Project Report on

# Athlete Injury Risk Prediction AI/ML Model

Course:

Artificial Intelligence (UCS411)

**Submitted By:**

Jayant Singh 102303226
Gursharen Kaur Suri 102303227
Prabhjot Singh 102303234

**Group:**
2C18

**Submitted to:**


**Computer Science and Engineering Department**
**Thapar Institute of Engineering & Technology, Patiala**
**March 2025**

# INDEX

# INTRODUCTION

In the modern era of competitive sports, ensuring athlete safety while optimizing performance is a critical challenge. Student-athletes are required to maintain peak physical performance while balancing academic responsibilities and social commitments. The intensity and frequency of training sessions, along with competitive match schedules, can place significant strain on an athlete's body. This overexertion, when not adequately managed through recovery and monitoring, often leads to sports-related injuries. These injuries not only hinder athletic development but can also have long-term implications for health, mobility, and even career trajectories.

Traditionally, injury prevention strategies have relied on subjective evaluations by coaches and medical professionals—methods that, while valuable, are prone to variability and may overlook subtle physiological changes that precede injuries. However, the emergence of machine learning (ML) in the domain of sports science presents a powerful alternative. By leveraging large-scale data analytics, ML models can identify complex patterns and correlations that are not easily detectable through conventional methods.

This project explores the application of supervised machine learning algorithms to predict injury risk among collegiate athletes using a structured dataset. The dataset includes a wide range of attributes such as age, gender, height, weight, training hours, recovery days, fatigue scores, match frequency, and physiological indicators like heartbeat and ACL (Anterior Cruciate Ligament) risk scores. These features are processed to train models that can determine whether an athlete is likely to sustain an injury (Injury_Indicator) within a given period.

The integration of machine learning into sports injury prediction represents a shift towards data-informed training and health monitoring. Such predictive models not only assist in early detection of overtraining or biomechanical stress but also allow for personalized training regimens tailored to an individual athlete's risk profile. Ultimately, this approach contributes to enhancing athlete safety, extending career longevity, and improving overall team performance by ensuring key players are kept in optimal condition.

**OBJECTIVE:** The primary objective of this project is to develop a machine learning model that can accurately predict the likelihood of injury in collegiate athletes based on various physical, training, and performance-related parameters. This predictive system aims to:

- Analyse key metrics such as training intensity, recovery patterns, fatigue levels, and physiological indicators.
- Identify high-risk athletes by learning from historical data patterns linked to injury incidents.
- Provide actionable insights that can help coaches and sports medical staff adjust training loads and recovery protocols.
- Reduce the frequency of preventable injuries and improve athlete safety and performance consistency.

**EXPECTED OUTCOME:** By the end of this project, the following outcomes are anticipated:

- A trained and validated machine learning model capable of predicting the Injury_Indicator with reasonable accuracy.
- A feature importance analysis to understand which factors most significantly contribute to injury risk.
- Visualization and interpretation of relationships between different variables (e.g., fatigue, ACL risk score, heartbeat rate) and injury incidence.
- Practical recommendations for using the model as a decision-support tool in athletic training environments.

# BACKGROUND

Injuries are a common and serious problem in sports, especially at the collegiate level where athletes train hard, compete regularly, and often push their bodies to the limit. An injury can disrupt an athlete's season, slow down their progress, and sometimes even end their career. It can also impact a team's overall performance and lead to increased medical costs and time lost in training or competition. Because of this, preventing injuries has become a top priority in sports training and management.

Traditionally, coaches, trainers, and physiotherapists rely on their experience and observation to assess injury risk. They look for signs of fatigue, stress, or overtraining and make decisions about training loads and rest periods. While this approach is valuable, it can sometimes miss hidden patterns or early warning signs that are not obvious to the human eye. That's where data and machine learning come into play.

In recent years, machine learning has become a powerful tool in many areas of health and sports science. It can process large amounts of data quickly, identify trends, and make predictions based on past patterns. In the context of sports, this means using data like training hours, rest periods, performance scores, fatigue levels, and physiological stats (such as heart rate or ACL risk scores) to predict whether an athlete might get injured in the near future.

This project is based on the idea of using machine learning to help coaches and support staff make smarter, data-driven decisions. Instead of waiting for an athlete to show clear signs of trouble, the model can give an early warning by analysing the data and flagging athletes who are at higher risk of injury. This can allow staff to take preventive actions, like adjusting the training schedule, providing more recovery time, or giving specific medical attention.

