seminário de iniciação científica
ISNN: 2358-6052

DESENVOLVIMENTO DE UM SISTEMA PARA PREVISÃO DE RESULTADOS DE JOGOS DE FUTEBOL BASEADO EM NOTÍCIAS

Júlio César Machado Álvares[1]; Marcos Roberto Ribeiro[2]

1 Bolsista (CNPq, FAPEMIG ou IFMG), Engenharia de Computação, IFMG Campus Bambuí, juliocmalvares07@gmail.com
2 Orientador, IFMG - Campus Bambuí, GPSisCom, marcos.ribeiro@ifmg.edu.br

RESUMO

O futebol é um esporte altamente difundido e praticado por todo o mundo, no Brasil, considerado uma paixão nacional. Existem diversos trabalhos na literatura que propõe a predição de resultados de jogos de futebol por meio de estatísticas sobre os jogos e ou jogadores. Porém, estes trabalhos em geral não consideram alguns fatos que podem contribuir para a mudança do resultado de um jogo, por exemplo, a contusão de um jogador do time. Neste trabalho, é proposto o enriquecimento de um sistema já presente na literatura, além do desenvolvimento de alguns sistemas utilizando inteligência artificial, onde foi buscado melhorar o resultado de predição inserindo análise de sentimentos de notícias sobre os times, esperado que esta consiga prover para o modelo os “dados faltantes” de um jogo, trazendo uma maior confiabilidade na predição. Dessa forma, foi possível desenvolver alguns sistemas de software livre e montar uma base de dados que engloba os vinte times do Campeonato Brasileiro, onde foi possível efetuar a melhoria do sistema de predição estatística e alcançar resultados com os modelos de aprendizagem de máquina em torno dos 60\% de acurácia.

INTRODUÇÃO:

O futebol é um esporte mundialmente conhecido e considerado, no Brasil, uma paixão nacional (QUEIROZ *et al*. 2019). Tal esporte sempre foi alvo de especulações quanto a campeonatos, rebaixamentos de clubes e outros fenômenos. Assim, foram criados os primeiros modelos de previsão de resultados de jogos, utilizando a estatística (FILHO et al. 2017; ARAÚJO et al. 2015; AHZARI; WIDYANINGSIH; LESTARI, 2018; SANTOS,2019; QUEIROZ et al. 2019). A previsão de resultados de jogos pode trazer resultados interessantes para sistemas de aposta e, principalmente, para os times que participam de um campeonato. Outro tema de pesquisa bastante explorado nos últimos anos é a Análise de Sentimentos (GODBOLE; SRINIVASAIAH;SKIENA, 2007; BALAHUR et al. 2010; JAI-ANDALOUSSI et al. 2015). Basicamente, a análise de sentimento verifica se um determinado texto transmite informações positivas ou negativas (JAI-ANDALOUSSI et al. 2015). Desta maneira, a análise de sentimento pode ser aplicada nas mais diversas áreas como marketing, ciências sociais e política.

Os trabalhos sobre predição de resultados de jogos utilizando estatística consideram informações sobre os jogadores, histórico de jogos e mando de campo. Contudo, existem outros fatores que podem influenciar no resultado dos jogos como uma lesão que pode tirar um jogador do próximo jogo. Uma alternativa para que acontecimentos dessa espécie sejam considerados é utilizar técnicas de análise de sentimento sobre notícias relacionadas aos times no período que antecede a partida. O presente trabalho utiliza essa proposta e combina informações da análise de sentimento com métodos estatísticos existente para melhorar a taxa de acerto do resultado das partidas.

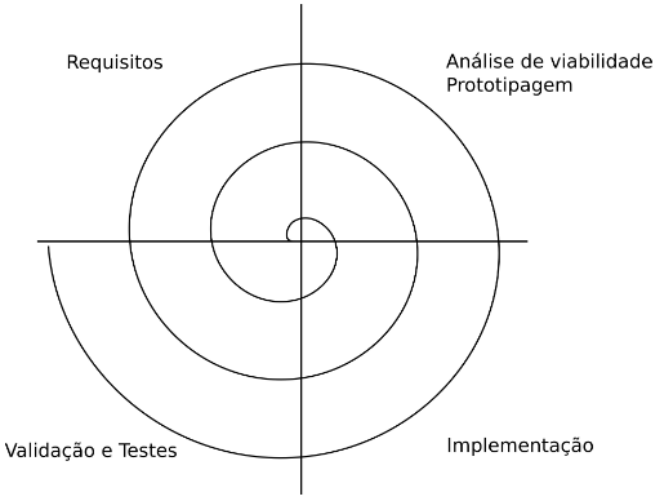
METODOLOGIA:

