# **Objective & Dataset Description**

For this project, I selected two healthcare-related datasets focused on **sports injury prediction and athlete performance**. I chose these datasets because I was a professional badminton athlete, and improper rest and recovery led to a shoulder injury. This personal experience motivated me to explore how data can help athletes understand when they should rest, rehab, or train, and how factors such as fatigue, workload, training hours, and physiological responses impact injury risk.

1. **Dataset 1: College Sports Injury Detection**  
   This dataset contains information on collegiate athletes, including **demographics** (age, gender, height, weight), **training metrics** (training intensity, training hours per week, match count per week, recovery days), and **performance indicators** (fatigue score, performance score, team contribution score, load balance score, ACL risk score). It also includes an **injury indicator** showing whether an athlete sustained an injury. The dataset consists of **approximately 200 rows and 17 columns**, capturing both subjective assessments (scores) and objective measurements to study factors contributing to injury risk among athletes.
2. **Dataset 2: Athlete Injury and Performance Dataset**This dataset records **session-level sports performance and health metrics** for athletes across multiple sports. It includes **demographics** (Athlete ID, sport type), **physiological measurements** (heart rate, respiratory rate, skin temperature, blood oxygen level), **training/physical activity metrics** (impact force, activity type, duration, cumulative fatigue index), and an **injury risk score**. The target variable is **Injury Occurred**, indicating whether an athlete sustained an injury during a session. The dataset has **over 1,000 rows and 13 columns**, allowing for analysis of the relationships between physiological load, activity type, and injury risk.Both datasets contain well over 100 rows (Dataset 1=200 rows, Dataset 2=1000 rows) and provide comprehensive insight into injury-related patterns across sports.

# **Data Collection / Sources**

# Both datasets were obtained from **Kaggle**, a reputable platform for public datasets used in machine learning, healthcare, and sports analytics.

* **Dataset 1: College Sports Injury Detection**  
  **Source:** Kaggle ([link](https://www.kaggle.com/datasets/ziya07/college-sports-injury-detection))
* **Dataset 2: Athlete Injury and Performance Dataset**  
  **Source:** Kaggle ([link](https://www.kaggle.com/datasets/ziya07/athlete-injury-and-performance-dataset))

These datasets are **publicly available for educational and research purposes**. No API, text mining, or web scraping was required; the data was downloaded directly in CSV format.

# **Data Cleaning & Wrangling**

Dataset 1: Collegiate Athlete Injury Dataset

1. Loaded the dataset using **pandas** to examine its structure and contents.
2. Displayed the first five records in a transposed format for better readability and to observe all features clearly.
3. Generated summary statistics for all variables (transposed) to assess central tendencies, variability, and potential anomalies.
4. Verified dataset metadata with df.info() to confirm column counts, data types, and non-null value distribution.
5. Checked for missing values using df.isnull().sum() to identify any columns requiring imputation or removal.
6. Detected duplicate rows with df.duplicated() to ensure all records are unique.
7. Validated data types by comparing actual column types with expected types (numeric vs. categorical).
8. Ensured categorical fields such as **Gender**, **Position**, and **Athlete\_ID** were stored as object type variables.
9. Verified numeric fields, including **Age**, **Height**, **Training Hours**, performance scores, and risk indicators, were stored as integers/floats suitable for statistical analysis.
10. Created a new variable **BMI**, calculated as weight (kg) divided by height (m) squared, to provide a measure of body composition. This column was added after **Weight\_kg** and rounded to one decimal for consistency.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A graph with blue squares

AI-generated content may be incorrect.

A blue and white rectangular graph

AI-generated content may be incorrect.

A blue graph with white text

AI-generated content may be incorrect.

A graph with text overlay

AI-generated content may be incorrect.

A graph with numbers and lines

AI-generated content may be incorrect.

Dataset 2: Sports injury detection dataset

**Dataset 2: Sports Injury Detection Dataset**

1. Loaded the dataset using **pandas** to inspect structure, verify correct file ingestion, and assess overall integrity.
2. Displayed the first five records in a transposed format (df.head().T) to clearly view feature-level patterns and improve readability of each record.
3. Generated summary statistics (df.describe().T) for all numerical variables to examine distributions, detect irregularities, and identify potential outliers.
4. Verified dataset metadata with df.info() to confirm column names, data types, non-null counts, and schema consistency.
5. Checked for missing values using df.isnull().sum() to determine if imputation or additional cleaning was required.
6. Identified duplicate records using df.duplicated() to ensure all entries were unique.
7. Validated data types by comparing stored formats with expected types, ensuring categorical, integer, and float features were correctly defined.
8. Ensured categorical variables, including **Athlete\_ID**, **Sport\_Type**, and **Session\_Date**, were stored as object types for proper grouping, filtering, and classification.
9. Confirmed physiological and activity-based metrics such as **Heart\_Rate\_BPM**, **Respiratory\_Rate\_BPM**, **Impact\_Force\_Newtons**, **Cumulative\_Fatigue\_Index**, and **Injury\_Risk\_Score** were stored in appropriate numeric formats for statistical and predictive modeling.
10. Performed outlier detection using the **Interquartile Range (IQR) method** to flag extreme or physiologically implausible values that could affect analysis.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A diagram of a heart rate

AI-generated content may be incorrect.

