Republic of Yemen

Ministry of High Education & scientific research

Al – Rayan University

Faculty of Higher Studies

Comparative Study Between K-Means and Genetic algorithm As a Technique for Web Search Ranking

Thesis Submitted To Al-Rayan University To complete requirements of obtaining a Master's degree in Information Technology

By

Mohammed Abdorabo Mohammed Ko'adan

Supervisors

Supervisor	Supervisor
Dr. Khalid Qaid Shafal	Dr. Mohammed Abdullah Bamatraf

1443/2021

Republic of Yemen

Ministry of High Education & scientific research

Al – Rayan University

Faculty of Higher Studies

Comparative Study Between K-Means and Genetic algorithm As a Technique for Web Search Ranking

Thesis Submitted To Al-Rayan University To complete requirements of obtaining a Master's degree in Information Technology

By

Mohammed Abdorabo Mohammed Ko'adan

Supervisors

Supervisor	**Supervisor**
Dr. Khalid Qaid Shafal	**Dr. Mohammed Abdullah Bamatraf**

1443/2021

Approval of the Proofreader

I certify that the master's dissertation titled ,

(Comperative study Between K-Means and Genetic algrithm as a technique for web Search Ranking)

submitted by the student Mohammed Abara ba Kaadan

has been linguistically reviewed under my supervision and has become in scientific style and clear from linguistic errors and for that I sign.

Proofreader : Abdullah Amer AL-Kathiri

Academic Title : Assitant Professor

University : Al-Rayan University

Signature : ....

Date : 14/7/2020

Approval of the Scientific Supervisor

I certify that this master's dissertation titled ,
(Comparative study Between K-means and Genetic As) Techniques for web Search ranking
submitted by the student , Mohammed A. M. Ko'adan

has been completed in all its stages under my supervision and so
I nominate it for discussion.

Supervisor : Dr. Mohd A. Bamatraf

Signature : ....

Date : 27/6/ 2021

The Discussion Committee Decision

Based on the decision of the President of the University No. (   ) in the year _______ regarding the nomination of the committee for discussing the master's thesis entitled (**Comparative Study Between K-Means And Genetic Algorithm As Technique for Web Search Ranking**) for the researcher **Mohammed Abdurabu Mohammed Ko'adan** We, the head of the discussion committee and its members, acknowledge that we have seen the aforementioned scientific thesis and we have discussed the student in its contents and what related to it.

Chairman of the Committee

Name: Saeed Mohammed Baneamoon

Signature

Committee member

Name: Mazin A. Bahashwan

Signature

committee member

Name: Mohammed A. Bamatraf

Signature

N.B :

Please write the academic title of each member of the discussion committee .

Professor Dr.

Associate Professor Dr.

Assistant Professor Dr.

قال تعالى:

﴿ وَعِندَهُۥ مَفَاتِحُ ٱلْغَيْبِ لَا يَعْلَمُهَآ إِلَّا هُوَ ۚ وَيَعْلَمُ مَا فِى ٱلْبَرِّ وَٱلْبَحْرِ ۚ وَمَا تَسْقُطُ
مِن وَرَقَةٍ إِلَّا يَعْلَمُهَا وَلَا حَبَّةٍ فِى ظُلُمَٰتِ ٱلْأَرْضِ وَلَا رَطْبٍ وَلَا يَابِسٍ إِلَّا فِى
كِتَٰبٍ مُّبِينٍ ﴾

(الأنعام - ٥٩)

Dedication

I dedicate this research to:

my parents who instilled in the spirit of steadfastness.

The soul of my father who I long for and how I wished he was by my side at such moments.

My university which allowed me to continue my higher studies

My friends who were my support and help in overcoming difficulties.

Those whom I have not met, and for whom I have great credit for providing the support, without which I would not have completed this

research

To all of these, thank you

Acknowledgment

First of all, I thank God for giving me the strength and the ability to complete this study.

Then, I would like to express my deep gratitude to my thesis supervisors:

Dr. Khalid Qaid Shafal his expertise, understanding, and patience added considerably to my graduate experience.

Dr. Mohammed Abdullah Bamatraf for his support through my studied and his advices to reach to the end of this research.

Also, I would like to thank **Student welfare association** and **Selah foundation for development** for them support, and help in scholarship.

Mohammed Abdorabo Ko'adan

Abstract

Nowadays, search tools are crucial to search for information on the Web. However, during this time, the amount of demand information that is available to us on the Web is changing and growing continuously, and those search technologies continually have been pushed to the absolute limit. Therefore, various new techniques have arisen to enhance the search process, and one of them is the analysis of query logs. Query logs have a mechanism to track the history of queries sent to the search engines and the pages favored after a search, among other data.

Classification of web pages is the first step of web page ranking (or we can call it indexing), one of the most ways to achieves indexing process is clustering those pages into groups as per the similarity, whenever the misclassification is too much less, the result will be perfect. Not far away, clustering is a collection of algorithms that divide the data into groups related to each other. For this job, we chose the (Microsoft learn to rank) dataset to achieve the analysis and model building on it, this dataset was designed especially for researches in this field, and it has enormous and different information about the ranking process. Because of a quantity of information, we chose randomly 16015 observations only from MSLR-WEB30K_2 _ fold 1; in this study, according to the ability of our hardware and the algorithms of analysis, some of the algorithms which used in the analysis (determine the optimal number of clusters) can't handle the massive quantity of observations. In this thesis, I will use clustering analysis to improve the web search ranking using comparative analysis between k-means and genetic algorithms as a technique used for this goal.

This process including the clustering analysis to reduce the features using Principal component analysis (PCA) with root main square error as feature reduction technique to compute the error rate and the accuracy of the model result to get the best number of attributes, this process achieved with cross-validation approach using extreme gradient boost algorithm as a training model to estimate the sum of errors during a training operation. After that, we will use various methods to determine the optimal number of clusters to ensure high-quality distribution of the data in the clusters. And to test the result of building models with k-means and genetic algorithms.

Contents **Page**

Contents **Page**

List OF Abbreviation

Symbols	**Nomenclatures**
MSLR-WEB	Microsoft Learn To Rank WEB
PCA	Principle Component Analysis
RMSE	Root Mean Square Error
XGBoost	eXtreme Gradient Boost
WCSS	within-cluster sum of square
NbClust	Number Of Clusters
GA	Genetic Algorithm
ASW	average silhouette width
CH -Index	Calinski-Harabasz Index

Chapter One: Introduction

1.1 Introduction

In the era of big data, and due to the massive amount of data in the network, the difficulty of getting the correct information is significantly increased. More and more people are beginning to realize the importance of network resource standardization and reorganization.

Clustering is unsupervised learning; it finds the natural grouping of instances given unlabeled data. Clustering is the process of gathering related objects into classes.(Padmaja & Sheshasaayee, 2016)

Web clustering technology has attracted much attention as an essential part of network resource management. It is mainly used in data extraction, knowledge discovery, data mining, and information fusion.(Z. Zhang et al., 2018)

A fundamental task of Web search engines is to rank the set of pages returned according to a user's query. So, the most relevant pages to the query; appear in the first places of the answer. Early search engines used similarity measures between queries and text in Web pages to rank pages. Variations of the vector model used for this purpose. However, these approaches did not achieve good results because the words in a document do not always capture properties such as semantics, quality, and relevance of the document; this process is not efficient when Web pages size reached half a billion pages. A further generation of ranking algorithms appeared in the late nineties. They take into account not only the text of Web pages but also Web hyperlinks. These algorithms explore the fact that links in the Web represent a source of human annotations about the quality of pages.(Baeza-Yates et al., 2004)

The critical factor for the popularity of today's web search engines is the cordial user interfaces, they provide the simplicity with which information can be retrieved, actually search engines allows users to post queries simple as phrases as a frame for a list of keywords, following the traditional approach of information retrieval systems.(Moreira et al., 2015)(Singh, 2013)

This process is just like this scenario. If you want to have an unforgettable dinner, the best friend who advises you with that restaurant to find your favorite food.

1.2 Motivation

A cluster analysis is a multivariate technique aimed at identifying web pages clustering. This approach offers several advantages to researchers, such as usefulness in web pages based on similarities among users, classifying users' groups, generating hypotheses about these users' groups and testing a concept to determine whether specific types of customers are present in the dataset, and it is possible to analyses a vast number of respondents effectively.

This technique has been used in the web search industry to identify topic clusters that are mutually exclusive and exhaustive.

Actually, we can express it as three motivations: ideal motivation, achievement motivation and self-expression motivation, that push the research to the end.

Many methods and algorithms were set to do the web pages clustering, some of those provide the less efficiency and some of them can keep the quality of clustering, and could not make stable models. So, it is necessary to develop more and new experiments with more model's evaluation metrics to measure quality, and efficiency.

1.3 Problem Statement

The major problem of in web search ranking is the misclassification of the topics, which cause the wrong ranking, and the wasting time in the big data.

1.4 Objectives:

The main objective is:

1- make clear comparison between k-mean and genetic algorithms as clustering algorithms.

2- and see how it is interacting in web pages clustering to develop the ranking with search engines results in indexing stage according to topic clustering fundamentals.

3- Provide a best of both algorithms in the web ranking field.

1.4 Scope and limitation

This research focuses only on web page ranking and comparison between clustering algorithms k-means algorithm and clustering Genetic Algorithm, using clustering mechanism here to let the machine make the decision in huge quantity of data.

1.5 Thesis organization:

Chapter one gives an overall introduction about the clustering techniques, research problem, and the thesis objective.

Chapter two interduce a general background of web pages clustering as method to rank the topics correctly, in addition, it discusses the related work in the same filed.

Chapter three in this chapter we explain the comparison, we started by describing the details of used dataset in the research, then we explain the data analysis and preprocessing, until we built the clustering models.

Chapter four in this section we discuss the results of our models according to the clustering metrics.

Chapter five this chapter summarize the main results we get from the experiments, and showing the conclusion and the future work.

Chapter Two: Background

2.1 Overview and background:

In fact, it is a simple interaction mechanism that proved to be successful for searching on the Web. A list of keywords is not always an excellent solution to the descriptor of the demand information of users. One reason for this is an ambiguity that shows up in many language expressions, such as ambiguity aroused because of polysemous words.(Singh, 2013)

The available data is a set of user logs from which we extract query sessions. Figure 2.1 shows the relationships between the different entities that participate in the process induced by using a search engine. Our approach focuses on the semantic relationship between queries (a notion of query similarity will define that) and the preferences/feedback of users about Web pages.(Baeza-Yates et al., 2004)

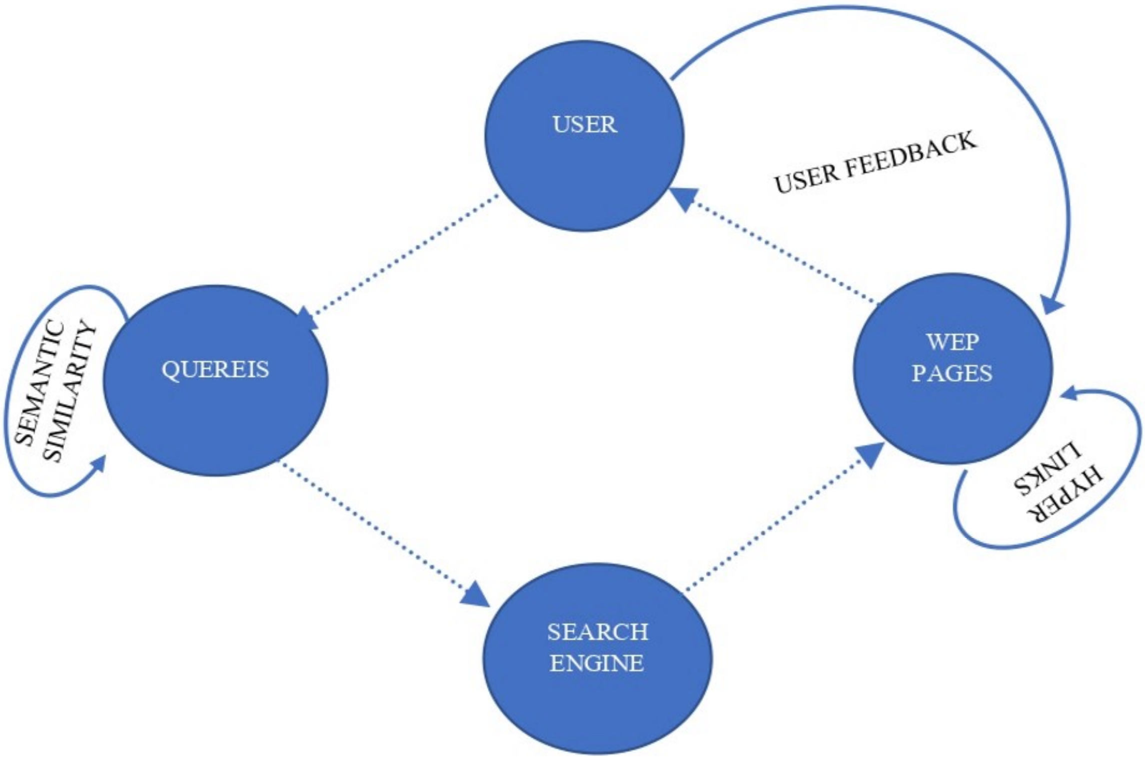

Figure. 2.1 Relationship between entities in the process of a search

The ranking algorithm is based on a clustering process that defines neighborhoods of similar queries according to user preferences. A single query (list of terms) may be submitted to the search engine several times, each defining a query session. This approach uses a simple notion of query session similar to the notion introduced by Wen et al.(Yadav & Jyotiyadavcs, 2016) which consists of a query, along with the URL's clicked in its answer:

QuerySession:=(query,(clickedURL)∗)

The selected URL’s are those the user that submitted a query clicked on until she/he submit another query. A more detailed notion of query session may consider the rank of each clicked URL and the answer page in which the URL appears, among other data that can be considered for other versions of the algorithm.

One of the research questions that motivates this work is concerned with the possibility of learning to rank approaches being effectively used in the context of expert search tasks to combine different estimators of expertise in a principled way to improve on the current state-of-the-art methods as shown in figure 2.2.

