Predicting Injury Outcomes in Soccer Using Statistical and Machine Learning Techniques

Danielle Naggar

Pace University, Seidenberg School of CSIS

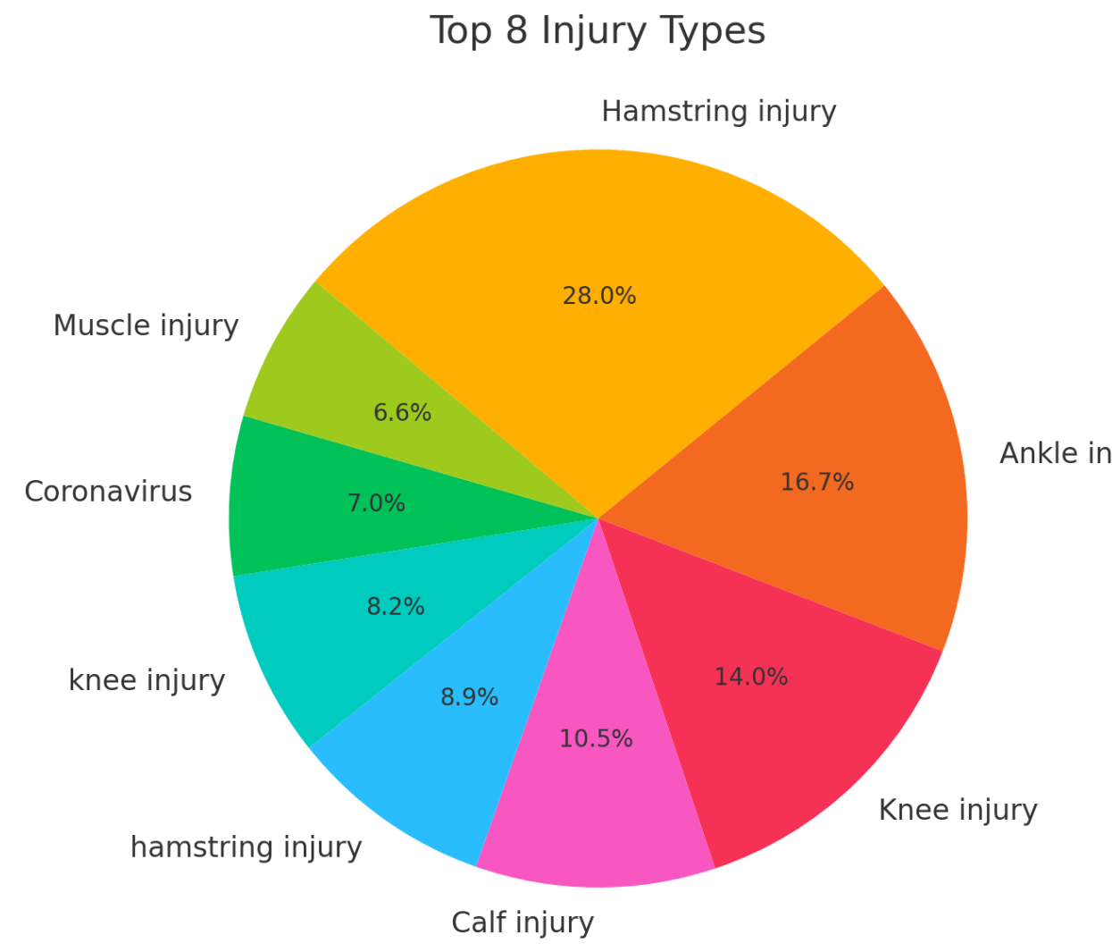
Abstract

This project investigates the relationship between professional football players' characteristics and their injury outcomes, with a focus on injury severity and type. Using a dataset of 656 recorded injury cases, we examined how traits such as age, position, and FIFA rating correlate with both the nature and seriousness of injuries. We then applied machine learning models—including Random Forest, Gradient Boosting, and Logistic Regression—to assess the predictability of injury outcomes from player features. Model performance was evaluated using accuracy, F1 score, and multi-class ROC curves. Although model accuracy varied by injury category, the study demonstrates that basic player profile data can offer meaningful insights into injury patterns and potential predictive value.

