

CSE523 Machine Learning

Weekly Report 5

**Athlete Profiling for Division I Basketball Players**

Submitted to faculty: Mehul Raval

Date of Submission: 09-03-24

Student Details

| Roll No. | Name of the Student |
| --- | --- |
| AU2140024 | Aagam Shah |
| AU2140133 | Dhairya Shah |
| AU2140159 | Aayushi Shah |
| AU2140182 | Aanal Dobariya |

2023-2024 (Winter Semester)

### **Aim:**

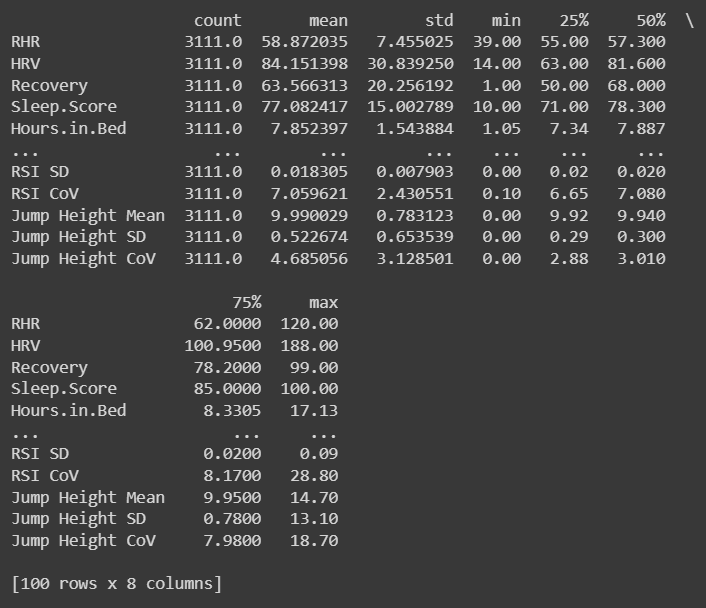
To do factor analysis and feature engineering

