Data Mining sobre Mundiais de Futebol

Classificação, Clustering e Regras de Associação

Docente: Rui Fernandes

Grupo Nº II

Trabalho Elaborado por:

ANTÓNIO FERREIRA – 9657

MAFALDA BARÃO – 20446 23033

RÚBEN DIAS –

GONÇALO GOMES – 23039

JOÃO MOARAIS - 23041

INSTITUTO POLITÉCNICO DO CÁVADO E DO AVE
ESCOLA SUPERIOR DE TECNOLOGIA

Conteúdo

Índice Tabelas

Lista de Abreviaturas e Siglas

- **AFC** — Asian Football Confederation (Confederação Asiática de Futebol)
- **CAF** — Confédération Africaine de Football (Confederação Africana de Futebol)
- **CONCACAF** — Confederation of North, Central America and Caribbean Association Football
- **CONMEBOL** — Confederación Sudamericana de Fútbol (Confederação Sul-Americana de Futebol)
- **OFC** — Oceania Football Confederation (Confederação de Futebol da Oceânia)
- **UEFA** — Union of European Football Associations (União das Associações Europeias de Futebol)
- **CSV** — Comma-Separated Values (valores separados por vírgulas)
- **DM** — Data Mining (Mineração de Dados)
- **ETL** — Extract, Transform, Load (Extrair, Transformar, Carregar)
- **ML** — Machine Learning (Aprendizagem Automática)
- **K-Means** — K-Means Clustering (agrupamento por centróides)
- **PCA** — Principal Component Analysis (Análise de Componentes Principais)
- **SSE** — Sum of Squared Errors (Soma dos Erros Quadráticos; usada como "inércia")
- **CV** — Cross-Validation (Validação Cruzada)
- **GridSearchCV** — Pesquisa exaustiva de hiperparâmetros com validação cruzada
- **RF** — Random Forest (Floresta Aleatória)
- **LR** — Logistic Regression (Regressão Logística)
- **GB** — Gradient Boosting (Boosting por Gradiente)
- **F1-macro** — Média do F1-score calculado por classe (peso igual por classe)
- **Accuracy** — Proporção de previsões corretas
- **LogLoss** — Logarithmic Loss (Perda Logarítmica)
- **TP/FP/FN/TN** — True Positive / False Positive / False Negative / True Negative
- **Apriori** — Algoritmo para descoberta de itemsets frequentes e regras de associação
- **Support** — Suporte (frequência de ocorrência de um itemset/regra)
- **Confidence** — Confiança (probabilidade do consequente dado o antecedente)

- **Lift** — Lift (força da associação; >1 indica associação positiva)
- **NaT** — Not a Time (valor nulo em colunas de data/hora no pandas)
- **Top10 / 26–50** — Tiers de ranking FIFA (intervalos de rank)
- **WinsTier / PointsTier / RankTier** — Discretização ("tiers") de vitórias, pontos e ranking
- **GoalDiff** — Goal Difference (diferença de golos marcados – sofridos)
- **Monte Carlo** — Simulação por amostragem repetida para estimar probabilidades
- **n_sims** — Número de simulações (na simulação do torneio)

1. Enquadramento e objetivos

Este trabalho aplica três abordagens clássicas de **Data Mining / Machine Learning** a dados históricos dos Campeonatos do Mundo de Futebol e ao **ranking FIFA**: (i) **classificação** do resultado de jogos, (ii) **segmentação** de seleções por desempenho e (iii) descoberta de **padrões** através de regras de associação.

Objetivos principais:

- Construir um conjunto de dados coerente a partir de múltiplas fontes (jogos, ranking FIFA e vencedores).
- Treinar e comparar modelos de classificação para prever o resultado (**vitória fora, empate, vitória casa**).
- Agrupar seleções por perfis de performance (por **equipa e ano**) com **K-Means** e interpretar os clusters.
- Extrair regras de associação (**Apriori**) para identificar relações frequentes entre características de performance e resultados.

2. Conjuntos de dados

Foram utilizados três datasets em formato CSV (na pasta data/), combinando informação dos jogos do Mundial, edições e ranking FIFA.

