APRENDIZAGEM COMPUTACIONAL

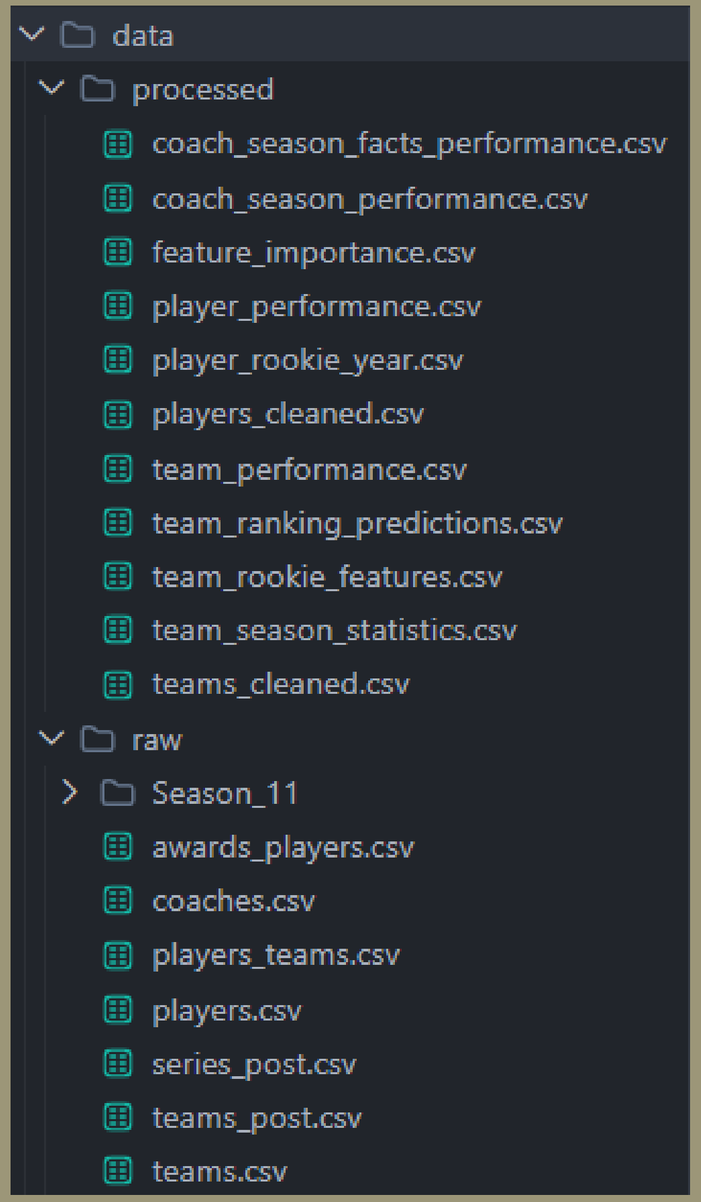
up202208296 - Lucas Greco do Espírito Santo Jorge

Grupo 71

DESCRIÇÃO DE DOMÍNIO

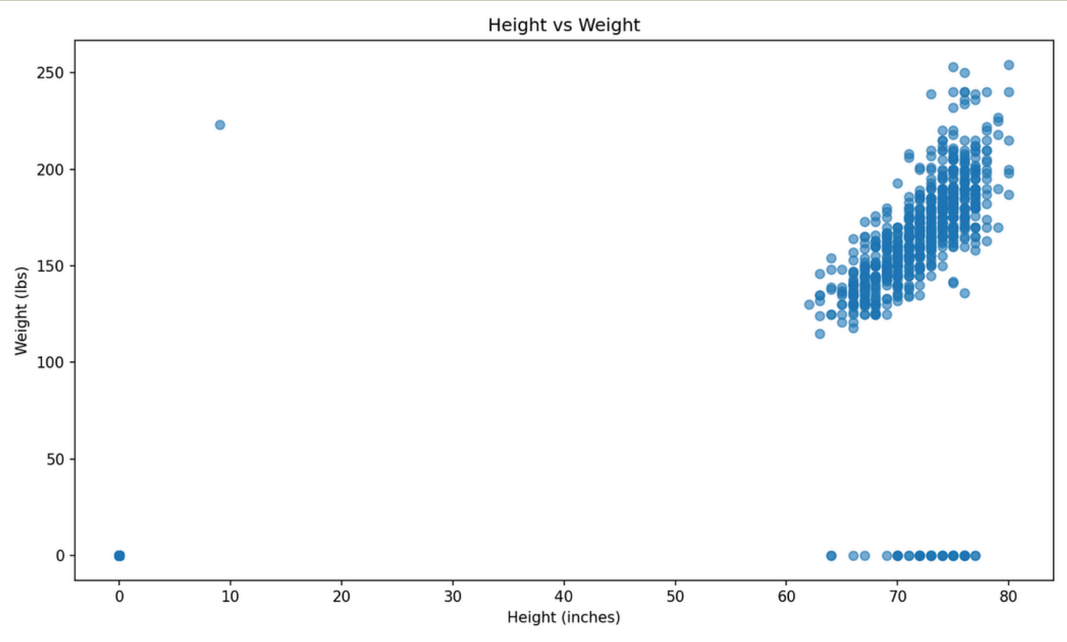
PREVER O RANKEAMENTO FINAL DAS EQUIPES DE CADA CONFERÊNCIA DE UM ANO FUTURO.
VARIAVÉL ALVO: “RANK”

- **DATASET:**
  - **RAW: CSVS QUE RECEBEMOS**
    - **INCLUSIVE A PASTA SEASON_11, QUE CONTÉM OS DADOS DE TESTE, COMO O TIME, OS JOGADORES E OS TÉCNICOS.**
  - **PROCESSED: CSVS LIMPOS E PROCESSADOS**



EXPLORAÇÃO DE DADOS

- VÁRIOS JOGADORES APRESENTAM POSIÇÕES INDEFINIDAS.
- EXISTEM REGISTROS COM ALTURAS E PESOS INCOMPATÍVEIS.
- ALGUMAS COLUNAS DE TEAMS, COMO “DIV ID”, ESTÃO VAZIAS.

