Predictive Analytics for Injury Prevention: Machine Learning Approaches in Competitive Sports
Faisal Shaikh
CIS 3715 Final Project

Introduction:

Athlete injuries in highly competitive sports present significant challenges, often with long-term consequences for individuals and their teams. The *motivation* for my project is to apply the predictive capabilities of machine learning (ML), specifically using supervised and classification models to identify the specific types of athletes that are more susceptible to injury and predict the main contributors of recovery time so that there can exist more specialized care and preventative measures and strategies to better the overall health of these such athletes. Such a system could significantly reduce injury rates, aid in resource allocation, and optimize athletes' training strategies, ultimately protecting athletes' health and improving performance.[1,2]

The fact that injury prevention can significantly impact an athlete's long-term career underscores the importance of this effort. Moreover, understanding injury patterns through data analytics allows for more effective allocation of training and medical interventions, especially for high-risk athletes. These modifications are necessary to minimize time lost to injury and to ensure that athletes perform at their best without risk of injury. The *problem* at hand is the current approach to sports injuries, which often results in missed prevention opportunities and inadequate resource allocation[1,2].

My *contribution* to this project was to apply exploratory data analysis, supervised and classification learning methods to a dataset consisting of player-specific information such as age, weight, height, and training intensity, and analyzing past injury occurrences, where ultimately, we propose the identification of these high-risk athletes and predict main factor that affects recovery time.

Description of Data:

Before discussing the approaches and results, all the data was retrieved from a Kaggle dataset of 1000 values and 7 features (target = likelihood of injury). Athlete profiles included numerical and standardized values of the athlete's age, weight, height, previous injuries (0 = no previous injury, 1 = had previous injury), training intensity, recovery time, and the likelihood of injury (0 = not likely for injury, 1 = likely for injury). The data was also balanced, and no missing values were present.

Approach: Exploratory Data Analysis (EDA):

In predictive analytics for injury prevention, exploratory data analysis (EDA) is critical to understanding athlete profiles important for modeling. EDA provides important statistics such as average, standard deviation, and range that quickly show an athlete's profile, including age, weight, height, previous injuries, training, and recovery. The mean value represents the typical athlete profile in the dataset, while the standard deviation shows the variability of data. Data ranges and quantiles help to analyze data spread and distribution, which is important for pattern recognition and outlier detection.

Regarding the goals of this project, the EDA correlation analysis examines relationship variables, such as how athlete size affects injury risk/recovery, and whether injury history is associated with training intensity. Additionally, graphical tools including histograms and scatterplots provide visual insight into the data, highlight outliers and potential errors, and help pre-model the data.

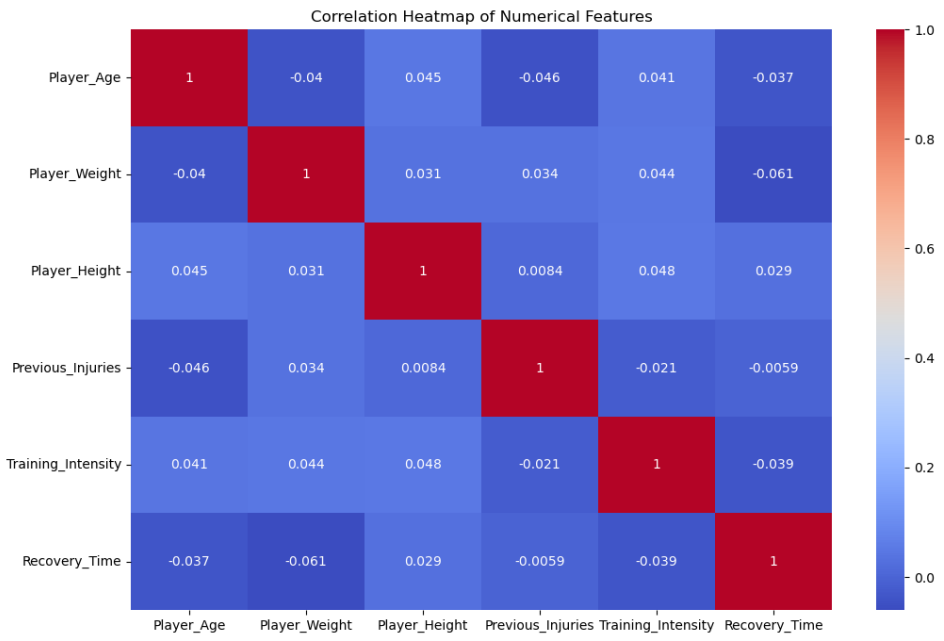
Results: Statistical Summary Insight:

	Player_Age	Player_Weight	Player_Height	Previous_Injuries
count	1.000000e+03	1.000000e+03	1.000000e+03	1.000000e+03
mean	-2.646772e-16	3.819167e-17	-1.037392e-15	-2.486900e-17
std	1.000500e+00	1.000500e+00	1.000500e+00	1.000500e+00
min	-1.565544e+00	-3.499553e+00	-3.486982e+00	-1.030464e+00
25%	-9.534655e-01	-6.928144e-01	-6.792785e-01	-1.030464e+00
50%	-3.534754e-02	2.287886e-02	2.868162e-02	9.704368e-01
75%	8.827704e-01	6.582545e-01	6.886870e-01	9.704368e-01
max	1.647869e+00	3.019492e+00	2.788121e+00	9.704368e-01

	Training_Intensity	Recovery_Time
count	1.000000e+03	1.000000e+03
mean	-8.881784e-18	-1.065814e-16
std	1.000500e+00	1.000500e+00
min	-1.714813e+00	-1.450376e+00
25%	-8.722396e-01	-8.622270e-01
50%	-2.316612e-02	3.140718e-01
75%	8.385730e-01	9.022212e-01
max	1.773211e+00	1.490371e+00

The statistical summary reveals a simplified version of the large dataset which has been normalized so that the mean values are standardized at zero. The median aligns with the mean, indicating a symmetric distribution and excessive examination of quantile ranges allows detection of outliers. This data confirms that the data is ready for predictive where regression and classification models can target the 'likelihood_of_injury' column. Ultimately, EDA is important for project integrity and reliability of predictive models, which contributes to injury prevention efforts.

Results: Correlation Analysis Insight:

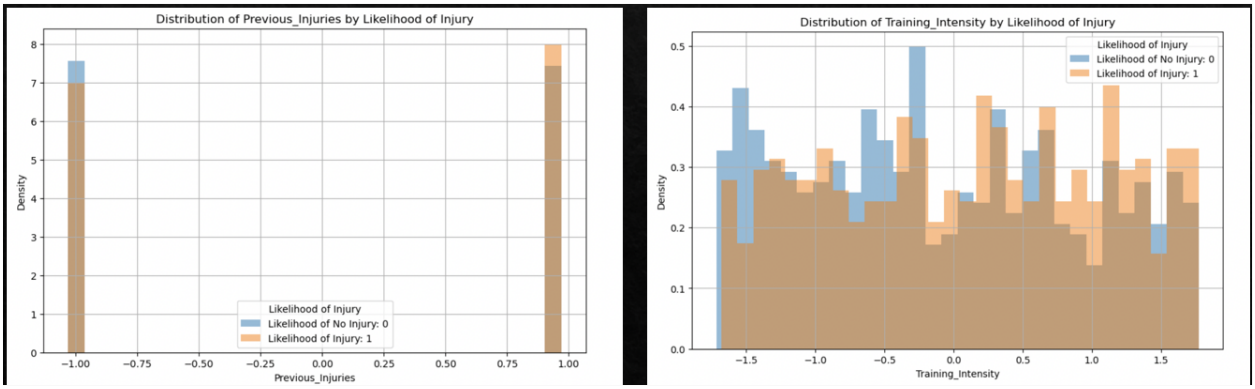


Correlation Analysis Methods include Correlation heatmaps which are an important component of injury analysis, providing insights into the relationships between data-set features and guiding feature selection for building models. The strongest correlation on the map (-0.061) reveals that although small, a negative

correlation exists between ‘Player_Weight’ and ‘Recovery_Time’ which indicates that heavier players have a slightly shorter recovery time, a hypothesis requiring further investigation. Additionally, 'Previous_injuries' shows no strong correlation with 'recovery_time' or 'training_intensity', indicating a non-linear relationship or other influencing factors. All weak correlations are beneficial, reducing concerns about multicollinearity that may affect model reliability.

The heat map facilitates the development of hypotheses, supporting sophisticated machine-learning techniques that can recognize complex and non-linear patterns. This study is a milestone in EDA, highlighting the complexity of injury statistics and the need for advanced modeling techniques to aid in injury prediction and inform prevention strategies.

