Análise de fatores influentes no valor das transferências de jogadores de futebol: Uma abordagem *Big Data*

Diogo Silva
Departamento de Engenharia Informática
Instituto Superior de Engenharia do Porto
Lixa, Portugal
1231420@isep.ipp.pt

Filipe Vinhas
Departamento de Engenharia Informática
Instituto Superior de Engenharia do Porto
Vila do Conde, Portugal
1231425@isep.ipp.pt

Gil Almeida
Departamento de Engenharia Informática
Instituto Superior de Engenharia do Porto
Espinho, Portugal
1231426@isep.ipp.pt

***Abstract*— O mercado de transferências no futebol é dinâmico e muitas vezes difícil de prever. Este trabalho propõe uma arquitetura de *Big Data* para prever o valor de transferência de jogadores com base nas suas características. Utilizando dados do Transfer Market, a proposta inclui desde o processamento de grandes volumes de dados até a construção de modelos preditivos com técnicas de machine learning, além de uma camada de visualização para análise dos resultados. A abordagem é aplicada a um conjunto de dados históricos, com foco na integração de tecnologias como SQL Server, Apache Spark, Power BI, TensorFlow, entres outras.**

Keywords*— *Big Data, Football Transfer Market, Machine Learning, Predictive Analytics, Apache Spark, Data Visualization, Transfer Fee Prediction.

I. INTRODUÇÃO

O mercado de transferências de jogadores de futebol movimenta milhares de milhões de euros todos os anos e desempenha um papel fundamental na estratégia de negócios de clubes, agentes e investidores. O valor de mercado de um jogador é influenciado por uma vasta gama de fatores, desde a idade e o clube em que atua, até às suas estatísticas de desempenho em campo, estado físico, entre outros. Prever com precisão o valor de transferência de um jogador não só permite uma melhor negociação para os clubes, mas também oferece uma vantagem competitiva no mercado.

Este projeto propõe-se a explorar o uso de técnicas de *Big Data* e *machine learning* para enfrentar o desafio de prever valores de transferência no futebol. Através da análise de grandes volumes de dados históricos e da construção de modelos preditivos robustos, este projeto visa oferecer uma solução escalável e eficiente para a previsão de valores de mercado, utilizando tecnologias como Apache Spark para processamento distribuído e SQLServer para armazenamento de dados estruturados.

O objetivo deste estudo é não só construir um modelo preditivo com elevada precisão, mas também demonstrar como a arquitetura de *Big Data* proposta pode ser aplicada de forma eficaz no domínio do desporto, especificamente no futebol. A proposta integra componentes como a recolha, armazenamento, processamento, análise e visualização de dados, promovendo um ciclo completo de gestão e análise de dados em larga escala.

II. DEFINIÇÃO DO PROBLEMA

A previsão do valor de mercado de um jogador de futebol envolve diversas variáveis, incluindo atributos pessoais (como idade, altura e posição), estatísticas de desempenho (como golos, assistências e jogos disputados) e fatores contextuais (como clube atual e campeonato em que joga). Para estabelecer relações entre estas variáveis e o valor de transferência, serão empregues técnicas de *machine learning* e análise de dados avançada, que têm a capacidade de identificar padrões em dados históricos e prever valores futuros com precisão.

O objetivo central deste projeto é desenvolver uma solução que processe grandes volumes de dados de maneira eficiente e extraia insights valiosos para modelagem preditiva de alta precisão. Um dos principais desafios reside em projetar uma arquitetura que não apenas suporte o processamento em larga escala, mas que também ofereça agilidade e flexibilidade na integração e análise dos dados, garantindo resultados consistentes para a previsão dos valores de transferência no futebol.

Este problema insere-se no campo da análise preditiva e *machine learning* aplicados a dados desportivos, com a proposta de contribuir para decisões mais informadas no mercado de transferências. Utilizando dados históricos provenientes do Transfermarkt, o objetivo é criar um modelo preditivo capaz de estimar o valor de transferência de um jogador de futebol, levando em consideração as suas características individuais e desempenho passado.