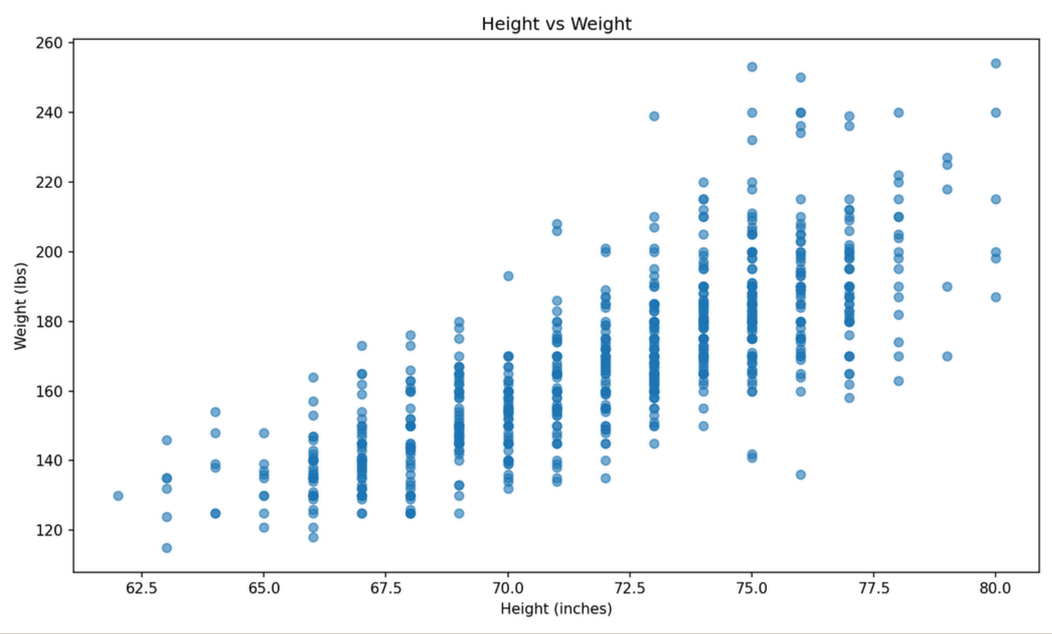


- HÁ COLUNAS COM VALORES CONSTANTES ENTRE TODOS OS TIMES.
- FORAM IDENTIFICADAS DUAS CONFERÊNCIAS DISTINTAS NO MESMO ANO.
- AS POSIÇÕES NO BASQUETE VARIAM AMPLAMENTE ENTRE ATLETAS.
- O TEMPO DE JOGO NÃO É UNIFORME ENTRE OS JOGADORES.
- ROOKIES NÃO POSSUEM HISTÓRICO PARA COMPARAÇÃO ESTATÍSTICA.
- TREINADORES PODEM TROCAR DE EQUIPE DURANTE A TEMPORADA.

- TABELA QUE RELACIONA A ALTURA E O PESO DAS JOGADORAS ANTES DA LIMPEZA

EXPLORAÇÃO DE DADOS E LIMPEZA

- LIMPEZA DA TABELA PLAYERS
  - REMOÇÃO DE REGISTROS DUPLICADOS.
  - POSIÇÕES VAZIAS SÃO CLASSIFICADAS COMO UNKNOWN.
  - APLICAÇÃO DE LIMITES MÍNIMOS: PESO ≥ 60 KG E ALTURA ≥ 24 CM.



- LIMPEZA DA TABELA TEAMS
  - VALIDAÇÃO DA CONSISTÊNCIA: WON + LOST = GAMES PLAYED.
  - CAMPOS CATEGÓRICOS VAZIOS SÃO PREENCHIDOS COM UNKNOWN (DIVID, CONFID, ARENA).
  - ELIMINAÇÃO DE COLUNAS SEM UTILIDADE, ISTO É, COM VALORES IDÊNTICOS EM TODAS AS LINHAS.

