O Estado da Arte em Machine Learning para a Previsão de Resultados de Futebol: Uma Análise de 2017 ao Presente

Resumo

Este relatório apresenta uma análise exaustiva do estado da arte das técnicas de *machine learning* aplicadas à previsão de resultados de jogos de futebol, com um foco específico em metodologias e dados relevantes para a previsão do campeão da Primeira Liga portuguesa. A análise, que abrange o período de 2017 até à atualidade, está estruturada em três domínios principais: uma revisão crítica da literatura científica e académica, um levantamento analítico de aplicações comerciais de previsão e uma análise de implementações de código aberto. As principais conclusões indicam a primazia de modelos baseados em árvores de decisão com *gradient boosting* sobre arquiteturas de *deep learning* para este tipo de problema com dados tabulares, o papel crítico da engenharia de características sofisticada — particularmente métricas temporais e de força das equipas — e a disponibilidade de conjuntos de dados públicos robustos. O relatório sintetiza estas conclusões num conjunto de recomendações estratégicas para o desenvolvimento de um modelo preditivo de alto desempenho, abordando escolhas metodológicas, fontes de dados e protocolos de avaliação, em alinhamento com as melhores práticas académicas e práticas estabelecidas.

1. Introdução

1.1. Contexto e Motivação

A intersecção entre a ciência de dados e a análise desportiva tem registado um crescimento exponencial, consolidando o futebol como um domínio de elevado interesse académico e comercial.[1] Este projeto, que visa prever o vencedor da Primeira Liga portuguesa, insere-se neste contexto dinâmico.[3] A tarefa é inerentemente complexa, dada a natureza "probabilística

e caótica" do desporto.[1] Fatores como o baixo número de golos por jogo, o elevado número de variáveis intervenientes e a imprevisibilidade intrínseca de cada partida tornam a previsão de resultados um desafio computacional e estatístico notável, constituindo um problema de *machine learning* de grande relevância.[4]

1.2. Formulação do Problema

O objetivo principal de prever o campeão de uma liga é decomposto numa série de tarefas subordinadas de *machine learning*. A abordagem fundamental consiste na classificação multiclasse dos resultados de jogos individuais (Vitória da Equipa da Casa, Empate, Vitória da Equipa Visitante). Os resultados destas previsões individuais são subsequentemente agregados através de uma simulação da temporada para projetar a classificação final da liga e determinar a equipa com a maior probabilidade de se sagrar campeã.[3] Este trabalho deve, adicionalmente, cumprir os requisitos formais académicos estipulados pelo regulamento da unidade curricular de Projeto, que exige uma abordagem estruturada, incluindo a contextualização do trabalho, a descrição de trabalhos semelhantes e a comparação de resultados com outras soluções existentes.[3]

1.3. Âmbito e Estrutura do Relatório

Este relatório está estruturado para fornecer uma base sólida para o desenvolvimento do projeto. A primeira parte foca-se na análise do panorama científico e académico, dissecando as metodologias, os algoritmos e os conjuntos de dados mais relevantes. A segunda parte investiga o ecossistema de aplicações comerciais, analisando as suas funcionalidades e modelos de negócio. A terceira parte examina projetos de código aberto para extrair padrões de arquitetura e implementação. Finalmente, o relatório culmina numa secção de síntese que consolida as conclusões e apresenta um conjunto de recomendações estratégicas para a implementação do projeto, em conformidade com as diretrizes do orientador.[3]

Parte I: Panorama Científico e Académico na Previsão de Resultados de Futebol

Esta secção constitui o núcleo técnico do relatório, avaliando criticamente as metodologias, os conjuntos de dados e os quadros de avaliação apresentados na literatura científica desde 2017.

2.1. Paradigmas Metodológicos Dominantes: Uma Perspetiva Evolutiva

A previsão de resultados de futebol evoluiu de modelos estatísticos simples para algoritmos de *machine learning* cada vez mais sofisticados. A literatura recente demonstra uma clara trajetória de desenvolvimento, mantendo, no entanto, a relevância dos modelos mais simples como importantes pontos de referência.

2.1.1. Modelos Estatísticos e Probabilísticos Fundamentais

Apesar do avanço de técnicas mais complexas, os modelos estatísticos continuam a ser um pilar na investigação, servindo frequentemente como *baseline* para avaliar o desempenho de abordagens mais avançadas. Entre estes, os modelos baseados na distribuição de Poisson são proeminentes. Estes modelos tratam a marcação de golos como um evento aleatório que segue uma distribuição de Poisson, permitindo modelar o número de golos marcados por cada equipa e, consequentemente, as probabilidades dos resultados dos jogos.[5] Outra abordagem relevante são os métodos Bayesianos, que têm sido aplicados com sucesso na criação de sistemas de *rating* dinâmicos para as equipas, onde a força de cada equipa é atualizada após cada jogo com base no resultado.[6] A persistência destes modelos na literatura sublinha a sua utilidade para estabelecer um limiar de desempenho que os modelos mais complexos devem superar.

