HDSC Summer ’22 Capstone Project [Neural Network]:

Injury Predictions for Athletes

25/10/2022

1.0 INTRODUCTION

Generally, in sports, injuries are disliked but unavoidably common. It doesn't matter whether a sport is played individually or as a team, injuries can cause significant damage physically, mentally, psychologically, financially and in extreme cases severe health issues and death. It is therefore essential to understand the key factors that contribute to players fitness and performance which include the health level of athletes, emotional status, magnitude of exercise and its physical intensity.

The ability to forecast the manifestation of a provocative injury event is difficult due to the factors related to the individual athlete and external factors and unforeseeable circumstances such as contact with other players. Therefore, we have focused our model to predicting injuries based on the training data gathered.

2.0 **PROBLEM STATEMENT**

In sports, injuries are common things that athletes sustain. Machine learning (ML) approaches could be used to improve injury prediction and allow development of proper approaches to injury prevention.

3.0 **AIMS AND OBJECTIVES**

The aim of our study was therefore to develop a machine learning model that can predict injuries based on the training regimen of an athlete. This can be used to make injury prevention approaches easier to develop.

4.0 Flow Process

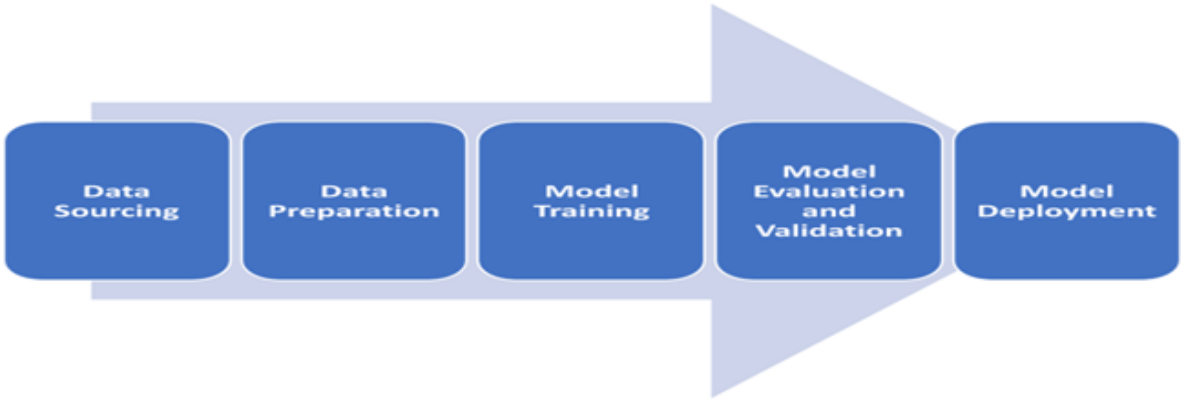

Figure 1: Project flow process

4.1 Data Source

The data used in this project was obtained from the Kaggle website. There were two approaches in data collection; daily approach which collected data on the daily training schedule of athletes for 7 days and the weekly approach which collected weekly data on the training schedule of athletes.

4.2 **Data Preparation**

Preliminary explorations that involved checking for null values and some visualizations were performed on the datasets. It was observed that there were no missing values on both weekly and daily dataset. Hence, the data is reliable for further visualisation and analysis purposes.

4.3 Exploratory Data Analysis (EDA)

4.3.1 EDA **on the Daily Approach**

There was a total of 42766 records and 73 columns. No null values were recorded in any of the columns. The daily approach contained information about athletes for every 7 days, and there are 10 unique features that define each session of an athlete. These features were repeated 7 times resulting in 70 columns. The columns were:

1. Number of sessions an athlete trained
2. Total distance
3. Sum of distance in Z3-Z4
4. Sum of distance in Z5, T1 and T2
5. Distance sprinting
6. Number of strength sessions
7. Hours alternative training
8. Perceived exertion
9. Perceived training success
10. Perceived recovery
11. Athlete ID