Para a realização deste projeto, uma metodologia de três fases foi adotada, sendo estas a construção e testes dos sistemas de recuperação de informação, construção e validação da base de dados e, por fim, construção e testes dos sistemas de predição. Como metodologia para o desenvolvimento dos sistemas, foi adotado o modelo de desenvolvimento espiral (BOEHM, 1988). A ideia deste modelo é dividir o desenvolvimento em vários estágios, que são trabalhados de forma incremental até a versão final do software.

A Figura 1 descreve como foi o desenvolvimento espiral do presente trabalho. Na primeira etapa, foram desenvolvidos os objetivos a serem alcançados pelo protótipo, além das limitações do mesmo. A segunda etapa é a análise de viabilidade e a prototipação, qualquer mudança necessária no sistema é efetuada nesta etapa, a fim de serem implementados na etapa seguinte. A terceira etapa consiste na

implementação do protótipo, sendo que o mesmo deve atender os requisitos que foram impostos na primeira etapa. E, por fim, a quarta etapa consiste na fase de testes e validação do protótipo. Ao fim de cada ciclo, um novo protótipo foi desenvolvido, até que, ao fim do processo, o sistema atenda todos os requisitos estabelecidos.

Figura 1 – Modelo espiral proposto para o presente trabalho.



Fonte: Os Autores (2021).

Todas as implementações apresentadas foram feitas utilizando a linguagem de programação Python em sua versão 3.8. O paradigma de programação adotado nas implementações foi o de orientação a objetos. Além disso, os dados utilizados em toda a aplicação foram persistidos utilizando arquivos JavaScript Oriented Notation (JSON), Comma-Separeted Values (CSV) ou Pickle. Os gráficos gerados foram plotados utilizando o *framework* matplotlib e, para controle de listas e matrizes foi utilizado os *frameworks* Numpy e Pandas. Para auxiliar nas implementações de modelos de inteligência artificial foi utilizado o *framework* SciKit Learn e, por fim, os diagramas de classe apresentados foram construídos no software livre Draw.io.

Quanto à tecnologia utilizada para codificar e testar o sistema, foi utilizado o editor de texto VisualStudio Code, em sistema operacional Linux. Os testes, feitos utilizando o interpretador padrão do sistema operacional. Além disso, todas as versões de protótipos gerados foram persistidas no GitHub, a fim de ter versionamento de todo o sistema, além de questões de segurança e backup.

Para o sistema de análise de sentimentos, foi utilizado o trabalho de Almeida (2018), um sistema de software livre denominado LeIA, onde este é uma implementação do trabalho de Gilbert (2014). O sistema de Gilbert (2014), denominado VADER, sigla para *Valence Aware Dictionary for Sentiment Reasoning*, onde seu funcionamento consiste na identificação de palavras chaves no corpus, onde é atribuído um valor de polaridade para o conjunto ao fim do processamento. O autor demonstra algumas vantagens do sistema, sendo estas a falta de necessidade de treinamento do sistema, sendo que o mesmo é baseado em dicionários externos, é rápido o suficiente para ser utilizado em *streaming* de dados e a sua ampla gama de domínios dos textos a serem analisados. É um sistema SISO, ou seja, funciona com uma entrada e uma saída, sendo essa 4 valores: a porcentagem de positividade, neutralidade e negatividade do texto e um valor denominado *compound*. O *compound* trata-se de um valor de domínio de [−1,1] e este está ligado com a positividade ou negatividade da entrada, ou seja, o quão positivo ou o quão negativo a entrada é.

