Sports Injury Risk Prediction

Saanvi R Prabhu (1RVU22CSE136)

Chinmay Umesh (1RVU22CSE042)

Deekshith R Prabhu (1RVU22CSE046)

Durga Prasad C Reddy (1RVU22CSE051)

ABSTRACT

Sports injuries have been a concern in the world of sports for a long time. Injuries can affect an individual player’s career and even the team performance if it is severe. The project tries to use data to analyze different factors that cause injuries and how to prevent them. The project proposes a Machine Learning model based solution to predict injury risk factors and estimate the recovery time in case of injury. The dataset taken is from American Football as it is a contact sport and injuries are common. Predicting injuries before they occur can have a huge impact on a team’s performance, planning and strategy as they can plan beforehand and try to rest the players at a risk of injury. The project also tells the body part at risk of injury and gives an estimated time of recovery which can help teams and players to take preventive measures. For access to the project's codebase, please refer to this.

INTRODUCTION

Sports industry is characterized by huge investments aimed at boosting the performance of players and reducing injuries. Multiple factors contribute to the possibility of injury including, but not limited to playing conditions (such as field type, weather conditions, and temperature), individual movements, positions, turf and speed. As there are numerous factors that can cause serious injuries, it becomes extremely important to understand what those factors are and in what ways they can be prevented. As the demand for Data Science and Machine Learning continues to grow, the market is also growing. More and more teams are investing huge amounts of money to try and approach the game in a statistical method giving a lot of scope to the application of Machine Learning in the field of sports.

Our project analyzes the occurrence of injuries relying on a dataset that encompasses these different aspects. It then becomes possible to predict if an injury would occur by employing a classifier. Furthermore, our regression model predicts how long it will take before an injury is healed or fully recovered. Besides, we have multi-class classification models predicting probable body parts that can be injured most likely during a sport game or practice. Through these predictive models, we aim to provide valuable insights for injury prevention and player safety in the sports industry.

RELATED WORK

A journal published by Hans Van Eetvelde et al. titled “Machine learning methods in sport injury prediction and prevention: a systematic review” searched PubMed for studies on Machine Learning in sports injury prediction and prevention, then had two reviewers independently screen and assess the articles for eligibility and bias. They evaluated methodological quality using the Newcastle–Ottawa Scale and graded study quality with the GRADE methodology. In essence, they rigorously analyzed existing research to understand how Machine Learning contributes to sports safety. [1]

Another journal published by Alessio Rossi et al. titled “A Narrative Review for a Machine Learning Application in Sports: An Example Based on Injury Forecasting in Soccer" talks about prediction of injuries and player performance in soccer. The researchers employed a narrative review to examine statistical methodologies and machine learning models for sports injury prediction. They categorized models into "real" algorithms like decision trees and baseline models for comparison. Methodologies included validation through techniques like 10-fold cross-validation and an evolutive scenario approach. Data preprocessing involved sampling (e.g., SMOTE) and feature selection (RFECV), with hyper-parameter tuning via Grid Search and Random Search. Model interpretation emphasized explainable AI methods like SHAP. Evaluation metrics such as precision, recall, and AUC were used to assess prediction performance. [2]

“Modeling the Risk of Team Sport Injuries: A Narrative Review of Different Statistical Approaches” is a journal by Joshua D. Ruddy et al. which discusses reducing the risk of injuries in team sports. To clarify statistical methodologies used to identify risk factors for injury in sports, the researchers performed a narrative review. The distinction between association and prediction was well illustrated in their work, which also noted how important it is to exercise caution when interpreting data using statistical approaches. It further showed that injuries are analyzed through binary classification, relative risks, odds ratios, sensitivity, and specificity among others. Their aim in employing these methods was to improve the ability of researchers and practitioners to accurately determine risk factors for injuries thus helping them develop better strategies for preventing injuries during sports activities. [3]

METHODOLOGY

For the Sports Injury Prediction project, the methodology involves three main tasks: a classification problem, a regression problem, and a multi-class classification problem.

