Predicting Soccer Injuries with Machine Learning

Jennifer Lawless
DASC 9311: Data Science Project
September 17, 2025

The Problem

Core Problem	• Predict whether a soccer player will suffer an injury in their next season
Why This Matters	• Player safety and career longevity • Identifying key injury risk factors • Preventative healthcare • Training program optimization
Technical Challenges	• **Dataset Size:** Limited samples increase overfitting risk • **Generalization:** Model must work on unseen players • **Feature Engineering/Selection:** Must avoid multicollinearity and data leakage
Goals	• **Primary Goal:** >85% accuracy with strong generalization • **Secondary Goal:** Outperform existing Kaggle solutions • **Practical Goal:** Provide actionable insights for injury prevention

Dataset Overview

Samples

800 Chinese university soccer players aged 18-24 participating in collegiate and provincial leagues

Features

18 features including physical characteristics, soccer-specific metrics, physical fitness assessments, lifestyle factors, and training compliance

Target

Injury_Next_Season: binary classification value where an injury is defined as training/competition-related injury causing ≥ 7 consecutive days of absence

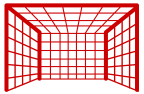
Data Quality

Multi-source verification (medical records, coach reports, student surveys), **well-balanced** dataset, and **minimal missing data**

Link to Kaggle: https://www.kaggle.com/datasets/yuanchunhong/university-football-injury-prediction-dataset/data

Solution Strategy





Comprehensive EDA	Feature Engineering	Model Development	Validation Framework	Multi-Metric Evaluation
• Perform an analysis of injury patterns by position, fitness level, and lifestyle factors	• Create position-relative metrics	• Start with simpler baseline models and then progress to more sophisticated ensembles	• Use nested cross-validation to prevent overfitting	• In addition to Accuracy Rate, use Recall, Precision, F1-Score, and AUC-ROC

Modeling Approach

1

Baseline Models

Logistic Regression, Random Forest XGBoost

2

Advanced Models

TabNet, Gaussian Process Classifier, SVM, LightGBM

3

Ensemble Methods

Stacking Ensemble, Voting Classifier, Bayesian Model Averaging

Project Timeline

Week 1 (9/1-9/7): Dataset acquisition, initial exploration, environment setup

Week 2 (9/8-9/14): Review of existing Kaggle solutions, methodology research

Weeks 3-4 (9/15-9/28): EDA, feature analysis & engineering, data preprocessing

Weeks 5-6 (9/29-10/12): Baseline model implementation, initial performance benchmarking

Weeks 7-10 (10/13-11/2): Advanced and ensemble model development, model fine-tuning

Weeks 11-13 (11/3-11/23): Final model selection, results analysis

Weeks 14-15 (11/24-12/7): Documentation, report-writing

Accomplishments to Date (through 9/17)

Kaggle Notebook Review

Analyzed 8+ existing Kaggle notebooks for this dataset to see that XGBoost achieves 85%+ accuracy in almost all solutions

Environment Setup

Configured Python with necessary packages (scikit-learn, XGBoost, PyTorch, TabNet)

Planning

Designed my model portfolio, established timelines and milestones

Exploratory Data Analysis

Starting basic EDA on the dataset this week

Thank You!