Research Article

Predicting Injury Recurrence and Recovery Patterns (Draft)

Victor Pastore Marqueti 1\*

|  |
| --- |
| Academic Editor: Firstname Lastname?  Received: date  Revised: date  Accepted: date  Published: date  **Citation:** To be added by editorial staff during production.  **Copyright:** © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/). |

1 student; vimarqueti@hotmail.com

**Abstract**

Injury recurrence is a constant problem in any professional sports, often affecting athlete performance, long term health, and team goals. This project wants to explore patterns in professional soccer players injury recurrence and investigate how machine learning models can help in predictions about injury outcomes. Using a dataset of player injury histories, we approached the problem with two prediction tasks: estimating the number of days until the same injury recurs (regression) and predicting whether an injury will recur (classification). Several models were tested and kept the relevant ones, including Ridge Regression, Gradient Boosting, and Logistic Regression. Gradient Boosting outperformed other models in both tasks, showing strong predictive accuracy and recall. Ridge Regression also performed with good results in estimating recurrence timing. These results suggest that data modeling can provide meaningful insight into injury trends and recurrence risks. In future work for this project, having access to bigger datasets, this could contribute more player health management strategies in professional soccer players, and team management.

**Keywords:** injury recurrence; machine learning; regression; classification; soccer analytics

**JEL Classification: ?**

1. Introduction (No citation Yet)

Injuries are an unavoidable aspect of professional soccer and have an impact on athletes’ careers and team performance. Understanding how and when injuries recur is important to improve player health, planning training schedules, and guiding medical staff decisions. The increase of the availability in soccer injury data, data methods have become a valuable tool in sports analytics, offering new insights into injury patterns and risks, becoming something important for the success of the team.

Athlete injuries have been studied in sports science, the analysis of recurrence patterns based on individual player histories can also be studies and contribute for the sport. This project focuses on identifying which injury types are more likely to recur and how long it typically takes for recurrence to occur. This project also examines which player characteristics, such as age, position, and previous recovery time could be linked to higher recurrence risk.

To address these questions, several machine learning models were applying to a curated dataset of professional athlete injuries. To generate meaningful visualizations and predictive insights that can support athletic performance and injury management decisions. This project demonstrates what could be done in soccer health analytics and can show areas for future exploration in the field that could benefit a lot of people…

**\*Deeper into show the problem, maybe some number on why it is important, or just change some information with citation confirmation\***

2.1 Data

The data for this project was extract using programming language R, using the worldfootballR package [1]. This package allows access to publicly available football data from websites like Transfermarkt. Using this tool, extracted detailed injury histories for a range of professional soccer players. The dataset includes variables such as player names, age at injury, injury type, position, number of games missed, and injury duration. To allow for the analysis of injury recurrence, each injury was linked to a date, enabling the calculation of how much time passed between injuries of the same type for the same player.

To prepare the dataset, cleaning steps were made. **… Continue explaining the cleaning part…**

Two target variables were made:

* days\_until\_recurrence: a numeric value representing the days until the same injury type recurred for the same player.
* recur\_same\_injury: a binary value indicating whether the same injury happened again (1) or not (0).

2.2 Methods

To analyze the recurrence of injuries, two different machine learning problem types were defined: a regression to estimate the number of days until recurrence, and a classification to predict whether an injury would recur or not.

For the regression target, several models were tested, however some models had better results and were use in the continuity of the project, including Linear Regression, Ridge and Lasso Regression, Gradient Boosting Regressor.

For the classification target, the following models were selected using the same, Logistic Regression, Gradient Boosting Classifier, AdaBoost Classifier, LightGBM Classifier.

**\*Explain each model and why use it, what test what made and performance that it was check\***

3. Results

This section presents the performance results of the classification and regression models applied to the injury dataset. We evaluation metrics and visualizations to each model’s predictive ability and interpret their effectiveness.

3.1 Classification Results

To predict whether an injury would recur, tested four classification models: Logistic Regression, Gradient Boosting, AdaBoost, and LightGBM. These models were evaluated using classification metrics...

