CSE 523 ML Project

Faculty mentor:
Prof. Mehul Raval

Project Number: 9
Project Title: Data-driven imputation scheme for human-subject-based dataset
Group name: Syntellect
Group member details:

ENROLLMENT NUMBER	NAME
Kashish Jethmalani	AU2140029
Srushti Thakar	AU2140117
Priyal Patel	AU2140204
Riya Patel	AU2140214

Progress report : (Week 5)
Date: 8-02-2024 to 14-02-2024

Feature Significance:

After applying feature selection, we found that 'rsi' (Relative Strength Index) and 'hrv' (Heart Rate Variability) emerged as the most significant features in predicting the target variable. These features were selected based on their high scores, indicating their strong correlation with the target variable. 'rsi' represents the relative strength of the player, which is a crucial factor in determining their performance. 'hrv', on the other hand, reflects the variability in heart rate, which can be indicative of the player's physiological state and readiness.

K-Means Clustering

```python
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt

# Read the dataset
data = pd.read_csv("season 3.csv")  # Replace "athlete_dataset.csv" with your actual dataset filename

# Select relevant columns
selected_columns = ["RSI.Mean", "RSI.SD", "RSI.CoV", "HRV"]
data_selected = data[selected_columns]

# Handle missing values (if any)
data_selected.dropna(inplace=True)

# Standardize the data
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data_selected)

# Apply k-means clustering
kmeans = KMeans(n_clusters=4, random_state=42)
kmeans.fit(data_scaled)
clusters = kmeans.predict(data_scaled)

# Add cluster labels to the dataframe
data_selected['Cluster'] = clusters

# Calculate the number of dimensions and adjust the number of subplots accordingly
num_dimensions = len(selected_columns)
num_rows = num_dimensions
num_cols = num_dimensions

# Visualize the clusters
fig, axs = plt.subplots(num_rows, num_cols, figsize=(12, 12))

# Plot clusters for all combinations of dimensions
for i, dim1 in enumerate(selected_columns):
    for j, dim2 in enumerate(selected_columns):
        ax = axs[i, j]
        ax.scatter(data_selected[dim1], data_selected[dim2], c=data_selected['Cluster'], cmap='viridis')
        ax.set_xlabel(dim1)
        ax.set_ylabel(dim2)
        ax.set_title(f'Clusters based on {dim1} and {dim2}')

plt.tight_layout()
plt.show()
```

1. Data Preparation

Data Loading: The code uses Pandas to load the dataset from a CSV file named "Season 3.csv".

Feature Selection: It selects specific columns ("RSI.Mean", "RSI.SD", "RSI.CoV", "HRV") from the dataset, which are deemed relevant for the clustering analysis.

2. Data Preprocessing

Handling Missing Values: Rows with missing values in the selected columns are dropped to ensure the quality of the data.

Data Standardization: The selected data is standardized using StandardScaler to bring all features to a similar scale, which is a common practice in clustering algorithms.

3. Clustering

Applying K-Means: The standardized data is then clustered using the KMeans algorithm with `n_clusters=4` to identify four distinct clusters. The random state is set for reproducibility.

