísticos, capaz de lidar com grande quantidade de dados, e é baseado no teorema de Bayes. O teorema de Bayes permite calcular a probabilidade de um evento, assumindo que os valores dos atributos de um objeto são independentes entre si [Barbosa et al., 2014].

O artigo está organizado nas seguintes seções: a seção 2 apresenta os trabalhos relacionados; a seção 3 aborda os métodos utilizados para o desenvolvimento da pesquisa; a seção 4 ilustra o desenvolvimento e os resultados obtidos; a seção 5 apresenta a conclusão.

2. Trabalhos Relacionados

Para verificar as publicações relacionadas nessa área, foi realizada uma bibliometria quantitativa, utilizando os seguintes termos de pesquisa: match, result, assistant, football,

prediction, bet, analysis e machine learning. O período utilizado para a busca foi entre 2014 e 2020, com os termos em inglês e em todas áreas do conhecimento, nas bases de dados Web of Science, Scielo, Scopus e Science Direct. Essa pesquisa foi realizada no portal de periódicos da CAPES.

A Tabela 1 apresenta um comparativo entre as principais técnicas utilizadas e objetivos propostos dos trabalhos selecionados. Os critérios definidos para comparação foram: objetivo, campeonato, variáveis verificadas e técnicas utilizadas.

Tabela 1. Comparativo dos trabalhos relacionados

Artigo	Objetivo	Campeonato	Variáveis verificadas	Técnicas e Validação
Baboota e Kaur (2018)	Desenvolver um modelo de previsão para a *English Premier League*	*English Premier League*	33 (partida em casa e fora, chutes a gol em casa e fora, gols, escanteios, entre outras)	*Gradient Boosting* (maior eficácia)
Constantinou e Fenton (2017)	Descobrir dados que são realmente importantes para realizar previsões	*English Premier League*	Pontos na liga, transferência de jogadores, troca de gerentes, participação em outra liga, lesões	*Redes Bayesianas*
Carpita *et al*. (2015)	Selecionar, entre centenas de covariáveis, as que têm maior influência na previsão	Série A (campeonato Italiano)	1 variável (derrota, empate ou vitória) e 481 covariáveis (estatísticas coletadas durante o jogo)	*Random Forest*
Martins *et al*. (2017)	Aplicar um algoritmo polinomial que utiliza dados coletados por observadores	*English Premier League*, La Liga e Campeonato Brasileiro	Gols, chutes a gol, escanteios, cartões amarelo e vermelho.	Classificador polinomial (método supervisionado)
Macedo e Silva (2014)	Entender fatores que podem ter influenciado na previsão do Campeão Brasileiro de 2012	Campeonato Brasileiro	Recuperação (dias) entre um jogo e outro, *ranking* da CBF e tamanho do gramado	Técnica estatística de regressão logística binária
Este trabalho	Analisar quais variáveis têm mais influência na previsão de uma partida de futebol	Campeonato Brasileiro	Forma, supremacia, pontos ganhos, vitórias, derrotas, empates, entre outras	*Naïve Bayes e Poisson*

Para este trabalho, foi escolhido o algoritmo Naïve Bayes, pois, conforme Zhang (2004), é um dos mais eficientes e eficazes algoritmos de aprendizado indutivo utilizado em aprendizado de máquina e mineração de dados. E o cálculo de Poisson por utilizado no ramos de apostas.

3. Métodos utilizados

O autor Chagas (2016) cita que é possível dividir o ramo de apostas disponibilizadas pelas operadoras na internet em dois modelos: o primeiro é o de adesão, no qual as pessoas apostam em determinado evento e que as probabilidades de resultados são determinadas pela própria casa de apostas, ou seja, o cliente aposta diretamente contra ela. Como exemplos, podem ser citados os sites Sportingbet (https://sports.sportingbet.com/pt-br/sports) e bet365 (https://www.bet365.com/pt/). O segundo modelo funciona sob o formato de betting exchanges (bolsa de apostas), similar ao mercado de ações, cujos preços são determinados pelos próprios apostadores de acordo com a regra de oferta e demanda, e a operadora funciona apenas como uma espécie de corretora. Como exemplo, pode ser citado o site Betfair (https://www.betfair.com/br). Para validar o resultado encontrado, esse trabalho utilizou o primeiro modelo citado, apostando em vitória/empate/derrota contra a casa de apostas.