The expert finding problem can be formalized as follows. Given a set of queries Q = {q1;….; qm} and a collection of experts E = {e1; ….; en}, each of which is associated with specific documents that describe his topics of expertise, a training corpus for learning to rank is created as a set of query-expert pairs (qi; ej) ∈ Q × E, upon which a relevance judgment that indicates the match between qi and ej is assigned by a labeler. This relevance judgment is a binary label that indicates whether the expert ej is relevant to the query topic qi or not. For each instance (qi; ej), a feature extractor produces a vector of features that contains statistical values that are related to qi and ej. The features can range from classical IR estimators computed from the documents associated with the experts (e.g., term frequency, inverse document frequency) to link-based features that are computed from networks that encode relations between the experts (e.g., PageRank). The inputs to the learning algorithm comprise training instances, their feature vectors, and the corresponding relevance judgments. The output is a ranking function, h, which produces a ranking score for each candidate expert ej so that, when sorting experts for a given query according to these scores, more relevant experts appear on the top of the ranked list.(Baeza-Yates et al., 2004)

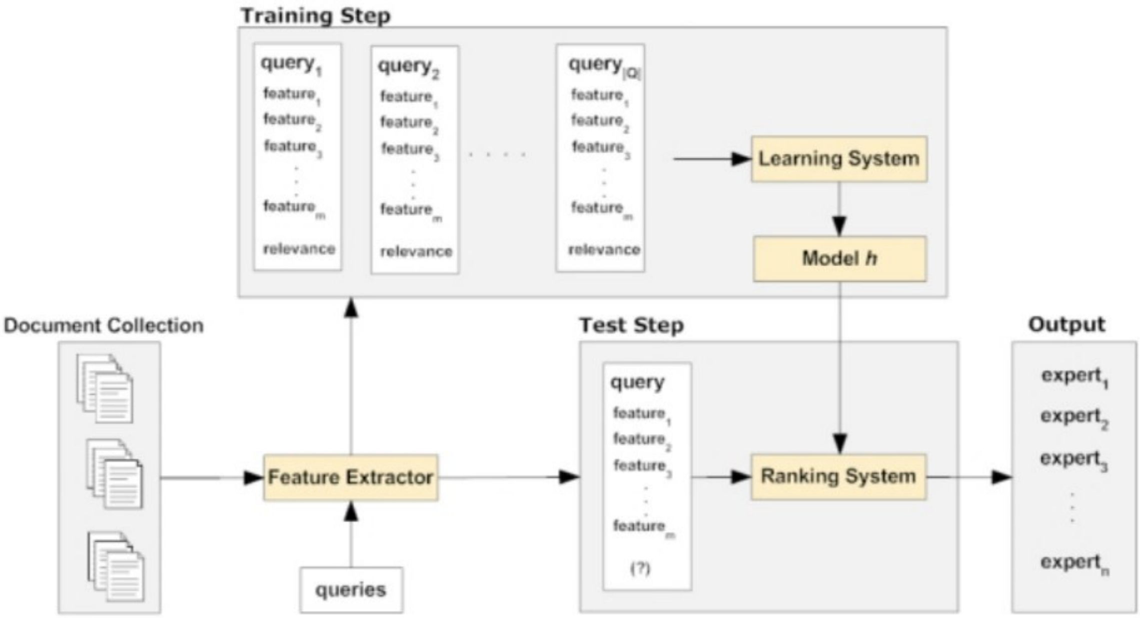

Figure 2.2: The learning to rank framework for expert finding

Various ranking factors that affect in a website if will show on top depends on the content which related to the search term, or the quality of backlinks pointing to the page.

2.1.1 Different search engines – different rankings:

A URL's ranking for keywords or keyword combinations defers much more than we imagine from search engine to search engine.

For example, in Bing search engine, a domain may rank for a particular keyword in the top three. However, not the same idea in Google search results for the same keyword, Bing, Google, Yahoo, and every other search engine uses its ways and methods for computing rankings and therefore ranks websites differently.

Nowadays, language or country can also vary when rankings using different methods of the same search engine, such as Google.

The goal of good rankings in the search results is to gain as much traffic as possible from the organic search channel, pages that have high-ranking results for a search query will get a higher chance is that the researcher will click on this result. And that will explain the direct connection between high rankings and increased traffic.

This relationship between rankings and clicks (and traffic) is most robust amongst the top 3 search results as shown in figure 2.3.(Padmaja & Sheshasaayee, 2016)

Figure2.3: The relationship between rankings and clicks

Unsupervised clustering is a type of clustering where items, based on how similar they are, agglomerated together automatically.

We use unsupervised clustering to help us find out the topic of a site and how that topic affects its traffic.

2.1.2: Understanding Internet Usage Patterns with Site Categories:

We find out what sites are about by categorizing their topics. For example, our data can tell us that sites about data science get ten times as much traffic as sites about a famous restaurant.

It can confidently say there are many more data science fans if I see the same number of panelists visiting a data science-themed site as a themed restaurant site.

While categorizing sites might sound easy (and it is, for a *single* site), things get complicated when we want to do this for every site on the web.

A couple of other things that make categorizing sites by topics tricky:

1- Huge data of possible topics.

2- The site can have multiple topics, and they may not be related in a meaningful way.

We can deal with the first one by compacting the topics by clustering, which represents the proper solution as a machine learning term for agglomerate similar things together. As an example, the cluster of sports topics would include things like baseball, football, soccer, and underwater handstands.

The second difficulty with categorizing site topics: a site can have multiple topics, and they may not be related in a meaningful way. For example, Contrast espn.com with sportsauthority.com. They are both about sports, but one is a news aggregator (among other things), and the other is a store. We deal with this issue by letting a site belong to multiple clusters.

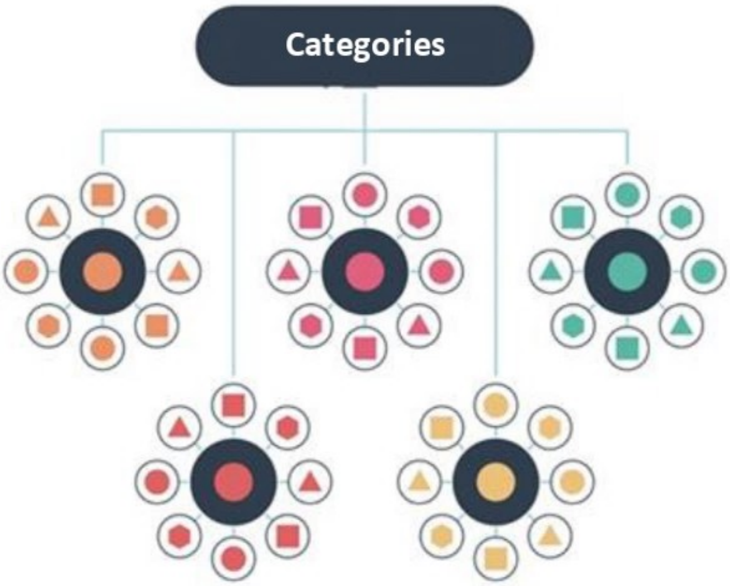

Figure 2.4: categorizing site topics

To simplify this idea, let us look at a travel website; it might have pages giving a general overview of different countries. Also, it may have pages talking about different cities. Then, topic clusters come where each page about a country contains links to different pages about cities in that country.

Moreover, it is possible to link the city pages to "places to visit" within each city page. On and on like, that is how it works. Basically, this will organize the ecosystem.h and we can consider that website as a web.

A single "pillar" page acts as the central hub of content for an overarching topic, and multiple content pages related to that same topic link back to the pillar page and each other. This linking action signals to search engines that the pillar page is an authority on the topic, and over time, the page may rank higher and higher for the topic it covers. (Baeza-Yates et al., 2004).

The term topic clusters presented toward reliance on topics occurred with Google in a project called (RankBrain) update.

Launched in 2015, (RankBrain) is Google's machine-learning algorithm designed to understand the context of people's search queries. It associates past searches with similar themes and pulls multiple keywords and phrases associated with the search query to find the best results.

2.2 Related Work:

- Ashok Kumar D, Usha T. A, Sivaranjani C. in this research, the authors collate The source data of K-means clustering and genetic algorithm on the basis of their functioning principle, advantage and disadvantage using Mushroom, Irish, Soybeans dataset with 119 items for experimentation. A set of association rules are determined by using K-Means and a Genetic Algorithm. By analyzing the data and providing various support and techniques. In this study, two types of techniques were used to determine the support, confidence, memory space, execution time accuracy of mushroom, Irish, soybeans data set. High accuracy achieved through the K-Means technique compares to GA.

  The result of the experiment is:

  - K-Means is an intuitively simple and effective clustering technique, but it may get stuck at suboptimal solutions, depending on the choice of the initial cluster centers.
  - A GA-based clustering technique is expected to provide an optimal clustering, more superior to that of the K-Means algorithm, but with a bit more time complexity.(Ashok et al., 2016)

- Shashi Mehrotra, Shruti Kohli. focus on the scope of improvement for the search result of a website using clustering approaches for improving web elements analysis; the main focus is to optimize the search result of a website using the clustering approach. Thus the proposed hybrid model will be using K-Means and a Genetic algorithm to overcome the drawbacks of K-Means. The evaluation parameters, accuracy in terms of object placement in the correct cluster, relevancy, speed, and user satisfaction are the main parameters considered for the study.

  The study is organized as follows: Section I: covers an introduction to Web Analytics, clustering, necessary clustering steps, and where it is being used. Section II: gives an overview of some significant work done so far, which includes some popular clustering algorithms, tools, and some approaches proposed used by the researcher. Section III: includes experimental analysis.

  They use several tools to analyze the data as available tools for social media and business, which are listed in the table like:

- Google Analytics.
- Yahoo! Web Analytics.
- Website Optimizer.
- Optimizely.
- Kissinights.

The used data were collected from the Similar Web Data repository. The table contains search keyword data derived from the Jabong website. Data is clustered on the search term. The following are the data sets used for the experiment.

File: SearchOrganiCompetitors File description: No of Instances: 3338. No of attributes: 5, Attribute names: Domain, Category, Global Rank, Affinity, and AdSense.

The result of this experiment is:

- There is no single clustering algorithm that can solve all problems.
- A clustering algorithm should include some characteristics to make it more efficient such as being able to use a big volume of data and high dimensional features.
- Outlier should be handle properly.(Mehrotra, 2015)

- Divyashree G, Gayathri Rayar: the study focuses on cosine similarity, Jaccard, Pearson Coefficient, and K-Means algorithm will be optimized by using the genetic algorithm in the proposed work.

  The performance metric such as purity, entropy, and F-measure will be evaluated for the K-means clustering algorithm and genetic algorithm, and the result is expected to possess' higher score of purity, a lower score of entropy, and maximized F-measure value.

  They suppose that Text mining is a part of data mining; its aim is to extract high-quality information from the given text. The extraction of high-quality information can be done through statistical pattern learning. Text mining includes information retrieval, lexical analysis, pattern recognition, information extraction, data mining techniques, association analysis, visualization, and predictive analytics.

They obtain important objects to achieve:

1. The problem of efficiency: Text document clustering must be efficient because it should be able to do clustering on ad-hoc collections of documents, e.g., ones found by a search engine through keyword search.
2. The problem of effectiveness: Text document clustering must be effective, i.e., it should relate to documents that talk about the same or a similar domain.
3. The problem of explanatory power: Text document clustering should be able to explain to the user why a particular result was constructed. Lack of understandability may pose a much bigger threat to the success of an application that employs text document clustering than a few percentage points decrease the accuracy.
4. The problem of user interaction and subjectivity: Applications that employ text document clustering must be able to involve the user. The results should be focusing one's attention on particularly on relevant subjects. For example, a search for "health" might turn up food-related issues that a user might want to explore in details relevant for him, such as "fruits," "meat," "vegetables," and others.(Divyashree & Rayar, 2015)

- Kunnuri Lahari et al.: enhanced reduce the local minima using evolutionary and population-based methods like Genetic algorithm and teaching learning-based optimization. The data sets iris and wine are used, and the experimental results are compared with the Genetic algorithm and teaching-learning-based optimization-based clustering with the k-means algorithm. The performance of the evolutionary-based clustering method compared with some existing clustering methods.(Lahari et al., 2015)

- Agustin Blas et al.: described the performance of the grouping Genetic algorithm in clustering, started with proposed encoding and different modification of crossover and mutation operation, and also initiated the local search include with the island model to improve the performance of the difficult situation. The real data sets like iris and wine were used and compared the results with the classical approaches such as DBSCAN and K-means, and obtaining the perfect results in proposed grouping based methodology the evolutionary approach such as Genetic algorithm. The performance of the algorithm was measured by using the different fitness functions.(Agustín-Blas et al., 2012)

- Rajashree Dash et al.: discussed comparative analysis of K-means and Genetic algorithm based on clustering. With respect to the idea were improved the cluster quality from K-means clustering using a Genetic algorithm. Large-scale clustering problems in data mining also address by this method. The best results are achieved by using this method. He found K-Means is an intuitively simple and effective clustering technique, but it may get stuck at suboptimal solutions, depending on the choice of the initial cluster centers. In contrast, GA is a randomized search and optimization technique guided by the principles of evolution and natural genetics and having a large amount of implicit parallelism. So it provides near-optimal solutions for the objective or fitness function of an optimization problem. Under limiting conditions, a GA-based clustering technique is expected to provide an optimal clustering, more superior to that of the K-Means algorithm, but with little more time complexity.(Dash & Dash, 2012)

- Anusha et al. depicted an enhanced K-means Genetic algorithm for optimal clustering. The author overcomes the disadvantage of local optima with a suitable dataset, and the algorithm fails in computational time. It is inferred that the technique produced more than 90% accuracy for a real-life dataset. The author also adopted a neighborhood knowledge strategy for optimizing multi-objective troubles. This algorithm used k means Genetic algorithm to find the smallness of the clusters. It is noted that the algorithm could produce a minimum index value for the maximum datasets.(Anusha & Sathiaseelan, 2014)

Chapter Three: Research Methodology

3.1 Methodology

Data mining is one of the most developed technologies in recent years. To research in this field, we follow the steps of data mining modeling to perform the comparison.

So before explaining the research methodology, here we have to overview the used tools in this research; first of all, the used language is R statistics with RStudio: which one of the most popular IDE to develop the data mining models and do good research, because it is an open-source language and have variant libraries about all methods and algorithms we ever imagine to develop any model or make any machine learning research.

R also has variant functions to computes the most clustering analysis calculus and most powerful visualization methods to present the results of the models as graphs, and that's why we prefer to choose it to performs this study.