4. Visualization

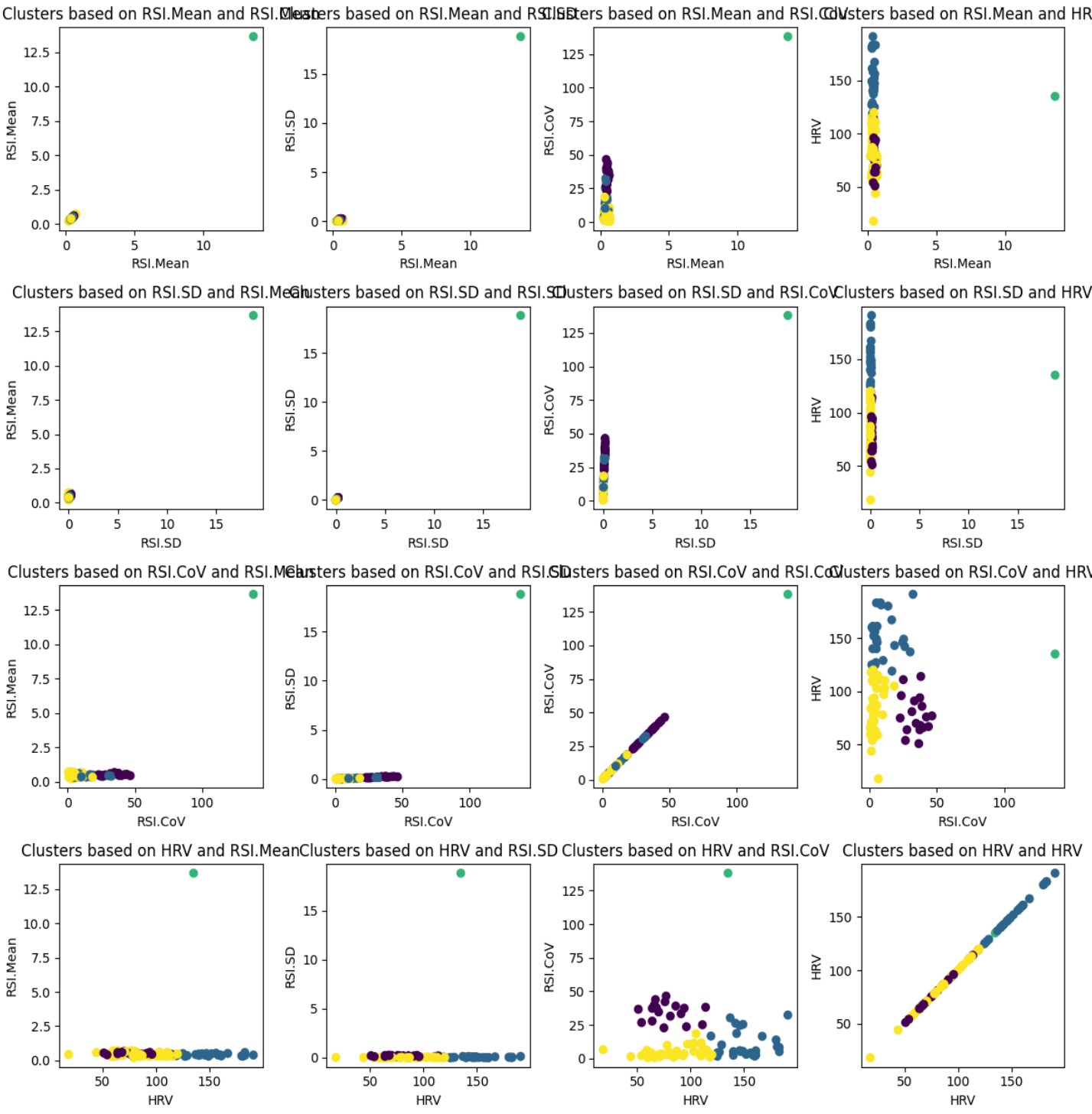
Visualizing Clusters: The clusters are visualized using a scatter plot matrix, where each subplot represents a combination of two dimensions from the selected columns. Data points are colored according to their assigned cluster, allowing for a visual understanding of the clustering results.

5. Results

- The clustering results show that the data can be grouped into four distinct clusters based on the selected features.
- Interpretation of the clusters can be further analyzed to understand each cluster's characteristics and implications.

6. Conclusion

- The clustering analysis using k-means provides insights into the underlying structure of the data, revealing distinct patterns and similarities among data points.
- Further analysis and interpretation of the clusters can lead to valuable insights for decision-making and future research.


Clusters based on RSI.Mean and RSI.Mean
Clusters based on RSI.Mean and RSI.SD
Clusters based on RSI.Mean and RSI.CoV
Clusters based on RSI.Mean and HRV
Clusters based on RSI.SD and RSI.Mean
Clusters based on RSI.SD and RSI.SD
Clusters based on RSI.SD and RSI.CoV
Clusters based on RSI.SD and HRV
Clusters based on RSI.CoV and RSI.Mean
Clusters based on RSI.CoV and RSI.SD
Clusters based on RSI.CoV and RSI.CoV
Clusters based on RSI.CoV and HRV
Clusters based on HRV and RSI.Mean
Clusters based on HRV and RSI.SD
Clusters based on HRV and RSI.CoV
Clusters based on HRV and HRV
RSI.Mean
RSI.SD
RSI.CoV
HRV