

# **Injury Risk Prediction in University Football Athletes: A Machine Learning Approach Based on Physical, Behavioral, and Historical Indicators**

**Author:** Fernanda Schenker Pieri

# Abstract

Injury prevention is a central challenge in competitive sports, where physical overload, inadequate recovery, and individual biomechanical factors contribute to elevated risk. This study presents a data-driven framework for predicting next-season injury probability in university football athletes. Using a dataset containing physical metrics, neuromuscular indicators, workload information, behavioral habits, and injury history, three machine learning models were evaluated. Feature engineering was performed to integrate multi-dimensional aspects of athletic readiness. Results indicate that Logistic Regression achieved the highest recall (0.97), while Random Forest offered superior interpretability through nonlinear importance and SHAP-based analysis. The findings highlight the combined impact of readiness, stress, sleep quality, and prior injuries as major contributors to injury risk. The study provides a reproducible pipeline for risk assessment and decision support in sports performance environments.

# 1. Introduction

Injury prediction has become a strategic priority in modern football programs due to its implications for athlete health, team performance, and long-term development. Traditional methods rely heavily on subjective evaluation or isolated metrics, which fail to capture the complex interactions between physical conditioning, behavioral factors, and cumulative load.

Machine Learning (ML) offers the ability to integrate multidimensional data into predictive risk models, providing actionable information for performance staff and enabling proactive interventions.

**Objective:**

To build, evaluate, and interpret a predictive model capable of estimating the probability that an athlete will suffer an injury in the following season.

## 2. Materials and Methods

### 2.1 Dataset

Dataset:

<https://www.kaggle.com/datasets/yuanchunhong/university-football-injury-prediction-dataset>

The dataset contains 800 football athletes and 19 variables across domains:

- **Physical:** Age, Height, Weight, BMI
- **Neuromuscular:** Knee strength, flexibility, balance, sprint speed, agility, reaction time
- **Workload:** Training hours, matches played
- **Behavioral:** Sleep, nutrition, warmup adherence, stress
- **History:** Previous injuries

No missing values were detected.

### 2.2 Preprocessing

A unified preprocessing pipeline ensured consistency and reproducibility:

- Standardization (StandardScaler)
- One-Hot Encoding (Position)
- Median imputation for numeric values
- Most-frequent imputation for categorical values

All steps were implemented using `sklearn.ColumnTransformer`.

### 2.3 Feature Engineering

Composite variables included:

- **Readiness\_Strength**
  - Strength + Flexibility + Balance → neuromuscular readiness indicator.
- **Workload\_Index**
  - Weighted combination of training hours and match exposure.
- **Prep\_Score**
  - Integrates sleep quality, nutrition, and warmup adherence.
- **Injury\_History\_Weight**
  - Captures accumulated injury susceptibility using previous injuries + stress.

These features greatly improved interpretability and predictive signal.

## 2.4 Modeling and Evaluation

Models trained:

- Logistic Regression
- Random Forest
- Gradient Boosting

Train-test split: 75/25 (stratified)

Primary metric: **Recall** (to minimize false negatives)

Additional metrics: Accuracy, Precision, F1-score, ROC AUC.

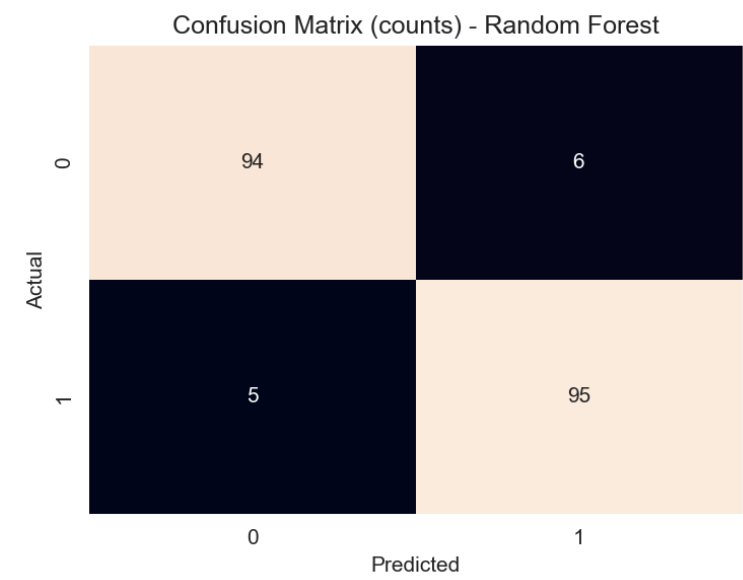
Random Forest was selected for final interpretability analysis.

