CPS 803 - Assignment 4

1) Background :-

The dataset used in this is “Rocket League Skill-shots”, which contains data on various in-game performance metrics of players in the esports tournament of the game Rocket League. This dataset provides quantitative measures of a player’s interaction with the game environment, including metrics such as BallAcceleration, PlayerSpeed, and Distance from objects (Wall, Ceiling, Ball) during gameplay.

I chose this dataset because it gives valuable insight not only the players but also the game developers about player behaviours identifying strengths and weaknesses, their play styles and areas of improvement of the game. Also, as a fellow player of the game, there are a few problems with the game physics for example which part of the car hits the ball and where it goes etc. ,that I would like to see changed so I’ve chosen this as my topic.

Key Concepts:-

1. BallAcceleration, PlayerSpeed, and BallSpeed : These features represent the dynamics of player and ball movement during gameplay, which are critical in high-paced matches.

2. Distance Metrics : DistanceWall, DistanceCeil, and DistanceBall provide spatial context, allowing us to understand positional strategies used by players.

3. Clustering : K-Means and DBSCAN clustering techniques are applied to group similar player behaviours or play styles.

2) Methodology:-

1. Data Loading

2.Data Preprocessing (Missing values, Feature Selection, Scaling)

3. Clustering i.(K-Means,Elbow Method, Assign Clusters) ii.(DBSCAN, Assign Clusters)

4. Visualization (Elbow Curve, Scatterplots, Heat maps)

5. Evaluation (Silhouette Score for K-Means and DBSCAN)

1. Load the dataset using pandas.

2) Data Preprocessing :-

Objective: Clean and prepare the data for clustering.

1. Handle Missing Values: missing numerical values filled with column medians.
2. Select Relevant Features: Chose numerical columns relevant to clustering, such as BallAcceleration, PlayerSpeed, and Distance metrics.

3. Scale the selected features.

3) Clustering : Two techniques: K-Means and DBSCAN.

a) K-Means Clustering:

Objective: Partitioned the data into k distinct clusters.

1.Tested k values from 1 to 10.

2.”Used the Elbow Method to determine the optimal k by plotting the distortion (sum of squared distances from each point to its cluster centre) against the number of clusters.[2]”

3.Find optimal K and assigned cluster labels

b) DBSCAN Clustering:

1. Used DBSCAN with specified eps and min_samples.

2. Assigned cluster labels, including -1 for noise points.

4) Evaluation for Clustering:

•Silhouette Score : Measures how similar points in a cluster are to each other, compared to other clusters. “Evaluates cluster quality by measuring the cohesion within clusters and separation between clusters[1]”:

S = b-a / max(a ,b)

where:

a: Mean intra-cluster distance : average distance between a point and others in its cluster

b: Mean nearest-cluster distance : average distance between a point and the nearest other cluster.

“Higher Silhouette Scores indicate better-defined clusters. Scores near +1 suggest tight clustering, while scores near 0 indicate overlapping clusters.[3]”

3) Results :-

1) Elbow Method (K-Means) :-

A line plot (Elbow Curve) showing the distortion values for k values ranging from 1 to 10. “The optimal number of clusters (k) is identified at the “elbow” point, where the distortion begins to level off.[4]”

2) K-Means Clustering :-

a) Silhouette Score for K-Means indicates the quality of the clusters.

b) A scatter plot of BallAcceleration vs. PlayerSpeed, with points colour-coded by their K-Means cluster assignments.

c) Assign Cluster columns their labels.

3) DBSCAN Clustering :-

a) Silhouette Score for DBSCAN clustering, indicates the quality of the clusters.

b) A scatter plot of BallAcceleration vs. PlayerSpeed, with points colour-coded by their DBSCAN cluster assignments. Noise points are labeled -1.

Metrics for silhouette :-

1) K-Means: A high value (close to +1) indicates well-separated clusters.

2) DBSCAN: “A higher score indicates better-defined density-based clusters.[1]”

4) Conclusion :-

The Elbow Method determines the optimal number of clusters (k) by analyzing the distortion values plotted against k. In this case, “the “elbow” suggests that k=3 is the optimal choice, therefore dataset split into 3 distinct clusters.[2]” The heat map of cluster feature averages reveals the distinct characteristics of each cluster. For instance:

One cluster might exhibit high BallAcceleration and PlayerSpeed, indicative of aggressive or fast-paced players.

The Silhouette Score for K-Means is higher compared to DBSCAN, indicating that K-Means is better suited for this dataset. Scatterplots observations for both K-Means and DBSCAN:

A) K-Means shows clear and well-separated clusters.

b) DBSCAN highlights noise points and groups, but clusters are less distinct.

Summary :-

K-Means clustering is the more effective method for this dataset, segmenting players into three well-defined clusters with distinct play styles. DBSCAN is valuable for detecting noise. The combination of Silhouette Scores and visualizations (Elbow Curve, Scatterplots, Heat maps) supports these conclusions.

References :-

[1] Hastie, T., Tibshirani, R., & Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media, 2009.

[2] https://towardsdatascience.com/the-elbow-method.

[3] https://fiveable.me/key-terms/data-visualization/silhouette-score

[4] https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans