DOKUZ EYLÜL UNIVERSITY

ENGINEERING FACULTY

DEPARTMENT OF COMPUTER ENGINEERING

CME4416 - INTRODUCTION TO DATA MINING

STEAM USER ANALYSIS FINAL REPORT

by

2017510078 İhsan Batuhan UZ

2017510034 Furkan GÖKCAN

January, 2022

İZMİR

INTRODUCTION

Today, Steam is the most used environment among frequently used Gaming Platforms. Digital rights management (DRM), server hosting, video streaming, and social networking services are all available through Steam. It also includes community features such as friends lists and groups, cloud storage, and in-game voice and chat functions, as well as game installation and automatic updates. Here, we are using a steam dataset to accomplish an accurate and multifunctional analysis as much as possible, as well as consistent with real life. Clustering is one of the methods used in order to organize and optimize the given data set. The instances are clustered so that the intraclass similarities are maximized and the interclass similarities are minimized. This is indicated according to the criteria defined on the attributes of the instances. In target data, after the datasets are arranged and data reduction processes are performed, then the known clustering methods k-means, Agnes and DBScan will be used. The dataset used is quoted via
“https://www.kaggle.com/calven22/steam-recommender-system/data”

ALGORITHMS

The method that we will follow to study steam user data is also to compare the results on a tool using three different algorithms. The three algorithms to be used are the K-Means algorithm, Agnes and DBSCAN, which are included in the clustering structure. K-means clustering method is to partition a data set consisting of N data objects into K clusters given as input parameters. AGNES, standing for Agglomerative Nesting, means that some clusters are merged until all clusters have been merged into one big cluster containing whole objects.DBSCAN algorithm can identify the outliers.Its full form is Density based spatial clustering of application with noise.DBSCAN has two basic features; epsilon and min points. Python will be used as the tool while the algorithms are utilized. Using python with the dataset that we have, we aim to use these three algorithms to plot the results both in a table and in a graph.

DATA REDUCTION

The “user id” and “other” attributes contained in the dataset will be removed because they are not necessary for the analysis to be performed. The behaviors attribute contains 'purchase' and 'play'. The value specifies the degree to which the behavior was performed - in the case of 'purchase', "hours_played" attribute value is always 1,in the case of 'play' the value represents the number of hours the user has played the game as well. Therefore, the “purchase” statements contained in the Behavior attribute will not be included in the dataset review.

TOOLS AND ALGORITHMS

Python was used as a tool to perform our study. We have tried to test three algorithms that we have determined via the Pycharm ide.These algorithm techniques are; Kmeans, Dbscan and Agnes.Before applying these techniques, it was necessary to make sense of dataset because the conclusion we wanted to reach was to examine whether people who have a lot of games are addicted to games or not.Because we cannot conclude our theory by using the attributes contained in this dataset alone. In order to make sense of the dataset, we tried to group the variables contained in the dataset before applying the three cluster algorithms we have specified.

In order to be able to clearly demonstrate these three algorithms in pycharm, we have divided the data into six parts for the data set to display the K Means algorithm.A train set was created and data was executed using the k means algorithm. For the Dbscan algorithm, the min point and epsilon parameters were set.As before, this algorithm was tested in the new train set created in the same way. Finally, when applying the Agnes algorithm, we divided the train set into 3 labels and visually converted the result into a graph.

TESTING AND RESULTS

When we examined the data set that we had determined earlier in the testing process, there was not much clarity about the variable names (Figure 1).

	A	B	C	D	E
1	151603712,"The Elder Scrolls V Skyrim",purchase,1.0,				
2	151603712,"The Elder Scrolls V Skyrim",play,273.0,0				
3	151603712,"Fallout 4",purchase,1.0,0				
4	151603712,"Fallout 4",play,87.0,0				
5	151603712,"Spore",purchase,1.0,0				
6	151603712,"Spore",play,14.9,0				
7	151603712,"Fallout New Vegas",purchase,1.0,0				
8	151603712,"Fallout New Vegas",play,12.1,0				
9	151603712,"Left 4 Dead 2",purchase,1.0,0				
10	151603712,"Left 4 Dead 2",play,8.9,0				
11	151603712,"HuniePop",purchase,1.0,0				
12	151603712,"HuniePop",play,8.5,0				
13	151603712,"Path of Exile",purchase,1.0,0				
14	151603712,"Path of Exile",play,8.1,0				
15	151603712,"Poly Bridge",purchase,1.0,0				
16	151603712,"Poly Bridge",play,7.5,0				

steam-200k

Figure 1 Raw data

Attribute names have been created for the instances included in the data in order to fulfill the goals we have set (Figure 2).

```
df = pd.read_csv("steam-200k.csv", header=None, index_col=None, names=['UserID', 'Game', 'Action', 'Hours', 'Other'])
```