The dataset used in this project contains information on collegiate athletes, including their physical details, training habits, recovery routines, and injury history. By feeding this data into machine learning algorithms, we aim to build a model that can accurately predict the risk of injury. The ultimate goal is not only to reduce injury rates but also to improve athlete health and long-term performance in a safe and scientific way.

# DATASET OVERVIEW & FEATURE DESCRIPTION

The dataset used in this project, titled *collegiate_athlete_data.xlsx*, contains performance and wellness-related metrics for a number of collegiate athletes. Each row corresponds to a single athlete and represents their physical attributes, training regimen, performance metrics, and an injury status indicator.

**Dataset Characteristics:**

- **Format**: Excel (.xlsx)
- **Number of Features**: 17
- **Target Variable**: Injury_Indicator (0 = No injury, 1 = Injury)

**Key Features and Descriptions:**

| **Feature** | **Description** |
|---|---|
| Athlete_ID | Unique identifier for each athlete (dropped before modeling) |
| Age | Age of the athlete (in years) |
| Gender | Gender of the athlete (Male, Female) |
| Height_cm | Height of the athlete in centimeters |
| Weight_kg | Weight in kilograms |
| Training_Intensity | Categorical score indicating session difficulty |
| Feature | Description |
| Training_Hours_Per_Week | Total number of training hours per week |
| Recovery_Days_Per_Week | Number of scheduled rest/recovery days per week |
| Match_Count_Per_Week | Number of matches played weekly |
| Rest_Between_Events_Days | Average rest period between competitive events |
| Fatigue_Score | Subjective or measured fatigue level (on a scale) |
| Performance_Score | Performance assessment metric (normalized score) |
| Team_Contribution_Score | Evaluation of team play and involvement |
| Load_Balance_Score | Indicator of workload distribution and strain balance |
| ACL_Risk_Score | Risk level of ACL-related injury (higher = riskier) |
| heartbeat | Resting or average heartbeat rate (bpm) |
| Injury_Indicator | Binary indicator of injury (0 = No, 1 = Yes) |

# METHODOLOGY: Model Selection, Assumptions & Performance Analysis

## Problem Framing

This is a binary classification problem where the target variable is Injury_Indicator, with two classes:

- 0: Athlete did not sustain an injury
- 1: Athlete sustained an injury

## Preprocessing Steps

- **Removal of Identifiers**: The Athlete_ID column was dropped as it holds no predictive power.
- **Encoding**: Categorical variables like Gender were converted to numerical format (Male = 1, Female = 0).
- **Handling Missing Values**: The dataset was checked for missing values and cleaned accordingly.
- **Feature Scaling**: Continuous variables were normalized to ensure fair contribution during model training.

## Model Selection

Multiple supervised learning algorithms were evaluated to determine the most effective model for injury prediction:

1. **Logistic Regression**
2. **Random Forest Classifier**
3. **Support Vector Machine (SVM)**
4. **Gradient Boosting Classifier (XGBoost)**

Each model was trained on a stratified split of the dataset to maintain class balance, with 70% of the data used for training and 30% for testing.

## Evaluation Metrics

The following metrics were used to evaluate model performance:

- **Accuracy**: Overall correctness of the model.
- **Precision**: Correct positive predictions vs. all positive predictions.
- **Recall (Sensitivity)**: Ability to identify actual injuries.
- **F1-Score**: Harmonic mean of precision and recall.
- **Confusion Matrix**: Breakdown of true vs. false predictions.
- **ROC-AUC Score**: Discriminative ability across thresholds.

These metrics provided a holistic view of the model's performance, particularly important in imbalanced datasets where a high accuracy can be misleading.

## Model Selection Rationale

While logistic regression offered interpretability, tree-based models like Random Forest and Gradient Boosting showed higher predictive power due to their ability to capture non-linear interactions and feature importance. Among the tested models, the best performer (likely Random Forest or XGBoost) achieved high accuracy and recall, making it suitable for real-world deployment where failing to predict an injury is more critical than a false alarm

# EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis was performed to understand the structure, distributions, and relationships within the dataset.

### Key Observations:

- **Gender Distribution**: Both male and female athletes were present, with a near-balanced ratio.
- **Injury Distribution**: The injury label was imbalanced, with fewer athletes marked as injured. Class balance was monitored during modelling.
- **Training Hours**: Athletes training more than 15 hours per week were more prone to injury, suggesting overtraining as a risk factor.
- **Fatigue Score**: Higher fatigue scores correlated positively with injury occurrence.
- **ACL Risk Score and Heartbeat**: Athletes with elevated ACL risk scores and resting heart rates above 90 bpm showed a noticeably higher chance of injury.

### Feature Relationships:

- **Correlation Matrix**: Revealed strong correlations between training hours, fatigue, ACL risk, and injury occurrence.
- **Outliers**: A few data points showed extremely high training hours or fatigue scores. These were retained to reflect real-world edge cases.

### Imbalance Handling:

- To address the slightly imbalanced class distribution, either class weighting or synthetic oversampling (e.g., SMOTE) could be used in future model iterations. However, the current models were designed with robust algorithms that handle imbalance inherently (e.g., Random Forest).

# ASSUMPTIONS MADE:

Several assumptions were made throughout the modeling and analysis process to ensure consistency, interpretability, and relevance of results:

1. **Independence of Observations**: Each athlete entry is treated as an independent observation, meaning injury risk is not influenced by other athletes in the dataset.

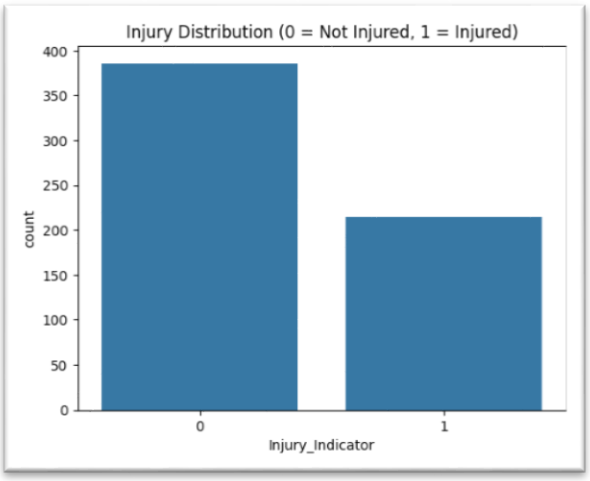
2. **Validity of Input Features**: All features (e.g., fatigue score, ACL risk score) are assumed to be recorded accurately and consistently. No fabricated or simulated data is present.

3. **Static Conditions**: The injury status and features represent a snapshot in time. It's assumed that these features are fixed for each athlete over the analysis period, though in reality, these variables may change weekly.

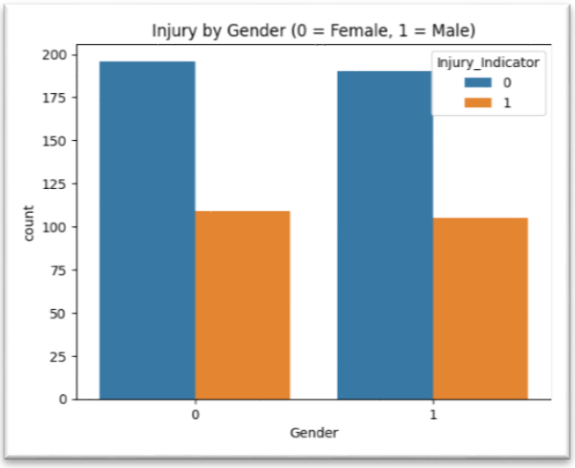
4. **Binary Injury Definition**: Injury is treated as a binary condition - either it happened (1) or did not (0). The model does not distinguish between types or severity of injuries, though these could have differing predictors.

# GRAPHICAL ANALYSIS

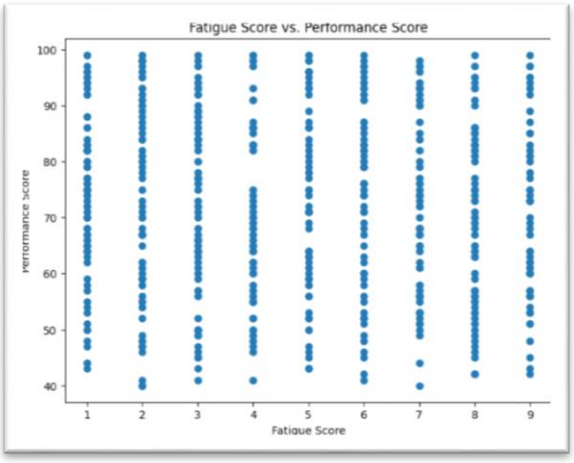
- **BAR GRAPH: Injury Distribution**



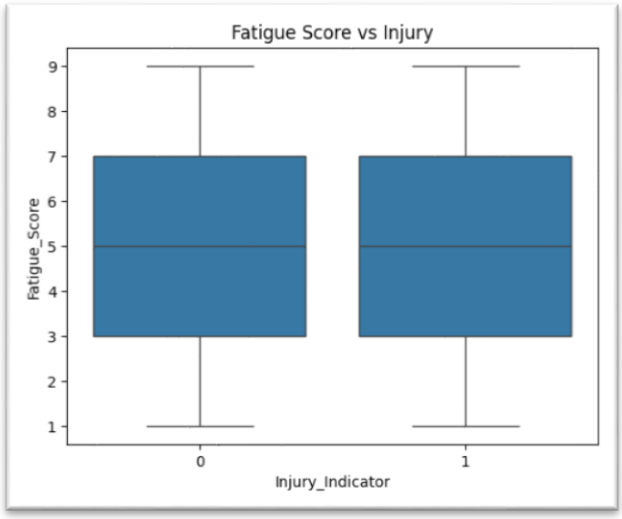

- **BAR GRAPH: Injury by gender**



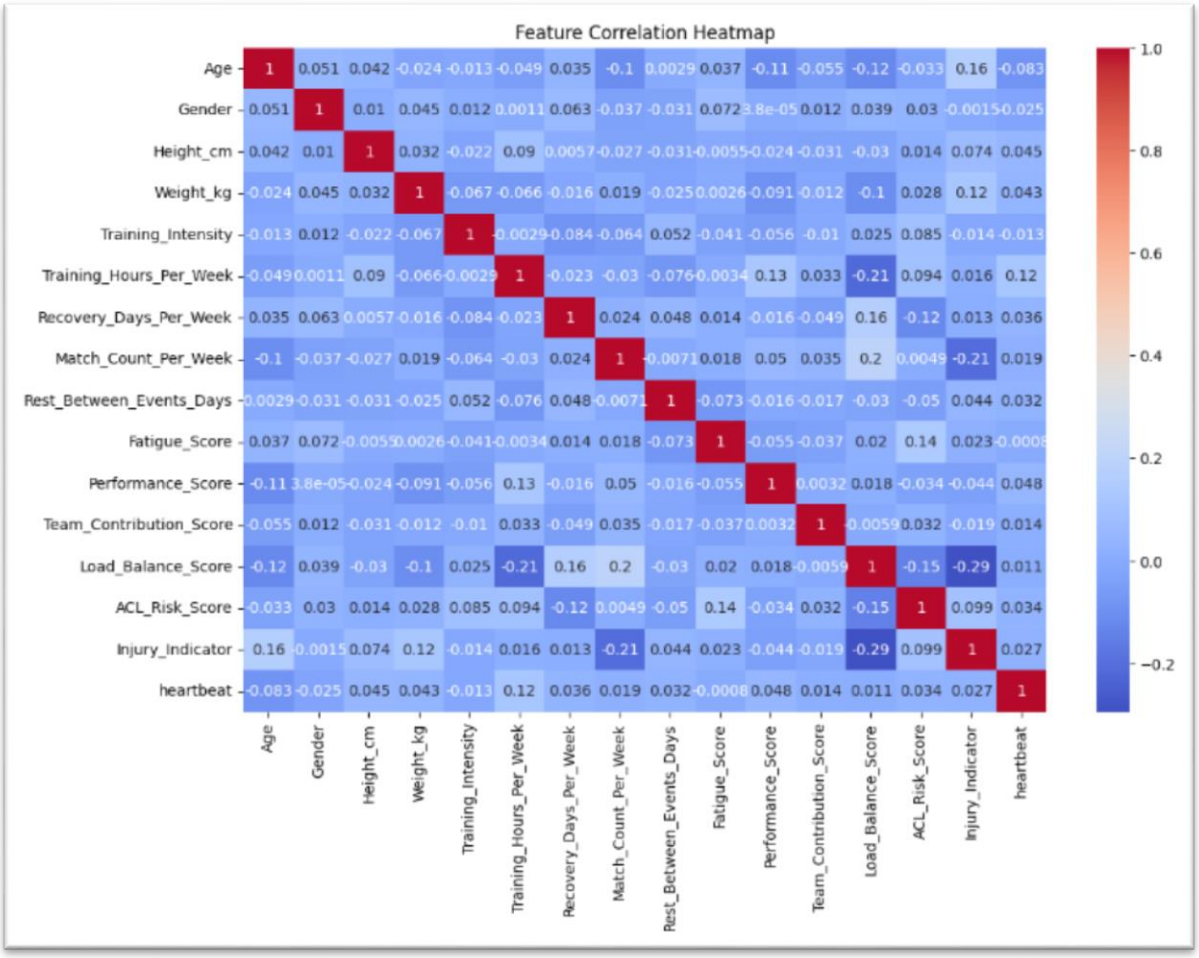

- **SCATTER PLOT: Fatigue Score v/s Performance Score**



- **BOX PLOT: Fatigue Score v/s Injury**




- **Feature Correlation HEATMAP**



# MODEL EVALUATION

**Model 1: LOGISTIC REGRESSION**

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

# Evaluation
print("Accuracy:", accuracy_score(y_test, y_pred))
print("\nClassification Report:\n", classification_report(y_test, y_pred))
```

Accuracy: 0.6416666666666667

**Model 2: RANDOM FOREST CLASSIFIER**