Now, the following description in the figure 3.1 demonstrates the methodology:

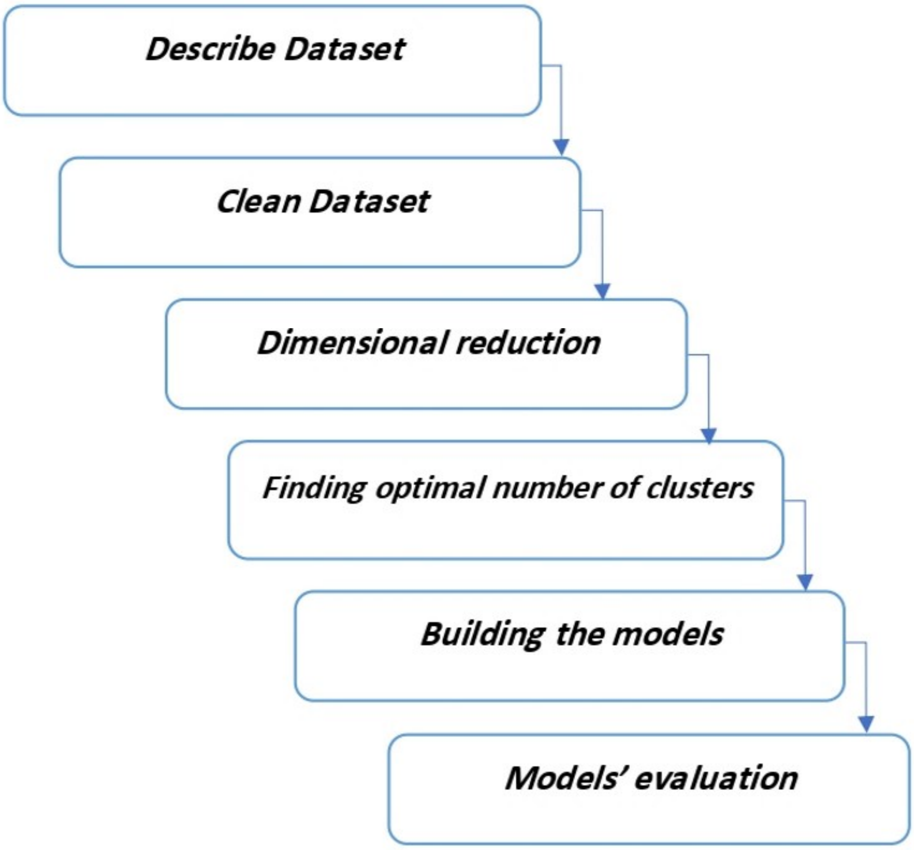

Figure 3.1. Methodology steps of research

Step1: Describe Dataset:

Defining the relationship between the attributes of our data, data can give much more answers about how its structure and how we can deal with it. Therefore, the first process we should be concerned about is discovering the correlation between each attribute to the others.

In this step, we determined the correlation, which gave a clear understanding of what type of algorithms can help the modeling process and the best way to clean the data and preprocessing process.

Step2: Clean Dataset:

Cleaning the data is the second most crucial step before doing anything with data. In most cases, the data represents the raw materials of data mining modeling, so in this step exactly, we could clean the data from the outlier results, which can give the error ratio in modeling. Then, we perform some algorithms to determine the unnecessary attributes according to the near-zero variance representing less than 0.1% variance. This step removes 16 attributes which reduces the consumption time of processing.

Step 3: Dimensional reduction:

In this step, we depend on the Principal Component Analysis algorithm to achieve this process because it is familiar with clustering algorithms and can handle the massive data with the Xgboost algorithm, representing the Queen of data mining modeling algorithms.

This process is done by regarding the use of all principal components. We need the correct number of those components to fit it with the model to reduce the features and reduce the processing time by computing the error ratio in dataset attributes.

Step 4: Finding the optimal number of clusters:

After we clean the data and reduce the features, we have to find the optimal number of clusters as the most important parameter to use in the k-means algorithm and clustering genetic algorithm.

We used many famous methods to find clusters, and we didn't depend on only one method to make sure the results represented the proper number.

This process is important as the first thing that must be done before building the clustering model to make sure our model's results are reliable.

Step 5: Building the models:

After we get the proper number of clusters, we can perform our models to compare the results. Here we fit the k-means algorithm and save the results, then fit the genetic clustering algorithm and save the results to the comparison. Within this step, we visualize the results of the models as graphs to give a clear understanding of the results.

Step 6: Models' evaluation:

After performing the models here, we evaluate the results several times and mention the best results to compare the models.

3.2 Data Description:

Search engines, or information retrieval at large, play an important role in the modern Internet. Given any query from the user, an ideal search engine should match the related web pages and rank them based on relevance with the query. Since very oftentimes, the user only takes a look at the top-ranked webs, and it is crucial to locate the most relevant web pages. Hence, it is interesting to learn the relevance between the query and web page by data mining algorithms. A line of works called Learning to Rank (LetoR)(Joachims, 2002),(Xu & Li, 2007),(Rudin & Schapire, 2009),(Burges, 2010)focused on this learning problem, and several algorithms have been proposed. Some of these algorithms are applied in commercial search engines such as Google, Bing, and Yahoo (T. Y. Liu, 2009).

To better research LetoR, Microsoft Research, and Yahoo have provided large-scale datasets: LetoR 3.0(Microsoft Research Asia, 2009), LetoR 4.0 (Qin & Liu, 2013), MSLR-WEB, Yahoo Challenge (Chapelle, 2011). Unlike Yahoo Challenge dataset, Microsoft Research has given descriptions on how to generate the features of all their datasets. Hence, Microsoft datasets(Microsoft Research Asia, 2009) ,(Qin & Liu, 2013) are more useful in research. However, there are two challenges with the datasets MSLR-WEB:

- Insufficient baselines reported. For LetoR 3.0 and LetoR 4.0, baseline results are posted on the website [18],[19], and they have extensive research works presenting experimental results on them (Minka & Robertson, 2008),(M. Zhang et al., 2009),(Chapelle, 2011),(Macdonald et al., 2013). For MSLR-WEB, the authors did not give baselines. To our knowledge, there are only two existing works reporting experiment results on MSLR-WEB(Suhara et al., 2013). However, in [24], only limited models and evaluation metrics are reported.
- Too many features. MSLR-WEB has 136 features. Note that each of the LetoR datasets has only less than 50 features. It is interesting to investigate the significance of each of the 136 features. This paper will try a feature selection. Another issue is that some features in MSLR-WEB are developed and privately owned only by Microsoft (e.g., the features on user click data, Boolean model, and language model). This will make the dataset not fully reproducible. By feature selection, we can evaluate the importance of these private features.

Tables 1, 2, and 3 are an adaptation of the feature list given by the MSLR-WEB website. All the features are numeric. The original datasets have the matched document for at most 30,000 queries, separated in 5 folds for five-fold cross-validation. However, in this report, we only use the Fold 1 data in order to reduce the training time to an appropriate scale.

Table No (3.1): Description of the MSLR-WEB dataset (Part 1)

Feature	No.	Description
covered query term number	1-5	How many terms in the user query are covered by the text. The text can be body, anchor, title, url and whole document (for features 1 - 5 respectively, similarly below).
Covered query term ratio	6 – 10	Covered query term number divided by the number of query terms
stream length.	11 - 15	Text length
IDF (inverse document frequency)	16 - 20	1 divided by the number of documents containing the query terms.
sum of term frequency.	21 - 25	Sum of counts of each query term in the ocument
min of term frequency	26 - 30	Minimum of counts of each query term in the document
max of term frequency.	31 - 35	Maximum of counts of each query term in the document
mean of term frequency	36 - 40	Average of counts of each query term in the document
variance of term frequency	41 - 45	Variance of counts of each query term in the document.
normalized sum of stream length	46 - 50	Sum of term counts divided by text length.
normalized min of stream length	51 - 55	Minimum of term counts divided by text length.
normalized max of stream length	56 - 60.	Maximum of term counts divided by text length
normalized mean of stream length	61 - 65	Average of term counts divided by text length.
normalized variance of stream length	66 - 70	Variance of term counts divided by text length.

Table No (3.2): Description of the MSLR-WEB dataset (Part 2)

Feature	No.	Description
sum of tf*idf	71 - 75	Sum of the product between term count and IDF for each query term
min of tf*idf	76 - 80	Minimum of the product between term count and IDF for each query term
max of tf*idf	81 - 85	Maximum of the product between term count and IDF for each query term
mean of tf*idf	86 - 90	Average of the product between term count and IDF for each query term
variance of tf*idf	91 - 95	Variance of the product between term count and IDF for each query term
boolean model.	96 - 100	Unclear. Privately owned by Microsoft
vector space model.	101 - 105	dot product between the vectors representing the query and the document. The vectors are privately owned by Microsoft
BM25	106 - 110	Okapi BM25
LMIR.ABS	111 - 115	Language model approach for information retrieval (IR) with absolute discounting smoothing
LMIR.DIR	116 - 120	Language model approach for IR with Bayesian smoothing using Dirichlet priors
LMIR.JM	121 - 125	Language model approach for IR with JelinekMercer smoothing
number of slashes in URL	126	e.g., “ucsb.edu/pstate/people” has 2 slashes.
length of url	127	The number of characters in the URL
Inlink number	128	The number of web pages that cite this web page.

Table No (3.3): Description of the MSLR-WEB dataset (Part 3)

Feature	No.	Description
Outlink number.	129	How many web pages this web cites.
PageRank	130	Evaluates the centrality of this web page based on web links over the Internet. This gives the success of Google.
SiteRank	131.	Site level PageRank. E.g., "ucsb.edu/pstat" and"ucsb.edu/math" share the same SiteRank
QualityScore	132	The quality score of a web page. The score is outputted by a web page quality classifier. Privately owned by Microsoft.
QualityScore2	133	The quality score of a web page. The score is outputted by a web page quality classifier, which measures the badness of a web page. Privately owned by Microsoft.
Query-url click count	134	The click count of a query-url pair at a search engine in a period. Collected and privately owned by Microsoft.
url click count	135	The click count of a url aggregated from user browsing data in a period. Collected and privately owned by Microsoft.
url dwell time	136	The average dwell time of a url aggregated from user browsing data in a period. Collected and privately owned by Microsoft.

3.3 Preparing Dataset:

Data mining is extremely helpful in information pre-processing and integration of databases. Data processing permits the researchers to spot co-occurring sequences and the correlation between any activities. Data visualization and visual data processing facilitate the data science with a transparent read of the data (Mohan, 2010).

Each search engine has to implement three stages to be optimized 1) indexing. 2) Sorting. 3) Storing. The indexing stage is our concern during this study to find the optimum result in search method, in web search engines, restoration process of data contains of the three important activities that made the search engine, which is cowling, dataset and the algorithm of search program. As short explanation about how this process happened. We can see it happen by calling the sites which related to the words or phrase that the user enters into searching interface. The extremely trick half is that the result ranking. The ranking is additionally what spend all time and effort attempting to affect (Ledford, 2015)

Nowadays the researches target to boost this field by decreasing the loss within the data to present higher results. Therefore, we decide to choose Cluster analysis because of it is a statistical technique within which a collection of objects or points with similar characteristics jointed together along in clusters. It encompasses a variety of various algorithms and strategies that are used for grouping objects of comparable types into various classes (Matt Young, 1990) (Chawla, 2016)

3.3.1 FEATURE SELECTION

Feature selection is a popular technique for processing steps used for pattern recognition, classification and compression schemes. In many data analytics problems, data redundancy is a challenge that must be avoided, so reduce dimensionality is an essential step before achieving any data analytics. The general criterion for reducing the attributes is the goal to present most of the related information of the original data accordingly to some optimality criteria. In some applications, it might be necessary to pick a subset of the features instead of finding a mapping that uses all of the features. The benefits of using this subset of features could reduce the usage of resources during computational of unnecessary features and costs of sensors (in physical measurements systems). (Lu et al., 2007) As an unsupervised technique Principal Component Analysis is a feature reduction algorithm for projection high dimensional data into a new subset of low dimensional, which represents the data as possible with minimum reconstruction error. PCA is a quantitatively rigorous method to perform this facilitation. The mechanism of this algorithm based on creating a new sub set of features called principle components, all of them are linear combination to the original features, and all of them are orthogonal to each other. That means no redundant in dataset. (Ahmad & Amin, 2015)

A. Principle Component Analysis(PCA):

In Conventional supervised FS evaluation methods working with various feature subsets using an evaluation function or metrics to choose only the features that have relation to the decision classes of the data.

However, in many data mining fields, decision class labels always unknown or incomplete, this indicates the meaning of unsupervised features selection.

In unsupervised learning, decision class labels are always unknown (Bartholomew, 2010)

Keeping the efficiency and accuracy in model building is a big challenge and too difficult especially with big number of features, so here we need to use PCA as an efficient feature reduction technique. Therefore, we are looking to linear transformation of a random vector with zero mean and variance matrix x to a lower dimension random vector

$X \in K^n$ With zero mean and covariance matrix $\sum_x$

x to a lower dimension random vector $Y \in K^q$, $q < n$

$$Y = A_q^T X \qquad 3.1$$

$$\text{With } A_q^T A_q = I_q$$

$$where\ I_q\ is\ the\ q \times q\ identity\ matrix$$

In PCA, A_q is a $n \times q$ matrix whose columns are the q orthonormal eigenvectors corresponding to the first q largest eigenvalues of the covariance matrix $\sum x$. these have been many perfect characteristics of the linear transformation; one of them is maximization the distribution of points between the axes in the graph which represents the avoiding the linearity of the data. Another important characteristic is minimization the mean square error between the prediction data to the original data. (Lu et al., 2007)

Here we recognize how to implement PCA in the real models thorough recognizing of PCA provides an essential for relativizing the fields of data mining a d dimensional reduction (Masaeli et al., 2010).

```
Algorithm : PCA
Input : Data Matrix
Output: Reduce set of features.
Step1: X     Create N x d data Matrix with one row vector Xn per data point.
Step2:  subtract mean $x$ from each row vector $xn$ in X.
Step3: Σ ←covariance matrix of X.
Step4: Find eigenvectors and eigen values of Σ.
Step5: PC's ←the M eigenvectors with largest eigen values.
Step6: Output PCs.
```

B. Root Mean Square Error(RMSE):

Root mean square error is the standard way to measure the error of a model in predicting quantitative data, Formally it is defined as :

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{3.2}$$

Where :

ŷ1,ŷ2,……….ŷn is predicted value.

y1,y2,……..yn is observed value.

n number of observations.(Chai & Draxler, 2014)

To understand why this measure of error makes sense for a mathematical perspective, we will ignore the division by n under square root, we can notice a similarity to the Euclidean distance formula.

$$distance(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad 3.3$$

If we keep n fixed and rescale the Euclidean distance by a factor of $\sqrt{1/n}$.

And consider our observed value determined by adding random error to all predicted value:

$$y_i = \hat{y}_i + \epsilon_i \text{ for } i = 1, \ldots, n$$

Where: $\epsilon_1, \ldots, \epsilon_n$

independent, identically distributed error.

Those errors, as random variables, might have Gaussian distribution with mean μ and standard deviation σ .

The mean of distribution error is μ when we want to estimate the standard deviation we can see that:

$$\mathbb{E}\left[\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}\right] \quad 3.4$$

$$= \mathbb{E}\left[\frac{\sum_{i=1}^{n} \epsilon_i^2}{n}\right] \quad 3.5$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\epsilon_i^2] \quad 3.6$$

$$= \mathbb{E}[\epsilon^2] \quad 3.7$$

$$= Var(\epsilon) + \mathbb{E}[\epsilon]^2 \quad 3.8$$

Where

$$= \sigma^2 + \mu^2 \quad 3.9$$

- E[..] is the expectation.
- Var (..) is the variance.(Chai & Draxler, 2014)(Madsen et al., 2005)

And because ε_i is already variable with the same ε, so we can replace it to E[ε_i^2] which is the average of the expectations.