Figure 2 Reading and grouping of data

We have previously stated that data reduction operations will be performed during the planning stage. This study is also necessary if we want to achieve an easier and cleaner result. For this purpose, the process of clearing the newly specified variable columns and values was performed (Figure 3).

```
df = pd.read_csv("../input/steam-200k.csv", header=None, index_col=None, names=['UserID', 'Game', 'Action', 'Hours', 'Other'])
del df['Other']
df.head()
```

	UserID	Game	Action	Hours
0	151603712	The Elder Scrolls V Skyrim	purchase	1.0
1	151603712	The Elder Scrolls V Skyrim	play	273.0
2	151603712	Fallout 4	purchase	1.0
3	151603712	Fallout 4	play	87.0
4	151603712	Spore	purchase	1.0

Figure 3 Deleting “Other” attribute in dataset

Some attribute were removed from the data set, while other attributes were combined to create a new group. At the beginning of the testing process, the “purchase” and “play” values included in the Action attribute were called a separate criterion under the name of purchased_games and played_games.

Then we calculate the number of games played by a user for user analysis, which is the main purpose of our topic (Figure 4).

```
user_counts = df_played_games.groupby('UserID')['UserID'].agg('count').sort_values(ascending=False)
```

Figure 4 Computing the number of games that user played

In the same way, the game time played by the user is calculated (Figure 5).

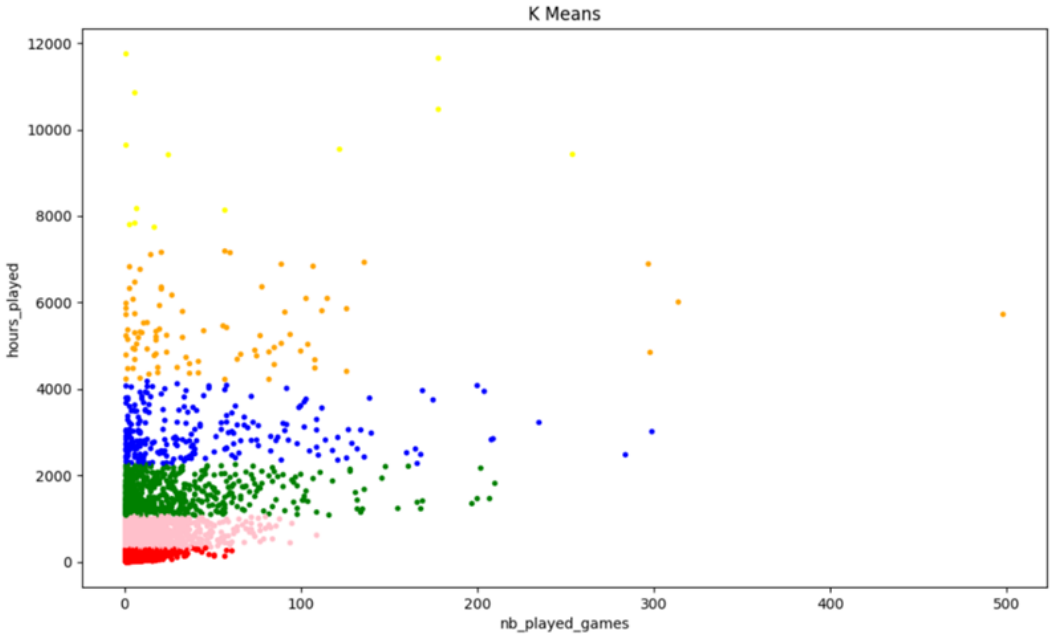
```
hours_played = df_played_games.groupby('UserID')['Hours'].agg(np.sum).sort_values(ascending=False)
```

Figure 5 The number of hours that has played

Then, a new data set is created by combining these two variables. At this point, we have designed the three cluster algorithms that we have determined with this data set so that we can test it.

