Project Report
on

USER REVIEW CATEGORISATION USING MACHINE LEARNING

Submitted in the partial fulfilment of the requirement for the award of degree of
Bachelor of Science In Computer Science

Submitted by
Purnendu Manna (423-1111-0393-19)
Aniket Bera (423-1111-0397-19)
Sandipan Dey (423-1111-0388-19)
Tiyasa Ghosh (423-1214-0385-19)

Under The Guidance Of
Mr. Rudra Prasad Chatterjee
S.A.C.T(Dept of Computer Science, PJC)

Prabhu Jagatbandhu College
Jhorehat, Andul-Mouri, Jhorehat, Andul
Howrah, West Bengal
711302

University of Calcutta
College St, College Square
Kolkata, West Bengal

CERTIFICATE

This is to certify that the project entitled **"User Review Categorisation Using Machine Learning"** has been developed by B.SC students of **Prabhu Jagatbandhu College** towards partial fulfillment of the requirements for the award of the degree of Bachelor of Science is a genuine record of the work carried out by **Purnendu Manna** (Roll no : 193423-21-0012 Registration no.: 423-1111-0393-19) **Aniket Bera** (Roll no : 193423-21-0015 Registration no.: 423-1111-0397-19) **Sandipan Dey** (Roll no : 193423-21-0007 Registration no.: 423-1111-0388-19) **Tiyasa Ghosh**(Roll no : 193423-11-0005 Registration no.: 423-1214-0385-19) under the Supervision and Guidance of **Mr. Rudra Prasad Chatterjee**, Department of Computer Science.

Date : 02.07.2022

……..…………………………………

Signature of Project Supervisor

……………………………………….

Signature of Head of the
Department

……………………………………….

Signature of Examiner

Acknowledgement

I thank the almighty for giving us the courage and perseverance in completing the this project, **“User Review Categorisation Using Machine Learning”**. This project itself is acknowledgements for all those people who have given us their heartfelt co-operation in making this project a grand success. I extend our sincere thanks to **Mr. Rudra Prasad Chatterjee** for providing sufficient infrastructure and good environment to complete our course and for her support and valuable suggestions regarding project work, and also for providing valuable guidance at every stage of this project work. With extreme jubilance and deepest gratitude, I would like to thank Head of the Computer Science Department, **Dr. Sumana Bandyopadhaya** for her constant encouragement. I am profoundly grateful towards the unmatched services rendered by him. Last but not least , we would like to express our deep sense of gratitude and earnest thanks giving to our dear parents for their moral support and heartfelt co-operation in doing the main project.

………………………………………………………….

Purnendu Manna (193423-21-0012)

………………………………………………………….

Aniket Bera (193423-21-0015)

………………………………………………………….

Sandipan Dey (193423-21-0007)

………………………………………………………….

Tiyasa Ghosh (193423-11-0005)

Index

Contents: ***Page No.:***

Abstract :-

Electronic commerce is an explosion in our modern business strategy. Nowadays people are purchasing and selling products in an easier and effective way through the network connection. Although it has become an impressive technique in trade, it has some drawbacks. All the images and advertisements cannot always give the right assurance. Sometimes the products aren't hopeful for the lack of touch and feel. This project will help the customers to overcome such kinds of drawbacks. The Analytical Hierarchy Process (AHP)has been followed. The data about the quality of a product which is adequate or inadequate according to some buyers are collected from their reviews/feedbacks and has been clustered in three categories or groups(Good, Average, Worst).Over 200 reviews have been classified into three categories depending upon positive and negative reviews. The k-means and EM clustering is used here as it helps to find out the best congestion. This will recommend to the customers what product is actually good and what is not. Customers can get an idea and assurance about the product before purchasing.

This project work acts as a deposition to the users who want to know the details and specifications of the Television before buying.

Introduction :-

With the elevation demand of necessary products, the online product marketing is dilating at an aggressive pace. With such a roaring in the digital industry, there needs to be some integrated review of the brand and the quality of the online products. With the evaluation of technology the demand for television is increasing at an alarming rate. There are many branded televisions present in the market. Some of them are dominant and Nowadays Amazon, Flipkart, Meesho, Snapdeal, eBay etc platforms are providing an opportunity for businesses to their customers to buy products easily.

For huge products manufacturing day by day of different brands, the relevant reviews for the consumers is needed. Customers mostly want user generated content over online to a considerable extent for decision making. Before buying the product, sincere customers always want to look up its reviews online and discuss it before making a decision.

So with the rapid changing or growing of digital marketing there needs some convenient methods to control the vast data generated by users.

For the type of “big data,” it requires the assumption of specific approaches for its right analysis and to automate this. So various sentiment analysis techniques are mostly used. Mining means minimizing large amounts of data processes, which are developing. Day by day it is becoming very worldly due to its convenient process of extracting out valid information from large datasets and managing information. It can be browsed from different angles. There are many types of mining modeling processes (Predictive, Descriptive).

Further traditional statistical methods will perform incompetently in situations where the volume of data is very large, or in situations of high dimension, which are typical in “vast data” contexts. Other problems may be encountered when the data are not in numerical format, but rather image, video, or audio files.

In this project, the data of Television Reviews have been extracted from Flip-cart. The number of reviews associated with Television or its brand is increasing at an alarming rate. The reviews have been classified using machine learning clustering like K-means clustering, EM-clustering etc . The reviews on the basis of positive feedback and negative feedback of the customers provide the result in better judgment. This will help the further customers to check the positive and negative feedback constructively and to get a proper decision as per their requirements before purchasing.

Literature Survey:-

Determining a consensus opinion on a product sold online is no longer simple as before. To address this issue, the use of Sentiment Analysis is getting more popular and more efficient using more and more powerful algorithms and techniques to give efficient and accurate recommendations. Clustering and Classification is a powerful way to extract features from raw data and extract useful information that can be useful in many ways to ease Sentiment Analysis. Sentiment Analysis is used in many varieties of wide fields such as marketing, business, industry, education, e-commerce and many more. From early 60’s, many related works had already been done using Sentiment Analysis and here are some of the following existing works,

Paper [1] provides a survey on lexicon-based approaches, together with cross domain and cross-lingual methods and some evaluation metrics. Research results show that machine learning methods, such as SVM and Naïve Bayes have highest accuracy and can be regarded as the baseline learning methods, while lexicon-based methods are very effective in some cases, which require few efforts in human-labeled documents. Researchers used various techniques in [2] like data consolidation, feature engineering, classifier selection, metrics for validation and finally segmentation of products to make an automated recommendation system from unstructured data and to provide users optimal recommendation. The problem moved from traditional classification to one of machine translation and showed that state-of-the-art machine translation (MT) models surpass previous classification approaches in [3] for categorizing products in two large real-world e-commerce datasets. In Paper [4] the mined datasets has been analyzed i.e related to top online grocery platforms and utilize different machine learning algorithms to classify multilevel product categories that are likely to contain very similar products. Paper [5] proposed a machine-learning approach to efficiently and effectively analyze customer needs of product ecosystems based on the user generated online product reviews and used analytical Kano model to categorize customer needs related to each topic quantitatively. A sentiment based recommendation and rating prediction model has been suggested in [6] which is to predict the rating from user reviews, fusing sentiment similarity, interpersonal sentiment influence, and item reputation similarity into a unified matrix factorization framework to achieve the rating prediction task. Paper [7] proposed a Long Short-Term Memory (LSTM) Model that digs into the polarity intensity such as whether the review is just positive or very positive and same for negative reviews and as a result it more efficiently and accurately gives detailed results on reviews. A sentiment analysis model has been proposed in [8] based on user's review on online product based websites using various Classification algorithms that predict recommendation for the product to user.

Paper [9] has been depicted that using multiple classifications, filtering products, and the notion of strong accuracies, helped the accuracy significantly and made sense in the context of categorizing products.In [10] a model has been discussed that uses product classification and product data consistency for web shops with application of Machine Learning (ML) classifiers and users' content as a moderating factor so that products can be filtered, mined, classified using ML techniques.Paper [11] presented a study that is to assist businesses in monitoring product sales and in better understanding their clients by gaining insight into their intents when purchasing a certain product and implementing focused marketing tactics to expand their customer base and profitability using various data visualization techniques and conducting a detailed sentiment analysis on product evaluations to better understand consumer interactions with e-commerce sites.In [12] a more accurate Machine Learning (ML) model compared to previous models using Sentiment Analysis that is used to identify and aggregate the sentiment of opinions expressed by the users, classifying the polarity of text in document or sentence in terms of the opinion expressed as positive, negative and neutral. Paper [13] mentioned a fundamental problem of sentiment analysis, Sentiment Polarity Categorization and tackles the problem

Methodology:-

In this paper we are going to analyze the reviews in e-commerce website, amazon.com about a product, ‘LG 80 cm (32 inches) HD Ready Smart LED TV’ and will use Sentiment Analysis technique to extract the information from the data and gain a meaningful outcome that can be used to recommend users to go for it or not. Figure 1 is a flowchart that depicts the process flow that is proposed in this paper to achieve the goal.