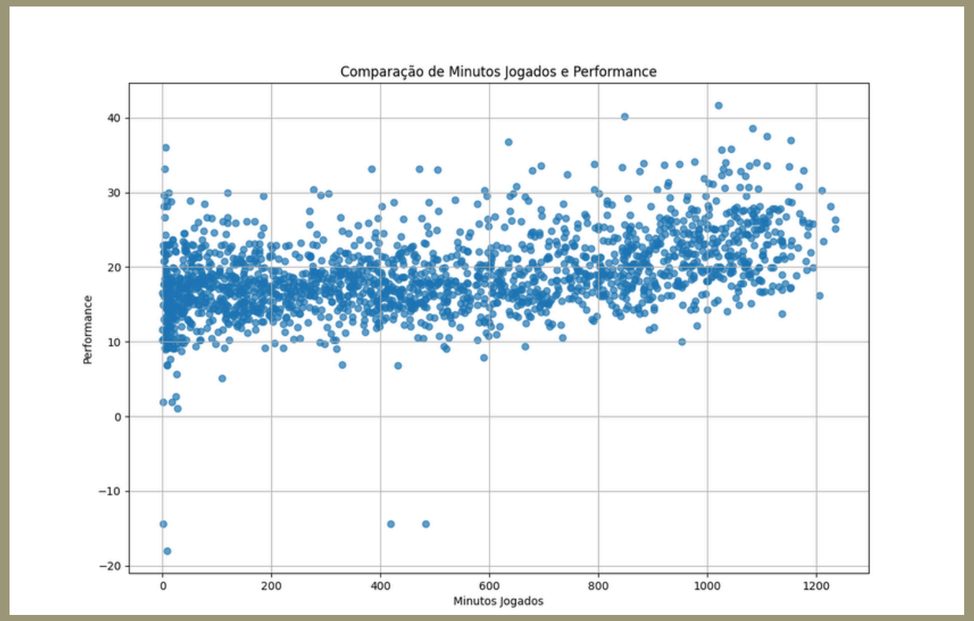
- TABELA QUE RELACIONA A ALTURA E O PESO DAS JOGADORAS APÓS A LIMPEZA

PREVISÃO DO RANKING

- DEFINIÇÃO DO PROBLEMA
  - DESENVOLVER UM MODELO PREDITIVO PARA ESTIMAR O RANKING DAS EQUIPES EM SUAS CONFERÊNCIAS, UTILIZANDO DADOS HISTÓRICOS E PADRÕES DE DESEMPENHO.
- VARIÁVEL-ALVO PARA PREVISÃO: “RANK”- DEFINIDA EM TEAMS.CSV.
  - DEVIDO À ALTA VOLATILIDADE ENTRE TEMPORADAS, O MODELO UTILIZA COMO FEATURES:
  - TENDÊNCIA RECENTE DA EQUIPE, DESEMPENHO DA EQUIPE, DESEMPENHO DO TÉCNICO E RANKINGS ANTERIORES.
- PIPELINE DO MODELO
  - LIMPEZA DE DADOS → CLEAN_PLAYERS.PY, CLEAN_TEAMS.PY
  - PERFORMANCE DOS JOGADORES → PLAYER_PERFORMANCE.PY
  - PERFORMANCE DAS EQUIPES → TEAM_PERFORMANCE.PY
  - PERFORMANCE DOS TÉCNICOS → COACH_SEASON_FACTS_PERFORMANCE.CSV
  - MODELO DE PREVISÃO DE RANKING → TEAM_RANKING_MODEL.PY
  - GERAÇÃO DE GRÁFICOS → MODEL_GRAPHICS.PY

PREVISÃO DO RANKING

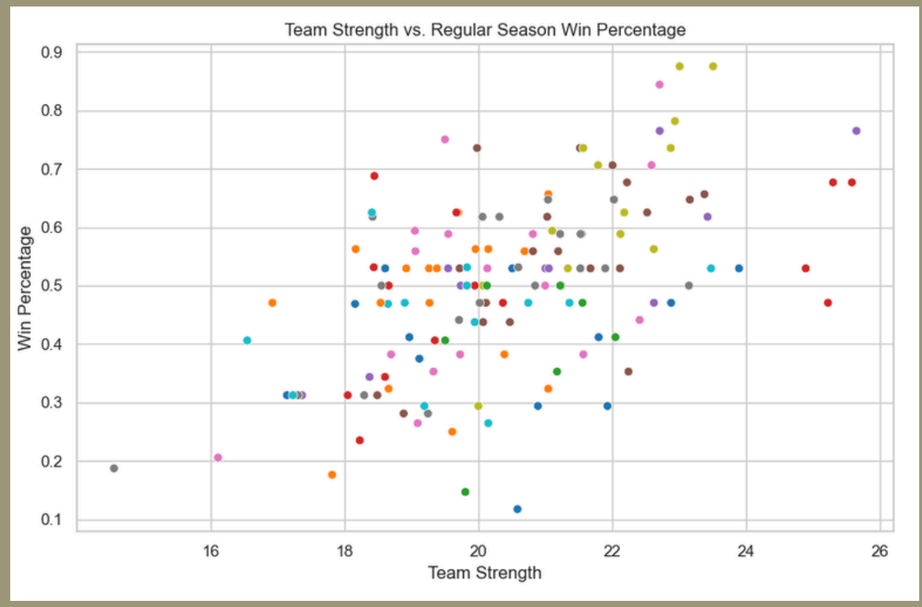
- PREPARAÇÃO DOS DADOS
  - CÁLCULO DA PERFORMANCE DOS JOGADORES BASEADO APENAS EM ESTATÍSTICAS INDIVIDUAIS DA TEMPORADA, INDEPENDENTEMENTE DO DESEMPENHO DA EQUIPE.
  - A PERFORMANCE É ESTIMADA POR MÉTRICAS PER-36, COMBINADAS POR MÉDIA PONDERADA CONFORME A POSIÇÃO.
  - PARA JOGADORES COM TEMPO DE QUADRA ABAIXO DO MÍNIMO ESTABELECIDO, APLICA-SE UM SISTEMA DE FALLBACK HIERÁRQUICO PARA SUBSTITUIR SUA PERFORMANCE.



