Data-driven imputation scheme for human-subject-based dataset

Aaditya Yadav, Shubham Apat, Helly Thakkar, Pal Patel

School of Engineering and Applied Science

Ahmedabad University

***Abstract*—This report discusses the problem of dealing with missing values in a multi-modal dataset of Division I basketball players. It describes a process that combines systematic data exploration, feature sensitivity analysis, and data imputation techniques. The findings show the progress gained in identifying key features, converting, and integrating datasets, and evaluating imputation algorithms. The discussion covers the problems encountered and the implications of the findings for improving dataset completeness and analytical accuracy.**

***Keywords*—Missing values, Multi-modal dataset, Feature sensitivity analysis, Data imputation, Data preprocessing.**

I. Introduction

In sports analytics, analysing multimodal datasets provides distinct challenges, particularly when dealing with missing information. This report focuses on a Division I basketball dataset that includes sleep habits, training details, emotional-mental states, and game scores, however missing data prevents thorough analysis. The major goal is to provide a data-driven imputation technique that can successfully handle missing values, increasing the dataset's usefulness for in-depth player performance and well-being analyses.

II. Methodology

The method takes a sequential approach, beginning with preliminary data exploration to better understand the dataset's structure and missing value patterns. Feature sensitivity analysis is used to prioritize features that are critical for imputation. Furthermore, data from various sources is transformed, structured, and combined to make a comprehensive dataset. Imputation strategies such as mean, median, and mode approaches are tested for their ability to fill missing values while respecting dataset features.

III. Result

We formatted the data and finally had 2 merged files of season 2 and season 3 then we performed data preprocessing using the mean imputation strategy then we did EDA and visualized the graphs of different features and found some features can be dropped due to collinearity issues. Then performed a Correlational analysis to find the correlation between the features and found this:

Biological Factors vs. Performance Metrics: Explore correlations between biological factors (e.g., RHR, HRV) and performance metrics (e.g., PTS, Jump Height). Understanding these correlations can provide insights into how physiological variables may impact athletic performance.

Training Load and Recovery:

Investigate correlations between training load metrics (e.g., Training Load Score, Cardio Load) and recovery-related metrics (e.g., Sleep Efficiency, Recovery Time). Strong correlations in this area can highlight the importance of balancing training intensity with adequate rest and recovery.

Game Performance and Heart Rate: Examine correlations between game performance metrics (e.g., PTS, AST) and heart rate data (e.g., HR avg, Time in Heart Rate Zones).

These correlations can shed light on how physiological responses during gameplay relate to performance outcomes.

IV. Discussion

The difficulties encountered when dealing with missing values originate from varying missingness patterns and fundamental causes such as athlete neglect or equipment malfunction. The findings emphasize the necessity of thorough data exploration, feature prioritizing, and careful imputation technique choice. Integrating data from many sources improves analysis capabilities while assessing imputation procedures reveals insights into balancing efficiency and accuracy.

V. Conclusion

The research continues by highlighting the importance of the findings in developing data preparation approaches for sports analytics. By resolving missing values in a data-driven manner, the dataset's completeness and analysis accuracy are improved. Future research directions may involve refining imputation algorithms and applying them to different sports datasets to enhance complete performance analysis and decision-making.

VI. References

1. Impact of sleep and training on game performance and injury in division-1 women's Basketball. Amidst the Pandemic. S Senbel, S Sharma, MS Raval, C Taber, J Nolan... - Ieee Access, 2022.

2. Thomas Reilly and Ben Edwards. "Altered sleep– wake cycles and physical performance in athletes". In: *Physiology & behavior* 90.2-3 (2007), pp. 274–284.

3. Jennifer Schwartz and Richard D Simon Jr. "Sleep extension improves serving accuracy: A study with college varsity tennis players". In: *Physiology & behavior* 151 (2015), pp. 541–544.