PROPOSTA DE TRABALHO
MAC0499 - Trabalho de formatura supervisionado
Março/2023

Aluno: Lorenzo Bertin Salvador
Responsável: Profª. Dra. Nina S. T. Hirata
Orientadora: Profª. Dra. Kelly Rosa Braghetto

Introdução:

O futebol é o esporte mais popular do mundo e está presente na vida de milhões de pessoas, dentro e fora do Brasil. Os jogadores quase sempre são os protagonistas e os valores envolvidos no mercado de transferências, durante e após todas as temporadas, são estratosféricos. A importância dada para determinado futebolista, que justifica seu valor de mercado, pode ser baseada em diversos elementos, sendo os principais as participações em gols (gols e assistências) e os títulos conquistados. Além disso, as discussões entre apaixonados pelo esporte são sempre em volta de comparações para decidir os melhores jogadores de cada clube, seleção, temporada ou até mesmo da história.

Desse contexto, surge a ideia de avaliar a qualidade individual dos jogadores a cada ano. Para isso, existem algumas maneiras já consolidadas dentro do mundo do futebol. As principais são os dois prêmios chamados de “Bolas de Ouro”, um da FIFA (Federação Internacional de Futebol) e o outro da revista francesa *France Football*. Como o critério das escolhas desses prêmios quase nunca é sistemático e, portanto, pode gerar incoerências, é possível comparar as estatísticas de cada jogador para chegar em conclusões mais precisas, utilizando diversos sites que mantêm extensas bases de dados para incluir a maioria das competições profissionais pelo mundo, atualizadas e públicas. O maior destaque é a plataforma alemã *Transfermarkt* (https://www.transfermarkt.com/), mas também existem brasileiros realizando esse trabalho, como é o caso do *OGol* (https://www.ogol.com.br/).

Além disso, há trabalhos científicos que tratam sobre ranqueamento de esportistas, inclusive em outros esportes diferentes do futebol, como *Koenigsberg, A., Pilgrim, J., & Baker, J. (2020)*[1], que trata do golf, e também *Xia, V., Jain, K., Krishna, A., & Brinton, C. G. (2018)*[2], que trata do basquete e futebol americano. Para este TCC, a fonte mais relevante é o artigo do professor de Educação Física na UFMA, Emanuel Péricles Salvador[3], que propõe uma fórmula para calcular a qualidade da carreira completa de um jogador, considerando tanto jogadores da atualidade, como também aposentados. De acordo com esse trabalho, os componentes da carreira de um jogador passíveis de análise são divididos em três grupos: as estatísticas individuais (gols, assistências, hat-tricks, etc.), os títulos

[1] Koenigsberg A, Pilgrim J, Baker J. Generational differences in the ranking pathways of top 100 ranked golfers. *J Sports Sci*.
[2] Xia, V., Jain, K., Krishna, A., & Brinton, C. G. (2018). A network-driven methodology for sports ranking and prediction. In *2018 52nd Annual Conference on Information Sciences and Systems, CISS 2018* (pp. 1-6)
[3] SALVADOR, E.P. et al. ModK: Formula for Determining the Best Season and Career of a Football Player by Objective Indicators. *The Open Sports Sciences Journal*, 2022, Volume 15.

(com diferença de peso de acordo com a importância do título) e o impacto do jogador no clube (isto é, ganhar um título em um clube que nunca ganhou é mais importante do que ganhar o mesmo campeonato em um clube multicampeão).

Objetivos:

O trabalho de conclusão de curso terá como principal objetivo o desenvolvimento de um sistema que deve permitir a aplicação da fórmula proposta pelo prof. Emanuel Péricles Salvador. Além disso, outro objetivo, que só poderá ser alcançado após a realização do primeiro, é a produção de uma análise comparativa entre os resultados obtidos a partir da fórmula e outras maneiras de se mensurar a performance de um jogador em comparação com outros futebolistas. Por fim, o último objetivo que deve ser atingido pelo projeto é a possibilidade de que usuários que não têm conhecimento em computação possam utilizar o sistema, de modo que profissionais de outras áreas, como a própria Educação Física ou até o Jornalismo, por exemplo, consigam interagir com a aplicação.

Metodologia:

O sistema a ser desenvolvido possui cinco grandes componentes. O primeiro componente se trata do extrator de dados, que utiliza fontes públicas para obter as informações relevantes da carreira de cada jogador. O segundo componente é o banco de dados operacional (BDop), que deve ser capaz de gerenciar as informações obtidas anteriormente. O terceiro componente é o código que realiza o processamento da fórmula, recebendo as informações do BDop e devolvendo as pontuações dos jogadores fornecidas pelo cálculo. O quarto componente é a interface de visualização, que exibe tanto as informações obtidas pelo extrator de dados, como as geradas pela fórmula. O quinto e último componente é o banco de dados analítico (BDan), que é fruto do transporte do BDop para um banco de dados de grafos. O BDan poderá permitir tanto a própria validação das pontuações obtidas pela fórmula, como também a realização de outras interpretações, que poderão ser feitas usando as próprias métricas fornecidas pela estrutura dos grafos. Quanto à linguagem de programação, a ideia é que a implementação da maioria dos componentes seja feita utilizando Python. A seguir, cada componente será descrito em maior detalhes.

