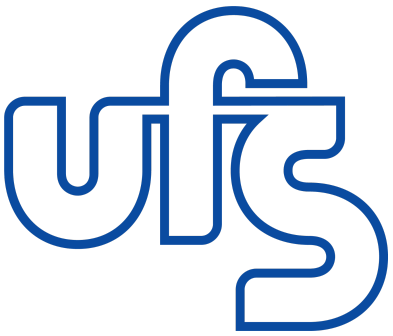

UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

Avaliação 3

Aprendizado de Máquina

JAIRO DE SANTANA DANTAS

JOÃO PAULO MENDONÇA

VITOR MANOEL SANTOS MOURA

São Cristóvão – Sergipe

2026

Resumo Técnico: Modelo de Predição de Trajetória de Jogadores (NFL Big Data Bowl 2026)

Visão Geral

Este relatório detalha a metodologia, a arquitetura e as estratégias de engenharia de dados empregadas no sistema de predição de movimentação para o NFL Big Data Bowl 2026. O objetivo central é estimar as coordenadas futuras (*X* e *Y*) de um jogador após *k frames*, utilizando como base seu estado e histórico recente.

Processamento e Estruturação de Dados.

A base para a modelagem é a consolidação e a ordenação estrita dos dados de rastreamento:

- **Agregação Global:** A concatenação de todos os arquivos de entrada e saída (`input_df` e `output_df`) permite que o modelo capture padrões de movimento que transcendem jogos ou jogadas individuais, otimizando a generalização.
- **Ordenação Temporal Crítica:** A sequência de ordenação (`game_id`, `play_id`, `nfl_id`, `frame_id`) é vital. A ausência dessa ordenação invalidaria as operações de *Lag* (memória histórica), misturando dados cinemáticos de forma incorreta.

Engenharia de Features: Traduzindo a Física para o Modelo

As *features* foram cuidadosamente projetadas para traduzir a inércia e a cinemática do futebol americano em variáveis processáveis pelo XGBoost.

Seção	**Feature**	**Descrição e Justificativa**
Memória Histórica	*Lag* (1, 2, 3 e 5 frames) para *x, y* e *s*	Permite ao modelo capturar a inércia, a

		aceleração e o *jerk* (derivada da aceleração), essencial para entender mudanças de direção súbitas.
Decomposição Vetorial	$vx = s \cdot \cos(dir)$ e $vy = s \cdot \sin(dir)$	Remove a descontinuidade numérica das grandezas circulares (*0°* e *360°* são iguais). A decomposição cartesiana é mais eficaz para modelos baseados em árvore.
Cinemática Progressiva	$v \cdot t$, k^2, $\frac{1}{2}at^2$	Features baseadas nas fórmulas de Movimento Retilíneo Uniformemente Variado (MRUV). Fornecem ao XGBoost uma **base física inicial**, permitindo que ele se concentre apenas no ajuste dos desvios (resíduos) da trajetória ideal.
Dinâmica com a Bola	`dist_ball`, `angle_ball`, `dot_ball_vel`	O Produto Escalar (`dot_ball_vel`) é crucial, indicando se o jogador está se movendo **na direção** ou **afastando-se** do ponto de queda previsto da bola.

Estratégias de Transformação Críticas

Duas transformações se destacaram como diferenciais técnicos:

Normalização do Alvo (Target Transformation)

Em vez de prever a posição absoluta (x_{target}, y_{target}) ou o deslocamento total $(\Delta x, \Delta y)$, o modelo prevê a **velocidade média necessária** para o deslocamento:

$$target_x = \frac{x_{target} - x_{input}}{k} \quad target_y = \frac{y_{target} - y_{input}}{k}$$

Impacto: Estabiliza o gradiente e garante que erros de predição para horizontes de tempo curtos (baixo *k*) tenham o mesmo peso relativo que erros de longo prazo, mitigando o problema da escala de campo.

Simetria de Campo (Direction Standardization)

Para reduzir a variância, todas as jogadas para a `left` são padronizadas para a direção `right`.

Ação: Inversão da coordenada *X* (*120* - *x*) e inversão da velocidade *vx* (multiplicada por *-1*).

Impacto: O modelo não precisa aprender duas representações para o mesmo movimento. Ele aprende um conceito unificado de "avançar no sentido da jogada", reduzindo pela metade a complexidade do espaço de dados.

Configuração do Modelo (XGBoost)

Foi utilizada uma abordagem de dois regressores `XGBRegressor` independentes (um para *X* e outro para *Y*), dada a diferença na dinâmica lateral versus vertical no futebol americano.

Hiperparâmetro	**Valor**	**Função no Modelo**
`n_estimators`	2000	Limite alto, controlado por *early_stopping* para evitar *overfitting*.
`max_depth`	6	Profundidade moderada para capturar interações

		entre *features* sem excesso de especialização.
`learning_rate`	0.05	Taxa conservadora para garantir uma convergência mais estável e precisa.
`subsample`, `colsample_bytree`	0.8, 0.8	Regularização estocástica: cada árvore é treinada em uma subamostra de dados e colunas, evitando a dependência excessiva de *features* únicas.
`tree_method`	"hist"	Otimização de performance (uso de histogramas) crucial para a escala de milhões de pontos de dados de rastreamento.

Estratégia de ValidaçãoGroupKFold (n_splits=5)

A validação é agrupada por **`game_id + play_id`**.

Motivo Crucial: Prevenir o *Data Leakage*. Se frames sequenciais da mesma jogada fossem divididos entre treino e validação, o modelo *decoraria* a trajetória em vez de *aprender a prevê-la*. O agrupamento garante que o teste ocorra em jogadas completamente inéditas.

Métrica de Desempenho

$$Metric = \sqrt{\frac{MSE_x + MSE_y}{2}}$$

A métrica representa a **distância euclidiana média** do erro, penalizando desvios significativos nas predições de trajetória.

Conclusões

Em suma, o modelo alcança precisão ao combinar o rigor da **física clássica** (integrada nas *features* cinemáticas) com a **robustez do XGBoost**, sendo a normalização pelo tempo (*k*) e a padronização direcional as chaves para sua performance estável em diferentes cenários e horizontes de predição.