A graph of blue dots

AI-generated content may be incorrect.

A graph of injury rate by activity type

AI-generated content may be incorrect.

A graph of a graph

AI-generated content may be incorrect.

# **Data Dictionary:**

| **Variable Name** | **Description** |
| --- | --- |
| **Athlete\_ID** | Unique identifier for each athlete |
| **Age** | Athlete age in years |
| **Gender** | Athlete gender (Male or Female) |
| **Height\_cm** | Athlete height in centimeters |
| **Weight\_kg**  **BMI** | Athlete weight in kilograms  Body Mass Index calculated as weight (kg) divided by height (m) squared, representing athlete’s body composition. |
| **Position** | Playing position (Guard, Forward, Center) |
| **Training\_Intensity** | Intensity score of training sessions (1–10) |
| **Training\_Hours\_Per\_Week** | Total weekly training hours |
| **Recovery\_Days\_Per\_Week** | Number of days allocated for recovery |
| **Match\_Count\_Per\_Week** | Number of competitive matches played per week |
| **Rest\_Between\_Events\_Days** | Rest period between events in days |
| **Fatigue\_Score** | Fatigue scale score (1–10) |
| **Performance\_Score** | Overall performance rating |
| **Team\_Contribution\_Score** | Contribution of athlete to team performance |
| **Load\_Balance\_Score** | Workload distribution score |
| **ACL\_Risk\_Score** | Anterior Cruciate Ligament injury risk score |
| **Injury\_Indicator** | Whether athlete sustained an injury (0 = No, 1 = Yes) |
| **Heart\_Rate\_BPM** | Heart rate during session (beats per minute) |
| **Respiratory\_Rate\_BPM** | Respiratory rate during session (breaths per minute) |
| **Skin\_Temperature\_C** | Body surface temperature in Celsius |
| **Blood\_Oxygen\_Level\_Percent** | Blood oxygen saturation (%) |
| **Impact\_Force\_Newtons** | Force experienced during activity in Newtons |
| **Cumulative\_Fatigue\_Index** | Accumulated fatigue over time |
| **Activity\_Type** | Type of physical activity (e.g., Running, Sprinting, Jumping) |
| **Duration\_Minutes** | Duration of the session in minutes |
| **Injury\_Risk\_Score** | Calculated risk of injury for the session |
| **Injury\_Occurred** | Whether injury occurred in the session (0 = No, 1 = Yes) |
| **Training\_Load\_Index (Created)** | Combined workload: Duration × Impact Force (or Intensity × Hours) |
| **Fatigue\_to\_Recovery\_Ratio (Created)** | Ratio of Fatigue Score to Recovery Days |
| **Weekly\_Stress\_Score (Created)** | Stress indicator: Training Intensity × Hours + Match Count |
| **Height\_Weight\_Ratio (Created)** | Athlete height divided by weight (cm/kg) |
| **High\_Risk\_Athlete\_Indicator** | Binary indicator for high ACL risk (1 = High Risk, 0 = Low/Normal Risk) |
| **Fatigue\_to\_Recovery\_Index (Created)** | Dataset 2 version: Cumulative Fatigue / Duration |
| **Oxygen\_Efficiency\_Ratio (Created)** | Blood Oxygen Level / Heart Rate |
| **High\_Risk\_Session (Created)** | Binary indicator for session risk (1 = High Risk, 0 = Low Risk) |
| **Stress\_Index (Created)** | Session stress: Cumulative Fatigue × Heart Rate |
| **Training\_Load (Created)** | Training Intensity × Training Hours |
| **Weekly\_Exposure (Created)** | Training Hours + Match Count |
| **Fatigue\_Category (Created)** | Categorical fatigue level: Low, Moderate, High |

### **Data Analysis:**

### **A. Research Questions and Statistical Answers**

### **Q1. Does higher training intensity increase the probability of injury?**

**Method:** Logistic regression  
**Result:**

* Training Intensity coefficient = **0.169**
* AUC = **0.610**, Accuracy = **0.933**  
  **Interpretation:**  
  Higher training intensity slightly increases injury probability, though the relationship is **weak-to-moderate**.  
  **Visualization:** Logistic regression probability plot.

