Topic: Sports Injury Prediction

Author: Era Ebhodaghe

Bellevue University

Project #1

DSC 680 Applied Data Science

September 25th, 2025

Introduction

This project analyzes a dataset of 800 Chinese university football players from collegiate and provincial leagues. It applies machine learning classification methods to predict whether a player is likely to sustain an injury in the upcoming academic season.

In addition to building predictive models, the project aims to identify key risk factors and early indicators of injury. These insights can help coaches, trainers, and medical staff improve training plans, manage workloads, and develop targeted injury-prevention strategies—ultimately enhancing both player safety and performance.

Business Problem

Sports injuries pose a major challenge for athletes and organizations, leading to high medical costs, lost playing time, and reduced team performance. This dataset of 800 Chinese university football players allows predictive analytics to identify players at risk, enabling proactive prevention strategies and individualized training programs. By managing injury risk, organizations can optimize coaching and medical resources, support long-term player development, and improve team performance, while also informing recruitment and selection decisions.

Background/History

Sports injuries are prevalent among athletes and can significantly impact performance and well-being. According to the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS), sports injuries are categorized into acute and chronic types. Acute injuries occur suddenly due to events like falls or collisions, while chronic injuries result from repetitive stress and overuse over time. Common examples include sprains, strains, tendinitis, and fractures (NIAMS, 2024).

Understanding these injury types is crucial for developing effective prevention strategies. By identifying risk factors such as improper technique, overtraining, and inadequate recovery, programs can implement targeted interventions to reduce injury rates among athletes.

Data Source

The dataset was carefully selected from Kaggle based on its data quality and usability rating, eliminating the need for supplemental sources. Nonetheless, some feature engineering was performed to enhance model performance. The dataset contains the following attributes:

Dataset Details

- **Samples**: 800 university football players
- **Features**: 18 input features + 1 target label
- **Task**: Binary classification (0 = No injury, 1 = Injury)
- **Balance**: Well-balanced dataset
- **Age Range**: 18-24 years (typical university students)

Feature Categories

Physical Characteristics (4 features)

- Age, Height, Weight, BMI
- Measured through standard university health checkups

Football-Specific Metrics (4 features)

- Playing position, training hours per week, matches played, injury history
- Collected from official records and coach evaluations

Physical Fitness Assessment (6 features)

- Knee strength, hamstring flexibility, reaction time, balance, sprint speed, agility
- Professional fitness testing using standardized protocols

Lifestyle Factors (3 features)

- Sleep hours, stress level, nutrition quality
- Self-reported surveys and validated questionnaires

Training Compliance (1 feature)

- Warm-up routine adherence (0=Poor, 1=Good)

Target Variable

- Injury_Next_Season: Binary classification where injury is defined as training/competition-related injury causing ≥7 consecutive days of absence, verified by university medical center and coaching staff.

Methods

The data analysis for this project was done using Python programming language and Jupyter Notebooks as the IDE. Some of the Python libraries used are:

1. Numpy for computations
2. Pandas for data analysis, manipulation and cleaning
3. Matplotlib and Seaborn for visualization and plotting
4. Scikit-learn for Machine Learning

In addition to the mentioned packages, I incorporated other libraries as needed for my project.

After cleaning and preprocessing the data, I implemented various classification models, including Decision Tree, Random Forest, and Logistic Regression. Below are the reasons for choosing these models:

1. **Decision Tree** – Provides easy interpretability and effectively handles missing values.
2. **Random Forest** – Combines multiple decision trees to reduce overfitting and improve accuracy.
3. **Logistic Regression** – A simple and efficient algorithm for modeling the linear relationship between input variables and categorical outcomes.

Analysis

Data preprocessing was critical to ensuring the quality and utility of the dataset for predictive modeling. Binary categorical variables, such as "Yes" and "No" responses, were converted into boolean values (True or False) to ensure compatibility with machine learning algorithms. Additionally, multi-class categorical features were transformed using label encoding, converting text labels into numeric form. This allowed the models to interpret categorical data without unnecessarily expanding the feature space.

To assess the generalizability of the models, the dataset was split into training and testing subsets using an 80:20 ratio. This approach preserved a representative sample for performance evaluation while reserving sufficient data for training. Feature scaling was applied to normalize the range of continuous variables. This step was especially important for algorithms sensitive to feature magnitude, such as Logistic Regression and Gradient Boosting, helping to ensure balanced learning.

Following preprocessing, four classification algorithms were selected for evaluation: Random Forest, Decision Tree, Gradient Boosting, and Logistic Regression. These models were chosen for their effectiveness in binary classification and their distinct strengths in interpretability, complexity handling, and robustness. Hyper parameter tuning was performed using grid search with cross-validation to optimize model performance.

Performance was measured using several key metrics: accuracy, precision, recall, and F1-score. While Random Forest achieved the highest overall accuracy, Logistic Regression demonstrated

the highest recall and ROC AUC score. This made it particularly valuable in this scenario, where predicting future injuries is far more critical than reducing false positives. After optimization, Logistic Regression yielded a recall of 0.975, ROC AUC of 0.997, and overall accuracy of 0.975. The confusion matrix confirmed further justifies the accuracy of the Logistic Regression model where it correctly identifies 78 injuries cases while missing 2.

Conclusion

The primary objective of this project was to develop a predictive model capable of determining whether a football player is likely to sustain an injury in the following season based on health-related variables. Given the significant short- and long-term consequences associated with sports injuries, early and reliable identification of at-risk individuals is essential.

