



(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

**DATA MINING IN RETAIL SECTOR FOR SALES  
BEHAVIOUR PREDICTION**

**A Project Report of Capstone Project - 2**

*Submitted by*

**ALOK PANDEY**

**(1613114008 / 16SCSE114051)**

*in partial fulfillment for the award of the degree  
of  
Bachelor of Technology*

**IN**

**Computer Science and Engineering**

**SCHOOL OF COMPUTING SCIENCE AND ENGINEERING**

**Under the Supervision of**

**Dr. SUDEEPT SINGH YADAV  
M.Tech., Ph.D, Professor**

**APRIL / MAY- 2020**



**SCHOOL OF COMPUTING AND SCIENCE AND ENGINEERING**

**BONAFIDE CERTIFICATE**

Certified that this project report **“DATA MINING IN RETAIL INDUSTRY FOR SALES BEHAVIOUR PREDICTION”** is the bonafide work of **“ALOK PANDEY (1613114008)”** who carried out the project work under my supervision.

**SIGNATURE OF HEAD**

Dr. MUNISH SHABARWAL,  
PhD (Management), PhD (CS)  
**Professor & Dean,**  
**School of Computing Science &  
Engineering**

**SIGNATURE OF SUPERVISOR**

Dr. SANJEEV KUMAR PIPAL,  
M.Tech., Ph.D.,  
**Professor**  
**School of Computing Science &  
Engineering**

## **TABLE OF CONTENTS**

- Abstract
- Introduction
  - (i) Overall description
  - (ii) Purpose
  - (iii) Motivations and scope
- Proposed model
- Existing System
- Implementation or Architecture Diagrams
- Output/Results
- Conclusion
- References

## ABSTRACT

Data mining refers to the process of extraction of useful information from a pool of data. Data mining has become one of the widely researched fields off late especially with the increase in technological advancements which have caused the volume of data to be stored and processed for query based applications or decision based processes to increase in great multiplicative measures. Data mining is also known as Knowledge Discovery in Databases (KDD) and enables data exploration, data analysis, and data visualization of huge databases at a high level of abstraction, without a specific hypothesis in mind. Big data and Data mining, both play important role in helping retailers to expand their market. Big Data is used to describe a collection of data that is huge in size and yet growing exponentially with time. Extremely huge data sets that may be analysed to reveal patterns, trends, and associations, especially relating to human behaviour and interactions. In this paper we are focused to find a proper and more relevant way to do data analysis. Data mining is proved to be one of the important tools for identifying useful information from very large amount of data bases in almost all the industries. Industries are using data mining to increase revenues and reduce costs. This article begins the concept of data mining that has emerged as a technique of discovering patterns to make better strategies and decisions. It also discusses standard tasks involved in data mining, reviews various data mining applications in different sectors. The algorithm proposed in this paper are classification and clustering in which classification has logistic regression, decision tree and KNN(k-nearest neighbor) method whereas clustering comprises of Centroid based, Distributive based, Connectivity based and Density based clustering. All these methods are used to find best way for selecting best way for data mining. This paper attempts, how data mining can be applied in retail industry to improve market campaign.

## **INTRODUCTION**

### **(i) Overall Description :**

Data Mining is the process of discovering interesting patterns and knowledge from large amount of data. The data sources can include databases, data warehouses, the web, other information repositories, or data that are streamed into the system dynamically. Data Mining is a process of extracting previously unknown, valid, potentially useful and hidden patterns from large data sets. Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Today retailer is facing dynamic and competitive environment, with increase in globalization and competitiveness retailers are seeking better market campaign. Retailer are collecting large amount of customer dialy transaction details. This data collected requires proper mechanisms to convert it into knowledge, using this knowledge retailer can make better business decision. Retail industry is looking strategy where in they can target right customers who may be profitable to them.

Knowledge discovery is the most valued output of computing. Finding new phenomenon or generating new patterns involves Data Mining and which includes a large set of data basically known as Big Data. Big Data comes into role because previous technologies were not able to handle that much amount of data efficiently. Financial data analysis, retail marketing, telecommunication industry, biological data analysis, scientific applications, intrusion detection, data mining applications in transportation, data

mining applications in medicine, data mining applications in health care and insurance, visual and audio data mining.

### DATA MINING AND ONLINE SHOPPING

Data mining is crucial for extracting and identifying useful information from a large amount of data that is why retailing companies operate purchase databases in a long way, such that all transactions are stored in arranged order.