III. REVISÃO DA LITERATURA

A. *Fatores Influenciadores do Valor de Mercado de Jogadores no Futebol*

A previsão de valores de mercado de jogadores no futebol tem-se tornado uma prática fundamental para a gestão e tomada de decisão dos clubes desportivos, especialmente no que diz respeito a contratações e vendas de atletas. Avaliar o valor de mercado de um jogador é uma tarefa complexa, que vai muito além da análise das suas habilidades desportivas.

Estudos recentes mostram que *Big Data* e machine learning desempenham um papel cada vez mais importante na análise de desempenho desportivo e na previsão de valores de mercado, permitindo prever com maior precisão o valor de mercado de jogadores através de uma análise abrangente de variáveis individuais [1].

Existem vários fatores que que influenciam o valor de mercado de um jogador de futebol, incluindo [2]:

- **Idade:** É um dos preditores mais fortes. Jogadores mais jovens geralmente têm valores mais altos devido ao potencial de desenvolvimento e à expectativa de uma

carreira longa, enquanto jogadores mais velhos tendem a uma valorização menor.

- **Posição em Campo:** A posição do jogador também é crucial. Jogadores em posições ofensivas costumam ter valores de mercado mais altos, enquanto defesas e guarda-redes geralmente têm valores mais baixos.
- **Desempenho e Estatísticas Anteriores:** Golos, assistências, minutos jogados e outras métricas específicas por posição são indicadores sólidos do valor. O desempenho recente pode impulsionar o valor de mercado de um jogador, refletindo a sua forma atual.
- **Popularidade e Presença Mediática:** Dados de redes sociais e audiências aumentam a visibilidade e o valor de mercado dos jogadores.
- **Condições Contratuais:** A duração restante do contrato e a presença de cláusulas de rescisão influenciam o valor. Contratos longos e cláusulas elevadas tendem a aumentar a avaliação de mercado.
- **Características dos Clubes:** Fatores como a situação financeira e a performance do clube comprador e vendedor afetam o poder de negociação e, por consequência, o valor do jogador.

Com a utilização de *machine learning*, é possível avaliar melhor estas variáveis, ajudando a compreender quais características influenciam mais o valor de mercado dos jogadores[3].

No entanto, a previsão de valores de mercado enfrenta desafios, como a variabilidade dos dados de entrada e a necessidade de ajustar os modelos conforme a posição de jogo. Por exemplo, variáveis como golos e assistências são mais relevantes para avançados, enquanto métricas como cortes e defesas são fundamentais para avaliar defesas e guarda-redes. Além disso, fatores externos, como o impacto da COVID-19 no mercado de transferências, introduzem incertezas adicionais, dificultando a precisão dos modelos preditivos[4].

B. *Arquiteturas de Big Data em Projetos na Área Desportiva*

O aumento exponencial dos dados desportivos, impulsionado pela proliferação de dispositivos da Internet das Coisas (IoT) e *wearables*, tem tornado as arquiteturas de *Big Data* essenciais para o processamento, análise e visualização de grandes volumes de informações. Esse crescimento demanda arquiteturas robustas e distribuídas, que suportem desde a recolha e armazenamento até ao processamento de dados em variados níveis de estrutura e velocidade.

As arquiteturas de *Big Data* no desporto geralmente consistem em três componentes principais: recolha e armazenamento, processamento e visualização/análise. A camada de recolha integra dados de várias fontes, incluindo históricos e dados em tempo real provenientes de dispositivos de monitoramento. Ferramentas como o Apache Kafka e o Flume são empregues para gerir esses fluxos contínuos, transferindo-os para sistemas distribuídos como o HDFS (*Hadoop Distributed File System*) e bases de dados NoSQL, garantindo a escalabilidade e a flexibilidade necessárias para lidar com dados não estruturados, como vídeos e sensores[5].

