CMP 7203 BIG DATA MANAGEMENT

EVALUATION OF BIG DATA PROCESSING PARADIGMS AND ANALYSIS OF THE “CATCH THE PINK FLAMINGO” GAME

BY

HARDIK NEERAJ JAIN
STUDENT NO**: 23111876**
MSc BIG DATA ANALYTICS

BIRMINGHAM CITY
University

SUBMITTED May 18, 2023

Contents

1 Introduction :-

The analysis looks at how well the users can use the data techniques like data processing, decision-making, and analytics in a real simulation. Big Data han- dles the challenges like velocity, volume, and variety the main components of it. Big data paradigms, exploratory data analysis (EDA), machine learning with two-algorithm classification and clustering analysis, and graph analysis to implement on the data set file of the 'Catch the pink flamingo' game report.

The report performance of the game and gain insight into the behavior. Notwith- standing these specialized viewpoints is additional research. The systematiza- tion of concepts of true and false behavior in relation to data, particularly to data referred to in the big data. The evaluation's findings demonstrated how well the big data concepts worked to raise their gaming performance. Making the best judgment and being able to forecast the condition of the game was made possible by the machine learning algorithms. The player's capacity for goal-achieving increased with the application of graph analysis.

2 Big Data Processing Paradigms :-

2.1 Big Data Background

In Big data may learn more about customer behavior, spot patterns, and make wiser decisions with the help of the data provided. For the examination of large amounts of data, a variety of tools and techniques are available like analysis, ML, and AI (Jain et al., 2016). These techniques may be used to identify data patterns and trends that are difficult to spot with the human eye. Information is an incredible asset that can be utilized in various ways. The ethical and responsible is that enor- mous amounts of information aren't utilized to oppress individuals and ensure that big data is utilized only for the benefit of the community (Casado et al., 2014).

The three main elements of big data are in Figure-1. Volume: Huge, informative collections can be particularly difficult to keep and manage due to their size (Casado & Younas, 2015). Due to factors including the expansion of digital devices, social media,and the digitalization of numerous processes, the volume of data has drastically expanded. A huge amount of patient information, sensor data, and medical imaging are produced by healthcare systems (Sagiroglu & Sinanc, 2013).

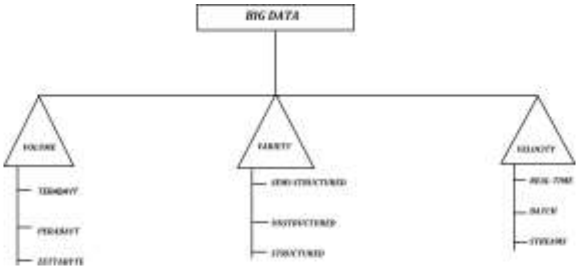


Figure 1: BIG DATA Elements

Variety: Since huge data sets might originate from a number of different sources, integrating and analyzing them can be difficult. A vast variety of organized, unstructured data, and semi-structured data, including posts on social media, sensor data, and more, are dealt with by the organization for the development of big data. Online reviews, customer service requests, and social media sources are all sources of consumer feedback (Hitzler & Janowicz, n.d.).

Velocity: Due to the speedy creation of enormous data sets, it is challenging to keep up with the rate of data collection. When businesses make rapid decisions and react quickly to changing conditions, velocity is critical. It centers around the pace of information ingestion. Companies in the transportation industry use data to optimize routes and monitor their location and status in real-time (Birke et al., n.d.).

Big data is notable for its volume, variety, and velocity. The size of a volume, variety of data types and sources, and speed of velocity that data is processed.

2.2 Batch Processing Paradigms

:-

In big data processing, where massive amounts of data are processed and evaluated utilizing distributed systems, batch processing is a popular technique. Batch processing is frequently employed for lengthy, computationally demanding tasks that call for the processing of massive data sets. Tools used by batch processing is a well-known open-source platform for distributed batch processing called Hadoop (Shahrivari, 2014). Large data sets may be processed in parallel over a cluster of computers using the distributed file system HDFS and the processing framework MapReduce provided by this system (Gheisari et al., 2022).

Pros of Batch-Processing:

! *Cost-effectiveness*: Large volumes of data can be processed efficiently and cheaply with batch processing.

! *Flexibility*: A variety of data processing operations, including data cleansing, transformation, analysis, and machine learning, can be carried out using batch processing frameworks.

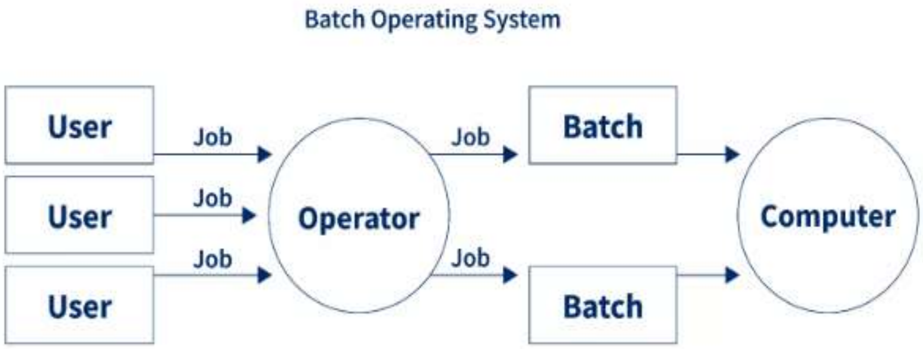
! *Scalability* : Horizontal scaling refers to the ability of batch processing frameworks like Apache Hadoop and Spark to handle massive data volumes by adding additional compute nodes to the cluster.

Cons of Batch Processing:

! *Limited interactivity*: Data is read in, processed, and written out in a batch process, which operates in a one-way fashion.

! *Resource requirements* : Batch processing frameworks need a lot of computational power to handle huge amounts of data effectively, such as powerful servers or computer clusters.

Uses of Batch Processing : - To handle a variety of data request kinds, batch processing systems are employed. It can be computationally expensive and wasteful to handle individual data transactions from data processing operations including backups, filtering, and sorting. Instead, data systems handle these jobs in batches at the end of each day and provide them to the team in charge of order fulfilment in a single batch (Deshpande & Rao, 2022).



Example of Batch Processing:- A batch processing system is being used by a social media platform to look for signs of abuse in user activity. The system stores data in a database from every user activity on the platform in real-time. The framework then runs a clump work consistently to investigate the information for indications of misuse. The batch job will remove any abusive content from the platform and notify the platform's abuse team if it finds any.

2.3 Real-Time Processing Paradigms

Real-time processing is the capacity to process data as it comes in, in close to real-time or in real-time, and to respond immediately depending on the outcomes (Voyvodic, 1999). Real-time processing denotes the immediate, uninterrupted processing and analysis of data. Continuous data entry, processing, and output are all part of real-time processing.

Benefits:

! Processing data happens with very little latency.

! Information is current and used right away.

! Fewer resources would be required for synchronization systems.

! Your uptime has risen. It aids in problem identification so you canrespond right away.

Drawbacks:

! Simple systems do not make it easy to implement.

! It cost a lot and necessitates powerful gear.

! When a system fails, it contributes an abundance of data.

Uses of Real-Time Processing :- All sectors in today's marketplacescan profit from real-time processing. With a rising emphasis on big data, this method of processing and gaining insights may propel businesses to new heights of success (J@zy, n.d.). Banking systems, data streaming, customer service systems, andwealth radars are a few examples of real-world uses in the process of real-time.

Example :- Real-time processing is critical for sectors like banking, traffic management, weather radars, and customer service organizations where prompt processing is required. It offers current knowledge that is instantly applicable. To sustain real-time insights, it does, however, need a constant flow of data input and output, which makes it more difficult than batch-processing systems (Ameri et al., 2014).

2.4 Hybrid Processing Paradigms

Introduction :- Hybrid processing began with the meaning of Lambda design, which has been executed in both classical and modern industrial studies in various ways (Beygelzimer et al., 2015). In big data, the hybrid processing paradigm refers to a strategy that combines real-time and batch processing. The half-handling worldwide has been utilized to make sense of many mental peculiarities, including consideration, memory, language, discernment, and thinking. Additionally, it has been using in the creation of cognitive disorders like schizophrenia (Yun & Epstein, 2012).

▪ Advantage and Disadvantage

Advantage:

! Hybrid processing enables organizations to reduce costs by utilizing a combination of different technologies.

! Hybrid processing also increases the accuracy of data processing operations, by utilizing a combination of different technologies.

Disadvantage:

! Implementing and maintaining specialized infrastructure and technologies for hybrid processing can be costly.

! Combination difficulties might emerge while consolidating batch and continuous processing frameworks, and potential issues.

! The overall architecture of data processing is made more complex by hybrid processing, which may necessitate the management of additional expertise and resources.

Lambda Architecture The Lambda arch. is a data processing which was made to handle real-time, large-scale data processing. It provides a scalable, flexible, huge data set for analyzing (Alexandre da Silva et al., 2016). A tool that is used to analyze this is Apache Spark based on the hybrid structure.

Example of Hybrid Processing: The lambda architecture is used on social media platforms to process and analyze numerous user generation content and interactions. Involvement and content references can occur immediately which is the processing in real time of users such as posts, shares, comments, and likes (Higgins, 1999).

2.5 Comparison of Paradigms

SPECIFICATION	BATCH PROCESSING	REAL-TIME PROCESSING	HYBRID PROCESSING
Data Source	Large amounts of data from multiple sources are collected and stored in a data lake or data warehouse.	Streaming information created continuously from sensors, gadgets, and applications.	Combination of sources combining real-time and batch data sources.
Freshness of Data	It often takes some time before data can be analyzed after bulk processing.	Data is processed as soon as it arrives, thus there is little time between data input and processing.	By processing data almost instantly or with a small delay, it is feasible to achieve a balance between batch processing and real-time processing.
Data Processing	Data is processed in huge batches, with results generated at the conclusion of each batch.	Real-time results are produced and data is processed as it arrives.	Depending on the particular needs, data can be handled in batch or real-time.
Data volume	From terabytes to petabytes, large data volumes can be handled by batch processing.	A continuous stream of data, typically of the order of gigabytes per second, is handled by real-time processing.	Huge batches of batch data, as well as ongoing streams of real-time data, can both be processed using hybrid processing.
Latency	High-latency processing is planned on a daily, weekly, or monthly basis.	Processing occurs in real-time or close to real-time with low-cost latency.	Depending on the data source and processing needs, latency might change.
Tools and Technologies	Hadoop, Apache Spark	Apache Kafka	Lambda Architecture.
Examples	The company may conduct nightly reports to analyse the previous day's sales information.	The business wants to monitor sales data as it develops and respond fast to any alterations. The next day, the findings are made accessible.	The retail organisation seeks to mix past study with ongoing observation. They mix group handling gets close with continuous handling draws near.

3 Exploratory Data Analysis (EDA) :-

Exploratory Data Analysis addresses as 'EDA'. It is a procedure of checking a data set to summarize its essential characteristics, habitually using real representations and different data insight systems. Finding patterns in the data and understanding how the factors relate to one another is the goal of EDA (Milo & Somech, 2020). Ex- ploratory Data Analysis (EDA) is an important part to analyze data. It helpsto understand the data better, choose the right statistical methods for analysis, and spot potential problems with the data. The relationship between variables and processes in statistics can be identified with data visualization (Chen, 2017).

3.1 Flamingo Data Overview

The game 'Catch the Pink Flamingo' is to capture the 'Pink Flamingo' provided on the map for each level as possible by performing missions that are provided by the real-time prompts. The levels get more mix up in guide intricacy and mission speed as the clients move from one level to another. It is a multi-user game. Consideration of the game and users can be made easier with help the provided data set used in the game. It can be used to make well-informed choices regarding the legitimize, promote, and evolution (Pascute & Engineering., 2002).

The Data sets implemented in the following games are the 'Ad-clicks' data file, 'Users' data file, 'Game-clicks' data file, 'Level-Events' data file, 'Team and Team-Assignments' data file, 'Buy-Clicks', and 'Combined-Data' data file are used for the EDA visualization process. 'Combined-Data' data set file used for the Machine learning methods to perform 'Classification' and 'Clustering' Analysis. Graph Analysis is performed on the 'Chat-Data' data set file to visualize the 'Team-Chat' which contains, 'Join-chat', 'Respond-chat', 'Leave-Chat', and 'Mentioned-Chat' data files are implemented.

3.2 Data Pre-processing

An information investigation is performed to figure out the distinguish and information any issues or irregularities in the information. 'Data pre-processing' is a radical stage in the information study, which includes, investigating, obtaining, and setting up the information for further analysis.

In the following data or game, some pre-processed took place like adding the 'Weekday' column data in the 'User-session' file and in the 'Users' data file 'Weekday', 'Year', 'Age', 'Time-period' is added to it by transforming from the 'timestamp' and 'Date-of-Birth' values to extract the data on which day users are mostly active. The 'Team' and 'Team-Assignments' files are merged to find out the count as per team. The file to find out the most source platform is a combination of 'Game-clicks' and 'User-Session' data files.

3.3 Exploratory Data Analysis (EDA) Visualization

Data visualization is in the work to create visualization and provide data analysis to others outside the sector. Exploratory Data Analysis visualization may be used to spot outliers, patterns, connections, and data trends between variables. EDA is a method of data analysis that make extensive use of visual tools in order to make the most of the data, and its structure, and create low-cost models. It is also utilized to communicate with others the findings of data visualization.

To perform the Exploratory Data Analysis Visualization the data set are used in the given Table-1

Series No.	DATA-SET CATEGORY	DATA REPRESENT
1	level-event	Any time a team initiates a level in the game a line is added to file
2	Buy-clicks	This file is updated with a new line each time an in-app purchase is made.
3	Ad-clicks	When a player clicks on an advertisement in the Flamingo game application, a line is added to this file.
4	Users	There is a line in this file for each person playing the game. Moreover, new sections with weekend days and time periods.
5	User-session	The record for each time a user starts or finishes a game is in this file.
6	Game-clicks	This document holds data from pretty much every one of the groups in the game.
7	Team	All of the game's information can be found in this file.
8	Team-assignment	Each time a user joins a team, a line is added to this file. A client can be in each group in turn.
9	Combined-data	All user data, including team, team level, game-click, buy-ID, average price, and player type, are combined in this file.

Table 1: Synopsis of the Data-set file.

3.3.1 Exploring squad circumstance and intensity.

The Figure-2 line chart depicts the time-varying team-level progression. In the year ‘2016-07’, team level ‘8’ reached the highest. In ‘2016-01’ team ‘1’ steadily increased until it was 8 in 2016-07 and for the remaining months the team 8 level remained top. The group-level movement with time shows that the group had to gain critical headway in a brief timestamp. The progression of the team’s level is a positive indication of the team’s future. The group level expanded by 7 out of 2016, July marked the team’s highest level, and January was noted as the lowest. The average level team is ‘6’, and the group level is reliable and consistent.

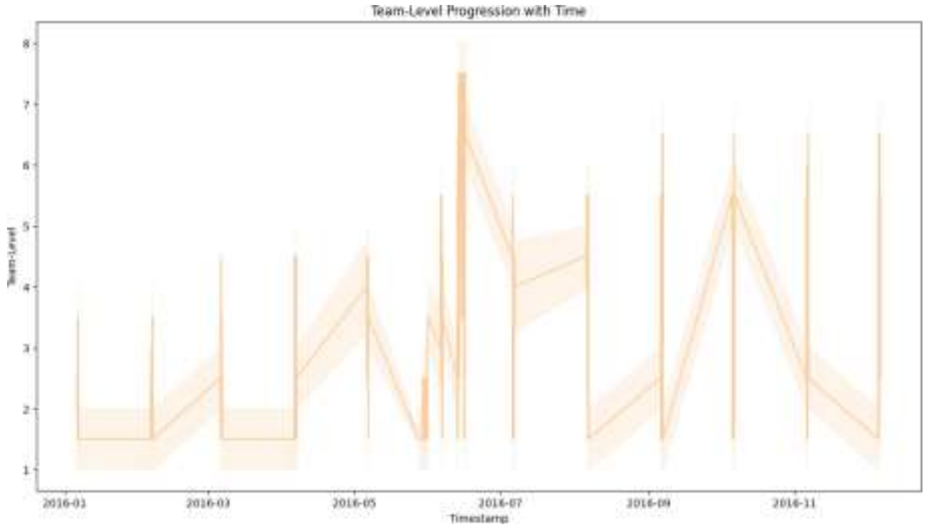


Figure 2: Team-Level Progression.

The team-level distribution is consistent with the team-building goals of the organization, the distribution is shown in Figure-3. The topmost ‘team-ID’ count is at level 2 with ‘216’, further goes with ‘212’ for team 3, and for team 4 it goes ‘201’, at level 5 with a ‘181’ count and to last with the lowest ‘51’ of team-level 8 but still has the reached highest progression. In this visualization, the data set used is a ‘level-event’ file that contains team-level, team ID, event Id, and event type of start and end.

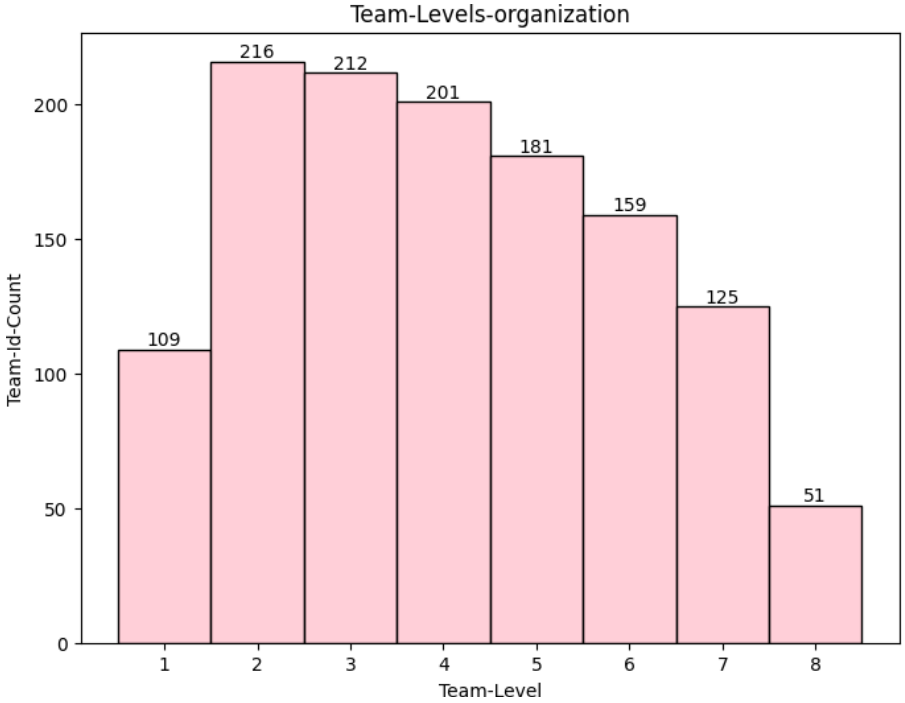


Figure 3: Team-Level-Organization.

3.3.2 Excessive Products Sold

The chart describes the spending for buy-id 0 to 5 in Figure-4, in which buy-id 5 is the highest followed by buy-id 4, 3, 2, and 0. To look at the chart buy-id ‘1’ is at the lowest in spending with ‘538’, and the buy-id with the most spending on the product is buy-id ‘5’. The x-coordinates show the ‘BUY-ID’ and the y-coordinates show the ‘Spendings’. The analysis is performed on the buy-clicks data file.

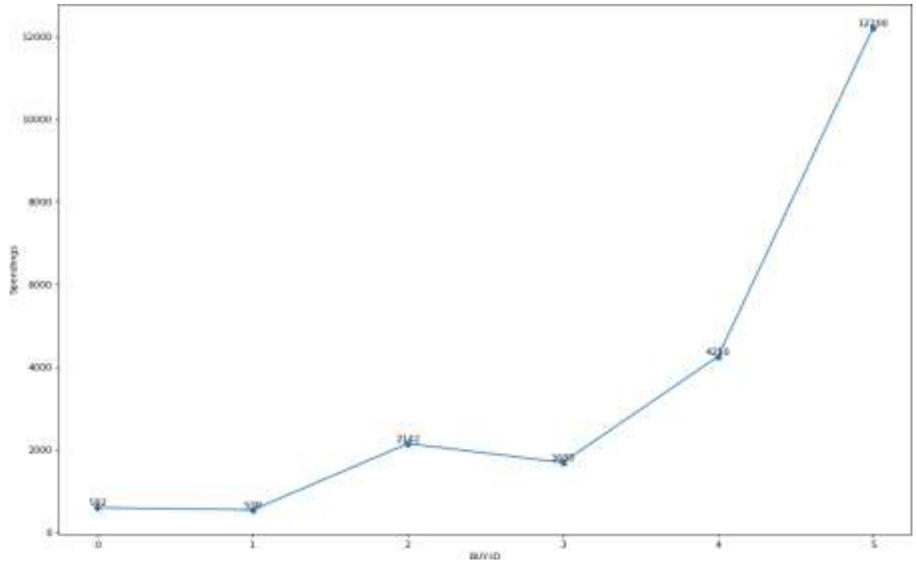


Figure 4: Returns generated by products.

3.3.3 Strike proportion on Various Sources.

Series	Source-Type	Strike’s Count	Overall Counts	Percent
1	iPhone	348976	3066988	0.1137
2	android	299576	2747372	0.1090
3	windows	113098	1042214	0.1085
4	linux	53910	498808	0.1080
5	mac	30582	276852	0.1104

Table 2: Source’s Stats on Strike n Count

Table-2 shows the values or stats of the strike count performed by the user. The source with the most clicks is ‘iPhone’, with a ‘348976’ strike count, The source with the second most strike is ‘Android’, with ‘299576’ clicks, followed by Windows, ‘Linux’ and last ‘MAC’ with a ‘30582’ strike count. To analyze this chart in the given Figure-5the ‘Game-click’ and ‘User-session’ are used to perform the visualization.

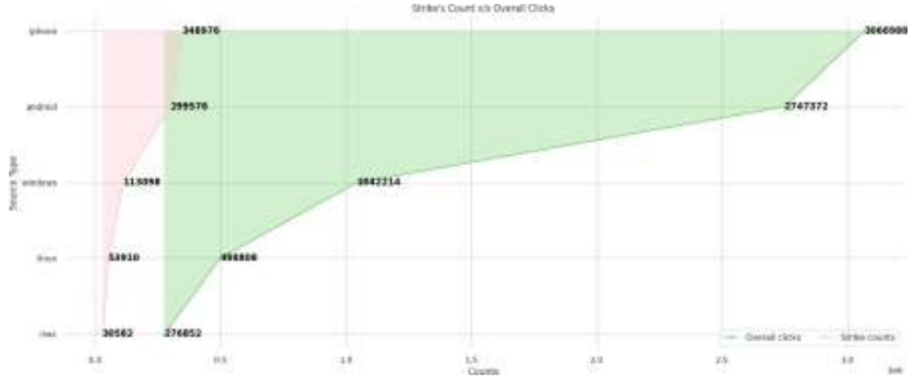


Figure 5: Source’s Strike Count n Click

3.3.4 Worldwide Source’s Console :-

In the given Table-3 stats of the Source used by the users for performing the game.

S.No	Source-Type	Average Rank	Source Count	Percentage
1	iPhone	4.3495	3974	41.88
2	Android	4.3885	3274	35.39
3	Windows	4.3645	1240	13.41
4	Linux	4.1825	504	5.45
5	mac	4.3799	358	3.87

Table 3: Source’s Stats

The analysis of the Figure-6 is carried out on the ‘user-session data set file the visualize the chart, in which the Lowest used source is ‘MAC’ with only ‘3.87’ percent of the user, and with the topmost user is on the ‘iPhone’ with ‘41.88’ percent. Other Sources like ‘Android’, ‘Windows’, and ‘Linux’ with 35.39 percent, 13.14 percent, and 5.45 percent is used. This shows the most used Source played by the users.

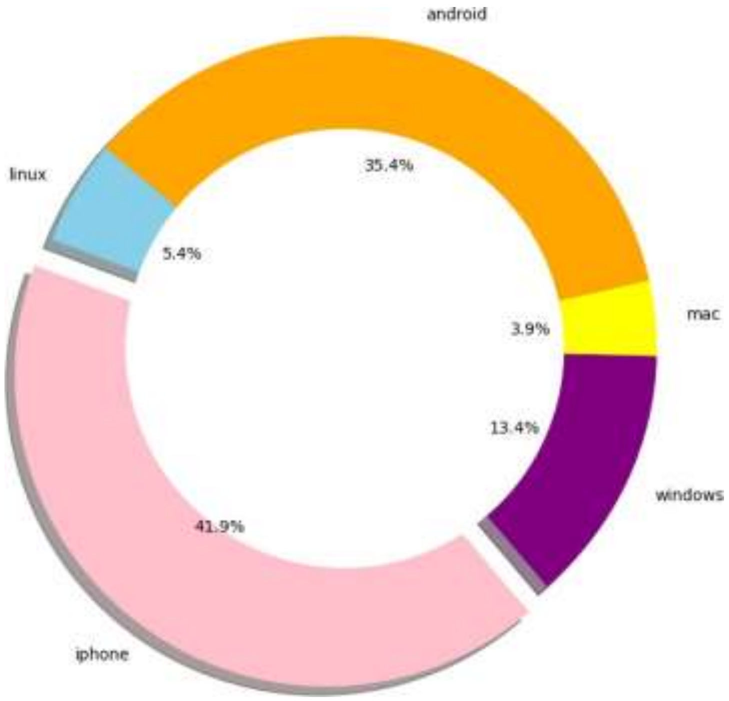


Figure 6: Source’s Strike

3.3.5 Users Age-Range on Week-Days :-

The Age-range with the most users is ‘25-32’ with users-count ‘608’, on specifically on Tuesday with ‘882’ users active. The second most age range from ‘33-40’ with ‘555’ users-count. The lowest user is the age of ‘65+’ with ‘186’ users-counts. The lowest days people performed on Mondays with ‘372’ active users in the given chart Figure-7. The file ‘users’ data set is implemented to perform the analysis of the user’s count on the weekdays.

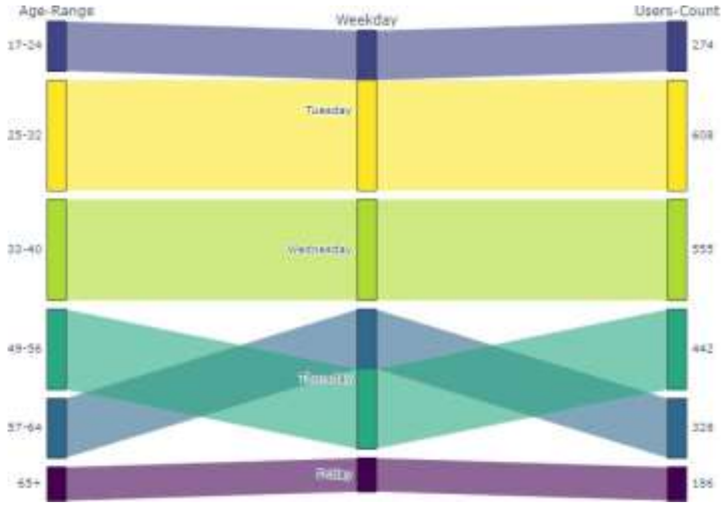


Figure 7: Users Age-Range on Week-Days

In Figure-8, looking into the time period of day, evening, and night, the most active users are in the ‘day’ time with ‘1174’ users which is ‘49’ percent of the users active time, in the evening time ‘806’ users are active with ‘34’ percent of them, and in night time only ‘413’ active user’s are present with ‘17’ percent of them. This shows most people are active in the day time because it gets time duration better to perform it.

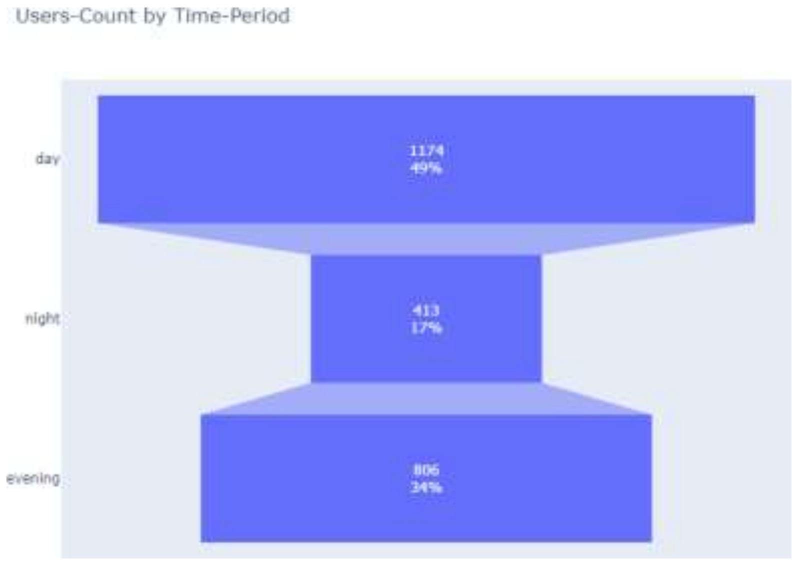


Figure 8: Users Time-Period

3.3.6 BOX

The data set implemented to visualize the plot is a 'buy-click' data file forFigure-9. The Plot shows the distribution of values for 'buy-Id' and 'prices'. The median price is '3.0', and for the 'buy-Id' the median price is '2.0' by the users, The values are displayed to observe the plot. The Price box plot iterates the median lines of the box plot with the value, and for the box plot iteratesthe median value to rectify the users with most buy-Id average and prices.

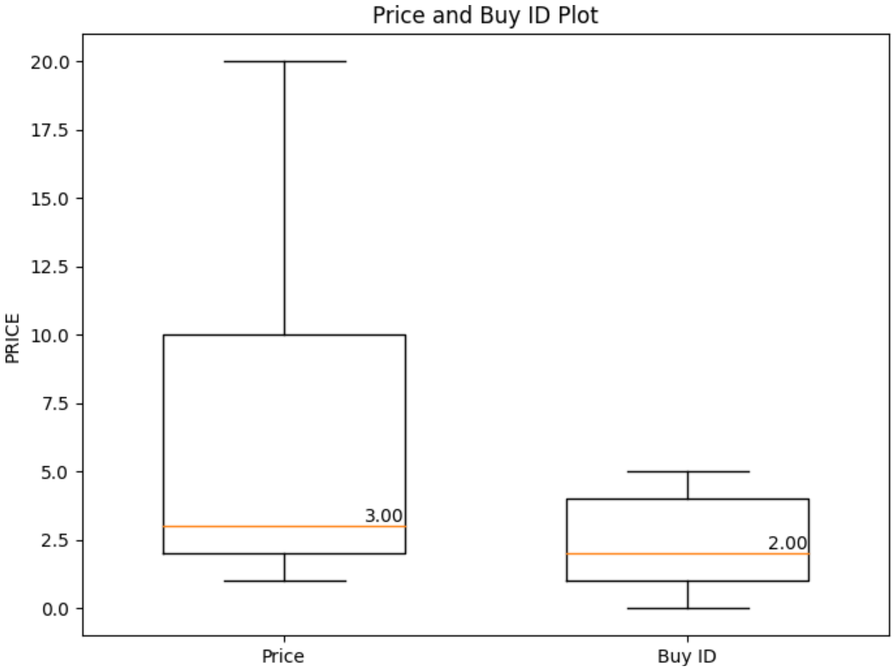


Figure 9: Price and Buy-Id of the Users

3.3.7 Game-clicks by the users is pro or noob

The median game-click count in this Figure-10 the 'NOOB' player is '65' of users and '146' for 'PRO' users players in the data set. The 'combined-data' data set file is implemented on this to analyze the user's player type if the user is 'NOOB' like a beginner to the game, and 'PRO' means the expert of this game. The average number of game clicks for each player type is the center of the distribution, to the discrepancy between the 'PRO' and 'NOOB' players in the game clicks count. It implies that PRO users are having a greater frequency of the game-clicks than NOOB users, which indicates a higher degree of participation in the game.

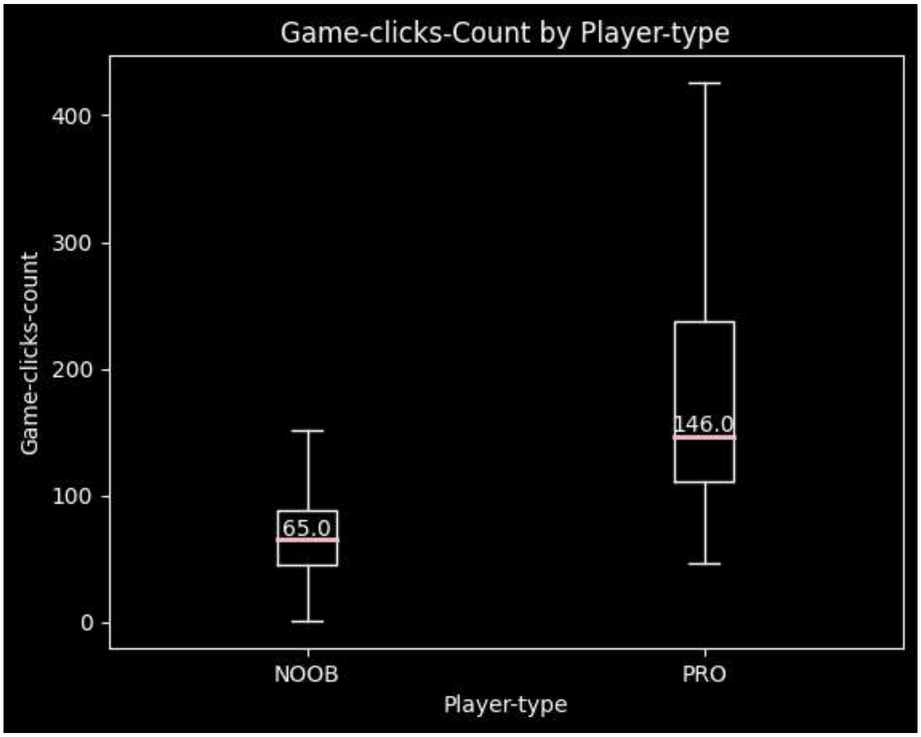


Figure 10: Player-Type (Users) 'PRO' and 'NOOB'

3.3.8 Assignment count per Team

The number of tasks finished by each team with their assignment. The 50 teams performed the assignment in which the most task were completed by team '94' and the least tasks were completed by team '139', the assignment count varies in the y-coordinates, and 'team' are in the x-coordinates.

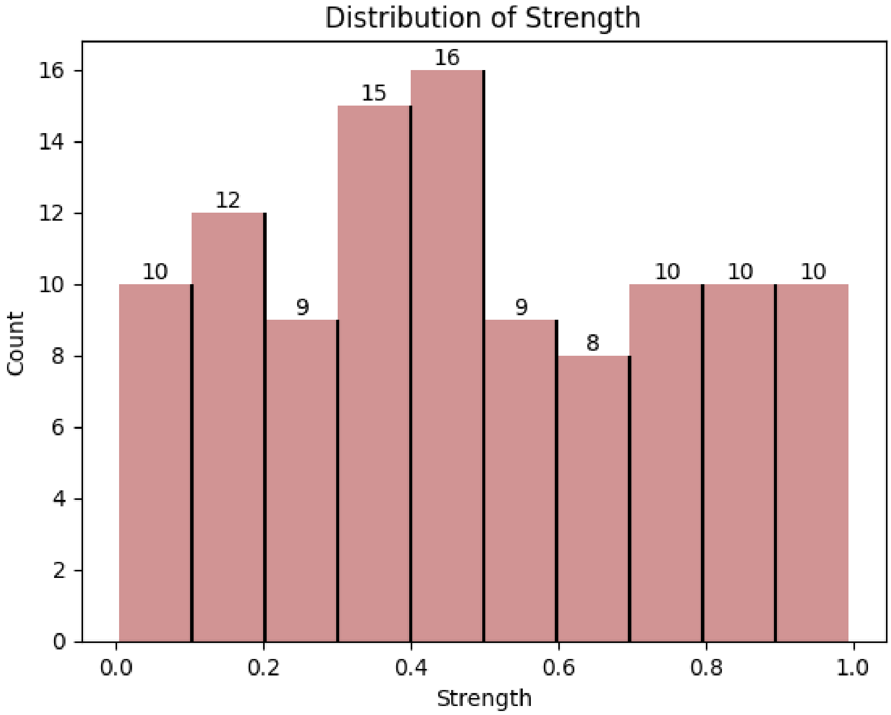


Figure 11: Teams and their Assignment Count.

The distribution of the variable ‘strength’ is present in Figure-11 indicating various ranges of strength levels. The strength values are shown along the x-coordinates, while the number of occurrences in each bar of the plot. The highest count is from ‘0.4-0.5’ with 16 task counts, and the lowest is at 8 between ‘0.6-07’ strength. The strength of 0.1, and from 0.7 to 1.0 is the count of 10. The data set implemented in this analysis is the ‘team and team-assignments data set file. The box plot Figure-12 shows a median range of line of strength with the value of ‘0.46’ for all the strength among the teams. In the following Table-4 it shows the total count of teamID, strength, and current level of the users with mean, minimum, and maximum values.

Concise	teamId	Strength	Current-Level
Count	109	109	109
Mean	56.889	0.483	1.0
Minimum	0	0.004	1
Maximum	171	0.994	1

Table 4: Synopsis of the Team Stats.

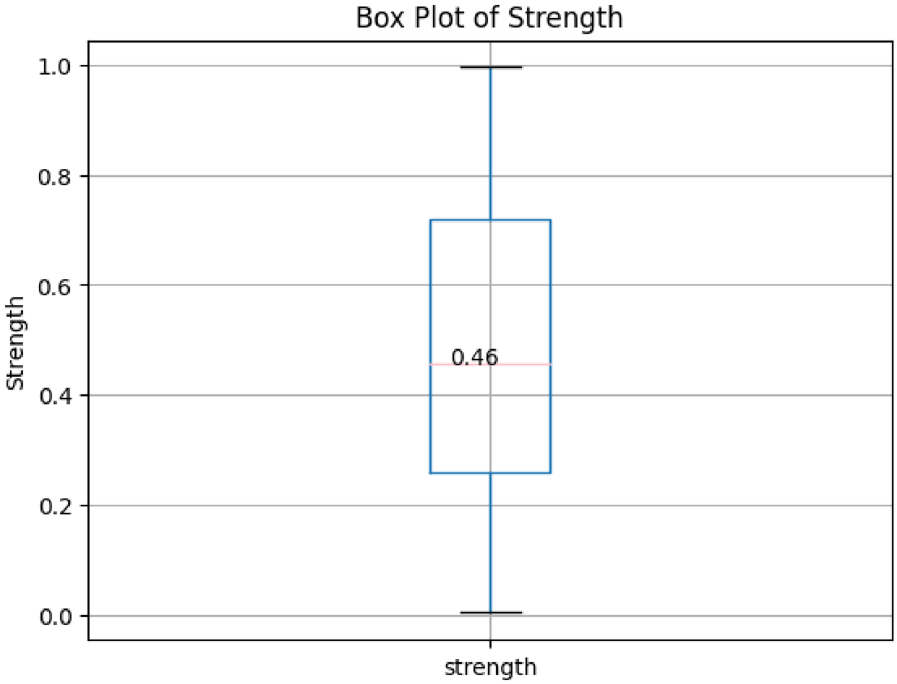


Figure 12: Teams and their Strength median range.

4 Machine Learning Models :-

The Algorithm used on the ‘combined-data’ file for the analysis with the help of Machine Learning. In the data set the values for avg-price and count-buy values NULL are replaced with 0 values. Two algorithms are used to perform the analysis which are Classification Analysis and Clustering Analysis. The column added to the data set is the player type which contains ‘PRO’ and ‘NOOB’ which represent the player is an expert or beginner to the game. The analysis is implemented using tools such as Apache Spark (Carbonell et al., 1983).

4.1 Classification Analysis:

Classification analysis has the potential tool for resolving a wide range of issues. In any case, it is crucial to remember that no calculation is perfect. Some data points will always be incorrectly classified (Guinand et al., 2002). Consequently, before applying the algorithm to production data, it is essential to evaluate its performance on a data set. It’s a tool to figure out the various problems to solve it. Algorithms to solve it are ‘Logistics Regression’, ‘Naves Bayes’, and ‘Super vector machines (SVM)’.

Logistic Regression :- Evaluating the effectiveness of the logistic regression in the classification analysis. Logistic regression calculation achieved an accuracy of ‘0.9004’, which shows that it accurately grouped the ‘90.04’ percent of the occurrence. The ‘NOOB’ and ‘PRO’ classes had high precision, recall, and F1 scores, indicating that correctly identifying instances of each class and minimizing wrong positives and negatives were achieved (Karthi et al., 2015).

Naves Bayes :- The Nave Bayes calculation performed well, with an accuracy of ‘0.9026’. For the NOOB and PRO classes, it demonstrated balanced performance with high precision, recall, and F1 scores. This shows that the nave bayes calculation complained about accurately ordering occurrences between the two classes.

The data set used to analyze this is a ‘combined-data’ file. In terms of accuracy and balanced classification yardstick, logistics regression performed the best. With high precision, recall, and F1 scores, this is able to correctly classify the instances. To perform the accuracy value of true positive and true negative sum up and divides with the true positive and negative(TRN) sum false positive and negative. To calculate precision is to a ratio of true positive sum up true positive(TRP) and false positive(FAP), next for recall true positive sum up with false negative(FAN), and for F1 score 2 times precision and recall and sum up the precision and recall in the given equation.

$$Accuracy = (TRP + TRN)/(TRP + TRN + FAP + FAN) \quad (1)$$

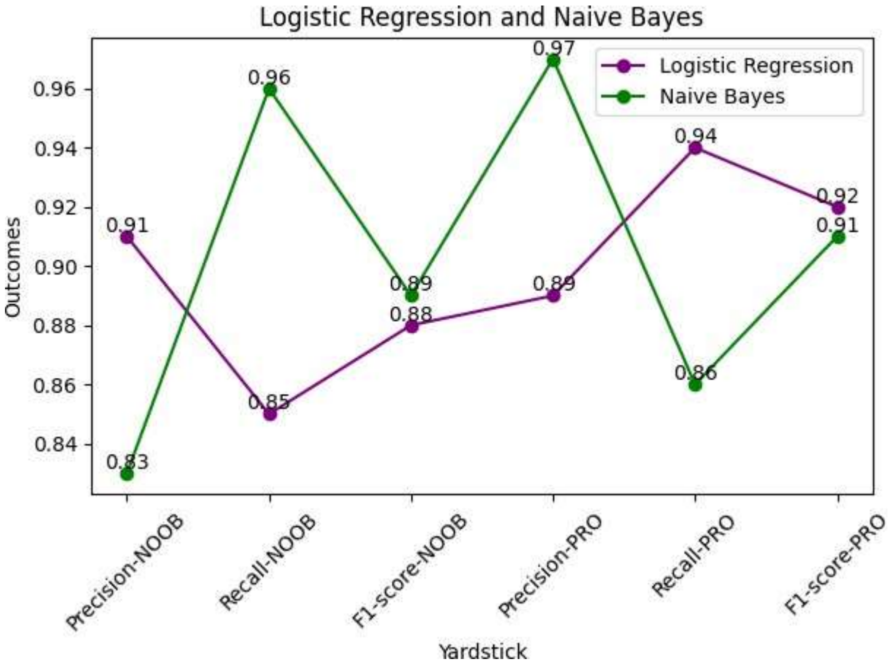


Figure 13: Logistic Regression and Naves Bayes.

The plot in Figure-13 compares the performance of the two classifiers using the various evaluation metrics, like accuracy, review, and F1 scores. The yardstick is represented by the x-coordinates, while the outcome is represented by the y-coordinates. Precision estimates the accuracy of the positive expectation. The PRO class, logistic regression outperformed Naves Bayes '0.97' in terms of precision of '0.89', and the NOOB class Naive Bayes had an accuracy of '0.83' compared with the calculated relapse of '0.91'.

Recall exemplifies the capacity to correctly identify positive instances. Naves Bayes shows a higher review for the NOOB class '0.96' compared with Logistic Regression '0.85'. Logistic regression had a higher recall of '0.94' than the Naves Bayes for the PRO class. F1 scores are the consonant mean of accuracy and review and give a fair measure. With F1 scores of '0.88' for NOOB and '0.92' for PRO, logistic regression outperformed both classes. Naves Bayes had marginally lower F1 scores with '0.89' for NOOB and '0.91' for PRO. In order to contrast, Logistic regression and Naves Bayes classify the PRO and NOOB classes, and the performance of the values is included in the figure.

4.2 Clustering Analysis:

Clustering analysis numerous issues can be effectively addressed. Unsupervised learning that uses similarity to group data points together is known as clusteringanalysis (Madhulatha, 2012). This can be used to find patterns in the data that would be hard tosee. Some algorithms that can be performed are ‘K-means’, and Density-based.

K-Means Clustering Analysis:- The characteristics of each cluster are based on the provided data (Xu & Lange, 2019). 7 distinct clusters were created by applying theK-means clustering algorithm to the data set. The calculation of the data setutilizes three highlights which are count-games-clicks, avg-price, and count-hits. (Von Luxburg, 2010)

The Cluster process that standardized the data in the following Table-5.

Cluster	Centroid	Size
0	[50.32, 5.63, 1.69]	1780
1	[382.32, 41.24, 1.92]	1388
2	[106.58, 11.93, 2.48]	733
3	[800.13, 82.78, 2.13]	356
4	[555.97, 60.29, 1.50]	207
5	[172.64, 18.96, 2.89]	109
6	[274.77, 29.87, 1.75]	46

Table 5: Clustering

Cluster 0 group has a centroid with the values [50.33,5.63,1.69]. It has the most data points, with 1780, and is the largest. Cluster 1 centroid is located at [382.32,41.24,1.93]. There are 1388 data points in it. Cluster 2 has a bunch of centroids of [106.59, 11.94, 2.48], with 733 data points made up.

Cluster 3 stands out because it has the highest values for the count-game clicks and count-hits, and its centroid is [800.13,82.78,2.13]. It is containing 356 data of interest. Cluster 4 is centroid [555.97, 60.29,1.50]. It has a high value for count-game-clicks and count-hits across its 207 data points.

Cluster 5 has a centroid of [172.65,18.96,2.90]. It has moderate centroid values for the three features and 109 data points. Cluster 6 centroid of [274.78, 29.88,1.76], this cluster has a moderate quality for the count-game-clicks and count-hits. It has 46 data points, making it the smallest cluster in Figure-14. The average values of count-game-clicks, count-hits, and avg-price for each cluster are represented by these centroid values. They shed light on the behavior and characteristics of the data points in each cluster.

For instance, the relatively low values of the data values in cluster 0 points to a group of users with lower engagement and lower average spending. In contrast, Cluster 3 has higher values for all three features, indicating a user group with high spending and collaboration.

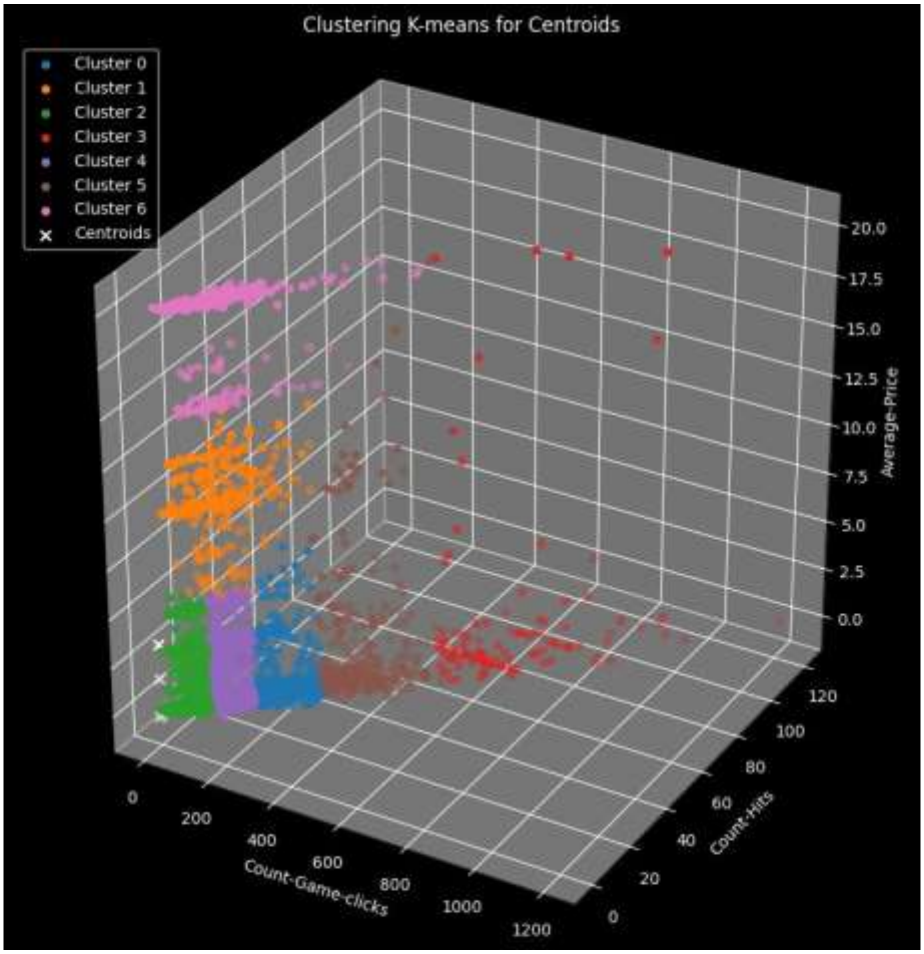


Figure 14: Logistic Regression and Naves Bayes.

K-means Clustering Analysis on Radar Chart :- The bunching values acquired from the K-means grouping calculation into the gathering of the information. A specific clustering value is assigned to each data point, indicating the group to which belongs. The three data that apply to k-means clustering from the data set are ‘count-game-clicks’, ‘count-hits’, and ‘Avg-price’. The calculation dived the information into a number of bunches which are 7 clusters. the clustering system includes information on the closet number of squares inside the bunch. The clustering e values, provide details about the assigned clusters. The clustering values can be determined which data points are grouped and to which they belong. For instance, data of interest with a grouping 1 shows that it has a clustering value of 1. In a similar way, a data point that has a clustering value of 0 is a member of Cluster 0. By looking at the appropriate grouping values, acquire experiences in the organization and attributes of each bunch in Figure-15.

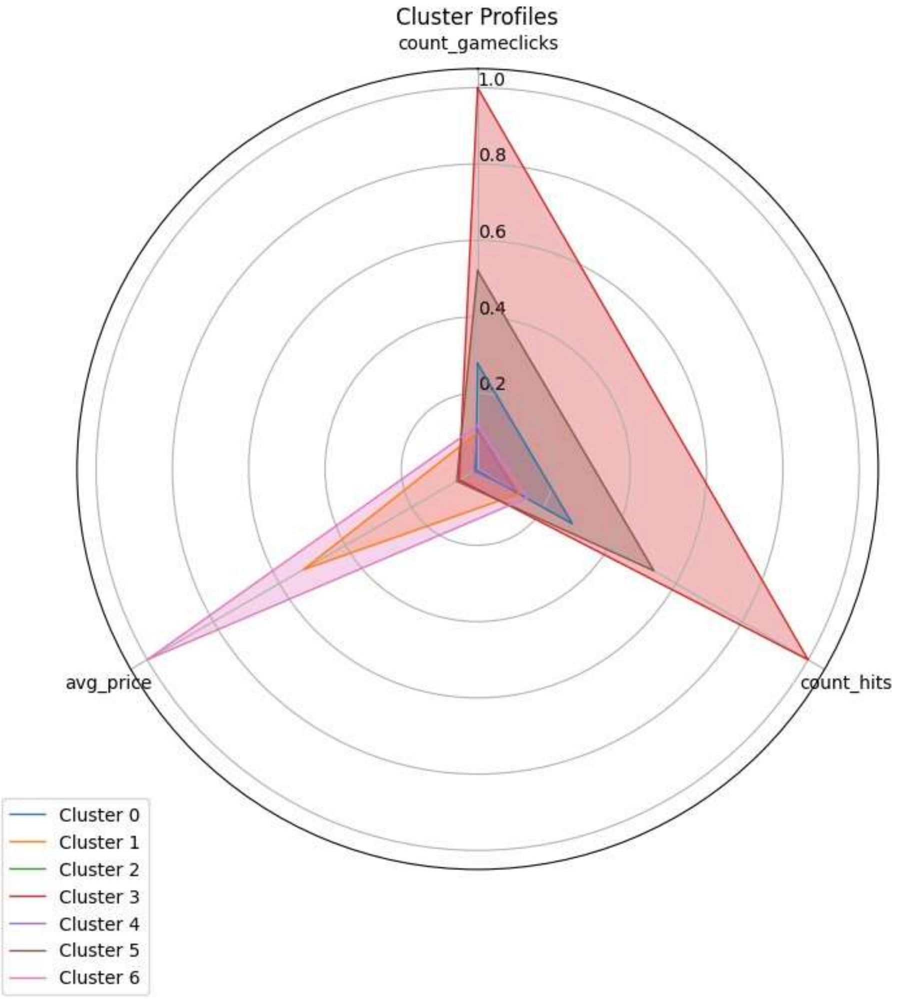


Figure 15: Logistic Regression and Naves Bayes.

5 Graph Analysis :-

A group of analytical techniques that describe and examine data using graph topologies is to be recommended as graph analytics. Using the algorithms, it rectifies relationships between items in a graph database, like between various individuals, and transactions (Sun et al., 2019).

DATA SET FILE (team-chat)	DATA'S NARRATION
chat-leave	From the user to the TeamChatSession, a border with the label "Leaves" is created. The User id, TeamChatSession id, and Leaves edge timestamp are the columns.
chat-join	From the user to the TeamChatSession, it creates an edge with the label "Joins." The sections are the Client id, TeamChatSession id and the timestamp of the Joins edge.
chat-respond	A line is added to this record when player with chatid2 answers a visit post by one more player with chatid1
chat-mention	Makes an edge marked "Referenced". Section 0 is the id of the ChatItem, segment 1 is the id of the Client, and section 2 is the timestamp of the edge going from the chatItem to the Client.

Table 6: Data set Stats for the Graph Analysis

(Needham & Hodler, n.d.)Graph analytics could be used to analyze the chatting activities of active users by capturing users, teams, team join sessions, team chat sessions, team respond sessions, and team mention sessions as nodes on the graph interaction between these nodes as edges. The data set used for the graph analysis to accomplish on the data file in the following Table-6 and bound in the toolwhich is 'Neo4j' to acquire proper graph figures.

5.0.1 Team-chat Entities

User-Interaction :- In this, the file chat-respond and chat-mentioned are used to visualize the graph analysis on the mention, chat time and respond in the following Figure-16.

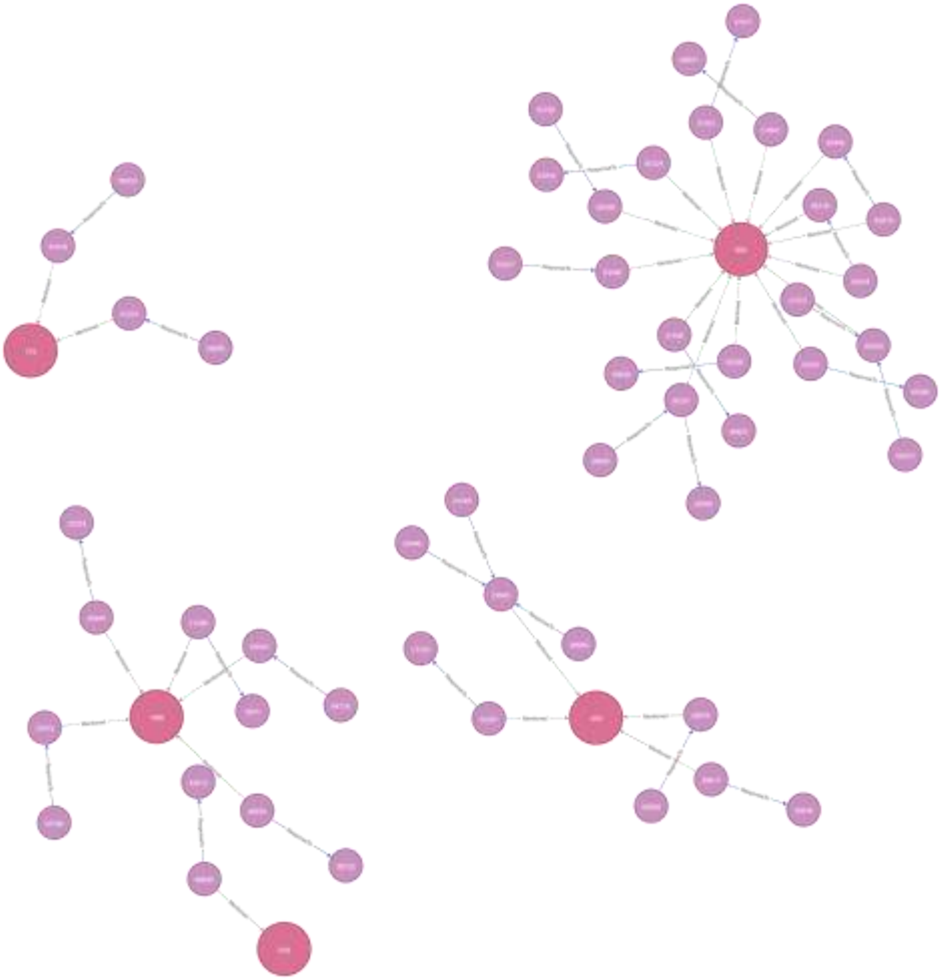


Figure 16: Interaction of Users for mention and respond chat.

RespondTo Entities :- The relation of the user chat with the response performed in the chat-respond data set file. The visualization shows the connection of the graph response to the user, in Figure-17 shows the analysis of the data performed.

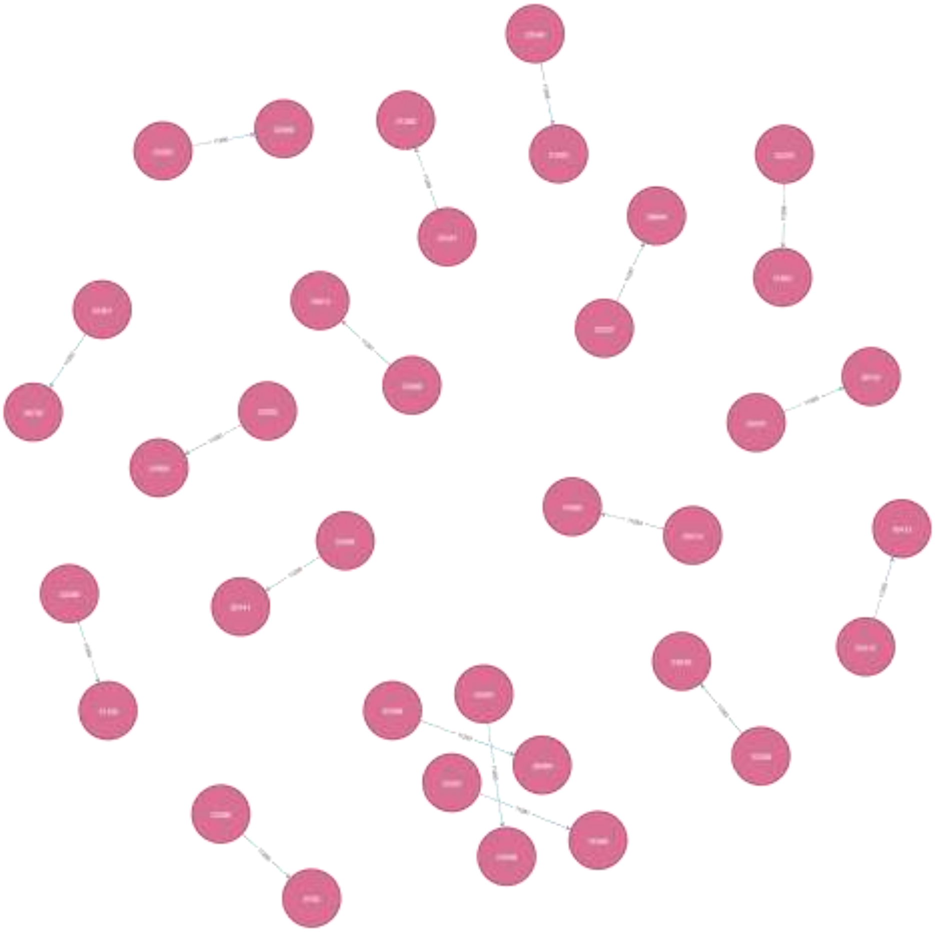


Figure 17: chats of user RespondTo in the graph.

Extensive Converse Series(chain) :- The longest converse chain to be found out in the analysis, with 10 nodes, and 9 relationships in Figure-18. It is a strategy to focus on the game flow rather than the chat box, with the highest messages passed from the nodes.

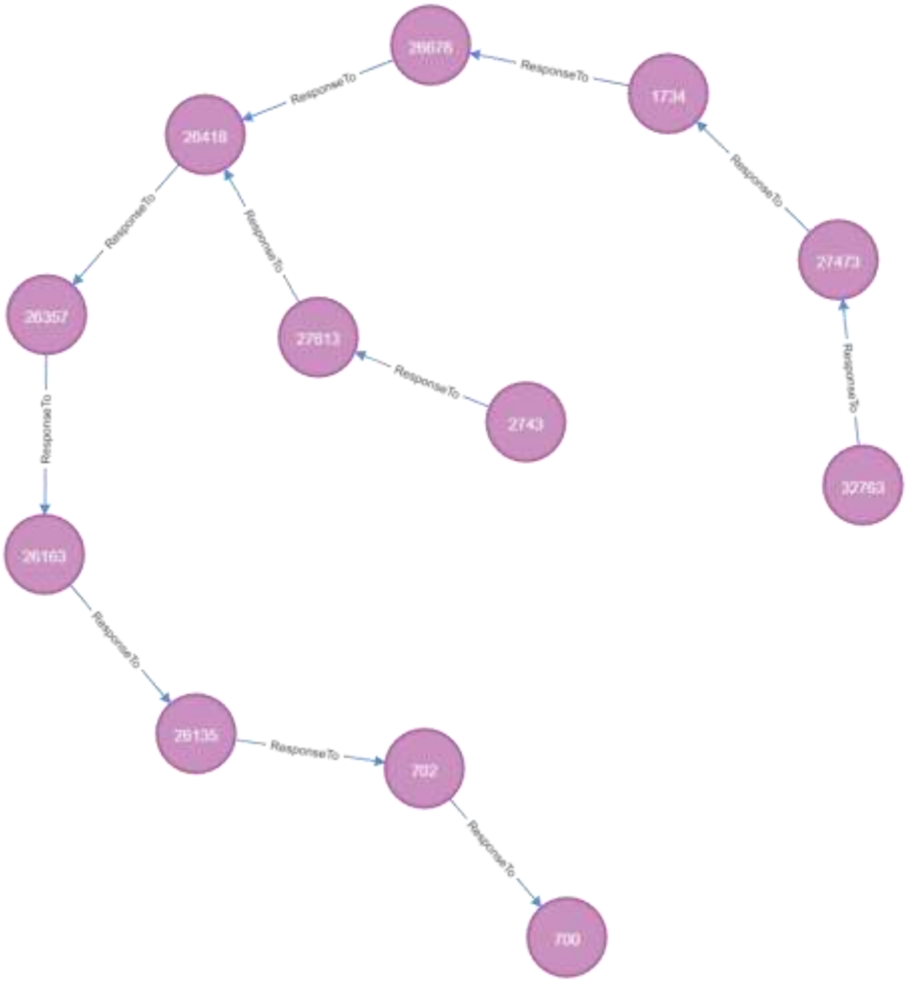


Figure 18: Extensive Conversation Flow.

Users Joins or Leaves :- The data set of chat-join and chat-leaves are performed to analyze which users are leaving the chat or which users are joining the chat-session in the given Figure-19 with nodes of chatSession and edges represent join, leave and mentioned.

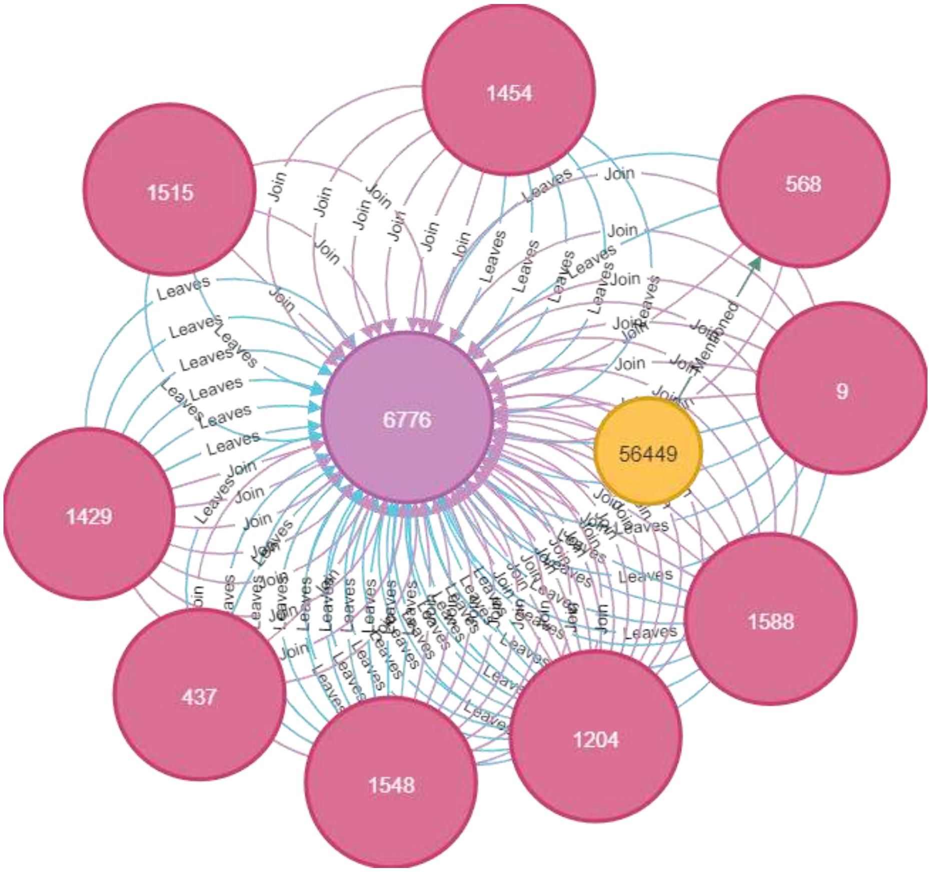


Figure 19: Users Joins, Leave and mention the chat-session.

User and chat-session for join and leave :- There are nodes with two different types which are teamSession and userId with an edge of join and leave. The data set file chat-join and chat-leave are performed in this analysis in Figure-20 shown below.

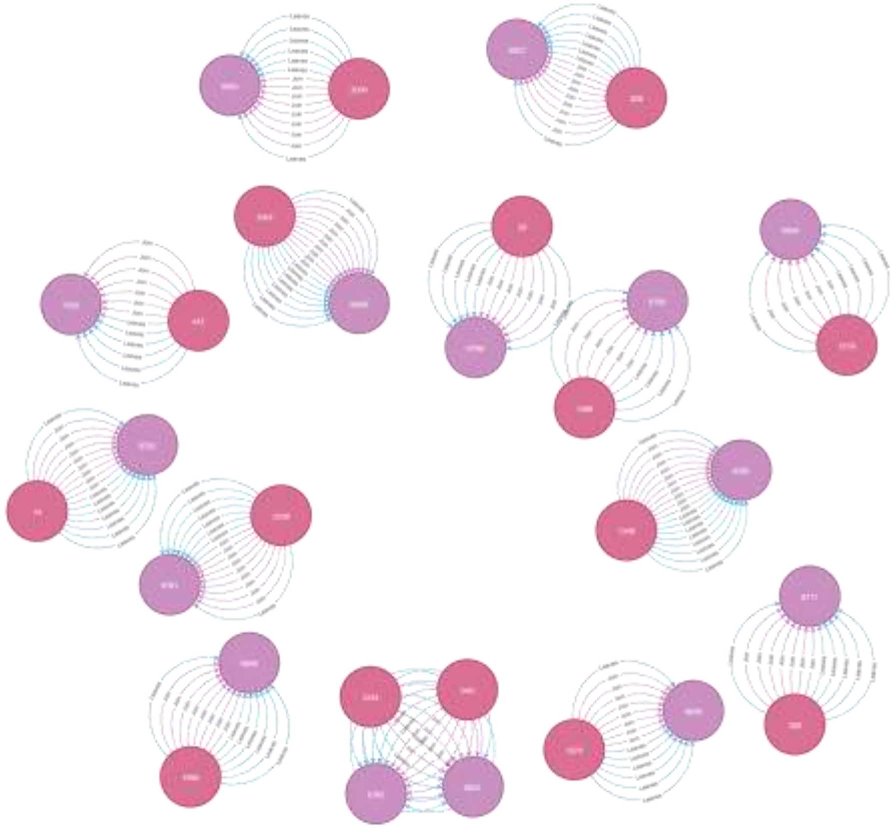


Figure 20: Users Join and Leave connection edge to the chat-session.

Mention-chat of team :- In the following Figure-21 a line gets adds to the chat mentioned. The edge of the timestamp is also mentioned with the chat item of the user. The chat-mention data set file is used to accomplish the analysis of the chat-mention data with the edges of mention to the chat session.

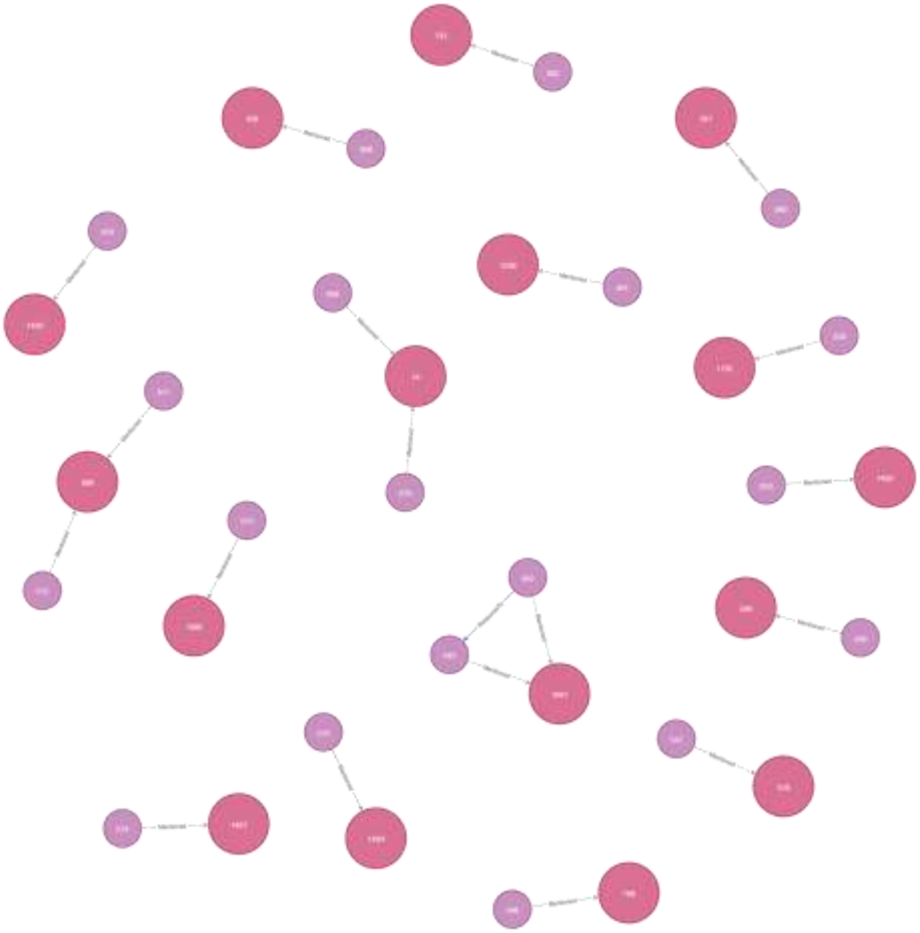


Figure 21: Graph of the chat-mention by the user.

6 Big Data Ethics :-

Big Data ethics is defined as crucial to any discussion of ethics and Big Data. As a rule, includes the test of leads that make advantage or cause mischief to others. The study of the ethical implication of big data collection uses, and analysis is called out to be the big data ethics. Moreover, there are a number of ethical issues that arise when big data is used, such as bias, discrimination, and privacy. In addition to better integrating data ethics into everyday life, case studies in data ethics serve as a foundation for ethical discussion. The vast amounts of data are generated by machines, devices, and people which is referred to as big data. Insights, predictions into human behavior, and improved decision-making can all be achieved with the help of this data (Herschel & Miori, 2017).

iHealth Ethics has a lot of potential to create a treatment that is more individualized and effective, the ethical analysis has described that there are various ethical issues (Anom, 2020). These are mostly the irresoluteness of independence and inquiry into how innovation might fit the necessities and assets of individual patients, the dangers to individual and information security, and the gamble of data classification and inclination (Rubeis, 2022). There are multiple levels at which ethical issues must be addressed. In the particular context of mental health, regula- tions must be implemented on a policy level to guarantee quality standards and control over one's own health data. The treatment, prediction of mental illness, and prevention of mental illness on the other side (Knoppers & Thorogood, 2017).

Big Data ethical analyze to follow individuals' exercise, development, and inclinations with the huge amount of data can be utilized that data can be used to create profiles of people, to target them with an advertisement like offering them a car loan (Mark, 2019). To discriminate against one person or a crowd of people can be used, big data can be used to decide which are eligible for a car loan to target them for particular demographics. The ethics of big data can be biased, which can be reflected in an outcome that is inaccurate and inequitable. Inthe event of an enormous amount of information is brought into play to foreseeprobably going to perpetrate wrongdoing, the assumption might be one-sided against a specific category.

7 Conclusion: -

After analyzing the report Big data refers to an enormous quantity of information obtained from several social media sensors, and transactions, with the help of data, get more information about consumer behavior and make better outputs. The data set provided to run the EDA visualization to examine and understand the data exploration, data preparation, and categorization of the data analysis. Two algorithms performed in the research are Classification and Clustering analysis of Machine Learning (ML). In the active session of the game, the graph analysis was implemented to the game to see the users joining, and responding to the chat session. These processes are about to get understanding the player’s visualization of the game.

7.1 Finding and Recommendation

- Utilize statistical methods to determine the most significant variables that influence player success and bird sightings. To discover relationships between variables, conduct correlation analysis.
- Optimize player strategies and actions with reinforcement learning algorithms. Reward players based on successful bird sightings or actions that led to the discovery of the pink flamingo, modeling the game environment.
- Utilize graph algorithms like shortest path or breadth-first search to examine the player's paths throughout the game. Determine the best patterns or routes that result in successful bird sightings.
- Use diagram based cooperative separating procedures to prescribe fruitful techniques or activities to players in light of the ways of behaving of comparable players or effective sightings previously.

8. References:

Alexandre da Silva, V., Julio C.S. dos, A., Edison Pignaton, de F., Thomas J., L., & Claudio F., G. (2016). Strategies for Big Data Analytics through Lambda Architectures in Volatile Environments. *IFAC-PapersOnLine*, *49*(30), 114–119. https://doi.org/10.1016/J.IFACOL.2016.11.138

Ameri, A., Kamavuako, E. N., Scheme, E. J., Englehart, K. B., & Parker, P. A. (2014). Real-time, simultaneous myoelectric control using visual target-based training paradigm. *Biomedical Signal Processing and Control*, *13*(1), 8–14. https://doi.org/10.1016/J.BSPC.2014.03.006

Anom, B. Y. (2020). Ethics of Big Data and artificial intelligence in medicine. *Ethics, Medicine and Public Health*, *15*, 100568. https://doi.org/10.1016/J.JEMEP.2020.100568

Beygelzimer, Y., Estrin, Y., & Kulagin, R. (2015). Synthesis of Hybrid Materials by Severe Plastic Deformation: A New Paradigm of SPD Processing. *Advanced Engineering Materials*, *17*(12), 1853–1861. https://doi.org/10.1002/ADEM.201500083

Birke, R., Bjoerkqvist, M., Chen, L. Y., Björkqvist, M., Smirni, E., & Engbersen, T. (n.d.). *This paper is included in the Proceedings of the 12th USENIX Conference on File and Storage Technologies (FAST ’14). Open access to the Proceedings of the 12th USENIX Conference on File and Storage Technologies (FAST ’14) is sponsored by (Big)Data in a Virtualized World: Volume, Velocity, and Variety in Cloud Datacenters (Big)Data in a Virtualized World: Volume, Velocity, and Variety in Cloud Datacenters*. 177. Retrieved May 18, 2023, from https://www.usenix.org/conference/fast14/technical-sessions/presentation/birke

Carbonell, J. G., Michalski, R. S., & Mitchell, T. M. (1983). AN OVERVIEW OF MACHINE LEARNING. *Machine Learning*, 3–23. https://doi.org/10.1016/B978-0-08-051054-5.50005-4

Casado, R., Practice, M. Y.-C. and C., & 2015, undefined. (2014). Emerging trends and technologies in big data processing. *Wiley Online Library*, *27*(8), 2078–2091. https://doi.org/10.1002/cpe.3398

Casado, R., & Younas, M. (2015). Emerging trends and technologies in big data processing. *Concurrency and Computation: Practice and Experience*, *27*(8), 2078–2091. https://doi.org/10.1002/CPE.3398

Chen, H. M. (2017). Chapter 1. An Overview of Information Visualization. *Library Technology Reports*, *53*(3), 5–7. https://www.journals.ala.org/index.php/ltr/article/view/6288/8214

Deshpande, K., & Rao, M. (2022). An Open-Source Framework Unifying Stream and Batch Processing. *Lecture Notes in Networks and Systems*, *336*, 607–630. https://doi.org/10.1007/978-981-16-6723-7_45/COVER

Gheisari, M., Shayegan, M. J., Ahvanooey, M. T., Liu, Y., Addobea, A. A., Li, Q., Amankona, I. O., Hou, J., & Cn, A. A. A.). (2022). A Batch Processing Technique for Wearable Health Crowd-Sensing in the Internet of Things. *Cryptography 2022, Vol. 6, Page 33*, *6*(3), 33. https://doi.org/10.3390/CRYPTOGRAPHY6030033

Guinand, B., Topchy, A., Page, K. S., Burnham-Curtis, M. K., Punch, W. F., & Scribner, K. T. (2002). Comparisons of Likelihood and Machine Learning Methods of Individual Classification. *Journal of Heredity*, *93*(4), 260–269. https://doi.org/10.1093/JHERED/93.4.260

Herschel, R., & Miori, V. M. (2017). Ethics & Big Data. *Technology in Society*, *49*, 31–36. https://doi.org/10.1016/J.TECHSOC.2017.03.003

Higgins, P. G. (1999). *Hybrid Intelligent Human-Computer Paradigm*.

Hitzler, P., & Janowicz, K. (n.d.). *Semantic Web 0 (0) 1 1 IOS Press Linked Data, Big Data, and the 4th Paradigm Editorial*. Retrieved May 18, 2023, from http://www.w3.org/TR/rdf-primer/.

Jain, P., Gyanchandani, M., & Khare, N. (2016). Big data privacy: a technological perspective and review. *Journal of Big Data*, *3*(1), 1–25. https://doi.org/10.1186/S40537-016-0059-Y/TABLES/5

J@zy, K. (n.d.). *The Real-Time Producer/Consumer construction Paradigm: A paradigm for the of efllcient, predictable real-time systems"*.

Karthi, M., Anu, M., Gladence, L. M., & Anu, V. M. (2015). *A statistical comparison of logistic regression and different bayes classification methods for machine learning*. *10*(14). www.arpnjournals.com

Knoppers, B. M., & Thorogood, A. M. (2017). Ethics and Big Data in health. *Current Opinion in Systems Biology*, *4*, 53–57. https://doi.org/10.1016/J.COISB.2017.07.001

Madhulatha, T. S. (2012). An Overview on Clustering Methods. *IOSR Journal of Engineering*, *02*(04), 719–725. https://doi.org/10.9790/3021-0204719725

Mark, R. (2019). Ethics of Using AI and Big Data in Agriculture: The Case of a Large Agriculture Multinational. *The ORBIT Journal*, *2*(2), 1–27. https://doi.org/10.29297/ORBIT.V2I2.109

Milo, T., & Somech, A. (2020). Automating Exploratory Data Analysis via Machine Learning: An Overview. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2617–2622. https://doi.org/10.1145/3318464.3383126

Needham, M., & Hodler, A. E. (n.d.). *The #1 Platform for Connected Data A Comprehensive Guide to Graph Algorithms in Neo4j*.

Pascute, L. C., & Engineering., Y. S. University. R. S. of. (2002). *A VHDL-based digital slot machine implementation using a complex programmable logic device /*. https://digital.maag.ysu.edu:8443/xmlui/handle/1989/6215

Rubeis, G. (2022). iHealth: The ethics of artificial intelligence and big data in mental healthcare. *Internet Interventions*, *28*, 100518. https://doi.org/10.1016/J.INVENT.2022.100518

Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. *Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013*, 42–47. https://doi.org/10.1109/CTS.2013.6567202

Shahrivari, S. (2014). Beyond Batch Processing: Towards Real-Time and Streaming Big Data. *Computers 2014, Vol. 3, Pages 117-129*, *3*(4), 117–129. https://doi.org/10.3390/COMPUTERS3040117

Sun, G., Li, F., & Jiang, W. (2019). Brief Talk About Big Data Graph Analysis and Visualization. *JBD*, *1*(1), 25–38. https://doi.org/10.32604/jbd.2019.05800

Von Luxburg, U. (2010). Clustering Stability: An Overview. *Foundations and Trends® in Machine Learning*, *2*(3), 235–274. https://doi.org/10.1561/2200000008

Voyvodic, J. T. (1999). Real-Time fMRI Paradigm Control, Physiology, and Behavior Combined with Near Real-Time Statistical Analysis. *NeuroImage*, *10*(2), 91–106. https://doi.org/10.1006/NIMG.1999.0457

Xu, J., & Lange, K. (2019). *Power k-Means Clustering* (pp. 6921–6931). PMLR. https://proceedings.mlr.press/v97/xu19a.html

Yun, X., & Epstein, S. L. (2012). A hybrid paradigm for adaptive parallel search. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7514 LNCS*, 720–734. https://doi.org/10.1007/978-3-642-33558-7_52/COVER