Research Questions

1. Which player traits (position, age, FIFA rating) are most associated with different injury types or severities?

2. Can ML models predict injury severity or injury type based on player characteristics?



Dataset

Source: Player injury records (CSV format)

Total Entries: 656 injured player cases

Key Features:

Player traits: Age, Position, FIFA Rating

Injury details: Injury Type, Date of Injury, Date of Return

Target Variables:

Injury Severity – categorized as Low, Moderate, or Severe based on days missed

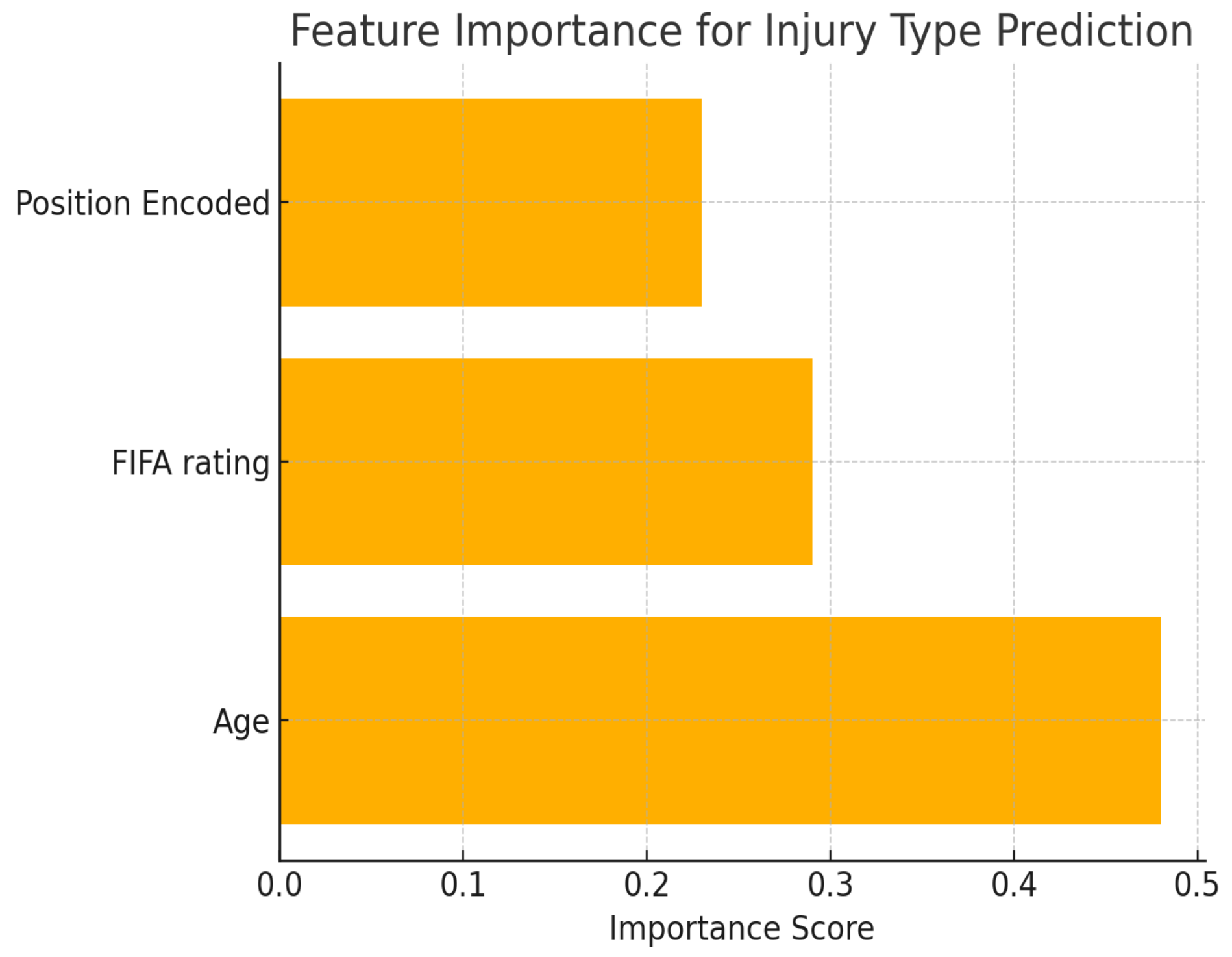
Injury Type – specific types such as hamstring strain, ACL tear, etc.