Na camada de processamento, *frameworks* como o Apache Spark são fundamentais para análises em tempo real e em *batch*, ambas essenciais no contexto desportivo, onde uma resposta rápida é um diferencial competitivo. O uso de algoritmos de *machine learning*, aplicados com o Apache Spark, permite prever eventos críticos, como lesões, e otimizar o desempenho dos atletas. Por exemplo, algoritmos de *clustering*, como o K-means, são utilizados para identificar padrões de sobrecarga física que podem levar a lesões, enquanto redes neurais são aplicadas para análise de padrões complexos em dados de performance. Esta abordagem com *machine learning* é facilitada pela utilização de containers, como Docker, que oferecem um ambiente flexível e escalável, proporcionando alta disponibilidade e baixa latência para análises críticas, algo que é essencial para monitorizações em tempo real [6], [7], [8].

A implementação de arquiteturas de *Big Data* no desporto visa analisar tanto dados estruturados, como estatísticas de jogo, quanto monitorizar em tempo real o desempenho dos jogadores. Dados estruturados permitem análises detalhadas, enquanto dados não estruturados, como vídeos e dados de movimento, requerem técnicas de processamento distribuído, como o *MapReduce*, para lidar com a sua complexidade [9]. A análise em tempo real possibilita a monitorização contínua do esforço físico dos atletas, permitindo que as equipas ajustem as cargas de treino para prevenir lesões. Esta prática é comum no desporto de alto rendimento, onde dados de IoT são utilizados para suportar decisões imediatas [7].

Embora a *Big Data* traga significativas vantagens para o desporto, a sua aplicação enfrenta desafios, especialmente no que diz respeito à privacidade e à segurança dos dados dos atletas. Como estes dados frequentemente incluem informações sensíveis de saúde e desempenho, recolhidos por dispositivos IoT, é essencial que existam políticas de proteção em conformidade com as regulamentações. Garantir a confidencialidade e integridade desses dados é fundamental para proteger a privacidade dos atletas e manter a confiança na tecnologia [6].

C. *Modelos de Machine Learning para Previsão no Futebol*

Modelos de *machine learning* têm sido amplamente utilizados para prever o valor e o desempenho de jogadores de futebol, destacando-se a aplicação de modelos de regressão e de classificação. Modelos como Regressão Linear, Lasso, *Elastic Net* e os baseados em árvores, como o *Random Forest* e o *Gradient Boosting*, são implementados para estimar o valor de mercado dos jogadores, analisando atributos como capacidades físicas, idade, posição e histórico de desempenho. Estudos demonstram que os modelos baseados em árvores frequentemente superam os modelos lineares, sugerindo que as relações não lineares entre os atributos são essenciais para previsões mais precisas. Além disso, técnicas de classificação, como o *Random Forest* e o *Support Vector Machines* (SVM), foram aplicadas para agrupar jogadores em posições específicas e prever métricas de desempenho, como número de golos e assistências, proporcionando uma análise detalhada do impacto potencial de cada jogador. Métodos de seleção de características e de equilíbrio de classes foram usados para melhorar a precisão, ao lidar com a elevada dimensionalidade dos dados e corrigir distribuições desiguais de jogadores por posição. Para otimizar a precisão, técnicas de ensaio e ajuste de Hiper parâmetros, como o *Gradient Boosting* com *Tree-structured Parzen Estimator* (TPE), ajudam a identificar os atributos mais influentes para cada posição, aprimorando a

performance dos modelos. A avaliação destes modelos é realizada com métricas como o *Mean Absolute Error* (MAE), o *Root Mean Squared Error* (RMSE) e o R^2, que quantificam a precisão e permitem uma comparação rigorosa da eficácia de diferentes abordagens preditivas, facilitando a sua aplicação no contexto desportivo [10], [11], [12], [13].

Os modelos de *deep learning* também são destacados em estudos, como " Predicting the Football Players' Market Value Using Neural Network Model: A Data-Driven Approach", que discute o potencial de redes neurais para capturar padrões complexos nos dados, tornando-se uma ferramenta promissora para análises preditivas de valores de mercado num contexto desportivo altamente dinâmico[14].