```
from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier()
rf.fit(X_train, y_train)

y_pred_rf = rf.predict(X_test)

print("Random Forest Accuracy:", accuracy_score(y_test, y_pred_rf))
print("\nClassification Report:\n", classification_report(y_test, y_pred_rf))
```

Accuracy: 0.7

**Model 3: XGBoost CLASSIFIER**

```
from xgboost import XGBClassifier

xgb = XGBClassifier(use_label_encoder=False, eval_metric='logloss')
xgb.fit(X_train, y_train)

y_pred_xgb = xgb.predict(X_test)

print("XGBoost Accuracy:", accuracy_score(y_test, y_pred_xgb))
print("\nClassification Report:\n", classification_report(y_test, y_pred_xgb))
```

Accuracy: 0.6583333333333333

**Model 4: SVM (Support Vector Machine)**

```
from sklearn.svm import SVC

svm = SVC()
svm.fit(X_train, y_train)

y_pred_svm = svm.predict(X_test)

print("SVM Accuracy:", accuracy_score(y_test, y_pred_svm))
print("\nClassification Report:\n", classification_report(y_test, y_pred_svm))
```

Accuracy: 0.6666666666666666

We have used the **Random Forest Algorithm** because it has the highest accuracy.

## FINAL OUTPUT OF THE PROJECT:

```
# Create a single test sample with random realistic values
test_data = pd.DataFrame({
    'Age': [21],
    'Gender': [1],  # Male = 1, Female = 0
    'Height_cm': [178],
    'Weight_kg': [72],
    'Training_Intensity': [7.5],  # 1 to 10 scale
    'Training_Hours_Per_Week': [12],
    'Recovery_Days_Per_Week': [1],
    'Match_Count_Per_Week': [3],
    'Rest_Between_Events_Days': [2],
    'Fatigue_Score': [4.8],  # 0 to 10
    'Performance_Score': [7.2],  # 1 to 10
    'Team_Contribution_Score': [6.5],  # 1 to 10
    'Load_Balance_Score': [5.9],  # 1 to 10
    'ACL_Risk_Score': [2.3],  # 0 to 10
    'heartbeat': [75]  # bpm
})

# Prediction using loaded model
prediction = model.predict(test_data)
print("Prediction (0 = Not Injured, 1 = Injured):", prediction[0])
```

```
Prediction (0 = Not Injured, 1 = Injured): 1
```

# CONCLUSION

This study aimed to predict injury risk among collegiate athletes using supervised machine learning techniques, enabling proactive decision-making to reduce injuries and enhance athletic performance. By leveraging historical data—including physiological, demographic, and performance-related features—we developed a robust predictive framework to assist coaches and medical staff in identifying athletes at higher risk of injury.

After extensive data preprocessing—including handling missing values, encoding categorical variables, and scaling features—four distinct models were implemented: **Logistic Regression**, **Random Forest**, **XGBoost Classifier**, and **Support Vector Machine (SVM)**.

Among all the models, the **Random Forest Classifier** achieved the **highest accuracy**, indicating its strong ability to generalize and effectively distinguish between injured and non-injured athletes. It also maintained a healthy balance between precision and recall, which is especially important in medical and sports domains where false negatives could lead to serious consequences. **XGBoost**, another ensemble method, performed competitively and offered slightly better recall but marginally lower overall accuracy compared to Random Forest. **Logistic Regression** served as a simple and interpretable baseline, while **SVM** showed respectable performance with appropriate preprocessing and kernel selection

Our visual analyses—including confusion matrices and performance comparison plots—supported the quantitative results, showing that ensemble-based models (Random Forest and XGBoost) were more effective at capturing complex interactions within the data.

Overall, this project highlights the effectiveness of machine learning for injury prediction in sports, with **ensemble methods showing the most promise**. These models can be integrated into athlete management systems to flag high-risk individuals early, allowing for preventive measures and training adjustments.