12. Injury
13. Date

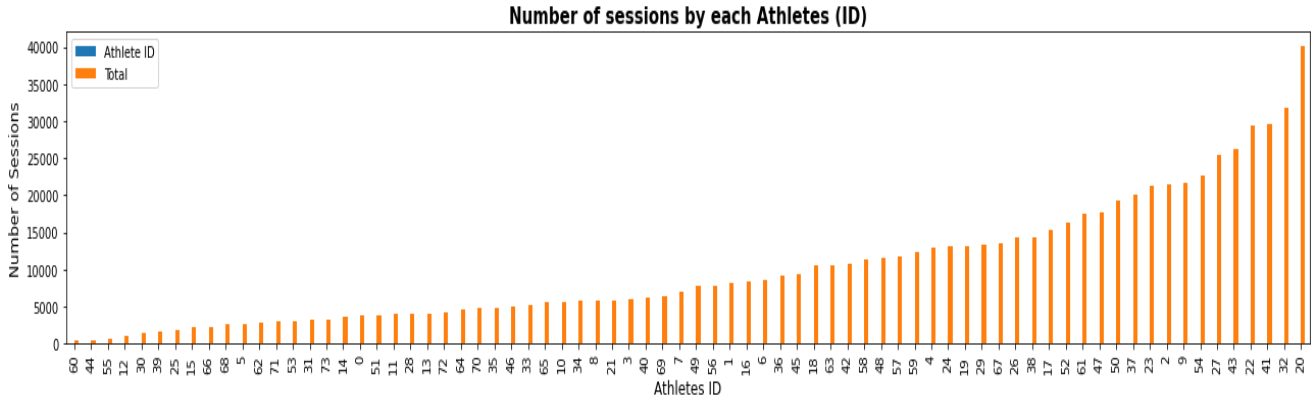

Figure 2: Plot of the total sessions for each Athlete.

In the figure above, athlete 60 had the least sessions in a week as opposed to athlete 20 with the most sessions.

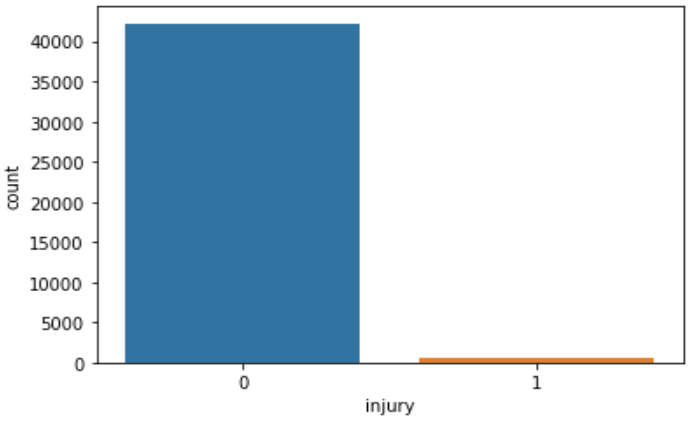

Figure 3: Count plot of occurrence of injury among athletes

In the above figure, there are a total of 42,183 times injury was not recorded among athletes, and 583 times when athletes got injured.

4.3.2 EDA on the Weekly Approach

Similar to the daily approach, there were a total of 72 columns and 42798 entries with no null values. There were 28 unique columns with 22 repeating 3 times denoting 3 weeks of training before the event of injury. The columns are:

1. Number of sessions
2. Number of rest days
3. Total distance
4. Max distance (1 day)
5. Max distance Z3-5, T1-2

6. Sessions in Z5-T1-T2
7. Sessions in Z3 or faster
8. Total distance Z3-4
9. Max distance Z3-4 (1 day)
10. Total distance Z5-T1-T2
11. Max distance Z5-T1-T2
12. Hours alternative training
13. Number of strength training sessions
14. Average exertion
15. Minimum exertion
16. Maximum exertion
17. Average training success
18. Minimum training success
19. Maximum training success
20. Average recovery
21. Minimum recovery
22. Maximum recovery
23. Total distance week1/week2
24. Total distance week0/week1
25. Total distance week0/week2
26. Athlete ID
27. Injury
28. Date