And to be remembered we will suppose our error distributed with μ =0.and by putting this in equation and add the square root of both sides then the output:

$$\sqrt{\mathbb{E}\left[\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}\right]} = \sqrt{\sigma^2 + 0^2} = \sigma \quad 3.10$$

And here we can observe the left side looks familiar, if we ignore the practice E[…]from the square root, it is definitely the formula of RMSE. (Jachner et al., 2007)

C. eXtreme Gradient Boost(XGBoost):

extreme Gradient Boost is a decision-tree-based ensemble data mining algorithm that uses a gradient boosting framework.

It includes efficient linear model solver and tree learning algorithm, and applied in predicting of unstructured data problems (Image, text, …etc), the algorithm differentiates in the :

•It can be employed in regression, classification, ranking, and so.

• It is a general portable algorithm that can run smoothly with all famous operating systems.

• It can support AWS, Azure and yarn cluster.

• It can handle various types of input data.

• It support customized objective function and evaluation function.

• It has higher performance on various datasets.

• It is Fast.(Chen et al., 2018)

The power of this algorithm represents in its scalability which can achieve fast learning through parallel and distributed computing and presents efficient memory usage. So, to understand however it's operating, this will explain the steps of the algorithm:

1. Fit a simple linear regression or decision tree on data

2. Computes error residuals. Actual aim value, minus predicted target value.

3. Fit a new model on error residuals as the aim variable with the same input variables.

4. Add the predicted residuals to the previous predictions.

5. Fit another model on residuals that is still left. and repeat steps 2 to 5, until it starts overfitting or the sum of residuals becomes constant and the figure 3.2 will show how it works.

Overfitting can be controlled by systematically checking accuracy on validation data.(Torlay et al., 2017)

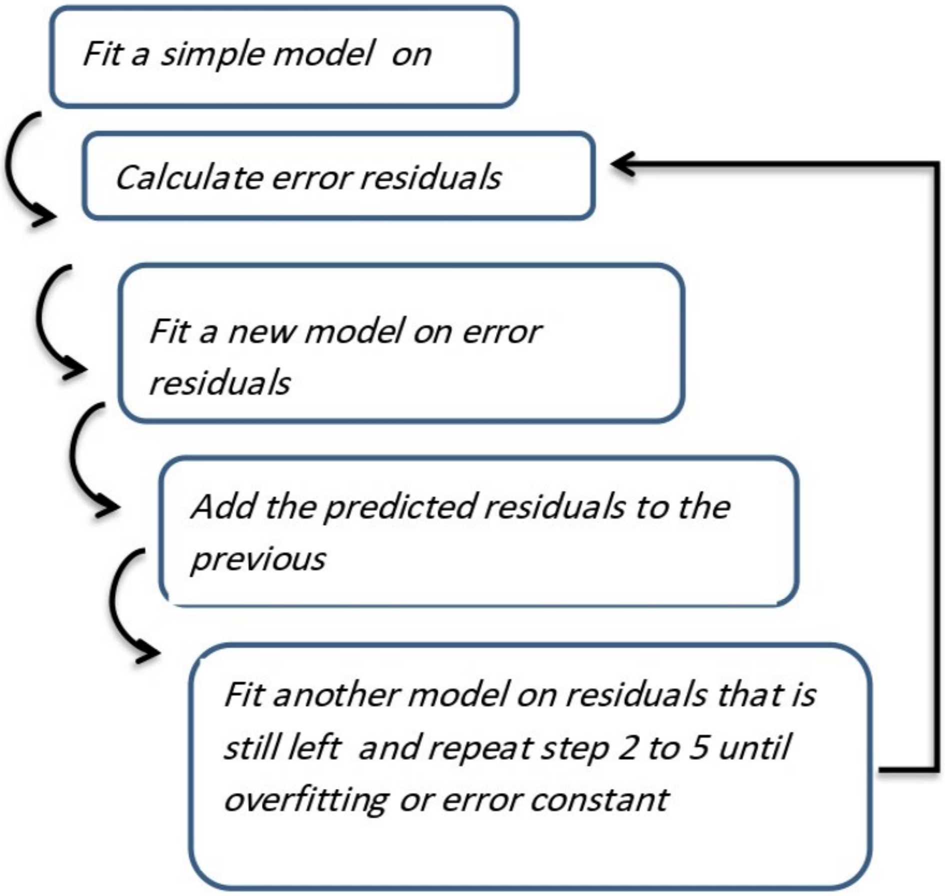
Fit a simple model on
Calculate error residuals
Fit a new model on error residuals
Add the predicted residuals to the previous
Fit another model on residuals that is still left and repeat step 2 to 5 until overfitting or error constant

Figure 3.2. Shows the XGBoost algorithm diagram

3.3.2 Feature Reduction Process:

Here we are aiming to clean the dataset from the variables, which are zero variance predictors, which means we have to eliminate those variables, which have very few unique values according to the number of samples, and the ratio of the frequency of the most common value against to the frequency of the second most common value is large. After that, we tend to transform the data via principal component analysis, to prepare the data for modeling. Actually, we use PCA here for two reasons: 1) to solve multicollinearity problem, which is shown in variables, that have high correlation between each other, 2) to reduce the features (feature selection process). After that we will use extreme gradient boost (which is called XGBoost function in R) to train the model after the previous two steps (cleaning and transforming data) using XGBoost here to compute the error rate during useing number of PCA component to find the proper number of these components which give stable results, instead of using it all, this step is to avoid the outlier problem which happed because of irregular objects that make the model unstable. So initially we will Error = 1, and PC from 1-5, during running the model, the model will run with 100 round each time to calculate the sum of root mean square error(RMSE), and gradually we will increment the PCA component 5 each time as below:

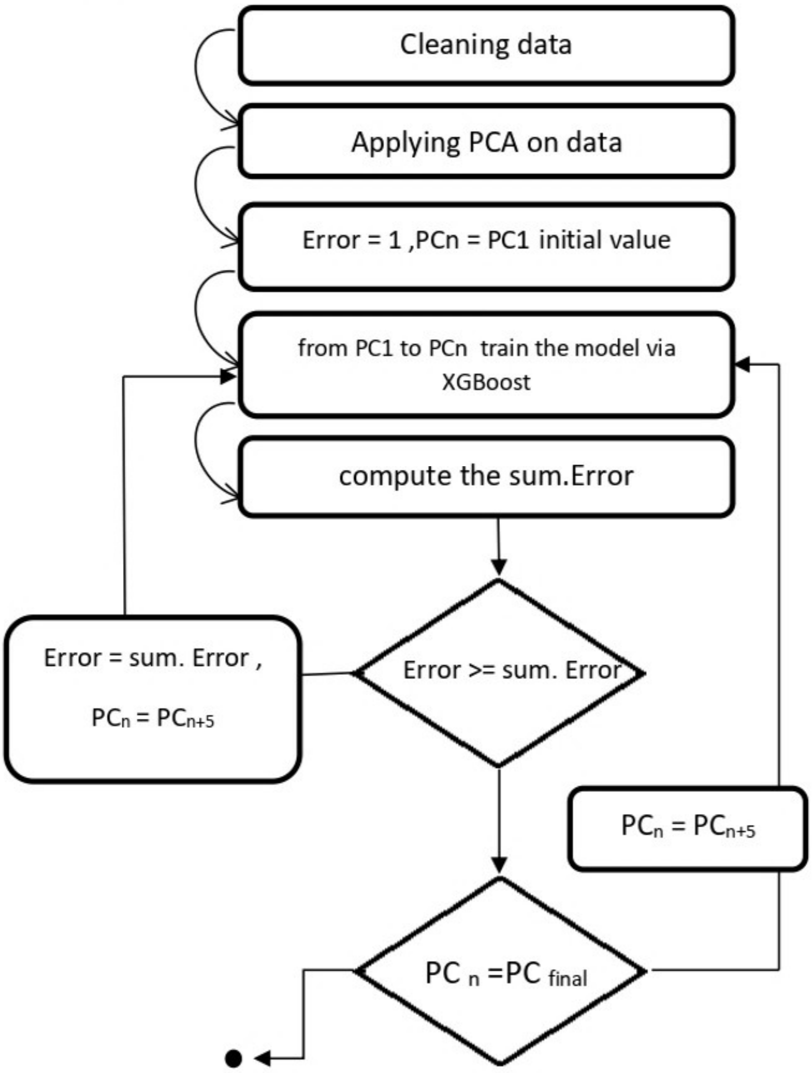

Figure3.3 feature reduction algorithm

As the figure (3.3) shown above we initiate the PCA component fromPC1 to PC5, to fit the XGBoost model for first time modeling, and compute the error rate of 100 round, second time we will increment the PCA components another 5 components (range will be from PC1 to PC10), and run the model to compute the error rate again, and so on. Each time we increment PCA components gradually 5 components to fit the model and compute the error rate, until we reach to final number of PCA components. As a result of these steps, we get the table below:

Table No (3.4) Error Estimation for XGBoost Model Table

RMSE	PCA Comp
0.0703	PC5
0.055	PC10
0.0501	PC15
0.0503	PC20
0.0449	PC25
0.0418	PC30
0.0412	PC35
0.0366	PC40
0.0374	PC45
0.0373	PC50
0.038	PC55
0.038	PC60
0.0381	PC65
0.0398	PC70
0.0399	PC75
0.0394	PC80
0.0398	PC85
0.0399	PC90
0.038	PC95
0.0372	PC100
0.0367	PC105
0.0368	PC110
0.0369	PC115
0.0368	PC120

Table No (3.4) describe the result of using Xgboost algorithm with PCA to reduce the features with RMSE as metric to determine the suitable number of features to fit in clustering model to check is our hypotheses valid or not. So, the next step is to experiment the result to build K-Means algorithm model, and check the stability of clusters by measuring the distance between and within clusters.

3.3 Clustering Modeling:

A cluster model is working to groups data, so that objects in the same group have similar characteristics (so they are "homogeneous") compared to those in other groups (which are "heterogeneous"). According to the 2017 KDnuggets poll[1], clustering is the second most popular machine learning method, and k-means is undoubtedly one of the most popular algorithms used. It is used in various ways, such as finding groups of similar customers, segmenting a market, optimizing a communication strategy and improving marketing effectiveness, positioning products, or selecting test markets.(Velchamy et al., 2011)

Before we start to build our model, we have to find the optimal number of clusters. This process is very important because it determines how the model is stable.

The first step to build a k-means clustering model we have to determine the optimal number of clusters.

3.3.1 The optimal number of clusters:

finding the optimal *k* should be one of the k-means algorithm's parameters, and this parameter is responsible for the stability of clusters and the whole model.

Therefore, to achieve this step, there are many metrics to find the optimal number of clusters What we used is as follows:

1) The Elbow method defines the clusters as that the total intra-cluster variation or total within-cluster sum of square (WCSS) is minimized. It measures the compactness of the clustering, and we want it to be as small as possible (Kassambara, 2017)

[1] https://www.kdnuggets.com/2017/12/top-data-science-machine-learning-methods.html

To make it simple, the Illustration of the K value on Elbow combination with K-Means was the graph of cluster relationship with error decreasing, increasing value of K then the chart will decrease slowly until the result of the value of K is stable. For example, the value of the cluster K = 2 to K = 3, then from K = 3 to K = 4 shows a drastic decrease in the elbow at point K = 3. Then the ideal cluster k is K = 3. The combined Elbow and K-Means Methods can determine the value of K at the best cluster.

The elbow method consists of plotting in a graph the WCSS(x) value (within-cluster sums of squares) on y-axis according to the number x of clusters considered on the x-axis, the WCSS(x) value being the sum for all data points of the squared distance between one data point x_i of a cluster j and the centroid of this cluster j (as written in the formula below), after having portioned the dataset in x clusters with the k-means method.

$$\mathrm{WCSS}(k) = \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in \text{cluster } j} \|\mathbf{x}_i - \bar{\mathbf{x}}_j\|^2, \tag{3.1}$$

where $\bar{\mathbf{x}}_j$ is the sample mean in cluster j

In figure (3.4) We can see that the optimal number which are given by elbow method is

K = 3 where is the knee of graph is indicated to 3.in addition of above the disadvantage of this method is sometimes can't give us the exactly number of k, so it's not the perfect method to depend only in clustering analysis, but it's useful to make the decision.

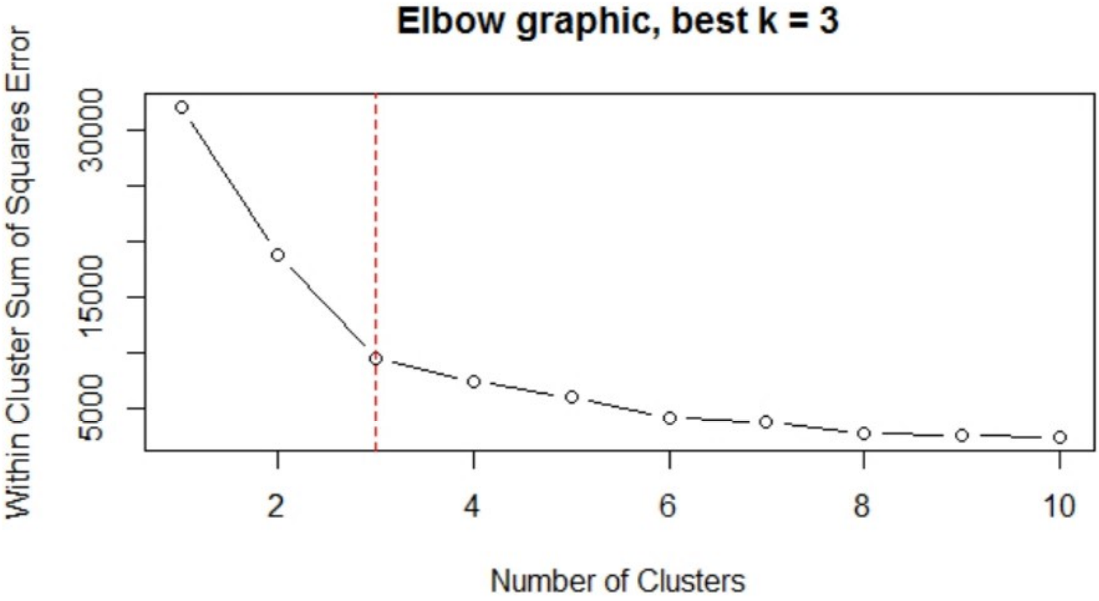

Figure3.4. Elbow representation of best K

2)NbClust: it's a package provides 30 indices for determining the optimal number of clusters, it's proposes the best clustering by variant methods to determine the best number, all used different assumed numbers of clusters, distance measures and clustering methods.(Charrad et al., 2014)

This method uses many distance metrics to determine the optimal number of clusters.

The following distance measures are written for two vectors x and y. They are used when the data is a d-dimensional vector arising from measuring d characteristics on each of n objects or individuals.