D. *Visualização e Análise de Resultados no Big Data*

A visualização de dados desempenha um papel fundamental na análise de grandes volumes de informações, especialmente no domínio do desporto, onde a rapidez e a precisão na tomada de decisões são essenciais. No entanto, a visualização de dados em *Big Data* enfrenta desafios significativos, como a escalabilidade percetual, que dificulta a extração de informações relevantes em grandes volumes de dados, e a necessidade de apresentação em tempo real, que exige que os sistemas consigam processar e visualizar informações rapidamente [15], [16]. Técnicas como gráficos de linhas, mapas e gráficos de bolhas são comuns para tornar dados complexos mais compreensíveis [15], [16].

Existem várias ferramentas disponíveis que facilitam a visualização de dados, cada uma com características distintas. As principais incluem [15], [16]:

1. **Tableau**: Uma das ferramentas mais populares, conhecida pela sua capacidade de criar *dashboards* interativos e gráficos complexos. É amplamente utilizada em diversas indústrias, incluindo o desporto, devido à sua facilidade de uso e robustez.
2. **Qlikview**: Este software é um forte concorrente do Tableau e é reconhecido pela sua capacidade de integração de análises em tempo real. É especialmente valorizado por empresas que necessitam de uma visualização dinâmica de dados.
3. **Microsoft Power BI**: Oferece uma interface intuitiva e permite a criação de relatórios interativos e *dashboards*. É uma boa opção para organizações que já utilizam produtos da Microsoft, dada a sua integração com outras ferramentas.
4. **Sisense**: Esta plataforma é escalável e capaz de lidar com grandes volumes de dados. É utilizada por grandes organizações e é especialmente eficaz em análises avançadas.
5. **D3.js**: Uma biblioteca JavaScript altamente personalizável para visualizações web. É ideal para desenvolvedores que desejam criar visualizações interativas e dinâmicas, embora exija conhecimentos técnicos.

Em resumo, Tableau e Microsoft Power BI são frequentemente considerados as melhores pela sua facilidade de uso e robustez. No entanto, a escolha da ferramenta "ideal" deve considerar fatores como o tipo de dados a serem visualizados, a complexidade da análise e as preferências da equipa de trabalho.

IV. ARQUITETURA DE *BIG DATA*

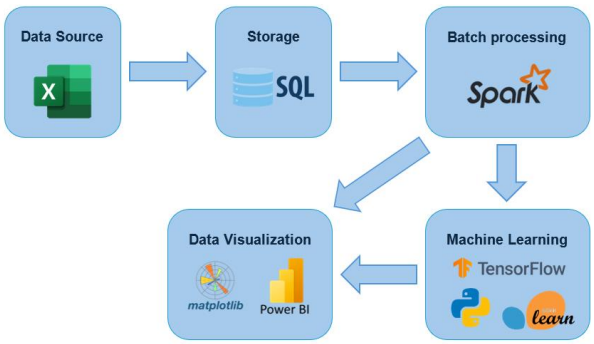

Figure 1 - Big Data Architecture.

A *figure 1* descreve a arquitetura de *Big Data* que iremos adotar para este projeto, assim como as ferramentas que serão utilizadas em cada componente da mesma.

A. *Fonte de Dados (Data Source)*

A arquitetura inicia-se com a recolha de dados estruturados provenientes de um dataset em formato tabular. Esse conjunto de dados contém informações detalhadas sobre jogadores de futebol, organizadas em atributos específicos como idade, posição, valor de mercado, entre outros. O objetivo inicial da fonte de dados é fornecer uma entrada padronizada para o pipeline, que pode ser inspecionada e carregada diretamente para o ambiente de armazenamento.

B. *Armazenamento (Storage)*

Após a recolha, os dados são enviados para uma camada de armazenamento relacional. Como o conjunto de dados possui uma estrutura tabular bem definida, optamos por uma base de dados relacional para organizar e consultar as informações de forma eficiente. Este armazenamento relacional é especialmente vantajoso para dados estruturados e fornece uma base escalável para eventuais expansões e aumento de complexidade no pipeline.

