

Construção de um software para análise de dados de partidas de futebol coletados através do processo de Scout

Gabriel de Araujo¹

Resumo: Uma das maiores dificuldades em se trabalhar com predições de partidas de futebol, é a falta de padronização das bases de dados existentes e o intuito com esse projeto é padronizar as bases de dados de acordo com o que o usuário deseja. As bases de dados utilizadas são do formato .csv. A visualização de informações do tipo são temas de vários outros trabalhos, o que também é o foco deste, onde o software visa a padronização dos dados e também permitirá o usuário compilar esses dados, projetá-los e facilitar a visualização dos dados compilados, provendo assim, uma ferramenta de visualização materializada.

Palavras-chave: Futebol. base de dados. padronização.

Development of software for analysis of football matches datas collected by scout process

Abstract: One of the biggest difficulties in working with football match predictions is the lack of standardization of existing databases, and the purpose of this project is to standardize the databases according to what the user wants. The databases used are of the .csv format. The visualization of such information is the subject of several other projects, which is also the focus of this, where the software aims at standardization of the data and will also allow the user to compile this data, design it and facilitate the visualization of the compiled data, providing thus, a materialized visualization tool.

Keywords: Football. databases. standardization.

¹Estudante de Ciência da Computação, IFTM, Campus Ituiutaba, gabrielaraujovni@gmail.com

1 INTRODUÇÃO

O futebol é uma das modalidades esportivas mais famosas e disputadas do mundo. Trata-se de um esporte cujo objetivo transpor uma bola entre as balizas, que são as extremidades do campo, utilizando basicamente toques com os pés. Vence a partida a equipe que atingir o objetivo – que são chamados de gols - mais vezes na partida (SFEIR, 2011). Uma das principais razões pelo futebol se tornar uma febre mundial é o fácil entendimento das regras, o baixo custo e o fato de ser uma das modalidades mais empolgantes no meio esportivo. A última copa do mundo que foi jogada na Rússia no ano de 2018 atingiu uma audiência televisiva de mais de 3,5 bilhões de pessoas, batendo recorde de audiência, e somente a final entre França e Croácia atraiu 1,1 bilhões de telespectadores (CHADE, 2018).

O alto nível de interesse das pessoas por esse esporte gera não apenas telespectadores, mas também muita movimentação financeira em torno dessa modalidade. A copa do mundo de 2018 teve lucro para a FIFA de 5,35 bilhões (FIFA, 2018). A movimentação financeira envolve patrocínios a clubes e seleções nacionais, venda de ingressos, produtos licenciados e transmissões por veículos de comunicação. Além é claro de transações de transferências envolvendo jogadores. Esse tipo de movimentação financeira atrai vários investidores que visam lucrar com o esporte. Sistemas computacionais que trabalham com a previsão de resultados e que auxiliam a minimizar os riscos e maximizar os lucros tornam-se então uma importante ferramenta de trabalho para o dia a dia do futebol (PERIN; VUILLEMOT; FEKETE, 2013).

O processo de scout é amplamente utilizado esportes, principalmente no futebol, para o registro, observação e análise do desempenho técnico e tático de equipes em partidas. Em esportes, o scout pode ser definido como uma técnica que consiste em analisar a partida, os momentos, os lances de um jogo para verificar o rendimento das equipes. O scout é objeto de estudo em várias modalidades esportivas: basquete, vôlei, handebol, futebol americano, beisebol e futebol (MARTINS et al., 2017). A coleta de dados é utilizada não apenas para conhecimento do rendimento da própria equipe, mas também para o estudo de táticas e técnicas de equipes adversárias. Esses dados possibilitam mensurar quais são as principais características de uma equipe, identificando quais são suas principais jogadas, seus principais jogadores, sua organização tática e técnica, porém não existe uma padronização de quais dados devem ser coletados, o que deixa o processo muito abrangente (DUARTE; SOARES; TEIXEIRA, 2015). O objetivo proposto é criar um software que seja capaz de acessar base de dados e compilar e projetar esses dados de maneira que auxilie o usuário a tomar decisões.

1.1 OBJETIVO GERAL

O objetivo deste trabalho é construir um software que consiga acessar diferentes bases de dados, compilar e projetar esses dados de maneira que auxilie o usuário no processo de tomada de decisão, assim facilitando e deixando o processo mais fácil de ser visualizado pelo usuário.