Results: Distribution Comparison Insights:



A histogram comparison categorized as ‘likelihood_of_injury’ reveals the distribution of player characteristics and their association with injury risks. The 'Player_Age', in 'Player_Height’ and 'Player_Weight' distributions are not strongly predictive of injuries, as the histograms and weak correlations indicated earlier, and were therefore not included in this report.

A difference was found in the presence of ‘previous_injuries’ and ‘training_intensity’, indicating a higher likelihood of future injury for players who have had an injury history. This demonstrated the potential predictive power of injury history. Differences in ‘training_intensity’ indicate an effect on injury risk, and also suggest that increases in intensity result in higher injury likelihood. These graphs were included in this section as shown above.

'Recovery_Time' presents a subtle relationship with the probability of injury, which indicates an optimal recovery time may exist. These insights are helpful for EDA, defining feature selection, and robust transformation construction for predictive models. They highlight the complexity of sports injury prediction and the need for sophisticated modeling techniques to refine the workflow for proposing effective injury prevention strategies in sports.

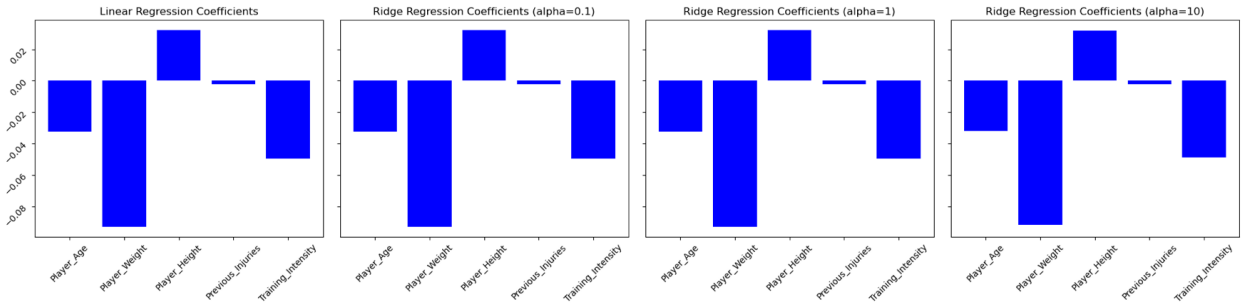
Approach: Supervised learning methods (regression models):

Through supervised learning, linear regression, data normalization, and encoding, can result in the potential prediction of 'recovery_time'. Performance is measured by MAE, MSE, and RMSE, with lower values indicating better accuracy. The table shows the effect of each variable on recovery time, with negative coefficients indicating a decrease and positive coefficients indicating an increase in recovery time with changes in the value of the variable.

Ridge regression is also used to refine the prediction model by penalizing large coefficients through an alpha hyperparameter and handling overfitting and multicore collinearity. By examining different alpha levels (0.1, 1, 10), the project balances bias and variance to improve generalization. As alpha increases, the coefficients decrease, demonstrating the role of ridge regression in increasing model stability and preventing overfitting.

Both these models are used for predicting a continuous variable, which in this project is 'Recovery_Time'. Thus these 2 models are used to evaluate recovery times based on certain athlete characteristics.

Results: Linear and Ridge Regression Insights:




```
Linear Regression Model bias (intercept): -0.001323336191312837
Linear Regression Model coefficients: [-0.03223537 -0.09304579  0.03228919 -0.00242463 -0.04959785]
Linear Regression - Training set - MAE: 0.8738, MSE: 0.9948, RMSE: 0.9974
Linear Regression - Testing set - MAE: 0.8689, MSE: 0.9888, RMSE: 0.9944

Ridge Regression with alpha = 0.1:
Intercept (bias): -0.001323336191312837
Coefficients: [-0.03223084 -0.09303417  0.03228421 -0.00242445 -0.0495921 ]
MAE: 0.8689, MSE: 0.9887, RMSE: 0.9944

Ridge Regression with alpha = 1:
Intercept (bias): -0.001323336191312837
Coefficients: [-0.03219016 -0.09292968  0.03223945 -0.00242281 -0.04954042]
MAE: 0.8689, MSE: 0.9887, RMSE: 0.9943

Ridge Regression with alpha = 10:
Intercept (bias): -0.0013233361913128367
Coefficients: [-0.03178889 -0.09189763  0.03179829 -0.00240656 -0.04902948]
MAE: 0.8690, MSE: 0.9884, RMSE: 0.9942
```

*The coefficient order is as follows Player_Age Coefficient, Player_Weight Coefficient, Player_Height Coefficient, Previous_Injuries Coefficient, Training_Intensity Coefficient

The trained linear regression model shows a bias close to zero, indicating an accurate 'recovery_time' prediction when the predictor features are at their mean. The analytical parameters MAE, MSE, and RMSE reflect the prediction accuracy and potential outliers but reflect the overall model, as reflected in their consistency between the training and test data.