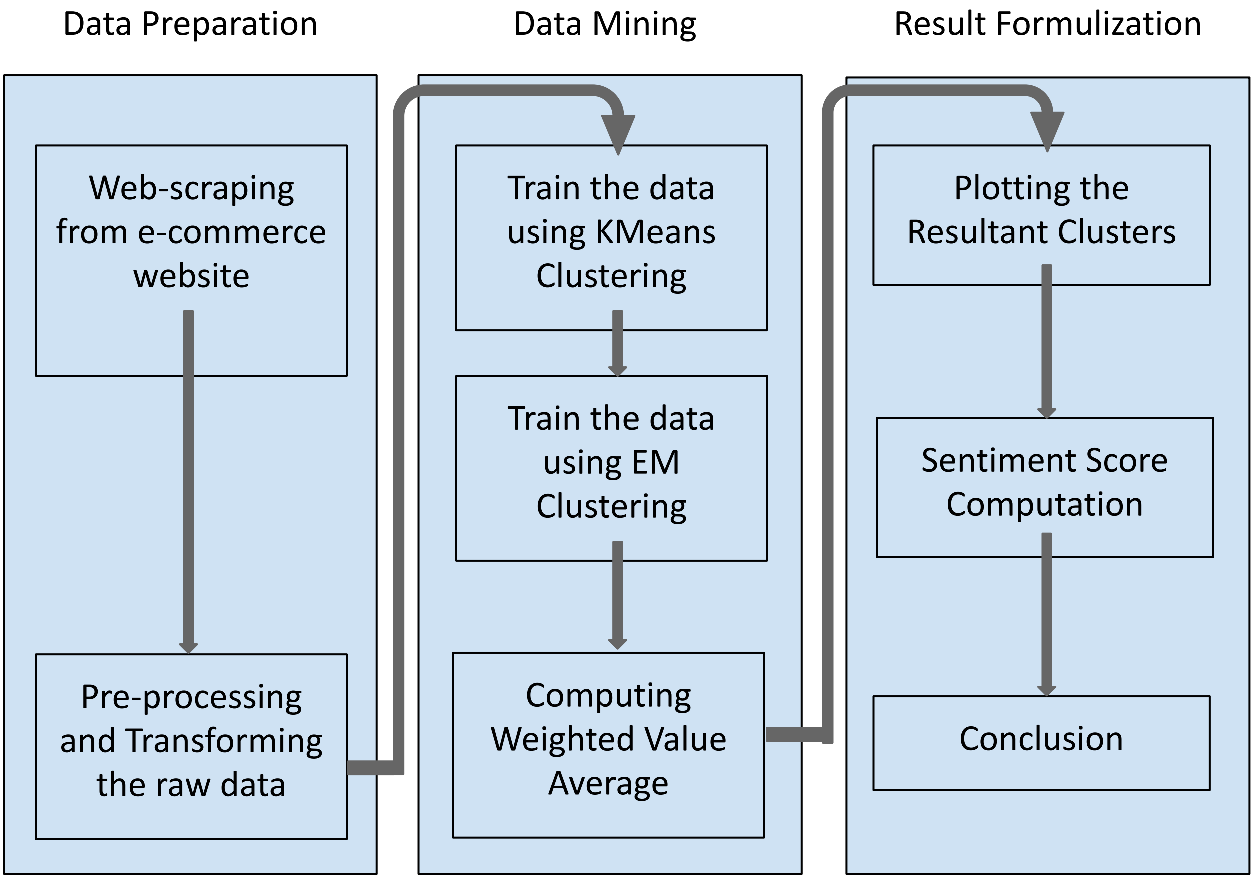
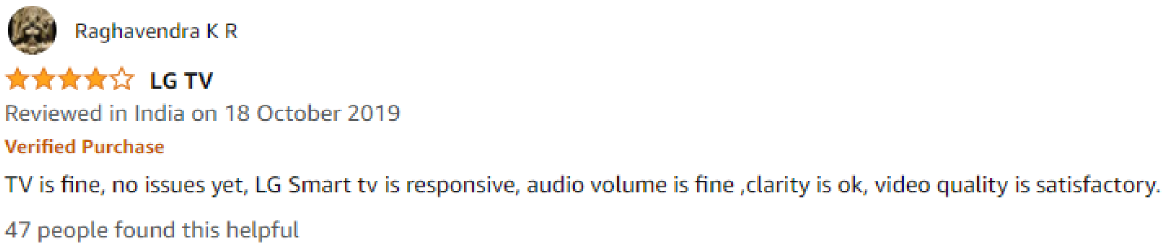

Data Preparation
Data Mining
Result Formulization
Web-scraping from e-commerce website
Pre-processing and Transforming the raw data
Train the data using KMeans Clustering
Train the data using EM Clustering
Computing Weighted Value Average
Plotting the Resultant Clusters
Sentiment Score Computation
Conclusion

Figure 1

Phase 1: Data Preparation-

Web-scraping from e-Commerce Website:

Data used in this paper is a set of reviews collected from amazon.com of ‘LG 80 cm (32 inches) HD Ready Smart LED TV’. Those online reviews were posted by over 1k users till March 2022 and we collected a set of 200 reviews from the review section of the product. Each review includes the following information : 1) reviewer ID; 2) product ID; 3) rating; 4) time of the review; 5) helpfulness; 6) review text.



Raghavendra K R

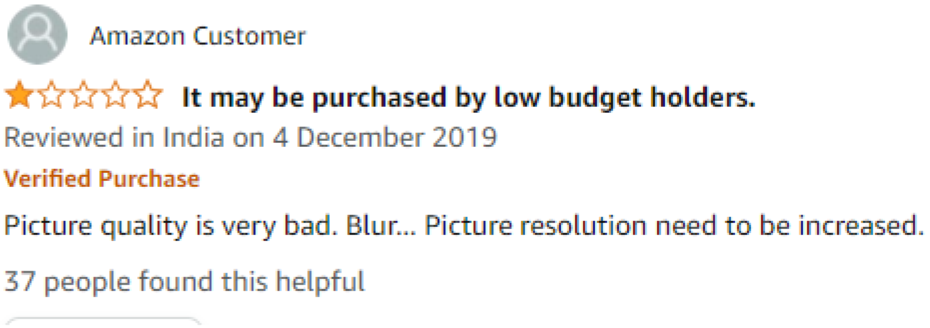
★★★★☆ **LG TV**

Reviewed in India on 18 October 2019

Verified Purchase

TV is fine, no issues yet, LG Smart tv is responsive, audio volume is fine ,clarity is ok, video quality is satisfactory.

47 people found this helpful



Amazon Customer

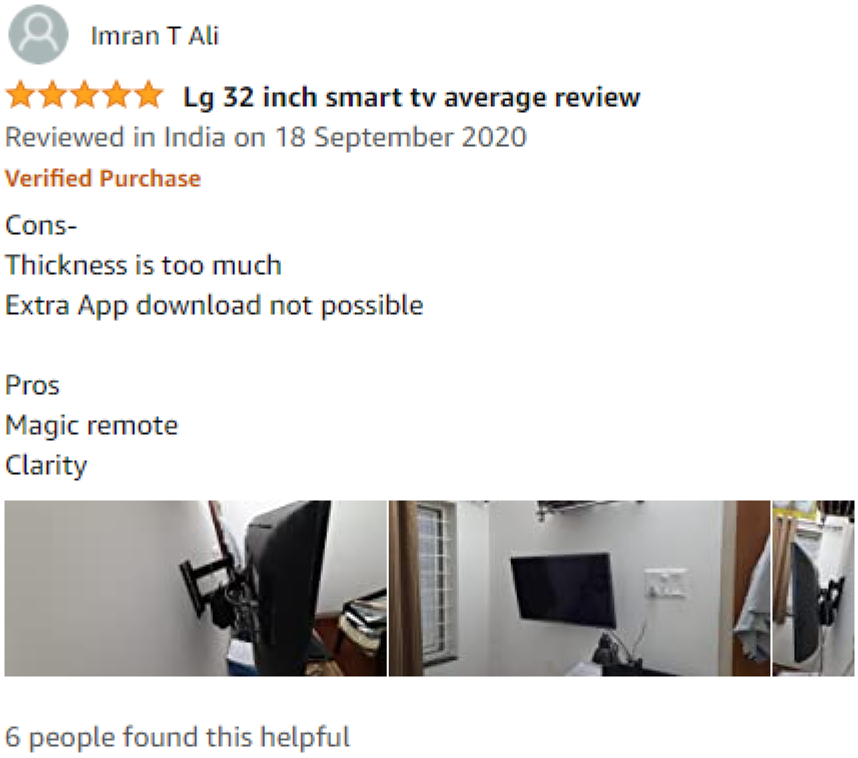
★☆☆☆☆ **It may be purchased by low budget holders.**

Reviewed in India on 4 December 2019

Verified Purchase

Picture quality is very bad. Blur... Picture resolution need to be increased.

37 people found this helpful



Imran T Ali

★★★★★ **Lg 32 inch smart tv average review**

Reviewed in India on 18 September 2020

Verified Purchase

Cons-
Thickness is too much
Extra App download not possible

Pros
Magic remote
Clarity

6 people found this helpful

Nowadays there are many judgemental criteria for televisions. We will use these criteria as categorizing factors. These criterions will help to extract the main motives and actual information from reviews that would lead to a good sentiment polarity categorization result. There are many judgemental criteria but we will use these main four criterions.

A. Picture Quality

Picture quality of a Television is one of the most important criteria. In the competitive market of TVs there are more TVs in its price range and every TV has its own signature picture quality. And TV is now used just not for content consumption but for luxurious crystal clear picture experience. If the other aspects are good like smart features, look but picture quality is compromised, then consumers will prefer the one with good picture quality and less features in the same price range. So this is a vital criteria for selection of TV for users. And that’s why we categorized the user experience of picture quality with the TV in 3 categories- 1) Highly Satisfactory; 2) Moderate; 3) Unsatisfactory.

B. Sound

Sound is also an important judgemental attribute in the selection process of the TV. The sound of the TV is measured in watts where higher the watts, the louder will be the maximum volume of the TV. The latest TV these days are equipped with pre-defined sound modes for gaming, sports, movies which automatically tune the volume of the TV accordingly to make TV viewing better. So Sound can differentiate a lot in buyers mind because sound can enhance the immersive experience of the user. So sound can play an important role in classifying the clusters.We also categorized the user experience of sound quality with the TV in 3 categories - 1) Nice; 2) Medium; 3) Poor.