Dataset	Descrição	Dimensão (linhas × colunas)
WorldCupMatches.csv	Jogos dos Mundiais (equipas, golos, fase, ano, etc.).	4572 × 20
WorldCups.csv	Edições do Mundial (vencedor, finalista, etc.).	22 × 10
fifa_ranking-2024-06-20.csv	Ranking FIFA por data (rank, pontos, confederação).	67472 × 8

Tabela 1- Conjuntos de dados utilizados (descrição e dimensão).

3. Preparação e limpeza de dados

As etapas de preparação foram comuns aos três notebooks, assegurando consistência nos identificadores e qualidade dos dados:

- **Normalização de nomes de equipas**, removendo espaços e harmonizando strings para permitir junções fiáveis entre datasets.
- **Conversão de colunas de datas** (rank_date e, quando aplicável, datas de jogo), garantindo tipos corretos para operações temporais.
- **Criação de variáveis derivadas**, como diferenças de ranking/pontos e métricas agregadas por equipa/ano.
- **Remoção de observações sem ranking FIFA válido** (quando necessário), para evitar ruído na modelação e inconsistências na criação de features.

4. Notebook 1 — Classificação do resultado de jogos

Problema: prever o desfecho de um jogo do Mundial a partir de informação de ranking FIFA e contexto.

O target foi codificado como **y ∈ {0,1,2}**, onde:

0 = vitória da equipa visitante

1 = empate

2 = vitória da equipa da casa

Construção do dataset

Foram considerados jogos a partir de **1994**, dado que o ranking FIFA é disponibilizado com maior consistência a partir dessa época.

Foi efetuado um **merge as-of por equipa**, associando a cada jogo o ranking mais recente anterior à data do jogo.

Na fase de merge, não houve datas inválidas (**NaT=0**) e obteve-se um subconjunto final de **355 jogos** com ranking (removidos **497** sem ranking).

Foram criadas features numéricas: *home_rank*, *away_rank*, *rank_diff*, *points_diff*, e flags de histórico como *champion_before*.
A fase do torneio (stage) foi tratada como variável categórica (one-hot).

Classe	Significado	Proporção
0	Away win	0.2042
1	Draw	0.2230
2	Home win	0.5728

Tabela 2 - Distribuição do target (resultado do jogo).

A distribuição mostra um desequilíbrio moderado a favor de vitórias da equipa da casa, o que tende a penalizar a previsão da classe "empate" e justifica o uso de métricas como **F1-macro**.

Divisão treino/teste

A separação foi feita por grupos (**GroupShuffleSplit**) usando o **ano** como grupo, de forma a reduzir fuga temporal e avaliar a capacidade de generalização para edições diferentes.

Conjunto	N amostras	N features
Treino	246	7
Teste	109	7

Tabela 3 - Dimensão dos conjuntos de treino e teste.

4.1 Modelos avaliados e resultados

Foram testados vários modelos supervisionados e comparados usando **Accuracy**, **F1-macro** e **LogLoss**.

Modelo	Accuracy	F1-macro	LogLoss
Baseline (classe mais frequente)	0.4587	0.2096	19.5099
Logistic Regression	0.5046	0.4083	1.0846
Random Forest	0.4495	0.3560	1.0973
Gradient Boosting	0.3945	0.3290	1.3083
Random Forest (tuned)	0.4954	0.4225	1.0580

Tabela 4 - Comparação de modelos (resultados no conjunto de teste)

O melhor desempenho em **F1-macro** foi obtido pelo **Random Forest afinado** (GridSearchCV), com **F1-macro = 0.4225** no teste e score médio de validação cruzada (3-fold) de **0.4478**.

Parâmetros escolhidos: {'clf__max_depth': None, 'clf__min_samples_leaf': 2, 'clf__min_samples_split': 2, 'clf__n_estimators': 300}.

4.2 Interpretação rápida

- rank_diff e points_diff são sinais fortes para o desfecho, pois representam diretamente a diferença de "força" entre seleções.
- A classe **empate (1)** tende a ser a mais difícil de prever (normalmente com recall inferior), sugerindo que seriam úteis features adicionais (ex.: rating ELO, forma recente, neutralidade do campo).

- O tuning melhora a estabilidade do modelo e reduz ligeiramente o **LogLoss**, indicando probabilidades mais consistentes/calibradas.