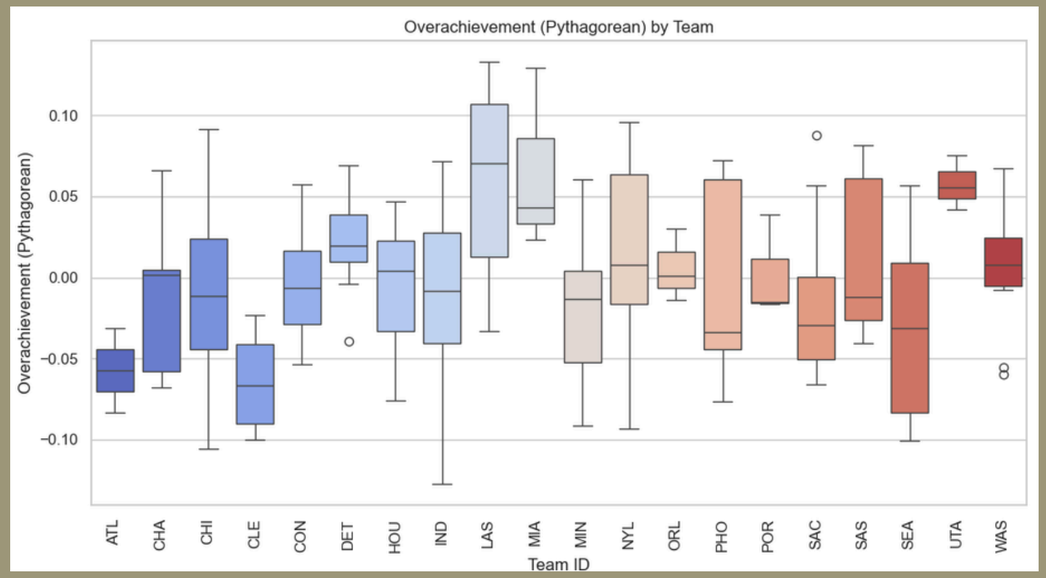
- TABELA QUE MOSTRA A RELAÇÃO DA PERFORMANCE COM OS MINUTOS JOGADOS, COMO PODEMOS VER, ESTÁ NORMALIZADA GRAÇAS AO PER 36, NÃO TEM UMA INCLINAÇÃO MUITO ACENTUADA.

PREVISÃO DO RANKING

- PREPARAÇÃO DOS DADOS:
  - TIME:
    - FORÇA DO TIME: MÉDIA PONDERADA DA PERFORMANCE DOS JOGADORES PELO TEMPO JOGADO, ESTIMANDO A QUALIDADE GLOBAL DO ELENCO.
    - ESTATÍSTICA PITAGÓRICA:PREVISÃO DA TAXA DE VITÓRIAS A PARTIR DOS PONTOS MARCADOS E SOFRIDOS
      - (PYTHAG_WIN_PCT = PF^EXP / (PF^EXP + PA^EXP)) .
        - PF : PONTOS A FAVOR
        - PA: PONTOS CONTRA
    - REGRESSÃO LINEAR: MODELO QUE RELACIONA A FORÇA DO ELENCO À EXPECTATIVA DE VITÓRIAS, ESTIMANDO QUANTAS VITÓRIAS O TIME “DEVERIA” ALCANÇAR.



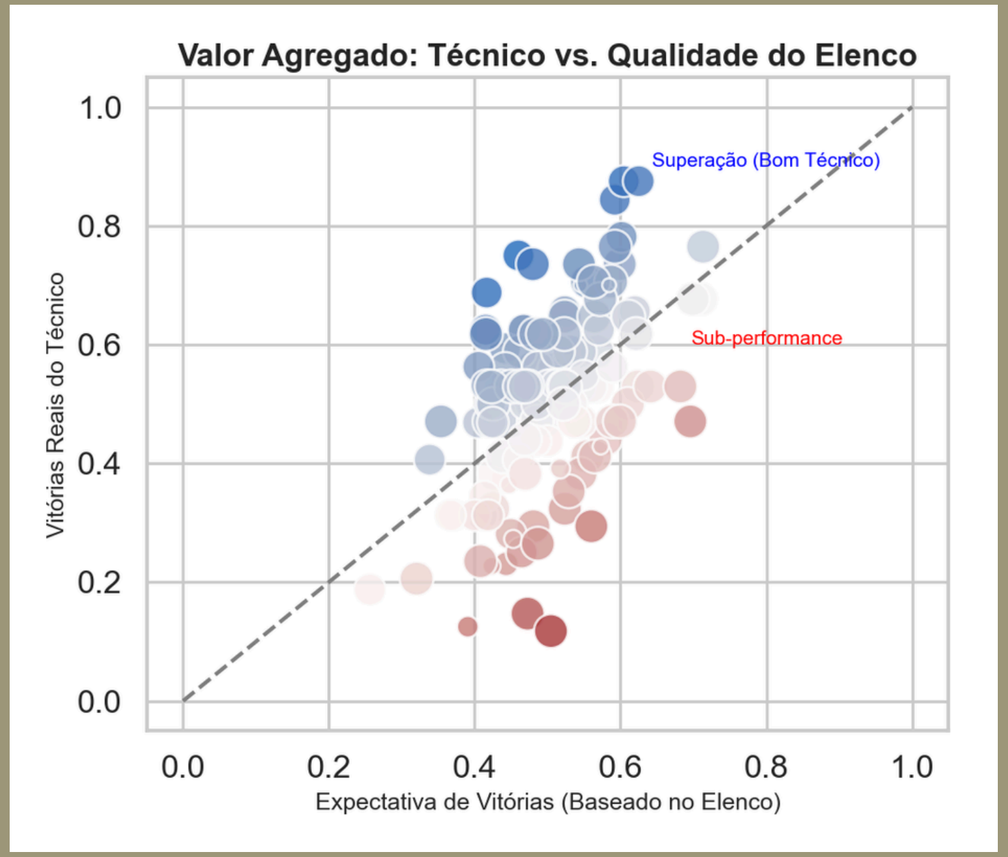
- A RELAÇÃO ENTRE A "FORÇA DO ELENCO" CALCULADA E A % DE VITÓRIAS REAL NA TEMPORADA.