Após o desenvolvimento completo dos sistemas, os mesmos passaram por etapas de testes e validação. Para desenvolver um sistema de aprendizagem de máquina, no caso de um sistema de aprendizagem supervisionada, o mesmo foi treinado com um conjunto de pontos conhecidos com suas respectivas classes, para que o mesmo possa concretizar sua estratégia de predição. Diferentes algoritmos apresentam diferentes estratégias de predição e, logicamente, diferentes estratégias de treinamento. Uma SVM (*Support Vector Machine*), por exemplo, constrói um hiperplano n-dimensional contendo todos os pontos $x \in R$ que satisfaça a condição de $h(x) = 0$, onde $h(x)$ é a função do hiperplano (ZAKI; MEIRA, 2014). Dessa forma, a SVM constrói uma função que separa a base de dados e, quando novos dados forem inseridos no modelo para predição, a mesma tem suas estratégias para definir a qual classe pertence aquele conjunto.

Ao fim do treinamento, é necessário medir a acurácia do modelo, ou seja, o quanto o modelo acerta na predição para novos conjuntos de dados que são inseridos nele. A acurácia de um sistema é a fração de predições corretas que o mesmo desempenha no conjunto de testes. Esta, é definida pela diferença de 100% − Error rate. O Error rate, ou taxa de erro, é definido por:

$$\textit{Error rate} = \frac{1}{n}\sum_{i=1}^{n} f(a_i \neq \hat{a}_i)$$

Como demonstra a equação, a taxa de erro representa a quantidade de observações erradas que foram preditas pelo sistema de aprendizagem de máquina. Por sua vez, a acurácia é definida por:

$$\textit{Accuracy} = \frac{1}{n}\sum_{i=1}^{n} f(a_i = \hat{a}_i) = 1 - \textit{Error Rate}$$

A Acurácia nos da uma estimativa da probabilidade de predições corretas, então, quanto maior. melhor o modelo é (ZAKI; MEIRA, 2014). Além destas, também foram utilizadas as medidas de *Precision*, *Recall* e *F1-score*, onde estas tem relação com as habilidades do modelo.

A *Precision* de um modelo é a proporção entre verdadeiros positivos e falsos positivos, ou seja, a habilidade do classificador de não rotular como positiva uma amostra negativa. Já o é a proporção entre verdadeiros positivos e falsos negativos, onde este demonstra a habilidade do sistema de encontrar amostras positivas. Por fim, o *F1-score* é uma média ponderada entre a *Precision* e o *Recall* (ZAKI; MEIRA,2014).

Para testar e validar os sistemas de predição, como se tratam de acontecimentos futuros, ou seja, deseja-se prever o resultado de um fenômeno que ainda não aconteceu, a metodologia de teste do sistema seguiu os seguintes requisitos:

1) Predizer todos os jogos da rodada x do Brasileirão, sendo que deve-se iniciar da rodada 2,
2) Medir a acurácia da rodada seguindo o método exposto nesta Seção,
3) Medir a acurácia do campeonato completo.

É necessário começar a partir da rodada 2 devido à falta de informações, como quantidade de gol sou chutes, para a predição da rodada 1. Caso a rodada 1 entrasse no sistema, a mesma teria teor completamente aleatório, ou mesmo não funcionaria nos métodos aqui propostos. No caso do sistema de predição estatístico, trabalho de Queiroz *et al*. (2019), assim como o autor apresenta, foram construídas as matrizes de probabilidade de 0 a 5 gols para todas as partidas do campeonato, além das suas somas de probabilidade de vitória do time mandante, empate ou derrota do time mandante. Para os sistemas de predição baseados em aprendizagem de maquina foram também adotados os requisitos expostos anteriormente.

RESULTADOS E DISCUSSÕES:

Como primeiro resultado deste trabalho, foi desenvolvida uma base de dados de estatísticas e notícias sobre o Campeonato Brasileiro de Futebol Série A 2018 (ÁLVARES; RIBEIRO, 2019). Tal base conta com estatísticas referentes aos 20 times participantes do campeonato, todos os seus jogos e, além disso, notícias de abril à dezembro de 2018 destes mesmos 20 times, somando mais de 18 mil notícias.

Partindo para os sistemas de predição, o primeiro sistema implementado e testado foi o de predição utilizando estatística e modelo Poisson, proposto por Queiroz et al. (2019). Tal sistema recebe uma partida e,utilizando das variáveis calculadas pelas médias de gols, constrói uma tabela de probabilidades de resultados dos jogos. Foi observado que o número máximo de gols da base de dados foi é 5, portanto, o sistema fez a predição para todos os placares de 0x0 até 5x5. A Tabela 1 demonstra as probabilidades do jogo Atlético-MG x Vitória, no dia 22/04/2018, pela segunda rodada do Campeonato Brasileiro.

