The Application of Machine Learning on the Injury Prediction of Soccer Players

Davronbek Malikov[1], Jaeho.Kim [2]

[1]Gyeongsang National University, Jinju-si , 53828 , South Korea
[2]Gyeongsang National University, Jinju-si , 53828 , South Korea

Abstract

Soccer player lives with a high risk of injury since soccer is one of the sports activities with relatively high injury incidence compared with many other sports. Injuries can be a huge influence not only player's career and financial situation but also it is the reason for soccer changing the coach's game tactics as well as for directors of clubs have to find a new player. In order to reduce the risk of getting injured by predicting the probability of soccer players' injuries for the new season, we conduct research in this paper. In this paper, we propose parameters that are connected each other and we collected using a non- technical way while most of the recent research provides technical ways such as GPS tracking technology or wearing devices. Moreover, we provide the accuracy of injury prediction and estimation of recovery time by using supervised classification machine learning models.

Keywords

non-technical data collecting, data analysis, soccer injury prediction, machine learning for a soccer injury

1. Introduction

Soccer players face numerous challenges throughout their careers, but one of the most significant is the risk of injury. Injuries can result in players having to take extended breaks from the game, incurring substantial rehabilitation costs, and in some cases, even ending their careers prematurely. Therefore, developing effective injury prevention and management strategies is crucial to ensure the long-term health and success of soccer players..In this paper, we propose an application of machine learning the predicting the likelihood that injury cases among professional soccer players while playing. Machine learning Naive Bayes model will be used in order to identify the player's risk of injury for the next soccer season and we check it by using training and testing data collected from professional sports websites. A purposing model will be created based on the connection of variables that play an important role in this research. In addition to the benefits for soccer players and coaches, our analysis and classification report can also have a positive impact on the overall performance and success of the team. By being able to predict the likelihood of injury, coaches can make informed decisions about player selection and substitution, adjust training regimes and playing tactics, and ultimately reduce the number of injuries sustained by their players.

2. Related Work

The training data and Machine Learning models are becoming a fundamental part of professional sports, especially in predicting the risk of sports injury. Numerous academic papers have been published on the topic of machine learning applications in the prediction of injuries in sports, including soccer.[13,14.]For example, the relationship between injury risk and the so-called monotony was defined in basketball and moreover, the ratio and the standard deviation of the session load recorded in the past week [1]. In the case of skating the product of the perceived exertion and the duration of the training session is proposed where the training session is measured by the session load. The skater can be caused by 'overtraining syndrome' and is considered a risk of injury when the session load is out of balance and the skater's ability to fully recover,before the next session [2]. Furthermore, the application of machine learning models has been playing a great role in order to conduct of research on predicting the risk of injury in Soccer and American Football where the linear regression model is used [3,4]. Machine learning is important for automatically learning and helps to improve small efforts of programming. Especially, in the case of performing tasks that are very complex and difficult for humans in between fixed time then the machine learning model is very useful [5,6]. Another example of applying linear regression model is applied to predicting England Premier League soccer players who play in the forward position [7]. According to the k-fold cross- validation, the testing and training scores can be checked for the risk of injury of a group of players for the next games and the more important focus of that research is the dependent variable which is the distance of players for each game [7]. Machine learning linear regression is not the only model for predicting injury and also decision trees also have been used to figure out non-contact soccer injuries [8] By using the 'ADASYN' oversampling technique [9] data imbalance problem is solved and 954 data recordings are collected by using GPS technology. ACWR, the ratio of mean and standard

1st International Workshop on Intelligent Software Engineering, 2022
davronbekmalikov96@gmail.com (Davronbek Malikov);
jaeho.kim@gnu.ac.kr (Jaeho Kim)

deviation(MSWR), and the exponential moving average(EWMA) of each training load variable included in this study in order to examine the classification model

3. Data Introduction and model environment

In this section, we provide an overview of the data used in our analysis, as well as a description of its features and the methods by which it was collected.

