Trabalho Prático Mineração de Dados

João Ferreira
Mestrado em Engenharia Informática
Universidade do Minho
Braga, Portugal
pg50461@alunos.uminho.pt

João Caldas
Mestrado em Engenharia Informática
Universidade do Minho
Braga, Portugal
pg50494@alunos.uminho.pt

André Martins
Mestrado em Engenharia Informática
Universidade do Minho
Braga, Portugal
pg50224@alunos.uminho.pt

***Abstract*—Este relatório apresenta um estudo detalhado sobre a aplicação de técnicas de mineração de dados para prever a classificação da Liga portuguesa de futebol tendo em conta os dados das últimas cinco temporadas. Com a crescente disponibilidade de dados no mundo do futebol, a mineração de dados tem se mostrado uma ferramenta valiosa para análises e previsões de desempenho das equipas. O modelo preditivo desenvolvido alcançou uma taxa de acerto razoável na previsão da posição final das equipas na tabela classificativa, superando as abordagens tradicionais baseadas em intuição e análise subjetiva.**

***Index Terms*—teste de hipóteses, data mining, futebol, previsão**

I. Contexto

O futebol é um desporto apaixonante que transcende fronteiras e une nações em torno de uma paixão comum. Desde os pequenos jogos da liga dos últimos até aos grandes confrontos nos melhores campeonatos mundiais, cada jogo é um espetáculo único, repleto de rivalidade, magia e estratégias táticas. No entanto, um dos maiores mistérios que envolve este desporto é a imprevisibilidade dos seus resultados. Ao longo de décadas, adeptos, especialistas e até apostadores tentaram desvendar a fórmula mágica para prever quem sairá vencedor de cada confronto. Agora, com o avanço da tecnologia e o poder do Machine Learning, estamos mais perto de desvendar este enigma e trazer mais insights para o mundo do futebol.

II. Motivação

A motivação para a realização deste relatório surge da crescente importância da análise de dados e da mineração de dados no mundo do desporto, em particular no futebol. Com a disponibilidade de dados cada vez mais abundante, surge uma oportunidade única para explorar essas informações valiosas e obter insights que podem impactar significativamente as estratégias das equipas, o desempenho dos jogadores e a previsão dos resultados.

Como três entusiastas do futebol, achamos que este seria um tema ideal para aplicarmos os conhecimentos adquiridos em Mineração de Dados e assim poder juntar duas das nossas paixões, e também poder observar se o nosso clube teria uma prestação diferente da que teve na realidade. Através da aplicação de técnicas de mineração de dados, é possível extrair conhecimento a partir dos dados recolhidos anteriormente, incluindo estatísticas de jogos, informações sobre os jogadores, táticas utilizadas e até mesmo fatores externos como condições climáticas e assistências.

III. Objetivos

Os objetivos que definimos são os seguintes:

1) Simulação da época 2022/23 da Primeira Liga Portuguesa através de uma simulação de Monte Carlo usando as probabilidades dadas pelo modelo treinado.
2) Estudo da influência do árbitro na percentagem de vitórias da equipa da casa.
3) Estudo da influência do número de espectadores na percentagem de vitórias da equipa da casa.
4) Estudo da influência do dia da semana na percentagem de vitórias da equipa favorita.
5) Estudo da influência do factor dia/noite na percentagem de vitórias da equipa favorita.

IV. Fontes de dados e métodos de acesso

A primeira fonte de dados foi o FBRef. Desenvolvemos um scraper para isto, que funciona em 2 etapas:

1) 1ª Etapa: Informação geral sobre os jogos, bem como link para o relatório pormenorizado de cada jogo.
2) 2ª Etapa: Consulta página a página do relatório de cada jogo e retirar as informações mais específicas (jogadores presentes, formação e estatísticas do jogo)

Ao longo da 2ª etapa encontraram-se problemas como a limitação de acesso por parte do website, que foram resolvidos através da introdução de um *sleep* no código desenvolvido ou como o erro em jogos atípicos (caso do Belenenses - Benfica da época 21/22, em que o Belenenses entrou em campo com 8 jogadores) resolvido através da introdução dos dados de forma manual.