É imprescindível deixar claro que, conforme citado por Chagas (2016), nas apostas esportivas os indivíduos realizam criteriosos juízos das possibilidades de ocorrência de cada situação. Nos jogos puramente de azar, ao contrário, os resultados dos eventos são ditados exclusivamente pelo acaso, isto é, pelas regras de probabilidade. Tratando-se de apostas esportivas, é realizada uma rigorosa análise dos fatos relacionados aos esportes, tais como o momento das equipes no campeonato, as prováveis escalações dos times, a posição dos adversários no ranking da modalidade, bem como as diversas outras estatísticas e informações disponíveis pelas mídias especializadas. O mesmo autor conclui que, desse modo, enquanto o ganho nos jogos de azar é determinado pela mecanicidade das máquinas ou pelo lançamento randômico dos dados, nas apostas esportivas o sucesso depende essencialmente da habilidade do apostador em fazer prognósticos precisos sobre os resultados dos eventos esportivos. Na Figura 1, visualiza-se um exemplo de um site de apostas.

Resultado Final Mudar Mercado			
Brasil - Brasileirão Série A			
Sab 08 Jun	1	X	2
16:30 Palmeiras v Atlético Paranaense	1.40	4.00	7.50
19:00 Cruzeiro v Corinthians	1.72	3.25	5.75
19:00 Grêmio v Fortaleza	1.36	4.33	9.50
19:30 Ceará v Bahia	2.40	3.00	3.20
21:00 Avaí v São Paulo	2.90	2.90	2.62
Dom 09 Jun	1	X	2
19:00 CSA v Botafogo	2.60	3.00	2.90
19:00 Fluminense v Flamengo	5.25	3.60	1.66
19:00 Santos v Atlético Mineiro	1.66	4.00	4.75

Figura 1. Exemplo de um site de apostas (Fonte: www.bet365.com)

Neste caso, por exemplo, a odd do Grêmio está baixa, pois a probabilidade do time sair vitorioso, segundo os cálculos da casa de aposta, é maior. A cada 1,00 apostado no Grêmio, a casa de apostas paga 1,36 centavos. Caso o 1,00 seja apostado no empate, a casa paga 4,33 centavos, e, finalmente, a cada 1,00 apostado no Fortaleza, a casa paga 9,50 centavos.

Os dados utilizados são do campeonato brasileiro (2010 a 2019) extraídos da base oficial da Confederação Brasileira de Futebol (CBF), dos sites https://www.ogol.com.br/ e http://www.football-data.co.uk/data.php, e de cálculos utilizados no ramo de apostas.

Nessa base de dados existem diversas variáveis para serem testadas e utilizadas, como, por exemplo: número de vitórias, empates e derrotas; número de gols feitos e sofridos; média de aproveitamento; número de pontos ganhos no campeonato; entre outras.

4. Desenvolvimento e Resultados

O foco e objetivo principal do trabalho foi compor diferentes conjuntos de variáveis do futebol para serem testadas no algoritmo, a fim de descobrir as variáveis de mais importância para a previsão de resultados, ou seja, as que tiverem uma melhor precisão de acerto. Essa previsão visa o resultado de uma partida de futebol, em vitória, empate ou derrota. O fluxo do sistema desenvolvido segue conforme a Figura 2.

