**Objective:** Injury prediction

**Process\_1 (Split\_Data - Decision\_Tree):**

* No Missing Values
* Histograms were plotted for: *Player\_Age, Player\_Weight, Player\_Height, Training\_Intensity, Days\_of\_Recovery* → Normal distribution → No Normalisation needed
* *Previous\_Injury*: conversion from integer to binominal
* *Injury\_No-Injury*: conversion from integer to binominal
* *Injury\_No-Injury*: set as a target variable
* Dataset split: 70% used for training, 30% for testing
* Model applied: Decision Tree Classifier

**Evaluation Metrics:**

All criteria, i.e. information gain, gain ratio, gini index and accuracy were assessed. The rest of the parameters of the decision tree were also modified, without nevertheless those modifications to significantly affect the outcome . Regarding the evaluation metrics, accuracy and classification error ranged from 47% to 49% and from 50% to 52% respectively.

**Process\_2 (Cross Validation - Decision\_Tree):**

For proper evaluation of the performance and robustness of the model (better estimation of the prediction error), a 10-fold cross validation operation was applied. Accuracy and classification error remained at approximately the same levels, with 47.7% and 52.3% respectively.

**Process\_3 (Cross Validation - Boosting):**

Attempting to improve the performance of the model, a boosting ensemble model was employed, AdaBoost, with a decision tree in it (as in the initial classification attempt). The accuracy increased marginally from 47.7% to 50.2%. No further preprocessing was done given that there were no missing values and that the variables were fairly balanced.

**Process\_4 (Cross Validation - Regression):**

Attempting to examine further the relationship between the included variables and the target variable, a logistic regression was applied. The logistic regression was selected due to the binominal type of the label attribute. All the variables were included in the model and only *training\_intensity* was found to have a statistically meaningful relationship with the target variable (Injury\_NoInjury) with a p value <0.05. The positive coefficient of 0.626 with an odds ratio of approximately 1.87 means that for each unit increase in *training\_intensity*, the odds of injury (compared to no injury) are multiplied by approximately 1.87. This finding aligns with the findings of the decision tree model as *training\_intensity* was selected for the first split at the root node, meaning that it is the most important feature in the dataset for predicting the target variable.

Overall, the employed models did not manage to produce a robust performance. A probable explanation for that is the fact that the used dataset was not a validated dataset. The features’ data points might not reflect actual measurements, something which could contaminate the results. This does not necessarily mean that the selected features do not play a role in the process of injury prediction. Training intensity, as it has been found here, is a meaningful selection nonetheless, which has also been reported in the bibliography as a risk factor.