1. Task 1: Data Cleaning

Aggregation of Data: Due to the dataset's high bias (267,005 playsKeys and only 105 injuries), aggregation was done using the GamelD (number = 5712) instead of playKey. This was done in order to solve the imbalance between number of plays and number of injuries in the dataset.

Resampling:

The class distribution is biased towards no injury; for this reason, data was resampled before being input to the classifier.

2. Task 2: Multi-Class Classification Problem

Data Preparation:

Injury Record.csv was combined with PlayList.csv using inner join on the 'GamelD' column. Null values were handled during the join.

Unnecessary columns were removed, and the data was oversampled to increase the number of rows.

Body Part Prediction:

Knee, Foot, Ankle, Heel, and Toes were the body parts considered, and the part most likely to be injured was predicted using Logistic Regression, Decision Tree, and Multi-Layer Perceptron (MLP) algorithms.This methodology provides a detailed approach to address each task in the project, ensuring clarity and reproducibility.

3. Task 3: Regression Problem

Data Modification:

The Injury column was modified to contain the sum of values, ranging from 0 to 4 where the 0 value implies no injury, and a non-zero value indicates an injury.

A new column was created to indicate the days of rest a player has to take in case of injury, with random values generated using Gaussian distribution.

Models Used:

Linear Regression, Support Vector Machine, Random Forest Regressor.

EXPERIMENTAL DETAILS

Different steps in injury prediction

The first step involves predicting the body part typically damaged by various attributes in which player position, game conditions, and environmental factors fall. This could be used for developing strategies geared towards injury prevention thus enabling teams to come up with specialized regiments or adjust their game settings so as to prevent predictable injuries. Once found, those specific areas which are highly sensitive should assist in generating preventive methods against them that are usually aimed at enhancing the safety of the players.

The second step is to predict how many days off an injured player will need. This projection is crucial for guiding the healing process leading to full recovery and guaranteeing outstanding performances upon resumption of matches. Team managers who make precise forecasts regarding the time they expect their players back may undertake preparatory steps like arranging rehabilitation sessions that would hinder any further damage.

Dataset details

A dataset from American Football is chosen as the sport is known for frequent injuries. The dataset is split into two parts: Injury Records and Playlist. The injury record file in .csv format contains information on 105 lower-limb injuries that occurred during regular season games over the two seasons. Injuries can be linked to specific records in a player's history using the PlayerKey, GamelD, and PlayKey fields. The playlist file contains the details for the 267,005 player-plays that make up the dataset. Each player is indexed by PlayerKey, GamelD, and PlayKey fields. Details about the game and play include the player's assigned roster position, stadium type,field type, weather, play type, position for the play, and position group.

Analysis of PlayList :

1)There are 250 players in the dataset.

2)There are 5712 games in the dataset.

3)There are 267005 plays in the dataset.

Analysis of Injury Record:

1)There are 105 injuries records in total

2)100 unique players have been injured and hence cases of multiple injuries to the same player are present.

3)28 Playkey values are missing.

Model architecture

1. **Multi-Class Classification Model for Predicting Body Part Injuries:**

Features: Merged injury and playlist data by the 'GameID' column.

Body Parts: Knee, Foot, Ankle, Heel, Toes.

Algorithms: Logistic Regression, Decision Tree, Multi-Layer Perceptron.

Purpose: Based on other attributes, such as the position of the player, conditions of the game, and environmental factors, this model determines which body part is more likely to be injured. The model is able to identify which parts of the body are more prone to injury under certain circumstances.

In preprocessing, the dataset is resampled in order to handle the class imbalance problem in the dataset. Then, both models are trained with the preprocessed data. For this dataset, we use metrics such as accuracy, precision, recall, and F1-score to compare the performance of each model for each task.

2. **Regression Model for Predicting Days of Rest:**

Features: RoasterPosition, PlayerDay, Player Game, Stadium Type, Player Position, Field Type, Temperature, Weather, Play Type.

Algorithms: Linear Regression, Support Vector Machine, Random Forest Regressor.