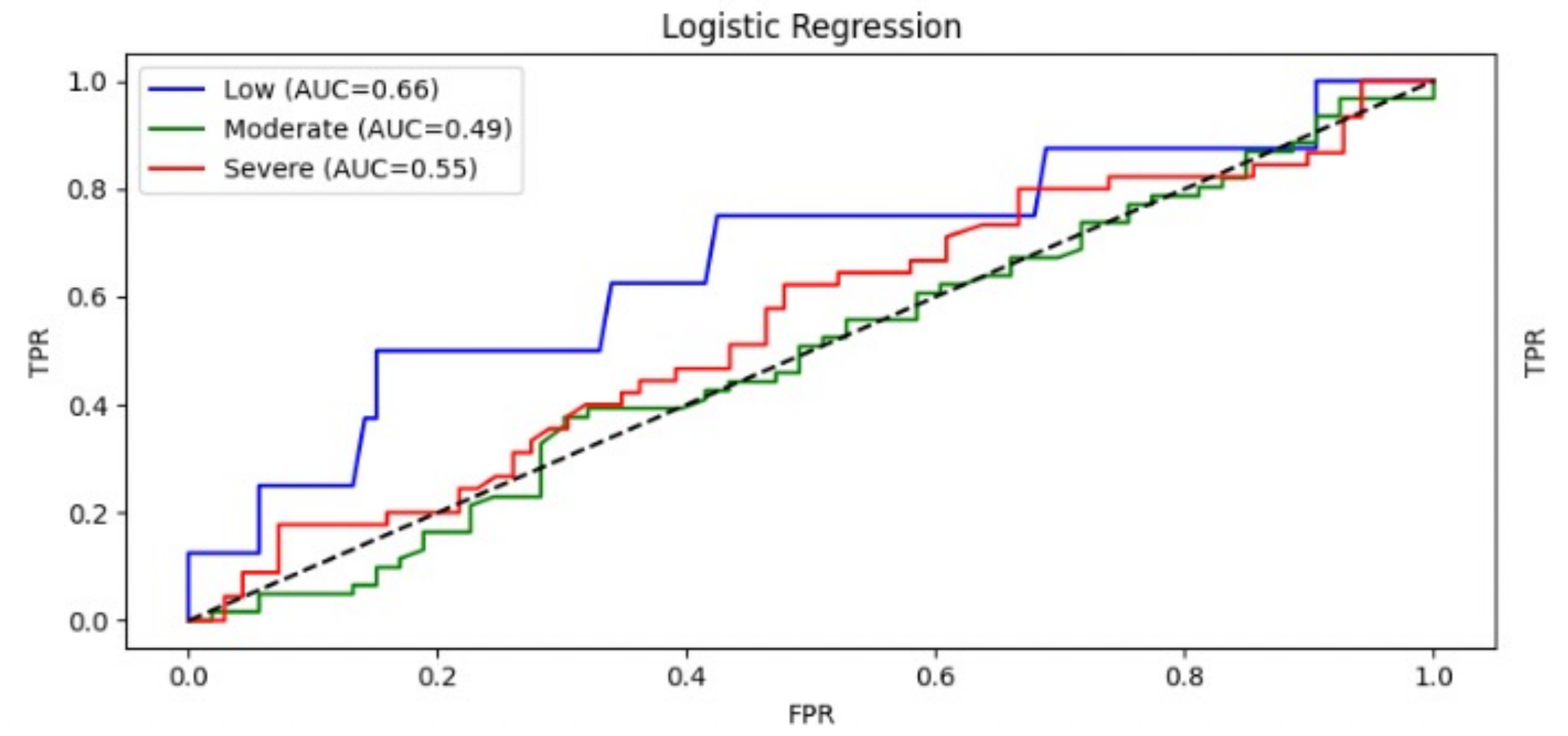
Methodology

The project began with data cleaning, where missing or inconsistent values were removed, and injury severity was calculated based on the number of days missed between the recorded injury and return dates. Severity was then categorized as low, moderate, or severe to allow for clearer comparison across cases. Categorical features, such as position and injury type, were encoded numerically to ensure compatibility with machine learning algorithms. Statistical methods were used to understand baseline relationships: ANOVA tested for differences in age and FIFA rating across injury outcomes, while chi-square tests examined any association between player position and injury category. These tests were selected because they are well-established for comparing groups and detecting associations in categorical and numerical data. For prediction tasks, five classification models were used—Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machines, and K-Nearest Neighbors. This range of models was chosen to compare linear and non-linear methods, and to evaluate performance under different assumptions. Model accuracy and F1 scores were used to assess classification success, with ROC curves providing additional insight into how well each model distinguished between classes.

Results

The analysis showed that age was significantly related to both injury severity and injury type. FIFA rating showed a weaker association, while player position did not appear to influence injury type in a meaningful way. When applying machine learning models, Logistic Regression performed best in predicting injury severity, particularly in identifying short-term injuries. For injury type, the Gradient Boosting model achieved the highest overall performance. ROC curve analysis supported these findings, with the strongest predictive ability seen in more frequent and less severe injury cases. These results suggest that even basic player data can help identify trends in injury outcomes, though prediction remains more difficult for less common injuries.