A record-of-transaction database typically contains the transaction date and the products bought in the course of a given transaction. Usually, each record also contains e-shopper ID, particularly when the purchase was made using a credit card or a frequent buyer card. Therefore, the purchasing sequence of an e-shopper in the database that has made repeated purchase can easily be determined.

This purchase sequence provides a description of the changes in an e-shopper' s preferences over time ,because a purchase sequence can reveal the changes of e-shopper' s preferences over time.

#### **( ii ) Purpose :**

The Amount of data generated and published over internet is drastically increasing day by day, the collection of Big Data is a set of large data which is too large and complicates to handle in a traditional data processing ways. Big data is the ability of processing data with the following properties velocity, variety and volume. Knowledge Discovery is the most desirable end Finding new phenomena or product of computing enhancing our knowledge about them has a greater value than optimizing production processes or and is second only to task that preserve our,inventories world and our environment. Data mining has the ability to extract hidden predictive information from very large databases of retail sales data. It is a powerful technology with great potential to help retailers focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviours, helps organisations to make proactive knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by

retrospective tools typical of decision support systems. Data mining tools can answer the questions that traditionally were time consuming to resolve. They prepare databases for finding hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. The extraction of hidden predictive information from large databases is a powerful tool with great potential to help organizations to define the information market needs of tomorrow. Data mining tools predict future trends and behaviors, allowing businesses to make knowledge-driven decisions that will affect the company, both short term and long term. The automated *prospective analysis* offered by data mining tools of today is much more effective than the analysis provided by tools in the past. Data mining answers business questions that traditionally were too time-consuming to resolve. Data mining tools search databases for hidden patterns, finding predictive information that experts may miss because it was outside their expectations.

**( iii ) Motivational and Scope :**

Retail industry is also realizing that it is possible to gain a competitive advantage utilizing data mining. Retailers have been collecting enormous amounts of data throughout the years, just like the banking industry, and now have the tool needed to sort through this data and find useful pieces of information. For retailers, data mining can be used to provide information on product sales trends, customer buying habits and preferences, supplier lead times and delivery performance, seasonal variations, customer peak traffic periods, and similar predictive data for making proactive decisions. Here are some examples of how the retail industry has been utilizing data mining effectively. Marketing: One of the most widely used areas of data mining for the retail industry, as in the banking industry, is marketing. 'Market basket analysis' is a marketing method used by many retailers to determine optimal locations to promote products. In telecommunication industry it can be useful for improving service quality, making better use of resources etc. In Financial Banking

With computerised banking everywhere huge amount of data is supposed to be generated with new

transactions. Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices that are not immediately apparent to managers because the volume data is too large or is generated too quickly to screen by experts.

Simply stated, it is the study of retail stock movement data recorded at a Point of Sale (PoS)—to support decisions on shelf space allocation, store layout, product location and promotion effectiveness.

- Fraud Detection
- Customer Acquisition and Retention
- Intrusion Detection:- Any action that will compromise the integrity and confidentiality of a resource is an intrusion. The defensive measures to avoid an intrusion includes user authentication, avoid programming errors, and information protection. Data mining can help improve intrusion detection by adding a level of focus to anomaly detection. It helps an analyst to distinguish an activity from common everyday network activity.



## **PROPOSED MODEL**

Nowadays Big data is playing a very big role in industrialization and sales industry and we can use that Data using Data Mining to extract more information and patterns to make it more valuable and useful. And key components of making it possible are described below.

### **Market Basket Analysis Market :**

Market basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns. It involves analyzing large data sets, such as purchase history, to reveal product groupings, as well as products that are likely to be purchased together. The adoption of market basket analysis was aided by the advent of electronic point-of-sale (POS) systems. Compared to handwritten records kept by store owners, the digital records generated by POS systems made it easier for applications to process and analyze large volumes of purchase data.

### **Consumer Behaviour Analysis :**

Before diving into the deep end of consumer behaviour analysis, it is important to understand its meaning. Consumer behaviour analysis is the study of how people make purchase decisions with regard to a product, service or organisation. Studying consumer behaviour would allow you to answer several questions, such as:

- How consumers feel about alternatives to their preferred brands;
- How consumers choose between the alternatives;
- How consumers behave while shopping;
- How consumer behaviour is swayed by their surrounding environment;
- How marketing campaigns can be improved to more effectively influence customer behaviour.

