Modelling and Prediction of Athletic Readiness based on Cognitive Data

Karan Kalani AU2444018
Meet Siddhapara AU2240243
Om Shah AU2240070
Souma Mazumdar AU24110004

***Abstract*—Player performance prediction and cognitive monitoring are essential components in modern collegiate athletics. This study presents a data-driven approach for preprocessing and feature selection on a cognitive performance dataset comprising various physiological, psychological, and training-related metrics. Motivated by recent research in game performance prediction and athlete readiness modeling, we apply a threshold-based correlation filtering method to identify and eliminate highly correlated features from the dataset. This preprocessing pipeline enhances data quality by removing redundancy, reducing multicollinearity, and retaining the most informative attributes. The cleaned dataset can then be effectively utilized in machine learning models such as Random Forests to predict game readiness, fatigue levels, or injury risk. Our approach aligns with hybrid interpretable models proposed in recent literature and serves as a foundational step in building robust athletic analytics systems.**

***Index Terms*—Data Cleaning, Feature Selection, Correlation Matrix, Random Forest, Python, Cognitive Dataset**

I. Introduction

In recent years, the integration of data analytics into sports science has revolutionized how athlete performance, fatigue, and readiness are monitored and predicted. Collegiate basketball, in particular, has seen increasing use of wearable devices, cognitive assessments, and physiological tracking to provide deep insights into player health and game-day preparedness. The abundance of such multidimensional data enables the development of predictive models that assist coaches and athletic staff in making informed strategic decisions.However, raw athletic datasets often come with challenges such as missing values, noise, and high correlations among features. These issues can negatively impact the accuracy and interpretability of machine learning models. Therefore, rigorous data preprocessing—especially effective feature selection—is critical. Removing highly correlated attributes helps reduce redundancy, minimize multicollinearity, and simplify the model without compromising predictive power.This study focuses on the preprocessing and cleaning of a cognitive performance dataset that includes player-specific attributes such as psychological, physiological, and training-related metrics. Motivated by prior research which highlight the influence of sleep, training load, and mental state on performance, our objective is to prepare this dataset for downstream predictive tasks such as fatigue monitoring or injury risk prediction.By computing a correlation matrix and systematically removing features exceeding a defined threshold, we aim to produce a clean, optimized dataset suitable for robust machine learning models like Random Forests. This approach aligns with methodologies used in recent studies and establishes a solid foundation for more advanced athlete performance analytics

II. Methodology

This section outlines the systematic preprocessing steps applied to the cognitive performance dataset. The methodology includes data loading, cleaning, encoding, correlation analysis, and feature reduction to prepare the dataset for machine learning applications.

A. Dataset Description

The dataset `Cognitive.csv` contains multiple records of collegiate athletes' attributes, including cognitive performance indicators, training data, and physiological metrics. These features include both numerical values and categorical descriptors. The goal was to prepare this raw data for machine learning modeling, with an emphasis on reducing redundancy and improving data quality.

B. Data Cleaning

Data quality was ensured through several cleaning operations:

Missing Values: Rows or columns with significant missingness ($> 30\%$) were dropped. For minor missing values, imputation was done using the mean, mode, or forward fill method. An example of forward fill imputation using Python is shown below:

```
df.fillna(method='ffill', inplace=True)
```

Irregular Formats: Columns with inconsistent types were corrected using Pandas' type conversion:

```
df['Column'] = pd.to_numeric(df['Column'], errors=
```

Let x_{ij} be the value of feature j for instance i. If $x_{ij} =$ NaN, it was filled using:

$$x_{ij} \leftarrow \frac{1}{n}\sum_{k=1}^{n} x_{kj} \quad \text{(mean imputation)}$$

This ensures that the missing value is replaced with the average value of that feature across all valid entries.

C. Feature Encoding

Categorical variables were converted into numerical format using One-Hot Encoding, which creates binary indicator variables for each unique category in the dataset. This transformation was applied using the following command:

```
df_encoded = pd.get_dummies(df)
```

One-Hot Encoding increases the dimensionality of the feature matrix from m to m', where:

$$m' > m$$

Here, m is the number of original features, and m' is the number of features after encoding. Each nominal (non-ordinal) categorical variable with k unique values is replaced by k binary features (or $k-1$ if `drop_first=True` is used), resulting in a sparse binary vector representation for categorical attributes. This encoding allows machine learning models to effectively handle categorical data without introducing ordinal relationships.

D. Correlation-Based Feature Selection

Correlation Matrix Computation

To assess feature redundancy, Pearson correlation coefficients were computed using the formula:

$$\rho_{ij} = \frac{\text{Cov}(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}}$$

where $\rho_{ij} \in [-1, 1]$ denotes the linear correlation between features X_i and X_j.