2.1.2. Abordagens Clássicas de Machine Learning

Os algoritmos de aprendizagem supervisionada clássicos formam a base de muitas investigações na área. Estudos publicados desde 2017 continuam a explorar e a refinar o uso de modelos como *Random Forests* [6], *Support Vector Machines* (SVM) [6] e Regressão Logística.[4] A popularidade destes algoritmos deve-se à sua robustez, boa capacidade de generalização em dados tabulares e um grau de interpretabilidade superior ao de modelos mais complexos, o que permite uma análise mais aprofundada da importância das diferentes características preditivas.

2.1.3. A Ascensão do Gradient Boosting

A análise da literatura mais recente revela um consenso claro: os modelos de *gradient boosting* emergiram como o estado da arte para a previsão de resultados de futebol com base em dados tabulares. Algoritmos como XGBoost, LightGBM e CatBoost são consistentemente reportados como os de melhor desempenho, superando tanto os modelos clássicos como as abordagens de *deep learning* em estudos comparativos.[10] O sucesso destes modelos pode ser atribuído a vários fatores: a sua capacidade de capturar interações complexas e não lineares nos dados, mecanismos de regularização internos que combatem

eficazmente o *overfitting* e uma elevada eficiência computacional.[10] A superioridade destes modelos é tão marcada que a sua escolha como algoritmo principal para um novo projeto nesta área é uma decisão fortemente fundamentada pela evidência empírica atual.

2.1.4. A Fronteira do Deep Learning

Apesar do seu sucesso transformador em domínios como a visão por computador e o processamento de linguagem natural, a aplicação de arquiteturas de *deep learning* na previsão de resultados de futebol permanece uma área mais experimental e com resultados mistos. A literatura documenta o uso de Redes Neuronais Artificiais (ANNs) [6], Redes Neuronais Convolucionais (CNNs) — por vezes aplicadas a dados tabulares codificados como imagens [13] — e Redes Neuronais Recorrentes (LSTMs) para capturar a natureza de série temporal do desempenho das equipas.[13] No entanto, um ponto crucial que emerge de revisões sistemáticas recentes é que os modelos de *deep learning* ainda não conseguiram superar consistentemente o desempenho dos modelos de *gradient boosting* neste domínio específico.[13] Esta lacuna de desempenho sugere que, para a estrutura e dimensionalidade dos dados tipicamente disponíveis, a capacidade indutiva dos modelos baseados em árvores de decisão é, atualmente, mais adequada.

2.2. Componentes Centrais dos Modelos Preditivos

O sucesso de um modelo preditivo não depende apenas da escolha do algoritmo, mas também, e talvez de forma mais crítica, da qualidade dos dados e da engenharia de características, bem como do rigor da sua avaliação.

2.2.1. Engenharia de Características e Representação de Dados

A literatura converge na ideia de que a engenharia de características (*feature engineering*) é um dos fatores mais determinantes para o sucesso de um modelo preditivo de resultados de futebol. A transformação de dados brutos em características informativas é o que permite aos algoritmos detetar padrões preditivos. As características utilizadas podem ser categorizadas da seguinte forma [6]:

- **Estatísticas Básicas do Jogo:** Incluem dados diretos como golos marcados e sofridos, remates, remates à baliza, cantos, faltas e cartões.[12]
- **Métricas de Desempenho Engenheiradas:** Para capturar a "forma" ou o momento de uma equipa, são comummente utilizadas médias móveis de estatísticas-chave (e.g., golos marcados nos últimos 5 jogos).[11] Métricas mais avançadas, como os *Expected Goals* (xG), que quantificam a qualidade das oportunidades de golo criadas, também são cada vez mais utilizadas.[15]

- **Ratings de Força das Equipas:** Sistemas de *rating* dinâmicos, como o sistema Elo, são amplamente empregues. Estes sistemas atualizam a pontuação de força de uma equipa após cada jogo, fornecendo uma medida contínua e relativa do seu poderio.[6]
- **Dados de Casas de Apostas:** As *odds* oferecidas pelas casas de apostas são uma característica poderosa, pois representam a expectativa do mercado e, frequentemente, encapsulam informações que não estão presentes nas estatísticas brutas (e.g., lesões de jogadores-chave, moral da equipa).[9]
- **Fatores Contextuais:** Variáveis como a vantagem de jogar em casa (*home advantage*), condições meteorológicas e lesões de jogadores são relevantes, embora a obtenção de dados fiáveis sobre estes últimos seja um desafio em conjuntos de dados públicos.[6]

A criação de um conjunto rico e diversificado de características, combinando dados históricos, métricas de forma e avaliações de força, é fundamental para construir um modelo robusto.

2.2.2. Análise de Conjuntos de Dados Públicos e Suas Limitações

A disponibilidade de dados históricos de alta qualidade é um pré-requisito para qualquer projeto de *machine learning*. Felizmente, existem várias fontes de dados públicos para as ligas europeias de futebol.