Tabela 1 – Tabela de probabilidades gerada pelo modelo Poisson puro

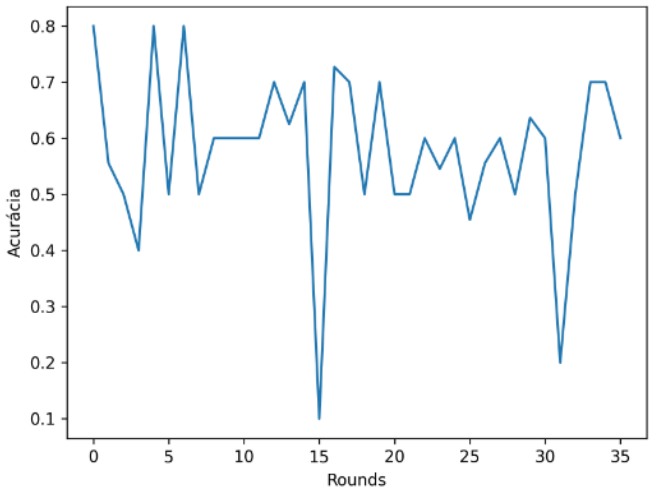
	0	1	2	3	4	5
0	3.019738	4.529608	3.397206	1.698603	0.636976	0.191093
1	6.039477	9.059215	6.794411	3.397206	1.273952	0.382186
2	6.039477	9.059215	6.794411	3.397206	1.273952	0.382186
3	4.026318	6.039477	4.529608	2.264804	0.849301	0.254790
4	2.013159	3.019738	2.264804	1.132402	0.424651	0.127395
5	0.805264	1.207895	0.905922	0.452961	0.169860	0.050958

Fonte: Os Autores (2021).

Ao observar a Tabela 1 é possível notar que, a diagonal principal, em amarelo, representa as probabilidades de empate do jogo. O quadrante inferior, em azul, representa a vitória do mandante e o quadrante superior, em vermelho,diz respeito a derrota do time mandante (ou vitória do time visitante). Dessa forma, o somatório de cada um dos quadrantes e da diagonal principal representam, respectivamente, a probabilidade de vitória do time mandante, derrota do time visitante e empate. Para esse exemplo, tem-se um somatório de 47.70% para vitória do time mandante, 21.61% para empate e 28.58% para derrota do time mandante. Dessa forma, o sistema acertou o resultado do jogo, onde o placar real deste foi de 2x1, vitória para o time mandante.

Ao final dos testes por rodada, o sistema foi avaliado como um todo, e o mesmo obteve acurácia de 57.77%. Como Queiroz et al. (2019) reforça em seu trabalho, um sistema de predição de resultados de futebol lida com três possíveis saídas, logo, um resultado ruim seria algo em torno de 33.3%. Considerando esta afirmação, nota-se que o modelo de Poisson alcança resultados satisfatórios para essa base de dados. Os resultados por rodada variaram de 10% a 80% de acurácia, assim como demonstra a Figura 2.

Figura 2 – Acurácia por rodada do preditor Poisson.



Fonte: Os Autores (2021).

Ainda observando a Figura 2, nota-se que a acurácia por rodada é bastante estável em torno dos 60%, valor bastante satisfatório para um sistema de predição. Entretanto, houve dois pontos no gráfico que demonstram resultados bastante incoerentes com o restante, sendo estes a rodada 15 e a rodada 31 do campeonato. Tal ocorrência provavelmente deve-se ao fato de que o modelo Poisson apenas leva em conta estatísticas puras para a predição do resultado, ou seja, caso o time venha se saindo bem nas ultimas rodadas,o modelo pré julgará que na próxima rodada, ele também se sairá bem, o que não acontece em todos os casos.