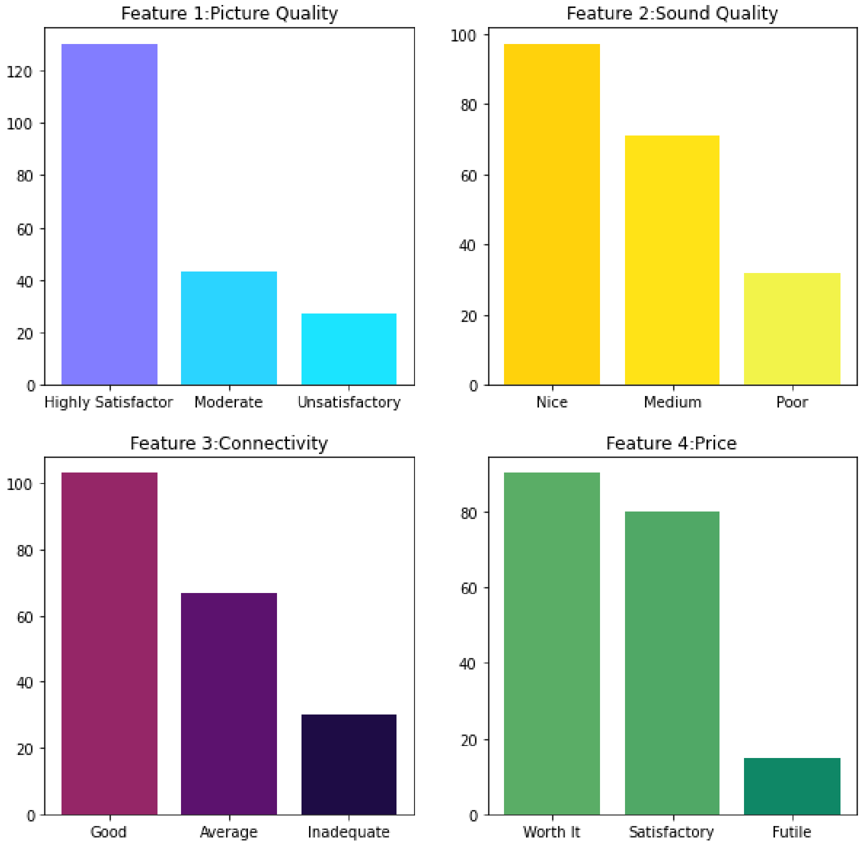
C. Connectivity

The connectivity option is also an attribute that the users desire most.The most common input connectivity options in modern-day TVs are USB port, HDMI ports, and Bluetooth Connectivity feature. Nowadays users also yearn for wifi connectivity, chrome cast, and android system for better experience in streaming online content from OTT platforms or video based platforms. And when it comes to output ports of the TV, users ache to have 3.5mm audio jack, HDMI arc, RCA audio output. So Connectivity is also a crucial feature of TV. The connectivity is also classified into 3 categories according to users - 1) Good; 2) Average; 3) Inadequate.

D. Price

The price of a TV is a very important criterion. In the ecommerce market many products compete with each other based on just this criteria, price of the product. Because most people like to filter TV depending on the price. This criterion reflects the economical aspect of a buyer. They yearn for a budget friendly feature loaded compact choice that can fulfill their demands without costing extra penny. So this criterion is also one of the superior criterions for TV selection. We can categorize this property from the user reviews in 3 categories - 1) Worth It; 2) Satisfactory; 3) Futile.

Now we analyzed each of the 200 reviews and categorize the judgemental criterias according to the user review and made an dataset called “dataset.csv”, having four columns named “Feature 1: Picture Quality”, “Feature 2: Sound Quality”, “Feature 3: Connectivity”, “Feature 4: Price”. This is our actual raw dataset which we got from web-scraping in amazon.com.

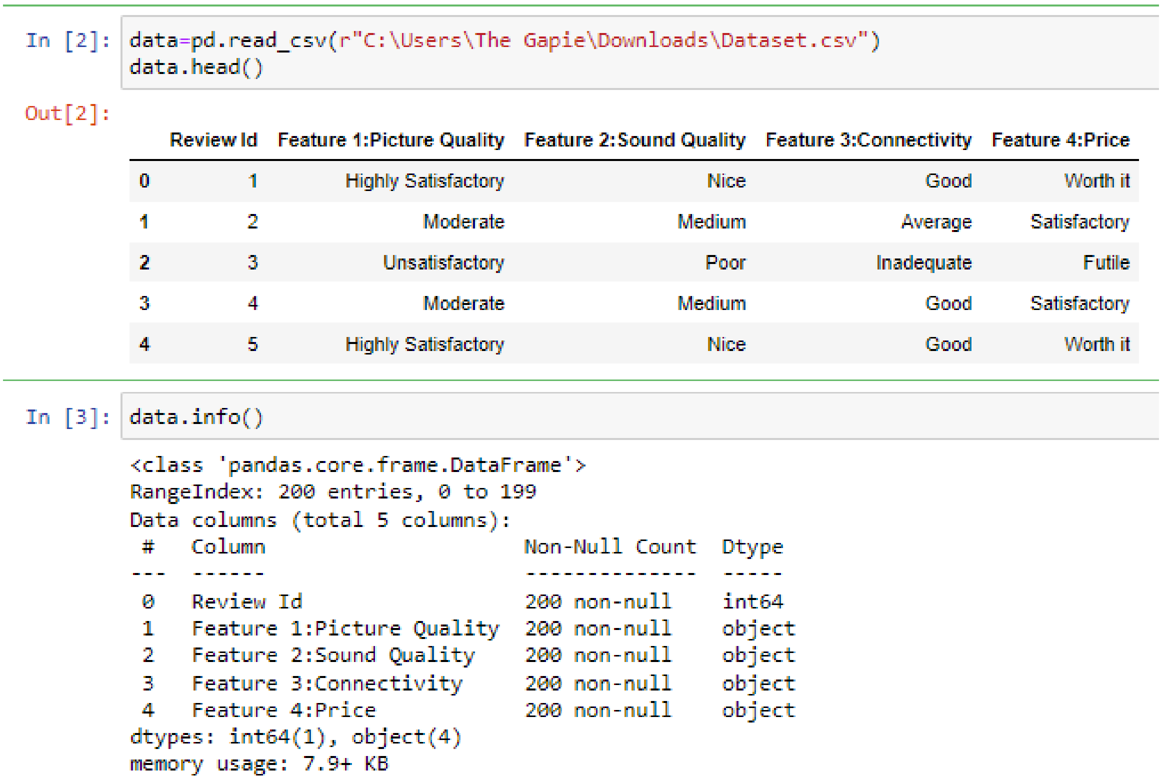


Pre-processing and Transforming the Raw Data:

The second step of phase 1 is to convert the raw data into computable data that can be used in the algorithm. Now the dataset of reviews obtained from amazon.com comes in a .csv format.

- **Unpacking and Preprocessing the data :**

A small python code has been implemented in order to read the dataset from dataset.csv and loaded into a dataframe that we are going to use in future.



```
In [2]: data=pd.read_csv(r"C:\Users\The Gapie\Downloads\Dataset.csv")
        data.head()
```

Out[2]:

	Review Id	Feature 1:Picture Quality	Feature 2:Sound Quality	Feature 3:Connectivity	Feature 4:Price
0	1	Highly Satisfactory	Nice	Good	Worth it
1	2	Moderate	Medium	Average	Satisfactory
2	3	Unsatisfactory	Poor	Inadequate	Futile
3	4	Moderate	Medium	Good	Satisfactory
4	5	Highly Satisfactory	Nice	Good	Worth it

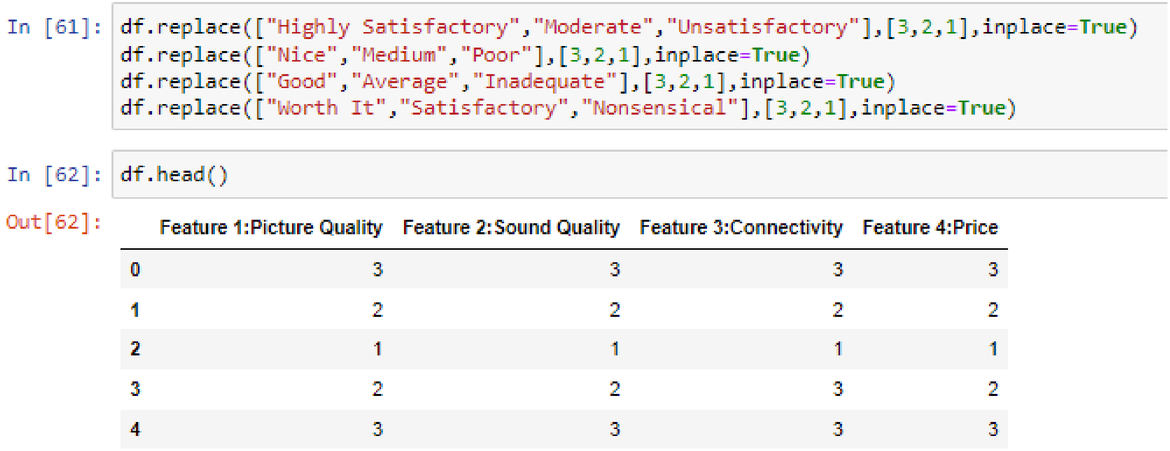
```
In [3]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Review Id                  200 non-null    int64
 1   Feature 1:Picture Quality  200 non-null    object
 2   Feature 2:Sound Quality    200 non-null    object
 3   Feature 3:Connectivity     200 non-null    object
 4   Feature 4:Price            200 non-null    object
dtypes: int64(1), object(4)
memory usage: 7.9+ KB
```

- **Data Transformation**

This is a vital phase of Data Preparation. The data that is collected from the ecommerce website, has been labeled with some values according to the sentiment weightage of the reviews, e.g. ‘Highly Satisfactory’, ‘Moderate’, ‘Unsatisfactory’ for ‘Picture Quality’ and so on. Now computers can’t calculate mathematical expressions or algorithms with these values, so we have to convert these values into weighted numbers so that the model can compute these values.

