Deliverable #1

Predicting Injury Risk in Elite Football Using Integrated GPS, Recovery, and Physical Performance Data

The Analytics IV: Santiago Botero, Sebastián de Wind, Thomas Arturo Renwick Morales, Santiago Ruiz

1. Introduction

Topic Overview

In recent years, the application of data analytics in sports has transformed player management, training optimization, and injury prevention strategies. However, predicting player injuries using integrated performance data remains a complex and underexplored challenge, especially at the elite football level. This project leverages detailed player workload, recovery, and physical capability data to address this critical gap.

Research Question

This project aims to use traditional machine learning techniques to predict the likelihood of injury in professional football players based on their workload, recovery status, and physical capability metrics.

2. Motivation and Significance

Problem Statement

Player injuries in elite football lead to significant financial loss, disrupted team performance, and extended rehabilitation periods. Despite advances in sports science, accurately forecasting injury risk based on objective performance and recovery data remains challenging.

Importance of the Study

Developing an injury risk prediction model could:

- Help teams like Chelsea FC personalize training loads.
- Reduce injury incidence through proactive interventions.
- Optimize player availability and performance over the season.

The project's outcomes could have significant real-world impact on player health management, sports team competitiveness, and data-driven decision-making in professional sports environments.

3. Intended Data Sources

Data Description

We intend to use Chelsea FC's Performance Vizathon datasets to carry out this project. The specific datasets we plan to use are listed and explained below:

- **GPS Data.csv**: Tracks daily player movement and workload metrics such as total distance covered, high-speed running distances, acceleration counts, session durations, and peak speeds.
- **Recovery Status Data.csv**: Includes composite scores for recovery across multiple domains (bio-markers, soreness, sleep, subjective recovery), aggregated into an overall daily recovery score.
- **Physical Capability Data_.csv**: Measures baseline physical qualities like sprint ability, jump capacity, and strength across isometric and dynamic expressions.

Accessibility

All datasets are provided through the publicly available Chelsea FC Performance Insights GitHub Repository. The data is synthetic but realistic, requiring no additional permissions.

A limitation is that the data currently pertains to a single player; however, the project will be framed for extrapolation to an entire squad scenario. Another limitation is that there are no injury event labels included in the datasets. In other words, the datasets do not contain any direct information indicating whether the player was injured on a given date. Therefore, to overcome this limitation we would have to contact the Chelsea FC's Performance Insights team to find out whether the player was injured and if that was the case, during which dates. Nevertheless, having briefly analyzed the GPS Data.csv dataset, it can be inferred that the player was perhaps injured, as well as the dates during which this potential injury took place. We just have to confirm these hypotheses.

4. Proposed Analytical Methods

We propose to study a classification problem, where we attempt to predict the likelihood of injury for a player given some features values (e.g., the aggregated distance the player ran, the average peak speed, etc.). For this classification problem, we first intend to use a logistic regression as a baseline model, and then once we have this baseline model, to use bagging and/or boosting techniques such as Random Forest or XGBoost to capture nonlinear interactions and improve predictive accuracy.

Overview of Methods

Data Cleaning/ Feature Engineering:

- Clean and analyze the datasets individually
- Classify injury in the GPS Data.csv dataset
- Merge all 3 datasets using the date as a primary key
- Feature selection via correlation analysis and deeper exploratory data analysis

Machine Learning Models:

- Baseline model: Logistic Regression (to predict injury probability)

- Advanced models: Random Forests and XGBoost (to capture nonlinear interactions and improve predictive accuracy)

Model Evaluation:

- Metrics: Recall, Precision, AUC-ROC, and F1-score
- We will use time-based validation, training the model on earlier sessions and testing on later sessions, to prevent data leakage and ensure realistic injury risk forecasting

Justification

Random Forests and XGBoost are appropriate due to their robustness to noisy sports data, ability to model complex feature interactions, and proven success in injury prediction studies.

5. Expected Outcomes

Anticipated Results

We anticipate developing a trained machine learning model capable of predicting the probability of a player sustaining an injury on any given day based on workload, recovery, and physical capability metrics. In addition, we plan to deploy a Streamlit application that allows users to input player data manually or upload CSV files, using the pre-trained model to generate injury risk predictions. Through this process, we aim to identify key workload, recovery, and physical traits that most significantly influence injury risk. Ultimately, the project will result in a flexible and scalable framework that could be applied across an entire football squad to support proactive injury management.

Potential Challenges

One of the primary challenges of this project is the significant class imbalance, as injuries are rare events compared to the number of non-injury observations. This can bias the model toward predicting the majority class unless properly addressed. Additionally, since the available dataset represents a single synthetic player, the model's ability to generalize to a full squad or real-world scenarios may be limited. We will address these challenges by applying a range of class balancing techniques. These include using class weights, SMOTE for oversampling the minority class, and evaluating model performance with metrics such as recall, precision, F1 score, and AUC-ROC to ensure robustness despite the imbalance. The project will be framed with the goal of extrapolating to a multi-player context in a real-world setting.

6. References

Chelsea FC Performance Insights. (2025). *Chelsea FC Performance Insights Vizathon – Data Overview*. Retrieved April 29, 2025, from https://chelsea-fc-performance-insights.github.io/Competition/#data

Chelsea FC Performance Insights. (2025). *Chelsea FC Performance Insights Vizathon – GitHub Repository*. Retrieved April 29, 2025, from https://github.com/Chelsea-Fc-Performance-Insights/Competition