- **Fonte Primária Recomendada:** O site football-data.co.uk destaca-se como a fonte mais completa e adequada para este projeto. Disponibiliza dados históricos detalhados para a Primeira Liga portuguesa em formato CSV, abrangendo múltiplas épocas desde 2017.[19] Os ficheiros incluem resultados finais e ao intervalo, estatísticas detalhadas do jogo (remates, cantos, faltas, cartões) e uma vasta gama de *odds* de várias casas de apostas, tanto de pré-jogo como de fecho de mercado.[19] A interpretação correta das abreviaturas nos ficheiros de dados é facilitada pelo ficheiro de notas (notes.txt) fornecido pelo site.[23]
- **Fonte Complementar Avançada (FBRef):** O site FBref.com é outra fonte de dados extremamente rica, focada em estatísticas detalhadas de jogadores e equipas.[36] A sua principal vantagem é a disponibilização de métricas avançadas, como *Expected Goals* (xG) e *Expected Assisted Goals* (xAG), que não se encontram facilmente noutras fontes gratuitas.[36] Embora não exista uma API pública oficial, os dados podem ser extraídos através de *web scraping*, existindo já bibliotecas e projetos de código aberto que facilitam este processo.[38] É crucial, no entanto, respeitar as políticas de utilização do site, que impõem limites de frequência aos pedidos para evitar sobrecarga dos seus servidores.[41] Vários trabalhos académicos já utilizam dados extraídos do FBRef, validando a sua utilidade para a investigação.[11] Esta fonte é ideal para complementar os dados de football-data.co.uk, permitindo a criação de características preditivas mais sofisticadas.
- **Fontes Alternativas e Suplementares:** Outras plataformas como o Kaggle [24], o

OpenFootball [25] e o football-data.org [26] também oferecem conjuntos de dados relevantes que podem ser utilizados para complementar ou validar a fonte principal.

- **Limitações dos Dados:** Uma limitação transversal a quase todos os conjuntos de dados públicos é a ausência de dados mais granulares. Informações como estatísticas individuais por jogador, dados detalhados sobre lesões ou dados de rastreamento espácio-temporal (que capturam a posição dos jogadores em campo) são raramente disponibilizadas publicamente, o que representa um teto para o potencial preditivo dos modelos.[1]

2.2.3. Métricas de Avaliação e Rigor Metodológico

A avaliação robusta de um modelo é crucial para garantir que o seu desempenho é real e generalizável, um ponto enfatizado nas diretrizes do projeto.[3]

- **Para Além da Acurácia:** A utilização da acurácia de classificação como única métrica é inadequada, especialmente devido ao desequilíbrio de classes presente nos resultados de futebol, onde os empates são significativamente menos frequentes do que as vitórias.[6] Este desequilíbrio pode levar a que um modelo que raramente prevê empates atinja uma acurácia enganadoramente alta. Métricas mais apropriadas incluem o F1-*score* por classe (que avalia o desempenho em cada resultado individualmente), a perda logarítmica (*log-loss*, que penaliza previsões excessivamente confiantes e incorretas) e o *Ranked Probability Score* (RPS), que é particularmente adequado para previsões probabilísticas de resultados ordenados.[10]
- **Prevenção de Fuga de Dados (*Data Leakage*):** É imperativo utilizar uma estratégia de validação que respeite a ordem cronológica dos dados. Modelos de previsão desportiva devem ser treinados com dados do passado para prever o futuro. Utilizar validação cruzada padrão (*k-fold cross-validation*) pode levar a que o modelo seja treinado com informações de jogos futuros, resultando numa avaliação de desempenho irrealisticamente otimista. A abordagem correta passa por uma divisão temporal fixa (e.g., treinar com dados até à época N-1 e testar na época N) ou por uma validação cruzada de séries temporais (*time-series cross-validation*).[20]
- **Tratamento do Desequilíbrio de Classes:** Conforme salientado pelo orientador [3], o desequilíbrio de classes, especialmente a sub-representação de empates, deve ser abordado. Uma dificuldade recorrente na literatura é, precisamente, a baixa capacidade dos modelos para prever empates corretamente.[6] Este desafio não é apenas uma questão técnica de desequilíbrio, mas reflete a complexidade das dinâmicas de jogo que levam a um resultado empatado. Técnicas como a ponderação de classes (*class weighting*), que atribui uma penalidade maior aos erros na classe minoritária durante o treino, ou a aplicação de métodos de reamostragem (e.g., SMOTE), podem ser utilizadas para mitigar este problema.

Tabela 1: Análise Comparativa de Algoritmos de Machine Learning na Literatura Recente (2017-Presente)

Categoria do Algoritmo	Algoritmo Específico	Características-Chave Utilizadas	Métrica e Valor de Desempenho Reportado	Pontos Fortes / Fracos Notados	Publicação de Origem
Gradient Boosting	CatBoost, XGBoost	Ratings de equipas (pi-ratings), estatísticas de jogo	RPS: 0.1989 (XGBoost), comparável às *odds*	Melhor desempenho em dados tabulares; supera DL	[10]
Gradient Boosting	XGBoost	Médias móveis de estatísticas de jogo	Acurácia: 53.61% (teste), 56% (semana de jogo)	Desempenho superior, mas dificuldade em prever empates	[11]
Deep Learning	LSTM, RNN	Sequências de desempenho, estatísticas de jogadores	Acurácia de classificação: até 98.63% (muitos-para-um)	Captura de dependências temporais; dados proprietários	[16]
Deep Learning	CNN, Transfer Learning	Dados tabulares convertidos em imagens	-	Desempenho inferior ao CatBoost no mesmo estudo	[13]
Modelos de Ensemble	Random Forest, SVM, Boosting	Atributos de jogo e de jogadores	Retorno económico: 1.58% por aposta	Abordagem robusta que combina múltiplos modelos	[9]
Regressão Logística	Multinomial	Estatísticas de jogo (remates, faltas, cartões)	Acurácia: 59.55%	Modelo simples, boa sensibilidade para vitórias em casa	[12]
Modelos Probabilísticos	Distribuição de Poisson	Golos marcados/sofridos, vantagem casa	-	Modelo interpretável, bom para modelar contagens de golos	[6]