3.1. Data Loading

Data is indeed the heart of every machine learning project and model. In order to develop accurate and effective predictive models, it is essential to have access to high-quality, relevant data that reflects the real-world phenomenon being studied. This data should be comprehensive, encompassing a range of variables and factors that are likely to impact the outcome being predicted..In this work, we study 22 professional football players history of career who have been playing in the top 5 five football. leagues in Europe, including, La Liga in Spain, the Premier League in England, Ligue 1 in France, the Bundesliga in Germany, and Serie A in Italy. The players in our data set were carefully selected to provide a representative sample of professional footballers playing in the Forward and Midfield positions, which are known to be particularly high-risk positions for injury. Players position also reason for getting injury as we seen from the soccer history[15]. Our dataset includes seven center-forwards, six left-wingers, four right-wingers, three central midfielders, and two attacking midfielders, reflecting a range of playing styles and positions. By focusing on these key positions, we were able to gain insights into the specific risk factors and injury patterns associated with each, providing a more nuanced understanding of injury likelihood among elite football players. Our analysis of these players' injury histories and risk factors will help inform injury prevention and management strategies for players in similar positions, ultimately promoting safer and more sustainable football practices. In this study, we collected data on professional football players from the transfermarkt.com website, which is a leading online platform for sports data and analysis. It provides a wealth of information on player statistics, career history, and injury records, making it an ideal source for our research.

3.2. Model Environement

Python has emerged as the preferred programming language for data science, data analysis, and machine learning projects due to its versatility, simplicity, and robust ecosystem of libraries and tools. Python provides a powerful set of libraries and frameworks for handling large datasets, statistical analysis, and visualization, including NumPy, Pandas, and Matplotlib. In case of machine learning there are many Integrated Development Environments (ID provides a range of powerful machine learning frameworks, such as Scikit-learn, TensorFlow, and PyTorch, which enable developers to build sophisticated machine learning models with ease. These frameworks provide a wide range of algorithms and techniques for supervised and unsupervised learning, as well as tools for model evaluation and optimization to run Python code, and data analysis, including; data cleaning, and data integration, and more importantly allows access to the Python machine learning libraries which will help use machine learning models.

For our data analysis and visualization tasks, we used the Jupyter Notebook as our integrated development environment (IDE) and Python version 3.9.12 as our programming language. Jupyter Notebook provides an interactive environment that allows us to write and execute code, visualize data, and communicate our findings in an organized and accessible format.

4. Model Evaluation

In this section, we will introduce the machine learning model that we used to predict injury likelihood among professional soccer players, as well as the Python library that we utilized to implement the model.

Initially, we introduce the model employed in our experiment and subsequently present the outcomes of our experiment with a classification report. We utilized two techniques to obtain the results; first, we provide the results for a group of professional soccer players, and secondly, we provide the outcomes of individual players based on our model. This approach helped us to evaluate the accuracy and efficiency of our model in identifying the risk of injuries for both groups of players. Additionally, this helped us to analyze the performance of our model in predicting the likelihood of injuries for specific players.

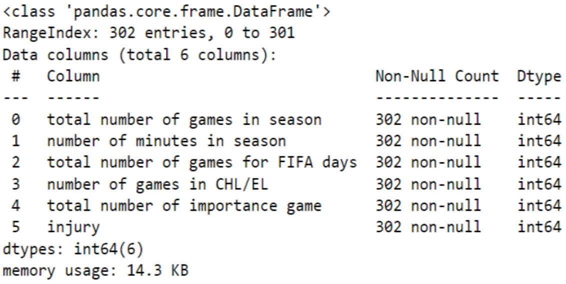
4.1. Model Introduction

In our research project, we opted to use a classification machine learning model due to the fact that our dataset possessed the necessary characteristics to enable classification. Classification is one of the fundamental problems in machine learning, and it involves assigning a class or category to a given data point. In this context, we used various classification machine learning models to classify data points based on their features. These models are Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Naive Bayes. Each classification model has its strengths and weaknesses, and the choice of which to use depends on the nature of the data and the problem at hand.