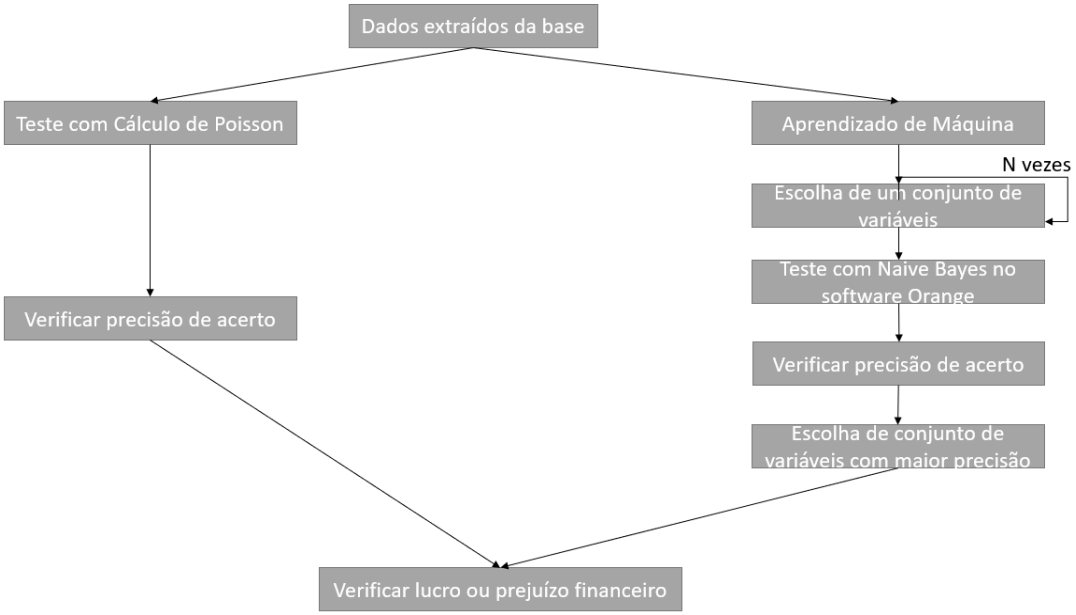

Figura 2. Fluxo do sistema

A ideia apresentada no fluxo é de extrair os dados (diversas variáveis disponíveis na base de dados), realizar a limpeza e preparação desses registros e, após isso, realizar o aprendizado de máquina. O próximo passo é escolher o conjunto de variáveis para serem testadas, e esse passo é feito inúmeras vezes. O conjunto de n variáveis escolhido é testado com o algoritmo Naïve Bayes e, após, comparado ao Cálculo de Poisson, documentando o resultado encontrado e a precisão de acerto de cada conjunto. Após encontrar o grupo de variáveis que apresentou a melhor precisão, foram analisadas partidas de futebol para aplicar o resultado encontrado e, se é obtido um lucro ou prejuízo com as apostas realizadas.

4.1 Base de dados

Para construir a base de dados, no processo de busca, os resultados das partidas foram preenchidos manualmente, conforme informação disponível nos sites esportivos. Para cada rodada do campeonato, a base de dados atualiza todas as variáveis, como, por exemplo, a colocação atual do time no campeonato, pontos feitos, gols pró e contra, porcentagem de aproveitamento, entre outras.

No processo de seleção das variáveis, foram removidos os dados duplicados, como colocação do time e posição na tabela, também sendo excluídas as variáveis como odds pagas pelo site de apostas. Na análise dos dados, eles diferenciam-se, basicamente, entre variáveis do time mandante e visitante. Uma das variáveis utilizadas é a "forma". Esta variável é uma fórmula empregada para avaliar o desempenho de cada time nos jogos anteriores, de modo a identificar quais equipes se encontram em melhor ou pior fase, o que pode influenciar no resultado de uma partida. Diferentes "formas" são calculadas por sites de apostas e variam desde os últimos quatro jogos até a temporada atual inteira. Para este trabalho, foi utilizado o cálculo definido pelos próprios autores (Equação 1).

$$\text{Forma} = 300 / (27/U9 + 18/U6 + 9/U3) \qquad (\textit{Equação 1})$$

Sendo 27 a quantidade de pontos possíveis nos últimos 9 jogos (U9), dividido por quantos pontos o time conquistou; 18 a quantidade de pontos possíveis nos últimos 6 jogos (U6), dividido por quantos pontos o time conquistou; 9 a quantidade de pontos possíveis nos últimos 3 jogos (U9), dividido por quantos pontos o time conquistou. O resultado é dividido por 300 para obter um resultado entre 0 a 100 e, assim, a porcentagem de desempenho nos últimos jogos. Desse modo, é possível obter um valor que apresente importância para os jogos mais recentes, assim como a um período mais prolongado, diminuindo a relevância de uma pequena sequência ruim ou boa de jogos de um time.

