Predicting Soccer Injuries with Machine Learning

Module 3: Progress Report

Jennifer Lawless
DASC 9311: Data Science Project
October 8, 2025

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

Today (10/8)

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

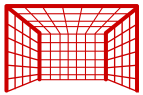
Weeks 7-10 (10/13-11/2): Advanced and ensemble model development, model fine-tuning

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

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

EDA Methods





Basic EDA	Injury Pattern Analysis	Correlation Analysis	Feature Engineering & Data Preprocessing	Feature Importance Tests
• Viewed descriptive statistics • Built histograms of numeric features • Built bar plots of categorical features	• Split variables into related groups • Ran appropriate statistical tests (t-test, Mann-Whitney U, chi-square) • Created boxplots/bar plots by injury status	• Generated full correlation matrix • Checked highly correlated pairs • Examined correlation with target variable • Visualized top correlations	• Created new features • Capped outliers using IQR method • Performed data scaling and one-hot encoding	• Ran F-test (ANOVA) and Mutual Information • Combined rankings • Visualized top features

Basic EDA

Shape

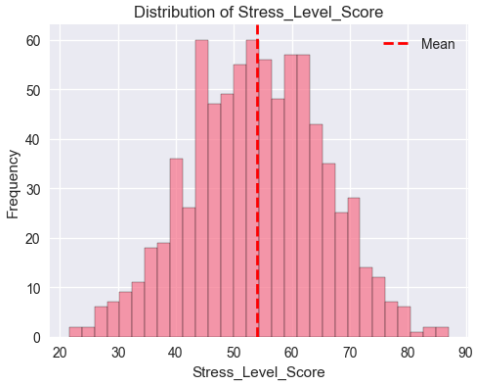
800 rows, 18 features (with 17 numerical and 1 categorical)
**NOTE: one of the numerical variables is a binary categorical variable

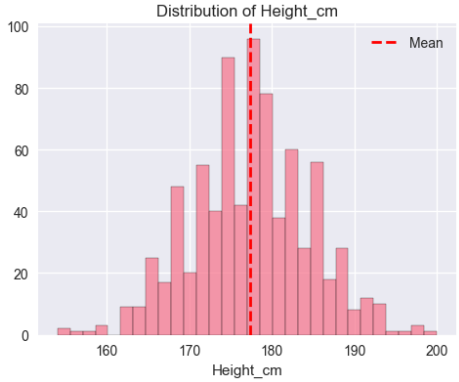
Target

Injury_Next_Season, perfectly balanced
(50% injury next season / 50% no injury next season)

Data Quality

No missing values, no anomalous fields
**NOTE: there are some outliers present






```
Data columns (total 19 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Age                         800 non-null    int64
 1   Height_cm                   800 non-null    int64
 2   Weight_kg                   800 non-null    int64
 3   Position                    800 non-null    object
 4   Training_Hours_Per_Week     800 non-null    float64
 5   Matches_Played_Past_Season  800 non-null    int64
 6   Previous_Injury_Count       800 non-null    int64
 7   Knee_Strength_Score         800 non-null    float64
 8   Hamstring_Flexibility       800 non-null    float64
 9   Reaction_Time_ms            800 non-null    float64
 10  Balance_Test_Score          800 non-null    float64
 11  Sprint_Speed_10m_s          800 non-null    float64
 12  Agility_Score               800 non-null    float64
 13  Sleep_Hours_Per_Night       800 non-null    float64
 14  Stress_Level_Score          800 non-null    float64
 15  Nutrition_Quality_Score     800 non-null    float64
 16  Warmup_Routine_Adherence    800 non-null    int64
 17  Injury_Next_Season          800 non-null    int64
 18  BMI                         800 non-null    float64
dtypes: float64(11), int64(7), object(1)
```