Este *scraper* da 2ª etapa do *FBRef* foi feito de modo a que possa apenas atualizar os datasets já existentes, o que permite agilizar os processos.

A segunda fonte de dados foi a Football-data, de onde se descarregou os datasets e se desenvolveu um *script* para juntar algumas informações destes datasets ao dataset previamente obtido do *FBRef*.

Para a junção foi desenvolvido 1 *script*, que primeiramente junta os datasets ano a ano e após isso concatena todos num só.

V. Análise dos dados e Testes de Hipóteses

A análise inicial dos dados passou pela observação da relação entre o resultado do jogo e o favorito antes desse mesmo jogo, mas o gráfico gerado não permitiu a deteção de padrões anormais que merecessem atenção. A definição do favorito fez-se com as *odds* obtidas do website BET365, onde se considerou:

1) *Odd* da equipa visitada inferior a 1.40 como **HH** (Equipa visitada muito favorita) - 174 jogos
2) *Odd* da equipa visitada inferior a 2.26 como **H** (Equipa visitada favorita) - 471 jogos
3) *Odd* da equipa visitante inferior a 2.26 como **A** (Equipa visitante favorita) - 226 jogos
4) *Odd* da equipa visitante inferior a 1.40 como **AA** (Equipa visitante muito favorita) - 59 jogos
5) Restantes jogos como sendo jogos de *empate* e caraterizados como **D** - 294 jogos

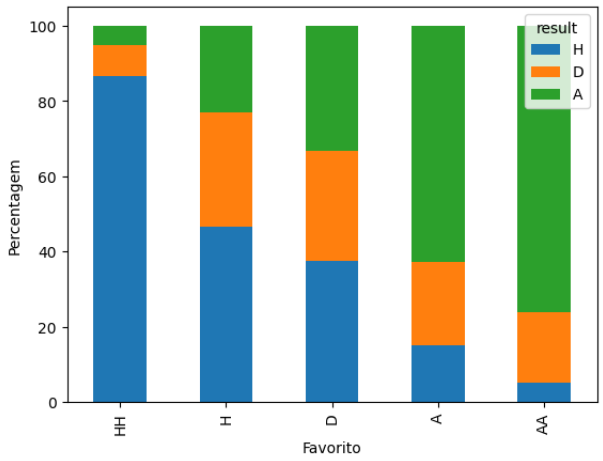

Fig. 1. Percentagem de acontecimento de cada resultado tendo em conta o favorito do jogo

Após isso, decidiu-se observar a influência dos espectadores nos resultados do jogos, ou seja, análise dos resultados dos jogos nas épocas em que os adeptos estavam interditos de aceder aos estádios derivado da pandemia. Pode-se observar que com a presença de público no estádio, o resultado das equipas que jogavam em casa era ligeiramente melhor, e fez-se um teste de hipóteses para verificar se essa possível influência poderia ser devido ao acaso.

- H0: média da diferença de golos com espectadores - média da dif de golos sem espectadores = 0
- H1: média da diferença de golos com espectadores - média da dif de golos sem espectadores > 0

p-value: 0.7645

Tendo em conta o resultado obtido, não se rejeita a hipótese nula, pelo que não é certo afirmar que a ligeira vantagem não se possa dever ao acaso.

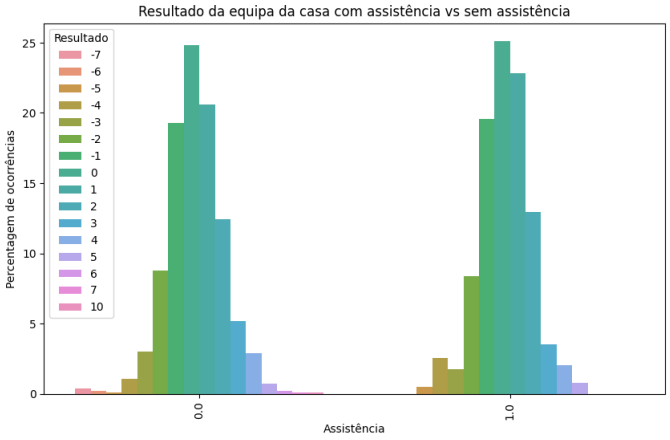

Fig. 2. Comparação dos resultados com e sem assistência

Depois, verificou-se se jogar à noite afetaria o resultado de um jogo a favor da equipa da casa, e realizou-se um teste de hipóteses, tendo-se obtido o seguinte resultado:

- H0: média da diferença de golos com jogo à noite
- H1: média da diferença de golos com jogo durante o dia

p-value: 0.5101

Mais uma vez, tendo em conta o resultado, não se rejeita a hipótese nula.