- A DIFERENÇA ENTRE QUANTAS VITÓRIAS O TIME TEVE E QUANTAS ELE DEVERIA TER TIDO BASEADO NO SALDO DE PONTOS

PREVISÃO DO RANKING

- PREPARAÇÃO DOS DADOS:
  - TÉCNICO
    - CÁLCULO DE MÉTRICAS BÁSICAS, COMO O PERCENTUAL DE VITÓRIAS (RS_WIN_PCT_COACH).
    - USO DE CÓDIGO STINT-AWARE PARA TRATAR SEPARADAMENTE PERÍODOS COM TROCA DE TÉCNICO.
    - APLICAÇÃO DE SUAVIZAÇÃO BAYESIANA EMPÍRICA PARA CORRIGIR AMOSTRAS PEQUENAS, AJUSTANDO O DESEMPENHO EM DIREÇÃO À MÉDIA DA LIGA.
    - AVALIAÇÃO DE VALOR AGREGADO: COMPARA-SE A TAXA DE VITÓRIAS DO TÉCNICO COM A FORÇA DO ELENCO (TEAM_STRENGTH) PARA IDENTIFICAR SE O TREINADOR ENTREGA RESULTADOS ACIMA DO ESPERADO.



- RELAÇÃO DAS VITÓRIAS DOS TÉCNICOS COM A QUANTIDADE DE VITÓRIAS ESPERADAS DO ELENCO
- TÉCNICOS BONS TRAZEM MAIS VITORIAS DO QUE ESPERADO

PREVISÃO DO RANKING

- PREVISÃO DE RANKING
- MODELO UTILIZADO: CLASSIFICATIVO, AO INVÉS DE REGRESSIVO.
- MÉTODO: PAIRWISE LEARNING-TO-RANK COM GRADIENT BOOSTING CLASSIFIER.
  - PREVÊ RANKINGS DE EQUIPES USANDO DADOS HISTÓRICOS E CARACTERÍSTICAS TEMPORAIS (MÉDIAS MÓVEIS E TENDÊNCIAS).
- SETUP EXPERIMENTAL:
  - ITERAÇÃO 1: TREINO (ANOS 1-6) | VALIDAÇÃO (ANO 7)
  - ITERAÇÃO 2: TREINO (ANOS 1-7) | VALIDAÇÃO (ANO 8)
  - TESTE FINAL: TREINO (ANOS 1-8) | TESTE (ANOS 9-10)

PREVISÃO DO RANKING

- FEATURES DO HISTÓRICO, CONSISTÊNCIA E QUALIDADE ESTRUTURAL
  - COACH_CAREER_RS_WIN_PCT_MA5: MÉDIA DE VITÓRIAS DO TÉCNICO NA CARREIRA (5 ANOS), REFLETINDO EXPERIÊNCIA.
  - TEAM_STRENGTH_MA5: FORÇA MÉDIA DO ELENCO NOS ÚLTIMOS 5 ANOS, INDICANDO POTENCIAL DO GRUPO.
  - CONF_WIN_PCT_MA5: MÉDIA DE VITÓRIAS NA CONFERÊNCIA (5 ANOS), MEDINDO COMPETITIVIDADE REGIONAL.
  - OFF_EFF_MA5: EFICIÊNCIA OFENSIVA MÉDIA EM 5 ANOS, CAPTURANDO QUALIDADE ESTRUTURAL.
  - DEF_EFF_MA5: EFICIÊNCIA DEFENSIVA MÉDIA EM 5 ANOS.
  - PYTHAG_WIN_PCT_MA5: PORCENTAGEM DE VITÓRIAS ESPERADAS (PYTHAGOREAN) MÉDIA EM 5 ANOS, AJUSTADA POR PONTOS.
  - POINT_DIFF_MA5: SALDO DE PONTOS MÉDIO DOS ÚLTIMOS 5 ANOS, REFLETINDO ATAQUE/DEFESA.

PREVISÃO DO RANKING

- FEATURES DE TENDÊNCIAS, ESTABILIDADE E RECÊNCIA
  - POINT_DIFF_TREND5: TENDÊNCIA DO SALDO DE PONTOS EM 5 ANOS, INDICANDO EVOLUÇÃO.
  - PYTHAG_WIN_PCT_TREND5: TENDÊNCIA DA PORCENTAGEM DE VITÓRIAS ESPERADAS EM 5 ANOS.
  - WIN_PCT_CONSISTENCY: ESTABILIDADE DA TAXA DE VITÓRIAS AO LONGO DO TEMPO.
  - PREV_RANK_MA5: POSIÇÃO MÉDIA DOS ÚLTIMOS 5 ANOS (CONSISTÊNCIA HISTÓRICA).
  - PREV_RANK_MA3: POSIÇÃO MÉDIA DOS ÚLTIMOS 3 ANOS (JANELA MAIS CURTA).
  - COACH_TENURE_PREV: TEMPO DO TREINADOR NO TIME NA TEMPORADA ANTERIOR (ESTABILIDADE TÉCNICA).
  - PREV_RANK_1: POSIÇÃO DO TIME NO ANO ANTERIOR (EFEITO IMEDIATO).