Para calcular a probabilidade de um evento acontecer, baseado nos anteriores, utilizando a Distribuição de Poisson, precisa-se obter a quantidade de eventos similares ocorridos previamente. Os modelos mais populares usados pelas casas de apostas baseiam-se no Fator de Ataque e Fator de Defesa dos times a se confrontar e na relação desses fatores com a quantidade prévia de gols ocorridos no campeonato.

O primeiro passo para chegar nestes fatores de ataque e defesa é calcular a média de gols dos times mandantes e dos times visitantes de todos os jogos realizados no campeonato ou no histórico de diversas edições da competição. Por exemplo, a média de gols dos mandantes no Campeonato Brasileiro de 2018 foi de 1,38. Enquanto que, a média de gols dos visitantes no Campeonato Brasileiro de 2018, foi de 0,79.

Em seguida, obtém-se a média de gols feitos e sofridos por um time jogando em casa e fora, dividindo os gols pela quantidade de partidas jogadas. Por exemplo, o Grêmio, no Campeonato Brasileiro de 2018, obteve os seguintes resultados:

MGFC = 1,89; MGLC = 0,74; MGFF = 0,63; MGLF = 0,68

Onde, MGFC = Média de gols feitos em casa; MGLC = Média de gols levados em casa; MGFF = Média de gols feitos fora de casa; MGLF = Média de gols levados fora de casa.

Cada um desses valores deve ser relacionado à média de gols dos mandantes e dos visitantes para que se obtenha o fator de gols feitos e sofridos, tanto em seu campo quanto no dos adversários, em comparação com a média do campeonato. Para isso, divide-se a

média de gols feitos em casa e de gols sofridos fora de casa pela média de gols dos mandantes, assim como, a média de gols sofridos em casa e dos feitos fora de casa pela média de gols dos visitantes. Por exemplo, o Grêmio, no Campeonato Brasileiro de 2018, obteve os seguintes resultados:

FGFC = 1,37; FGLC = 0,93; FGFF = 0,79; FGLF = 0,49.

Onde, FGFC = Fator de gols feitos em casa; FGLC = Fator de gols levados em casa; FGFF = Fator de gols feitos fora de casa; FGLF = Fator de gols levados fora de casa.

Por fim, para obter-se o Fator de Ataque (FA), soma-se o fator de gols feitos em casa e fora e divide-se por dois. Para o Fator de Defesa (FD), soma-se os gols levados em casa e fora e divide-se por dois. Quanto mais alto for o Fator de Ataque, maior a tendência de fazer gols, e quanto mais baixo o Fator de Defesa, maior as chances de não sofrer gols. Por exemplo, o Grêmio, no Campeonato Brasileiro de 2018, obteve: FA = 1,08 e FD = 0,71.

O último passo para a previsão de resultados da Distribuição de Poisson é relacionar os fatores de ataque e defesa dos times a se enfrentarem e a média de gols atual do campeonato. Para o time mandante, utiliza-se o Fator de Ataque do time mandante multiplicado pelo Fator de Defesa do time visitante, multiplicado pela média de gols dos mandantes. Para o time visitante, utiliza-se, da mesma forma, seu Fator de Ataque multiplicado pelo Fator de Defesa do mandante, porém, neste caso, multiplica-se pela média de gols dos visitantes.

A partir destes dois valores, aplica-se a Distribuição de Poisson de forma que se identifique a probabilidade de gols de cada time para todas as quantidades de gols, normalmente, relevante até 4 ou 5 gols por time.

Obtendo-se a probabilidade da quantidade de gols de cada time, verifica-se qual a porcentagem somada de resultados que indiquem a vitória de um time, o empate ou a vitória de outro. Essa porcentagem demonstra a probabilidade de cada resultado, assim como a cotação justa para que uma aposta tenha valor, conforme Figura 3.