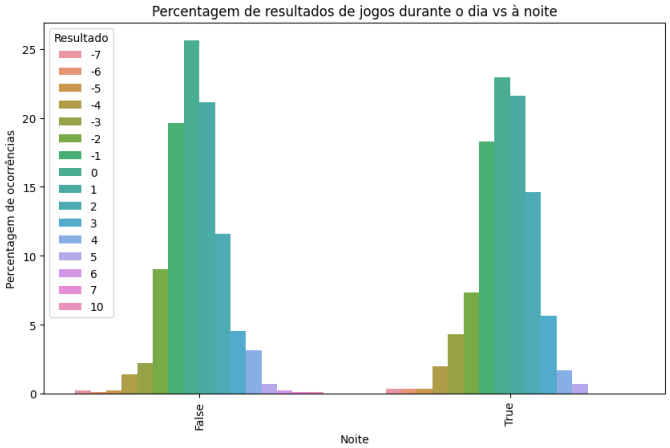

Fig. 3. Comparação dos resultados a jogar de dia ou de noite

À semelhança do exemplo anterior, verificou-se se jogar ao fim de semana ou durante os dias laborais teria influêmcia no resultado do jogo, e percebeu-se que a equipa visitada tem uma ligeira vantagem, mas que esta não é significativa.

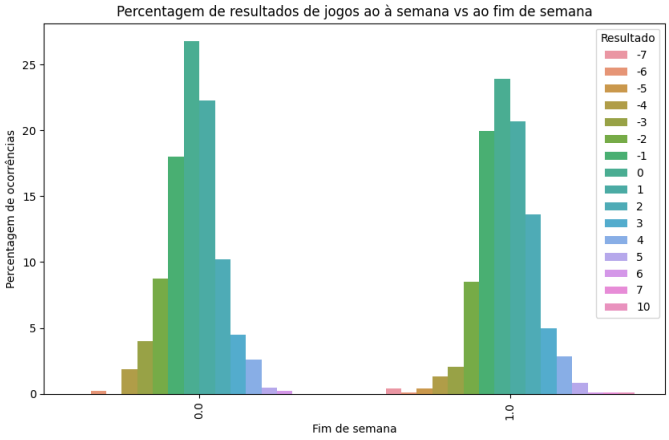

Fig. 4. Comparação dos resultados a jogar à semana ou ao fim de semana

Durante o trabalho, foram criadas colunas que registavam a forma das equipas nos últimos 5, 17 e 34 jogos (média de pontos) e após análise à forma das equipas face ao resultado do jogo seguinte, reparou-se que quando a equipa tinha apenas 0.2 pontos por jogo nos últimos 5 jogos, em mais de 50 por cento das ocasiões vencia o jogo seguinte, e decidiu-se testar com um Teste-U de Mann-Whitney.

- H0: A equipa da casa ganha após má forma
- H1: A equipa da casa não ganha

p-value: 0.1873

Tendo em conta o resultado, não se rejeita a hipótese nula.

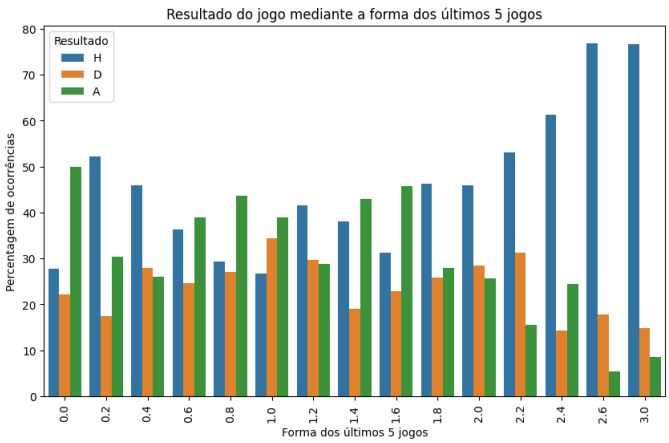

Fig. 5. Análise do resultado tendo em conta a forma dos últimos 5 jogos

Por fim, verificou-se a influência dos árbitros a favor das equipas visitadas:

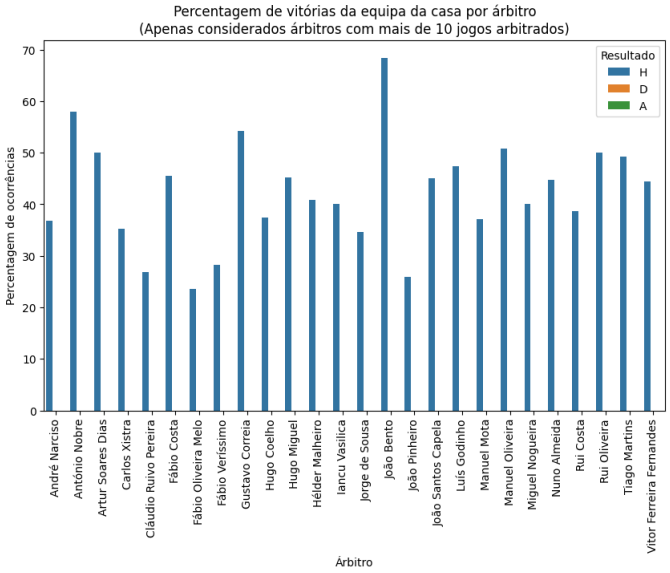

Fig. 6. Percentagem de vitórias da equipa da casa por árbitro

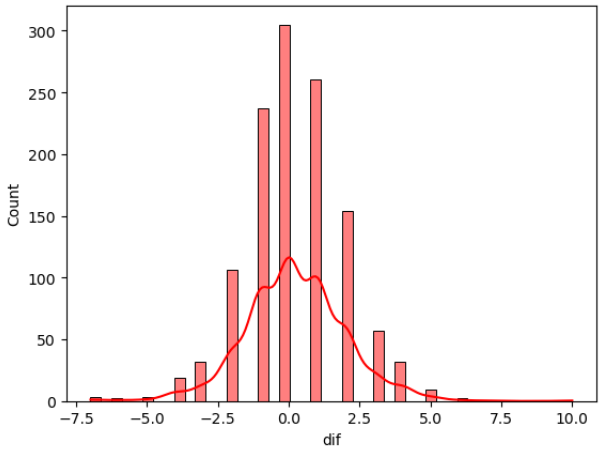

Fig. 7. Diferença de golos

Começou-se pela realização de um teste de Shapiro-Wilk à normalidade da diferença de golos.

Shapiro-Wilk p-value: 6.63e-18

Com o p-value obtido, conclui-se a não normalidade da diferença de golos, pelo que serão utilizados testes não paramétricos.

Analisando o gráfico da % de vitórias da equipa da casa por árbitro, verifica-se um valor anormalmente elevado para o árbitro João Bento, pelo que se procedeu a realização de um teste de hipóteses:

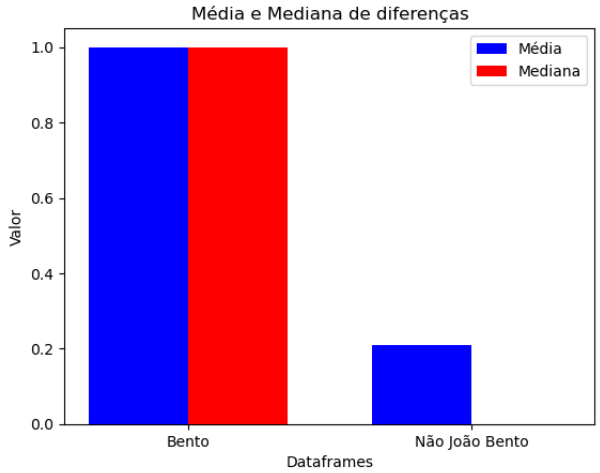

Fig. 8. Média e Mediana de diferença de golos com e sem o árbitro João Bento

- H0: média da diferença de golos com João Bento - média da diferença de golos sem João Bento = 0 (A equipa da casa não tem vantagem)
- H1: média da diferença de golos com João Bento - média da diferença de golos sem João Bento > 0 (A equipa da casa tem vantagem)

p-value: 0.0269

Depois da observação destes valores, e de termos concluído que a 95% de confiança, a equipa da casa era efetivamente beneficiada pelo árbitro João Bento, fez-se uma pesquisa e conclui-se que o árbitro João Bento tinha sido despromovido na época 2020/21!