FUNCIONAMENTO DA PREVISÃO DO RANKING

- A FUNÇÃO GENERATE_PAIRWISE_DATA CONSTRÓI PARES DE EQUIPAS E CALCULA DIFERENÇAS VETORIAIS ENTRE AS SUAS FEATURES ESTATÍSTICAS :

Equipa	Feature 1 (`win_pct`)	Feature 2 (`strength`)	Feature 3 (`tenure`)
A (Lakers)	0.60 (60%)	15.5	1 ano
B (Celtics)	0.55 (55%)	18.0	4 anos

- EXEMPLO DE VETORES DE FEATURES
- (E.G., EQUIPA A - EQUIPA B = [+0.05, -2.5, -3])

- O MODELO GRADIENTBOOSTINGCLASSIFIER APRENDE PADRÕES NESSAS DIFERENÇAS RELATIVAS, ESTIMANDO PROBABILIDADES DE SUPERIORIDADE P(A > B).
- NA FUNÇÃO PREDICT_RANKS_PAIRWISE, O MODELO REALIZA COMPARAÇÕES EXAUSTIVAS ENTRE TODAS AS EQUIPAS DE UMA CONFERÊNCIA. PARA CADA PAR, PREVÊ A PROBABILIDADE DE VITÓRIA (E.G., P(A > B) = 0.8).
- CÁLCULO DO SCORE FINAL:
  - SCORE(A) = Σ P(A > I) PARA TODAS AS EQUIPAS I
  - SCORE(B) = Σ P(B > I) PARA TODAS AS EQUIPAS I
- EXEMPLO: SCORE(A) = 1.7 → RANK 1; SCORE(B) = 0.8 → RANK 2
- VANTAGEM: ESTA ABORDAGEM PAIRWISE CAPTURA DIFERENÇAS ORDINAIS ENTRE EQUIPAS, ALINHANDO-SE COM O OBJETIVO DE CLASSIFICAÇÃO RELATIVA EM VEZ DE PREVISÃO ABSOLUTA DE DESEMPENHO.

FUNCIONAMENTO DA PREVISÃO DO RANKING

- PROCESSO DE PREVISÃO E RANKING
  - A FUNÇÃO PREDICT_RANKS_PAIRWISE EXECUTA COMPARAÇÕES EXAUSTIVAS ENTRE TODAS AS EQUIPAS DE UMA CONFERÊNCIA, CALCULANDO PROBABILIDADES DE SUPERIORIDADE PARA CADA PAR.
  - ALGORITMO DE SCORING:
  - PARA CADA EQUIPA I, O SCORE FINAL É DADO POR:
    - $\text{SCORE}(I) = \Sigma_j P(I > J)$
  - ONDE J REPRESENTA TODAS AS OUTRAS EQUIPAS DA CONFERÊNCIA.
- EXEMPLO ILUSTRATIVO:
  - EQUIPA A: $P(A>B) + P(A>C) = 0.8 + 0.9 = 1.7 \rightarrow$ RANK 1
  - EQUIPA B: $P(B>A) + P(B>C) = 0.2 + 0.6 = 0.8 \rightarrow$ RANK 2
  - EQUIPA C: $P(C>A) + P(C>B) = 0.1 + 0.4 = 0.5 \rightarrow$ RANK 3
- FUNDAMENTAÇÃO: ESTA ABORDAGEM PRIORIZA A ORDENAÇÃO RELATIVA SOBRE MÉTRICAS ABSOLUTAS. EM CONTEXTO COMPETITIVO, A CLASSIFICAÇÃO BASEIA-SE NA SUPERIORIDADE COMPARATIVA ENTRE EQUIPAS, NÃO EM VALORES NUMÉRICOS ISOLADOS DE DESEMPENHO.

PREVISÃO DO RANKING

- MÉTRICAS USADAS:
- MAE (MEAN ABSOLUTE ERROR): QUANTIFICA, EM TERMOS LINEARES, O DESVIO MÉDIO DAS POSIÇÕES ATRIBUÍDAS PELO MODELO.
- SPEARMAN (SPEARMAN'S RANK CORRELATION COEFFICIENT): INDICA O GRAU EM QUE A ORDEM RELATIVA DOS ELEMENTOS É PRESERVADA ENTRE O RANKING VERDADEIRO E O RANKING PREVISTO.
- NDCG (NORMALIZED DISCOUNTED CUMULATIVE GAIN): UTILIZA UM FATOR DE DESCONTO LOGARÍTMICO PARA PENALIZAR ERROS EM POSIÇÕES MAIS BAIXAS, NORMALIZANDO O RESULTADO DE MODO A FACILITAR COMPARAÇÕES ENTRE LISTAS.
- WITHIN-K ACCURACY: MEDE A PROPORÇÃO DE INSTÂNCIAS EM QUE O ITEM CORRETO APARECE ENTRE AS K PRIMEIRAS POSIÇÕES DO RANKING PRODUZIDO PELO MODELO. QUANTIFICA A CAPACIDADE DO MÉTODO EM POSICIONAR O ELEMENTO RELEVANTE PRÓXIMO AO TOPO, MESMO QUANDO NÃO ACERTA A POSIÇÃO EXATA.

```
============================================
TRAIN METRICS (AGGREGATED)
============================================
MAE_rank: 1.5826
Mean_Spearman: 0.4516
Mean_NDCG@10: 0.9164
n_groups: 16

============================================
VALIDATION METRICS
============================================

Validation Year 7:
  Conference              EA          WE
  MAE_rank            0.5714      0.2857
  Spearman            0.8929      0.9643
  NDCG@10             0.9894      0.9960
  Within-1 Acc        85.71%     100.00%
  Within-3 Acc       100.00%     100.00%
  n_teams                  7           7

Validation Year 8:
  Conference              EA          WE
  MAE_rank            1.3333      2.2857
  Spearman            0.6000     -0.1071
  NDCG@10             0.9174      0.7634
  Within-1 Acc        50.00%      42.86%
  Within-3 Acc       100.00%      71.43%
  n_teams                  6           7
```

```
============================================
TEST METRICS
============================================

Test Year 9:
  Conference              EA          WE
  MAE_rank            1.1429      1.7143
  Spearman            0.7857      0.1786
  NDCG@10             0.9651      0.9002
  Within-1 Acc        71.43%      57.14%
  Within-3 Acc       100.00%      85.71%
  n_teams                  7           7

Test Year 10:
  Conference              EA          WE
  MAE_rank            2.5714      0.6667
  Spearman           -0.1429      0.8286
  NDCG@10             0.7992      0.9913
  Within-1 Acc        28.57%      83.33%
  Within-3 Acc        71.43%     100.00%
  n_teams                  7           6

============================================
OVERFITTING DIAGNOSIS
============================================
Train-Test MAE gap: 0.0271
Val-Test MAE gap: -0.4444
```