In order to understand this concept better, let's take a look at the factors that affect consumer behaviour:

- **Psychological:** This is considered to be the most important factor that affects consumer behaviour. Traits like perception, motivation, personality, beliefs and attitude are important to decide why a consumer would buy a product.
- **Personal:** These are characteristics that are applicable to individuals and may not relate to other people in a group. These factors can include age, occupation, financial situation and lifestyle.
- **Social:** Social characteristics play an important role in consumer behaviour, and it can include family, communities and social interaction. These factors are difficult to assess while preparing marketing plans.
- **Geographical:** The location of consumers also play a role in how they purchase products. For example, a person living in warmer weather would be less likely to purchase winter clothing compared to someone living in temperate climate.

### **Consumer Behavior**

Consumer behavior means the study of individuals, groups or organizations about their process of selecting, securing, using and disposing the products, services, experiences or ideas to satisfy needs and the impact of these process on the consumer and the society. Behavior concerns either with the individual or the group (e.g. In college friends influence what kind of clothes a person should wants to wears) or a firm (peoples working in firm make decision as to which products the firm should use.) The use of product is often so important to the marketer because this may influence how a product is best positioned or how we can encourage increased consumption.

### **Algorithms Used :**

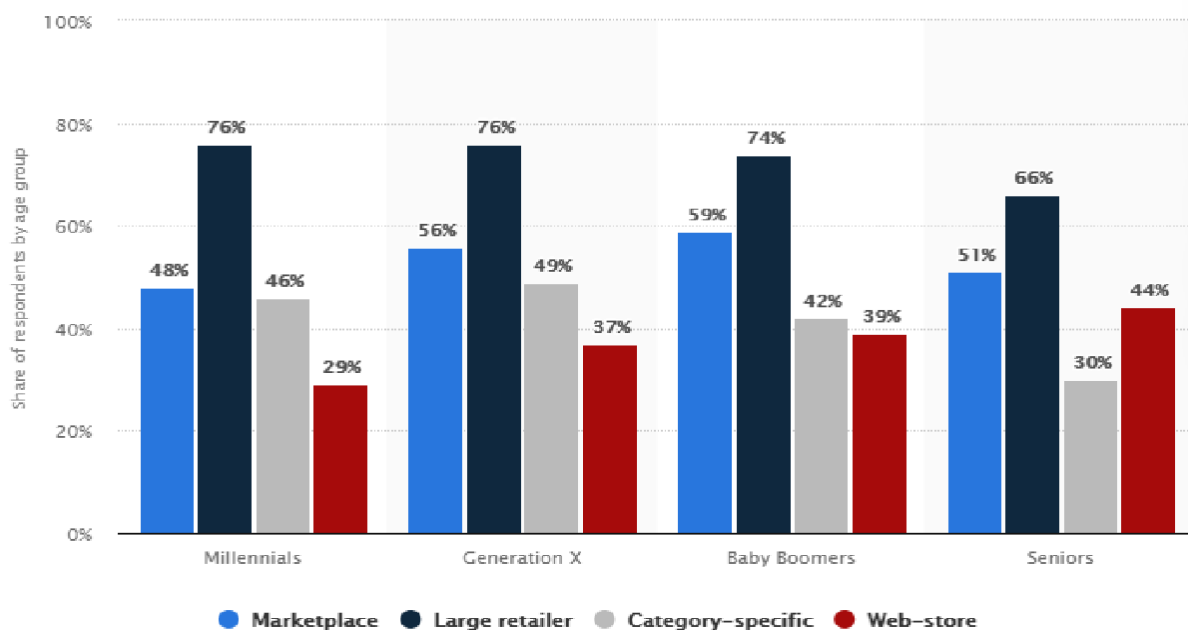
#### **a) Classification :**

It is a Data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts. Classification is the problem of identifying to which of a set of categories (subpopulations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known.

## Classification based on:

### Age Group based Classification

This statistic presents the shopping preference regarding retail types in the United States as of 2017, by age group. According to the findings, the strongest preferred online shopping channel across all age groups were large retailers, with 76 percent of Millennials having recorded this shopping method. Additionally, 74 percent of Baby Boomers held the same preference as Millennials.



### Client based Classification

In the context of customer segmentation, cluster analysis is the use of a mathematical model to discover groups of similar customers based on finding the smallest variations among customer within each group. These homogeneous groups are known as “customer archetypes” or “personas”.