4.3 Simulação de probabilidades de campeão

A partir do modelo final, foi implementada uma simulação do torneio (32 seleções) em formato **Monte Carlo**. Em cada jogo simulado:

1. o modelo estima probabilidades para {away, draw, home};
2. o resultado é amostrado de acordo com essas probabilidades;
3. repete-se o torneio para **N simulações**, estimando-se a frequência com que cada seleção se torna campeã.

Seleção	Vitórias (sim.)	P(campeão)
Portugal	65	0.130
Belgium	57	0.114
Netherlands	49	0.098
Argentina	35	0.070
Spain	34	0.068
Germany	29	0.058
Chile	24	0.048
Uruguay	19	0.038
Colombia	19	0.038
England	19	0.038

Tabela 5 - Top 10 probabilidades estimadas de campeão (N=500 simulações)

Nota: estas probabilidades são indicativas e dependem das features disponíveis, do modelo e do número de simulações.

5. Notebook 2 — Clustering (K-Means)

Objetivo: identificar perfis de seleções a partir de métricas agregadas por (**equipa, ano**), agrupando campanhas semelhantes nos Mundiais (1994–2022).

Construção do dataset team_year

- Transformação do dataset de jogos para formato longo (uma linha por equipa e jogo).
- Cálculo de métricas por equipa e ano: jogos, vitórias, empates, derrotas, golos marcados/sofridos, diferença de golos e pontos.
- Atribuição da melhor fase atingida (stage_reached) e enriquecimento com ranking FIFA (rank, total_points) e confederação.
- Resultado final: **174 observações** (team_year) com **21 colunas**.

5.1 Seleção do número de clusters (K)

Foram testados vários valores de K e comparadas duas métricas: **inércia (SSE)** e **silhouette score**.

k	inertia	silhouette
2	1306.908624	0.365402
3	1131.250753	0.186847
4	1023.186502	0.200218
5	926.959202	0.205342
6	843.748956	0.203451
7	775.701845	0.167126
8	702.205109	0.183804
9	656.733425	0.189699
10	632.441369	0.182152

Tabela 6 - Avaliação de K no K-Means (inércia e silhouette)

Justificação do K=4: optou-se por **K=4** por permitir uma segmentação suficientemente granular para distinguir padrões estáveis de desempenho, mantendo o modelo simples de interpretar e evitando uma fragmentação excessiva em grupos muito pequenos.

5.2 Perfis dos clusters

Cluster	N observações
0	58
1	21
2	25
3	70

Tabela 7 - Tamanho de cada cluster (K=4)

Resumo interpretativo:

- **Cluster 1:** rank médio mais baixo (melhor) e total_points mais alto; inclui seleções mais fortes e consistentes.
- **Cluster 2:** maior número médio de jogos e pontos conquistados, refletindo campanhas prolongadas (fases muito avançadas).
- **Cluster 0:** desempenho intermédio, com tendência para atingir fases a meio do torneio (ex.: oitavos/quartos).
- **Cluster 3:** rank médio mais alto (pior) e pontos conquistados mais baixos; associado a eliminações precoces e menor produção ofensiva.

5.3 Interpretação rápida

- Os clusters resumem padrões históricos: melhores rankings/pontos tendem a associar-se a fases mais avançadas.
- Embora a confederação não seja uma feature numérica direta, a composição por cluster sugere maior presença UEFA/CONMEBOL nos grupos mais fortes.

- Esta segmentação pode ser reutilizada como variável derivada (cluster_id) em análises futuras.

6. Notebook 3 — Regras de Associação (Apriori)

Objetivo: descobrir associações frequentes entre características de desempenho das seleções por ano, usando Apriori e métricas de **support**, **confidence** e **lift**.

Transações

- Cada transação representa uma campanha (**equipa, ano**).
- Variáveis numéricas foram discretizadas em tiers (RankTier, PointsTier, WinsTier).
- Incluíram-se itens como confederação, tier de ranking, tier de pontos, fase atingida, histórico de campeão e sinal da diferença de golos.
- Número de transações: **174**.

6.1 Ajuste de parâmetros

min_support	min_confidence	n_itemsets	n_rules
0.05	0.65	376	286
0.08	0.65	216	187
0.10	0.65	157	139
0.12	0.65	117	100

Tabela 8 -— Impacto de min_support no nº de itemsets e regras (min_confidence=0.65)

Valores de min_support entre **0.08 e 0.12** reduzem a quantidade de regras mantendo padrões suficientemente frequentes para interpretação.