- Euclidean distance: Usual square distance between the two vectors (2 norm).

$$d(x,y) = \left(\sum_{j=1}^{d} (x\mathrm{j} - y\mathrm{j})^2\right)^{1/2} \qquad 3.2$$

- Maximum distance: Maximum distance between two components of x and y(supremum norm)

$$d(x,y) = \sup|xj - yj| \quad 3.3$$

$$1 \leq j \leq d$$

- Manhattan distance: Absolute distance between the two vectors (1 norm).

$$d(x,y) = \sum_{i=1}^{d} |xj - yj| \quad 3.4$$

- Canberra distance: Terms with zero numerator and denominator are omitted from the sum and treated as if the values were missing.

$$d(x,y) = \sum_{j=1}^{d} \frac{|xj - yj|}{|xj| + |yj|} \quad 3.5$$

- Binary distance: The vectors are regarded as binary bits, so non-zero elements are "on" and zero elements are "off". The distance is the proportion of bits in which only one is on amongst those in which at least one is on.
- Minkowski distance: The p norm, the p^{th} root of the sum of the p^{th} powers of the differences of the components.

$$d(x,y) = \left(\sum_{j=1}^{d} |x_j - y_j|^p \right)^{1/p} \quad 3.6$$

The table below summarizes indices implemented in NbClust and the criteria used to select the optimal number of clusters.

Table No (3.5): summarizes indices implemented

No.	Index in NbClust	Optimal number of clusters
1	"kl" or "all" or "alllong"	Maximum value of the index
2	"ch" or "all" or "alllong"	Maximum value of the index
3	"hartigan" or "all" or "alllong"	Maximum difference between hierarchy levels of the index.
4	"ccc" or "all" or "alllong"	Maximum value of the index
5	"scott" or "all" or "alllong"	Maximum difference between hierarchy levels of the index.
6	"marriot" or "all" or "alllong"	Max. value of second differences between levels of the index
7	"trcovw" or "all" or "alllong"	Maximum difference between hierarchy levels of the index.
8	"tracew" or "all" or "alllong"	Maximum value of absolute second differences between levels of the index.
9	"friedman" or "all" or "alllong"	Maximum difference between hierarchy levels of the index.
10	"rubin" or "all" or "alllong"	Minimum value of second differences between levels of the index.
11	"cindex" or "all" or "alllong"	Minimum value of the index.
12	"db" or "all" or "alllong"	Minimum value of the index.
13	"silhouette" or "all" or "alllong"	Maximum value of the index.
14	"duda" or "all" or "alllong"	Smallest *nc* such that index > critical value.
15	"pseudot2" or "all" or "alllong"	Smallest *nc* such that index < critical value.
16	"beale" or "all" or "alllong"	*nc* such that critical value of the index >= alpha
17	"ratkowsky" or "all" or "alllong"	Maximum value of the index
18	"ball" or "all" or "alllong"	Maximum difference between hierarchy levels of the index.
19	"ptbiserial" or "all" or "alllong"	Maximum value of the index.
20	"gap" or "alllong"	Smallest *nc* such that critical value >= 0
21	"frey" or "all" or "alllong"	the cluster level before that index value < 1.00
22	"mcclain" or "all" or "alllong"	Minimum value of the index
23	"gamma" or "alllong"	Maximum value of the index
No.	Index in NbClust	Optimal number of clusters

24	"gplus" or "alllong"	Minimum value of the index
25	"tau" or "alllong"	Maximum value of the index
26	"dunn" or "all" or "alllong"	Maximum value of the index
27	"hubert" or "all" or "alllong"	Graphical method
28	"sdindex" or "all" or "alllong"	Minimum value of the index
29	"dindex" or "all" or "alllong"	Graphical method
30	"sdbw" or "all" or "alllong"	Minimum value of the index

Figure 3.5 shows us that number 3 is the optimal number of clusters by most of NbClust package measures.

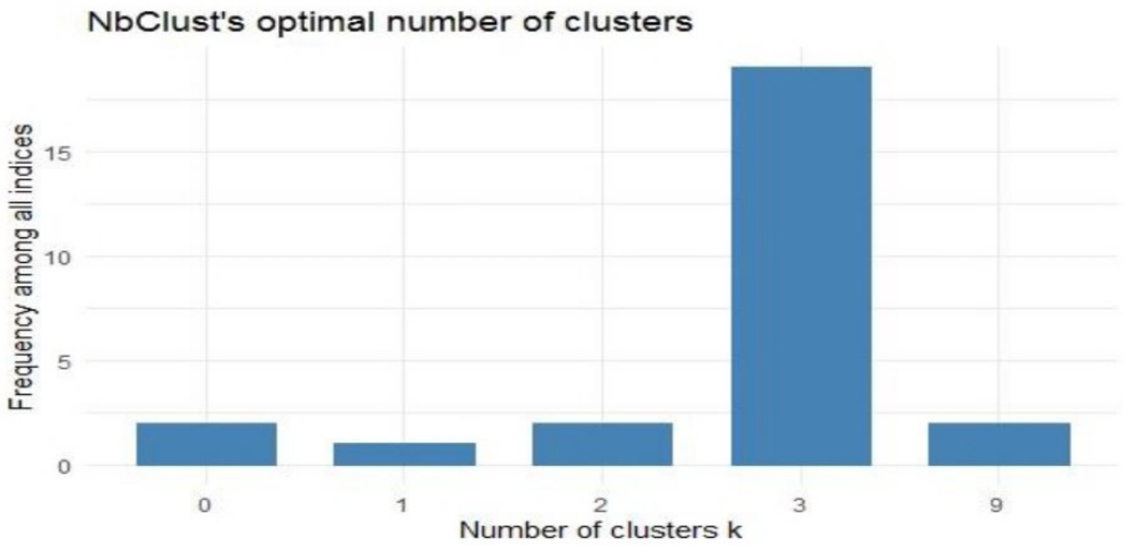

Figure 3.5 Show NbClust method

2) Gap statistics: it uses the comparative between the total intra-cluster variation and their expected values under null reference distribution of the data, the optimal value will be that maximize the gap. (Xu & Li, 2007)

   The gap statistic was developed by Stanford researchers Tibshirani, Walther and Hastie in their 2001 paper. The idea behind their approach was to find a way to standardize the comparison of $\log W_k$ with a null reference distribution of the data, i.e. a distribution with no obvious clustering. Their estimate for the optimal number of clusters K is the value for which $\log W_k$ falls the farthest below this reference curve. This information is contained in the following formula for the gap statistic:

$$Gap"_{n}(k) = E_{n}\{\log W_{k}\} - \log W_{k} \qquad 3.7$$

   Here in Figure (3.6) clearly we can see gap statistics method suggest the optimal number of clusters is two.

   According to previous methods, we recommend to choose number 3 as the optimal number of clusters because it is the estimated number of two methods of determining optimal number of clustering and that is enough for building model.

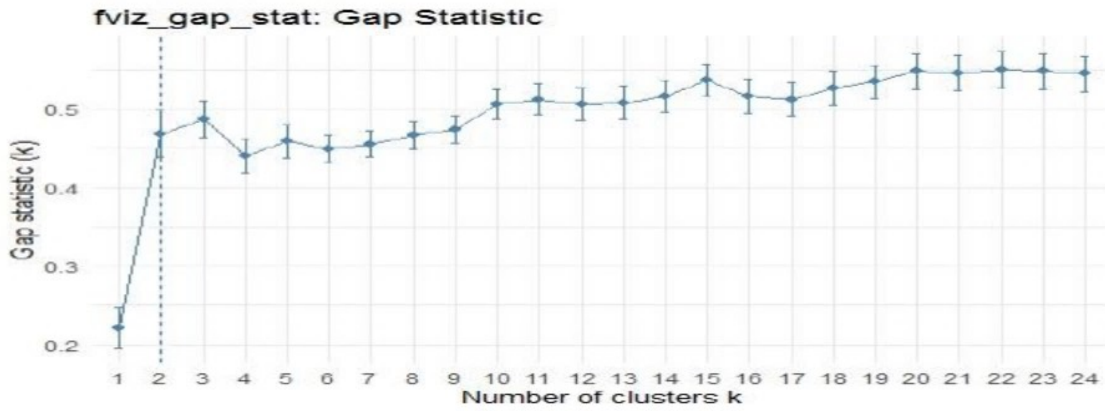

Figure 3.6 Shows Gap method results.

After we use many methods to determine the optimal number of clusters, we can strongly say the optimal number of clusters according to the previous, is three (3) to build model with strong and connected model.

3.3.2 K-Means Algorithm Model:

Being a clustering algorithm, k-Means takes data points as input and groups them into k clusters. This process of grouping is the training phase of the learning algorithm.

The result would be a model that takes a data sample as input and returns the cluster that the new data point belongs to, according to the training that the model went through. With this way, we can find how it is useful, in a very simplistic manner.

Websites may choose to put people in bubbles (i.e., clusters) with other people who share similar activities (i.e., features) on the website. By way, the recommended content will be somewhat on-point, as existing users with similar activities are likely to be interested in similar content. Moreover, as a new person goes into the ecosystem of the website, person will be placed within a particular cluster, and the content recommendation system takes care of the rest(Rajaraman & Ullman, 2012).

Building on that idea, k-Means is just a clustering algorithm. It uses the distance between points as a measure of similarity, based on k averages (i.e. means).(Dhanachandra et al., 2015)

A. K-means in brief:

Here we can see how the k-means algorithm work according to its mechanism:

```
Assign initial values for each u (from u=1 till u=k);
Repeat
         {
          Assign each point in the input data to the u that is closest to it in value;
          Calculate the new mean for each u;
          {if all u values are unchanged {break out of loop; }
        }
```

To explain this, initially k number of so-called centroids are chosen. A centroid is a data point (imaginary or real) at the center of a cluster, each centroid is an existing data point in the given input data set, picked at random, all centroids are unique (that is, for all centroids ci and cj, ci ≠ cj). These centroids are used to train a classifier, the resulting classifier is used to classify (using k = 1) the data and thereby produce an initial randomized set of clusters(Aurélien Géron, 2019). Each centroid is thereafter set to the arithmetic mean of the cluster it defines. The process of classification and centroid adjustment is repeated until the values of the centroids stabilize. The final centroids will be used to produce the final classification/clustering of the input data, effectively turning the set of initially anonymous data points into a set of data points, each with a class identity as shown in figure 3.7.(Bachem et al., 2016)

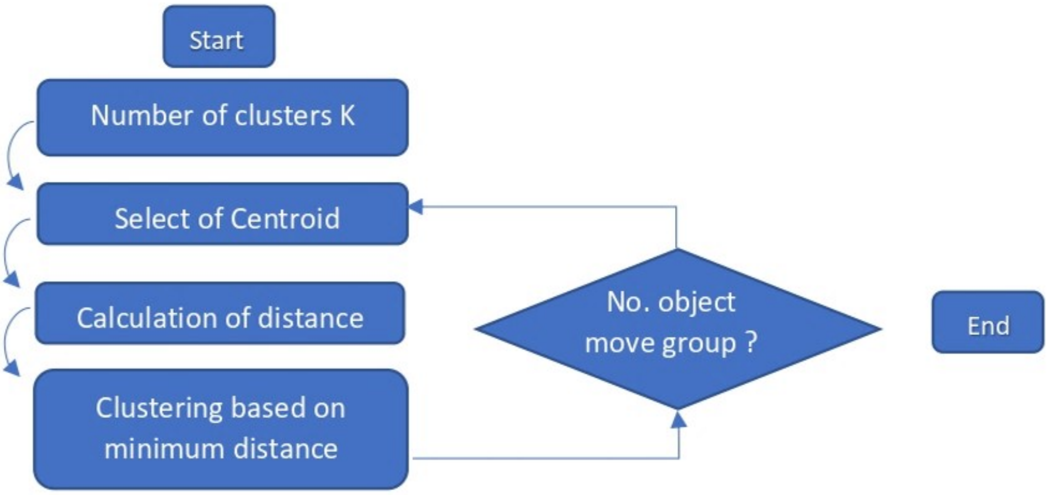

Figure 3.7 K means algorithm diagram

According to the previous chapter, we prepare the dataset for Model building process with PCA.

for this experiment we use the following parameters to perform the clustering:

Table No (3.6): parameters of K-means algorithm model

Parameters	Value
No. of clusters	3 clusters
No. iterations	30 iterations
No. start	25.

And to make it clearer we will explain what is this parameter means:

1- No. of clusters: is a numeric parameter represents the number of clusters which you want to customize, this number can be found by the optimal clusters number algorithms which help too much in this step, finding the optimal clusters number of k-means algorithm is important thing to do before we start model building, because this step determines the purity of the model and can limit the accuracy of our model.

2- No. iterations: which is represent to the parameter "iter.max" in "kmean" function in R language, iter.max parameter is the number of times the algorithm is run before results are returned.

   if we say iter.max=1 it could work but it's a nonsense as for the most times 1 iteration is not enough, because the algorithm works finding a minimum of a cost function through several iteration steps.

3- No. start: which is "nstart" parameter in "kmeans" function in R language, nstart represents the number of random data sets used to run the algorithm. It means that in order to get the algorithm initialized, it should feed the algorithm with the initials coordinates of the cluster centers, which it do not want to do manually. So nstart extract for a number of times (example nstart=10) k random numbers between 1 and the number of "observations" in your csv file (the number of lines) and it takes that line as a starting point for cluster 1 up to cluster k candidates centroids.

B. k-means model results:

After we determine the optimal number of clusters and fit the model the outputs of this experiment as follow:

the output distributed within 3 unequal clusters (C1, C2, C3) and the observations distributed as:

C1	C2	C3
6574	3444	5992

And the time consuming for modeling is: 0.64 sec.

C. k-means model visualization:

the final step is to represent the model visually to find out the outcomes of k-mean model

K-means algorithm Model

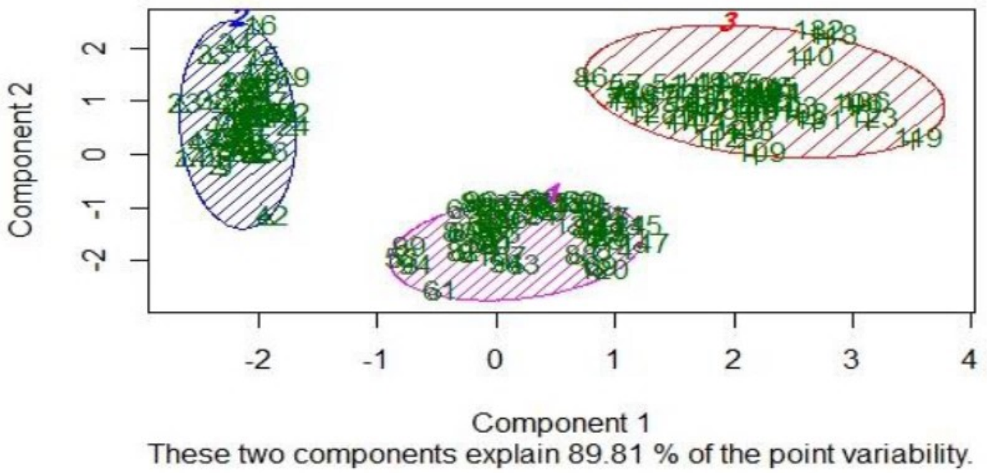

Figure 3.8 Shows the k-means algorithm model

According to figure 3.8 we can see how the observations allocated in its clusters according to the distance, and how the clusters are stable, according to the distance between each other, and how the connectivity inside each cluster which connect the element of the clusters.