# 3. Results

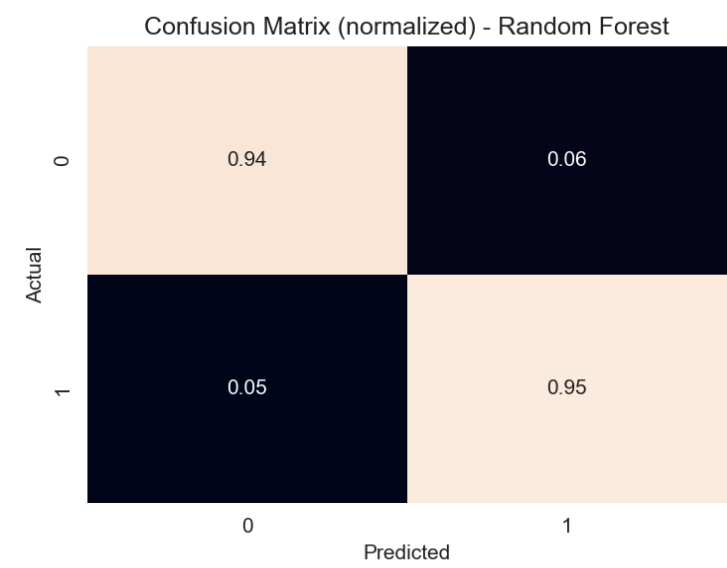
## 3.1 Model Performance

| Model               | Recall | F1-score | ROC AUC |
|---------------------|--------|----------|---------|
| Logistic Regression | 0.97   | 0.965    | 0.995   |
| Random Forest       | 0.95   | 0.945    | 0.991   |
| Gradient Boosting   | 0.92   | 0.915    | 0.983   |