The goal of cluster analysis in marketing is to accurately segment customers in order to achieve more effective customer marketing via personalization. A common cluster analysis method is a mathematical algorithm known as *k-means cluster analysis*, sometimes referred to as scientific segmentation. The clusters that result assist in better customer modeling and predictive analytics, and are also used to target customers with offers and incentives personalized to their wants, needs and preferences.

		R	F	M
Customers		Days Since Last Purchase	Number of Purchases (Past 12 Months)	Net Revenue (Past 12 Months)
High Spenders	922	9	4	\$154
Mid Spenders	581	54	3	\$121
Risk of Churn	807	192	2	\$70
Low Spenders	1,361	192	2	\$4
	<b>3,671</b>	<b>447</b>	<b>3</b>	<b>\$87</b>

Some of classification techniques are:

i. **Logistic Regression**

It estimates discrete values (**Binary values like 0/1, yes/no, true/false**) based on a given set of independent variable(s). Simply put, it basically, predicts the probability of occurrence of an event by fitting data to a *logit function*. Hence, it is also known as **logit regression**. The values obtained would always lie within 0 and 1 since it predicts the probability.

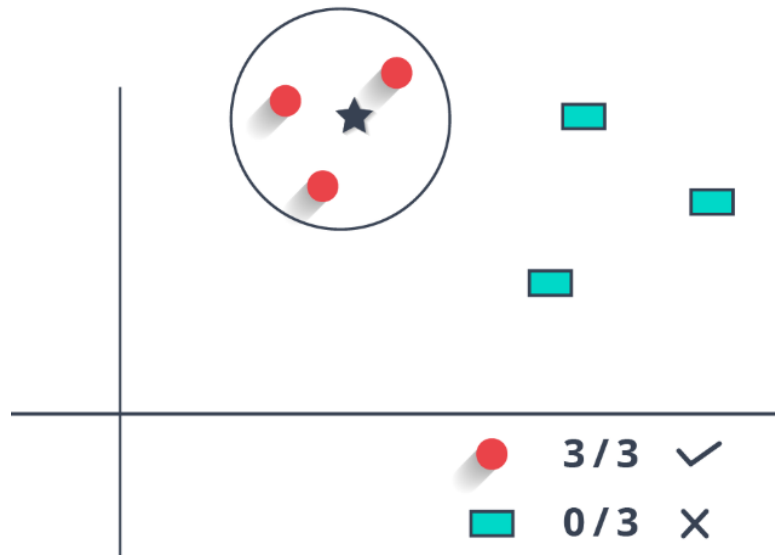
ii. **Decision Tree**

With versatile features helping actualize both categorical and continuous dependent variables, it is a type of supervised learning algorithm mostly used for classification problems. What this algorithm does is, it splits the population into two or more homogeneous sets based on the most significant attributes making the groups as distinct as possible.

iii. **KNN (k- Nearest Neighbour)**

**K nearest neighbors** is a simple algorithm used for both classification and regression problems. It basically stores all available cases to classify the new cases by a majority vote of its k neighbors. The case assigned to the class is most common amongst its K nearest neighbors measured by a distance function (Euclidean, Manhattan, Minkowski, and Hamming).

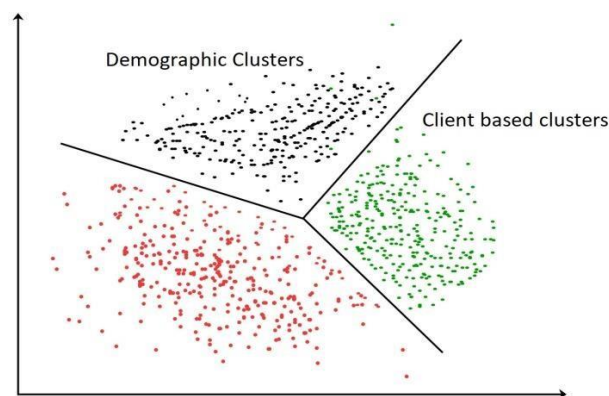
While the three former distance functions are used for continuous variables, Hamming distance function is used for categorical variables. If **K = 1**, then the case is simply assigned to the class of its nearest neighbor. At times, choosing K turns out to be a challenge while performing kNN modeling.



**KNN (k- Nearest Neighbour)**

**b) Clustering :**

Clustering is a Machine Learning technique that involves the grouping of data points. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group. In theory, data points that are in the same group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features. Clustering is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields.