For example, the word ‘Highly Satisfactory’ is signifying something ‘really good’, so we can replace it with ‘3’ as a sentiment score of the review. Whereas, the word ‘Unsatisfactory’ is signifying something ‘poor’ or ‘bad’ review, so we can replace the word with ‘1’ as a sentiment score of the review. And, the word ‘Moderate’ is indicating ‘Not good, Not bad’ or ‘Average’, so we can replace this with ‘2’ as a sentiment score from 1 to 3. And after iterating the same for the rest of the features we will get a result of a weighted sentiment table.



```
In [61]: df.replace(["Highly Satisfactory","Moderate","Unsatisfactory"],[3,2,1],inplace=True)
         df.replace(["Nice","Medium","Poor"],[3,2,1],inplace=True)
         df.replace(["Good","Average","Inadequate"],[3,2,1],inplace=True)
         df.replace(["Worth It","Satisfactory","Nonsensical"],[3,2,1],inplace=True)

In [62]: df.head()
```

Out[62]:

	Feature 1:Picture Quality	Feature 2:Sound Quality	Feature 3:Connectivity	Feature 4:Price
0	3	3	3	3
1	2	2	2	2
2	1	1	1	1
3	2	2	3	2
4	3	3	3	3

Now the model can use this table to compute and perform mathematical expressions and algorithms to achieve the goal that we want.

Phase 1: Data Mining-

K-Means Clustering

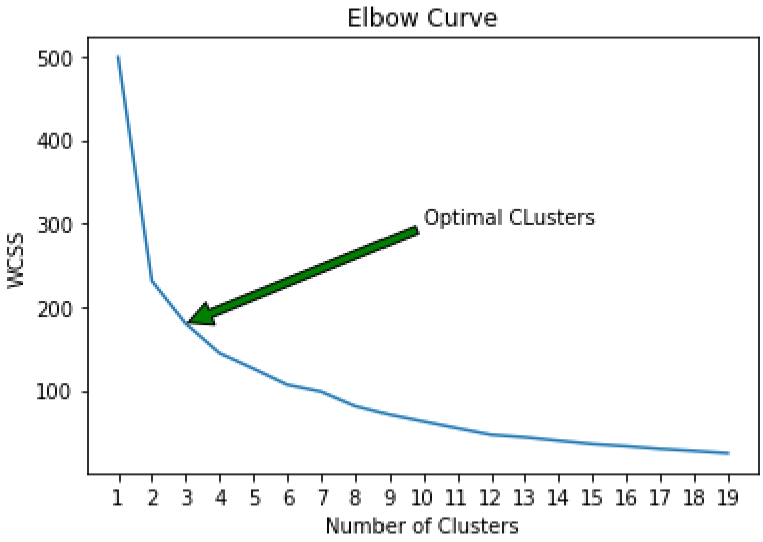
K-Means clustering is one of the simplest and popular unsupervised machine learning algorithms, used for data mining and pattern learning. It is a hard cluster centroid based algorithm whereby data items are clustered into K clusters such that each observation only belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labeled, outcomes. Aiming and minimizing cluster performance index, square error and error criterion are foundations of this algorithm. This algorithm can work on n-dimensional space, so that we can cluster data items with higher order dimension and with more feature attributes.

The K-Means clustering computes Euclidean distances between the data points and calculates centroid by extracting the mean of the Euclidean distances or within -cluster-sum-of-square (WCSS). Finding the number of optimal clusters is an important part of this algorithm. A commonly used method for finding optimal K-value is the Elbow Method.

Suppose that X and Z are two samples of pattern vectors, $X = (x_1, x_2, \cdots\cdots\cdots, x_n)^T$, $Z = (z_1, z_2, \cdots\cdots\cdots, z_n)^T$ and the Euclidean distance between X and Z is,

$$WCSS = ||X - Z|| = \sqrt{\sum_{i=1}^{n} (x_i - z_i)^2}$$

WCSS is the sum of squared distances between each point and the centroid in a cluster. When we plot the WCSS with the K value, the plot will look like an Elbow. As the number of clusters increases the WCSS value decreases and the variance between Euclidean distances will also decrease. The value of WCSS is largest when K=1 and WCSS is 0 when K=n because then every point would become an individual cluster. And when we analyze the graph the graph will rapidly change at a point and thus creating an elbow shape. From this point the plot starts becoming saturated and moves parallel towards the X-axis.



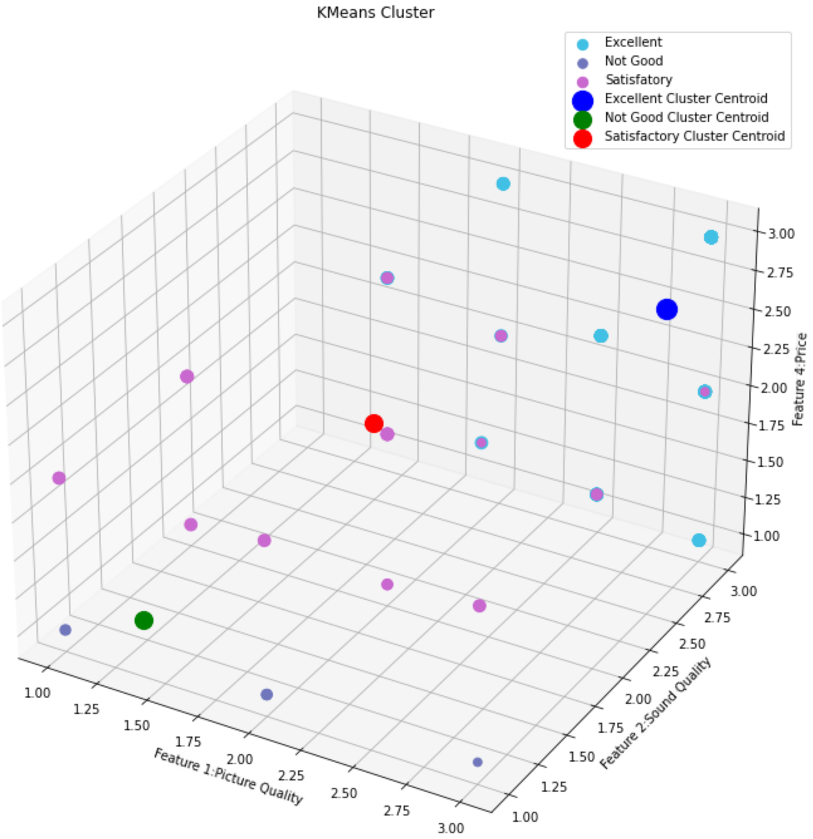
Here we can see the value of K is 3. That means, the number of clusters should be 3 for our dataset named as - Excellent, Satisfactory and Not Good clusters.

Algorithm-

The following steps will help us to understand how K-Means clustering algorithm works-

- *Step 1:* First, we need to provide the number of clusters, K=3, that need to be generated by this algorithm.
- *Step 2:* Next, choose K data points at random and assign each to a cluster. Briefly, categorize the data based on the number of data points.

- *Step 3:* The cluster centroids will now be computed.
- *Step 4:* Iterate the steps below until we find the ideal centroid, which is the assigning of data points to clusters that do not vary.
  - 4.1 The sum of squared distances between data points and centroids would be calculated first.
  - 4.2 At this point, we need to allocate each data point to the cluster that is closest to the others (centroid).
  - 4.3 Finally, compute the centroids for the clusters by averaging all of the cluster's data points.



Here, the cluster has formed in a 4D space where each feature is representing a certain axis- X-axis is representing 'Picture Quality', Y-axis is representing 'Sound Quality', Z-axis is representing 'Price' and the varying size of the data points representing 'Connectivity' and big three dots are the centroid of the three clusters.

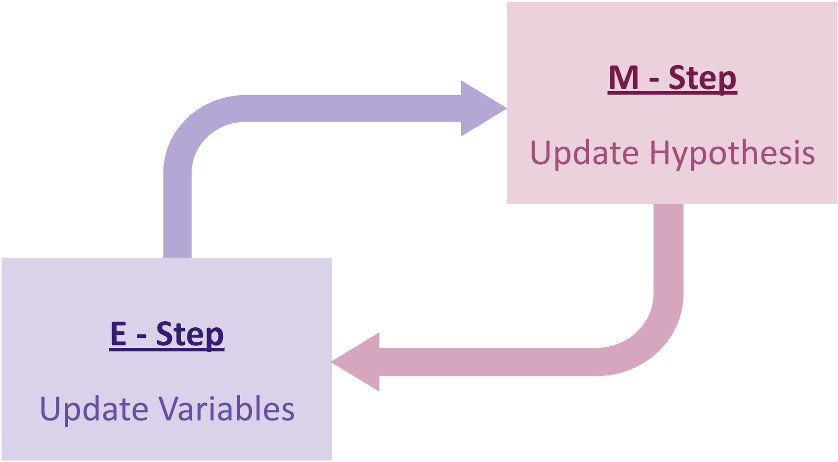
. K-Means clustering method is applied to extract suitable information through proper clustering to solve many complex real-life problems where data values are not labeled and the relation among data can be in n-dimensional space.

