Ahmedabad University

ML Project Weekly Report

Course Name: Machine Learning (CSE 523)
Week Number: 2 (3rd - 9th February 2024)
Group Name: White
Instructor's Name: Mehul S Raval
Project no.: 7

Project 7 | Athlete Statistics Visualization and Prediction

Introduction

This report details the progress made in the second week of our project. We focused on data cleaning, pre-processing, and initial exploration, following the Teaching Assistant's instructions. Our focus remained on understanding the dataset provided, which includes various factors crucial for athlete performance.

Weekly Activities

1. Data Pre-processing:

- We converted Vertical Jump Season 2.xlsx into a CSV format with a structure similar to Vertical Jump Season 3. csv .
- We merged Vertical Jump Season 2.csv and Season 2 with Polar.csv using "date" as the key, creating a single CSV with all modalities for each athlete.
- We repeated the merging process for Season 3 data, resulting in two comprehensive CSV files for further analysis.

2. Understanding the Dataset:

We spent time getting familiar with the dataset, which contains information on sleep patterns, training logs, cardiac rhythm measurements, emotional-mental state assessments, game performance scores, weekly readiness evaluations, and jump mechanics (RSImod) from Division I basketball athletes.

Ahmedabad University

Challenges Faced

1. Missing Values:

We encountered a significant number of missing values in the Season 2 with Polar.csv file and Season 3 file. We will be addressed this by imputing missing values using mean or median values for continuous features and mode for categorical features in future, while also considering the potential impact of missing data on our analysis.

2. Data Discrepancies:

We found inconsistencies in athlete names and data formats between different datasets. We carefully merged and corrected these inconsistencies to ensure data integrity and avoid merging errors.

Learnings

Through our work, we learned the importance of:

- Cleaning and preparing data to ensure its quality for analysis.
- Managing collinearity issues and removing redundant features to avoid bias in our analysis.
- Integrating and merging datasets effectively to create a comprehensive dataset for analysis.

Next steps

- We will further refine our approach to handle missing values by employing techniques such as imputation or removal based on their impact and prevalence. This step is crucial for ensuring the completeness and accuracy of our dataset.

- We aim to normalize features using techniques like min-max scaling or z-score normalization. By doing so, we can ensure consistency across different data scales, which is essential for accurate analysis and model development.

- We will explore dimensionality reduction techniques like Principal Component Analysis (PCA), Correlation Matrix or Weighted Rank-Sum [Hard Voting] to reduce the feature space while retaining essential information.

Conclusion

Our second week was mainly about the retrieval of the datasets and to make progress in preparing the dataset for analysis. We will continue refining the data pre-processing steps, including addressing any remaining missing values and outliers, to ensure the quality of our analysis.