Tabela 2: Visão Geral de Conjuntos de Dados de Futebol Públicos para Ligas Europeias

Fonte de Dados	URL	Cobertura Liga I (PT)	Intervalo Temporal	Conteúdo dos Dados	Custo	Limitações Principais
football-data.co.uk	https://www.football-data.co.uk	**Sim**	1990s-Presente	Estatísticas de jogo, resultados, *odds* de apostas	Gratuito	Ausência de dados de jogadores individuais
FBRef	https://fbref.com/	**Sim**	Variável	Estatísticas avançadas de jogadores e equipas (xG, xAG), resultados, dados de jogo	Gratuito	Sem API oficial; acesso via *scraping* com limites de frequência [39]
Kaggle Datasets	https://www.kaggle.com/datasets	**Sim** (vários)	Variável	Variável (jogos, jogadores, equipas, etc.)	Gratuito	Qualidade e manutenção variáveis entre os conjuntos de dados
OpenFootball	https://openfootball.github.io/	**Sim**	Variável	Resultados, equipas, jogos (formato texto)	Gratuito	Menos detalhe estatístico; formato não tabular
football-data.org	https://www.football-data.org/	**Sim** (API)	Recente	Dados em tempo real e históricos via API	Gratuito (Tier) / Pago	O acesso gratuito é limitado a ligas principais

---

Parte II: Análise de Aplicações Comerciais e Orientadas ao Consumidor

Esta secção analisa as aplicações existentes no mercado para compreender como as tecnologias preditivas são transformadas em produtos, respondendo a uma solicitação direta do orientador do projeto.[3]

3.1. Uma Tipologia de Aplicações de Previsão

O mercado de aplicações de previsão de futebol pode ser segmentado em duas categorias principais, com base na sua proposta de valor e no seu público-alvo.

3.1.1. Motores de Previsão Dedicados e Impulsionados por IA

Esta categoria engloba aplicações cujo principal objetivo é fornecer probabilidades geradas por algoritmos de IA para os resultados dos jogos. Exemplos notáveis incluem *Football Predictions AI* [28], *Soccer Predictions Football AI* [29], *NerdyTips* [30] e *Sports AI*.[31] O foco destas plataformas é o resultado algorítmico em si, apresentado como a principal ferramenta para o utilizador. Estas aplicações visam fornecer uma vantagem quantitativa, transformando dados complexos em previsões acionáveis.

3.1.2. Plataformas Integradas de Análise de Apostas

Esta segunda categoria é composta por plataformas mais abrangentes que combinam previsões algorítmicas com análises de especialistas, ferramentas de visualização de dados e integrações diretas com casas de apostas. Exemplos incluem *Action Network*, *VegasInsider* [32] e *BettingPros*.[33] Nestas plataformas, a previsão de IA é uma de várias funcionalidades concebidas para apoiar o processo de tomada de decisão do apostador. O valor não reside apenas na previsão, mas no ecossistema de ferramentas que a rodeia, como o acompanhamento de apostas, a análise de tendências de mercado e o acesso a uma comunidade de outros apostadores.

3.2. Funcionalidades Essenciais e Mercados de Previsão

As aplicações comerciais oferecem uma gama consistente de mercados de previsão, refletindo as opções mais populares nas casas de apostas. As previsões mais comuns incluem:

- **1X2:** Probabilidades para Vitória da Equipa da Casa, Empate ou Vitória da Equipa Visitante.
- **Over/Under:** Probabilidades de o número total de golos no jogo ser superior ou inferior a um determinado limiar (e.g., 2.5 golos).
- **BTTS (Both Teams to Score):** Probabilidades de ambas as equipas marcarem pelo menos um golo.
- **Correct Score:** Previsões do resultado exato do jogo, geralmente classificadas por probabilidade.[28]

Para além das previsões, estas aplicações enriquecem a experiência do utilizador com funcionalidades suplementares, como tabelas classificativas interativas, guias de forma das equipas, histórico de confrontos diretos e acompanhamento de resultados em tempo real.[28] Estas ferramentas contextuais ajudam o utilizador a interpretar as previsões algorítmicas e a tomar decisões mais informadas.

3.3. Discernimento da Tecnologia Subjacente e Modelos de Negócio

3.3.1. Tecnologia

Uma análise crítica da comunicação destas aplicações revela uma notável falta de transparência técnica. A linguagem de marketing utiliza frequentemente termos genéricos como "*advanced machine learning*" ou "*proprietary AI*" sem fornecer detalhes sobre os algoritmos, as características ou os métodos de avaliação utilizados.[28] Esta opacidade contrasta fortemente com a abordagem aberta e revista por pares da investigação académica. A aplicação *NerdyTips* [30] é uma rara exceção, fornecendo alguns detalhes sobre o seu motor "NT Apex", descrito como um sistema baseado em Java com um modelo de duas camadas. No entanto, a regra geral é que as alegações de alta precisão não são publicamente verificáveis. Isto sugere que, do ponto de vista do desenvolvimento do projeto, as metodologias validadas academicamente são uma base mais fiável do que as alegações de marketing das aplicações comerciais.

3.3.2. Modelos de Negócio

Os modelos de negócio predominantes neste mercado são:

- **Freemium com Publicidade:** As previsões básicas são oferecidas gratuitamente, sendo a plataforma monetizada através de publicidade na aplicação.[28]
- **Subscrição / Níveis VIP:** Funcionalidades avançadas, como previsões para mais mercados, análises mais aprofundadas ou uma experiência sem anúncios, são disponibilizadas através de subscrições mensais ou anuais.[29]

Este modelo de negócio dual permite que as aplicações atraiam uma base de utilizadores alargada com a oferta gratuita, enquanto monetizam os utilizadores mais empenhados através de subscrições *premium*.

Tabela 3: Análise de Funcionalidades e Tecnologia de Aplicações Comerciais de Previsão

Nome da Aplicação	Tipo Principal	Mercados de Previsão Oferecidos	Tecnologia Declarada	Modelo de Negócio	Cobertura Liga Portuguesa	Diferenciador Principal

Football Predictions AI	Motor de IA	1X2, Over/Under, BTTS, Score Correto	"Advanced machine learning"	Freemium com anúncios	Não especificado	Foco em múltiplos mercados de previsão
Soccer Predictions Football AI	Motor de IA	1X2, outros não especificados	"Machine learning algorithm"	Gratuito com subscrição VIP	Sim (92+ ligas)	Análise de mais de 10,000 pontos de dados por jogo
NerdyTips	Motor de IA	1X2, Over/Under, etc.	Motor "NT Apex" baseado em Java	Planos de subscrição	**Sim** (Primeira e Segunda Liga)	Transparência relativa na tecnologia e cobertura alargada
Sports AI	Motor de IA	1X2, outros desportos	"Advanced AI", "Machine learning algorithms"	Gratuito com Bot de apostas pago	Não especificado	Cobertura multidesportiva e bot de Telegram
BettingPros	Plataforma de Análise	1X2, Props, Parlays	"AI powered predictions"	Gratuito com subscrição *premium*	Sim (NFL, NBA, MLB, etc.)	Sincronização com casas de apostas e análise de *props*
Action Network	Plataforma de Análise	1X2, Props, etc.	Análise de especialistas e previsões baseadas em dados	Gratuito com subscrição	Sim (desportos dos EUA)	Integração direta com casas de apostas para colocação de apostas

---

Parte III: Estudo de Projetos e Implementações de Código Aberto

Esta secção oferece uma visão prática sobre como projetos semelhantes são construídos, fornecendo um modelo tangível para a implementação do projeto do utilizador.

4.1. Padrões Arquiteturais e Stacks Tecnológicos Comuns

A análise de repositórios no GitHub revela padrões consistentes na arquitetura e nas tecnologias utilizadas para projetos de previsão de futebol.[8] A abordagem mais comum consiste num *backend* desenvolvido em Python, que orquestra todo o fluxo de trabalho. As bibliotecas centrais incluem:

- **Pandas:** Para a manipulação e pré-processamento de dados.
- **Scikit-learn:** Para a implementação de modelos de *machine learning* clássicos e para tarefas de avaliação.
- **XGBoost / LightGBM:** Implementações otimizadas de algoritmos de *gradient boosting*.
- **Flask / Django:** *Frameworks* web leves para criar uma API ou uma interface web simples para servir as previsões.

O fluxo de trabalho típico segue uma sequência lógica:

1. **Ingestão de Dados:** *Scripts* para descarregar e processar dados de fontes como football-data.co.uk.
2. **Engenharia de Características:** *Scripts* dedicados à criação das características preditivas.
3. **Treino e Serialização do Modelo:** Treino do modelo com dados históricos e armazenamento do modelo treinado num ficheiro (e.g., usando pickle ou joblib).
4. **Inferência:** Uma API ou interface web que carrega o modelo serializado para fazer previsões sobre novos jogos.

Esta arquitetura modular e baseada em componentes reutilizáveis é uma prática recomendada para a organização do código do projeto.

4.2. Estudo de Caso: dagbolade/all_leagues-_prediction

O repositório dagbolade/all_leagues-_prediction no GitHub serve como um excelente estudo de caso, pois exemplifica a implementação prática dos conceitos discutidos nas secções anteriores.[35]

- **Estrutura:** O projeto adota uma estrutura de diretórios clara e lógica, separando as responsabilidades em pastas distintas como data, models e app. Esta organização facilita a manutenção e o desenvolvimento do código, representando uma boa prática de engenharia de software a ser seguida.
- **Tecnologia:** A presença da biblioteca lightgbm no ficheiro requirements.txt [35] valida diretamente as conclusões da literatura académica sobre a eficácia dos modelos de *gradient boosting*. A utilização desta biblioteca num projeto de código aberto demonstra que esta tecnologia de ponta é acessível e pode ser implementada num projeto académico.
- **Características:** A análise do código do projeto indica a implementação de características como os *ratings* Elo [35], o que demonstra a aplicação prática das técnicas de engenharia de características avançadas discutidas na Parte I.