Testou-se também para o árbitro António Nobre da seguinte forma:

- H0: média da diferença de golos com António Nobre - média da diferença de golos sem António Nobre = 0 (A equipa da casa não tem vantagem)
- H1: média da diferença de golos com António Nobre - média da diferença de golos sem António Nobre > 0 (A equipa da casa tem vantagem)

p-value: 0.0034

Também após a observação do valor obtido, é seguro dizer que com 99% de confiança, a equipa visitada sai beneficiada da presença do árbitro António Nobre.

VI. Desenvolvimento

Começou-se por estabelecer uma baseline mínima. Isto é, alguém que escolha arbitrariamente o vencedor de um jogo de futebol teria uma precisão de 33%. Após alguma análise, estabeleceu-se uma baseline mais racional, usando-se a percentagem de vitórias do favorito. Após análise, o valor fixou-se nos 54.2%.

Procedeu-se então à definição de médias a 5, 17 e 34 jogos de todas as variáveis, com o intuito de, mais à frente, usar um algoritmo de seleção de *features*, mas logo se percebeu que o estabelecimento de médias a 17 e a 34 jogos provocava uma perda grande de informação devido a *missing values*, pelo que se procedeu ao uso de médias apenas a 5 jogos.

Às variáveis dos jogadores, o tratamento dado passou pelo cálculo do número de titulares semelhantes ao jogo anterior.

Aplicou-se um *One-hot encoding* às formações, aos árbitros e ao dia da semana e procedeu-se ao estabelecimento dum algoritmo que calculava a posição de ambos os clubes no momento do jogo.

Foram excluídas as 5 primeiras jornadas de cada época, por serem momentos onde as classificações não estavam bem definidas e havia muita variação nos 11 iniciais.

Como citado em cima, o plano passaria pelo uso de uma rede neuronal, mas esta deu alguns problemas (ou dava overfit ou considerava as mesmas probabilidades para todos os eventos) pelo que se optou pelo uso do *XGBoost*.

Após isto, usou-se um Grid Search para procurar os parâmetros que beneficiavam mais o nosso modelo.

VII. Resultados e Discussão

Mesmo depois do fine tune do modelo, a *accuracy* ficou aquém do que esperávamos, obtendo um valor de 53.1%. Analisou-se ainda em termos de apostas desportivas e observou-se que o modelo daria uma perda de 4.6 unidades (assumindo a aposta de 1 unidade por jogo). Para efeitos de comparação, calculou-se o retorno da aposta no favorito em todos os jogos e observou-se um lucro de 35.45 unidades, que acabo por ser estranhamente elevado. À procura de explicações, percebeu-se que a percentagem de vitórias da equipa favorita foi anormalmente elevada nos dados de teste: 63.1% vs 54.2% nos dados de treino. Acredita-se que esta será uma das razões que levou ao resultado abaixo do baseline estimado.

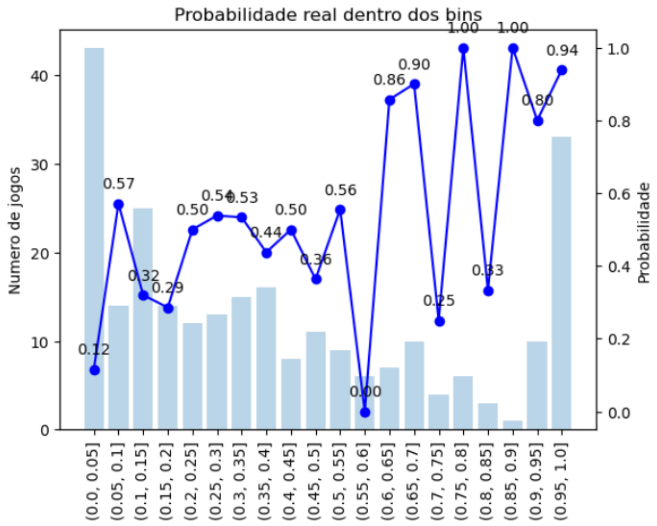

Fig. 9. Probabilidade real vs prevista relacionada ao número de jogos

Nesta imagem, onde relacionamos a probabilidade que o modelo estima para a vitória face à frequência com que