Embora os dados originais venham de um CSV estático, o uso de uma base de dados SQL permite preparar o pipeline para futuras necessidades de *Big Data*. Esta solução facilita a realização de consultas rápidas e complexas, proporcionando uma infraestrutura robusta e flexível, caso haja um aumento no volume de dados ou a necessidade de integração com outras fontes. A escolha de uma base de dados relacional também cria uma estrutura bem consolidada para integração com ferramentas de processamento e visualização, ampliando o potencial do pipeline.

Dado o carácter estruturado dos dados do projeto (tabelas com colunas fixas, como nome do jogador, posição, idade e valor de mercado), um Data Lake não é necessário neste contexto. Um Data Lake seria mais apropriado em cenários com dados não estruturados ou semi-estruturados, como logs ou imagens, que requerem armazenamento mais flexível e escalável para múltiplos formatos de dados [17].

C. *Processamento em Lote (Batch Processing)*

Após o armazenamento, os dados passam por uma etapa de processamento em lote, ou *batch processing*, que realiza transformações, limpeza e agregações de maneira eficiente. Esta etapa visa transformar os dados brutos em informações

mais trabalháveis, preparando-os para análises posteriores e potenciais modelos de *machine learning*. O processamento em *batch* permite também manipulações em grandes volumes de dados, onde operações de filtragem e agregação são essenciais.

D. Machine Learning

Com os dados transformados, a próxima etapa consiste na modelagem e análise preditiva. Nesta fase, são aplicados algoritmos de *machine learning*, desde modelos tradicionais até redes neurais, para identificar padrões e realizar previsões. Nesta fase serão exploradas diferentes abordagens, e encontradas as mais adequadas para prever o valor estimado de transferências de jogadores, com base nos atributos.

E. Visualização de Dados (Data Visualization)

A etapa final é a visualização dos dados. Este componente permite que as informações e previsões geradas ao longo do pipeline sejam transformadas em insights visuais, como *dashboards* e gráficos, facilitando a análise de resultados e tendências. A visualização serve tanto para exibir dados provenientes diretamente do processamento em *batch* quanto para exibir os resultados de previsões dos modelos de *machine learning*.

V. Explicação do Fluxo de Dados

1. **Data source (CSV):** Os dados de entrada vêm de um dataset em CSV extraído do Kaggle (Football Data from Transfermarkt: https://www.kaggle.com/datasets/davidcariboo/player-scores). Estes dados serão carregados para um *Jupyter Notebooks*, com a ajuda do *Pandas*, para uma limpeza inicial dos dados, se necessário, e serão depois armazenados numa base de dados relacional (SQL Server).
2. **Storage (SQL Server):** Aqui, os dados estruturados são guardados. Será possível realizar *queries* SQL para analisar as características dos jogadores e definir subconjuntos de dados para processamento adicional.
3. **Batch processing (Apache Spark):** Uma vez armazenados, os dados são processados em *batch* pelo Apache Spark. O Spark é usado para realizar operações de limpeza, transformação (como normalizar variáveis ou criar *features*) e preparar os dados para os modelos de *machine learning*.
4. **Machine Learning (TensorFlow/Scikit-learn)**: Os dados processados são passados para os modelos de *machine learning*. É nesta fase que serão testados diferentes algoritmos (Regressão Linear, *Random Forest*, Redes Neurais, …) para prever os valores de transferência. O TensorFlow será utilizado para implementar redes neurais mais complexas, enquanto o Scikit-Learn será empregue para modelos mais clássicos.
5. **Visualização (Power BI/Matplotlib):** É nesta fase que os dados serão visualizados, tanto as previsões feitas na etapa de *machine learning*, como os dados provenientes diretamente do *batch proccessing*. Será utilizado o Power BI para criar *dashboards* dinâmicos que mostram as previsões, tendências, e possíveis insights sobre os valores dos jogadores. Matplotlib ou Seaborn são usados para gráficos detalhados como distribuições de erro, gráficos de dispersão entre características dos jogadores e valores preditivos.