Este projeto serve como uma ponte entre a teoria académica e a implementação prática,

fornecendo um modelo arquitetural e tecnológico sólido e validado para o desenvolvimento do projeto do utilizador.

4.3. Análise de Outras Implementações Notáveis

O ecossistema de código aberto é diversificado, e a exploração de outros projetos no GitHub oferece uma perspetiva mais ampla sobre as diferentes abordagens possíveis.[8] Existem implementações que utilizam métodos Bayesianos [8], outras que exploram arquiteturas de *deep learning* com TensorFlow [8] e projetos desenvolvidos noutras linguagens de programação, como R, que também possui um ecossistema robusto para análise estatística e *machine learning*.[12] Esta diversidade mostra que, embora exista um consenso em torno dos modelos de *gradient boosting*, a exploração de abordagens alternativas continua a ser uma área ativa de desenvolvimento na comunidade de código aberto.

---

5. Síntese e Recomendações Estratégicas para o Projeto

Esta secção final sintetiza as conclusões das partes anteriores e traduz-as num conjunto de recomendações claras e acionáveis para o desenvolvimento do projeto, abordando diretamente os objetivos definidos no enunciado [3] e na reunião com o orientador.[3]

5.1. Desafios Consolidados e Lacunas de Investigação

A análise do estado da arte revela um conjunto de desafios persistentes e bem documentados:

- **O "Problema do Empate":** A dificuldade em prever corretamente os empates é uma limitação fundamental dos modelos atuais, decorrente tanto do desequilíbrio de classes como da complexidade inerente a este resultado.[6]
- **Limitações dos Dados Públicos:** A ausência de dados granulares (e.g., desempenho individual dos jogadores, lesões) nos conjuntos de dados públicos impõe um limite ao potencial preditivo dos modelos.[1]
- **Teto de Desempenho:** Os modelos de *machine learning*, embora superiores a abordagens ingénuas, raramente atingem níveis de acurácia drasticamente superiores aos que estão implícitos nas *odds* das casas de apostas, que já incorporam uma vasta quantidade de informação.

5.2. Um Roteiro Metodológico Recomendado

Com base na análise efetuada, recomenda-se o seguinte roteiro para a implementação do projeto:

- **Fonte e Preparação de Dados:**
  - Utilizar os ficheiros CSV históricos para a Primeira Liga portuguesa do site football-data.co.uk.[19]
  - Recolher os dados desde a época 2017-2018 até à época mais recente concluída, para garantir a relevância temporal exigida.
- **Estratégia de Engenharia de Características:**
  - Implementar um *pipeline* de engenharia de características robusto.
  - Começar com as estatísticas básicas do jogo.
  - Criar características mais complexas, dando prioridade a:
    1. **Médias móveis** de estatísticas-chave (e.g., golos marcados/sofridos, remates, cantos) calculadas numa janela de 5 a 10 jogos anteriores para capturar a forma da equipa.
    2. Um **sistema de *rating* Elo** dinâmico para cada equipa, atualizado após cada jogo.
- **Seleção de Modelos:**
  - **Modelo Principal:** Implementar um classificador **XGBoost**. Esta escolha é fortemente suportada pela literatura como o estado da arte para esta tarefa.[10]
  - **Modelo de Referência (*Baseline*):** Implementar um modelo mais simples de **Regressão Logística Multinomial**. Este servirá como um ponto de referência crucial para quantificar o ganho de desempenho obtido com o modelo XGBoost.
- **Protocolo de Avaliação:**
  - **Validação:** Utilizar uma divisão temporal estrita. Treinar os modelos em todas as épocas até à penúltima disponível e usar a última época completa como um conjunto de teste (*hold-out set*). Esta abordagem simula um cenário de previsão do mundo real e evita a fuga de dados.
  - **Métricas:** Avaliar os modelos com base num conjunto de métricas: acurácia geral, F1-*scores* por classe (para avaliar especificamente o desempenho na previsão de empates) e *log-loss* (para avaliar a qualidade das previsões probabilísticas).
- **Estrutura do Projeto:**
  - Organizar o código do projeto de forma modular, seguindo o exemplo do repositório dagbolade/all_leagues-_prediction.[35] Criar uma separação clara para o processamento de dados, engenharia de características, treino de modelos e avaliação.

5.3. Observações Finais

O estado da arte na previsão de resultados de futebol demonstra que o sucesso neste domínio não reside tanto na descoberta de um algoritmo revolucionário, mas sim na aplicação

rigorosa e meticulosa das melhores práticas estabelecidas. A atenção ao detalhe na preparação dos dados, a criatividade na engenharia de características e o rigor no protocolo de avaliação são os fatores que, em conjunto, determinam o desempenho e a validade de um modelo preditivo.

6. Bibliografia

Al-Tuhafi, A.Q.M., Al-Shargabi, A.A. and Al-Marridi, A. (2021) 'Predicting Football Outcomes by Using Poisson Model: Applied to Spanish Primera División', *Journal of Applied Science and Technology Trends*, 2(04), pp. 105–112.

Betting Alliance (2025) *Football Predictions AI*. Google Play Store. Disponível em: https://play.google.com/store/apps/details?id=com.siron.footballpredictionsai.