Conclusion & Future Work

This study explored how player characteristics relate to football injury outcomes and whether those traits can be used to predict injury severity and type. Age consistently emerged as a meaningful factor in both statistical analysis and machine learning models, while position and FIFA rating showed weaker associations. The models used were most effective at identifying lower-severity injuries and more common injury types. While the results are promising, they are limited by the scope of the dataset and the range of available features. In future work, expanding the dataset to include uninjured players and additional variables—such as training intensity, workload, or previous injuries—could improve the accuracy and practical use of injury prediction models.

Limitations

This study is limited by the scope of the dataset, which only includes players who were injured, preventing comparisons with uninjured individuals. Additionally, the injury type distribution is imbalanced, with some categories underrepresented, which may have affected model performance. The available features—age, position, and FIFA rating—provide only a partial view of injury risk, excluding critical factors like training load, match intensity, or medical history.

Sources

- Ayala, R. E. D., Granados, D. P., Gutiérrez, C. A. G., Ruíz, M. A. O., Espinosa, N. R., & Heredia, E. C. (2024). Novel Study for the Early Identification of Injury Risks in Athletes Using Machine Learning Techniques. *Applied Sciences*, *14*(2), 570. https://doi.org/10.3390/app14020570
- Amendolara A, Pfister D, Settelmayer M, et al. (September 28, 2023) An Overview of Machine Learning Applications in Sports Injury Prediction. Cureus 15(9): e46170. doi:10.7759/cureus.46170
- https://www.kaggle.com/datasets/amritbiswas007/player-injuries-and-team-performance-dataset : Data collected in the Premier League from seven clubs (Tottehnham, Ashton Villa Brighton, Arsenal, Brentford, Everton, Burnley and Manchestor United.) Date: Collected from 2019 to 2023| Geography: English Premier Leagues | Size: 600 records