Final Project Writeup: https://github.com/meelod/cs158-final-project/tree/main

Introduction:

I explored the Random Forest algorithm to predict injuries in Track and Field athletes. According to the National Library of Medicine, 1 in every 5.8 males and 7.5 females get injured. When looking at projects to find, I wanted to look into providing solutions for problems that have high rates of affecting a subset of people. Another reason is that I ran Track and Field in high school and my first year at Pomona College. During those years, I missed three full seasons due to injuries. I decided to use Random Forest because of its ability to track feature importance.

Experimental Setup:

I used a Kaggle dataset containing training logs from a professional Dutch running team spanning seven years, from 2012 to 2019. The data focused on middle and long-distance runners, ranging from 800 meters to the marathon. The reason for this is because of how the training programs of long-distance runners are traditionally structured, as that bracket of long-distance runners tends to follow similar training programs compared to short-distance runners or sprinters. Another key aspect of the dataset is that the same head coach oversaw all the training within the data throughout the seven years. This is a good indicator that the training program throughout the years remains consistent, as it was provided by the same head coach.

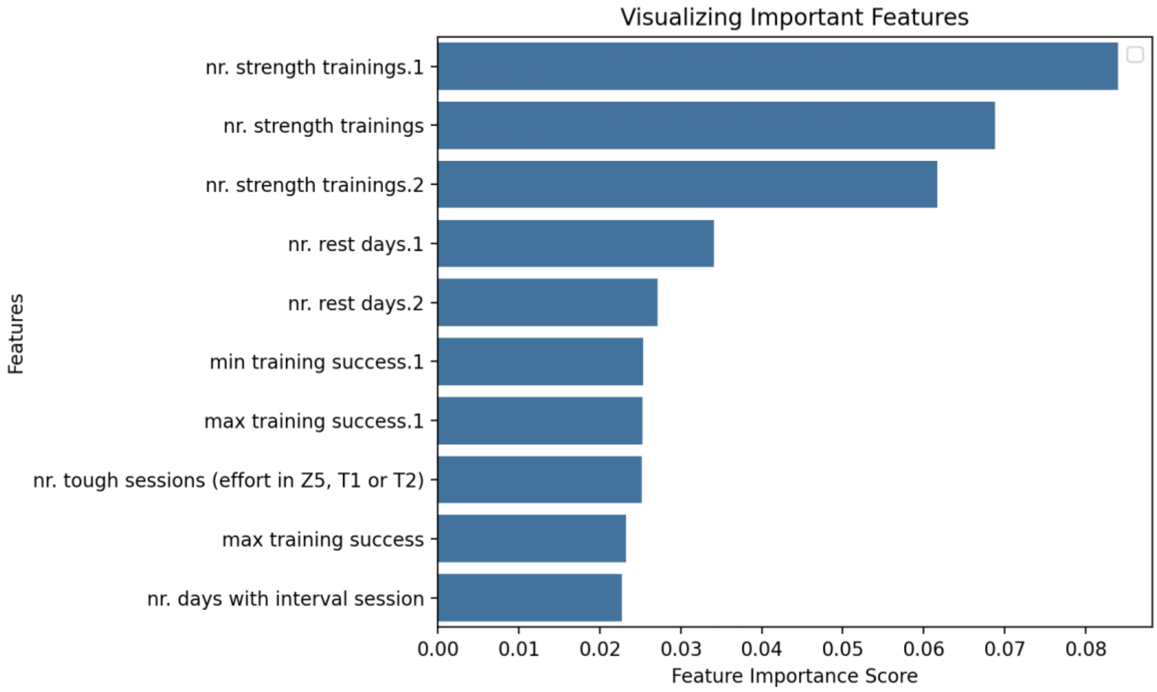
I then deployed a Random Forest Classifier to predict injuries and rank the most important features that cause injuries. Some of those key features include the number of strength training sessions, rest days, and total distance traveled—all in weekly intervals.

Facing an initial imbalance with over 200 examples of injuries for over 8000 examples of non-injuries. I addressed this by using the Synthetic Minority Over-sampling Technique

(SMOTE) to generate synthetic samples for the minority class (injured athletes). The dataset was split into training and testing sets using an 80-20 split.

Results:

With over 96% accuracy, it was able to predict injuries and rank the features with the highest importance in relation to causing injuries. The rankings show that strength training 1 day before has the highest probability of causing injury, followed by strength training in the week and 2 days before.



Conclusions:

With my algorithm, I was able to accurately predict injuries in mid to long-distance track and field runners and determine that the highest cause of injury is strength training and having rest days one or two days before training.