6.2 Exemplos de regras relevantes

Antecedente	Consequente	Support	Confidence	Lift
Stage=Quarter-finals	WinsTier=Mid	0.103	0.75	2.9
WinsTier=High	GoalDiff=Pos	0.132	0.958	2.647
Stage=Quarter-finals	GoalDiff=Pos	0.126	0.917	2.532

ChampionBefore=Yes	RankTier=Top10	0.132	0.742	2.436
Confed=UEFA, WinsTier=Mid	GoalDiff=Pos	0.103	0.783	2.161
WinsTier=Mid	GoalDiff=Pos	0.190	0.733	2.025
Confed=CAF, WinsTier=Low	Stage=Group/1st round	0.109	0.905	1.968
Confed=AFC, WinsTier=Low	GoalDiff=Neg	0.103	1.0	1.955
ChampionBefore=No, WinsTier=Mid	GoalDiff=Pos	0.138	0.686	1.894
RankTier=26-50, WinsTier=Low	Stage=Group/1st round	0.195	0.85	1.849

Tabela 9 - — Exemplos de regras de associação (Apriori)

6.3 Discussão

- As regras confirmam padrões plausíveis: **mais vitórias** e **diferença de golos positiva** tendem a associar-se a **fases mais avançadas**.
- **Lift > 1** indica associação positiva (o consequente torna-se mais provável quando o antecedente ocorre). Valores elevados sugerem relações mais informativas.
- Para reduzir regras evidentes e focar padrões mais específicos, pode-se aumentar min_support e/ou exigir um lift mínimo (ex.: ≥ 1.5).

7. Conclusões e trabalho futuro

Neste trabalho foram aplicadas três abordagens complementares de Data Mining a dados do Mundial e do ranking FIFA. No **Notebook 1**, o **Random Forest afinado** apresentou o melhor compromisso entre desempenho e robustez na previsão do resultado dos jogos, destacando-se quando avaliado com **F1-macro**, métrica mais adequada ao desequilíbrio entre classes e à maior dificuldade em prever **empates**. No **Notebook 2**, o **K-Means** permitiu obter uma visão compacta e interpretável das campanhas históricas por (equipa, ano), facilitando a comparação entre perfis de desempenho e a análise exploratória de padrões. No **Notebook 3**, o **Apriori** revelou associações frequentes coerentes com o domínio do futebol, formalizando relações entre variáveis de performance (ex.: vitórias, diferença de golos, fase atingida) e reforçando conclusões interpretativas.

Apesar dos resultados, existem limitações relevantes. O conjunto efetivo para classificação ficou relativamente pequeno após o *merge* temporal com o ranking FIFA, o que pode limitar a capacidade de generalização do modelo. Além disso, os **empates** são intrinsecamente difíceis de prever apenas com ranking e fase, sugerindo que a inclusão de features adicionais — como **ELO**, forma recente, histórico de confrontos diretos, neutralidade do campo ou contexto geográfico — poderá melhorar o desempenho. Por fim, a simulação do torneio assume **independência entre jogos** e não modela fatores externos importantes (lesões, rotação, dinâmica de grupo, vantagem de jogar em casa/continente), pelo que as probabilidades estimadas devem ser interpretadas como indicativas e condicionadas ao modelo e às variáveis disponíveis.

Como trabalho futuro, recomenda-se (i) enriquecer o dataset com variáveis contextuais e temporais, (ii) testar modelos com melhor calibração probabilística e/ou técnicas de balanceamento, e (iii) tornar a simulação mais realista, incorporando fatores externos e dependências entre jogos.

8. Checklist de conformidade com o enunciado

A entrega contempla:

- Estrutura de pastas conforme solicitado: data/ e notebooks separados por tarefa.
- Notebook 1: classificação com preparação, treino, avaliação e comparação de modelos + tuning.
- Notebook 2: clustering com K-Means, justificação do K e interpretação dos clusters.
- Notebook 3: regras de associação (Apriori) com discretização, métricas e discussão.
- README.txt com descrição do projeto e instruções para execução.

9. Como executar

Pré-requisitos sugeridos:

- Python 3.10+
- pandas, numpy, scikit-learn, matplotlib
- mlxtend (para Apriori)

Executar os notebooks pela ordem 1→2→3, garantindo que a pasta data/ está no mesmo diretório dos notebooks.