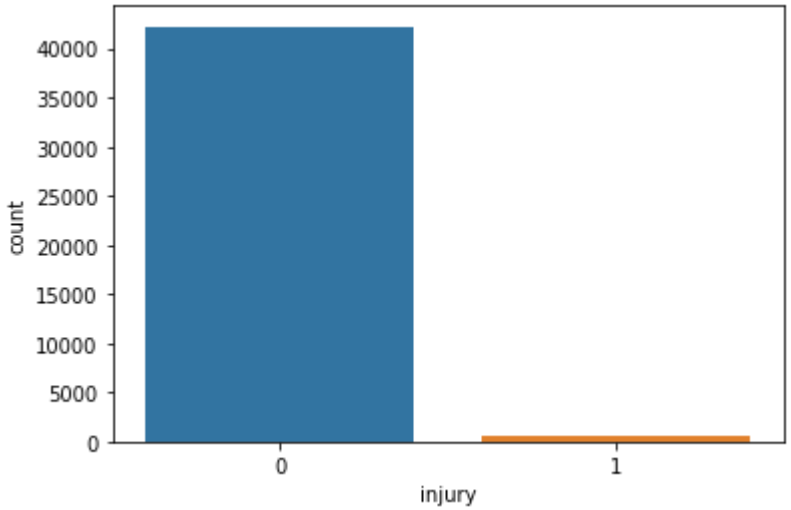

Figure 4: Count of injured and non-injured athletes from the weekly approach

The figure shows the count of athletes that were injured or not injured in the weekly approach. In the entire dataset, there are a total of 42223 athletes who were not injured, and 575 athletes got injured.

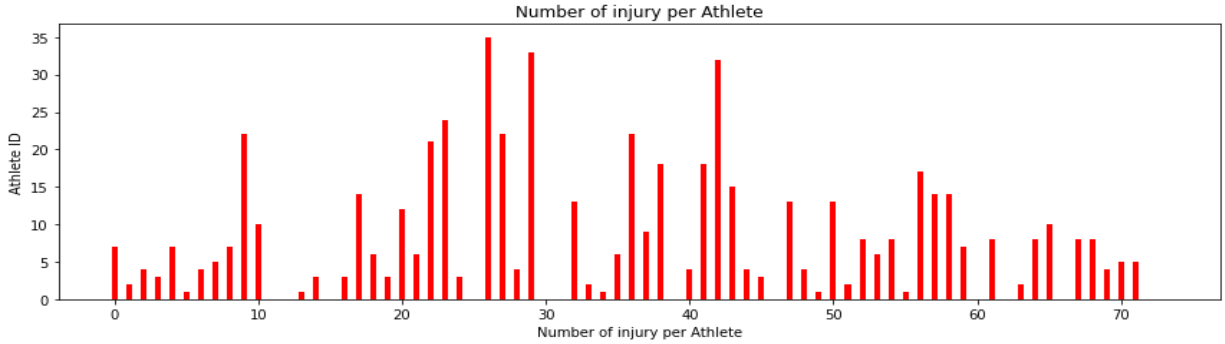

Figure 5: Bar Plot of the number of injuries per athlete.