Some of clusters are:

- Demographic Clusters
- Client Based Clusters
- Age Group Clusters
- Language Bases Clusters
- Location Based Clusters
- Sales Based Clusters

## Existing System

Data is raw material of information that can be understood as any facts, numbers, or text which can be processed by machines. Information is the data that has been given some meaning nu way of relational connections. For ex data collected from sales transaction can be used to analyze sales trends of particular years. Knowledge is application of data and information.it can be considered as general awareness of information, facts, ideas, truth or principle.

Across a wide variety of fields, data are being collected and accumulated at a dramatic pace. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. These theories and tools are the subject of the emerging field of knowledge discovery in databases.

At an abstract level, the KDD field is concerned with the development of methods and techniques for making sense of data. The basic problem addressed by the KDD process is one of mapping low-level data (which are typically too voluminous to understand and digest easily) into other forms that might be more compact (for example, a short report), more abstract (for example, a descriptive approximation or model of the process that generated the data), or more useful (for example, a predictive model for estimating the value of future cases). At the core of the process is the application of specific data-mining methods for pattern discovery and extraction.

Data Mining is basically used today by most of the companies with a very strong consumer focus retail, financial, communication, and marketing organizations, to drill down into their transactional data and determine pricing, client preferences and product related information, impact on the sales, client satisfaction and corporate profits. With the help of data mining, a retailer can use point-of-sale records of client purchases to develop products and promotions to



appeal to specific client segments.

**Knowledge discovery** and data mining have become areas of growing significance because of the recent increasing demand for KDD techniques, including those used in knowledge acquisition, machine learning, databases, statistics, data visualization, and high performance computing. Knowledge discovery and data mining can be very useful for the field of Artificial Intelligence in many areas, for example education, industry, commerce, government, and so on. The relation between Knowledge and Data Mining, and Knowledge Discovery in Database (KDD) process are presented in the paper. Data mining theory, Data mining tasks, Data Mining technology and Data Mining challenges are also proposed.

#### **Disadvantages of Existing Model :**

- Excessive work intensity may require investment in high performance teams and staff training.
- The difficulty of collecting data. Depending on the type of data that you want to collect can be a lot of work.
- The Existing Model is totally based upon the customers past behavior, and determine the future event totally based upon that.
- Now the main problem in this Model is that if Customer changes its taste and not buying the things from the past event in that case it won't be able to give accurate results.

## Implimentation or Architecture Diagram

### Case Study of Wal Mart :

Wal-Mart is often described as a pioneering leader in data mining and data management:

Wal-Mart captures point-of-sale transactions from over 2,900 stores in six countries and continuously transmits this data to its massive 7.5 terabyte data warehouse.

Wal-Mart allows more than 3,500 suppliers to access data on their products and perform data analyses. These suppliers use this data to identify customer-buying patterns at the store display level. They use this information to manage local store inventory and identify new merchandising opportunities.

Other companies supplement their customers' transactional information with external data such as postal codes to do a market basket analysis. Practically every retailer now records all the details of each POS (Point of Sale) transaction for stock keeping purposes. Sometimes these are supplemented by customer information. *Home Depot*, for example supplements the data with ZIP or postal code of the purchaser. Sometimes the cashier may also enter the sex and appropriate age of the customer into the cash register. Affinity cards and credit card numbers can be used to track repeat customers.

## **DATA SETS**

### **ONLINE SHOPPING DATA SET :**

The Dataset used is obtained from highly reputational online shopping agency which sells only online.

The dataset is composed of online ordering log file for three months. The dataset consists of 303 instances and 26 attributes

All data sets are given in the Table Below:-

Attribute Name	Description
Personal information	Include serial number, buyer name ,gender ,age
Educational level	Describes buyer educational level and it is classified into categories from (1 -10)  (1-3) Graduated,(4-6) Master,(7-10) PHD.
Brand	Describes product brand name.
Product name	Describes the product name.
Item description	Describes the product specification.
Category	Describes product Category.
Quantity	Describes product ordered quantity per order.
Price	Describes product price.
Item Type	Describes the product different types.
Payment Method	Describes order payment method which is classified here into three methods (COD):cash on delivery ,credit card, buyer web site account.