Variable Groupings

Age, Height_cm, Weight_kg, BMI

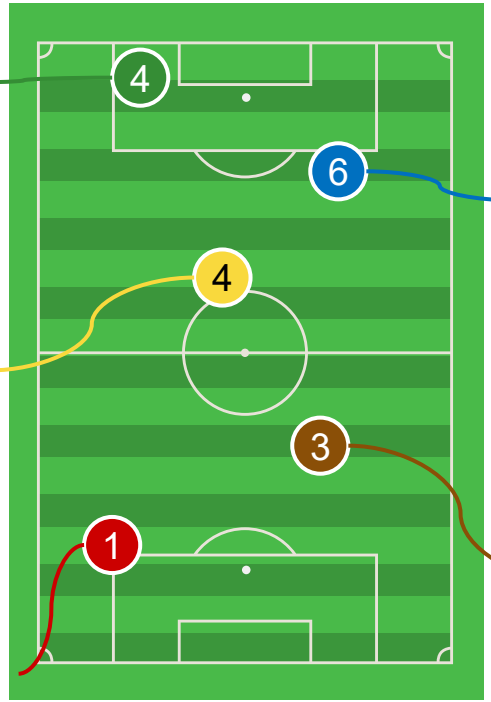
Physical Characteristics

Position, Training_Hours_Per_Week, Matches_Played_Past_Season, Previous_Injury_Count

Soccer-Specific Metrics

Warmup_Routine_Adherence

Training Compliance



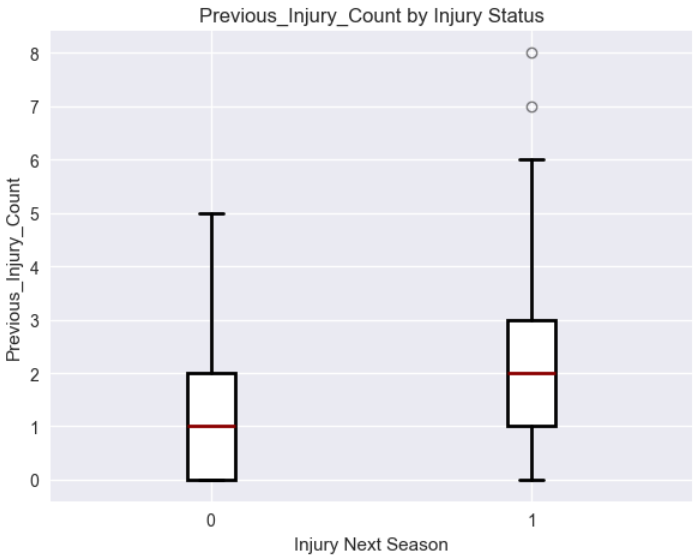
Physical Fitness Assessment

Knee_Strength_Score, Hamstring_Flexibility, Reaction_Time_ms, Balance_Test_Score, Sprint_Speed_10m_s, Agility_Score