Table 1 displays the scores for each model. Gradient Boosting achieved the best performance overall, with the highest values. Logistic Regression and LightGBM also performed reasonably well, showing consistent results. AdaBoost had comparatively lower performance**… Deeper**

To better visualize model performance, Figure 1 shows the ROC curves. The ROC curve for Gradient Boosting is above. **Maybe Figure 2?**

3.2 Regression Results

In the regression task, the goal was to estimate how many days it would take for a recurrence to happen. Tested four models: Linear Regression, Ridge Regression, Lasso Regression, and Gradient Boosting Regressor.

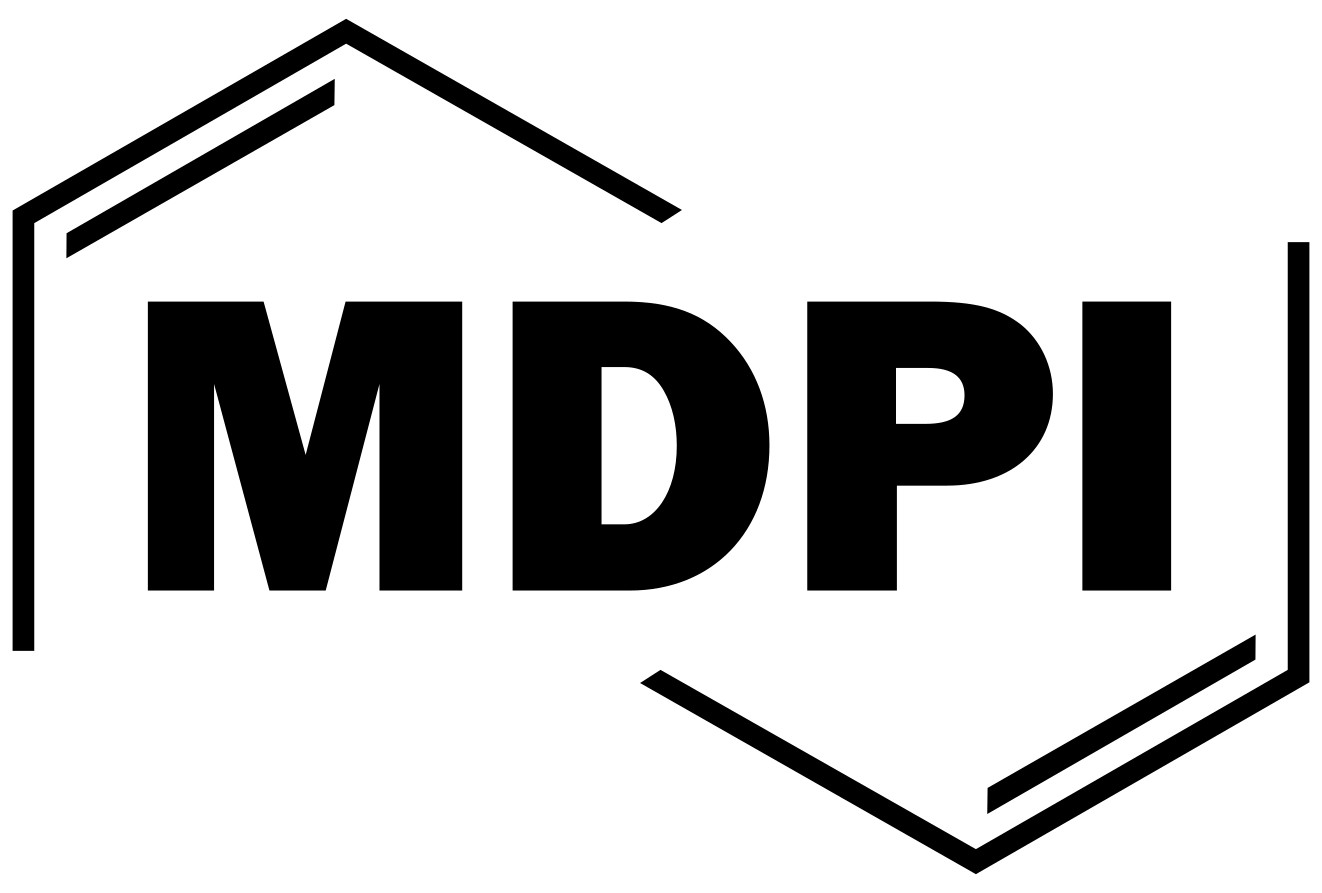
To evaluate performance, calculated the prediction errors (actual minus predicted values). Figure 3 shows boxplots of the error distributions for each model. Could also show more Figure if necessary.

