IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Athlete Statistics Visualization and Prediction

A. Patel [AU2140118][1], A. Prajapati [AU2140090][2], B. Shah [AU2140088][3], and S. Virani [AU2140139][4]

[1,2,3,4]are with School of Engineering and Applied Science, Ahmedabad University, Gujarat, India
(email: arpit.p@ahduni.edu.in, aryan.p3@ahduni.edu.in, buland.s@ahduni.edu.in, sujal.v@ahduni.edu.in).

ABSTRACT We have a multi-modal dataset of division I basketball players consisting of their physical and physiological characteristics and using methods of data cleaning and preprocessing, we have made the data suitable for further analysis and prediction using machine learning algorithms. The data was provided by the authors of the paper that is referenced the first in citations. That data in turn was collected during the pandemic-condensed season with high instances of unpredictability. While the overall objective of the project is to train models in order to predict 5 modality scores of sleep, training, cardiac rhythm, jump and cognitive, on the original dataset and representation of these predictions and data analysis through visualization in form of a dashboard being the conclusive requirement, we have been able to complete the preprocessing so far. This includes proper data imputation, interpretable feature set, data balancing and classifiers.

INDEX TERMS Basketball, collegiate athletes, data mining, sports analytics, machine learning, game performance, injury prediction

I. INTRODUCTION

The field of sports data analytics is increasingly being recognized for its importance across both collegiate sports and professional leagues, including Esports. The ability to derive insights related to athletic performance has the potential to not only enhance game outcomes but also to minimize injury risks, thereby playing a crucial role in a team's overall achievement. This multidisciplinary study has unified efforts from a diverse group comprising athletes, coaches, exercise scientists, engineers, and data analysts. It focuses on a Division-1 basketball team's season, which was notably affected by the COVID-19 pandemic. By analyzing the athletes' daily sleep patterns, training, cardiac rhythm, jump and cognitive patterns and integrating these with subjective training feedback and survey responses, this initiative has set the stage for predicting game outcomes and injury likelihood using machine learning techniques.

II. METHODOLOGY

The data that was collected was unusual and noisy; null values were in majority to a good extent for a group of features. Since the predictability of the machine learning model/s that would be applied on the data also depends on the quality of the data (Garbage In = Garbage Out), it is essential (and strictly mandatory!) for us to perform Data Preprocessing to help make out important assumptions and expressions of features and modalities, besides enhancing the accuracy of the machine learning model/s.

We followed the following steps for the task of preprocessing the provided data files(.csv), using Python:

1) Importing essential libraries for required data operations
2) Importing the datasets into the data structures compatible with the Python environment
3) Management of the null values
4) Encoding categorical data
5) Splitting the dataset into Test and Training datasets
6) Feature Scaling

The first two steps were easily implemented by going through the syntax of the programming language. Before proceeding to the third step, we studied and observed the data through rough glances to get an idea of the structure it possesses. As we found out that null values were present to a great extent, we employed Data Imputation.

A. *Data Imputation*

In Data Imputation, we retain the majority of the information by substituting null values with a different value. This needs to be done since it would be impractical to keep on removing data from a dataset every time, besides the reduction in the size of the dataset under consideration (which would further raise questions about bias and impairing analysis).

Our research for techniques under this domain gave valuable insight to the technique named MICE (Multiple Imputation by Chained Equation) which is itself based on the iterative nature of Machine Learning algorithms. We decided to go on with this technique in light of the learnings of our current curriculum. Here is a short insight regarding the methodology:

```
Fill in missing values from random draws of non-missing data
For each iteration
    For each variable v with missing values
        Optional: subset data where v was originally nonmissing
        Train model v ~ X where X are the other variables in the dataset
        Do one of:
            1) Replace missing values with predictions from model
            2) Replace missing values using mean matching
    End
End
```

FIGURE 1. **Visualization of the MICE (Multiple Imputation by Chained Equation) technique for filling missing values in datasets. The image illustrates the iterative process of imputing missing values using random draws of non-missing data. For each iteration, the algorithm trains a model for each variable with missing values and replaces them either with predictions from the model or using mean matching. This process iterates until all missing values are filled.**

B. *Correlation-based Feature Importance*

Feature selection is important to identify the most relevant variables contributing to a model's performance, ensuring that the model does not pick up patterns that seem to appear while testing but does not represent closeness anywhere near to the ground truth.

We decided to implement this as it selects subsets of features that are highly correlated with the target variable but have low correlation with each other. The underlying principle consists of a subset of features that can provide the maximum amount of information about the target variable while minimizing redundancy among the features.