Number of visits	Describes buyer visit number for the web site page.
duration of visit	describes buyer duration visit and it is measured by minutes.
<b>Rating</b>	describes product rating from the buyer and it's measured by scale from(1-5) (1)represent poor and (5) represent excellent.
User Satisfaction of the product	describes user satisfaction from the product and it is rated from(1 -100) .100 represent highly satisfied and I represent riot satisfied.
Best deal	Describes the best offer for the product I

---

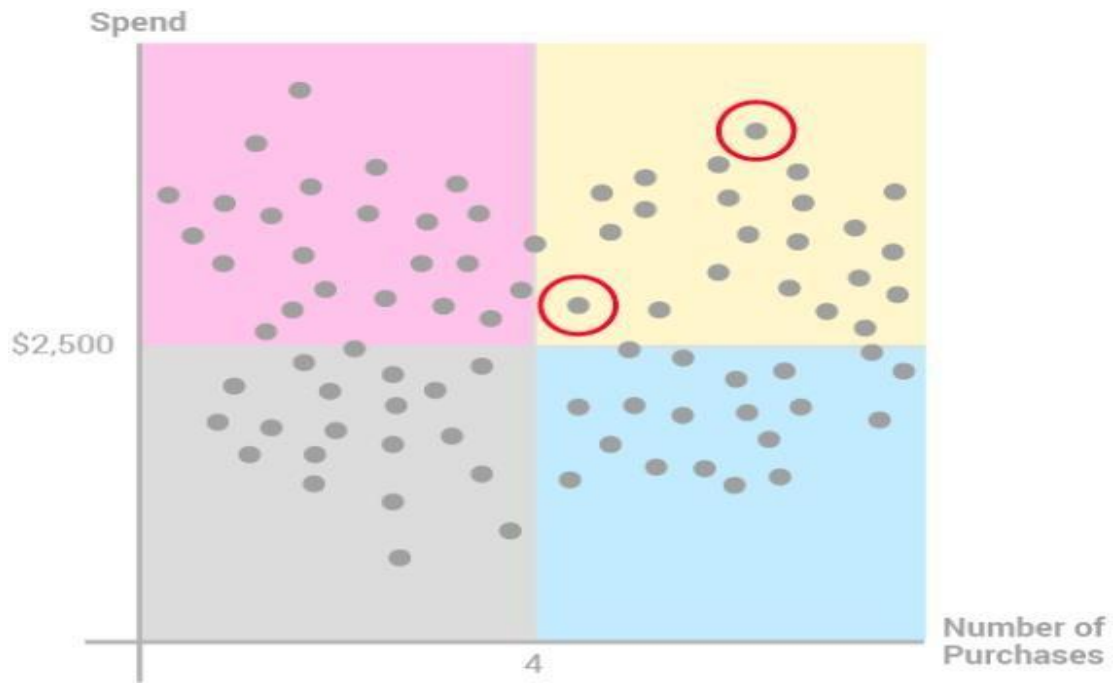
The test procedure is repeated 10 times .The final accuracy of an algorithm will be the average of the 10 trials.