3.4.3 Genetic Algorithm for Clustering (GA)

While clustering refers to assigning categories to describe a dataset and its points according to its similarities or dissimilarities. It separates the data into partitions while elements in the same partition are similar to each other and dissimilar to the elements in the other partitions, in agreement with some criterion or distance metric. It is widely used to classify datasets when the target class is not known.(Sultana, 2016)

Another way to do clustering is by using evolutionary approaches to find the best partitions. From an optimization perspective, clustering can be formally considered as a particular kind of *NP-hard* grouping problem (Falkenauer 1998). Under this assumption, a large number of evolutionary algorithms for solving clustering problems have been proposed in the literature. These algorithms are based on optimizing some objective function (i.e., the so-called fitness function) that guides the evolutionary search (Hruschka et al., 2009).

Following the taxonomy proposed by Hruschka et al. (2009), we aim to present Genetic Approach to Maximize (here we called it (gama)), a clustering criterion — an R package for evolutionary hard partitional clustering, by using guided operators (those shown by some information about the quality of individual clusters), for a fixed *a priori* known number of partitions, encoded as real-valued centroid based partitions.(Al-Subaihin & Sarro, 2020)

The main advantage of the proposed technique is to allow the user to realize clustering guided by the maximization of a chosen cluster validation criterion. Genetic search enables the algorithm to find a centroid configuration so good as more significant the number of generations used until the convergence for a local (reasonable) or global maximum if there are any.(Balato & Vitelli, 2014)

A. Genetic algorithm for clustering in brief

Here we can understand the genetic algorithm when it is working as clustering according to its mechanism asshown in figuer 3.9:

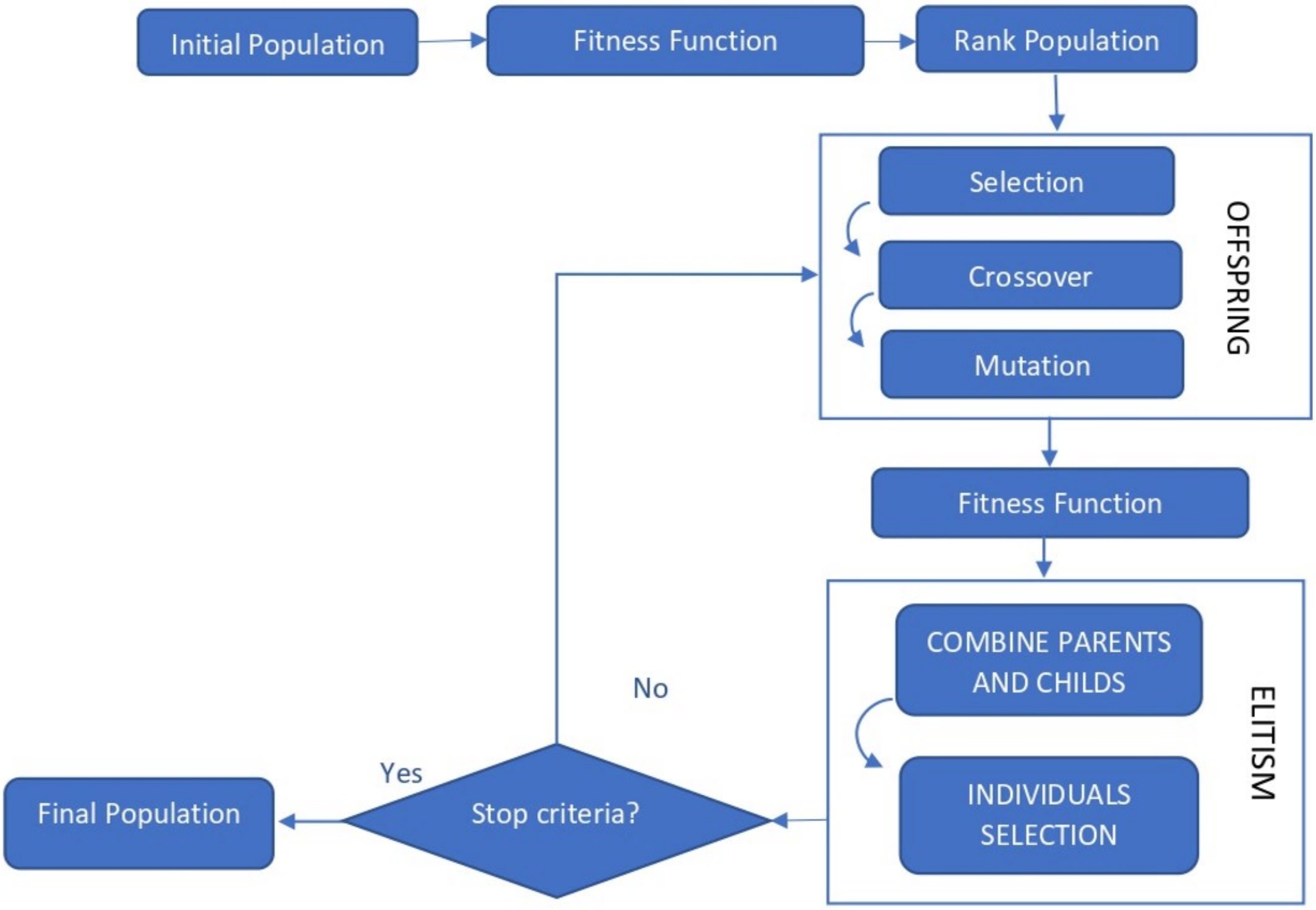

Figure 3.9 Genetic algorithm diagram

String representation for GA: Each string is a sequence of real numbers representing the K cluster centers. For an N-dimensional space, the length of a chromosome is N*K words. The first N positions (or, genes) represent the N dimensions of the first cluster center. The following N positions represent those of the second cluster center so on.(Akmaliyah, 2013)

Population initialization: The k cluster centers encoded in each chromosome are initialized to k randomly chosen points from the data set. This process is repeated for each of the P chromosomes in the population, where P is the size of the population.(Akmaliyah, 2013)

Fitness computation: In the fitness computation process, the constant number k is generated randomly, which is considered as the number of clusters to be formed, and the clusters are then created according to the centers encoded in the chromosome under consideration. This is done by calculating the sum of the distance of each data from each cluster center in a cluster, and this should be done for each individual. After this, by comparing the sum of each individual, the individual having minimum sum is identified, which means that the sum of all clusters in that individual is minimum, which will give us the fitness function. For example, suppose there are three clusters, i.e., C1, C2, and C3, formed according to the centers encoded in the chromosome, i.e., individuals. Data having minimum distance from the cluster center is assigned to the corresponding cluster Ci. Then the sum of distances between the data and the cluster center for each cluster is computed as follows where i vary from 1 to L/2, where L represents the length of chromosomes.(Akmaliyah, 2013)

Selection: The selection process selects chromosomes from the mating pool directed by the survival of the fittest concept of natural genetic systems. A chromosome is assigned several copies according to the selection process. The chromosomes having minimum fitness value (proportional to their fitness in the population) goes into the mating pool for further genetic operations. For this, the Roulette wheel selection process is used.(Akmaliyah, 2013)

Crossover: Crossover is a probabilistic process that exchanges information between two-parent chromosomes for generating two-child chromosomes. Single point crossover with a fixed crossover probability that is kc is used. For chromosomes of length L, a random integer, called the crossover point, is generated in the range [1, L-1]. The portions of the chromosomes lying to the right of the crossover point are exchanged to produce two offspring.(Akmaliyah, 2013)

Mutation: Each chromosome undergoes mutation with a fixed probability km that is called mutation probability. For the binary representation of chromosomes, a bit position (or gene) is mutated by simply flipping its value. Since we are considering floating-point representation, we use the following mutation. A number d in the range [0, 1] is generated with uniform distribution. The value at a gene position is X; then, if the value of d<0.5, then after mutation, it becomes X-1 otherwise, it becomes X+1.(Akmaliyah, 2013)

For this experiment, we use the following parameters to perform the clustering:

Table NO (3.7): parameters of Genetic algorithm model

Parameters	Value
No. of clusters	3 clusters
Cross over rate	9%
Mutation rate	1%
No. of Generations	30 generations
Fitness function	Average silhouette width "ASW"
Population size	25 observations

Before we show the results, we have to make a clear understanding of those parameters Parameters understanding help us to read the results correctly, which is:

1- **No. clusters**: it means the optimal number of estimated clusters according to the optimal number of cluster algorithms as above in the K-means algorithm.

2- **Crossover rate**: the probability of crossover between pairs of chromosomes.

3- **Mutation Rate**: the probability of mutation in a parent chromosome. Usually, a mutation occurs with a small probability.

4- **No. of Generations**: the number of generations to execute the search.

5- **Fitness function**: the key point of the genetic search. The algorithm will search for the ideal centroids that maximize one of the pre-specified criteria. The default value is "ASW" and will be assumed if the user does not supply value or put an invalid entry.

6- **Population size**: the number of individuals in the population. This argument has a significant impact on the performance of the search due to the increased number of matrix calculations when the population grows.

B. Genetic Model Results:

After we fit the model, the outputs of this experiment as follow:

The size of clusters is:

C1	C2	C3
5773	3356	6885

The time consumed to build the model is 11m:56.4 sec, and that is too much time can't confirm to build any model.

And to represent the model graphically, we can see the figure 3.10:

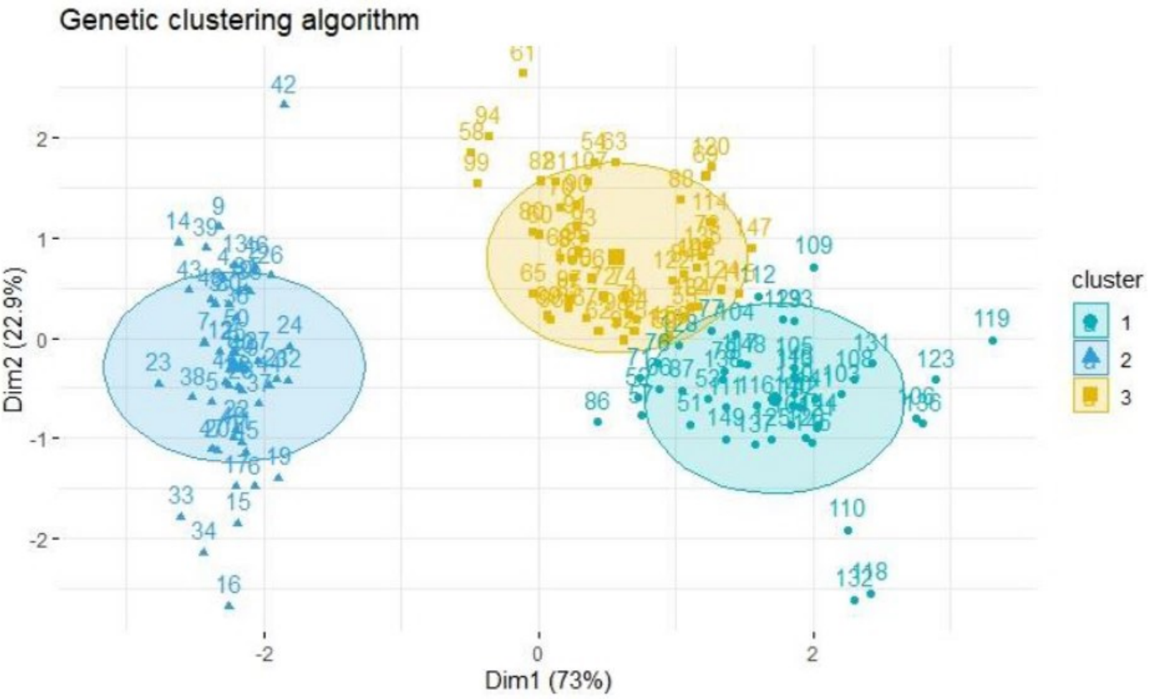

Figure 3.10 Genetic algorithm model

As we can see, the clusters located near each other accept cluster No. 2 is far much from the other clusters (1,3), and this told us about the relationship between the tight clusters that contained the closest data.

Chapter Four: Results & Discussion

4.1 Model evaluation

The selection of the optimal algorithm to construct a cluster model is a daunting task for the researcher conducting practical experiments. In addition, another problem is determining the optimal number of blocks that represent the data.

On the positive side, the calculations must be done carefully to ensure high-quality results. The main objective of the various set of clustering assessment criteria is to verify the validity of the clustering processes results (analysis and determination of the best performing clustering algorithm based on the results of practical experiments).(Brock et al., 2008)

The clustering results validation process is summarized into three groups:

Internal evaluation criteria, stability evaluation criteria, and biological criteria.

Whereas the internal evaluation criteria are concerned with ensuring the strength of the internal interconnection of the components that make up each tire and the extent of their closeness and suitability to each other. That is what we can call it clustering quality.

Stability metrics assess the stability of the clustering result by comparing it to groups obtained by removing one column at a time.

However, here our concern is to evaluate the algorithms results with the internal and stability metrics.(Everitt, 1980)

4.1.1Internal Evaluation:

To use internal standards, we must understand the nature of their work, which reflects the compactness, interconnectedness, and separation of groupings.

To clarify this matter, verification is based on the principle of communication between all the elements of each cluster separately, showing the correct positioning of these similar elements in the same group. This is done by measuring internal communication.

Internal measures take a clustering and the underlying dataset as the input and use information intrinsic to the data to assess the clustering quality. Using the same categorization for clustering methods, it can be grouped according to the particular notion of clustering quality that they employ.(Desgraupes, 2018)

To clarify the internal measures here, it is:

A. Connectivity:

It is a technique to assess clusters, which provide a measure of the quality of a particular division by measuring the interconnectedness of the elements of each cluster with each other based on the distance between each element within the group.(Desgraupes, 2018)

Let *N* denote the total number of observations (rows) in a dataset and *M* denotes the total number of columns, which are assumed to be numeric (e.g., a collection of samples, time points, etc.). Define *nni* (*j*) as the *j*th the nearest neighbor of observation *i*, and let *xi;nni*(*j*) be zero if *i* and *j* are in the same cluster and 1=*j* otherwise(Emmons et al., 2016). Then, for a particular clustering partition *C* = *{C*1*;….;CKg}* of the *N* observations into *K* disjoint clusters, the connectivity is defined as:

$$Conn(\mathrm{C}) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{inn_{i(j)}} \qquad 4.1$$

Where L is a parameter giving the number of nearest neighbors to use. The connectivity has a value between zero and one and should be minimized.(Brock et al., 2008)

After we do the experimental process, we found the following results:

When the numbers of clusters = 3:

Table No (4.1): shows connectivity results

	Elbow dis.	Manhattan dis.	Correlation
k-means algorithm	2308.851	2034.659	2401.38
Genetic algorithm	10.34881	11.55556	9.612302

According to the above result, the k-means algorithm represents the best performance than the genetic algorithm in the building that has strong inter-relation clusters.