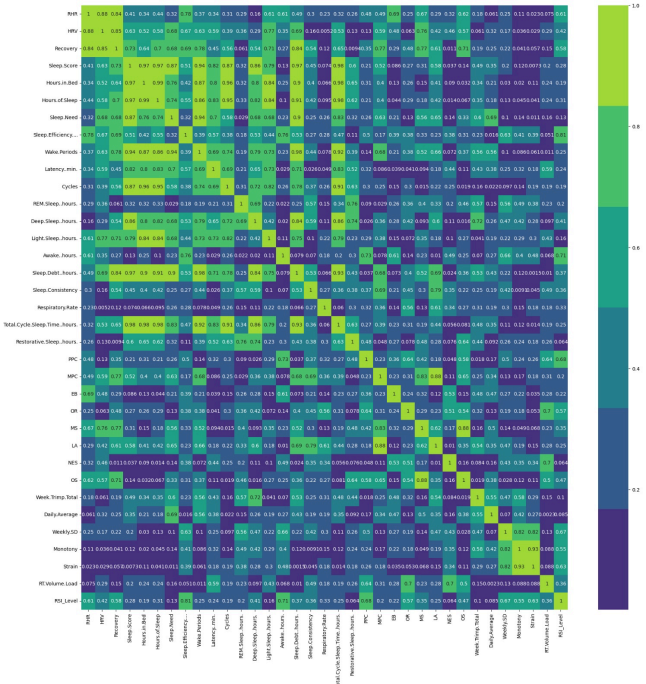

FIGURE 2. **The heatmap showcases the correlation between variables in the dataset. Each square denotes the correlation coefficient, with colors ranging from green to yellow based on the "viridis" color palette. Annotated values display correlation coefficients, aiding in understanding variable relationships.**

C. *Quartile Ranging for RSI categorization*

RSI Quartile Range used to categorize athletes into 4 groups/classes.

These techniques were employed in Python using a Python notebook on Google Colab.

III. DISCUSSION

We believed that the predictability testing and training of the machine learning model/s would be of the most significant complexity level to perform but upon commencement of data preprocessing, we understood that first-hand dealing with datasets is also important and of much complexity given the fact that it determines the degree of useful analysis and therefore, the performance of the prediction.

One of the key learnings included the lesson of how to deal with missing data without using the traditional simple mean/median methods but to go with imputing missing values by predicting them using other features from the dataset. Initially, we believed that the use of ML algorithms would be only in the training and testing phases and we did not even think about the fact that we can apply the iterative nature of these algorithms in the data preprocessing in itself.

With guidance from the paper, we learnt that the application of ENN (Edited Nearest Neighbor) technique on the oversampled dataset can try to fix the aforementioned problem.

It essentially implements the removal of those majority class instances whose neighbors belong to a different class completely. The identification of these neighbors is performed by KNN (K-Nearest Neighbor) algorithm.

A key point to note is that generating synthetic samples from the minority class does not always work well since it does not account for complete data variability. A researched method, ADASYN (Adaptive Synthetic Sampling), introduces adaptiveness into the classification boundary to address the challenging instances but this sometimes exhibits sensitivity to the minority data distribution. Therefore, SMOTE implementation with Edited Nearest Neighbor serves as a more comprehensive approach.

Coming to the future goals concerning the project which is the modeling and data prediction, the collected data needs to be evaluated against game statistics and injury reports. Papers for reference suggest the use of Bayesian linear regression model for player performance on win probability, accounting for its multi-dimensional parameters. The primary aim remains in the use of methodology that internally supports multiple boosted decision trees as the base models; models satisfying these are XGBoost and Random Forest. Random Forest is known for providing interpretability; it is easier to compute contribution of features to decision since Random Forest makes use of change in accuracy to find feature importance when the feature in consideration is excluded from decision.

IV. CONCLUSION

Data analysis of player performance in sports and prediction using the implementation of Machine Learning algorithms can revolutionize sports management by the outcomes of rostering composition, building gameplay strategies and managing player flow during game. One of the key aspect outputs also include injury prediction that can be used to avoid predicted mishappenings. The player efficiency crunched from the datasets also provided valuable insights for each player and can help the management to create formations and strategies based on that while the individual athlete response to these outputs can be in regards to training of their weak areas.

V. REFERENCES

[1] *Taber, C. B., Sharma, S., Raval, M. S., Senbel, S., Keefe, A., Shah, J., ... & Kaya, T. (2024). A holistic approach to performance prediction in collegiate athletics: player, team, and conference perspectives. Scientific Reports, 14(1), 1162.*

[2] *Senbel, S., Sharma, S., Raval, M. S., Taber, C., Nolan, J., Artan, N. S., ... & Kaya, T. (2022). Impact of sleep and training on game performance and injury in division-1 women's Basketball Amidst the Pandemic. Ieee Access, 10, 15516-15527.*

[3] *Ibáñez, S. J., Sampaio, J., Feu, S., Lorenzo, A., Gómez, M. A., & Ortega, E. (2008). Basketball game-related statistics that discriminate between teams' season-long success. European journal of sport science, 8(6), 369-372.*

[4] *Sarlis, V., & Tjortjis, C. (2020). Sports analytics—Evaluation of basketball players and team performance. Information Systems, 93, 101562.*