PREVISÃO DO RANKING

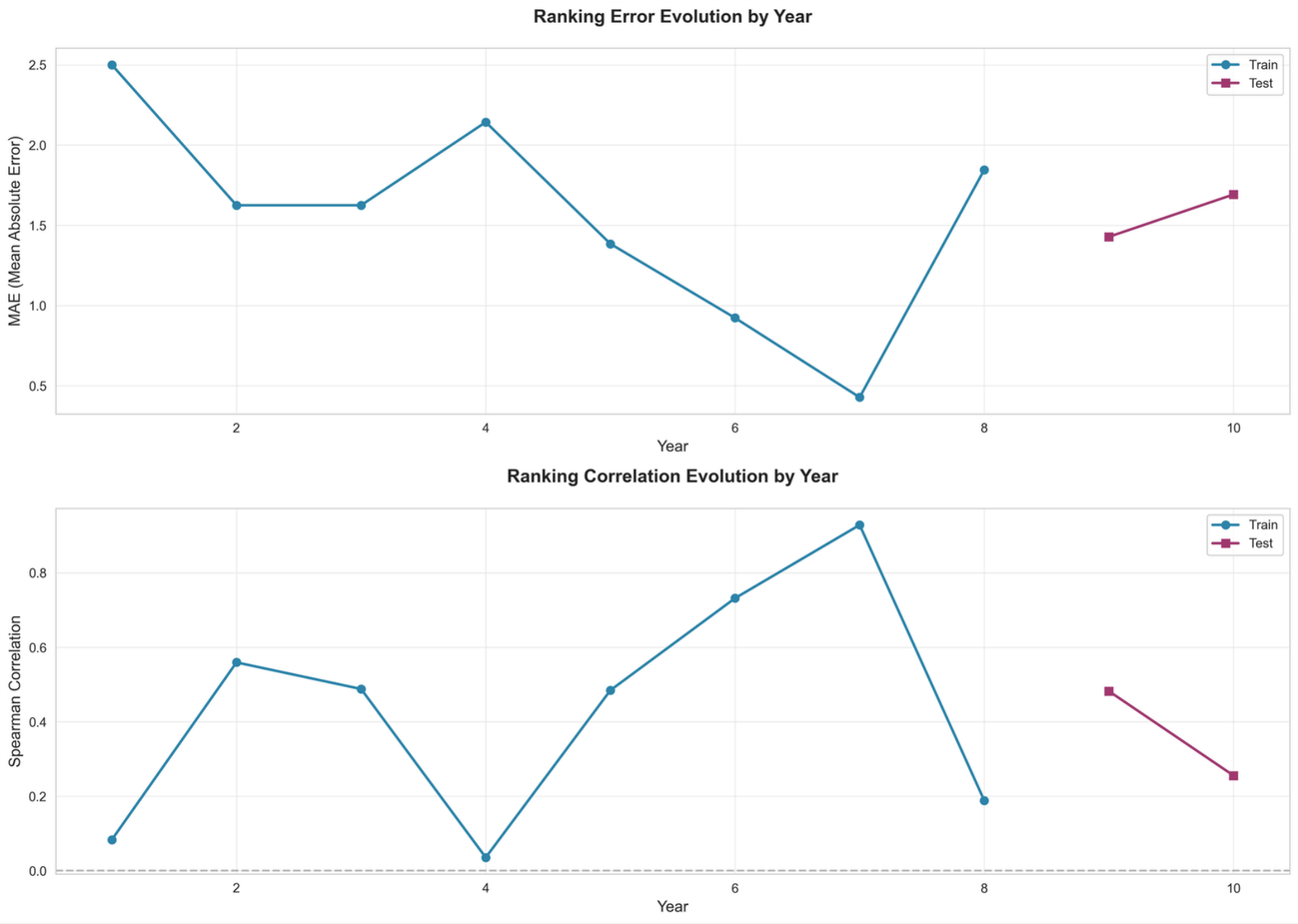
- RESULTADOS:
  - A DIFERENÇA DE DESEMPENHO ENTRE TREINO E TESTE É BAIXA, INDICANDO QUE O MODELO APRENDEU PADRÕES RELEVANTES SEM INCORRER EM OVERFITTING.
  - OBSERVA-SE VARIAÇÃO SIGNIFICATIVA ENTRE DIFERENTES CONFERÊNCIAS.
  - EMBORA O MODELO NÃO ACERTE EXATAMENTE O RANKING DE TODOS OS TIMES, APRESENTA ALTA PRECISÃO EM PROXIMIDADE, ALCANÇANDO WITHIN-3 ACCURACY DE 100% EM DIVERSOS CASOS.
  - DESTACAM-SE COMO OUTLIERS NEGATIVOS A CONFERÊNCIA EA NO ANO 10 E A CONFERÊNCIA WE NO ANO 8.