Expectation-Maximization or EM Clustering

In K-Means, we approached clustering as the problem of finding codebook vectors that minimize the total reconstruction error. Whereas, the approach of EM-clustering is probabilistic.. This algorithm looks for the component density parameters to find maximum likelihood or maximum a posteriori (MAP) in statistical models, where the model depends on unobserved latent variables.The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the *E* step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

The EM (Expectation Maximization) technique is like a softer version of the K-Means algorithm. Instead of assigning examples to clusters to maximize the differences in means for the continuous variables, the EM clustering algorithm computes probability of cluster memberships based on one or more probability distributions. The goal of the clustering algorithm then is to maximize the overall probability or likelihood of the data, given the (final) clusters.

Given the statistical model which generates a set of X of observed data, a set unobserved latent or missing values Z, and a vector of unknown parameters θ, along with a likelihood function.The maximum likelihood estimate (MLE) of the unknown parameters is determined by maximizing the marginal likelihood of the observed data,



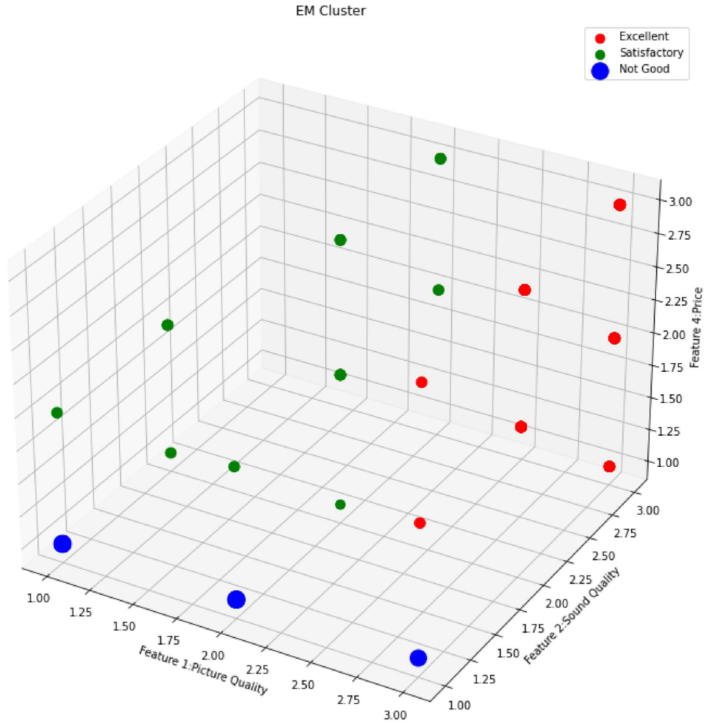
The EM algorithm seeks to find the MLE of the marginal likelihood by iteratively applying these two step:

Expectation step (E-Step) : Using the observed available data of the dataset, estimate (guess) the values of the missing data.

Maximization step (M-Step) : Complete data generated after the expectation (E) step is used in order to update the parameters.

Algorithm-

- Initially, a set of initial values of the parameters are considered. A set of incomplete observed data is given to the system with the assumption that the observed data comes from a specific model.
- The next step is known as “Expectation” – step or *E-step*. In this step, we use the observed data in order to estimate or guess the values of the missing or incomplete data. It is basically used to update the variables.
- The next step is known as “Maximization”-step or *M-step*. In this step, we use the complete data generated in the preceding “Expectation” – step in order to update the values of the parameters. It is basically used to update the hypothesis.
- Now, in the fourth step, it is checked whether the values are converging or not, if yes, then stop otherwise repeat *step-2* and *step-3* i.e. “Expectation” – step and “Maximization” – step until the convergence occurs.



Here, the cluster has formed in a 4D space where each feature is representing a certain axis- X-axis is representing ‘Picture Quality’, Y-axis is representing ‘Sound

Quality', Z-axis is representing 'Price' and the varying size of the data points representing 'Connectivity'.

We thus see that k-means clustering is a special case of EM applied to Gaussian mixtures where inputs are assumed independent with equal and shared variances, all components have equal priors, and labels are hardened. k-means thus pave the input density with circles, whereas EM in the general case uses ellipses of arbitrary shapes, orientations, and coverage proportions.

Weighted Average Value-

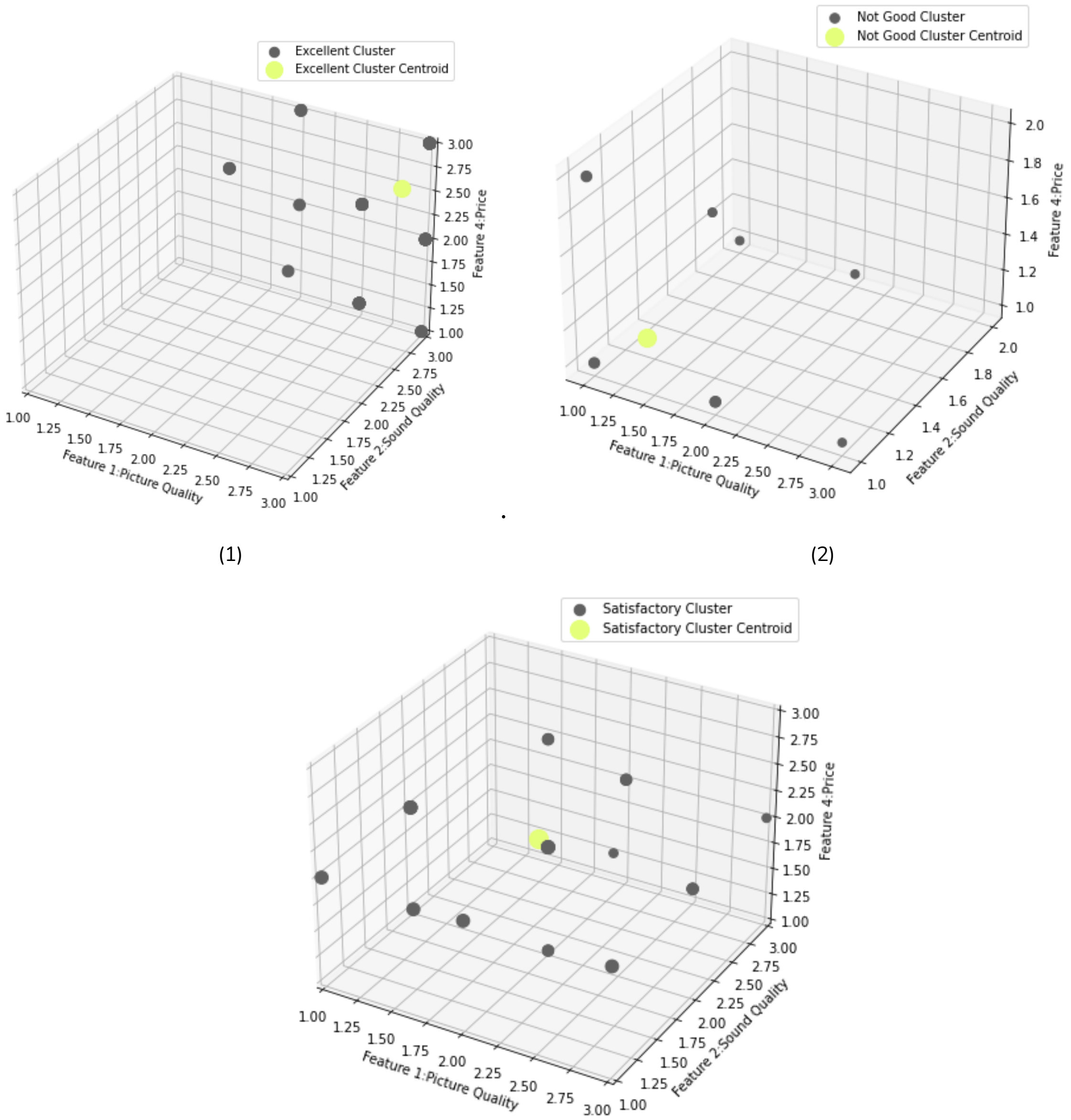
Weighted average is the calculation that takes into account the varying degrees of importance of the numbers in a dataset. In calculating a weighted average, each number in the dataset is multiplied by a predetermined weight before the final calculation is made.

$$weighted\ avg = \frac{\sum_{i=1}^{n} w_i x_i}{|X|}$$

Where X is a set of data points such as $X = (x_1, x_2, \cdots\cdots\cdots, x_n)$ and w_i is weight. now calculate weighted average for each feature attribute in the dataset.

COMPARATIVE STUDY :-

The process of K-Means is something like assigning each observation to a cluster and the process of EM(Expectation Maximization) is finding the likelihood of an observation belonging to a cluster(probability). This is where both of these processes differ. For instance, suppose that most of the observations truly belong to a small number of (unknown) subgroups, and a small subset of the observations are quite different from each other and from all other observations. Then since K- means force every observation into a cluster so that it converges every data point into a particular cluster. That means, every observation must be present in a cluster or not. That's why the clusters found may be heavily distorted due to the presence of outliers that do not belong to any cluster