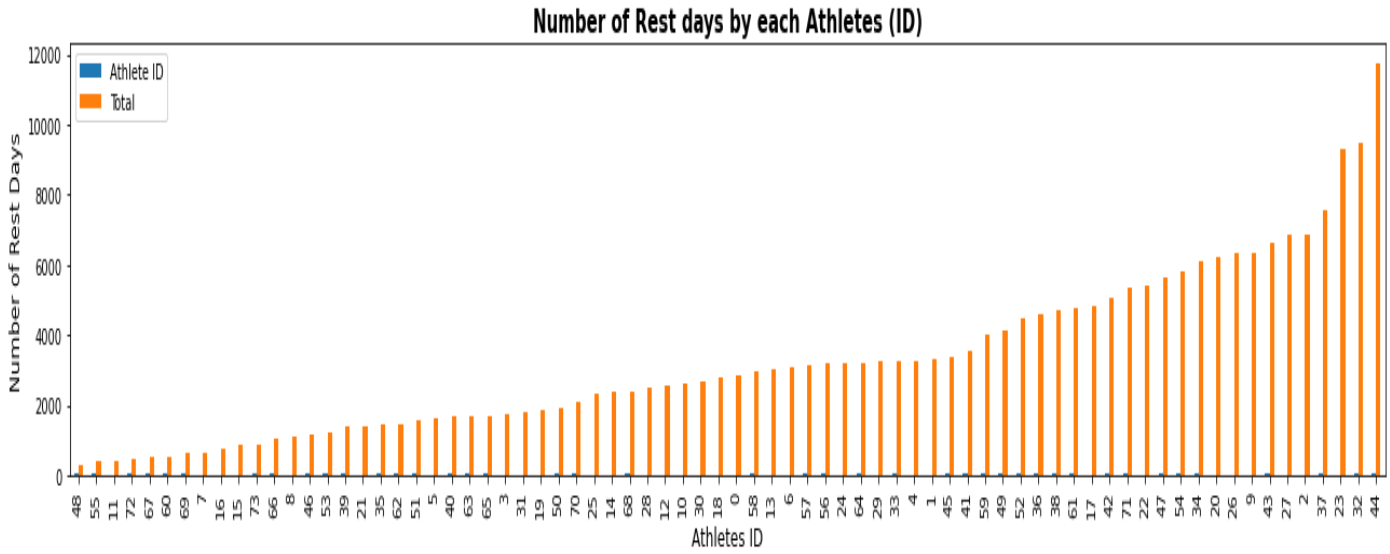

Figure 6: Total number of rest days by athlete ID

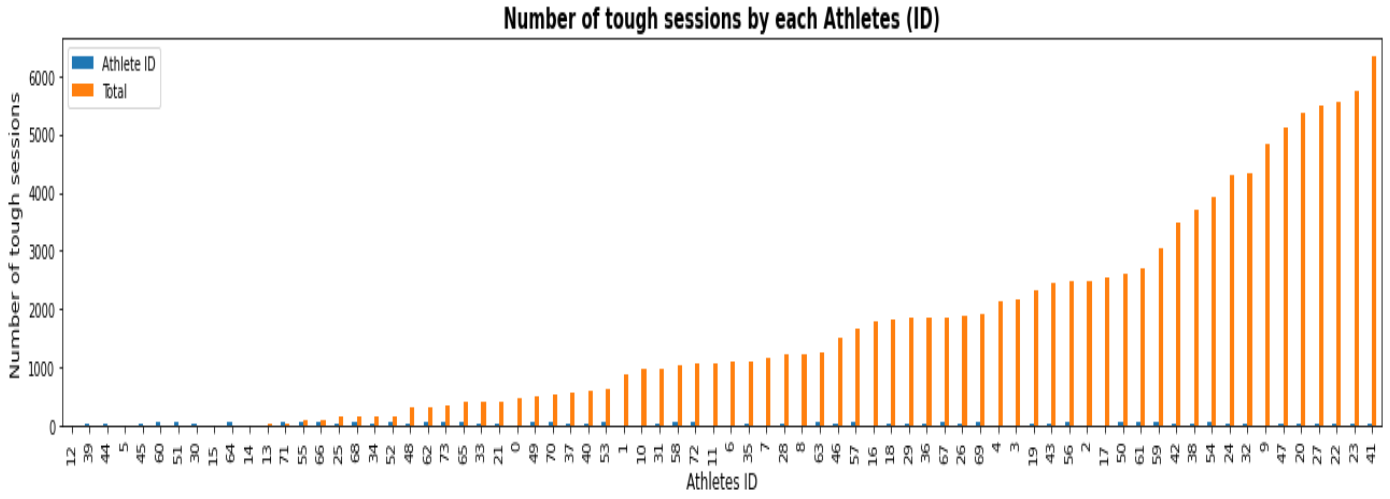

Figure 7: Total number of tough sessions by athlete ID.

5.0 Feature Engineering or Feature Selection

This process involved selecting and transforming the most important or relevant features from the dataset that will be used to create the predictive model. In this project, the features were selected based on importance. In the daily dataset ten (10) features were used in training the model. The features include the first 10 features in the daily approach repeating 7 times to give a total of 70 features. Similarly, in the weekly approach, all features except ‘Athlete ID’, ‘Injury’ and ‘Date’ were selected. The weekly features amounted to 69 features with 22 repeating 3 times.

6.0 Model Training

For this process, various models were trained and evaluated using different metrics. The four (4) algorithms employed in this project were Deep learning Neural Network, and K-Nearest Neighbours, XGBoost. The models were evaluated based on accuracy, recall, precision, F1 score and roc_auc. The imbalance in the data set was catered for by employing the Synthetic Minority Oversampling Technique (SMOTE).

6.1 **Performance on Test Data**

The performance of the various models were tested using metrics including accuracy, recall, precision, roc_auc and F1 score.