### Confusion Matrix (Counts)



### Confusion Matrix (Normalized)

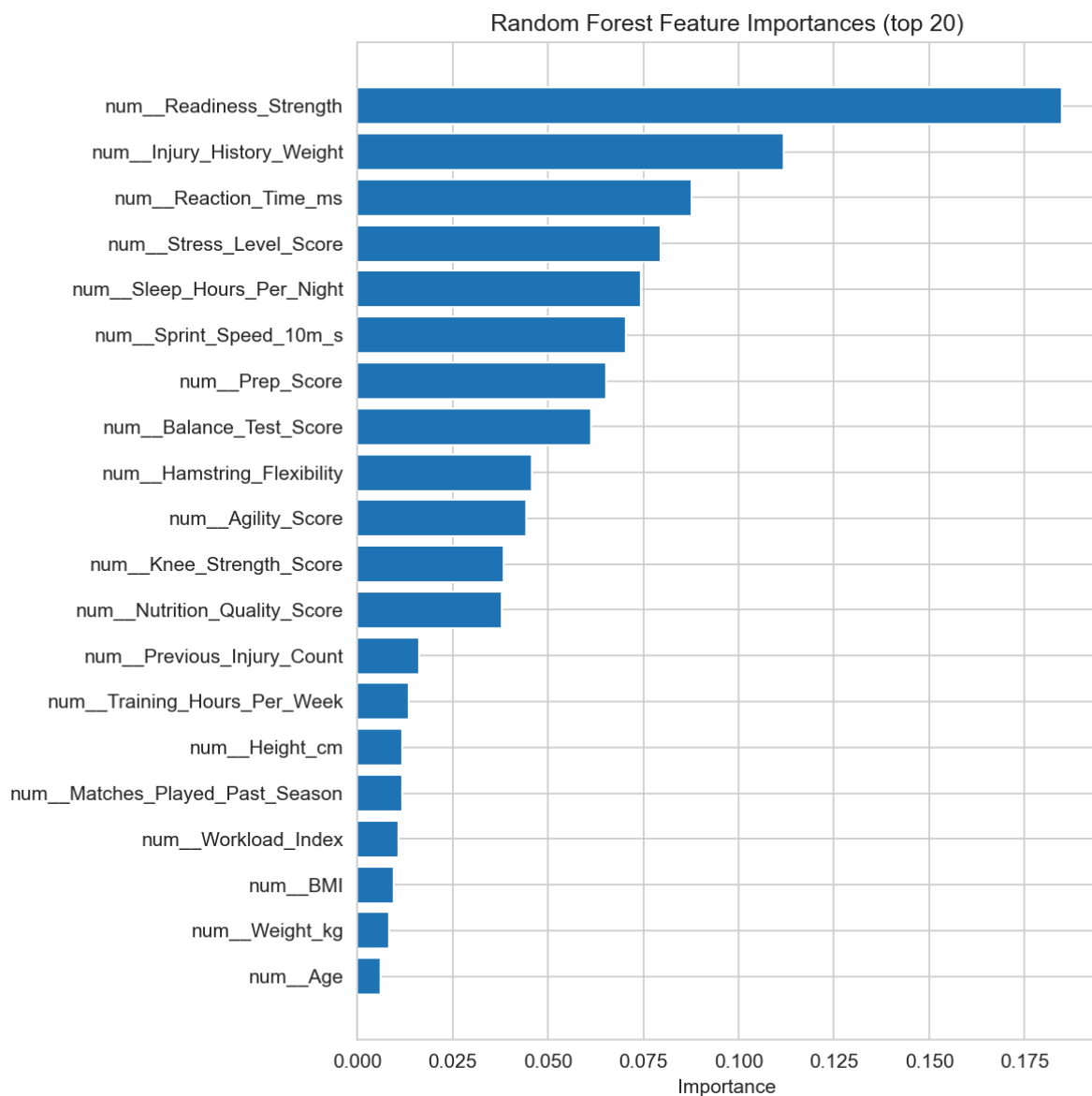


## 3.2 Feature Importance

Random Forest and SHAP identified consistent top predictors:

- Readiness\_Strength
- Injury\_History\_Weight
- Reaction\_Time\_ms
- Stress\_Level\_Score
- Sleep\_Hours\_Per\_Night