Ridge regression evaluations across alpha levels show consistent error parameters—MAE, MSE, and RMSE—indicating stable generalization. This model was also hyper-tuned by using multiple alpha values. An alpha of 0.1 introduces slightly more error than linear regression, indicating a poor fit. The coefficient reduction with alpha set to 1 and 10 is significant, but there is no significant change in the error parameters, indicating robustness against overfitting and a good bias-variance trade-off. The resistance of a dataset to overfitting or the features of prediction are determined by the position of the

error matrix in the alphas. This robustness against multicollinearity indicates the predictive strength of the components.

The coefficients of the models show the impact of each component, which is important for intervention strategies. Both models have almost identical accuracies and results which indicate that from the 5 coefficients, the strongest coefficient is an inverse (negative) relationship with players' weight indicating that heavy athletes have a shorter recovery time. Overall, with satisfactory metrics, the models skillfully predict 'recovery_time' and contribute to personalized care, supporting injury prevention, and athlete longevity.

In summary, the linear regression along with the alphas of ridge regression fine-tunes the 'recovery_time' prediction model, balances both complexity and feasibility, and contributes to the development of injury prevention strategies in sports.

Approach: Classification models (Logistic Regression):

The purpose of the Logistic Regression analysis in this work is to predict the probability of injury, which is a dichotomous problem. Logistic regression is particularly well suited for such problems because it predicts probability by fitting data to a logistic curve. Model effectiveness has been quantified with accuracy, F1 scores, recall, and accuracy parameters, and results are summarized in the confusion matrix and classification report. True/false Negative/positive is essential in understanding the model's performance.

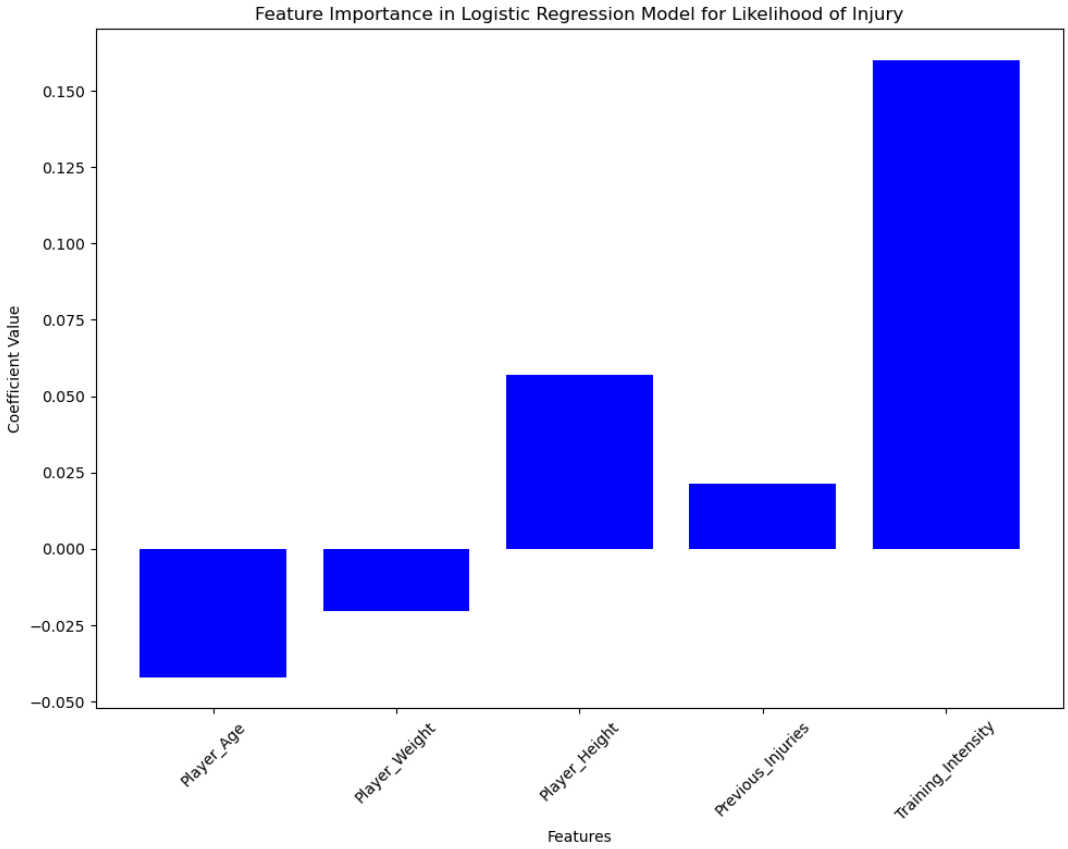
This model is used for predicting a binary outcome, which in this project is 'Likelihood_of_Injury'. Thus, this model is used to evaluate the likelihood of injury based on certain athlete characteristics.