The final model demonstrated strong performance, achieving a balanced accuracy of 0.975 and a recall of 0.975, underscoring its effectiveness in correctly identifying true injury cases. An analysis of feature importance revealed that variables such as Knee Strength Score, Hamstring Flexibility, Reaction Time, Balance Test Score, Sprint Speed, Sleep Hours per Night, and Stress Level Score were the most influential predictors. These results emphasize the importance of adequate rest, targeted flexibility training, and stress reduction strategies as key measures in mitigating injury risk.

This project highlights a practical application in which recall is prioritized over overall accuracy, reflecting the critical need to minimize false negatives when predicting injuries in athletic contexts. By ensuring that potential injury cases are not overlooked, the model aligns with the broader objective of safeguarding player health and performance.

Assumptions

This project assumes that the dataset is representative of the target population, that historical relationships between health variables and injuries predict future risk, and that all measurements, including self-reported factors like sleep and stress, are accurate. It also assumes a consistent definition of injury, independence of players' risk, relative stability of physical and health variables over time, and that no critical factors influencing injury risk are missing from the data.

Limitations

A key limitation of this project is the scope and quality of the dataset. With only 800 university athletes represented, the model's generalizability is restricted, and its predictions may not extend to other levels of competition or broader player populations. In addition, several important risk factors such as medical history, genetics, or external lifestyle influences were not captured, and some variables like sleep or stress levels may be unreliable due to self-reporting.

Another limitation relates to bias and practical use. If certain player groups are overrepresented in the data, predictions may skew unfairly toward those groups. Injury risk is also dynamic and influenced by sudden changes in training or external conditions that the model cannot easily account for. While the model offers strong predictive power, it should be used alongside expert judgment rather than as the sole basis for decisions.

Challenges

Key challenges include ensuring data quality and completeness, especially for subjective measures like sleep and stress, as inconsistencies can reduce model accuracy. Maintaining fairness is also critical, as overrepresentation of certain player groups could introduce bias. Additionally, injury risk is dynamic, and sudden changes in player health or external conditions may not be captured by the model. Finally, integrating predictive insights into decision-making without overreliance, while ensuring staff understand and use the system responsibly, presents practical and ethical challenges.

Future Uses/Additional Applications

Future work will focus on implementing advanced feature engineering, increasing player data and model optimization strategies to further enhance predictive performance and clinical applicability.

Recommendations

Based on the findings of this project, it is recommended that the university's sports program integrate predictive modeling into its athlete health management practices to proactively identify players who may be at higher risk of injury. Training and conditioning programs should place stronger emphasis on improving knee strength, hamstring flexibility, and balance, while also monitoring additional factors such as sprint speed, reaction time, sleep quality, and stress levels. Wellness initiatives within the program should also promote adequate rest and effective stress management strategies to further reduce injury susceptibility.

Given the model's high recall performance, athletes flagged as at risk should be prioritized for preventive interventions, including personalized training adjustments and modified workloads. However, decisions related to athlete participation and team selection must continue to rely on a holistic evaluation of performance and ability, rather than predictive results alone, to maintain fairness and integrity. By adopting these measures, the university's sports program can strengthen injury prevention efforts, safeguard athlete well-being, and enhance overall team performance across seasons.

Implementation Plan

Implementation should begin with a structured system for collecting accurate, consistent player health data across all positions and seasons. The predictive model can then be trained and periodically updated with new data. A user-friendly interface should allow coaches and medical staff to review predictions and design preventive strategies, such as customized training or recovery plans. Clear guidelines must ensure that the model supports, rather than dictates, decisions, while ongoing monitoring and staff training maintain accuracy, ethical use, and program adoption.

Ethical Assessment

The results of this analysis are intended exclusively as a proactive measure to inform modifications in players' training, exercise, and dietary regimens, with the overarching objective of minimizing injury risk. Ethical implementation of these findings is essential to ensure that no biases are introduced into player selection for the upcoming season. Decisions regarding player selection should be grounded in a comprehensive evaluation of overall performance and capabilities, rather than being influenced solely by the outcomes of this study.

10 Questions for the audience

1. How does the predictive model account for sudden, unexpected injuries that are not related to measurable risk factors?
2. Why was Logistic Regression prioritized over other models, even though Random Forest had slightly higher overall accuracy?
3. How can the model's predictions be integrated into the coaching staff's day-to-day decision-making without overreliance?
4. Are there plans to expand the dataset to include players from different universities, regions, or age groups for better generalizability?
5. How are self-reported factors, such as sleep and stress, validated to ensure the data is reliable?
6. Could the model inadvertently introduce bias against certain player positions or demographics? How is fairness ensured?
7. How frequently should the predictive model be retrained with new data to maintain accuracy?
8. Are there specific interventions recommended for athletes identified as high-risk, and how are these tailored individually?
9. How do ethical considerations influence the use of predictive modeling in player selection and participation decisions?
10. Can the methods used in this project be applied to other sports or injury types beyond university football?

References:

1. National Institute of Arthritis and Musculoskeletal and Skin Diseases “Sports Injuries,” accessed September 14, 2025, https://www.niams.nih.gov/health-topics/sports-injuries
2. Kaggle. (2025). “University Football Injury Prediction Dataset” [Dataset]. Kaggle. https://www.kaggle.com/datasets/yuanchunhong/university-football-injury-prediction-dataset?resource=download