### **Algorithm Sampling:-**

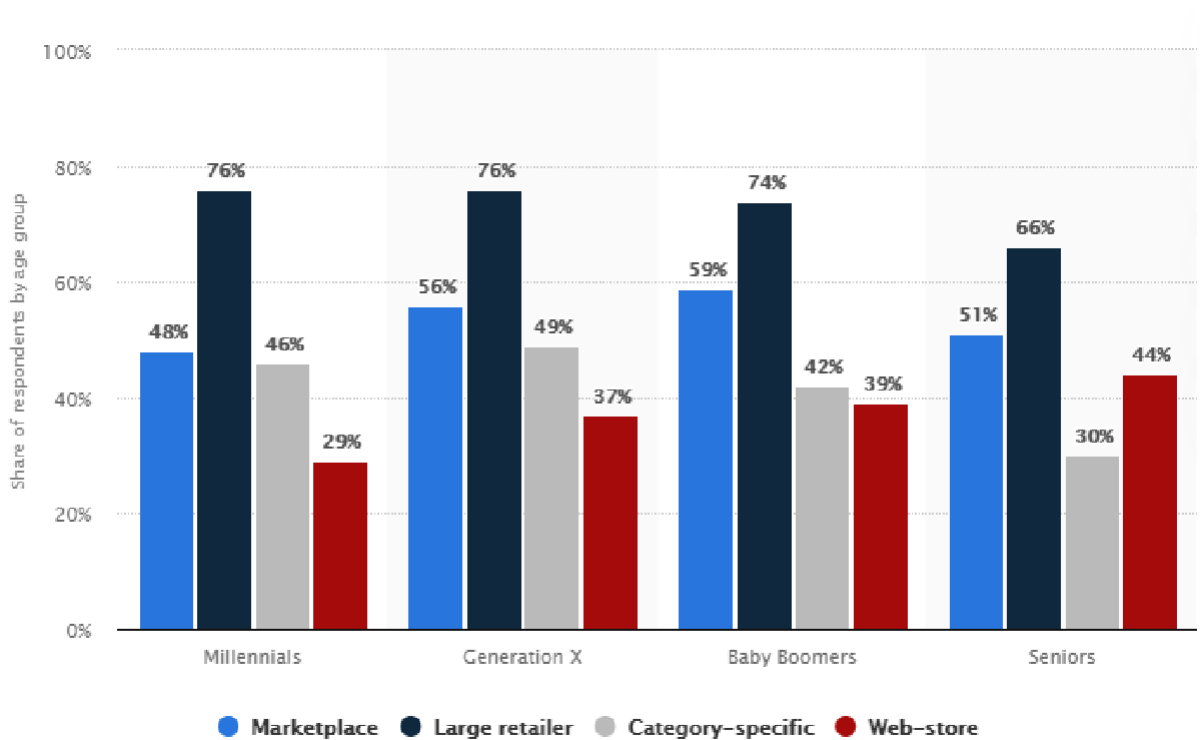
### **Classification Analysis:**

In the context of customer segmentation, cluster analysis is the use of a mathematical model to discover groups of similar customers based on finding the smallest variations among customers within each group. These homogeneous groups are known as “customer archetypes” or “personas”. The goal of cluster analysis in marketing is to accurately segment customers in order to achieve more effective customer marketing via personalization. A common cluster analysis method is a mathematical algorithm known as *k-means cluster analysis*, sometimes referred to as scientific segmentation. The clusters that result assist in better customer modeling and predictive analytics, and are also used to target customers with offers and incentives personalized to their wants, needs and preferences.

The process is not based on any predetermined thresholds or rules. Rather, the data itself reveals the customer prototypes that inherently exist within the population of customers.



**Age based Classification**



**Client based Classification**

		R	F	M
Customers		Days Since Last Purchase	Number of Purchases (Past 12 Months)	Net Revenue (Past 12 Months)
High Spenders	922	9	4	\$154
Mid Spenders	581	54	3	\$121
Risk of Churn	807	192	2	\$70
Low Spenders	1,361	192	2	\$4
	<b>3,671</b>	<b>447</b>	<b>3</b>	<b>\$87</b>

## Output / Results

### Types of Outputs

Generally , applications of data mining can generate outputs such as:

- Buying patterns of customers; associations among customer demographic characteristics;
- Predictions on which customers will respond to which mailings;
- Patterns of fraudulent credit card usage
- identities of “loyal” customers; credit card
- spending by customer groups;
- Predictions of customers who are likely to change their credit card affiliation;
- Predictions on which customers will buy new insurance policies; behavior patterns of risky customers;
- Expectations of fraudulent behavior;
- Characterizations of patient behavior to predict frequency of office visits.

The Result investigate the performance of selected classification algorithms that was previously described by using WEKA. Figures represent snap in for the result is composed of two sectors **WEKA explorer tool** interface that include classification algorithms analysis and results and WEKA knowledge flow interface that include the knowledge flow charts analysis for each algorithm respectively.



WEKA Explorer Tools results are given below:-

Classifier

Choose **KStar -B 20 -M a**

Test options

Use training set

Supplied test set

Cross-validation Folds

Percentage split %

(Nom) class

Result list (right-click for options)

17:06:08 - lazy.KStar

Classifier output

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	154	50.8251 %
Incorrectly Classified Instances	149	49.1749 %
Kappa statistic	0.2371	
Mean absolute error	0.2499	
Root mean squared error	0.4696	
Relative absolute error	74.9141 %	
Root relative squared error	115.0949 %	
Total Number of Instances	303	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC
	0.29	0.149	0.333	0.29	0.31	
	0.386	0.133	0.466	0.386	0.422	
	0.719	0.408	0.621	0.719	0.667	
	0.16	0.065	0.182	0.16	0.17	
Weighted Avg.	0.508	0.263	0.49	0.508	0.496	