Lifestyle Factors

Sleep_Hours_Per_Night, Stress_Level_Score, Nutrition_Quality_Score

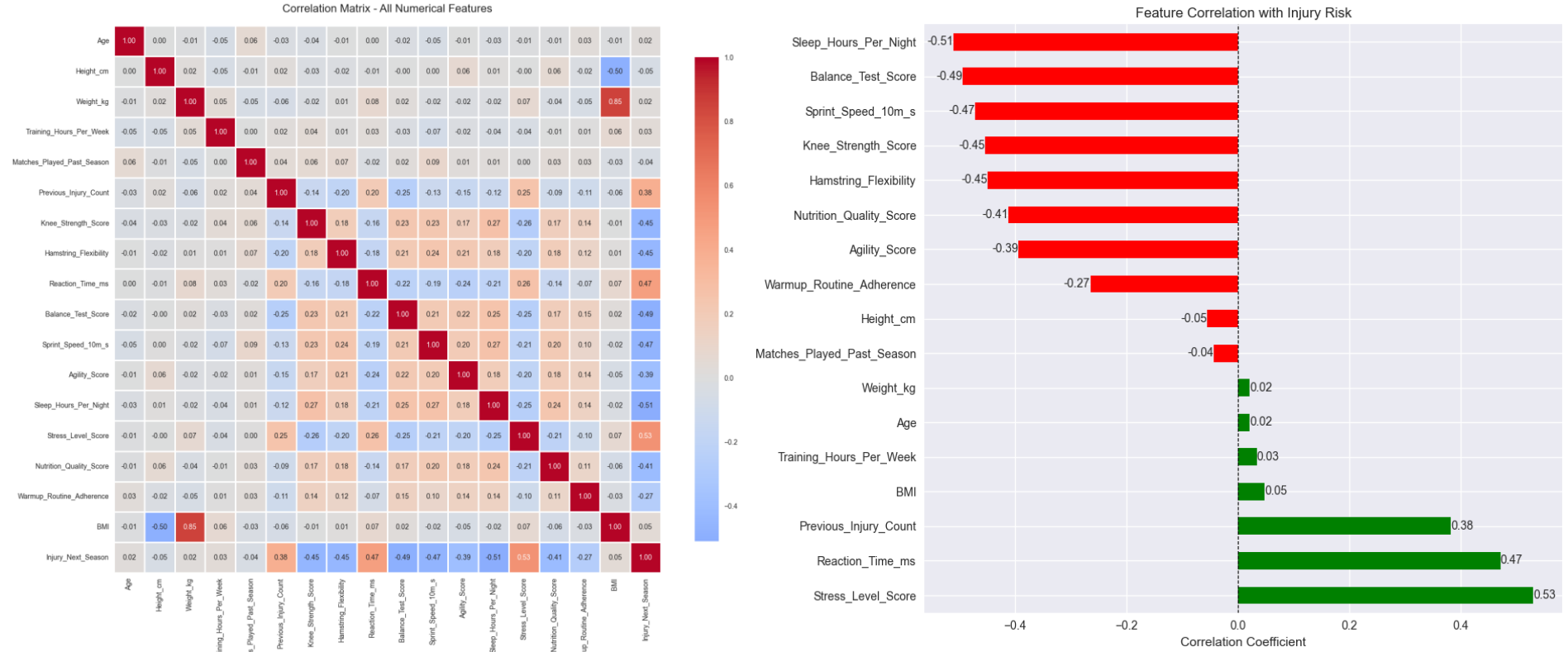
Injury Pattern Analysis




```
Previous_Injury_Count
   Normality p-values -> Injured: 0.0000, Not Injured: 0.0000
   Mann-Whitney U -> statistic=114072.5000, p-value=0.0000
   Significant difference: Injured players have higher average Previous_Injury_Count.
```

Physical Characteristics	• Age, height, weight, BMI have no significant differences, so they are not strong predictors
Soccer-Specific Metrics	• Previous injury count is strongly predictive ($p < 0.001$) • Training load & position not significant
Fitness Assessments	• Injured players had lower strength, flexibility, balance, agility and slower reaction/sprint speed • Poor fitness greatly increases injury risk
Lifestyle Factors	• Injured players had less sleep, higher stress, poorer nutrition • Lifestyle is strongly linked to risk
Training Compliance	• Poor warmup adherence strongly associated with injury ($p < 0.001$)

Correlation Analysis



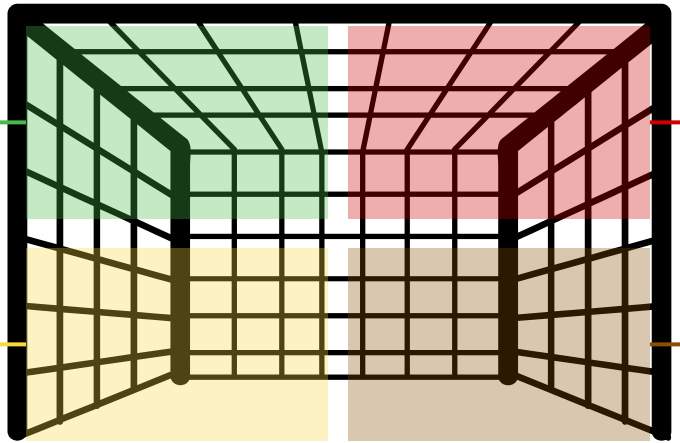
Feature Engineering

Binning & Categories

- **Age_Group** (U20, U22, U25)
- **BMI_Category** (Underweight, Normal, Overweight, Obese)

Ratios & Interactions

- **Stress_Sleep_Ratio** (Stress_Level_Score / Sleep_Hours_Per_Night)
- **Workload_Recovery_Ratio** (Matches_Played_Past_Season / Sleep_Hours_Per_Night)
- **Age_x_Training** (multiplying age with training hours)