### Feature Importance

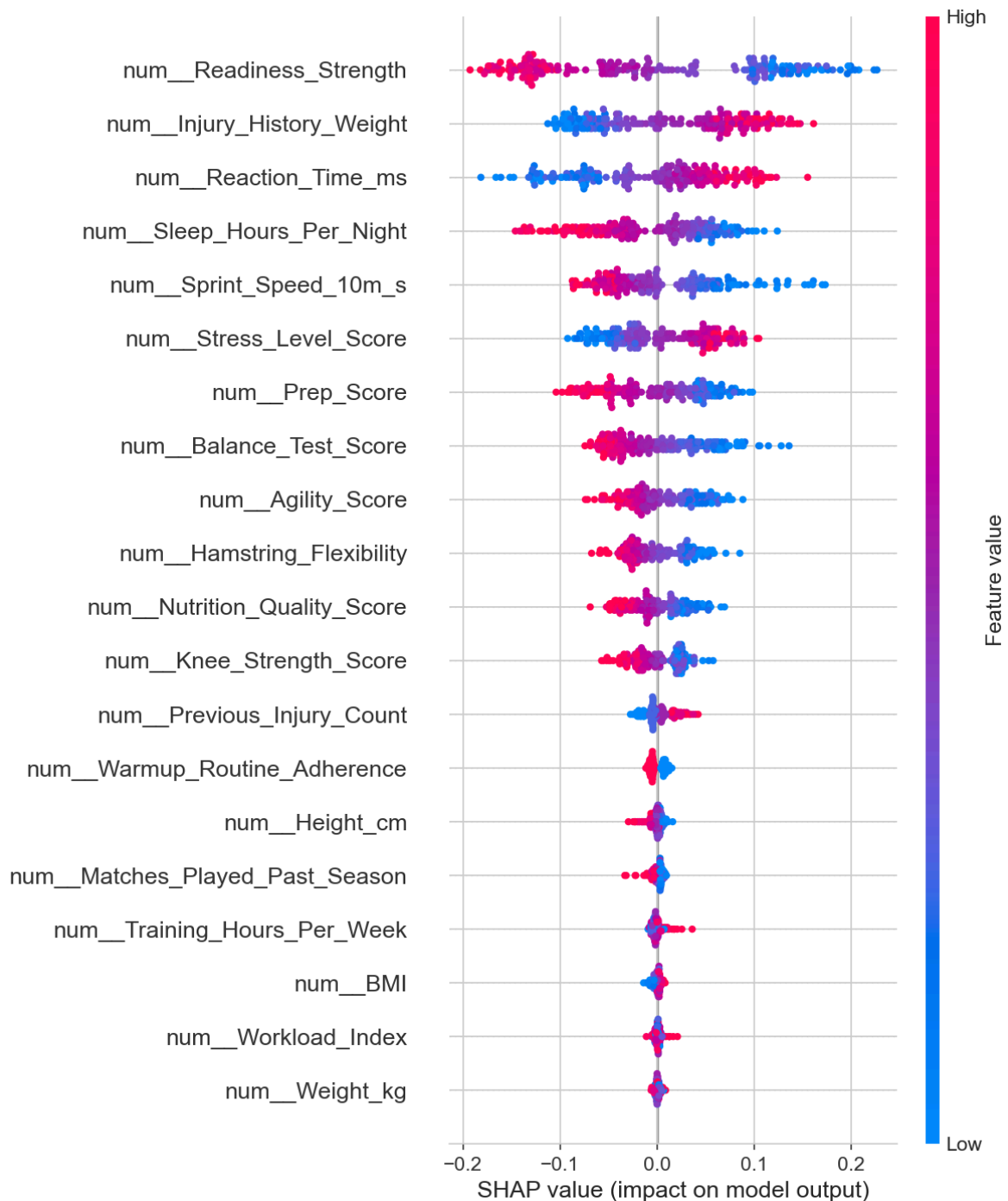


## 3.3 SHAP Explanations

Global insights:

Athletes with low readiness or high stress showed consistently elevated SHAP values. Behavioral consistency and recovery (sleep, prep routine) shifted risk downward.

