Machine Learning and Sports

Executive Summary

By Alejo Alegre Bustos

Background

This analysis investigates weekly training data from a cohort of athletes to identify potential predictors of injury. The primary goal was to leverage performance metrics to understand which training patterns might lead to higher injury risk, and to build a predictive model that could help anticipate and reduce those injuries in future training cycles.

The dataset includes variables such as:

* Maximum kilometers run in a single day
* Total kilometers for the week
* Acute and chronic training loads
* Number of running days
* Injury status (binary)

These metrics were collected weekly per athlete, offering a granular view into how training decisions affect physical outcomes.

Key Findings

1. Strong Correlation Between Rest days and Max training success

This chart highlights a clear negative relationship between the number of weekly rest days and the athlete’s ability to reach high training performance. Weeks with fewer rest days (0–2) show the highest average training success, suggesting that consistent training is key to maintaining performance levels.

As rest days increase—especially beyond 4 days—the performance drops significantly. A full week of rest (7 days) shows zero training success, likely due to a loss of training continuity, physical readiness, or focus.

This implies that while rest is necessary for recovery, too much rest may lead to undertraining and reduced effectiveness. Coaches should aim for a balanced training schedule, typically with 1 to 3 rest days per week, depending on the athlete’s condition and goals.

These insights are valuable for planning weekly workloads, optimizing rest periods, and preventing performance loss during low-activity weeks.

A graph with a line

AI-generated content may be incorrect.

1. Resting days and Injuries

A graph with a red line

AI-generated content may be incorrect.

This chart shows an inverse relationship between the number of rest days and the weekly injury rate. As rest days increase, the likelihood of injury decreases consistently. Weeks with zero rest days have the highest injury rate, while those with six or seven rest days show a much lower injury occurrence.

This pattern strongly suggests that insufficient rest increases the risk of injury, possibly due to accumulated fatigue, inadequate recovery, or overtraining. Even small increases in rest days—from 0 to 2 or 3—lead to noticeable reductions in injury rate, highlighting the protective effect of structured recovery.

For training plans, this insight emphasizes the need to balance training intensity with adequate rest. Coaches should avoid scheduling full training weeks without rest, especially during high-load periods, and consider using rest strategically to minimize injury risk. Athletes showing early signs of fatigue or under-recovery could particularly benefit from adjusted rest schedules.

Predictive Modeling

A Random Forest Classifier was trained to predict injury presence based on the weekly metrics. The data was split into training and testing sets, and the model’s performance was evaluated using standard classification metrics:

* Accuracy: > 98%
* Precision/Recall: Balanced, suggesting minimal false positives or false negatives
* F1-Score: High consistency in prediction

The model's output isn’t just a binary flag — it provides probability scores that can be integrated into early warning systems or dashboards.

Importance of Findings

These findings validate a long-held belief in sports science: unmanaged training volume leads to injury. But this project goes further — it quantifies the risk and demonstrates how data-driven thresholds can guide training plans. Rather than relying on intuition or fixed schedules, teams can:

* Monitor dynamic thresholds for load increases
* Use predictive risk scores weekly to flag at-risk athletes
* Adjust training for individuals instead of applying general rules

Impact on Future Training

The integration of this model into athlete management systems would offer substantial benefits:

* Proactive Adjustments: Coaches can intervene before injury happens, adjusting volume or intensity based on real-time data.
* Athlete Longevity: Fewer injuries mean more consistent participation and improved long-term performance.
* Objective Feedback: Training decisions can now be grounded in measurable risk, reducing guesswork.

Notebook: https://colab.research.google.com/drive/1JyRzgJsPxYd7i9u\_nXVrrI-fVCKDlR9o?authuser=1#scrollTo=iGjTKz2\_G3AA