6.1.1 **Deep Learning Neural Network (DNN)**

A

	precision	recall	f1-score	support
0	0.85	0.66	0.74	12667
1	0.72	0.88	0.79	12667
accuracy			0.77	25334
macro avg	0.78	0.77	0.77	25334
weighted avg	0.78	0.77	0.77	25334

B

	precision	recall	f1-score	support
0	0.93	0.84	0.88	12655
1	0.86	0.93	0.89	12655
accuracy			0.89	25310
macro avg	0.89	0.89	0.89	25310
weighted avg	0.89	0.89	0.89	25310

Figure 8: Performance of DNN model for the weekly and daily approaches

Key:

A. **Weekly approach**
B. **Daily approach**

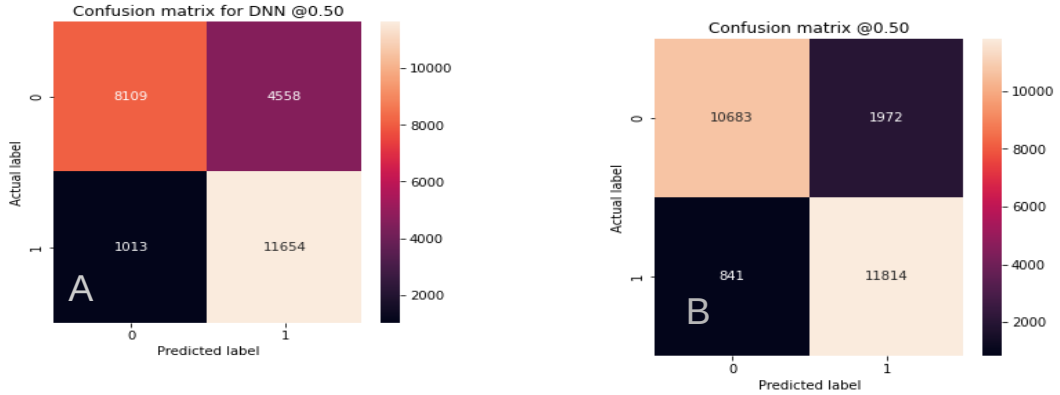

Figure 9: Confusion matrix of the weekly and daily approaches

Key:

A. **Weekly approach**
B. **Daily approach**

6.1.2 Extreme Gradient Boosting (XGBoost)

A

	precision	recall	f1-score	support
0	0.99	0.86	0.92	12667
1	0.03	0.32	0.06	173
accuracy			0.85	12840
macro avg	0.51	0.59	0.49	12840
weighted avg	0.98	0.85	0.91	12840

B

	precision	recall	f1-score	support
0	0.99	0.81	0.89	12655
1	0.02	0.29	0.04	175
accuracy			0.81	12830
macro avg	0.50	0.55	0.47	12830
weighted avg	0.97	0.81	0.88	12830

Figure 10: Performance of xgboost model for the weekly and daily approaches

Key:

A. **Weekly approach**
B. **Daily approach**

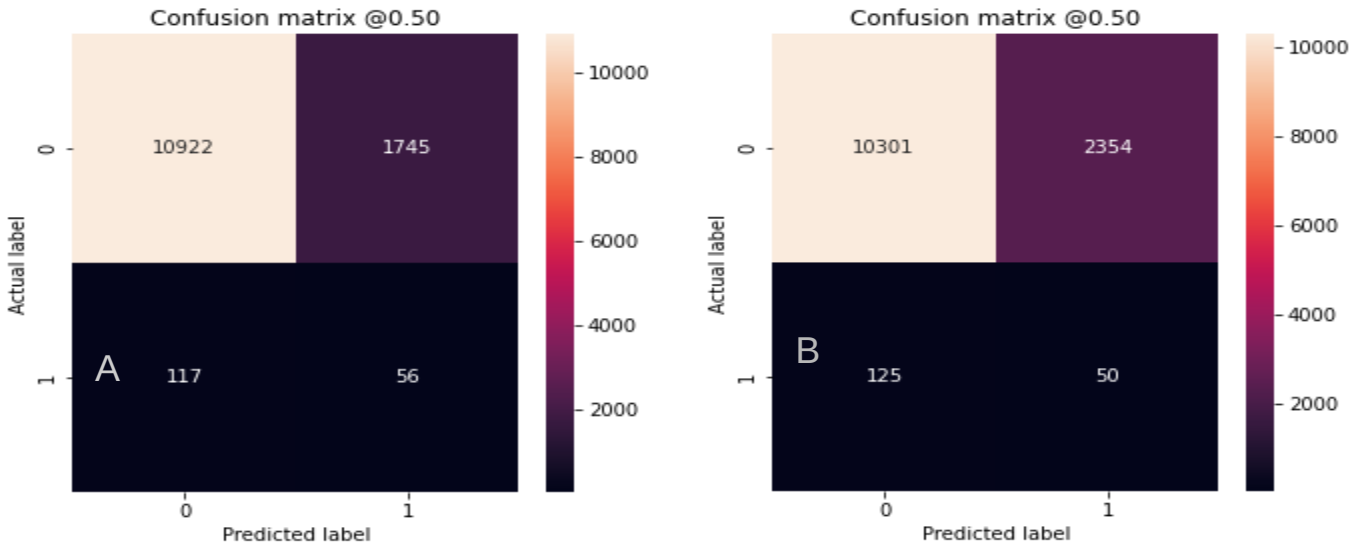

Figure 11: Confusion matrix of the weekly and daily approaches

Key:

A. **Weekly approach**
B. **Daily approach**

6.1.3 Model 3 - K - Nearest Neighbours

A

	precision	recall	f1-score	support
0	0.99	0.96	0.97	12667
1	0.96	0.99	0.97	12667
accuracy			0.97	25334
macro avg	0.97	0.97	0.97	25334
weighted avg	0.97	0.97	0.97	25334

B

	precision	recall	f1-score	support
0	0.99	0.94	0.97	12655
1	0.94	0.99	0.97	12655
accuracy			0.97	25310
macro avg	0.97	0.97	0.97	25310
weighted avg	0.97	0.97	0.97	25310

Figure 12: Performance of KNN model for the weekly and daily approaches

Key:

A. Weekly approach
B. Daily approach

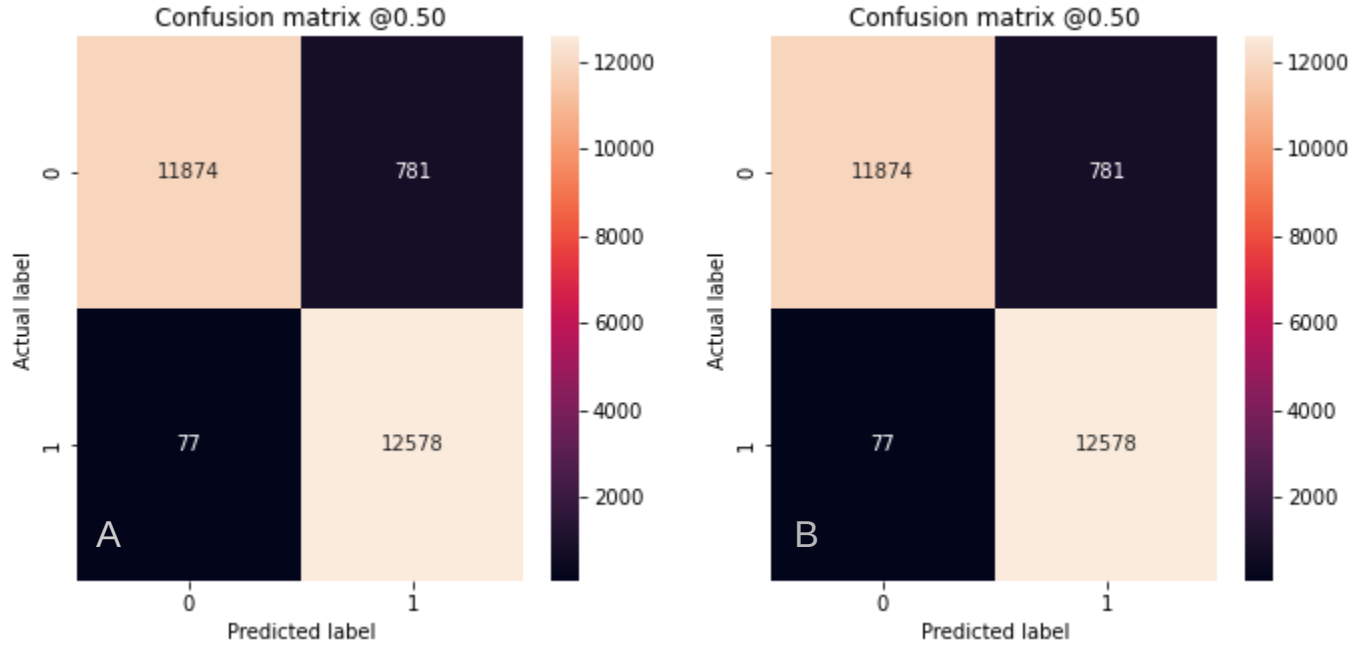

Figure 13: Confusion matrix of the weekly and daily approaches

Key:

A. Weekly approach
B. Daily approach

Table 1: Summary of performance of all models

Approach	Model	Type 1 error	Type 2 error	Score accuracy	Precision	Recall	F1-Score
Day	DNN	841	1972	0.89	0.86	0.93	0.89
	KNN	103	769	0.97	0.94	0.99	0.97
	XGboost	125	2354	0.81	0.02	0.29	0.04
	DNN	1488	4347	0.79	0.72	0.88	0.77
Week	KNN	167	506	0.97	0.96	0.99	0.97
	XGboost	117	1745	0.85	0.03	0.32	0.06

7.0 Machine Learning Model Deployment

A web application was built to illustrate the daily injury prediction from runners training routine or schedule the KNN model. The web application takes in imputations of 10 training activities or routines of runners over a course of 7 days and predicts if a runner will have an injury or not when following such plan/routine. The application was created and deployed on streamlit cloud, ready for consumption.

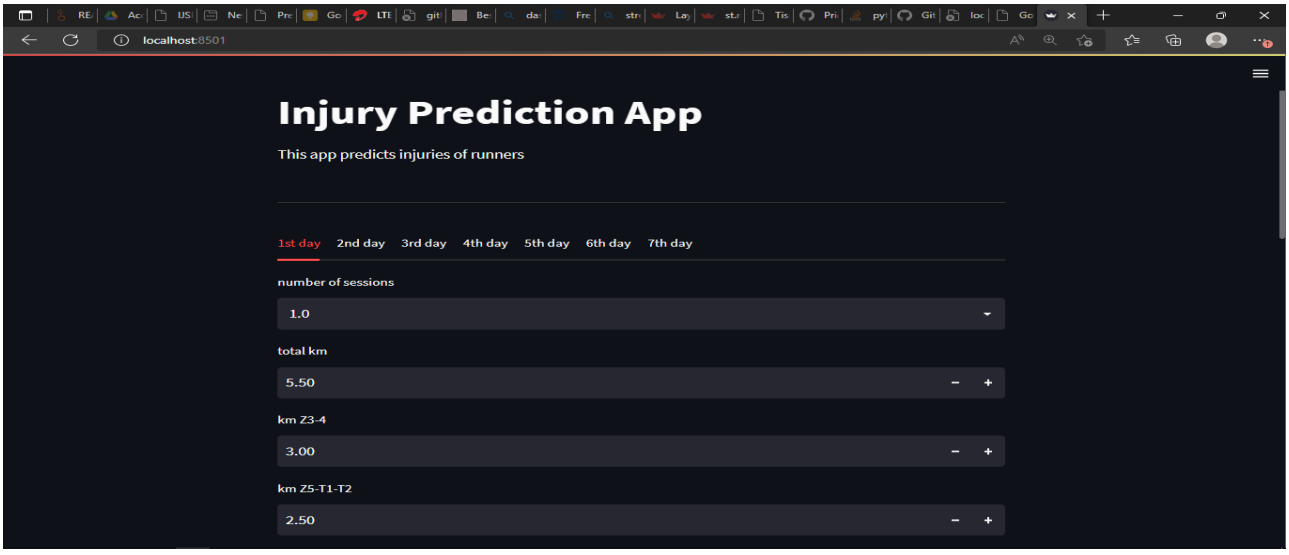

Figure 15a: Deployment dashboard

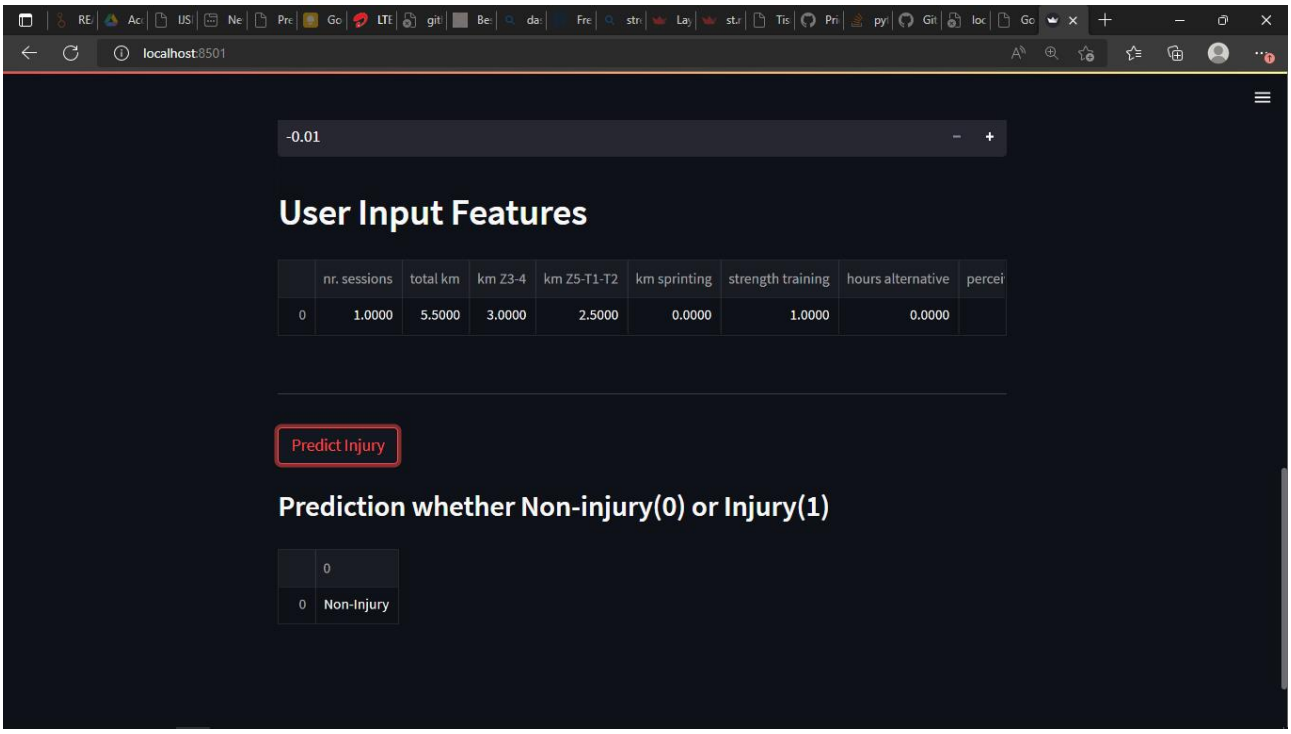

Figure 15b: Deployment dashboard

8.0 **Limitations**

The entire process of working on the machine learning models was not without challenges. The most relevant ones being;

- Huge Data Imbalance: There was a huge imbalance in the data, especially the target variable. This was overcome by employing the Synthetic Minority Oversampling Technique (SMOTE) where the minority class is oversampled synthetically. This creates a balanced dataset used in training the models.
- Collaboration: This posed a challenge because the group members were stretched across the globe with different time zones and varying schedules. This made getting an harmonized model difficult. However, we bypassed the constraint by syncing code and models using git as a collaboration tool.