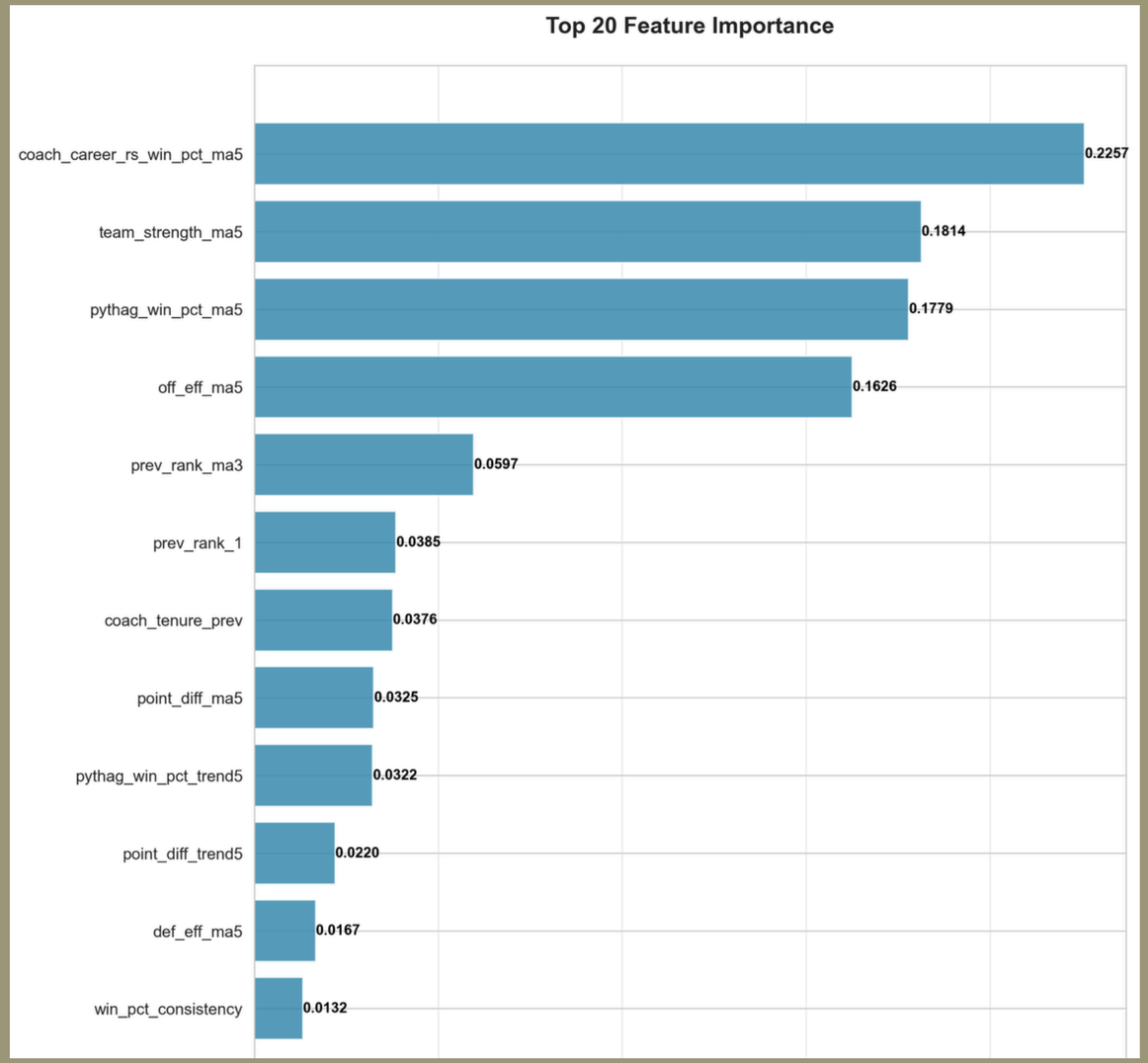
```
============================================
TRAIN METRICS (AGGREGATED)
============================================
MAE_rank: 1.5826
Mean_Spearman: 0.4516
Mean_NDCG@10: 0.9164
n_groups: 16

============================================
VALIDATION METRICS
============================================

Validation Year 7:
  Conference              EA          WE
  MAE_rank            0.5714      0.2857
  Spearman            0.8929      0.9643
  NDCG@10             0.9894      0.9960
  Within-1 Acc        85.71%     100.00%
  Within-3 Acc       100.00%     100.00%
  n_teams                  7           7

Validation Year 8:
  Conference              EA          WE
  MAE_rank            1.3333      2.2857
  Spearman            0.6000     -0.1071
  NDCG@10             0.9174      0.7634
  Within-1 Acc        50.00%      42.86%
  Within-3 Acc       100.00%      71.43%
  n_teams                  6           7
```

```
============================================
TEST METRICS
============================================

Test Year 9:
  Conference              EA          WE
  MAE_rank            1.1429      1.7143
  Spearman            0.7857      0.1786
  NDCG@10             0.9651      0.9002
  Within-1 Acc        71.43%      57.14%
  Within-3 Acc       100.00%      85.71%
  n_teams                  7           7

Test Year 10:
  Conference              EA          WE
  MAE_rank            2.5714      0.6667
  Spearman           -0.1429      0.8286
  NDCG@10             0.7992      0.9913
  Within-1 Acc        28.57%      83.33%
  Within-3 Acc        71.43%     100.00%
  n_teams                  7           6

============================================
OVERFITTING DIAGNOSIS
============================================
Train-Test MAE gap: 0.0271
Val-Test MAE gap: -0.4444
```

PREVISÃO DO RANKING

- **GRÁFICOS:**





PREVISÃO DO RANKING

- **ANO 11:**
  - **TREINAR DO ANO 1 AO 10:**
  - **TESTAR ANO 11:**