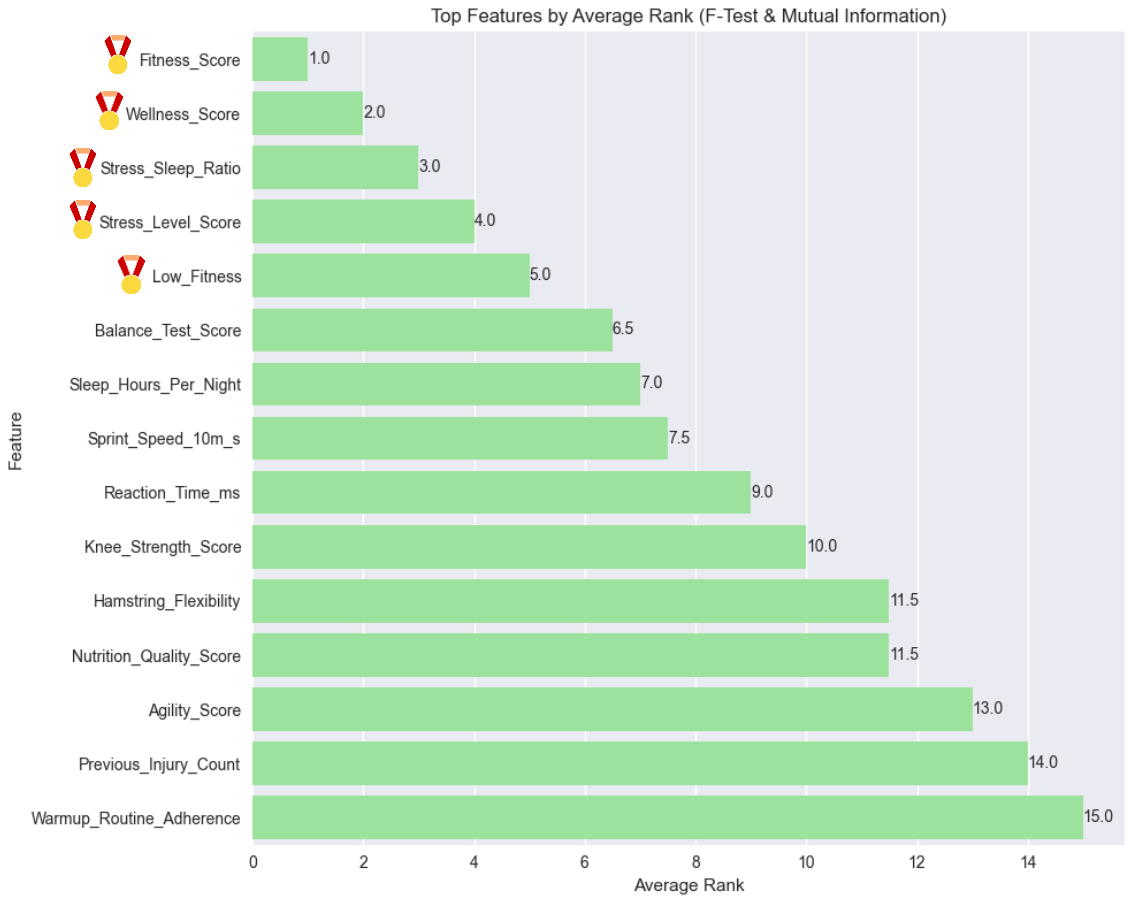
Composite Scores

- **Fitness_Score** (combined strength, flexibility, sprint speed, agility, balance)
- **Wellness_Score** (combined sleep, stress, nutrition)
- **Training_Intensity** (weighted measure of training hours, match load, and recovery)

Risk Flags

- **High_Risk** (binary indicator if multiple risk thresholds are exceeded (high stress + low sleep))
- **Low_Fitness** (binary indicator if fitness composite falls below a cutoff)

Feature Importance



Top Predictors

- **Fitness Score** - strongest overall predictor (higher fitness = lower risk)
- **Wellness Score** - captures sleep, stress, recovery balance
- **Stress metrics** (Stress/Sleep Ratio, Stress Level) - highlight recovery imbalance
- **Low Fitness flag** - confirms poor conditioning as a clear risk factor