K-MEANS

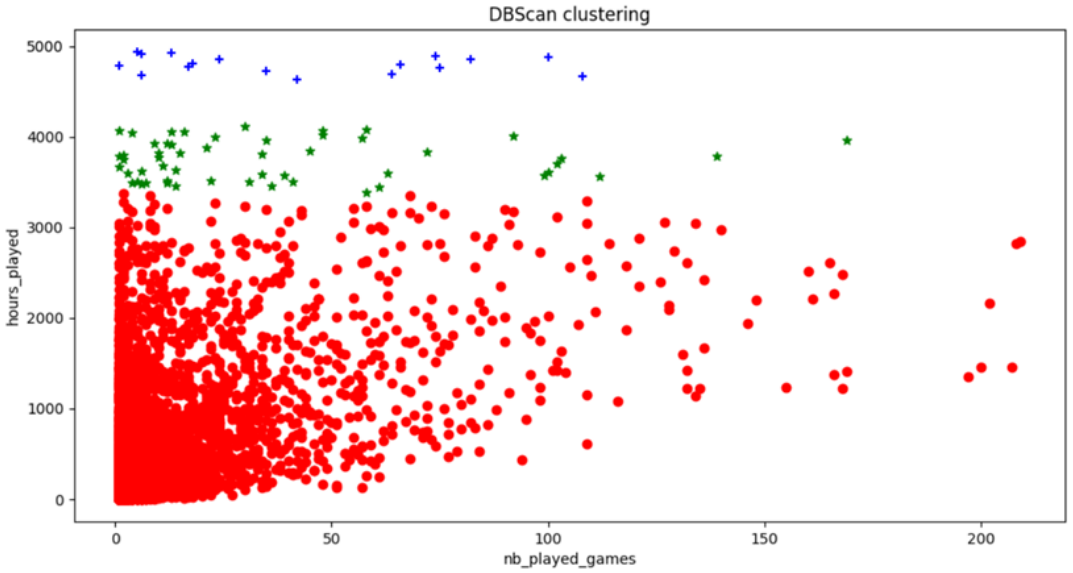
According to the working mechanism of the K-means algorithm, K objects are randomly selected to represent the center point or mean of each cluster. The remaining objects are included in the clusters with which they are most similar, taking into account their distance from the mean values of the clusters. Then, by calculating the average value of each cluster, new cluster centers are determined and the distances of the objects to the center are examined again. The algorithm continues to repeat until there is no change. The following results belong to k-means algorithm of number of played games and time spent to the games by users in the used dataset.



DBSCAN

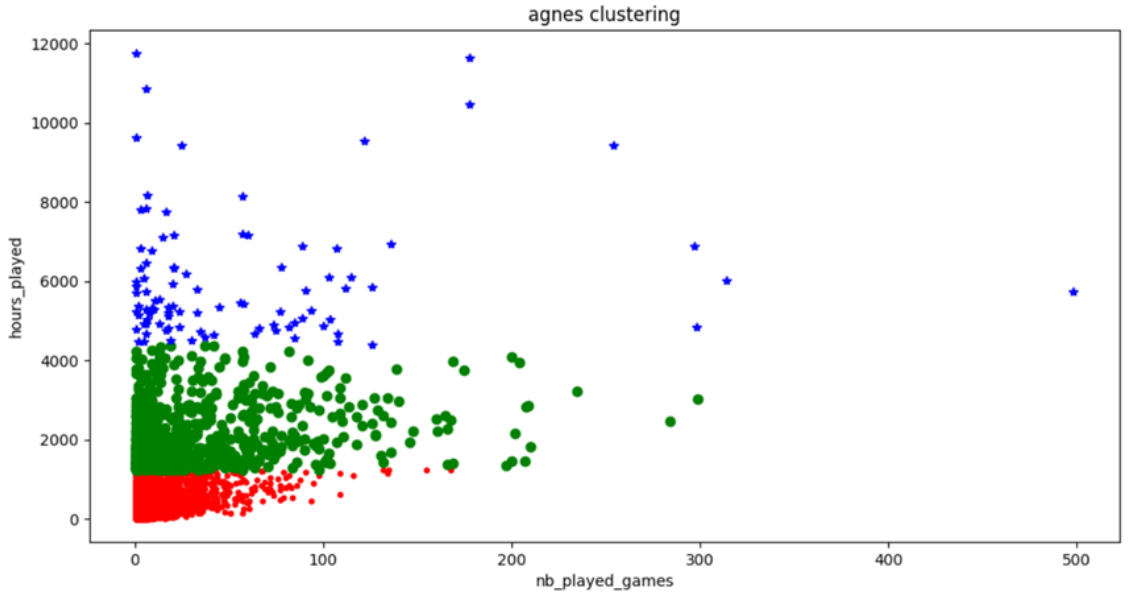
MinPts and Eps parameters must be declared for the DBSCAN algorithm to work.
The algorithm first chooses a random point p. MinPts and Eps to point p
the density according to its values finds all accessible points, if p is the seed point
If it satisfies the condition, a new cluster is discovered. density to point p
The same procedure is applied by taking all accessible points one by one, if any
If a point does not satisfy the seed point condition, this point is the boundary point of the set.
When none of the examined points satisfy the kernel point condition
the boundaries of the set are determined. The algorithm is the same by choosing a new random point.
repeats operations. If the randomly selected point satisfies the kernel point condition

If not, this point is defined as noise or exception. The DBScan algorithm results are implemented below



AGNES(Agglomerative Nesting)

The AGNES algorithm follows a bottom-up construction structure. Initially, each object is considered a separate set. At each subsequent step of the algorithm, those atomic clusters with similar properties are combined. After each merge operation, the total number of clusters decreases by one. Following plot is inserted for the AGNES algorithm.



The k-means algorithm is an algorithm that gives good results when working with large data, spends little time and consumes less memory. But, this algorithm is very sensitive to outliers and noise. It is needed to give a predetermined k value. As a result, K-Means and AGNES algorithms have the similar results, but, DBScan algorithm has a result different from other algorithms.