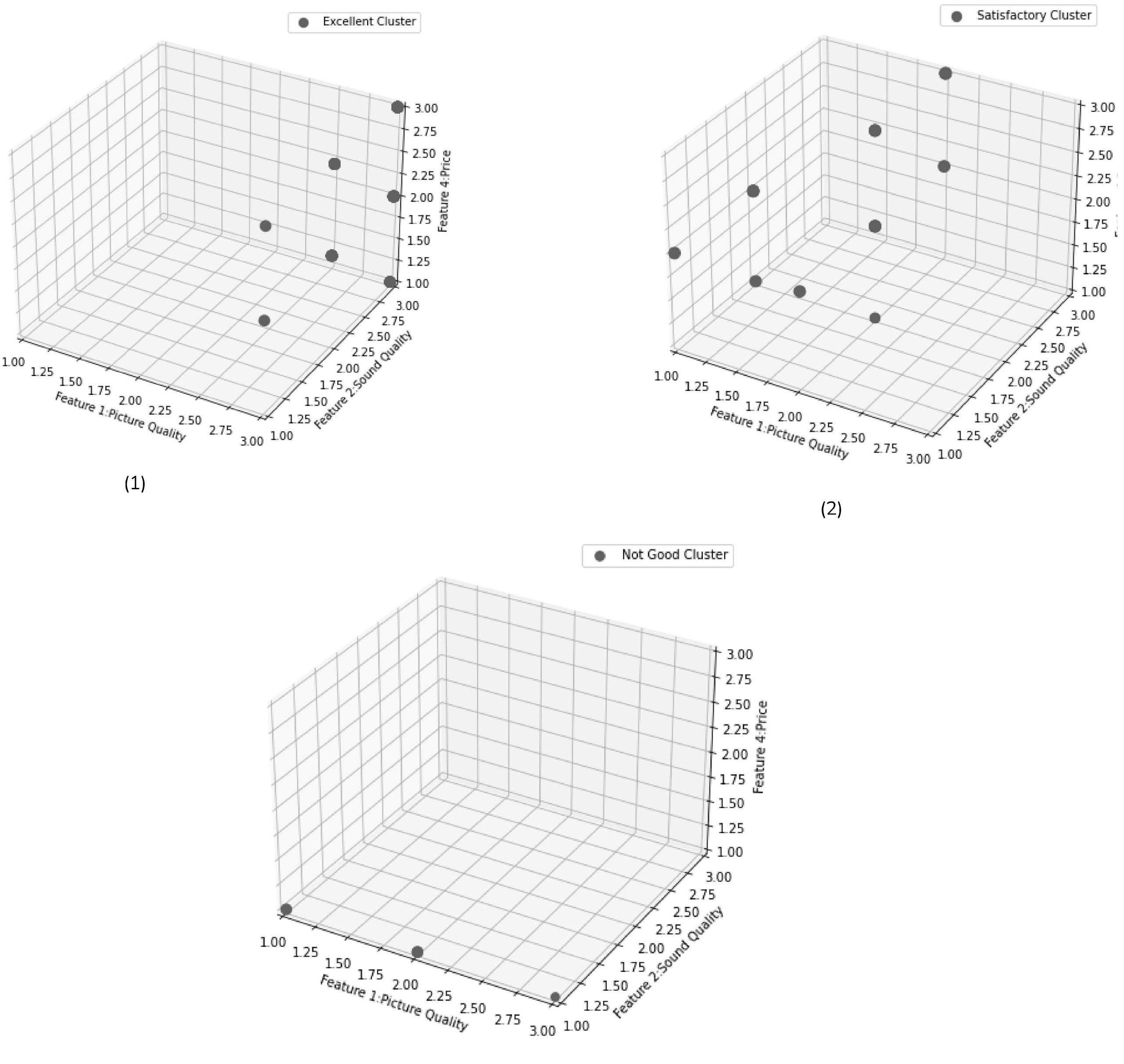
(1)

(2)

(3)

outliers in clusters formed by K-Means

Mixture models are an attractive approach for accommodating the presence of such outliers. EM (Expectation Maximization) clustering algorithm is a soft version of K-means clustering. The approach of EM-clustering is probabilistic. This algorithm looks for the component density parameters to find maximum likelihood or maximum a posteriori (MAP) in statistical models, where the model depends on unobserved latent variables. The goal of the clustering algorithm then is to maximize the overall probability or likelihood of the data, given the (final) clusters. Unlike K-Means, it doesn’t calculate the reconstruction error of the vectors but calculates the probability of likelihood to present the observations in the cluster and that’s how it decrease the chance of presence of outliers in the clusters.



(1)

(2)

(3)

outliers in clusters formed by EM Cluster

EXPERIMENTAL ANALYSIS :-

The values are follows -

Feature1	Feature2	Feature3	Feature4
Highly Satisfactory(3)	Nice(3)	Good(3)	Worth It(3)
Moderate(2)	Medium(2)	Average(2)	Satisfactory(2)
Unsatisfactory(1)	Poor(1)	Inadequate(1)	Nonsensical(1)

Now we shows our dataset with values as follows -

Review Id	Feature 1:Picture Quality	Feature 2:Sound Quality	Feature 3:Connectivity	Feature 4:Price
1	3	3	3	3
2	2	2	2	2
3	1	1	1	1
4	2	2	3	2
5	3	3	3	3
6	1	1	1	1
7	3	3	3	3
8	3	3	3	3
9	3	3	3	2
10	1	1	1	1
11	1	2	3	2
12	3	3	3	3
13	3	3	2	2
14	1	2	2	2
15	3	3	3	3
16	3	3	3	3
17	3	2	3	3
18	3	2	2	3
19	1	2	1	1
20	3	3	2	2
21	1	1	1	1
22	2	2	2	3
23	1	1	1	1
24	2	2	1	1
25	2	2	3	3
26	2	1	1	1
27	3	3	3	3
28	3	3	3	2
29	3	3	3	3
30	3	3	2	2
31	2	1	1	2
32	3	1	2	2
33	2	2	2	2
34	3	2	2	3
35	3	3	3	2
36	3	3	2	3
37	2	3	2	2
38	3	2	3	3
39	3	3	3	3
40	2	2	2	2
41	3	3	3	3
42	3	3	3	2
43	3	2	3	2
44	2	2	2	2

45	3	3	3	3
46	3	3	3	3
47	2	2	3	3
48	3	2	2	2
49	2	3	3	3
50	1	1	1	1
51	3	3	3	2
52	3	3	2	2
53	3	3	3	2
54	2	1	2	2
55	3	3	3	1
56	3	2	3	3
57	3	3	2	3
58	3	2	3	2
59	2	3	3	2
60	3	3	3	3
61	3	3	2	3
62	3	2	3	2
63	3	2	3	2
64	1	2	2	2
65	1	1	1	1
66	3	2	2	3
67	3	1	3	2
68	2	2	2	2
69	3	2	3	3
70	3	3	3	2
71	3	3	3	3
72	3	3	2	2
73	3	2	3	3
74	3	3	3	3
75	3	2	3	2
76	1	2	3	2
77	3	3	1	3
78	3	3	1	2
79	2	3	1	2
80	3	3	3	3
81	3	3	3	3
82	2	1	3	2
83	2	2	2	2
84	3	3	2	2
85	1	1	1	1
86	3	3	2	3
87	3	3	3	2
88	3	2	2	2
89	3	3	3	1
90	3	3	2	3
91	2	1	2	1
92	3	1	3	3

93	3	3	2	2
94	3	2	3	2
95	2	2	3	2
96	3	2	3	3
97	3	1	1	1
98	2	3	3	3
99	3	3	3	3
100	3	2	3	2
101	2	2	3	2
102	1	2	2	2
103	1	1	3	2
104	2	3	3	3
105	3	2	3	2
106	3	3	3	1
107	3	3	3	3
108	2	1	2	2
109	3	2	2	3
110	3	2	3	3
111	3	3	3	3
112	2	2	1	2
113	1	1	2	2
114	3	3	2	3
115	3	3	2	3
116	3	3	2	2
117	3	3	3	3
118	3	3	2	3
119	3	2	2	2
120	3	3	3	1
121	3	3	3	2
122	3	3	3	3
123	2	2	1	2
124	3	2	2	3
125	2	2	2	2
126	3	3	2	3
127	3	3	3	3
128	3	3	3	3
129	3	3	2	2
130	3	3	3	3
131	3	3	2	3
132	3	3	3	2
133	3	3	2	3
134	3	3	3	3
135	3	3	3	3
136	2	2	2	1
137	3	2	3	2
138	3	3	3	3
139	2	1	1	2
140	1	2	3	1

141	1	1	1	1
142	3	3	2	3
143	3	2	3	3
144	3	2	2	3
145	1	2	2	2
146	2	2	2	1
147	3	3	3	2
148	3	3	3	3
149	3	3	3	2
150	3	3	2	2
151	3	3	3	3
152	2	3	1	2
153	2	2	2	2
154	3	2	2	3
155	3	2	3	3
156	1	2	3	2
157	1	1	2	1
158	3	3	3	2
159	2	2	2	2
160	2	2	2	2
161	3	3	2	3
162	3	2	3	3
163	3	3	3	2
164	2	2	3	2
165	2	2	2	3
166	3	3	3	2
167	2	3	2	2
168	3	2	2	3
169	1	1	1	1
170	3	3	3	3
171	1	1	1	1
172	2	2	3	3
173	3	3	3	3
174	3	2	2	3
175	2	1	2	1
176	3	3	2	3
177	3	3	3	3
178	3	2	3	3
179	2	2	2	2
180	1	1	1	1
181	3	1	1	3
182	3	3	2	3
183	3	3	3	3
184	2	1	1	1
185	1	1	1	1
186	3	2	2	3
187	3	3	3	3
188	3	3	3	3

189	3	3	3	3
190	3	2	3	3
191	1	1	1	1
192	3	3	3	3
193	2	2	1	2
194	3	3	2	2
195	3	2	3	2
196	3	3	3	3
197	3	3	3	3
198	1	1	1	1
199	3	3	3	2
200	3	2	3	3

Here we show our clustering results as follows-

K-Means Clustering Results :-

```
Data written to the working file.
4 variables and 200 cases written.
Variable: Picture_Quality    Type: Number  Format : F1
Variable: Sound_Quality      Type: Number  Format : F1
Variable: Connectivity       Type: Number  Format : F1
Variable: Price              Type: Number  Format : F1

Substitute the following to build syntax for these data.
/VARIABLES=
Picture_Quality F1
Sound_Quality F1
Connectivity F1
Price F1
```

Quick Cluster

```
[DataSet1]
```

Iteration History[a]

	Change in Cluster Centers		
Iteration	1	2	3
1	1.525	1.074	.607
2	.188	.240	.144
3	.085	.068	.098
4	.139	.171	.056
5	.022	.081	.104
6	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 6. The minimum distance between initial centers is 2.828.