Tabela 2 – Tabela de probabilidades gerada pelo modelo Poisson com *compound*

	0	1	2	3	4	5
0	2.873124	4.370970	3.324844	1.686061	0.641264	0.195114
1	5.827960	8.866251	6.744247	3.420075	1.300765	0.395778
2	5.910834	8.992329	6.840150	3.468709	1.319262	0.401406
3	3.996591	6.080134	4.624945	2.345356	0.892015	0.271409
4	2.026711	3.083296	2.345356	1.189353	0.452349	0.137634
5	0.822212	1.250856	0.951482	0.482506	0.183512	0.055836

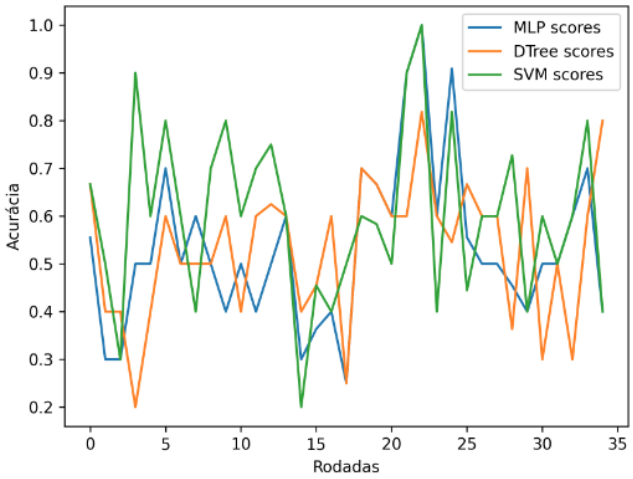
Fonte: Os Autores (2021).

Esperando que o modelo Poisson pudesse sofrer modificações ao adicionar o valor de polaridade sentimental das notícias, foi implementada uma função, denominada by_round_with_sentiment, através da qual o valor de polaridade sentimental era multiplicado nos parâmetros λ de cada um dos times. A Tabela 2 demonstra os resultados obtidos para o mesmo jogo. Ao observar os resultados, nota-se que não há uma diferença marcante. O que ocorreu foi que algumas probabilidades diminuíram enquanto algumas aumentaram, de forma bastante equilibrada. Não houve uma concentração de probabilidade nos resultados esperados e, este fenômeno ocorreu em toda a base de dados. Em geral, a acurácia geral do sistema não foi afetada, mesmo utilizando os comentários, que deixam os valores de *compound* mais dispersos.

Partindo para os sistemas de aprendizagem de máquina, foi utilizado 3 modelos, sendo estes a *Multi-Layer Perceptron, Support Vector Machines* e as árvores de decisão, como já comentado anteriormente. Para rodar os sistemas, foi dado como entrada os valores de somatório das tabelas de probabilidade geradas pelo modelo Poisson, juntamente com os valores de *compound* das notícias de cada time. A Figura 3 demonstra a acurácia por rodada dos três modelos.

seminário de iniciação científica
ISNN: 2358-6052

Figura 3 – Acurácia por rodada dos sistemas MLP, SVM e Árvore de Decisão.



Fonte: Os Autores (2021).

Observando a Figura 3, nota-se que, primeiramente, o modelo SVM teve leve vantagem sobre os outros modelos. Tal fato provavelmente deve-se à forma como uma SVM é treinada, buscando sempre reduzira dimensionalidade dos dados a fim de buscar um hiperplano que os separe linearmente. Ao comparar as Figuras 2 e 3, a acurácia do modelo Poisson é bastante estável em torno dos 60%, como foi comentado, e esse comportamento não se repete tanto para os sistemas de aprendizagem de máquina. É notório também que, assim como foi obtido varias observações com altas acurácias, mas também, mais observações com baixas acurácias, comparadas ao modelo Poisson.

Em geral, os modelos de aprendizagem de máquina foram bem na predição dos resultados de jogos de futebol, porém, o único que ultrapassou o modelo Poisson, para a arquitetura aqui apresentada, foi a SVM. Tanto a MLP como a árvore de decisão, obtiveram acurácia final de 53.30%, mas a SVM alcançou 59.84%, cerca de 2% a mais que o modelo Poisson. A Tabela 3 demonstra a acurácia de todos os modelos desenvolvidos neste projeto.

Tabela 3 – Métricas de avaliação de todos os modelos trabalhados

Modelo	Acurácia (%)	*Precision* (%)	*Recall* (%)	*F1-score* (%)
Poisson	57.77	49.44	49.44	49.44
Poisson + *compound*	57.77	49.44	49.44	49.44
MLP	53.30	41.72	41.72	41.72
Árvore de Decisão	53.30	46.66	41.72	44.05
SVM	59.84	47.05	31.78	37.94

Fonte: Os Autores (2021).