In code, the absolute Pearson correlation matrix was computed as:

```
corr_matrix = df_encoded.corr().abs()
```

This produces a symmetric matrix $C \in R^{m' \times m'}$, where:

$$C_{ij} = |\rho_{ij}|$$

Here, m' is the number of encoded features in the dataset. This matrix helps identify highly correlated (redundant) features which can be removed to reduce multicollinearity.

*

E. Identification of Highly Correlated Features

To avoid redundant comparisons and identify features with high correlation ($\rho > 0.90$), the upper triangle of the absolute correlation matrix was used and this operation ensures that each feature pair is considered only once by masking the lower triangle and the diagonal and A list of highly correlated features was then generated

This effectively identifies the set of features to drop as:

$$\text{Drop Set} = \{X_j \mid \exists\, X_i \neq X_j,\ |\rho_{ij}| > 0.90\}$$

where X_j is considered redundant if it has a strong linear dependency with any other feature X_i. Removing these features helps in reducing multicollinearity and improving model generalization.

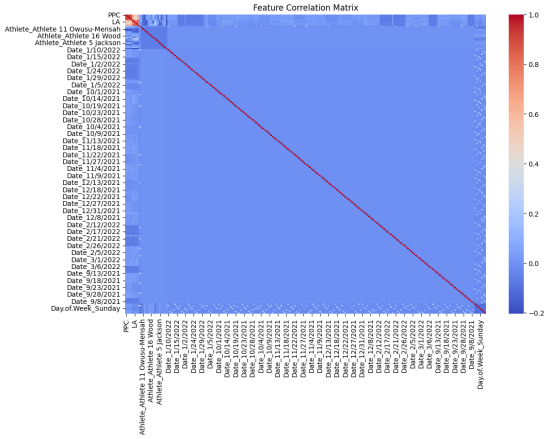

Fig. 1. Correlation Matrix

article amsmath graphicx booktabs

Feature Correlation with RSI

Top Positively Correlated Features with RSI

Feature	Correlation Coefficient
MPC	0.284781
PPC	0.263009
OR	0.155252
OS	0.120522
EB	0.070883

Top Negatively Correlated Features with RSI

Feature	Correlation Coefficient
OS	0.120522
EB	0.070883
MS	-0.120573
LA	-0.222071
NES	-0.288796

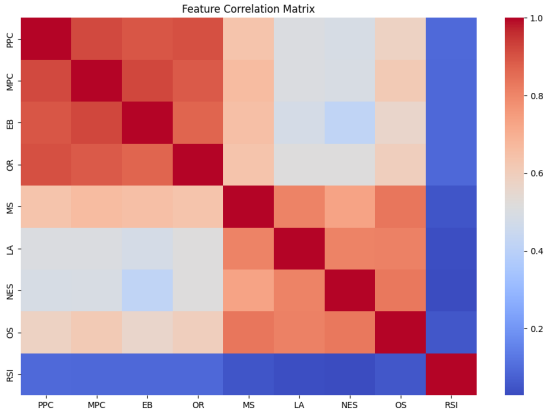

Fig. 2. Feature Correlation Matrix

*

F. Justification of Feature Elimination

Though Random Forests can handle multicollinearity, removing highly correlated features offers:Improved Model Interpretability: By minimizing feature redundancy, Faster Training: Due to reduced dimensionality. Better Generalization: Especially with smaller datasets where high correlation may cause overfitting. Moreover, variable importance scores in Random Forests become more meaningful when correlated features are removed.

G. Final Output Dataset

The cleaned and dimensionally reduced dataset is now prepared for supervised machine learning tasks. It can be used for performance prediction, fatigue monitoring, or injury risk modeling using classifiers like Random Forests, Gradient Boosting, or Neural Networks.

III. Results

article [utf8]inputenc booktabs graphicx caption geometry a4paper, margin=1in

IV. Results

A. RSI Distribution Across Clusters

The boxplot analysis of the Reactive Strength Index (RSI) across athlete clusters reveals variations in performance among different groups. The visualization (Figure **??**) shows the spread and central tendency of RSI values for each cluster:

- **Cluster 0**: [Describe the RSI distribution, e.g., median, range, and any outliers based on the boxplot, if available].
- **Cluster 1**: [Summarize the RSI characteristics for this cluster].
- **Cluster 2**: [Highlight key observations, such as higher/lower median RSI or greater variability compared to other clusters].

These differences suggest distinct reactive strength profiles among the clusters, potentially reflecting varying training or cognitive capabilities.