By leveraging a suitable classification algorithm, we aimed to predict the likelihood of injury among professional soccer players. In our work we use the Naive Bayes model which is a supervised classification machine learning model. The Naive Bayes model, which is a type of probabilistic algorithm, is based on low conditional probability and is highly interpretable. This algorithm is suitable for both binary and multi-class classification problems, making it a versatile choice for predicting injury likelihood among professional soccer players. Additionally, Naive Bayes is known for its relatively fast training and prediction times, making it a popular choice in many machine learning applications. In our work we have a binary class classification problem such that 0 for not injury and 1 is injury and find predictive value for soccer player risk of injury. In this research , we use sci-kit-learn library [11] that is available in python alongside Naive Bayes. We create 2 parts of data that will use for fitting and predicting values. Our research methodology involved the use of both dependent and independent variables, which were incorporated into our training and testing datasets. By including both types of variables in our analysis, we aimed to develop a robust machine learning model that could accurately predict injury likelihood among professional soccer players. In our case independent variables such as the total number of games in the season, the number of minutes in a season, the total number of games in FIFA days, the number of games in CHL/EL, and the total number of an important game. We considered the total number of games including such games in the top five football leagues CHL/EL and Europe Conference League group stages, World Cup, and Olympic Games as well as playoff rounds. Furthermore, dependent variable by default of machine learning models and in our case, we are involving the relationship between independent variables with a single dependent variable which is injury

4.2. Evaluation of model and classification report

To develop our machine learning model, we began by splitting our dataset into two separate sets: a training dataset and a testing dataset. This approach enabled us to train our model on a portion of the data and then test its performance on an independent subset. Fig 1 illustrates the variables of our data set alongside their characters of them. We can see from the Figure that it includes 6 columns including dependent and independent variables, and also an index of data from 0 to 301.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 302 entries, 0 to 301
Data columns (total 6 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   total number of games in season   302 non-null    int64
 1   number of minutes in season       302 non-null    int64
 2   total number of games for FIFA days  302 non-null    int64
 3   number of games in CHL/EL         302 non-null    int64
 4   total number of importance game   302 non-null    int64
 5   injury                            302 non-null    int64
dtypes: int64(6)
memory usage: 14.3 KB
```

Figure 1: **Characteristics of dataset**

In addition to providing insight into the variables included in our dataset, Figure 1 also highlights some important characteristics of the data itself. Specifically, it indicates that our dataset is complete, with no null or missing values, and that all variables have been recorded as integers. Additionally, the memory usage of our dataset is relatively low, at just 14.4 KB. These findings suggest that our dataset is well-organized and suitable for use in developing our machine learning model.

Moreover in the figure 2 we can see the injury indicator that is machine learning model refers to a variable or feature that is used to predict the likelihood or risk of injury. According to the our dataset and the number of injury history we can see the deep inside of information of players

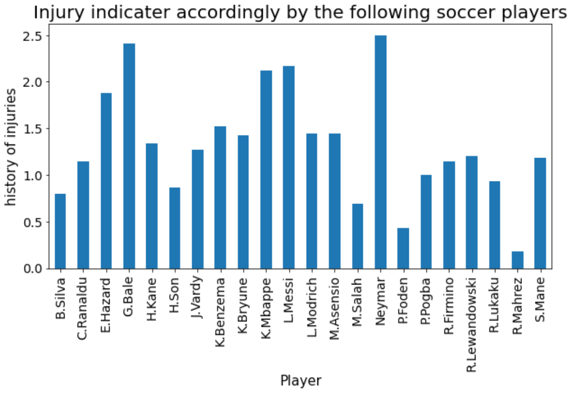


Figure 2. **Injury indicator of soccer players.**

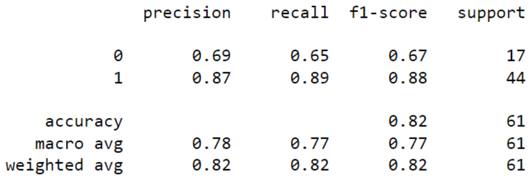
Furthermore, to split data into training and testing we import the train_test_split function from sklearn.cross_validation library and then 80% of the total data for the training set and 20% of data for testing . Then 241 samples were used for training and other remaining 61 samples for testing

Gaussian Naive Bayes is a popular classification algorithm that works well with continuous data and assumes that the features are normally distributed. It is based on Bayes' theorem and assumes independence between features. The algorithm works by calculating the probabilities of each class for a given set of features, and then choosing the class with the highest probability as the predicted class. In our case, we used Gaussian Naive Bayes to train our model due to its effectiveness in training data and its ability to handle continuous data.

Moreover, an important part of the model is Evaluating Process. In this process, we test our data and we take a result with the accuracy of our model.

In Fig.3, the classification report of the model is given. The precision and recall score of the model is 87 and 89 percent for the injury sample (1) respectively where the precision score is injury identification and recall are the proportion of actual injury cases that are correctly classified, f1-score is the mean of the precision and recall. We can see by f1-score the performance of these two classifiers in our case is 88 percent while the support is the number of samples for testing data with 61 total numbers 17 for not-injury and 44 for injury cases since we used 20 percent of data for testing our model.

Moreover, the accuracy of our model is 82 percent. The macro average of the model is the most straightforward among other types of averaging methods and for precision is 78 percent and with the same 77 percent for recall and f1-score one of the differences between a macro and a weighted average is considering the proportion of each support of classes is considered in the weighted average and our case weighted average is same 82 percent for precision and recall and f1-score.