### SHAP Summary Plot

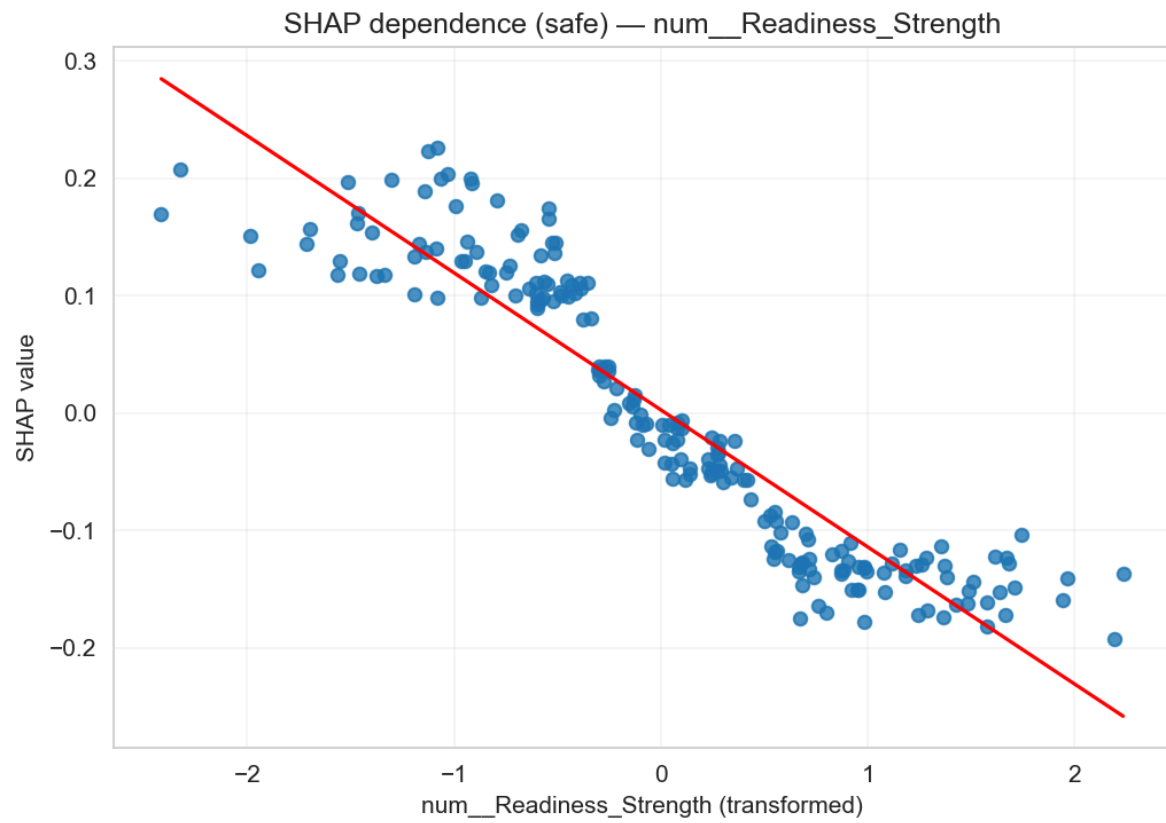


### Local example:

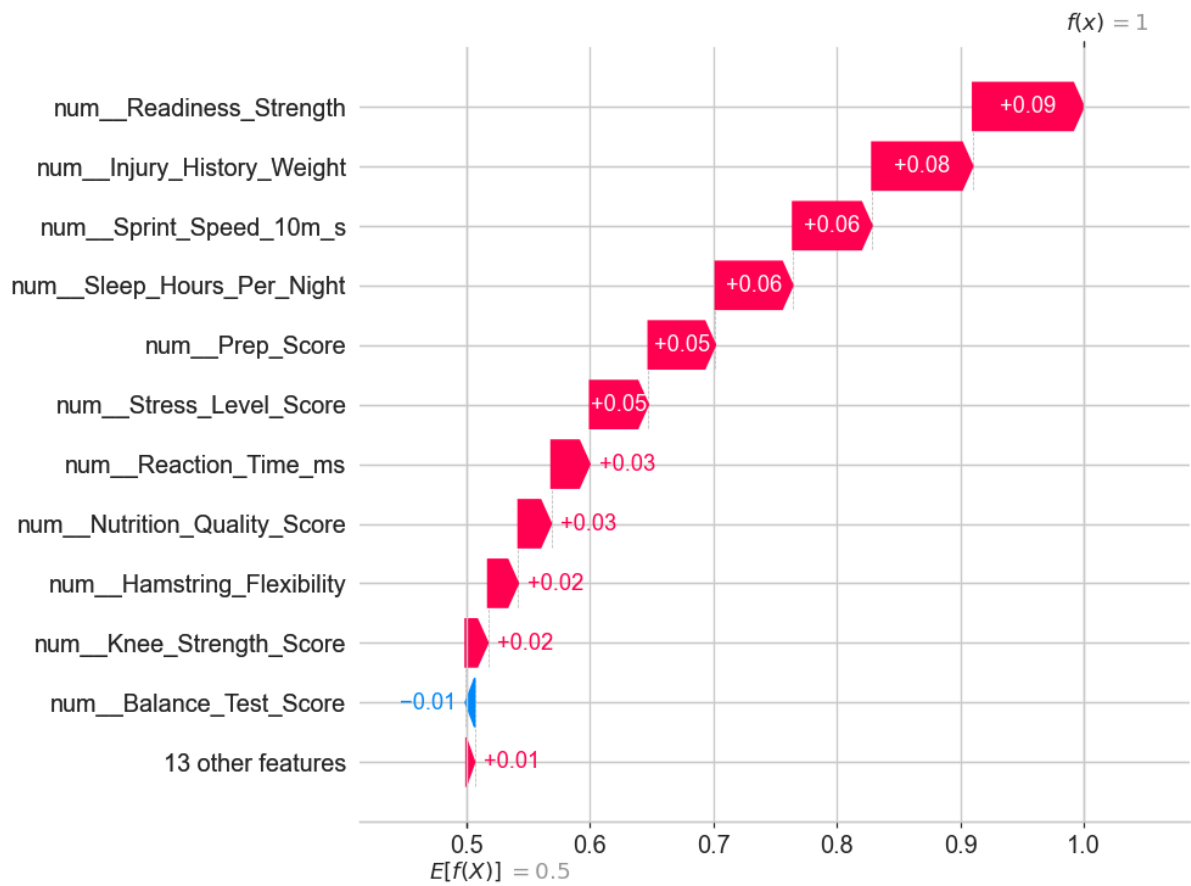
The highest-risk athlete showed large positive SHAP contributions from:

- high stress
- low readiness
- previous injuries
- suboptimal prep score

### SHAP Dependence Plot



## SHAP Waterfall Plot



## 4. Discussion

The results highlight the multifactorial nature of injury risk in football athletes. Rather than relying on isolated biomarkers, the model integrates physical capacities, behavioral habits, and cumulative stress exposure — mirroring the multidimensional frameworks commonly used in sports performance science.

A key insight is the prominence of **Readiness\_Strength** as the strongest predictor. This engineered feature combines knee strength, flexibility, and balance — three pillars often associated with neuromuscular control and joint stability. Athletes presenting low readiness values consistently exhibited positive SHAP contributions, suggesting greater vulnerability to overload or biomechanical compensation.

Similarly, the influence of **Injury\_History\_Weight** reinforces findings from sports medicine literature: prior injuries, compounded by elevated stress levels, enhance susceptibility to subsequent injuries. The SHAP summary plot shows a clear monotonic relationship: as historical vulnerability increases, so does the predicted injury risk.

### **Behavioral factors also emerged as meaningful contributors.**

Sleep\_Hours\_Per\_Night and Prep\_Score (integrating sleep, nutrition, and warmup adherence) displayed protective effects. The dependence plots demonstrate that even moderate improvements in nightly sleep or adherence to warmup routines can meaningfully shift risk downward. These findings align with well-established principles in athlete monitoring, where recovery quality and daily consistency play active roles in sustaining physical resilience.

Finally, **Reaction\_Time\_ms** and **Stress\_Level\_Score** bridge physiological and psychological dimensions. Slower reaction times and elevated stress — both commonly associated with fatigue or reduced cognitive sharpness — contribute strongly to injury likelihood in the model. The combination of mechanical (strength), cognitive (reaction time), and psychosocial (stress) indicators offers a holistic understanding of athlete readiness.

The confusion matrices show that the model is particularly effective at correctly identifying injured athletes (high recall), which is crucial in applied contexts — false negatives could mean missing a high-risk athlete who needs support or tailored recovery.

## 5. Conclusion

This study demonstrates that machine learning can meaningfully support injury risk monitoring in football environments when fed with well-structured, multi-domain athlete data. Models combining physical, behavioral, neuromuscular, and historical features successfully captured nonlinear patterns that are difficult to detect through manual analysis alone.

The strongest insights emerged from engineered features, especially those summarizing **global readiness** and **historical vulnerability**. SHAP-based interpretability further revealed that athlete risk profiles arise from the interaction of daily behaviors (sleep, prep score), physiological readiness, and accumulated load — not from isolated metrics.

From a practical standpoint, the model is well-suited to serve as a **decision-support tool** for coaching, strength & conditioning, and sports medicine teams. Rather than dictating interventions, the system can highlight individuals whose profile warrants closer monitoring, recovery prioritization, or modification in training load.

While the dataset is synthetic and lacks external validation, the methodological pipeline is robust and transferable to real-world athlete monitoring systems. Future implementations could integrate GPS-based load tracking, longitudinal analysis, and dynamically updated risk probabilities to create a fully adaptive monitoring ecosystem.

Overall, the results reinforce an important principle: **injury risk prediction is most effective when technical data science methods are combined with sports science knowledge and contextual understanding**. The integration of AI-driven interpretability with domain expertise provides a powerful foundation for informed decision-making in athlete health and performance.