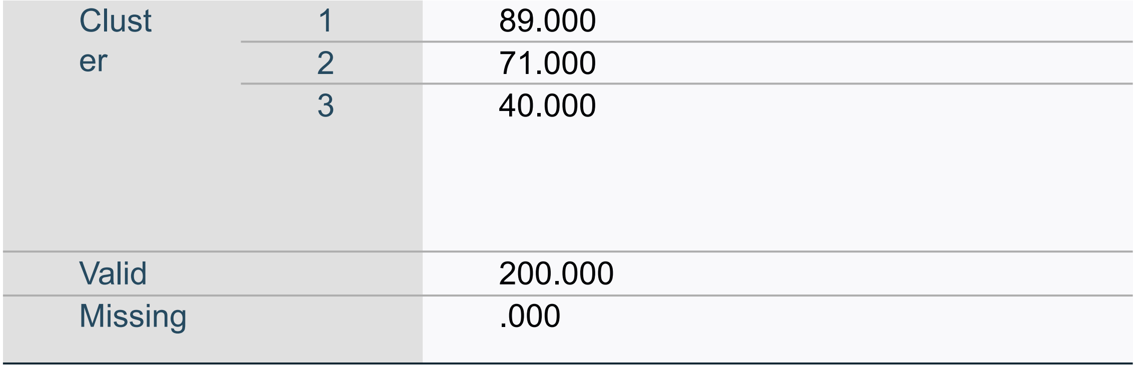
Final Cluster Centers

	Cluster		
	1	2	3
Picture Quality	3	3	1
Sound Quality	3	2	1
Connectivity	3	3	1
Price	3	2	1

Initial Cluster Centers

	Cluster		
	1	2	3
Picture_Quality	3	3	1
Sound_Quality	3	3	1
Connectivity	1	3	1
Price	3	1	1

Number of Cases in each Cluster



Cluster	1	89.000
	2	71.000
	3	40.000
Valid		200.000
Missing		.000

E-Means Clustering Results :-

```
=== Run information ===

Scheme:         weka.clusterers.EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10
-num-slots 1 -S 100
Relation:     Dataset1
Instances:    200
Attributes:   5
Review Id
Feature 1:Picture Quality
Feature 2:Sound Quality
Feature 3:Connectivity
Feature 4:Price
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


EM
==

Number of clusters selected by cross validation: 3
Number of iterations performed: 2
```

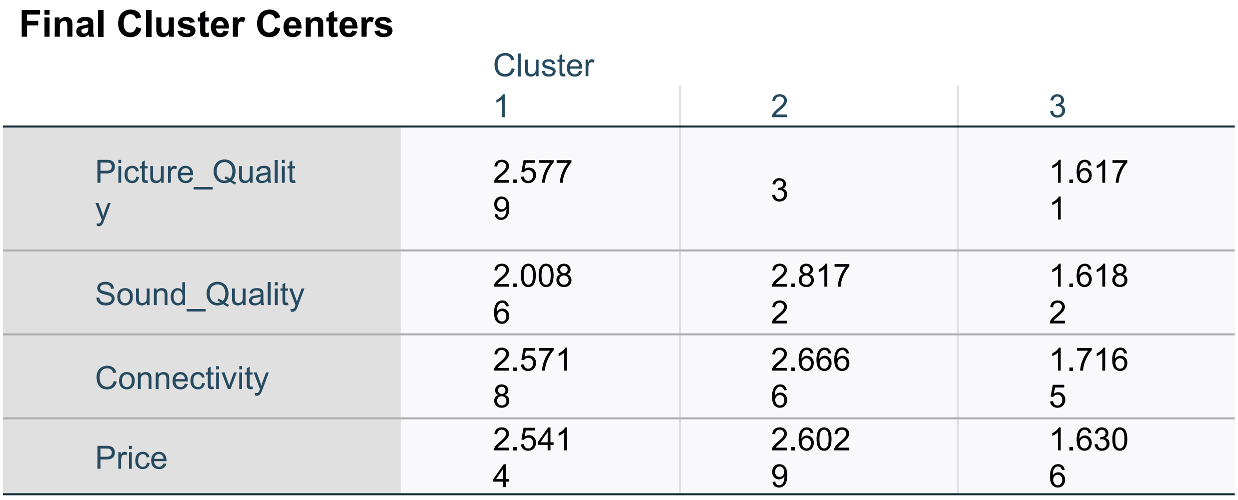
```
Cluster
Attribute                    0      1      2
(0.16)   (0.54)   (0.3)
======================================================

Feature 1:Picture Quality
mean                  2.5779      3   1.6171
std. dev.             0.5017   0.0031   0.6021

Feature 2:Sound Quality
mean                  2.0086   2.8172   1.6182
std. dev.             0.1139   0.4151    0.701

Feature 3:Connectivity
mean                  2.5718   2.6666   1.7165
std. dev.             0.4974   0.5094   0.7512

Feature 4:Price
mean                  2.5414   2.6029   1.6306
std. dev.             0.4996   0.5602   0.5972
```



Final Cluster Centers

	Cluster		
	1	2	3
Picture_Quality	2.5779	3	1.6171
Sound_Quality	2.0086	2.8172	1.6182
Connectivity	2.5718	2.6666	1.7165
Price	2.5414	2.6029	1.6306

```
Time taken to build model (full training data) : 0.75 seconds

=== Model and evaluation on training set ===
Clustered Instances

0       20 ( 10%)
1      127 ( 64%)
2       53 ( 27%)

Log likelihood: -5.97174
```

Weight Value Average:-

Feature1(picture quality)

Highly Satisfactory=130

Moderate=43

Unsatisfactory=27

Weight value=((130*3)+(43*2)+(27*1))/200=2.515

Feature2(sound quality)

Nice=97

Medium=71

Poor=32

Weight Value=(97*3)+(71*2)+(32*1)/200=2.325

Feature3(connectivity)

Good=103

Average=67

Inadequate=30

Weight Value=(103*3)+(67*2)+(30*1)/200=2.365

Feature4(price)

