WORKSHOP #3
KAGGLE SIMULATION – NFL IMPACT DETECTION

INTEGRANTS
NAHIN JOSE PEÑARANDA MEJIA (njpenarandam@udistrital.edu.co)
NICOLAS FELIPE PULIDO SUAREZ (nfpulidos@udistrital.edu.co)
ANDERSON DAVID ARENAS GUTIERREZ (adarenasg@udistrital.edu.co)

SYSTEMS ANALYSIS & DESIGN
UNIVERSIDAD DISTRITAL FRANCISCO JOSE DE CALDAS
JUNE – 2025
BOGOTA – COLOMBIA

Simulation Report:

The simulation aims to model and evaluate the performance of a machine learning system designed to detect player contacts in NFL games. The system integrates multimodal data from sensors, video feeds, and metadata, processed through microservices for ingestion, preprocessing, feature extraction, and inference. Chaos theory insights, including random perturbations and data losses, were incorporated to simulate real-world uncertainties.

Methodology

The simulation uses a chaos-engineered, modular methodology combined with machine learning pipeline prototyping. It introduces controlled disruptions, like noise and missing data, to test system resilience under real-world uncertainties. The design is modular, with separate microservices for data ingestion, preprocessing, feature extraction, and model training, ensuring clarity and maintainability.
The approach integrates machine learning by training a Random Forest Classifier, evaluating performance with F1-Score, and engineering features like distance and relative speed. It is iterative, running multiple scenarios to assess patterns in runtime, accuracy, and memory usage. The focus on data quality through preprocessing highlights its data-centric nature, making it a systematic and robust methodology for stress-testing and performance evaluation.

1. **Data Ingestion:**
    - CSV files for labels, tracking data, and helmet positions were loaded. Only a subset of relevant game plays was used.
2. **Preprocessing:**
    - Numeric values in tracking data were interpolated to handle missing values.
    - The data was merged to form a complete dataset for training and evaluation.
3. **Feature Extraction:**

    - Distance and relative speed between players were calculated as features.
    - Contact labels served as the target variable for machine learning.
4. **Model Training:**
    - A Random Forest Classifier was trained to predict contact events.
    - F1-Score was used to evaluate model performance.
5. **Chaos Simulation:**
    - Noise was added to player positions, and a fraction of tracking data rows was dropped to test system resilience.

Results

- **F1-Score:** The model achieved an average F1-Score of 0.13 to 0.17 across multiple simulations. This indicates that while the system can identify some patterns, its predictive performance is currently limited, likely due to the quality of the features or the data disruptions introduced during chaos simulation.
- **Runtime:** Each simulation took approximately 14 to 15 seconds, which exceeds the desired processing time. This suggests the need for optimization in the preprocessing pipeline or model training process.

Discussion and Findings

- The low F1-Score points to potential issues in feature engineering or model selection. Features like distance and relative speed may not fully capture the complexity of player interactions, or the dataset may require more robust preprocessing to handle disruptions.
- The runtime of 14 to 15 seconds poses a significant bottleneck, particularly for real-time applications. Optimization in feature extraction and data merging could improve this.