A linha diagonal se refere aos resultados de empate, acima dessa linha são as porcentagens a favor da vitória do Grêmio, enquanto abaixo da linha são os resultados favoráveis ao Palmeiras. A distribuição de Poisson divide os 100% em probabilidades encontradas em partidas que já aconteceram. Neste caso, o Palmeiras está jogando como mandante e o Grêmio como visitante, e o modelo diz que o resultado mais provável é de 1 a 0 a favor do Palmeiras.

Outra variável utilizada é a supremacia, que é um valor para estimar a qualidade do time em relação ao campeonato, e que consiste em (Equação 2):

Supremacia = (Fator de ataque x Média de gols do campeonato)
– (Fator de defesa x Média de gols do campeonato) (*Equação 2)*

Palmeiras \ Grêmio	0	1	2	3	4	5	6	7	8	9	10
0	12,4%	7,3%	2,1%	0,4%	0,1%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
1	18,7%	10,9%	3,2%	0,6%	0,1%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
2	14,0%	8,2%	2,4%	0,5%	0,1%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
3	7,0%	4,1%	1,2%	0,2%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
4	2,6%	1,5%	0,4%	0,1%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
5	0,8%	0,5%	0,1%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
6	0,2%	0,1%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
7	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
8	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
9	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
10	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%

Figura 3. Modelo de probabilidade do Poisson

4.2 *Software* Orange

O *software Orange* possui *widgets* (interface de comunicação) que se conectam, fazem a leitura de um arquivo de treinamento e geram a previsão a partir de um arquivo de teste, como está ilustrado na Figura 4. É uma ferramenta de fácil entendimento e possui o algoritmo Naïve Bayes implementado, o que proporciona tempo para a preparação dos dados, testes e análises. O *Orange* também possui um tempo de resposta computacional muito rápido, uma vez que o fluxo dos *widgets* está configurado e funcionando, o aprendizado e a previsão são realizados rapidamente.

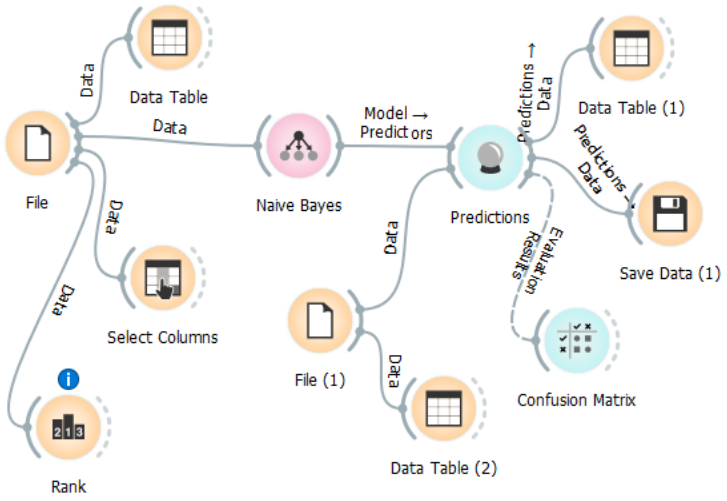

Figura 4. Modelo de previsão

No *widget file* à esquerda, é feito o *upload* do arquivo de treinamento, nesse caso, dos anos de 2010 a 2017. Esse arquivo passa pelo *Naïve Bayes* e pela previsão, gerando os valores encontrados, tendo como base o arquivo de teste, anos de 2018 e 2019, que é colocado no *widget file* (1) mais à direita. Um *widget* interessante na aplicação é o *rank*, em que se pode visualizar quais variáveis, para determinado conjunto, o Orange considerou serem as mais relevantes, conforme a Figura 5. O conjunto utilizado como exemplo na Figura 5 considera as variáveis J (Número do jogo), Ano (Ano do campeonato

corrente), R (Número da rodada no campeonato corrente), FAH1 (Força de Ataque histórica do time 1), FDH1 (Fator de Defesa histórica do time 1), FAH2 (Força de Ataque histórica do time 2), FDH2 (Fator de Defesa histórica do time 2). E o Orange considera que as variáveis mais relevantes, para esse conjunto, são FAH2 e FAH1.