BettingPros (2025) *BettingPros: Sports Betting*. Google Play Store. Disponível em: https://play.google.com/store/apps/details?id=com.bettingpros.app.play.

Bunker, R., Yeung, C. and Fujii, K. (2024) *Machine Learning for Soccer Match Result Prediction*. Disponível em: https://arxiv.org/abs/2403.07669.

Bunker, R.P. and Susnjak, T. (2022) 'The application of machine learning techniques for predicting football match outcomes: a review', *Applied Artificial Intelligence*, 36(1).

Dagbolade, B. (2024) *all_leagues-_prediction*. GitHub. Disponível em: https://github.com/dagbolade/all_leagues-_prediction.

Football-Data.co.uk (2025a) *Notes for Football Data*. [online] Disponível em: https://www.football-data.co.uk/notes.txt.

Football-Data.co.uk (2025b) *Portugal Football Results, Liga I*. [online] Disponível em: https://www.football-data.co.uk/portugalm.php.

football-data.org (2025) *Coverage*. [online] Disponível em: https://www.football-data.org/coverage.

Harish, S. et al. (2023) 'Expected Goals Prediction in Football using XGBoost', *ESP Journal of Engineering & Technology Advancements*, 3(1), pp. 21–26.

Kaggle (2025) *Liga Portugal Prediction Cup*. [online] Disponível em: https://www.kaggle.com/competitions/liga-portugal-prediction-cup.

KUKI APP D.O.O.E.L (2025) *Soccer Predictions Football AI*. Apple App Store. Disponível em: https://apps.apple.com/us/app/soccer-predictions-football-ai/id1089469095.

Lewandowski, M. and Chlebus, M. (2021) 'Predicting football results with machine learning methods', *Faculty of Economic Sciences, University of Warsaw*.

Matos, P. (2024) 'Predicting Portuguese Primeira Liga match outcomes with Machine Learning', *Medium*. [online] Disponível em: https://medium.com/@paulo.matos16/predicting-portuguese-primeira-liga-match-outcomes-with-machine-learning-1af84c06041a.

NerdyTips (2025) *NerdyTips - Football AI Predictions*. [online] Disponível em: https://nerdytips.com/.

Oliveira, D. (2025) *Enunciado do Projeto: Machine learning aplicado à previsão desportiva*.

Castelo Branco: Instituto Politécnico de Castelo Branco.
OpenFootball (2025) *football.db*. [online] Disponível em: https://openfootball.github.io/.
Pechta, P. (2021) *football-prediction-model*. GitHub. Disponível em: https://github.com/pawelp0499/football-prediction-model.
Schulte, J., Kück, J. and Mühlich, B. (2020) 'Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics', *Applied Sciences*, 10(1), p. 46.
Shah, S.H.H. (2017) *Football Match Prediction using Deep Learning*. Tese de Mestrado, Chalmers University of Technology.
Silva, A. (2025) *Reunião 1 - Projeto de Fim de Curso*. Reunião realizada a 17 de Outubro.
Sports AI (2025) *Sports AI - Most accurate sports predictions*. [online] Disponível em: https://www.sports-ai.dev/.
Sportshandle.com (2025) *Best Sports Pick Apps in 2025*. [online] Disponível em: https://sportshandle.com/betting-guides/best-sports-betting-picks-apps/.
Unidade Técnico Científica de Informática (2023) *Regulamento das Unidades Curriculares de Projeto de Fim de Curso*. Castelo Branco: Escola Superior de Tecnologia, IPCB.
Valadés-Cruz, D., Mezquita, Y. and Iglesias, C.A. (2024) 'Machine Learning Applied to Professional Football: Performance Improvement and Results Prediction', *Future Internet*, 16(3), p. 85.

Trabalhos citados

1. Machine Learning Applied to Professional Football: Performance Improvement and Results Prediction - MDPI, acesso a outubro 18, 2025, https://www.mdpi.com/2504-4990/7/3/85
2. Machine learning application in soccer: a systematic review - PMC, acesso a outubro 18, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC9806754/
3. Enunciado do Projeto.pdf
4. Football Match Prediction Using Machine Learning - Doria, acesso a outubro 18, 2025, https://www.doria.fi/bitstream/handle/10024/187628/sjoberg_fredrik.pdf?sequence=2
5. Match predictions in soccer: Machine learning vs. Poisson approaches - arXiv, acesso a outubro 18, 2025, https://arxiv.org/pdf/2408.08331
6. Predicting Football Match Outcomes with eXplainable Machine ..., acesso a outubro 18, 2025, https://arxiv.org/pdf/2211.15734
7. (PDF) Predicting Football Outcomes by Using Poisson Model: Applied to Spanish Primera División - ResearchGate, acesso a outubro 18, 2025, https://www.researchgate.net/publication/357359461_Predicting_Football_Outcomes_by_Using_Poisson_Model_Applied_to_Spanish_Primera_Division
8. football-prediction · GitHub Topics, acesso a outubro 18, 2025, https://github.com/topics/football-prediction?o=desc&s=stars
9. Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics - MDPI, acesso a outubro 18, 2025, https://www.mdpi.com/2076-3417/10/1/46