Primeiramente, com relação à extração dos dados, a informação poderá ser proveniente de pelo menos duas das seguintes fontes: o site transfermarkt.com, o site ogol.com.br e diversas tabelas contendo estatísticas de jogadores que não estão presentes ou completas nos sites. Para as duas primeiras fontes, será necessário realizar web scraping em diversas URLs desses sites, enquanto que para a última fonte será necessário inserir as informações diretamente no BDop.

Quanto à fórmula, é importante notar que o artigo que a descreve será usado como bibliografia, portanto, a ideia não é alterá-la e nem analisar suas componentes, apenas aplicá-la com os dados obtidos e armazenados no BDop. As informações utilizadas pela fórmula podem ser divididas em cinco elementos principais e um secundário. Os principais são os jogadores, clubes, campeonatos, países e continentes. O elemento secundário, presente em diversos dos anteriores, é a temporada (período de um ano no futebol), que

descreve a variação ao longo do tempo das importâncias dadas para campeonatos, países e continentes.

Com relação aos bancos de dados, dois tipos de BDs serão utilizados neste trabalho: os já mencionados BDop (operacional) e BDan (analítico). As principais diferenças entre os dois são o formato e a finalidade, enquanto a semelhança é o conteúdo de ambos. O BDop será o principal, e é nele que as informações serão inicialmente armazenadas, advindas tanto da extração dos dados quanto do cálculo da fórmula. Nele será necessário registrar, atualizar e remover os dados obtidos, sendo o principal alvo da interação entre usuário e sistema. A ideia é que o BDop seja um banco de dados orientado a documentos, para que a interação entre linguagem de programação (Python) e o BD aconteça mais facilmente. O BDan, por outro lado, deve ser uma "cópia" do BDop que utiliza a estrutura de grafos para representar os dados. A ideia é que os nós dos grafos sejam os jogadores, ou clubes, e as arestas poderão ser a atuação conjunta entre dois jogadores em algum clube em uma determinada temporada, por exemplo. A finalidade do BDan é permitir a análise das informações utilizando o próprio contexto fornecido pela estrutura dos grafos.

Para que seja possível visualizar e manipular todos os dados, é necessário que seja desenvolvida uma interface, em formato de *dashboard,* que seja intuitiva para alguém que não seja da área de computação. A interface é responsável por exibir os dados, possivelmente em formato de gráficos e tabelas, e possibilitar o registro, a atualização e a remoção das informações, interagindo diretamente com o BDop. Além disso, a interface deve permitir a extração dos dados (a princípio em formato CSV), tanto na forma bruta, como também aplicando filtros para essa ação. Ela também será responsável pela visualização do BDan, que é muito mais rica que a visualização do BDop.

A fim de que todos os componentes citados anteriormente se comportem como o esperado, a criação de testes será necessária. Dessa maneira, o comportamento do sistema pode ser monitorado, gerando maior estabilidade e eficiência ao longo e após o processo de implementação. Outro ponto importante é que, como há informações que serão advindas de fontes externas (páginas da web), é possível que a estrutura desses sites seja alterada durante o desenvolvimento do sistema. Por esse motivo, os testes na extração dos dados serão essenciais para assegurar que a informação correta está sendo registrada no banco de dados.

Por fim, a última consideração a ser feita é que, para embasar as decisões de projeto realizadas ao longo do trabalho, como a escolha das ferramentas de interface ou a escolha dos sistemas gerenciadores de bancos de dados, serão feitas análises comparativas entre as opções disponíveis, de modo que fiquem claros os critérios que justificam as escolhas tomadas durante o processo de desenvolvimento.

Atividades previstas:

A principal atividade do TCC é a implementação do sistema com todas as propriedades e funcionalidades descritas em **Objetivos** e **Metodologia**. As etapas necessárias são: modelagem e implementação do banco de dados operacional, criação dos procedimentos para a extração automática dos dados das fontes e carga no banco de dados, codificação da fórmula que calcula uma pontuação para as carreiras dos jogadores e o desenvolvimento da interface de visualização e consulta.

A atividade seguinte, após a população do banco de dados operacional, é a criação dos procedimentos de transporte desses dados para o BD de grafos. Por fim, a última

atividade relacionada ao sistema, que apenas pode ser iniciada quando todas as outras forem completadas, é a análise do grafo que foi gerado na etapa anterior.

É importante notar que, após finalizar cada uma das atividades, a monografia deve ser atualizada, a fim de que o trabalho de escrita não se concentre apenas após o término da implementação do sistema. No fim do ano, o processo de escrita seria apenas concluído, revisando e complementando o texto já produzido anteriormente.

Cronograma:

Período para execução / **Atividades**	MAR	ABR	MAI	JUN	JUL	AGO	SET	OUT	NOV	DEZ
Decisões de projeto, entregas iniciais (resumo, proposta de trabalho, etc.)	X									
Extrator de dados (web scraping e fonte manual)		X	X							
Modelagem e criação do Banco de dados			X	X	X					
Codificação da fórmula				X	X					
Interface Gráfica						X	X			
BD Grafos (procedimento de transporte)								X		
BD Grafos (análise)								X	X	
Escrita da monografia			X	X	X	X	X	X	X	
Apresentação										X