	#	Gain ratio	Gini
FAH2		0.005	0.004
FAH1		0.004	0.004
FDH1		0.004	0.003
FDH2		0.002	0.002
R		0.001	0.001
ANO		0.001	0.001
J		0.001	0.001

Figura 5. *Rank* das variáveis

O modelo sempre sugere determinada porcentagem para cada variável buscada, ou seja, ele divide 100% entre derrota, empate e vitória, conforme a Figura 6.

	Naive Bayes	J	ANO	R	MAN	VIS
1	0.23 : 0.27 : 0.50 → 3	3041.0	2018.0	1.0	CRU	GRÊ
2	0.32 : 0.24 : 0.45 → 3	3042.0	2018.0	1.0	VIT	FLA
3	0.11 : 0.22 : 0.67 → 3	3043.0	2018.0	1.0	SAN	CEA
4	0.35 : 0.24 : 0.41 → 3	3044.0	2018.0	1.0	AMÉ	SPO
5	0.40 : 0.26 : 0.34 → 0	3045.0	2018.0	1.0	VAS	AMG
6	0.22 : 0.25 : 0.53 → 3	3046.0	2018.0	1.0	COR	FLU
7	0.18 : 0.28 : 0.54 → 3	3047.0	2018.0	1.0	INT	BAH
8	0.26 : 0.25 : 0.50 → 3	3048.0	2018.0	1.0	APR	CHA
9	0.22 : 0.23 : 0.56 → 3	3049.0	2018.0	1.0	BOT	PAL
10	0.23 : 0.23 : 0.54 → 3	3050.0	2018.0	1.0	SÃO	PAR

Figura 6. Porcentagem de previsão pelo Naïve Bayes

Nesse caso, os números sublinhados em verde são a porcentagem de chance do time mandante ser vitorioso, em vermelho é a chance do empate, e por último, em azul, a chance do time visitante ganhar a partida. Como exemplo, para o primeiro jogo da primeira rodada do ano de 2018, o Orange está prevendo que o Cruzeiro tem 50% de chance de ganhar do Grêmio, 27% de chance do empate e 23% de chance de vitória do Grêmio, utilizando o conjunto de variáveis 5. Uma das dificuldades encontradas pelo *Naïve Bayes* foi a previsão de empates, pois ele tende a prever a vitória para alguma das equipes. Isto acontece, porque, para o empate ser previsto, a chance deve ser maior que 33% e, historicamente o time da casa possui um valor aproximado de 50% de vitória.

4.3 Validação

Para validação dos resultados encontrados, foram comparados os resultados das partidas que já aconteceram (anos de 2018 e 2019) com o gerado pelo *software Orange*. Para cada acerto, foi somado um na contagem (para verificar a quantidade de acerto no total de 680 partidas). Na Tabela 2 podem ser verificados os resultados encontrados pelo *Naïve Bayes* para os 14 conjuntos de variáveis testados.

Tabela 2. Comparativo de resultados do Naïve Bayes

Conjunto	Método	Taxa acerto
1	Naïve Bayes	50%
2	Naïve Bayes	50%
3	Naïve Bayes	49%
4	Naïve Bayes	50%
5	Naïve Bayes	53%
6	Naïve Bayes	52%
7	Naïve Bayes	51%

Conjunto	Método	Taxa acerto
8	Naïve Bayes	49%
9	Naïve Bayes	51%
10	Naïve Bayes	51%
11	Naïve Bayes	50%
12	Naïve Bayes	50%
13	Naïve Bayes	50%
14	Naïve Bayes	52%

Na Tabela 3 podem ser verificados os resultados encontrados pelo Cálculo de *Poisson*. Nesse caso, o *Poisson* normal é o cálculo utilizando os dados do campeonato atual; o *Poisson* histórico utiliza dados de todos campeonatos (entre 2010 e 2017); e o *Poisson* + Forma utiliza dados do campeonato atual e a variável Forma, levando em conta os resultados dos últimos 6 jogos de determinado time.