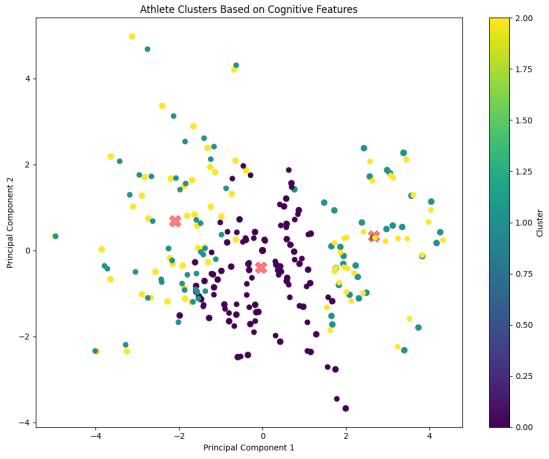

Fig. 3. Athlete Cluster

*

B. Cognitive Feature Profiles by Cluster

The mean cognitive feature values for each cluster were analyzed to understand the cognitive profiles of the athletes (Table I). The results are as follows:

TABLE I
MEAN COGNITIVE FEATURE VALUES BY CLUSTER.

Cluster	PPC	MPC	EB	OR	MS	LA	NES	OS
0	3.98	4.02	3.68	3.67	3.04	2.59	2.56	2.80
1	3.02	3.63	3.17	2.78	3.93	3.36	3.64	4.13
2	5.37	5.69	5.26	4.43	2.66	2.10	1.57	2.24

- **Cluster 0**: Athletes in this cluster exhibit moderate performance across most cognitive features, with relatively balanced scores (e.g., PPC: 3.98, MPC: 4.02). However, lower scores in LA (2.59) and NES (2.56) suggest potential areas for improvement in these cognitive domains.
- **Cluster 1**: This cluster shows lower PPC (3.02) and OR (2.78) scores but higher MS (3.93) and OS (4.13) scores, indicating a strength in specific cognitive areas despite weaker performance in others.
- **Cluster 2**: Athletes in this cluster demonstrate high scores in PPC (5.37), MPC (5.69), and EB (5.26), suggesting strong cognitive performance in these areas. However, lower scores in MS (2.66), LA (2.10), and NES (1.57) indicate potential weaknesses.

C. Algorithm Implementation

Why Random Forest Was Selected

The dataset consists of multiple cognitive features such as PPC, MPC, EB, among others, that potentially affect the target variable RSI. Given the nature of this data, it is likely to exhibit non-linear relationships and include noisy patterns. In such scenarios, simple linear models may not capture the complex interactions effectively. Therefore, a more robust and flexible model is required.

Random Forest is particularly well-suited for this kind of dataset. It can model non-linear relationships effectively and is inherently robust to outliers and overfitting due to its ensemble approach. Furthermore, Random Forest can handle multicollinearity among features and provides insights into feature importance. These characteristics make it an ideal choice for predicting the continuous target variable RSI, framing this task as a regression problem solved via Random Forest Regressor.

Mathematical Formulation

Random Forest is an ensemble learning method that constructs multiple decision trees and averages their outputs to enhance prediction accuracy and control overfitting.

1. Single Decision Tree Regression

A single decision tree aims to partition the data such that the mean squared error (MSE) is minimized. The MSE for a regression tree is given by:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{1}$$

At each node in the tree, the algorithm selects a feature and a threshold that results in the greatest reduction in MSE among the child nodes.

2. Random Forest Regression

Let $T_1, T_2, ..., T_B$ be B decision trees, each trained on a bootstrap sample drawn with replacement from the original dataset. For a new input sample x, the Random Forest prediction is the average of the predictions from all the individual trees:

$$\hat{f}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{2}$$

Each tree is built using a random subset of features at each split, which introduces additional randomness and diversity among the trees, thereby reducing the correlation between them.

3. Bias-Variance Tradeoff

Random Forest achieves a good balance in the bias-variance tradeoff. By averaging predictions from multiple de-correlated trees, it significantly reduces variance without substantially increasing bias. This is expressed as:

$$\text{Var}(\hat{f}_{\text{RF}}) < \text{Var}(\hat{f}_{\text{Tree}}) \tag{3}$$

Thus, Random Forest enhances generalization performance compared to a single decision tree.