CONCLUSÕES

O presente trabalho teve como objetivo desenvolver uma melhoria para um sistema de predição de resultados de jogos de futebol utilizando notícias sobre os times. O tema abordado conta com alguns pontos de importância para a sociedade, ciência e negócios.

Observando os resultados obtidos, pode-se afirmar que as notícias modelam o comportamento da performance de um time no Campeonato Brasileiro de 2018. Além disso, também é possível afirmar que todos os objetivos levantados neste trabalho foram alcançados. Foi possível construir dois sistemas de recuperação de informação, que foram capazes de construir uma base de dados de estatísticas e notícias sobre os vinte times do campeonato. Os métodos de predição existentes na literatura foram analisados para a escolha do trabalho de Queiroz et al. (2019) como o trabalho que seria aprimorado. Além disso, foi possível aplicar análise de sentimentos na base de dados de notícias e, assim, obter resultados bastante interessantes. Por fim, foi possível mesclar por meio de aprendizagem de máquina, informações provindas

de um sistema de predição baseado em estatística com as informações da análise de sentimentos, criando um modelo com cerca de 2% mais acurácia que o modelo considerado, chegando a aproximadamente 60% de acurácia.

REFERÊNCIAS BIBLIOGRÁFICAS:

AHZARI, H. C.; WIDYANINGSIH, Y.; LESTARI, D. Predicting Final Result of Football Match Using PoissonRegression Model.IOP Conf. Series: Journal of Physics, 2018.

ALMEIDA, R. J. A.LeIA - Léxico para Inferência Adaptada. GitHub, 2018.https://github.com/rafjaa/LeIA.

ÁLVARES, J. C. M.; RIBEIRO, M. R. SoccerNews2018: a dataset of statistics and news of the 2018 BrazilianSoccer Championship. In: SIMPÓSIO Brasileiro de Banco de Dados (SBBD) - Dataset Showcase Workshop(DSW). Fortaleza, CE, Brasil, 2019.

ARAÚJO, C.et al. Modelagem Estatística para Previsão de Jogos de Futebol: Uma aplicação no CampeonatoBrasileiro de 2014.Revista de Estatística da Universidade de Ouro Preto, 2015.

BALAHUR, A.et al. Sentiment Analysis in the News.Proceedings of the Seventh International Conferenceon Language Resources and Evaluation (LREC'10), 2010.

BOEHM, B. W. A Spiral Model Of Software Development and Enhancement.Computer, n. 5, p. 61–72, 1988.
FILHO, C. et al. Uma Abordagem Bayesiana para Previsão de Resultados de Jogos de Futebol: Uma Aplicação ao Campeonato Inglês.Revista Brasileira de Biometria, 2017.

GILBERT, C. H. E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In:EIGHTH International Conference on Weblogs and Social Media (ICWSM-14). 2014.

GODBOLE, N.; SRINIVASAIAH, M.; SKIENA, S. Large-Scale Sentiment Analysis for News and Blogs.Icwsm,v. 7, n. 21, p. 219–222, 2007.

JAI-ANDALOUSSI, S.et al. Soccer Events Summarization by Using Sentiment Analysis.International Conference on Cumputational Science and Computational Intelligence, p. 398–403, 2015.

QUEIROZ, E. R.et al. Da (Im)previsibilidade do futebol sob a ótica da distribuição de Poisson. In:SANTOS, J. M. A. d. Previsões de resultados em partidas do campeonato brasileiro de futebol. 2019.Tese (Doutorado) – FGV.

ZAKI, M. J.; MEIRA, W.Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press, 2014.

Participação em Congressos, publicações e/ou pedidos de proteção intelectual:

Com a construção da base de dados para este projeto, denominadaSoccerNews2018, até onde se sabe,é a única base de dados que contém dados de partidas e notícias sobre campeonatos brasileiros. Tal fato proporcionou a publicação de um artigo no XXXIV Simpósio Brasileiro de Banco de Dados, na categoria *Dataset Showcase Workshop*, que ocorreu em Outubro de 2019 na cidade de Fortaleza, Ceará, Brasil.