Results: Logistic Regression Insights:

```
Logistic Regression - Accuracy: 0.5650, F1 Score: 0.5348, Recall: 0.4762, Precision: 0.6098
Confusion Matrix:
 [[63 32]
 [55 50]]
Classification Report:
              precision    recall  f1-score   support

           0       0.53      0.66      0.59        95
           1       0.61      0.48      0.53       105

    accuracy                           0.56       200
   macro avg       0.57      0.57      0.56       200
weighted avg       0.57      0.56      0.56       200
```



The logistic regression model reports an accuracy of 0.5650, an F1 score of 0.5348, a recall of 0.4762, and a precision of 0.6898. These indicate reasonably accurate but limited recall, indicating a lack of actual damage. At 0.57, the equilibrium accuracy is slightly above chance.

As mentioned, logistic regression yielded in ineffective data for accuracy, F1 score, and Recall. Even after consulting the professor attempting regularization (L1/L2) and checking for overfitting, the results still yielded just above 50%. After looking into other projects that used this same dataset, they also reported these readings using different classification methods like LGBMClassifier, AdaBoostClassifier, ExtraTreesClassifier, NuSVC, ExtraTreeClassifier. Thus, it can be safe to conclude that this is an intrinsic characteristic of this dataset and something worth addressing.

The model predicts ‘no injury’ to be more reliable than ‘injury’, as indicated by more false negatives in the confusion matrix. Feature significance analysis reveals that ‘training intensity’ is a strong predictor, with ‘past injuries’ also positive but to a lesser extent, and ‘player age’ correlated inversely with injury probability.

The model also identified 'training intensity' as a significant predictor of injury potential results, indicating that athletes with high-intensity training require more attention to prevent potential injuries.

In conclusion, the model is insightful but requires modification for accuracy. This information is important for improving training strategies and injury prevention.

Note:

For the 3 models mentioned in this project, 80% training/20% testing split, standardization, regularization, performance metrics display, and select Lambda values were applied to the appropriate models to yield the most clear and concise results.

Conclusion:

This project, "Predictive Analytics for Injury Prevention: Machine Learning Approaches in Competitive Sports," has been able to identify specific types of athletes that are more injury-prone, addressing the primary goal of identifying preventive care and attention to specific high-risk individuals vulnerable to injury.

EDA revealed that the likelihood of injury does not depend solely on individual factors such as age or weight but is composed of many variables. Notably, athletes with a history of previous injury and higher-intensity workouts emerged as a group at high risk for future injury.

Supervised learning models such as linear and ridge regression provided a predictive basis for 'recovery_time', an important factor in athlete management after injury. Specifically, both models suggested that Player_Weight had the most significant impact from all other features in that dataset indicating that the negative correlation showed how heavier athletes trend towards quicker recovery. The logistic regression results revealed a need for model refinement for higher accuracies but provided a predictive basis for 'Likelihood_of_Injury', an important insight into athlete management after injury. This model indicated that athletes with higher-intensity training tend to have a higher likelihood of injury. This information can result in future projects explaining why this may be the case.

By identifying these key risk factors and athlete issues, the project results provide a means to prioritize and individualize athlete care. High-risk athletes can now be targeted for increased monitoring, customized training modifications, and specific rehabilitation programs to reduce the risk of injury. This targeted method aims not only to improve the care of individual athletes but also contributes to the broader goal of expanding sporting careers by reducing the impact of injury and its incidence in sports.

Acknowledgements:

References:

1. https://jeo-esska.springeropen.com/articles/10.1186/s40634-021-00346-x
2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10613321/

Dataset:
https://www.kaggle.com/datasets/mrsimple07/injury-prediction-dataset