realmente acontece consegue-se perceber que cerca de metade das estimativas de probabilidades se inserem ou no intervalo [0, 0.05] ou no intervalo [0.95, 1.00]. No intervalo [0.95, 1.00], a probabilidade real parece ajustada à prevista pelo modelo (94%), já no intervalo [0, 0.05], a probabilidade está desajustada, já que o erro médio para esse intervalo será de no mínimo 7% . Nos restantes intervalos existe um número reduzido de números, pelo que não se descarta que a diferença nos valores não seja apenas variância.

Equipa	Pnts Previstos	Pnts Reais	Dif	Pos Pred	Pos Real
Porto	71.2	73.0	1.8	4	1
Benfica	78.2	72.0	-6.2	1	2
Sp Lisbon	76.9	67.0	-9.9	2	3
Sp Braga	71.6	65.0	-6.6	3	4
Guimaraes	39.1	47.0	7.9	5	5
Arouca	26.0	47.0	21.0	16	6
Famalicao	27.6	40.0	12.4	13	7
Chaves	26.7	38.0	11.3	15	8
Rio Ave	28.7	35.0	6.3	11	9
Vizela	35.1	35.0	-0.1	6	10
Boavista	32.6	35.0	2.4	8	11
Casa Pia	28.6	33.0	4.4	12	12
Gil Vicente	30.4	32.0	1.6	9	13
Estoril	30.3	28.0	-2.3	10	14
Maritimo	33.9	26.0	-7.9	7	15
Pacos Ferreira	25.0	23.0	-2.0	18	16
Portimonense	27.5	22.0	-5.5	14	17
Santa Clara	25.5	18.0	-7.5	17	18

Fig. 10. Tabela classificativa real vs prevista

Após efetuada a simulação de Monte Carlo (1000 iterações), gerou-se a seguinte tabela, para efeitos de comparação. Verificou-se que as 3 maiores diferenças entre o real e o previsto são de equipas que oscilaram muito durante a época (Arouca, Famalicão e Chaves), algo que já se esperava ao não haver a possibilidade de usar médias a 17 e a 34 jogos. Pensa-se que com mais dados e o uso de variáveis temporais mais longas se poderia obter um melhor desempenho do modelo. Algo mais que reforça este ponto é a valorização que o modelo faz da consistência do Benfica.

VIII. Conclusão

Neste trabalho, explorámos a aplicação da mineração de dados na previsão da classificação da Liga Portuguesa de futebol. Os resultados obtidos demonstraram que a abordagem que tivemos neste trabalho foi de encontro ao que foi lecionado e desenvolvido na UC de mineração de dados. O modelo preditivo alcançou uma taxa de acerto razoável, conseguindo prever uma classificação final próxima da que foi real.

Além disso, o trabalho permitiu identificar variáveis-chave que influenciam o desempenho das equipas, destacando a importância de fatores como a forma dos jogos anteriores, informações sobre as equipas e jogadores e árbitros, e outros elementos externos. Estas descobertas têm o potencial de auxiliar na tomada de decisões estratégicas, contribuindo para um planeamento mais eficiente e fundamentado.

Concluindo, consideramos este trabalho como algo positivo, tendo atingido os objetivos propostos, tanto por nós como pelo corpo docente, podendo assim pôr em prática a matéria lecionada na Unidade Curricular de Mineração de Dados.

IX. Trabalho Futuro

O trabalho futuro passará pela incorporação de mais dados estatísticos ao dataset, bem como da expansão temporal do dataset para permitir a existência de mais entradas nos dados de treino. Terá de se levar em conta também o uso de outros modelos, como redes neuronais, que pensa-se que serão treinadas de forma adequada com a existência de mais entradas nos dados.