Pattern Discovery of Injury Risk Factor in Under-25 Professional Footballers through Association Rules and Clustering.

Alexander San Agustin-Melendez
College of Computer and Software Engineering
Kennesaw State Univeristy
Kennesaw, USA
asanagus@students.kennesaw.edu

***Abstract*—The incidence of serious injuries among players under the age of 25 has become a major challenge for coaching and technical staffs in professional soccer. A poorly managed injury can result in a player missing crucial matches or, in the worst-case scenario, being sidelined for the entire season. This study aims to uncover hidden associations between workload metrics, biomechanical profiles, and injury incidence among soccer players in this age group. The research focuses on discovering latent patterns and non-obvious correlations using unsupervised learning techniques. By utilizing an integrated dataset of performance metrics and medical records, the study employs clustering algorithms to segment risk profiles and identify a priori combinations of factors that present a statistically significant increase in injury risk. The results seek to characterize signs of fatigue and vulnerability thresholds, providing a scientific basis for knowledge discovery in elite sports talent management.**

I. Introduction

During 2024, a total of 4,123 injuries were recorded across the five major European leagues (LaLiga, Serie A, Ligue 1, Bundesliga, and the Premier League), resulting in an economic loss of €732 million, according to a study by the British insurance group Howden. Notably, the report highlights that injuries among players under the age of 21 have surged by 187% since the 2020-2021 season [1]. Despite advances in sports medicine, the traditional approach to injury management has remained largely reactive. Consequently, there is an urgent need to shift toward a knowledge discovery framework. Rather than merely predicting injury likelihood, this project proposes the use of data mining techniques to reveal the underlying structures and association rules governing injury events in players under 25. By identifying specific relationships between workload, field position, and accumulated fatigue, this research seeks to establish distinct risk profiles.

II. Dataset Description

The project database is constructed by integrating two complementary sources of information.

A. Transfermarkt

A web scraping engine developed in Python will be implemented, using the **BeautifulSoup** library to extract detailed injury histories for players under 25 across the five major European leagues over the last 5 seasons. Key attributes to be captured include injury type (e.g., muscular or ligamentous), duration of absence, and date of occurrence. This data source provides the primary events for pattern analysis.

B. StatsBomb Open Data

The StatsBomb repository will be utilized to obtain high-fidelity event data. Unlike aggregated statistics, StatsBomb provides access to the precise location of events, defensive pressure exerted, and action intensity. These data points will serve as the contextual foundation for association rule mining, enabling the characterization of physical exertion and tactical strain preceding an injury [2].

C. Discovery Questions

- **Question 1:** Which combination of factors represents a significant elevation in the occurrence of injuries among players under the age of 25?
- **Question 2:** Are there any association rules linking previous injury history with patterns of accumulated fatigue among young players?

III. Data Procesing

Given that the data sources originate from disparate domains, the following preparation phases will be implemented:

A. Entity resolution

A name-matching algorithm will be implemented to ensure that StatsBomb event records align precisely with Transfermarkt injury profiles, accounting for variations in phonetic spelling and naming conventions.

B. Feature Engineering (Workload)

Granular event data will be transformed into quantifiable workload metrics. Temporal variables, such as cumulative minutes played over the previous 14 days, high-intensity distance covered, and sprint frequency per match, will be synthesized.

C. Handling Missing Values

Statistical imputation techniques will be applied to address incomplete performance records. Conversely, instances without reported injuries will be utilized as control cases (healthy profiles) to facilitate comparative clustering analysis.

TABLE I
ATRIBUTOS SELECCIONADOS PARA EL DESCUBRIMIENTO DE PATRONES

Atributo	Fuente	Tipo
Player_Age	Transfermarkt	Numérico
Injury_Type	Transfermarkt	Categórico
Minutes_30d	StatsBomb	Numérico
Rest_Days	StatsBomb	Numérico

IV. PLANNED TECHNIQUES

A. Association Rule Mining

Utilizing the **mlxtend** library, this study will extract tules in the form of

$$\{A, B\} \rightarrow \{C\}$$

where A represents a tactical position, B denotes a fatigue threshold, and C signifies a specific injury type. The objective is to identify rules with a **Lift** $>$ **2.0**, indicating a relationship significantly stronger than would be expected by chance.

B. Clustering for Vulnerability Profiles

K-Means or **DBSCAN** algorithms will be applied to categorize players based on their physiological response to physical load. This approach will facilitate the identification of "high vulnerability" clusters, determining if specific young players, due to their unique biomechanical profiles or playing styles (as captured by StatsBomb), exhibit higher risk factors than peers of the same age and position.

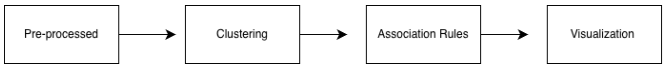

Fig. 1. Planned Data Mining Analysis Pipeline.

V. PRELIMINARY TIMELINE

The project will be executed in four distinct phases, aligning with the course milestones. Each phase is designed to build upon the previous discovery, ensuring a robust knowledge discovery process.

TABLE II
PROJECT MILESTONES AND DELIVERABLES

Milestone	Key Tasks	Status
M1: Proposal	Definition of discovery questions, toolstack setup (Fedora/Miniforge), and data source identification.	**Complete**
M2: EDA	Web scraping of Transfermarkt, data cleaning, and Exploratory Data Analysis to identify initial distributions.	In Progress
M3: Mining	Implementation of Clustering (K-Means/DBSCAN) and Association Rule Mining using the *mlxtend* library.	Planned
M4: Final	Interpretation of discovered patterns, visualization of "injury signatures," and final report delivery.	Planned

The schedule takes into account the complexity of merging disparate data sets (Transfermarkt and StatsBomb). A significant portion of the time in March will be devoted to *Entity Resolution* to ensure that medical records accurately match tactical event data. The final weeks will focus on qualitative evaluation of the discovered rules to ensure that they provide useful information for sports medicine.

REFERENCES

[1] "Howden's 2023/24 Men's European football injury index," Howden Insurance - Group, Oct. 17, 2024. https://www.howdengroupholdings.com/reports/2023-24-mens-european-football-injury-index

[2] statsbomb, "GitHub - statsbomb/open-data: Free football data from StatsBomb," GitHub, 2018. https://github.com/statsbomb/open-data/tree/master