VI. Tecnologias e Frameworks

Esta seção justifica a seleção das tecnologias e *frameworks* para cada etapa do pipeline, oferecendo uma visão mais detalhada das alternativas e dos critérios usados para a decisão.

A. Fonte de dados: Ficheiro CSV

O conjunto de dados inicial será fornecido em formato CSV, amplamente utilizado para armazenar dados estruturados em formato tabular, sendo adequado para as necessidades de processamento e análise deste projeto.

Este dataset contém dados estruturados sobre jogadores de futebol, incluindo características como idade, posição, desempenho, clube, entre outros.

Como o dataset de transferências do Transfermarkt é fornecido em formato CSV, o uso do Excel torna-se útil para inspecionar rapidamente os dados. Utilizaremos o Pandas em Python para carregar o ficheiro CSV diretamente do Excel, o que permite transformar o CSV num DataFrame que pode ser facilmente manipulado e passado para etapas posteriores do pipeline.

O CSV foi escolhido por ser um formato simples, de fácil manipulação e leitura, além de compatível com ferramentas de processamento como SQL Server e Apache Spark. Esta escolha facilita a ingestão dos dados no pipeline de *Big Data* e simplifica a integração inicial no processo de análise e modelagem.

B. Storage: SQL Server

O SQL Server é uma base de dados relacional robusta e confiável, ideal para o armazenamento de dados estruturados, como o nosso conjunto de dados. A escolha de um sistema de base de dados relacional, como o SQL Server, permite organizar e consultar os dados com mais eficiência do que seria possível ao usar apenas um simples CSV.

O SQL Server possibilita consultas rápidas e integra-se bem com ferramentas de visualização como o Power BI, otimizando o processo de criação de relatórios. Além disso, oferece importantes recursos para ambientes de *Big Data*, como transações ACID (Atomicidade, Consistência, Isolamento e Durabilidade), recursos de segurança e indexação para otimizar consultas, essenciais em contextos onde a eficiência e a confiabilidade são prioritárias.

Optamos pelo SQL Server em detrimento de outras opções, como o PostgreSQL, também conhecido pela escalabilidade e integração em ambientes de *Big Data*, devido à familiaridade e ao conforto da equipa de trabalho com esta ferramenta específica, o que nos permite uma curva de aprendizagem mais rápida e uma implementação mais eficiente [18].

C. Batch Processing: Apache Spark

O Apache Spark foi escolhido para o componente de processamento em lote (batch processing) devido à sua eficiência e escalabilidade no tratamento de grandes volumes de dados estruturados.

Suporta processamento distribuído e em memória, tornando-o ideal para manipulações e transformações de dados de forma rápida e eficiente num contexto de *Big Data*.

Permite processar dados em larga escala de forma paralela e em batch, o que é essencial para transformar os dados para as próximas etapas do pipeline.

Oferece API de integração direta com Python (PySpark) e SQL, facilitando o uso de algoritmos de *Machine Learning*, transformações complexas e manipulação de dados em larga escala antes de enviar os dados para a etapa de *Machine Learning* ou visualização [19].

D. Machine Learning: TensorFlow e Scikit-Learn

TensorFlow é uma biblioteca robusta e amplamente utilizada para *deep learning*, ideal para construir modelos mais complexos. É especialmente eficaz em projetos de previsão devido ao seu suporte avançado para GPUs, que acelera significativamente o treino de modelos. Além disso, TensorFlow oferece um ecossistema completo, incluindo o Keras, que simplifica a construção de redes neurais através de uma API de alto nível e é uma ferramenta que a equipa de trabalho já está mais familiarizada.

Por outro lado, o Scikit-Learn é uma ferramenta amplamente adotada para machine learning tradicional e modelos preditivos clássicos, como regressão e árvores de decisão. A sua simplicidade e integração com outras ferramentas em Python facilitam o desenvolvimento rápido de protótipos. Dado que o tamanho inicial do dataset de transferências de jogadores não é considerado excessivamente grande, decidimos utilizar o Scikit-Learn nesta fase do projeto.