Final Feature Selection

Selected Features

Feature	Reasoning
`Stress_Level_Score`	Strong correlation (+0.53), top-ranked, statistically significant
`Sleep_Hours_Per_Night`	Strong negative correlation (–0.51), top-ranked, significant
`Wellness_Score`	Top-ranked, captures multiple recovery dimensions
`Sprint_Speed_10m_s`	Strong negative correlation, top-ranked, significant
`Reaction_Time_ms`	Strong positive correlation, top-ranked, significant
`Knee_Strength_Score`	Strong negative correlation, top-ranked, significant
`Hamstring_Flexibility`	Strong negative correlation, top-ranked, significant
`Balance_Test_Score`	Strong negative correlation, top-ranked, significant
`Agility_Score`	Strong negative correlation, top-ranked, significant
`Nutrition_Quality_Score`	Strong negative correlation, top-ranked, significant
`Previous_Injury_Count`	Statistically significant, moderately correlated
`High_Risk`	Engineered feature, moderately ranked
`Fitness_Score`	Top-ranked, likely aggregates multiple physical metrics
`Stress_Sleep_Ratio`	Top-ranked, captures interaction between stress and recovery
`Warmup_Routine_Adherence`	Statistically significant (Chi-square), moderately correlated

Model 1: Logistic Regression

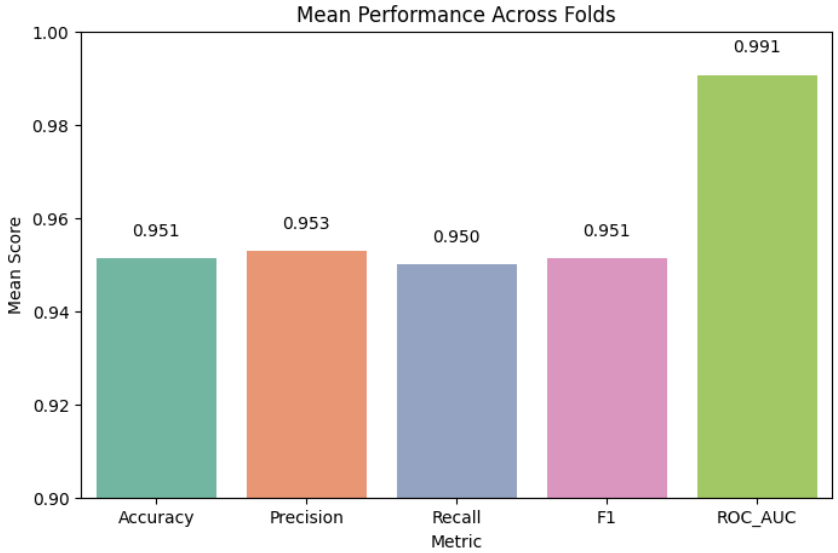
Model Setup

- **Algorithm:** Logistic Regression (binary classification)
- **Solver:** (supports L1 & L2 penalties)
- **Max Iterations:** 1000 (ensures convergence)
- **Hyperparameters tuned:**
  - Regularization strength = [0.01, 0.1, 1, 10]
  - Penalty = L1 (Lasso) or L2 (Ridge)

Validation Strategy

- Nested Cross-Validation for unbiased evaluation
- Outer loop (5-fold Stratified CV): evaluates generalization
- Inner loop (3-fold Stratified CV): tunes hyperparameters with GridSearchCV

Scoring Metrics: Accuracy, Precision, Recall, F1, ROC-AUC



Comparison to Existing Kaggle Solutions (Accuracy Rate)

- My Solution: 0.951
- Kaggle Solution 1: 0.9458
- Kaggle Solution 2: 0.950

Next Steps

Complete baseline models

- Build Random Forest and XGBoost models
- Performance benchmarking

Advanced and ensemble model development

- Build TabNet, GPC, SVM, LightGBM, Stacking Ensemble, Voting Classifier, Bayesian Model Averaging models

Model Fine-Tuning & Evaluation

- Fine-tune the models
- Evaluate and analyze each model
- Determine the best-performing model

Documentation

- Document all steps taken
- Write the final report and prepare for presentations

Thank You!