Worth It=90

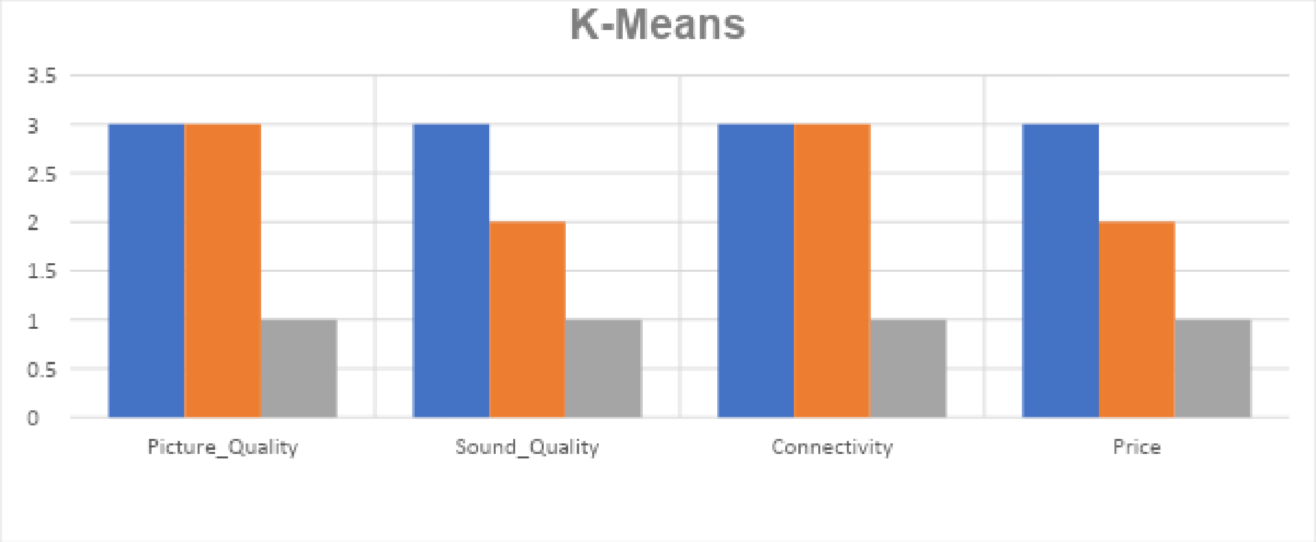
Satisfactory=80

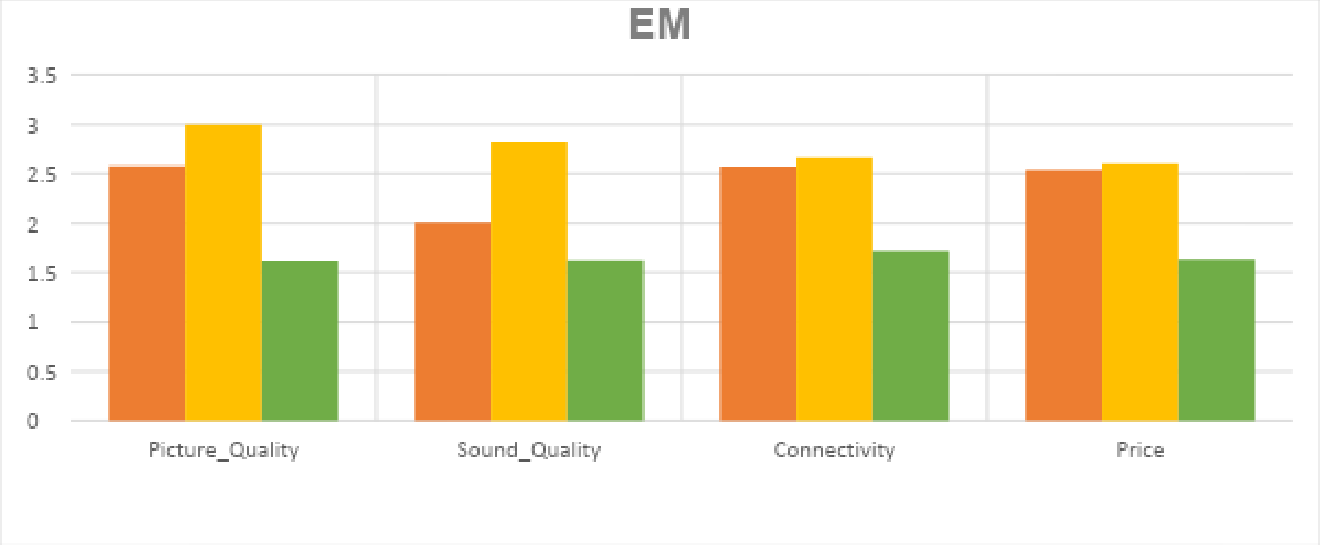
Nonsensical=30

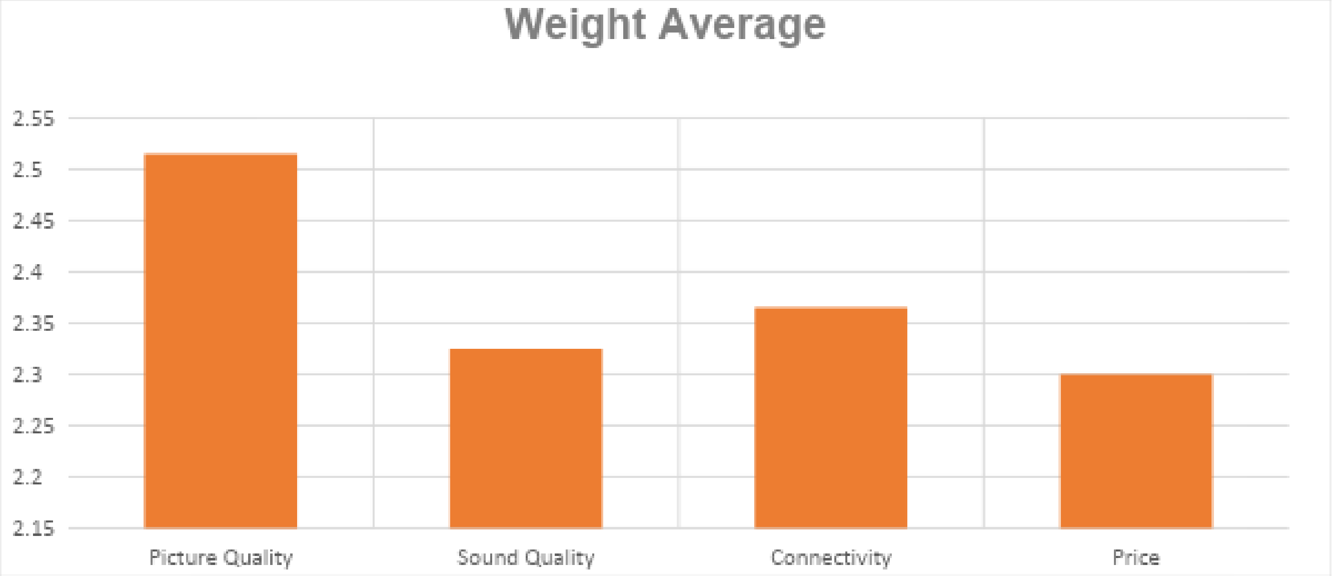
Weight Value=(90*3)+(80*2)+(30*1)/200=2.300

Picture Quality	2.515

Sound Quality	2.325
Connectivity	2.365
Price	2.3


K-Means
3.5
3
2.5
2
1.5
1
0.5
0
Picture_Quality
Sound_Quality
Connectivity
Price


EM
3.5
3
2.5
2
1.5
1
0.5
0
Picture_Quality
Sound_Quality
Connectivity
Price


Weight Average
2.55
2.5
2.45
2.4
2.35
2.3
2.25
2.2
2.15
Picture Quality
Sound Quality
Connectivity
Price

Future Scope:-

In future the customers would efficiently analyze their needs of product based on the user-generated online product reviews. This paper is illustrated with the case study of amazon.com product furthermore other platforms can be included. Not only k-means or EM-clustering, but also hierarchical clustering or Agglomerative and Optics clustering algorithms can be used.

Here it is only categorized the customer needs related to the topics extracted from the product level. Therefore, more detailed analysis of individual products and its services should be added for understanding their individual performance in the future.

We can also work with different kinds of referential fields and can reveal the product attributes with its optimality. Also, it can be implemented in many different ways.

In future ,this project can be extended to implement the type of customers and their preferences. Along with favorable time of choosing a product it would analyze the security of a product.

Hence ,more research is needed to further develop.

CONCLUSION :

In present times E-commerce (electronic commerce) is the buying and selling of goods and services, or the transmitting of funds or data, over an electronic network, primarily the internet. These business transactions occur either as business-to-business (B2B), business-to-consumer (B2C), consumer-to-consumer or consumer-to-business. The terms e-commerce and e-business are often used interchangeably. The term e-tail is also sometimes used in reference to the transactional processes that make up online retail shopping.

In this paper, we proposed a machine-learning approach to effectively and efficiently analyze customer needs of product ecosystems based on the user generated online product reviews on **LED TV** which is manufactured by **LG** Corporation. We have used four features, including **Picture Quality**, **Sound Quality**, **Connectivity** and **Price as** a part of our experimental analysis. Two different clustering techniques namely K-means and EM have been applied on our input data and the results has been compared with weighted value average technique also. In E-means clustering, the results can't help to reach the final decision for recommending the product to the customers specifically. Similarly in weighted average technique the generated value is also not too satisfactory. Finally, the K-means is used to categorize customer needs related to each feature quantitatively. The proposed clustering method was illustrated with a case study of our product and the results demonstrated the potential of the proposed method.

REFERENCES:-

[1] Vishal A. Kharde, S. S. Sonawane, *"Sentiment Analysis of Twitter Data: A Survey of Techniques"*, *International Journal of Computer Applications (0975-8887)*, Volume 139 – No. 11, April 2016.

[2] Vijay Nair, Supti Kanta Mohapatra, Reema Malhotra, Nivedith Maknoor, *"A Machine Learning Algorithm for Product Classification based on Unstructured Text Description"*, *International Journal of Engineering Research & Technology (IJEERT),* ISSN: 2278-0181, Vol. 7 Issue 6, June-2018.

[3] Maggie Yundi Li, Stanley Kok, Liling Tan, *"Don't Classify, Translate: Multilevel E-Commerce Product Categorization via Machine Translation".*

[4] Hadi Jahansahi, Ozan Ozyegen, Mucahit Cevik, Beste Bulut, Deniz Yigit, Fahrettin F. Gonen, Ayse Basar, *"Text Classification for Predicting Multi-level Product Caategories", CASCON,*September 7, 2021.

[5] Feng Zhou, Jackie Ayoub, Qianli Xu, X. Jessie Yang, *"A Maachine Learning Approach to Customer Needs Analysis for Product Ecosystems", Journal of Mechanical Design,* Vol. 142/011101-1, January 2020.

[6] Kuncherichen K Thomas, Sarath P Anil, Ebin Kuriakose, Neema Geroge, *"Sentment Analysis in product reviews using Natural Language Processing and Machine Learning", International Journal of Information System and Computer Sciences,* Vol. 8, No.2, ISSN: 2319-7595, March-April 2019.

[7] Shweta Dhabekar, Dr. M. D. Patil, *"Implementation of Deep Learning Based Sentiment Classification and Product Aspect Analysis", ITM Web of Conferences,* 2021.

[8] Anusuya Dhara, Arkadeb Saha, Sourish Sengupta, Pranit Bose, *"Sentiment Analysis of Product-Based Reviews Using Machine Learning Approaches*", RCC Institute of Information Technology, 2017-2018.

[9] Sushanta Sarkar and Irving Lin, *"Applying Machine Learning to Product Categorization", Stanford University*.

[10] Pawlowski, M. (2021), *"Machine Learning Based Product Classification for eCommerce", Journal of Computer Information Systems.*

[11] Ms. Swapna Kadam, Dr. Sarika Chouhan, *"Consumer Buying Behavior Analysis Using Machiine Learning Algorithm", Pradnyaa International Journal of Multidisciplinary Research,* Vol. 01, Issue No. 01.

[12] D. Elangovan, V. Subedha, *"Product Review Based on Machine Learning Algorithms", Smart Intelligent Computing and Communication Technology,* 2021.

[13] Xing Fang and Justin Zhan, *“Sentiment Analysis using Product Review Data”, Journal of Big Data (2015) 2:5.*

[14] Zeeina Singla, Sukhchandan Randhawa, Sushma Jain, *“Sentiment Analysis of Customer Product Reviews Using Machine Learning”, International Conference on Intelligent Computing and Control, 2017.*

[15] Rosa Arboretti, Riccardo Ceccato, Luca Pegoraro, Luigui Salmaso, *“Design of Experiments and Machine Learning for Product Innovation: A Systematic Literature Review”, Quality and Reliability Engineering International,* Vol. 38, ISSN-1131-1156, 25 October, 2021.

[16] Kaustuv Deb, Sonali Banarjee, Rudra Prasad Chatterjee, Atanu Das, Rajiv Bag, *“Educational Website Ranking Using Fuzzy Logic and K-Means Clustering Based Hybrid Method”, International Information and Engineering Technology Association,* Vol. 24, No. 5, October, 2019, pp. 497-506.