No entanto, se durante a implementação novos dados forem adicionados e o volume de dados aumentar significativamente, estaremos abertos a considerar a utilização do MLlib do Apache Spark. Esta biblioteca foi projetada para funcionar num ambiente distribuído, permitindo o treino de modelos em conjuntos de dados muito maiores do que os suportados pelo Scikit-Learn. Embora a curva de aprendizagem seja mais acentuada, a capacidade do MLlib de integrar-se com o ecossistema Spark e trabalhar eficientemente com grandes volumes de dados torna-o uma escolha vantajosa em cenários onde o volume de dados é uma preocupação.

A combinação de TensorFlow para *deep learning*, Scikit-Learn para protótipose, potencialmente, MLlib para *Big Data* permite uma abordagem abrangente, cobrindo desde modelos preditivos tradicionais até redes neurais complexas, otimizando assim todo o processo de treino, teste e ajuste dos modelos[20].

E. Data Visualization: Power BI e Matplotlib

O Power BI é ideal para a criação de *dashboards* interativos e visualização de dados. O Power BI recebe os dados previamente processados pelo Apache Spark, o que permite a visualização de informações já transformadas e otimizadas para análise [21], [22].

O Matplotlib é excelente para visualizações em Python, especialmente em ambientes de Jupyter Notebook, permitindo a criação de gráficos personalizados durante o desenvolvimento e análise de modelos de *machine learning*.

Ambas as ferramentas combinam versatilidade e acessibilidade. O Power BI permite criar visualizações interativas que facilitam a análise e a apresentação dos dados, enquanto Matplotlib permite visualizações rápidas durante a análise e experimentação no Jupyter Notebook[23].

VII. Conclusão

Este projeto teve como objetivo explorar e definir uma arquitetura de *Big Data* eficiente para a previsão do valor de transferências de jogadores de futebol, considerando um conjunto diversificado de variáveis. A escolha de uma arquitetura baseada em SQL Server para armazenamento e Apache Spark para processamento em *batch* garante uma infraestrutura escalável e preparada para lidar com volumes crescentes de dados de forma robusta.

Optámos por ferramentas de *machine learning* adequadas à flexibilidade e precisão necessárias para a análise preditiva no desporto, com TensorFlow e Scikit-Learn a oferecerem suporte para treino e tese de modelos com diferentes níveis de complexidade. A visualização com Power BI e Matplotlib facilita a interpretação e apresentação de resultados, aproximando a análise dos utilizadores finais e promovendo insights acessíveis e dinâmicos.

A pesquisa realizada e as ferramentas selecionadas estabelecem uma base sólida para o desenvolvimento futuro de modelos preditivos no contexto desportivo. Esta arquitetura e o conjunto de tecnologias escolhidas visam proporcionar uma infraestrutura ágil e escalável, que responde tanto às necessidades imediatas como a futuras exigências em análise preditiva no mercado de transferências de futebol.

VIII. Referências

[1] C. Li, S. Kampakis, and P. Treleaven, “Machine Learning Modeling to Evaluate the Value of Football Players.” [Online]. Available: https://www2.deloitte.com/uk/en/pages/sports-business-group/articles/deloitte-football-money-

[2] M. Franceschi, J. F. Brocard, F. Follert, and J. J. Gouguet, “Determinants of football players’ valuation: A systematic review,” *J Econ Surv*, vol. 38, no. 3, pp. 577–600, Jul. 2024, doi: 10.1111/joes.12552.

[3] C. Li, S. Kampakis, and P. Treleaven, *Machine Learning Modeling to Evaluate the Value of Football Players*. 2022. doi: 10.48550/arXiv.2207.11361.

[4] M. Elahi, S. Pandey, and S. S. Malhi, “Market Value Prediction Of Football Players,” 2024. [Online]. Available: https://ssrn.com/abstract=4485449

[5] J. Wang, Y. Yang, T. Wang, R. Simon Sherratt, and J. Zhang, “Big data service architecture: A survey,” 2020, *Taiwan Academic Network Management Committee*. doi: 10.3966/160792642020032102008.