1.2 OBJETIVOS ESPECÍFICOS

- comparar as bases de dados existentes e compilar essas informações de maneira a projetar uma base de dados única;
- fornecer acesso ao usuários de informações referentes aos quatro elementos do processo de scout: espaço, tempo, jogador e fundamento;
- permitir ao usuário projetar e visualizar os relatórios com a compilação das informações proveniente da base de dados.

2 REFERENCIAL TEÓRICO

Em uma partida de futebol o número de anotações que se pode fazer para que seja possível sua correta descrição é bastante elevado (CONSTANTINO; FENTON; NEIL, 2013). Se for levado em consideração outros fatores extracampo temos novamente um alto número de componentes envolvidos. Dessa maneira a definição de quantas e quais características serão utilizadas para se realizar a predição dos resultados de partidas se torna bastante complexa (TAX; JOUSTRA, 2015). Na literatura é possível encontrar trabalhos que levam em consideração uma série de fatores tais como dados de fundamentos obtidos por scout como em Sfeir (2011), outros como em Owramipur e Mozneb (2013) são utilizados fatores de logísticas tais como distância percorrida entre duas partidas subsequentes (TAX; JOUSTRA, 2015) ou fatores psicológicos como em Duarte, Soares e Teixeira (2015).

O processo scout apesar de bastante difundido e utilizado não é padronizado no que diz respeito ao número de fundamentos que serão coletados (PENDHARKAR; KHOSROWPOUR; RODGER, 2000). Em Brooks (BROOKS; KERR; GUTTAG, 2016) é utilizado um conjunto de 7 fundamentos obtidos por scout e um conjunto de 6 características de logísticas para se obter uma correta predição. Já em Ulmer e Fernandez (2013) são utilizados 9 fundamentos obtidos por scout. Em Igiri (2015) são utilizadas como características 23 fundamentos coletados por scout. Em Hucaljuk e Rakipović (2011) são utilizados 10 fundamentos obtidos por scout. Tax e Joustra (2015) utiliza 19 características obtidas por scout, 12 de logística e 12 baseadas em sites de apostas. Enquanto que Duarte, Soares e Teixeira (2015) utiliza um conjunto de 12 características obtidas por scout e 12 características envolvendo aspectos psicológicos das equipes.

É importante ressaltar mais uma vez que não existe uma padronização para as características envolvidas com os trabalhos existentes na literatura. Pode-se afirmar, entretanto que as características estão ligadas ao tipo de informação que se necessita coletar sobre um ou mais aspectos que envolvem uma partida de futebol (TAX; JOUSTRA, 2015). Não é possível ainda afirmar que uma outra metodologia é superior a outra, mas que são complementares. Para o caso desse software o mesmo deverá ser capaz de: comparar as bases de dados existentes e compilar essas informações de maneira a projetar uma base de dados única.

3 DESENVOLVIMENTO

Machine Learning, termo em inglês para aprendizagem de máquina, é a capacidade de computadores aprenderem a tomarem decisões sem que sejam previamente programados para isso. Esse aprendizado, assim como acontece boa parte do aprendizado humano, se dá através de erros e acertos, que são obtidos através de algoritmos matemáticos. Esses algoritmos matemáticos trabalham classificando variáveis, por exemplo, queremos que um computador seja capaz de distinguir um ser humano de um gavião. Para que o computador seja capaz de fazer essa distinção, fornecemos variáveis que serão utilizadas pelos algoritmos matemáticos para realizar as classificações. Nesse caso, fornecemos variáveis como, possui asas, capacidade de falar, capacidade de voar, capacidade de caminhar, possui bico, põe ovos, entre outras. Assim, após o algoritmo analisar os dados, o computador poderá distinguir com facilidade humanos de gaviões. É importante frisar que a qualidade da classificação se dá pela qualidade das variáveis, porque se utilizarmos variáveis como, possui olhos, respira, possui dedos ou ate mesmo se alimenta, nesse exemplo isso seria inútil para o algoritmo, uma vez que ambos os elementos comparados possuem essas características, assim, além de ser desprezível para o algoritmo, poderia até ser prejudicial para sua execução. O objetivo desse projeto é criar um sistema que ajude no tratamento de dados para predição de partidas de futebol, sendo assim, utilizamos durante a execução do projeto bases de dados disponíveis no site football-data (*CITE*), e algumas das variáveis que utilizadas foram:

- FTHG = Full Time Home Goals
- FTAG = Full Time Away Team Goals
- FTR = Full Time Result
- HTHG = Half Time Home Team Goals
- HTAG = Half Time Away Team Goals
- HTR = Half Time Result
- HS = Home Team Shots
- AS = Away Team Shots
- HST = Home Team Shots on Target

- AST = Away Team Shots on Target
- HHW = Home Team Hit Woodwork
- AHW = Away Team Hit Woodwork
- HC = Home Team Corners
- AC = Away Team Corners

Essas variáveis apresentam importância significativa para os algoritmos, uma vez que podem determinar o andamento da partida e assim o resultado final.