10. Working Papers - WNE UW, acesso a outubro 18, 2025, https://www.wne.uw.edu.pl/download_file/813/494
11. Predicting Portuguese Primeira Liga match outcomes with Machine Learning - Medium, acesso a outubro 18, 2025, https://medium.com/@paulo.matos16/predicting-portuguese-primeira-liga-match-outcomes-with-machine-learning-1af84c06041a
12. pawelp0499/football-prediction-model: Multinomial logistic regression model for predicting the outcomes of football matches - GitHub, acesso a outubro 18, 2025, https://github.com/pawelp0499/football-prediction-model
13. Machine Learning for Soccer Match Result Prediction - arXiv, acesso a outubro 18, 2025, https://arxiv.org/pdf/2403.07669?
14. Machine Learning for Soccer Match Result Prediction - ResearchGate, acesso a outubro 18, 2025, https://www.researchgate.net/publication/388544286_Machine_Learning_for_Soccer_Match_Result_Prediction
15. Expected Goals Prediction in Football using XGBoost - ESP JETA, acesso a outubro 18, 2025, https://www.espjeta.org/jeta-v3i1p104
16. Football Match Prediction using Deep Learning - Chalmers ODR, acesso a outubro 18, 2025, https://odr.chalmers.se/server/api/core/bitstreams/4ad5c2f4-f2bc-477d-9397-8b179b5a893f/content
17. Football match prediction using deep learning - SciSpace, acesso a outubro 18, 2025, https://scispace.com/pdf/football-match-prediction-using-deep-learning-12hf3ykdpb.pdf
18. Machine Learning for Soccer Match Result Prediction - arXiv, acesso a outubro 18, 2025, https://arxiv.org/pdf/2403.07669
19. Football Results, Statistics & Soccer Betting Odds Data, acesso a outubro 18, 2025, https://www.football-data.co.uk/data.php
20. Thesis Project in Data Science - DiVA portal, acesso a outubro 18, 2025, http://www.diva-portal.org/smash/get/diva2:1965104/FULLTEXT02.pdf
21. Predicting sport event outcomes using deep learning - PMC, acesso a outubro 18, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12453701/
22. Portugal Football Results and Betting Odds - Football-Data.co.uk, acesso a outubro 18, 2025, https://www.football-data.co.uk/portugalm.php
23. Notes - Football-Data.co.uk, acesso a outubro 18, 2025, https://www.football-data.co.uk/notes.txt
24. Football Data European Top 5 Leagues - Kaggle, acesso a outubro 18, 2025, https://www.kaggle.com/datasets/kamrangayibov/football-data-european-top-5-leagues
25. Open Football Data: Welcome - football.db, acesso a outubro 18, 2025, https://openfootball.github.io/
26. football-data.org - ur src for machine readable football data, acesso a outubro 18, 2025, https://www.football-data.org/
27. Football Coverage - Football-Data.org, acesso a outubro 18, 2025,

https://www.football-data.org/coverage
28. Football Predictions AI - Apps on Google Play, acesso a outubro 18, 2025, https://play.google.com/store/apps/details?id=com.siron.footballpredictionsai
29. Soccer Predictions Football AI - App Store, acesso a outubro 18, 2025, https://apps.apple.com/us/app/soccer-predictions-football-ai/id1089469095
30. NerdyTips: AI Football Predictions with over 75% Accuracy, acesso a outubro 18, 2025, https://nerdytips.com/
31. Sports AI - Most accurate sports predictions, acesso a outubro 18, 2025, https://www.sports-ai.dev/
32. Best Sports Picks Apps for October 2025 - SportsHandle, acesso a outubro 18, 2025, https://sportshandle.com/betting-guides/best-sports-betting-picks-apps/
33. BettingPros: Sports Betting on the App Store, acesso a outubro 18, 2025, https://apps.apple.com/us/app/bettingpros-sports-betting/id1468109182
34. BettingPros: Sports Betting - Apps on Google Play, acesso a outubro 18, 2025, https://play.google.com/store/apps/details?id=com.bettingpros.app.play
35. dagbolade/all_leagues-_prediction: ⚽ AI-powered open ... - GitHub, acesso a outubro 18, 2025, https://github.com/dagbolade/all_leagues-_prediction
36. FBref.com: Football Statistics and History, acesso a outubro 18, 2025, https://fbref.com/en/
37. Scraping FBRef for Data-Driven Scouting and Enhanced Player Profiling - Ricardo Heredia, acesso a outubro 18, 2025, https://ricardoheredia94.medium.com/scraping-fbref-for-data-driven-scouting-and-enhanced-player-profiling-464acad83270
38. Extracting data from FBref • worldfootballR - GitHub Pages, acesso a outubro 18, 2025, https://jaseziv.github.io/worldfootballR/articles/extract-fbref-data.html
39. Scraping Fbref —Creating a Pipeline | by Henrik Schjøth - Medium, acesso a outubro 18, 2025, https://medium.com/@henrik.schjoth/scraping-fbref-creating-a-pipeline-f5c9c23ba9da
40. adamcorren/fbref_football_player_data_scraper: Creates csv files containing football data scraped from the website www.fbref.com - GitHub, acesso a outubro 18, 2025, https://github.com/adamcorren/fbref_football_player_data_scraper
41. FBR API, acesso a outubro 18, 2025, https://fbrapi.com/