Overall, results…

\*Not delete for example purpose for final report. Do not consider BELOW\*

3.2. Figures, Tables and Schemes

All figures and tables should be cited in the main text as Figure 1, Table 1, etc.

****

**Figure 1.** This is a figure. Schemes follow the same formatting.

**Table 1.** This is a table. Tables should be placed in the main text near to the first time they are cited.

|  |  |  |
| --- | --- | --- |
| **Title 1** | **Title 2** | **Title 3** |
| entry 1 | data | data |
| entry 2 | data | data 1 |

1 Tables may have a footer.

The text continues here (Figure 2 and Table 2).

|  |  |
| --- | --- |
| C:\Users\martin\Downloads\testFigure.tif | C:\Users\martin\Downloads\testFigure.tif |
| (**a**) | (**b**) |

**Figure 2.** This is a figure. Schemes follow another format. If there are multiple panels, they should be listed as: (**a**) Description of what is contained in the first panel; (**b**) Description of what is contained in the second panel. Figures should be placed in the main text near to the first time they are cited.

**Table 2.** This is a table. Tables should be placed in the main text near to the first time they are cited.

|  |  |  |  |
| --- | --- | --- | --- |
| **Title 1** | **Title 2** | **Title 3** | **Title 4** |
| entry 1 \* | data | data | data |
| data | data | data |
| data | data | data |
| entry 2 | data | data | data |
| data | data | data |
| entry 3 | data | data | data |
| data | data | data |
| data | data | data |
| data | data | data |
| entry 4 | data | data | data |
| data | data | data |

\*Not delete for example purpose for final report. Do not consider ABOVE\*

4. Discussion

The classification models, Gradient Boosting, strong performance in identifying recurrence risk, which shows that could be important in integrating machine learning into athlete monitoring systems. However, regression models struggled to accurately predict the exact number of days until recurrence. This limitation maybe shows because more that it is need in this variable**...**

These results support the idea that data informed decisions may improve player care and team planning. Future research should aim to include more future and detailed datasets to improve model accuracy and reliability.

**\*Initial Text\***

5. Conclusions

This project explored the use of machine learning models to analyze injury recurrence in professional soccer players. Classification models performed better, showing potential to help teams in identifying players at higher risk of re-injury. Regression models had more difficulty**…? Again same 4. ?**

These results show the potential of data to support athlete and performance planning. Future work should focus on collecting additional features and expanding the player database to improve model accuracy and reliability.

**\*Initial text\***

6. Patents (Ask about it)

This section is not mandatory but may be added if there are patents resulting from the work reported in this manuscript.

**Supplementary Materials:** The following supporting information can be downloaded at: https://www.mdpi.com/article/doi/s1, Figure S1: title; Table S1: title; Video S1: title.

**Author Contributions:** For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used “Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.” Please turn to the [CRediT taxonomy](https://img.mdpi.org/data/contributor-role-instruction.pdf) for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

**Funding:** Please add: “This research received no external funding” or “This research was funded by NAME OF FUNDER, grant number XXX” and “The APC was funded by XXX”. Check carefully that the details given are accurate and use the standard spelling of funding agency names at https://search.crossref.org/funding. Any errors may affect your future funding.

**Institutional Review Board Statement:** In this section, you should add the Institutional Review Board Statement and approval number, if relevant to your study. You might choose to exclude this statement if the study did not require ethical approval. Please note that the Editorial Office might ask you for further information. Please add “The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval).” for studies involving humans. OR “The animal study protocol was approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval).” for studies involving animals. OR “Ethical review and approval were waived for this study due to REASON (please provide a detailed justification).” OR “Not applicable” for studies not involving humans or animals.

**Informed Consent Statement:** Any research article describing a study involving humans should contain this statement. Please add “Informed consent was obtained from all subjects involved in the study.” OR “Patient consent was waived due to REASON (please provide a detailed justification).” OR “Not applicable.” for studies not involving humans. You might also choose to exclude this statement if the study did not involve humans.