- Time: Interns involved in this project were always faced with a tight schedule due to other engagements. With the clock ticking, we had to spend extra hours just so the project could be completed in time.

9.0 **Conclusion**

In conclusion, the project prediction was based on three models on the weekly approach dataset namely, the DNN, K-NN, and XGBoost algorithm while on the daily approach, the project implemented four models, namely, DNN, K-NN and XGBoost. Thus, the best-performing model on the daily dataset was K-NN with an accuracy of 97%. The weekly dataset on the other hand, k-NN was the best model with about 97% accuracy (f1-score). Furthermore, the feature with the highest influence on injury status was the 'sum of max km' with about 34.5 to 4.7 km per day according to explorations on the data.

Finally, to demonstrate consumption of the developed model, the generated k-NN model was deployed using a streamlit cloud framework. The web application uses 10 training activities over 7 days. Thus, the overall system was a success as the complete system was able to provide the required prediction according to the 70 different features. The system which is invariably a model for future expansion can be provided with other features for robustness and more functionality enhancement. Features such as high-level security, Artificial Intelligent (AI) model, and Portable mobile application software to enhance accessibility.

Team members

1. Egbaidomeh Daniel Ehiz
2. Vishal Garg
3. Solomon Nkwor
4. Abdullahi Isa Ahmed
5. Muhammed Balogun
6. Daniel Kurui
7. Joshua Owoicho
8. Abdulhamid Ibrahim
9. Abdullateef Ogundipe
10. Oluwatomiwa Adeleke
11. Sharon Omovie