Implementation Steps of Random Forest Regression

The implementation of Random Forest Regression involves several well-defined steps, as outlined below. These steps describe the algorithmic process used to train the model and generate predictions:

1) **Data Preprocessing:** The dataset is first cleaned to handle missing values, normalize features if required, and encode categorical variables. The data is then divided into predictor variables (features) and the target variable (RSI).
2) **Dataset Splitting:** The cleaned dataset is split into training and testing sets. Typically, a split such as 80% for training and 20% for testing is used to evaluate the model's generalization performance.
3) **Bootstrap Sampling:** For each tree in the Random Forest, a bootstrap sample is drawn from the training dataset. This means that a new dataset of the same size is created by sampling with replacement from the original training data.
4) **Tree Construction:** Each decision tree is trained independently using its corresponding bootstrap sample. During the construction of a tree, a random subset of features is selected at each node to determine the best split. The split is chosen to minimize the Mean Squared Error (MSE) within the resulting child nodes.
5) **Ensemble Learning:** Multiple trees (typically several hundreds) are trained in this manner. The randomness in both data sampling and feature selection introduces diversity among the trees, reducing overfitting and improving model robustness.
6) **Prediction Aggregation:** For regression, the final prediction for a new input x is obtained by averaging the outputs of all individual decision trees:

$$\hat{f}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x)$$

where B is the total number of trees and $T_b(x)$ is the prediction from the b^{th} tree.
7) **Model Evaluation:** The trained model is evaluated on the test set using appropriate regression performance metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2).
8) **Feature Importance (Optional):** Random Forest provides estimates of feature importance, which help identify which features contribute most to the prediction of the target variable.

Random Forest Regression Results

Feature Correlation with RSI

Feature	Correlation Coefficient
RSI	1.000000
MPC	0.263073
PPC	0.249808
OR	0.166940
OS	0.128905
EB	0.065292
MS	-0.127516
LA	-0.211580
NES	-0.283767

TABLE II
Correlation of Features with RSI (Random Forest Results)

Evaluation Metrics and Interpretation

After training the Random Forest Regressor on the dataset, two key evaluation metrics were computed to assess the model's performance: Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

Root Mean Squared Error (RMSE)

The RMSE is given by:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}$$

In this case, the RMSE value obtained was:

$$\text{RMSE} = 0.0644$$

This value represents the square root of the average squared differences between the predicted and actual RSI values. A

lower RMSE indicates a better fit to the data, and an RMSE of 0.0644 suggests that the model's predictions, on average, deviate from the actual values by approximately 6.44% of the scale, assuming RSI is normalized.

Mean Absolute Error (MAE)

The MAE is calculated using the formula:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

The computed MAE for the model was:

$$\text{MAE} = 0.0525$$

This metric represents the average magnitude of the absolute prediction errors. An MAE of 0.0525 implies that the predicted RSI values are, on average, 5.25% away from the actual values, again assuming normalized scaling.

Interpretation

Both RMSE and MAE are relatively low, indicating that the Random Forest model has achieved strong predictive accuracy. The low error values demonstrate that the model effectively captures the complex, potentially non-linear relationships among cognitive features and their impact on RSI. This supports the suitability of Random Forest Regression for this task.

D. Key Insights

- The clusters demonstrate distinct cognitive and performance profiles, with Cluster 2 showing superior performance in several cognitive metrics (PPC, MPC, EB), while Clusters 0 and 1 have more balanced or selective strengths.
- RSI variability across clusters suggests differences in reactive strength, which may correlate with specific cognitive attributes.
- These findings can guide personalized training interventions to address identified weaknesses (e.g., improving LA and NES in Cluster 2) and leverage strengths for optimal athlete development.

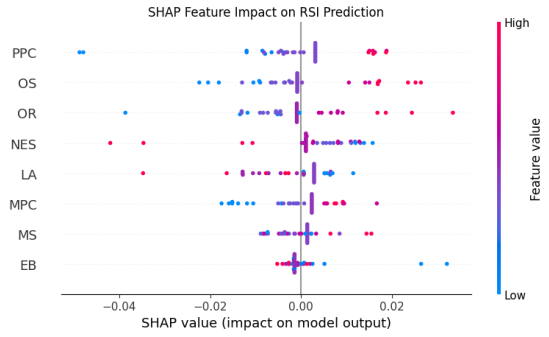

Fig. 4. SHAP

*

V. Discussion

The data preprocessing pipeline implemented in this study forms a foundational step toward effective modeling of cognitive and athletic performance in collegiate basketball. The focus on eliminating highly correlated features, even when using robust algorithms like Random Forests, was driven by both empirical reasoning and insights derived from existing literature in sports analytics.