[6] Y. M. Özgüven, U. Gönener, and S. Eken, “A Dockerized Big Data Architecture for Sports Analytics,” Jun. 22, 2021. doi: 10.21203/rs.3.rs-524005/v1.

[7] J. Reece and S. Y. Hong, “Big data analytics for smart sports using apache spark,” *Issues in Information Systems*, vol. 22, no. 3, pp. 1–13, 2021, doi: 10.48009/3_iis_2021_1-15.

[8] B. Chambers and M. Zaharia, *Spark: The Definitive Guide Big Data Processing Made Simple*, 1st ed. O’Reilly Media, Inc., 2018.

[9] . IEEE Staff, *2012 IEEE Aerospace Conference*. IEEE, 2012.

[10] V. H. Sayan and E. Hançer, “Techno-Science A SURVEY ON FOOTBALL PLAYER PERFORMANCE AND VALUE ESTIMATION USING MACHINE LEARNING TECHNIQUES,” 2018. [Online]. Available: www.pesstatsdatabase.co

[11] L. Gijs, “Predicting Transfer Value of Professional Football Players Based on Player Skills and Characteristics Using Multiple Linear Regression, Support Vector Regression, and Random Forest Regression,” 2022. [Online]. Available: https://www.kaggle.com/datasets/stefanoleone992/fifa-21-complete-player-dataset.

[12] N. M. Watanabe, S. Shapiro, and J. Drayer, “Big data and analytics in sport management,” *Journal of Sport Management*, vol. 35, no. 3, pp. 197–202, Oct. 2021, doi: 10.1123/JSM.2021-0067.

[13] S. Mahadevan, “Predicting Market Value of Football Players using Machine Learning Algorithms.”

[14] V. Steve Arrul, P. Subramanian, and R. Mafas, “Predicting the Football Players’ Market Value Using Neural Network Model: A Data-Driven Approach,” in *IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics, ICDCECE 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICDCECE53908.2022.9792681.

[15] L. T. Mohammed, A. A. A. Alhabshy, and K. A. Eldahshan, “Big Data Visualization: A Survey,” in *HORA 2022 - 4th International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/HORA55278.2022.9799819.

[16] S. M. Ali, N. Gupta, G. K. Nayak, and R. K. Lenka, “Big data visualization: Tools and challenges,” in *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, 2016, pp. 656–660. doi: 10.1109/IC3I.2016.7918044.

[17] Alexander Yu, “Database vs. Data lake vs. Data warehouse - Redpanda.” Accessed: Oct. 25, 2024. [Online]. Available: https://www.redpanda.com/blog/database-data-lake-data-warehouse-differences

[18] Microsoft, “Databases - SQL Server | Microsoft Learn.” Accessed: Oct. 25, 2024. [Online]. Available: https://learn.microsoft.com/en-us/sql/relational-databases/databases/databases?view=sql-server-ver16

[19] Spark, “Apache SparkTM - Unified Engine for large-scale data analytics,” 2024.

[20] baratchandar venkatapathy, “Comparsion of Scikit Learn ,Tensor flow ,Pytorch.” Accessed: Oct. 26, 2024. [Online]. Available: https://medium.com/@barathchandarcse/comparsion-of-scikit-learn-tensor-flow-pytorch-ee9f43bcc5d6

[21] A. Schneeweiss, “Master Degree Program in Data Science and Advanced Analytics Optimizing Football Coaching Through Data-Driven Insights: A Power BI Dashboard Approach,” 2023.

[22] Arian Forouhandeh, “Power BI for Sports Science.” Accessed: Oct. 27, 2024. [Online]. Available: https://www.scienceforsport.com/power-bi/

[23] The Matplotlib development team, “Matplotlib: Visualization with Python.” Accessed: Oct. 26, 2024. [Online]. Available: https://matplotlib.org