```
              precision    recall  f1-score   support

           0       0.69      0.65      0.67        17
           1       0.87      0.89      0.88        44

    accuracy                           0.82        61
   macro avg       0.78      0.77      0.77        61
weighted avg       0.82      0.82      0.82        61
```

Figure 3: Classification Report

In our experiment we have determined injury values with classification report for all 22 professional soccer players that are in our dataset. However individual players can be selected.

According to our proposed model, we divided players into two groups, high-risk injury players and low-risk of injury players. The player who is a maximum number of injury levels that is based on the player's history of injuries can be selected individually. G. Bale's injury level is higher than other players with 8 and while P. Foden with R.Mahrez is minimum level 1 as given in Fig.3.

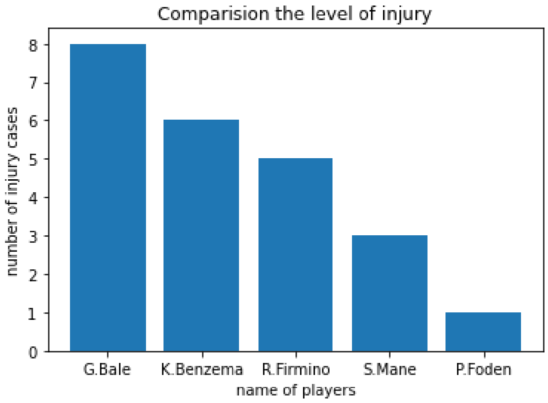


Figure 4: Player's injury-level representation

We can see from Figure 4 that the frequency of injury is given in the y-axis while the name of the player whose level of injury lies between G.Bale and P.Foden who are max and min of level injury in our data respectively is given in the x-axis.Then G. Bale's likelihood of injury next season with 87 percent and P. Foden has a less injury chance

5. Conclusion

The field of machine learning has made remarkable advancements, particularly in the area of sports medicine and sports science. With the help of machine learning algorithms, researchers and practitioners are now able to analyze large datasets and gain insights into athlete performance and injury risk. These insights can be used to develop personalized training programs, prevent injuries, and improve overall athletic performance in the different sports.

Predicting the likelihood of the risk of a soccer injury is the focus of our research, and in order to achieve this goal, we have proposed a classification machine-learning model - learning model to predict the likelihood of the risk of a soccer injury. In order to collect the necessary data for our research, we utilized a well-known and reputable soccer website called transfermarkt.com. This website is widely recognized as a reliable source for obtaining valuable information on soccer players, such as their career history, transfer fees, market value, and more. By using this website, we were able to access and gather data on the 22 professional soccer players who have played in the top five leagues in Europe, including La Liga in Spain, the Premier League in England, Ligue 1 in France, the Bundesliga in Germany, and Serie A in Italy. This data set enabled us to conduct a thorough analysis of each player's injury history and assess their level of risk for future injuries. The classification report is a useful tool to evaluate the performance of a classification model. It helps to understand the accuracy of the model by providing metrics such as precision, recall, and F1 score for each class. In our study, we used a classification report to evaluate the performance of our machine learning model for predicting the likelihood of injury in soccer players. In addition, we evaluated the accuracy of our model to determine how well it fits our research objectives. To provide a more comprehensive evaluation, we reported two types of averages, namely macro and weighted. The macro-average calculates the unweighted mean of the precision, recall, and F1-score across all classes, while the weighted average takes the average weighted by the number of samples in each class. By reporting both averages, we can gain insights into how well our model performs overall and how well it performs for individual classes.

To improve the accuracy and applicability of our model, we plan to further develop our approach in future work. Our focus will be on extending our study to include a larger sample size of soccer players. Additionally, we will explore the possibility of predicting and preventing specific types of injuries in our experiment.