Tabela 3. Comparativo de resultados do Cálculo de Poisson

Conjunto	Método	Taxa acerto
1	Poisson Normal	51%
2	Poisson Histórico	52%
3	Poisson + Forma	52%

Com os resultados encontrados foi possível testar, utilizando as *odds* dos campeonatos de 2018 e 2019, se o lucro é ou não obtido. As *odds* são calculadas da seguinte maneira: 1 dividido pela porcentagem de chance do time ganhar. Por exemplo, 1/0,75 = 1,33. Sendo assim, 75% é a porcentagem de chance do time ganhar e 1,33 é a *odd* justa a se pagar. Da mesma maneira, pode-se descobrir a porcentagem de chance do time ganhar usando a *odd*. Nesse caso, 1/1,33 = 0,75, sendo 1,33 a *odd* paga e 75% a chance do time ganhar. É possível dizer que apostar em todos os jogos do campeonato não é lucrativo, mesmo porque a casa de aposta retira uma porcentagem em cada uma delas.

Cada conjunto de variáveis teve um ótimo desempenho para determinado *range* de *odds*. No *range* entre 1,10 e 1,30 um dos conjuntos obteve melhor resultado; entre 1,30 e 1,50 outro conjunto. Consequentemente, conclui-se que escolher diferentes conjuntos de variáveis para determinadas *odds* é sim lucrativo. As Figuras 7 e 8 ilustram essa análise. O conjunto 6 obteve lucro entre as *odds* 1,10 e 1,24, considerando a porcentagem média de acerto de 52%. A imagem ilustra que, caso fosse apostado nos 680 jogos utilizando esse conjunto, para esse *range* de *odds*, o valor é positivo, portanto seria obtido lucro. O conjunto 14 obteve lucro entre as *odds* 1,70 e 2,40, considerando a porcentagem média de acerto de 52%.

As porcentagens se referem à quantidade de acerto até determinada *odd*. O terceiro campo mostra o valor somado de lucro ou prejuízo pelo valor apostado, ou seja, no conjunto 14 e na *odd* de 1,90, se o valor apostado fosse 10 reais em cada aposta, o valor

recebido seria de R$ 151,40, sendo R$ 141,40 de lucro. Isso não quer dizer somente para um único jogo, mas sim que, apostando em todos os jogos do campeonato com *odds* até 1,90, este seria o lucro.

Conjunto 6		
52%		
1,10	88%	1,15
1,11	87%	2,22
1,12	88%	5,27
1,13	87%	5,83
1,14	81%	3,75
1,15	81%	4,80
1,16	78%	4,19
1,17	77%	3,54
1,18	75%	1,85
1,19	76%	4,13
1,20	77%	5,73
1,21	75%	5,45
1,22	76%	6,78
1,23	74%	4,50
1,24	71%	1,34
1,25	70%	-0,25

Figura 7. *Range* de *odds* com lucro

Conjunto 14		
52%		
1,50	60%	-1,00
1,60	71%	-1,15
1,70	71%	0,10
1,80	68%	3,63
1,90	62%	15,14
2,00	60%	6,85
2,20	57%	9,86
2,40	54%	3,59
2,60	53%	-10,80

Figura 8. *Range* de *odds* com lucro

Ainda é possível concluir que, enquanto o Cálculo de *Poisson* é praticamente fixo ao Fator de Ataque e Fator de Defesa, com poucas possibilidades de variações, o *Naïve Bayes* pode trabalhar com infinitos conjuntos diferentes de variáveis, gerando sempre resultados diferentes, com lucros em determinadas porções.

Comparando aos trabalhos estudados e referenciados (seção 2), foi interessante encontrar diferentes conjuntos de variáveis que geram lucro para determinadas faixas de *odds*, utilizando aprendizado de máquina e o *software* Orange. O lucro se mostrou superior ao encontrado pelo Cálculo de *Poisson*.