A graph with a red line

AI-generated content may be incorrect.

### **Q2. Does accumulated fatigue correlate with injury risk?**

**Method:** Pearson correlation  
**Result:**

* Pearson r = **0.292**
* p < **0.0001**  
  **Interpretation:**  
  There is a **significant positive correlation** between fatigue and injury. Higher fatigue strongly increases injury risk.

### **A graph of injury and injury AI-generated content may be incorrect.**

### **Q3. Do athletes with fewer recovery days have more injuries?**

**Method:** Group mean comparison (bar chart)  
**Result:**  
Mean injury rates by recovery days:

* **1 day:** 0.179
* **2 days:** 0.014
* **3 days:** 0.016  
  **Interpretation:**  
  Athletes with only **1 recovery day per week** have dramatically higher injury rates.

**Visualization:** **A graph with a bar

AI-generated content may be incorrect.**

### **Q4. Is high heart rate associated with higher injury occurrence?**

**Method:** Logistic regression  
**Result:**

* Coefficient = **0.00246**
* Accuracy = **0.617**, AUC = **0.517**  
  **Interpretation:**  
  Heart Rate does **not** meaningfully predict injury. Relationship is minimal.

**Visualization:**

A graph with a red line

AI-generated content may be incorrect.

### **Q5. Does high impact force predict injury in Dataset 2?**

**Method:** Pearson correlation  
**Result:**

* Pearson r = **0.126**, p < **0.0001**  
  **Interpretation:**  
  Impact Force has a **small but significant** positive relationship with injury likelihood.

**Visualization:**

A graph of a number of objects

AI-generated content may be incorrect.

### **Q6. Does training load correlate with fatigue score?**

**Method:** Linear regression  
**Dataset 1 Result:**

* Coefficient = **–0.00041**
* Interpretation: Training load does **not** predict fatigue (relationship negligible).  
  **Dataset 2 Result:**
* Coefficient = **9.98 × 10⁻⁷**
* Interpretation: Almost zero correlation.

**Visualization:**

A graph of loading and load

AI-generated content may be incorrect.

### **Q7. Are certain positions (Guard/Forward/Center) more injury-prone?**

**Method:** Chi-square test  
**Result:**

* χ² = **1.168**, p = **0.558**  
  **Interpretation:**  
  There is **no significant difference** in injury rates between positions.

Position,0,1

Center,63,3

Forward,65,5

Guard,58,6

### **Q8. Which activity types show higher injury risk?**

**Method:** Group means + bar chart  
**Injury rates by activity:**

* **Sprinting:** 0.661
* **Dribbling:** 0.639
* **Jumping:** 0.619
* **Running:** 0.612
* **Jogging:** 0.550  
  **Interpretation:**  
  High-intensity explosive activities (Sprinting, Dribbling, Jumping) have the highest injury risk.

**Visualization:** **A graph of injury rate by activity type

AI-generated content may be incorrect.**

### **Q9. Does athlete body composition (height/weight) influence fatigue?**

**Method:** Multiple regression  
**Result:**

* Height coefficient = **–0.0057**
* Weight coefficient = **–0.0086**  
  **Interpretation:**  
  Taller/ heavier athletes show **slightly lower fatigue**, but the effect is small.

**Visualization:**

### **Q10. Is weekly exposure (hours + matches) a predictor of injury?**

**Method:** Logistic regression  
**Result:**

* Coefficient = **0.0734**
* AUC = **0.640**, Accuracy = **0.933**  
  **Interpretation:**  
  Weekly exposure **increases injury probability**, making this a useful risk factor.

**Visualization:** Logistic regression curve, exposure histogram.

A graph with a red line

AI-generated content may be incorrect.

**Basic EDA:**

**A chart with different colored squares

AI-generated content may be incorrect.**

**A chart with different colored squares

AI-generated content may be incorrect.**

**A diagram of a number of dots

AI-generated content may be incorrect.**

**A colorful dots on a white background

AI-generated content may be incorrect.**

**A graph of blue dots

AI-generated content may be incorrect.**

**A chart with blue dots

AI-generated content may be incorrect.**

# **Software Used**

* **Python** – Cleaning, analysis, visualization.
* **VS Code** – Main coding environment.
* **pandas** – Data manipulation tasks.
* **numpy** – Numerical computations support.
* **matplotlib** – Plotting visual charts.
* **seaborn** – Statistical visual graphics.
* **scikit-learn** – Machine learning models.
* **Excel** – Initial dataset inspection.

.