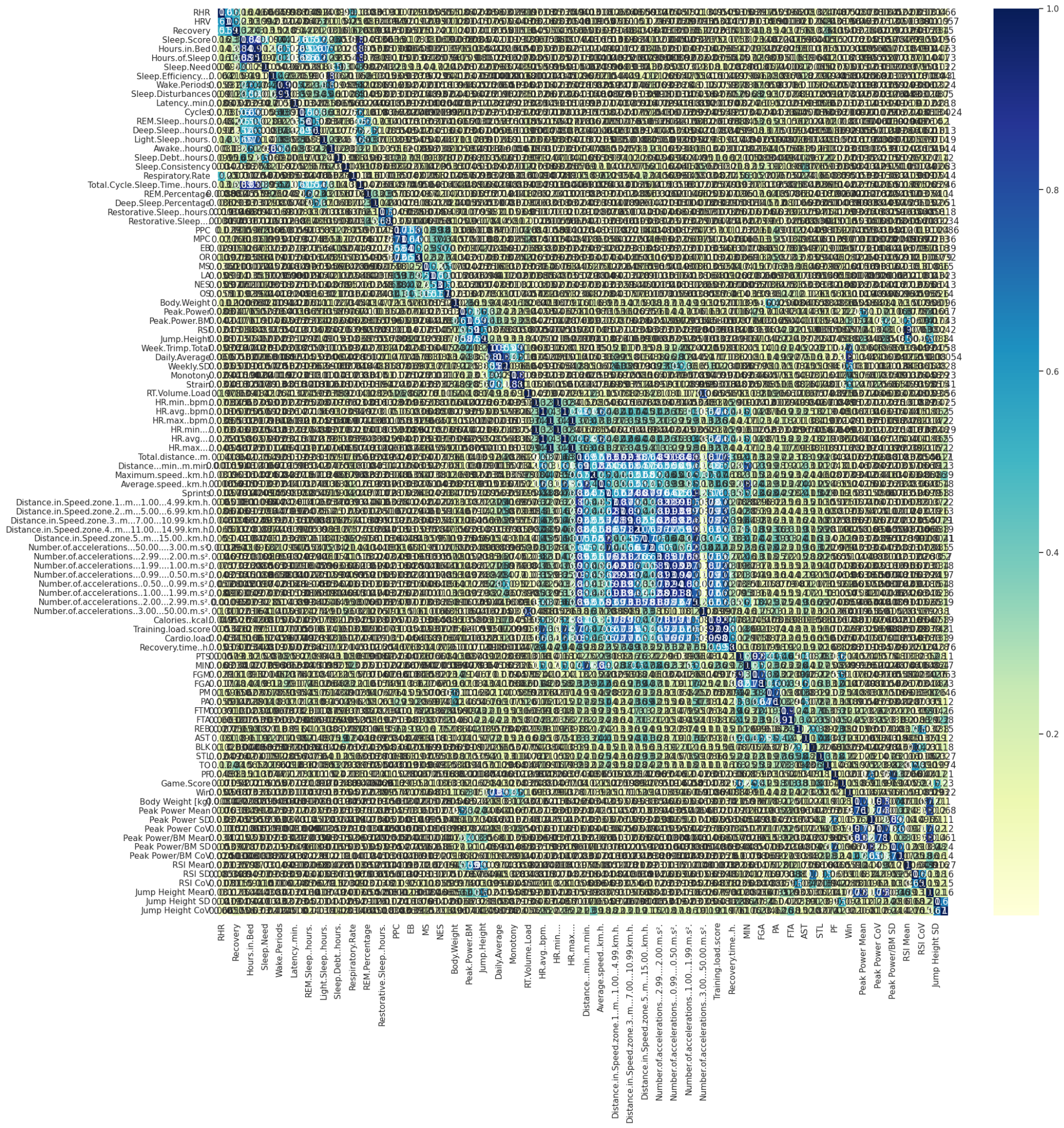
### **Introduction:**

We currently have more than 100 features (115, to be specific), with a lot of features having multiple none values. Therefore, we need to do factor analysis in order to reduce the complexity and null values of the data by finding latent factors that contribute to the observed correlations between variables. This dimensionality reduction would help us to manage and interpret the dataset better. Sometimes, the features don’t have a very high correlation but when they are merged, they together may have a high correlation with the label. This process is called feature engineering (transforming or creating new features from existing data). Specifically, we will be focusing on feature clustering.

### **Work Completed:**

* We have used MICE (Multivariate imputation using chained equations) for imputation using random forest regressor as an estimator.
* After imputation, as a part of the exploratory data analysis (EDA), we have produced a descriptive statistics report using data.describe() function in python.



* Moving ahead in EDA, we have produced a heatmap for every feature and found a very low correlation between some features with the labels. Instead of completely removing such features, we have decided to do factor analysis, i.e. combining the less correlated features, and check whether their correlation increases or decreases with the target label.
* Further more, to categorize RSI values into RSI\_Level, we have find the quantile values using SelectNonCollinear(0.6,scoring=f\_regression) function, which selects non-collinear features with a threshold of 0.6 and uses f\_regression scoring for evaluation, commonly employed in feature selection tasks.

### **Next steps and goals:**

* To identify the features that have a very high correlation with the label.
* To test the decision tree algorithm on our data.

### **Conclusion:**

Our dataset had a bunch of 115 features, and most of them had missing values. So, we did factor analysis to simplify stuff by finding hidden factors linking the features. At the same time, we did some feature tweaking (feature clustering) to fix missing information and spot connections between features. Mixing factor analysis and feature clustering helped us get a better grip on the data, making our analyses more focused and boosting the overall effectiveness of our data-driven moves.

Link to the colab: <https://colab.research.google.com/drive/1WMOik8eT5PZ26DLBYIn0FNncvzSAK5GA?usp=sharing>

### **References:**

1. 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.

<https://ieeexplore.ieee.org/abstract/document/9690164/>

1. Taber, C.B., Sharma, S., Raval, M.S. et al. A holistic approach to performance prediction in collegiate athletics: player, team, and conference perspectives. Sci Rep 14, 1162 (2024). <https://doi.org/10.1038/s41598-024-51658-8>
2. Sharma, S. U., Divakaran, S., Kaya, T., & Raval, M. (2022, July). A Hybrid Approach for Interpretable Game Performance Prediction in Basketball. In 2022 International Joint Conference on Neural Networks (IJCNN) (pp. 01-08). IEEE. <https://ieeexplore.ieee.org/abstract/document/9892583/?casa_token=Ye3GQJ1JpD0AAAAA:GSlyds24pIa__7Od6UBSNs8nugbEwLvCbI8vG6w-YMYrFX2O-TUlUDo4xej3ulJUAvmO_4ij0J36>