5. Conclusão

A partir do estudo realizado, pode-se verificar que o uso da técnica de aprendizado de máquina e de algoritmos de previsão é fundamental para previsão de resultados de partidas de futebol. Tal técnica e algoritmos consistem em verificar resultados de jogos anteriores, a fim de prever resultados futuros. Também se observa a dificuldade na precisão dessas previsões, pois, além do futebol ser um esporte coletivo e com diversos fatores que podem influenciar no resultado, o empate é um grande obstáculo para uma análise mais efetiva.

Com o desenvolvimento do trabalho, preparando os dados necessários para testes e escolhendo as variáveis a serem testadas, verificou-se que é possível chegar a um aproveitamento satisfatório na previsão de resultados utilizando técnicas de aprendizado de máquina, principalmente utilizando o algoritmo *Naïve Bayes*, que obteve resultado superior ao cálculo estatístico de *Poisson*. Também foi possível verificar a diferença entre dados do atual campeonato com dados históricos, de anos anteriores. O resultado máximo alcançado foi de 53% de acerto nas previsões, ressaltando-se que há três possibilidades

de resultado, bem como foram encontrados conjuntos de variáveis que geram lucro para determinadas faixas de *odd*s. O trabalho também se mostrou eficaz na utilização da ferramenta Orange, que já possui o algoritmo de *Naïve Bayes* implementado. A aplicação é de fácil entendimento e simples, auxiliando muito bem na solução do problema proposto. As variáveis que mais influenciaram o resultado de uma partida foram Força de Ataque e Força de Defesa. Conclui-se que é possível melhorar a previsão de resultados de partidas de futebol com o uso de técnicas inteligentes, baseadas em aprendizado e conhecimento útil, a partir da análise de um conjunto de variáveis.

6. Referências

Baboota, R.; Kaur, H. Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal of Forecasting. Vol 35, Nro 2. 2018.

Barbosa, R.M.; Nacano, L.R.; Freitas, R.; Batista, B.L.; Barbosa Jr, F. The use of decision trees, naïve Bayes algorithms, and trace element patterns for controlling the authenticity of free-range-pastured hens' eggs. Journal of food science, Wiley Online Library, v. 79, n. 9, p. C1672–C1677, 2014.

Carpita, M.; Sandri, M.; Simonetto, A.; Zuccolotto, P. Discovering the Drivers of Football Match Outcomes with Data Mining, Quality Technology & Quantitative Management, Vol. 12:4, p. 561-577, 2015.

Chagas, J. M. A (im)possibilidade de regulamentação das apostas esportivas no ordenamento jurídico brasileiro. Florianópolis: UFSC. 2016. (Trabalho de Conclusão)

Constantinou, A. C; Fenton, N. E. Towards smart-data: Improving predictive accuracy in long-term football team performance, Knowledge-Based Systems, 124 (2017), p. 93-104, 2017.

Esme, E.; Kiran, M. S. Prediction of Football Match Outcomes Based on Bookmaker Odds by Using k-Nearest Neighbor Algorithm, International Journal of Machine Learning and Computing, Vol. 8, No. 1, 2018.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth P. From Data Mining to Knowledge Discovery in Databases. AI Magazine, Vol 17, No. 3, p. 37-54. 1996.

Keogh, F., Rose, G. Football betting - the global gambling industry worth billions, BBC, 2013. Disponível em: http://www.bbc.com/sport/football/24354124. Acessado em: 12/03/2019.

Macedo, P.A.P.; Silva, C.D.; Predição de resultados no Campeonato brasileiro 2012 série A, Revista Brasileira de Futebol, p.35-41, 2014.

Rahman, M. H. A. A.; Mustapha, A.; Fauzi, R.; Razali, N. Bayesian Approach to Classification of Football Match Outcome, International Journal of Integrated Engineering: Special Issue 2018: Data Information Engineering, Vol. 10 No. 6 (2018) p. 155-158, 2018.

Rezende, Solange Oliveira. Sistemas Inteligentes – Fundamentos e Aplicações. São Paulo: Manole, 2003.

Zhang, H. The Optimality of Naïve Bayes, Faculty of Computer Science, University of New Brunswick, Canada, 2004.