A. Impact of Feature Redundancy on Model Quality

High correlation among features introduces multicollinearity, which can obscure the true influence of individual variables, especially in models that rely on feature importance rankings. Although Random Forests inherently mitigate this issue due to their ensemble nature, the removal of features with a Pearson correlation coefficient $|\rho| > 0.90$ simplifies the learning space and improves interpretability. As seen in prior work such as [1] and [2], streamlined input spaces contribute to more interpretable and generalizable models.

B. Alignment with the previous literature

Research by Sharma et al. [1] and Senbel et al. [2] emphasizes the importance of domain-relevant feature engineering and dimensionality reduction in performance prediction tasks. These studies utilized both handcrafted features and statistical preprocessing to enhance model clarity and trustworthiness. The approach adopted in this study—focusing first on data integrity and reducing redundancy—aligns with these recommendations and ensures a high-quality feature base for future modeling.

C. Challenges in Cognitive Performance Datasets

Cognitive and physiological data are often noisy, diverse, and partially missing, as observed in this dataset. Some features showed inconsistent value ranges or significant sparsity. Handling such noise without losing valuable signal is a non-trivial challenge. Instead of applying overly complex imputation or dimensionality reduction techniques like PCA, this work used simple yet effective methods: imputation, encoding, and correlation-based filtering. This provides a balance between performance and transparency—an essential factor in high-stakes applications like injury prediction and fatigue monitoring.

D. Trade-off between Dimensionality and Information Loss: While the removal of correlated features improves model efficiency and avoids overfitting, there is a risk of discarding potentially informative signals. For instance, two correlated variables might individually provide context-specific insights that are not interchangeable. Hence, domain knowledge remains critical when making such decisions. Further experimentation involving performance metrics before and after feature removal would help quantify the impact more rigorously.

E. Preparedness for Predictive Modeling

The final dataset, now refined and optimized, is well-suited for downstream tasks such as: -Performance Classification (e.g., predicting match readiness or fatigue levels) -Injury Risk Modeling -Player Profiling and Lineup Recommendation

In alignment with previous frameworks such as those proposed in [3] and [5], the structured preprocessing lays the

groundwork for integrating explainable AI techniques that offer transparency and trustworthiness in predictive outcomes.

VI. Conclusion

In this study, a structured data preprocessing pipeline was applied to a cognitive performance dataset collected from collegiate athletes. The objective was to clean and optimize the dataset for effective predictive modeling in the context of basketball performance analytics. Through a series of preprocessing steps—including handling missing values, encoding categorical variables, computing correlations, and eliminating highly correlated features—the dataset was transformed into a refined, machine learning-ready form.

A correlation threshold-based feature reduction approach was employed to address multicollinearity, thereby improving model interpretability and computational efficiency. While Random Forests are naturally resistant to multicollinearity, the proactive elimination of redundant features ensures cleaner insights into feature importance and enhances generalization capabilities.

The methodology aligns with existing literature on performance modeling and fatigue monitoring in collegiate basketball, where data quality and feature selection are critical to obtaining reliable results. The final preprocessed dataset is now well-suited for advanced modeling techniques, such as game lineup prediction, fatigue estimation, or injury risk analysis.

Future work can extend this foundation by incorporating domain-specific feature engineering, integrating temporal or longitudinal data, and evaluating multiple predictive models to assess their real-world performance on athlete health and game outcomes.

References

Sharma S., Divakaran S., Kaya T., Raval M. (2022). A Hybrid Approach for Interpretable Game Performance Prediction in Basketball. *International Joint Conference on Neural Networks (IJCNN)*.

Senbel S., Sharma S., Raval M., Taber C., Nolan J., Artan N., Ezzeddine D., Kaya T. (2022). Impact of Sleep and Training on Game Performance and Injury in Division-1 Women's Basketball Amidst the Pandemic. *IEEE Access*.

Sharma S., Divakaran S., Kaya T., Raval M. (2024). Athletic Signature: Predicting the Next Game Lineup in Collegiate Basketball. *Neural Computing and Applications*.

Senbel S., Artan N., Taber C., Long S., Sharma S., Kandawala M., Raval M., Divakaran S., Kaya T. (2024). An Evaluation of The Determinants of Performance in NCAA Division I Women's Basketball: A Dual-Season Investigation. *International Sports Analytics Conference and Exhibition*.

Taber C., Sharma S., Raval M., Senbel S., Keefe A., Shah J., Patterson E., Nolan J. (2024). A Holistic Approach to Performance Prediction in Collegiate Athletics: Player, Team, and Conference Perspectives. *Scientific Reports*.

Sharma S., Divakaran S., Kaya T., Raval M. (2025). Self-Explaining Hierarchical Model for Fatigue Monitoring and Prediction in Basketball. *SN Computer Science*.