We can't ignore something in the results. The correlation represents a higher result, which means the correlation is one of the important indexes that must be fucosed on during clustering analysis. And that will give k-means algorithm an additional point in Comparison.

B. Silhouette Width:

One of the most common methods is silhouette display, which depends in its mechanism on two criteria: separation, which means the average distance between the elements from different groups, and the second element is merging, which means the average distance between the elements within each group separately, Silhouette width is calculated of every object of the classification thus indicating how well they fit into their respective cluster. The cluster-wise or the global mean of objects can be used to assess the distinctness of specific clusters or the validity of the total classification, respectively. Due to the compactness criterion involved as average within-cluster distance, silhouette prefers spherical cluster shapes.(Lengyel & Botta-Dukát, 2018),With well-clustered observation shaving values near one and poorly clustered observations having values near -1 for observation i, it is found as:

$$S(i) = \frac{b_i - a_i}{maax[b_i - a_i]}$$

where a*i* is the average distance between i and other elements in the same group, and b*i* is the average distance between i and the elements in the nearest neighboring group.(Everitt, 1980)(Wang & Xu, 2019).

So, after practicing the formula above, we found the following results:

In total average silhouette width (ASW) as:

Table No (4.2): Shows ASW results

	ASW
k-means algorithm	0.813333
Genetic algorithm	0.56

To give more details, we can calculate the ASW per cluster for each algorithm to clarify the Comparison between the algorithms, so for the k-means clustering, the results are:

	C1	C2	C3
C1	0	0.79	0.77
C2	0.79	0	0.88
C3	0.77	0.88	0

And for genetic algorithm the results are:

	C1	C2	C3
C1	0	0.84	0.28
C2	0.84	0	0.56
C3	0.28	0.56	0

C. Calinski-Harabasz Index:

The Calinski-Harabasz index, also known as the Variance Ratio Criterion, is the ratio of the sum of between-clusters dispersion and inter-cluster dispersion for all clusters. The results are often higher when the groups are well separated, which relate to a standard concept of a cluster(Emmons et al., 2016).

In a simple way, The index depends on the dispersion, and the within-cluster dispersion is the sum of the squared distances between the observations $M_i^{\{k\}}$ And the barycenter $G^{\{k\}}$ of the cluster. Finally, the pooled within-cluster sum of squares (WGSS) is the sum of the within-cluster dispersions for all the closers:

$$WGSS = \sum_{k=0}^{K} WGSS^{\{k\}} \tag{4.3}$$

Geometrically, this sum is the weighted sum of the squared distances between the $G^{\{k\}}$ and G, the weight being the number n_k of elements in the cluster C^k:

$$BGSS = \sum_{k=0}^{K} n_k \left\| G^{\{k\}} \; - G \right\|^2 \tag{4.4}$$

Using the notations of the above equations, the Calinski-Harabasz index is defined like this:(Desgraupes, 2018)(Everitt, 1980)

$$C = \frac{BGSS/(K-1)}{WGSS/(N-K)} = \frac{N-K \;\; BGSS}{K-1\, WGSS} \tag{4.5}$$

The results of applying this metric are impressive, which proves the excellence of k-mean clustering in this area (particularly with this simple dataset)(Wang & Xu, 2019). To compare the performance of any clustering algorithms, it's good to use The Calinski-Harabasz index,

Has the following results:

Table No (4.3): Shows CH-index results

	CH index
k-means algorithm	556.8795
Genetic clustering algorithm	420.9526

To understand the results, the high output is the best, which indicates to k-means.

D. C- Index:

It is an internal evaluation metric that compares the total internal dispersion in the cluster. Ideally, a value for a number of clusters that reduces this indicator to a minimum will be considered the optimal number of clusters.

In the cluster *Ck*, there are *nk*(*nk* - 1)=2 pairs of distinct points (the order of the points does not matter). Let us denote by *NW* the total number of such pairs:

$$N_W = \sum_{k=1}^{K} \frac{n_k(n_k - 1)}{2} \qquad 4.6$$

The total number of pairs of distinct points in the data set is

$$N_T = \frac{N(N-1)}{2} \qquad 4.7$$

Let us consider the distances between the pairs of points inside each cluster. The numbers N_W and N_T have been defined. One computes the following three quantities:

- S_W is the sum of the N_W distance between all the elements within each cluster.
- S_{min} is the sum of the N_W smallest distances between all the pairs of points in the entire data set. There are N_T such pairs one takes the sum of the N_W smallest values.

- S_{max} is the sum of the N_W most considerable distances between all the pairs of points in the entire data set. There are N_T such pairs: one takes the sum of the N_W most significant values.(Desgraupes, 2018)(Bezdek et al., 2016)

The C index is defined like this:

$$C = \frac{S_{W-} S_{min}}{S_{max} - S_{min}} \qquad 4.8$$

After we apply this metric, we found the following results:

Table No (4.4): Shows C index results

	C index
k-means algorithm	0.05123
Genetic clustering algorithm	0.9534

E. Dunn Index:

It is an evaluating metric for clustering algorithms. It is considered one of the internal measures where its results are based on the data itself, like all other indicators that aim to determine the interconnectedness of elements belonging to the same cluster so that the variance is in its lowest state. The clusters are separated from each other sufficiently, as compared to the within-cluster variance.(Gupta & Panda, 2019)

The higher the Dunn index value, the better is the clustering. The number of clusters that maximize the Dunn index is taken as the optimal number of clusters k.

$$Dunn\ index\ (U) = \min_{1\le i\le c}\{\min_{1\le j\le c\ j\neq i}\{\frac{\delta(X_i, X_j)}{\max_{1\le k\le c}\{\Delta(X_k)\}}\}\} \qquad 4.9$$

Where:

$$\delta(X_i, X_j)\ is\ the\ intercluster\ distance\ (the\ distance\ between\ cluster\ X_i\ X_j)$$

$\Delta(X_k) is\ the\ intracluster\ distance$ (within the cluster X_k).(Brock et al., 2008)(Desgraupes, 2018)

And hereafter we apply the Dunn- index we found the following results:

Table No (4.5): Shows Dunn -index results

	Dunn index
K -means algorithm	0.053834
Genetic clustering algorithm	0.00056497

As results show, the performance of the k- means algorithm is better than a genetic algorithm.

4.1.2 External evaluation

In external evaluation, we evaluate the model results according to the data which not used for clustering, like the class labels and external benchmark. These types of evaluation methods measure how close the clustering is to the predetermined benchmark classes. However, it has recently been discussed whether this is adequate for actual data or only on synthetic data sets with factual ground truth. Since classes can contain internal structure, the attributes present may not allow separation of clusters, or the classes may contain anomalies.

As an internal evaluation, there are several metrics to evaluate the clustering models externally, and here we will talk about the Jaccard index and rand index as external evaluation metrics.(Souto et al., 2012)

The basic idea is, any two objects or any two data points which belongs to the same cluster must have the same class label, and that we give us matrix call congestion matrix, which represented as:

- Ideal class similarity matrix.
- Ideal cluster similarity matrix.(H. Liu et al., n.d.)

Then we join these two matrices, and we get the conjunction table based upon it, and then we calculate the measure.

According to the above, if the data points belong to the same class, we assign it with 1 in the matrix. Otherwise, we assign it with 0 logically, and we can define it as:

- TP (True Positive) if the points in the same class same cluster.
- TN (True Negative) if the points in the same class different cluster.
- FP (False Positive) if the point in the different class same cluster.
- FN (False Negative) if the point in a different class different cluster.

A. Jaccard index:

The Jaccard index (which is called the similarity index) compares the elements of two clusters to find out which of the elements are common between the two clusters and which are distinctive. The range of this metric ranges between 0% to 100%, as the higher the percentage, the greater the similarity between the two clusters. Although It is effortless to clarify this, it may lead to erroneous results if the samples are too small or the observed data set is missing.(Tang et al., 2020)

We can assign the Jaccard equation as:

$$J(A\&B) = \frac{TP}{TP + FP + FN} \tag{4.10}$$

Jaccard index is used to quantify the similarity between two datasets, so it takes on a value between 0 and 1. An index of 1 means that the two datasets are identical, and an index of 0 indicates that the datasets have no common elements.

About the experiment result, we fund the following results:

Table No (4.6): Shows Jaccard index

	Jaccard index
k-means algorithm	0.99568
Genetic clustering algorithm	0.63407

B. Rand Index:

Rand index is a measure of similarity between two clusters to calculate how they identical to each other. Simply, we can define the rand index that is adjusted for chance grouping the elements, which call the adjusted rand index.(Robert et al., 2021)

Mathematically, the Rand index is related to the clustering accuracy and can apply even if the class label is missing or not used.

Suppose $a + b$ is the number of agreements between X and Y and $c + d$ is the disagreements between X and Y, the Rand index represents the frequency of occurrence of agreements over the total pairs, or the probability that X and Y will agree on a randomly chosen pair : $\binom{n}{2}$ which is calculated as $n\,(n-1)/2$.(Osamor & Osamor, 2020)

Similarly, one can also view the Rand index as a measure of the percentage of correct decisions made by the algorithm. It can be computed using the following formula:

$$RI = \frac{TP + TN}{TP + FP + FN + TN} \qquad 4.11$$

The result which we found after applying this index is:

Table No (4.7): Shows Rans index results

	Rand – index
k-means algorithm	0.98677
Genetic clustering algorithm	0.6334

Here we can see how it close to the result found by the Jaccard index before, and that gives us a clear understanding of the excellence of the k-means algorithm in this field. It should not absent in topic cluster modeling in the future. At least the k means algorithm can give a big hand to suggest the web search process pattern to fit the proper machine learning model.

Generally, all used evaluation metrics above prove that k means can present better performance than the genetic clustering algorithm in this field, which indicates that the future must be focused on enhancing the web ranking results.

Chapter Five: Conclusion & Future Work

5.1 Conclusions:

According to the previous we got the following set of conclusions.

- Data can tell more than we imagine: the most significant advantage of analyzing the row data is to define the relationship between the attributes to create and build the model. So the first thing we get from the analysis is the nature of the model and the proper algorithm for the model. According to this research dataset, it shows the linear relationship between most attributes. The first problem we faced with this is the multi co-linearity which makes difficulties in model computing, co-linearity in this research cause outlier results in clustering. According to this research, we found one of the best ways to remove the co-linearity by return the data to its principles via PCA(principal components analysis), which effective method to rearrange the dataset for error computations.
- Root mean square error (RMSE) computation: computing RMSE helps us much more to clear dataset, solve outlier data problem, and the misclassification errors, which is the most problem of clustering models.
- Dimensional reduction process: this process is one of the success factors in the research because it reduces the attributes according to the principal component analysis, which is familiar with clustering algorithms, but what we did actually not just feature extraction in this research we aggregate between feature extraction and feature selection as two stages done in preprocessing chapter.
- In model building, the dimensional reduction method enhanced the k-means results according to the graphs, showing the difference between using it.
- There is no effect in the genetic algorithm results when we use the dimensional reduction method because of its nature.
- The evaluation method is more effective in clustering evaluation the metric that can calculate both ways (external and internal distance), which can present the perfect results about the connectivity inside each cluster and how far each cluster form the others.

By the end of the research, the results of the k-means algorithm are better than the genetic clustering algorithm in model building and the evaluation metrics, the most effective evaluation metric in the search process is time. At the same time, the genetic clustering algorithm consuming too much to do the same job of k-means algorithm, here we can say the time consuming is incomparable, and that metric is enough to outweigh the excellence of the k-means algorithm as a clustering topics algorithm in the search engine to enhance the ranking methods.

5.1 Future work:

- Adding some attributes to customize the search process: Adding location and language are among the essential things that help in the indexing process, especially and determining the rank for each site based on these characteristics, knowing that, in light of the progress achieved, the positioning has become one of the primary data on which the search is based on the Internet and through which search services can be provided. Based on the user's location, as an example, instead of asking the user about the best places to eat, the search engine can provide a suggestion of what type and where to eat your next meal, language can be an excellent addition to customize the culture.
- To enrich this topic, it is preferable to compare all clustering algorithms with different databases to present development proposals for individual clustering algorithms and the effect of each algorithm on improving the Rank results.
- The dimensional reduction method used in this research needs more implementation in other clustering algorithms to evaluate it to measure the effectivity and the efficiency with other models.
- We found many difficulties to choose the proper evaluation metrics to compare between the results, which lead us to find out the evaluation which applicable in both algorithms in the study, so to avoid the problem in the future, we have developed that metrics as a part of the research, or make it individual research to help in this field.

References:

Agustín-Blas, L. E., Salcedo-Sanz, S., Jiménez-Fernández, S., Carro-Calvo, L., Del Ser, J., & Portilla-Figueras, J. A. (2012). A new grouping genetic algorithm for clustering problems. *Expert Systems with Applications*, *39*(10), 9695–9703. https://doi.org/10.1016/j.eswa.2012.02.149

Ahmad, I., & Amin, F. (2015). Towards feature subset selection in intrusion detection. In *2014 IEEE 7th Joint International Information Technology and Artificial Intelligence Conference, ITAIC 2014*. https://doi.org/10.1109/ITAIC.2014.7065007

Akmaliyah, M. (2013). SEARCH ENGINE TUNING WITH GENETIC ALGORITHMS. *Journal of Chemical Information and Modeling*, *53*(9), 1689–1699.

Al-Subaihin, A. A., & Sarro, F. (2020). Exploring the Use of Genetic Algorithm Clustering for Mobile App Categorisation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12420 LNCS*, 181–187. https://doi.org/10.1007/978-3-030-59762-7_13

Anusha, M., & Sathiaseelan, J. G. R. (2014). An enhanced K-means genetic algorithms for optimal clustering. *2014 IEEE International Conference on Computational Intelligence and Computing Research*, 1–5. https://doi.org/10.1109/ICCIC.2014.7238422

Ashok, K. D., Usha, T. A., & Sivaranjani, C. (2016). *A Comparative Study on K-Means And Genetic Algorithm For Data Clustering*. *12*(11), 1–9.

Aurélien Géron. (2019). Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems. In *O'Reilly Media*. https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/

Bachem, O., Lucic, M., Hassani, S. H., & Krause, A. (2016). Fast and provably good seedings for k-means. *Advances in Neural Information Processing Systems*, *Nips*, 55–63.