Purpose: Based on various characteristics of players, games, and environments, it forecasts the number of days of rest the player needs after an injury.

Injury Prediction

Predicting Body Part:

Logistic Regression: The non-numeric attributes were converted to binary and the total number of features fed into the model were 57. Thus the data becomes too sparse and highly multi-dimensional for a logistic regressor to handle. The model fails to converge. The model fails to optimally fit this non-linearly separable data.

Decision Tree Classifier: It showed better performance compared to Logistic Regression, probably because it is able to learn a more complex relationship present in the data. However, it overfitted, which created the difference in performance while training and testing the data. These are indications of how crucial regularization techniques and ensemble methods are in mitigating overfitting, hence improving generalization performance. Moreover, better performance by the Decision Tree with unseen data could also be improved by fine-tuning hyperparameters like maximum tree depth or minimum sample split.

Multi Layer Perceptron: Our Multi Layer Perceptron performed the best with the ROC curve. This is because of the use of non-linear activation functions that allows the model to capture complex non-linear relations in the data**.**

Finding Recovery Days:

Linear Regression: We used Linear Regression as a baseline for our regression problem. The mean squared error was highest for this model. As the data is highly dimensional, a linear regressor does not yield satisfactory results.

Support Vector Regressor: Essentially it creates a single optimal hyperplane based on the support vectors thus producing a better result compared to

Linear Regression. However the problem is better addressed by ensemble learning classifiers.

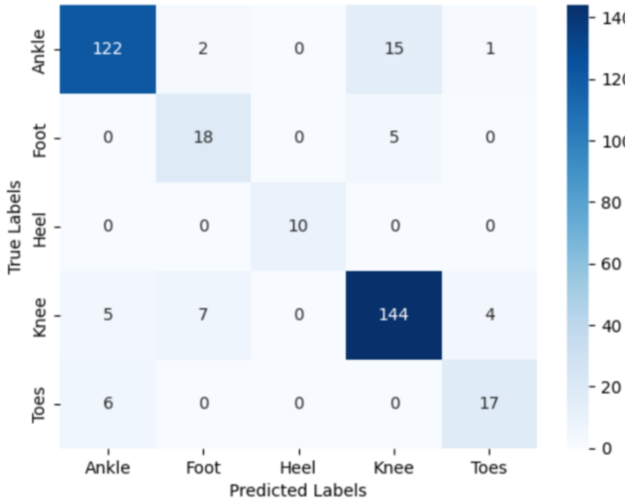
Random Forest Regressor: This ensemble learning algorithm helped us in achieving the best results that gave minimum error.

RESULT

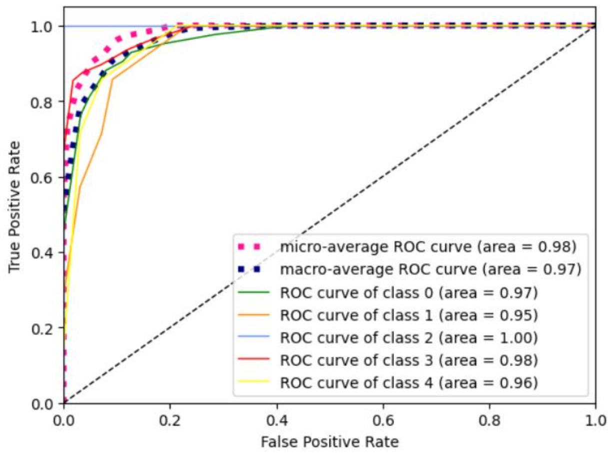
The Decision Tree algorithm gave the best result for finding body parts at the risk of injury. This is because the Decision Tree algorithm has the ability to capture complex relations.