Para auxiliar esse processo de escolha de variáveis, foi criado um sistema que funciona da seguinte maneira: a página inicial é composta por dois elementos, sendo um botão do tipo input, que permite ao usuário a escolha da base de dados presente na máquina, e um outro botão do tipo submit que confirma o envio da base de dados para a pasta do projeto, para que seja utilizada nas demais telas. A segunda página é composta por uma tabela que lista todos os dados da base de dados, permitindo a visualização completa desses dados. Nessa tela também estão presentes um elemento select multiple, que permite que o usuário escolha as colunas que irá usar para a construção de uma nova base de dados apenas com os dados selecionados pelo usuário. A terceira página é responsável por listar os dados que foram selecionados na página anterior e gerar a base de dados limpa.

4 CONCLUSÃO

Referências

- BROOKS, J.; KERR, M.; GUTTAG, J. Using machine learning to draw inferences from pass location data in soccer. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, Wiley, v. 9, n. 5, p. 338–349, jun 2016.
- CHADE, J. *China faz Copa de 2018 ter audiência recorde de 3,5 bilhões de pessoas*. 2018. Disponível em: <https://esportes.estadao.com.br/noticias/futebol, china-faz-copa-de-2018-ter-audiencia-recorde-de-3-5-bilhoes-de-pessoas,70002654539>. Acesso em: 17 jun 2019.
- CONSTANTINOU, A. C.; FENTON, N. E.; NEIL, M. Profiting from an inefficient association football gambling market: Prediction, risk and uncertainty using bayesian networks. *Knowledge-Based Systems*, Elsevier BV, v. 50, p. 60–86, sep 2013.
- DUARTE, L. M. da S.; SOARES, C.; TEIXEIRA, J. *Previsão de resultados de jogos de futebol*. Dissertação (Mestrado) — Faculdade da Engenharia da Universidade do Porto, 2015.
- FIFA. *FIFA Financial Report 2018*. 2018. Disponível em: <https://resources.fifa.com/image/upload/xzshsoe2ayttyquuxhq0.pdf>. Acesso em: 17 jun 2019.
- HUCALJUK, J.; RAKIPOVIĆ, A. Predicting football scores using machine learning techniques. *2011 Proceedings of the 34th International Convention MIPRO*, may 2011.
- IGIRI, C. P. Support vector machine–based prediction system for a football match result. *IOSR Journal of Computer Engineering*, v. 17, n. 3, p. 21–26, jun 2015.
- MARTINS, R. G. et al. Exploring polynomial classifier to predict match results in football championships. *Expert Systems with Applications*, Elsevier BV, v. 83, p. 79–93, oct 2017.
- OWRAMIPUR, P. E. F.; MOZNEB, F. S. Football result prediction with bayesian network in spanish league-barcelona team. *International Journal of Computer Theory and Engineering*, October 2013.
- PENDHARKAR, P. C.; KHOSROWPOUR, M.; RODGER, J. A. Application of bayesian network classifiers and data envelopment analysis for mining breast cancer patterns. *Journal of Computer Information Systems*, v. 40, n. 4, p. 127–132, 2000.
- PERIN, C.; VUILLEMOT, R.; FEKETE, J. D. Soccerstories: A kick-off for visual soccer analysis. *IEEE Transactions on Visualization and Computer Graphics*, v. 19, n. 12, p. 2506–2515, Dec 2013. ISSN 1077-2626.
- SFEIR, M. N. Laws of the game (adapted from fifa 2010-11). *World Literature Today*, v. 85, n. 3, p. 38–39, may 2011.
- TAX, N.; JOUSTRA, Y. Predicting the dutch football competition using public data: A machine learning approach. Unpublished, 2015.
- ULMER, B.; FERNANDEZ, M. *Predicting Soccer Match Results in the English Premier League*. Tese (Doutorado) — Stanford University, 2013.
- WEKA. Disponível em: <https://www.cs.waikato.ac.nz/ml/weka/>.