Baeza-Yates, R., Hurtado, C., & Mendoza, M. (2004). Query clustering for boosting web page ranking. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *3034*, 164–175. https://doi.org/10.1007/978-3-540-24681-7_19

Balato, M., & Vitelli, M. (2014). A new control strategy for the optimization of Distributed MPPT in PV applications. *International Journal of Electrical Power & Energy Systems*, *62*, 763–773. https://doi.org/https://doi.org/10.1016/j.ijepes.2014.05.032

Bartholomew, D. J. (2010). Principal components analysis. *International Encyclopedia of Education*, 374–377. https://doi.org/10.1016/B978-0-08-044894-7.01358-0

Bezdek, J., Moshtaghi, M., Runkler, T., & Leckie, C. (2016). A Generalized C Index for (Internal) Fuzzy Cluster Validity. *IEEE Transactions on Fuzzy Systems*, *24*, 1. https://doi.org/10.1109/TFUZZ.2016.2540063

Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). ClValid: An R package for cluster validation. In *Journal of Statistical Software* (Vol. 25, Issue 4, pp. 1–22). https://doi.org/10.18637/jss.v025.i04

Burges, C. J. C. (2010). *MSR-TR-2010-82.pdf*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.180.634&rep=rep1&type=pdf%0Apapers3://publication/uuid/0B8C568E-291D-431D-8873-EF4B0A31E356

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chapelle, O. (2011). Yahoo! Learning to Rank Challenge Overview. *JMLR: Workshop and Conference Proceedings*, *14*, 1–24.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, *61*, 1–36. https://doi.org/10.18637/jss.v061.i06

Chawla, S. (2016). A novel approach of cluster based optimal ranking of clicked URLs using genetic algorithm for effective personalized web search. *Applied Soft*

Computing, *46*. https://doi.org/10.1016/j.asoc.2016.04.042

Chen, T., He, T., & Benesty, M. (2018). xgboost : eXtreme Gradient Boosting. *R Package Version 0.71-2*, 2–5.

Dash, R., & Dash, R. (2012). Comparative Analysis of K-Means And Genetic Algorithm Based Data Clustering. *International Journal of Advanced Computer and Mathematical Sciences*, *3*(2), 257–265.

Desgraupes, B. (2018). *clusterCrit: Clustering Indices. R package version 1.2.8.* (p. 10). https://cran.r-project.org/web/packages/clusterCrit/clusterCrit.pdf

Dhanachandra, N., Manglem, K., & Chanu, Y. J. (2015). Image Segmentation Using K-means Clustering Algorithm and Subtractive Clustering Algorithm. *Procedia Computer Science*, *54*, 764–771. https://doi.org/10.1016/j.procs.2015.06.090

Divyashree, G., & Rayar, G. (2015). *Comparison Between K-Means and Genetic Algorithm in Text Document Clustering*. *3*(14), 1–10.

Emmons, S., Kobourov, S., Gallant, M., & Börner, K. (2016). Analysis of Network Clustering Algorithms and Cluster Quality Metrics at Scale. *PLOS ONE*, *11*(7), e0159161. https://doi.org/10.1371/journal.pone.0159161

Gupta, T., & Panda, S. P. (2019). Clustering Validation of CLARA and K-Means Using Silhouette & DUNN Measures on Iris Dataset. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 10–13. https://doi.org/10.1109/COMITCon.2019.8862199

Jachner, S., Van Den Boogaart, K. G., & Petzoldt, T. (2007). Statistical methods for the qualitative assessment of dynamic models with time delay (R package qualV). *Journal of Statistical Software*, *22*(8), 1–30. https://doi.org/10.18637/jss.v022.i08

Joachims, T. (2002). Optimizing search engines using clickthrough data. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 133–142. https://doi.org/10.1145/775047.775067

Kassambara, A. (2017). *Practical Guide to Cluster Analysis in R, Unsupervised Machine*

Learning (A. Kassambara (ed.); first edit). published by STHDA. https://play.google.com/books/reader?id=plEyDwAAQBAJ&hl=ar&printsec=frontcover&pg=GBS.PR6

Lahari, K., Murty, M. R., & Satapathy, S. C. (2015). Partition Based Clustering Using Genetic Algorithm and Teaching Learning Based Optimization: Performance Analysis. In S. C. Satapathy, A. Govardhan, K. S. Raju, & J. K. Mandal (Eds.), *Emerging ICT for Bridging the Future - Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2* (pp. 191–200). Springer International Publishing.

Ledford, J. L. (2015). *Search Engine Optimization Bible*. Wiley. https://books.google.com/books?id=2Gz-CAAAQBAJ

Lengyel, A., & Botta-Dukát, Z. (2018). Silhouette width using generalized mean – a flexible method for assessing clustering efficiency. In *bioRxiv*. https://doi.org/10.1101/434100

Liu, H., Wang, J., & Jing, L. (n.d.). *Cluster-wise Hierarchical Generative Model for Deep Amortized Clustering*. 15109–15118.

Liu, T. Y. (2009). Learning to rank for Information Retrieval. *Foundations and Trends in Information Retrieval*, *3*(3), 225–231. https://doi.org/10.1561/1500000016

Lu, Y., Cohen, I., Zhou, X., & Tian, Q. (2007). Feature selection using principal feature analysis. In *Proceedings of the 15th international conference on Multimedia, MUL-TIMEDIA '07*. https://doi.org/10.1145/1291233.1291297

Macdonald, C., Santos, R. L. T., & Ounis, I. (2013). The whens and hows of learning to rank for web search. In *Information Retrieval* (Vol. 16, Issue 5). https://doi.org/10.1007/s10791-012-9209-9

Madsen, H., Pinson, P., Kariniotakis, G., Nielsen, H. A., & Nielsen, T. S. (2005). Standardizing the performance evaluation of short-term wind power prediction models. *Wind Engineering*, *29*(6), 475–489. https://doi.org/10.1260/030952405776234599

Masaeli, M., Yan, Y., Cui, Y., Fung, G., & Dy, J. G. (2010). Convex principal feature selection. *Proceedings of the 10th SIAM International Conference on Data Mining, SDM 2010*, *2*, 619–628. https://doi.org/10.1137/1.9781611972801.54

Matt Young. (1990). *The technical writer's handbook*. https://onlinelibrary.wiley.com/doi/epdf/10.1002/mrd.1080250118#citedby-section

Mehrotra, S. (2015). *Comparative Analysis of K-Means with other Clustering Algorithms to Improve Search Result*. 309–313.

Microsoft Research Asia. (2009). *LETOR: A Benchmark Collection for Learning to Rank for Information Retrieval (an incomplete draft)*. 1–19. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/08/letor3.pdf

Minka, T., & Robertson, S. (2008). Selection bias in the LETOR datasets. *Learning to Rank for Information Retrieval*, *April*, 48. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.155.4576&amp;rep=rep1&amp;type=pdf#page=55

Mohan, A. (2010). *An Empirical Analysis on Point-wise Machine Learning Techniques using Regression Trees for Web-search Ranking*. *January*.

Moreira, C., Calado, P., & Martins, B. (2015). Learning to rank academic experts in the DBLP dataset. *Expert Systems*, *32*(4), 477–493. https://doi.org/10.1111/exsy.12062

Osamor, I. P., & Osamor, V. C. (2020). OsamorSoft: clustering index for comparison and quality validation in high throughput dataset. *Journal of Big Data*, *7*(1), 48. https://doi.org/10.1186/s40537-020-00325-6

Padmaja, S., & Sheshasaayee, A. (2016). Clustering of user behaviour based on web log data using improved K-means clustering algorithm. *International Journal of Engineering and Technology*, *8*(1), 305–310.

Qin, T., & Liu, T.-Y. (2013). *Introducing LETOR 4.0 Datasets*. http://arxiv.org/abs/1306.2597

Rajaraman, A., & Ullman, J. D. (2012). Clustering. *Mining of Massive Datasets*, 213–251.

https://doi.org/10.1017/cbo9781139058452.008

Robert, V., Vasseur, Y., & Brault, V. (2021). Comparing High-Dimensional Partitions with the Co-clustering Adjusted Rand Index. *Journal of Classification*, *38*(1), 158–186. https://doi.org/10.1007/s00357-020-09379-w

Rudin, C., & Schapire, R. E. (2009). Margin-based ranking and an equivalence between AdaBoost and RankBoost. *Journal of Machine Learning Research, 10*, 2193–2232.

Singh, N. (2013). *Query Recommendation employing Query Logs in Search Optimization*. *1921*, 1917–1921.

Souto, M. C. P. de, Coelho, A. L. V, Faceli, K., Sakata, T. C., Bonadia, V., & Costa, I. G. (2012). A Comparison of External Clustering Evaluation Indices in the Context of Imbalanced Data Sets. *2012 Brazilian Symposium on Neural Networks*, 49–54. https://doi.org/10.1109/SBRN.2012.25

Suhara, Y., Suzuki, J., & Kataoka, R. (2013). Robust online learning to rank via selective pairwise approach based on evaluation measures. *Transactions of the Japanese Society for Artificial Intelligence*, *28*(1), 22–33. https://doi.org/10.1527/tjsai.28.22

Sultana, R. (2016). *Application of Genetic Algorithm in Multi-objective Optimization of an Indeterminate Structure with Discontinuous Space for Support Locations*. 821. http://scholarworks.gvsu.edu/theseshttp://scholarworks.gvsu.edu/theses/821

Tang, M., Kaymaz, Y., Logeman, B., Eichhorn, S., Liang, Z. S., Dulac, C., & Sackton, T. B. (2020). Evaluating single-cell cluster stability using the Jaccard similarity index. *BioRxiv*, 2020.05.26.116640. https://doi.org/10.1101/2020.05.26.116640

Torlay, L., Perrone-Bertolotti, M., Thomas, E., & Baciu, M. (2017). Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. *Brain Informatics*, *4*(3), 159–169. https://doi.org/10.1007/s40708-017-0065-7

Velchamy, I., Subramanian, R., & Vasudevan, V. (2011). Cluster analysis research design model, problems, issues, challenges, trends and tools. *International Journal on Computer Science and Engineering*, *3*, 3064–3070.

Wang, X., & Xu, Y. (2019). An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. In *IOP Conference Series: Materials Science and Engineering* (Vol. 569, Issue 5). https://doi.org/10.1088/1757-899X/569/5/052024

Xu, J., & Li, H. (2007). AdaRank: A boosting algorithm for information retrieval. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'07*, *49*, 391–398. https://doi.org/10.1145/1277741.1277809

Yadav, S., & Jyotiyadavcs, M. (2016). *A Novel Technique to Recommendation of Query in Search Engine*. 35–42. https://doi.org/10.141079/IJITKM.2016.914

Zhang, M., Kuang, D., Hua, G., Liu, Y., & Ma, S. (2009). Is learning to rank effective for Web search ? *SIGIR Workshop*.

Zhang, Z., Zhao, J., & Yan, X. (2018). A web page clustering method based on formal concept analysis. *Information (Switzerland)*, *9*(9). https://doi.org/10.3390/info9090228

الملخص

في الوقت الحاضر، تعتبر أدوات البحث ضرورية للبحث عن المعلومات على الويب. ومع ذلك ومن وقت لآخر، يتغير مقدار المعلومات المتاحة لنا على الويب ويزداد باستمرار، وقد تم دفع تقنيات البحث هذه باستمرار إلى الحد الأقصى المطلق. لذلك، ظهرت العديد من التقنيات الجديدة لتعزيز عملية البحث، وأحدها هو تحليل سجلات الاستعلام. تحتوي سجلات الاستعلام على آلية لتتبع محفوظات الاستعلامات المرسلة إلى محركات البحث والصفحات المفضلة بعد البحث، من بين بيانات أخرى.

يعد تصنيف صفحات الويب الخطوة الأولى في ترتيب صفحات الويب (أو يمكننا تسميتها بالفهرسة)، ومن أكثر الطرق لتحقيق عملية الفهرسة تجميع تلك الصفحات في مجموعات وفقًا للتشابه، فكلما كان التصنيف الخاطئ أقل كلما كانت النتيجة مثالية. ليس بعيدًا، عملية التجميع باختصار، هي مجموعة من الخوارزميات التي تقسم البيانات إلى مجموعات مرتبطة ببعضها البعض. لهذه الوظيفة، اخترنا مجموعة بيانات (Microsoft Learn to Rank) لتحقيق التحليل وبناء النموذج عليها، وقد تم تصميم مجموعة البيانات هذه خصيصًا للباحثين في هذا المجال، وتحتوي على معلومات هائلة ومختلفة حول عملية التصنيف. بسبب كمية المعلومات، اخترنا عشوائياً ١٦٠١٥ ملاحظة فقط من MSLR-WEB30K_2 _ fold 1 في هذه الدراسة، وفقًا لقدرة أجهزتنا وخوارزميات التحليل، حيث أنه لا تستطيع بعض الخوارزميات المستخدمة في التحليل (تحديد العدد الأمثل للمجاميع) التعامل مع الكمية الهائلة من الملاحظات.

في هذه الأطروحة، سأستخدم التحليل التجميعي لتحسين ترتيب بحث الويب باستخدام التحليل المقارن بين الوسائل (k- means) والخوارزميات الجينية (Genetic algorithm) كأسلوب مستخدم لهذا الهدف.

تتضمن هذه العملية تحليل المجموعات لتقليل الميزات باستخدام تحليل المكون الرئيسي (PCA) مع احتساب خطأ المربع الرئيسي للجذر (RMSE) كتقنية لتقليل الميزات لحساب معدل الخطأ ودقة نتيجة النموذج للحصول على أفضل عدد من المجموعات، وقد تم تحقيق هذه العملية باستخدام نهج التحقق المتبادل باستخدام خوارزمية تعزيز التدرج الشديد (eXGBoost)كنموذج تدريب لتقدير مجموع الأخطاء أثناء عملية التدريب.

بعد ذلك، تم استخدام طرقًا مختلفة لتحديد العدد الأمثل للمجموعات لضمان توزيع عالي الجودة للبيانات في المجموعات. واختبار نتيجة بناء النماذج وفقا لمعايير تقييم نماذج التجميع المتعارف عليها.

الجمهورية اليمنية

وزارة التعليم العالي والبحث العلمي

جامعة الريان

كلية الدراسات العليا

دراسة مقارنة بين خوارزمية K-Means و خوارزمية Genetic كأسلوب لتصنيف مواقع الأنترنت لعمليات البحث

رسالة مقدمة إلى جامعة الريان

لاستكمال متطلبات نيل درجة الماجستير لتخصص تقنية معلومات

إعداد الباحث

محمد عبدربه محمد كعدان

إشراف

المشرف الأول

د. خالد قايد شعفل

المشرف الثاني

د. محمد عبدالله بامطرف

٢٠٢١/١٤٤٢

جامعة الريان
AL-RAYAN UNIVERSITY

الجمهورية اليمنية

وزارة التعليم العالي والبحث العلمي

جامعة الريان

كلية الدراسات العليا

دراسة مقارنة بين خوارزمية K-Means و خوارزمية Genetic كأسلوب لتصنيف مواقع الأنترنت لعمليات البحث

رسالة مقدمة إلى جامعة الريان

لاستكمال متطلبات نيل درجة الماجستير لتخصص تقنية معلومات

إعداد الباحث

محمد عبدربه محمد كعدان

إشراف

المشرف الأول	المشرف الثاني
د. خالد قايد شعفل	د. محمد عبدالله بامطرف

1442هـ /2021 م