Written informed consent for publication must be obtained from participating patients who can be identified (including by the patients themselves). Please state “Written informed consent has been obtained from the patient(s) to publish this paper” if applicable.

**Data Availability Statement:** We encourage all authors of articles published in MDPI journals to share their research data. In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Where no new data were created, or where data is unavailable due to privacy or ethical restrictions, a statement is still required. Suggested Data Availability Statements are available in section “MDPI Research Data Policies” at https://www.mdpi.com/ethics.

**Acknowledgments:** In this section, you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments). Where GenAI has been used for purposes such as generating text, data, or graphics, or for study design, data collection, analysis, or interpretation of data, please add “During the preparation of this manuscript/study, the author(s) used [tool name, version information] for the purposes of [description of use]. The authors have reviewed and edited the output and take full responsibility for the content of this publication.”

**Conflicts of Interest:** Declare conflicts of interest or state “The authors declare no conflicts of interest.” Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. Any role of the funders in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results must be declared in this section. If there is no role, please state “The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results”.

Abbreviations

The following abbreviations are used in this manuscript:

|  |  |
| --- | --- |
| MDPI | Multidisciplinary Digital Publishing Institute |
| DOAJ | Directory of open access journals |
| TLA | Three letter acronym |
| LD | Linear dichroism |

Appendix A (Ask about it)

Appendix A.1

The appendix is an optional section that can contain details and data supplemental to the main text—for example, explanations of experimental details that would disrupt the flow of the main text but nonetheless remain crucial to understanding and reproducing the research shown; figures of replicates for experiments of which representative data is shown in the main text can be added here if brief, or as Supplementary data. Mathematical proofs of results not central to the paper can be added as an appendix.

**Table A1.** This is a table caption.

|  |  |  |
| --- | --- | --- |
| **Title 1** | **Title 2** | **Title 3** |
| entry 1 | data | data |
| entry 2 | data | data |

Appendix B

All appendix sections must be cited in the main text. In the appendices, Figures, Tables, etc. should be labeled starting with “A”—e.g., Figure A1, Figure A2, etc.

References

References must be numbered in order of appearance in the text (including citations in tables and legends) and listed individually at the end of the manuscript. We recommend preparing the references with a bibliography software package, such as EndNote, ReferenceManager or Zotero to avoid typing mistakes and duplicated references. Include the digital object identifier (DOI) for all references where available.

Citations and references in the Supplementary Materials are permitted provided that they also appear in the reference list here.

In the text, reference numbers should be placed in square brackets [ ] and placed before the punctuation; for example [1], [1–3] or [1,3]. For embedded citations in the text with pagination, use both parentheses and brackets to indicate the reference number and page numbers; for example [5] (p. 10), or [6] (pp. 101–105).

1. Wragg, J. (2023). worldfootballR: Data from the World of Football (Soccer). Comprehensive R Archive Network (CRAN). <https://cran.r-project.org/web/packages/worldfootballR/readme/README.html> (Changing Place)
2. Author 1, A.; Author 2, B. Title of the chapter. In *Book Title*, 2nd ed.; Editor 1, A., Editor 2, B., Eds.; Publisher: Publisher Location, Country, 2007; Volume 3, pp. 154–196.
3. Author 1, A.; Author 2, B. *Book Title*, 3rd ed.; Publisher: Publisher Location, Country, 2008; pp. 154–196.
4. Author 1, A.B.; Author 2, C. Title of Unpublished Work. *Abbreviated Journal Name* year, *phrase indicating stage of publication (submitted; accepted; in press)*.
5. Author 1, A.B. (University, City, State, Country); Author 2, C. (Institute, City, State, Country). Personal communication, 2012.
6. Author 1, A.B.; Author 2, C.D.; Author 3, E.F. Title of Presentation. In Proceedings of the Name of the Conference, Location of Conference, Country, Date of Conference (Day Month Year).
7. Author 1, A.B. Title of Thesis. Level of Thesis, Degree-Granting University, Location of University, Date of Completion.
8. Title of Site. Available online: URL (accessed on Day Month Year).

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.