In the professional soccer world, it is a well-known fact that aspiring soccer stars often have to face prolonged periods of time out of games due to soccer injuries. Hence, to achieve success and become a soccer star, it is crucial to avoid such injuries. For instance, the examples of Ronaldo and Messi, who have managed to steer clear of serious soccer injuries that require lengthy recovery periods, highlight the importance of injury prevention in achieving success in this sport.

The primary objective of our research was to assist aspiring soccer players in achieving success and reaching the pinnacle of their sport, just like Cristiano Ronaldo and Lionel Messi. By doing so, we aimed to contribute towards the advancement of soccer as a whole and provide soccer fans with an even more exciting and competitive game.

In addition, the proposed machine learning model can help coaches in developing preventive measures to avoid player injuries. By identifying players with a high risk of injury, coaches can create personalized training programs and make adjustments to their playing tactics to minimize the chances of injury. This could not only benefit the team's performance, but also save the team money by avoiding medical expenses and lost playing time.

Furthermore, the model can be used to compare the injury risk of players across different teams and help in the scouting process for potential new players. In addition to being beneficial for coaches and players, our model can also provide value to team scouts. By analyzing specific players before making decisions about their recruitment, scouts can more easily assess the player's likelihood of consistent performance in the upcoming season. This not only saves time and money for teams, but it also allows them to make more informed decisions about player recruitment. Overall, our model has the potential to significantly impact the decision-making process of teams and improve their overall performance.

References

[1] Anderson L,Triplet-McBride T,Foster C,Duberstein S, Brice G.Imact of traing patterns on the incidence of illness and injury during a woman;s collegiate basketball season. The Journal of Strength&Conditioning Research.2003;17:734-738.

[2] Foster C. Monitoring training in the athletes with reference to overtraining syndrome. Med Sci Sports Exerc. 1998;30:1164-1168.pmid966290

[3] K.Pelechrines and E. Papalexakis,’The Anatomy of American Football:Evidence from 7 years of NFL Game Data’.PlosOne. 22-Dec-2016

[4] D.Memmert and R.Rein, ‘Match Anaysis,Big Daya and Tactics: Current Trends in Elite Soccer’, Research Gate, March 2018

[5] R.J. Bar and J.F.X. De Souza , ‘Tracking Plasticity: Effects of long-term Rehearsal in Export Dancers Enoding Music to Movement’’, Plosone 20 Jan-2016

[6] B.Pang, L.Lee and S.Nithyanatahan, Thumbs Up? Sentiment Classification using Machine Learning’ Association for Computational Lingusitics, July-2022

[7] Amiel S, Richard, P. Injury Prediction for Soccer Players Using Machine Learning. World Academy of Science ,Enigeeri ng, abd Technology International Journal of Sport and Halth Sciences, Vol:16,No:3,2002

[8] Rossi A, Pappalardo L, Cinta P,et all.Effective injury forecasting in football with GPS training and machine learning PlosOne.2018;13(7)1322-1328

[9] He.H, Bai Y Arcia EA,Li S.ADASYN:adaptive synthetic sampling approach for imbalanced learning.I:Proceddings of the International Joint Conference on Neural Networks.2008;pp1322-1328

[10] Daniel B.Baye’s Theorems and Naïve Bayes Classifier. The Open University. DOI:10.1016/B978-0-12-809633-8.20473-1

[11] Gavin H.Mastering Machine Learning with scikit-learn.PACKT publishing.Birmingham-Mumbai.Open Source community experience distilled

[12] Transfermarket.com

[13] P.Wong, Y Hong. Soccer Injury in the lower extremities. Br J Sports Med 2005; 39:473–482. doi: 10.1136/bjsm.2004.015511

[14] Oliver JL, Ayala F, De Ste Croix MBA, Lloyd RS, Myer GD, Read PJ. Using machine learning to improve our understanding of injury risk and prediction in elite male youth football players. J Sci Med Sport. 2020; 23(11):1044–1048

[15] Dick U, Brefeld U. Learning to Rate Player Positioning in Soccer. Big Data. 2019; 7(1):71–82.