-	Precision	Recall	F score
Micro	0.8735	0.8735	0.8735
Macro	0.8469	0.8586	0.8516

The above table gives us micro and macro values in decision tree

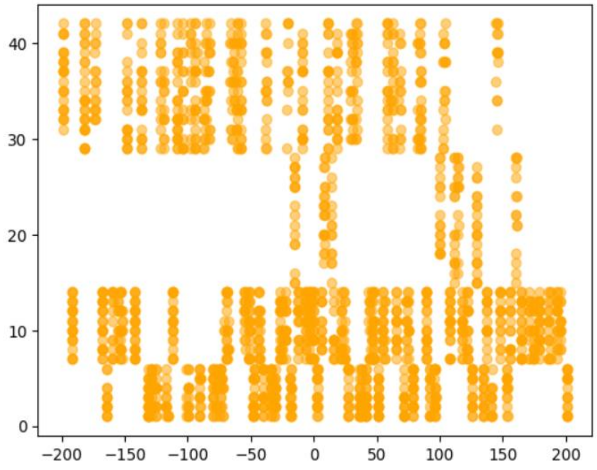


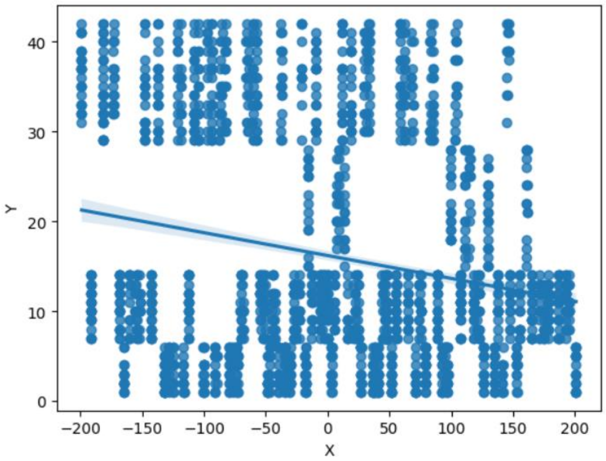
The above image gives us the heatmap predicting the body part at highest risk of an injury.



The above ROC curve gives us True Positive Rate v/s False Positive Rate for Decision Tree

To find out the number of recovery days required, Random Forest Regressor was used as it gave the best performance compared to other Machine Learning models.





The above scatter plot predicts the amount of days required to recover

CONCLUSION

The project used real player injury data, which was highly imbalanced. We applied a number of pre-processing techniques for removal of NA values and used a correlation matrix to drop highly correlated features, applied dimensionality reduction for faster convergence before feeding it to the models. Moreover, to tackle the imbalanced dataset, we applied various re-sampling techniques and did a certain number of tweaks while combining the datasets. Various baseline models like Logistic Regression, and Decision Trees were tested with different parameters.

An important step was creating a model which tells us which body parts are more prone to getting injured, given the set of features related to the field, weather, and player's health. This was a multi-class problem. For the baseline of this problem, logistic regression and decision tree were used. Finally, a multi-layer perceptron classifier was used to capture the complex relations in the data. We observed that the MLP Classifier gave most of the satisfactory results for this dataset.

Last but not least, the final phase of this project was to find the amount of days required for recovery. For the Regression problem, various models such as Linear Regression, SVM, and Random Forest Regressor were used to gain minimum MSE loss. Random Forest Regressor along with optimal model parameters performed the best among the tested models.

The findings of the project are key to helping teams identify players at the risk of injury and prevent it from happening. A few possible limitations are imbalanced/enormous datasets which can affect model performance and using complex models that can lead to overfitting. When applying the model to predict injuries, these limitations should be kept in mind to ensure a good model performance.

References

[1] H. M. L. D. C. S. R. T. T. Van Eetvelde, "Machine learning methods in sport injury prediction and prevention: a systematic review," *Journal of Experimental Orthopaedics,* vol. 8, no. 1, 2021.

[2] L. P. P. C. Alessio Rossi, "A Narrative Review for a Machine Learning Application in Sports: An Example Based on Injury Forecasting in Soccer," *Artificial Intelligence in Sports Injury and Injury Prevention,* 2021.

[3] S. J. C. R. W. M. D. W. R. G. T. D. A. O. Joshua D. Ruddy, 2019.