=== Confusion Matrix ===

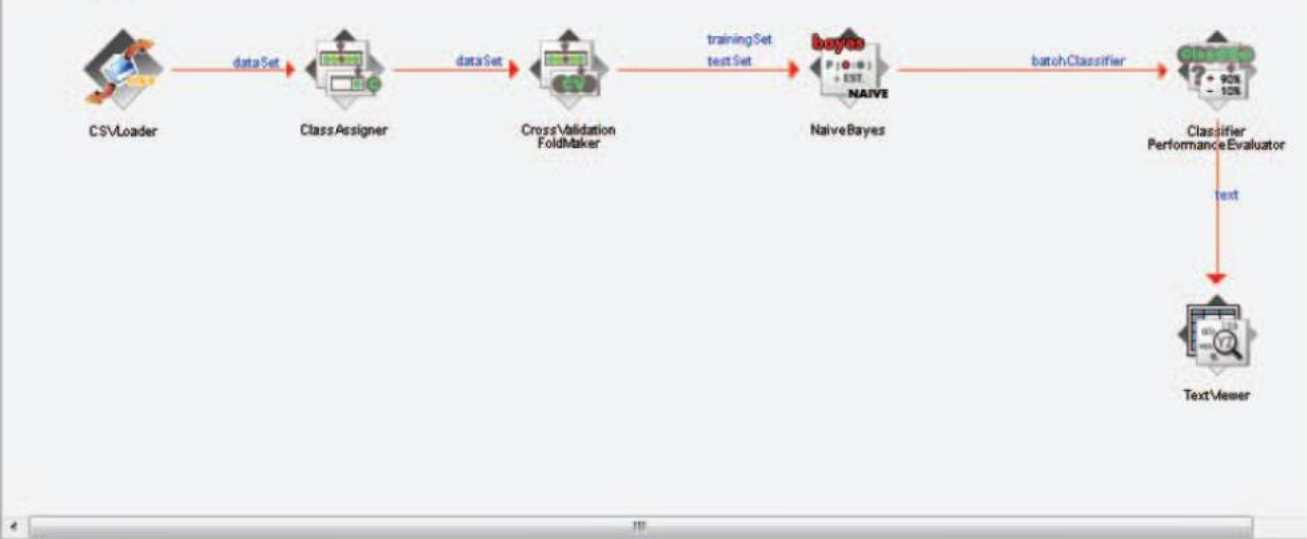
|||

Status  
OK

Visualization

Data Visualizer | Scatter PlotMatrix | Attribute Summarizer | Model PerformanceChart | CostBenefit Analysis | **Text Viewer** | Graph Viewer | Strip Chart

Knowledge Flow Layout



Status [Log](#)

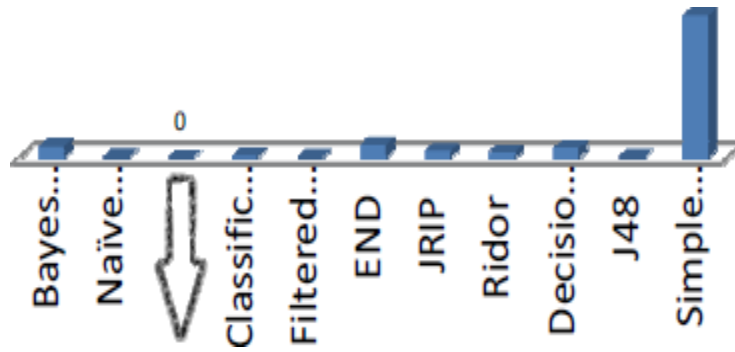
Component	Parameters	Time	Status
[KnowledgeFlow]		00:38:09	Welcome to the Weka Knowledge Flow

## CLASSIFIERS EXECUTION TIME

<b>Classifier Name</b>	<b>Time taken to build model per</b>
Classification	0 Second
Filtered Classifier	0.01 Second
J48	0.01 Second
Naïve Bayes	0.02 Second
Classification via Clustering	0.05 Second
Ridor	0.15 Second
JRIP	0.17 Second
Decision Table	0.24 Second
Bayes Net	0.28second
END	0.34 Second
Simple Cart	3.7 Second

## Classifiers Time taken to build model

- ❖ Time taken to build model per Second



Classification

## **Conclusion**

Data mining is neither a magic bullet nor a simple process and as such presents challenges that go well beyond the technical. Data mining should be used with carefully outlined criteria, which should not be based on any kind of prejudice or stereotype assumptions. Otherwise the outcome of datamining will not give the desired result. Many data management challenges remain, both technical and societal. Large online databases raise serious societal issues. To cite a few of the societal issues: Electronic data interchange and data mining software make it relatively easy for a large organization to track all of your financial transactions.

By doing that, someone can build a very detailed profile of your interests, travel, and finances. Is this an invasion of your privacy? Indeed, it is possible to do this for almost anyone in the developed world.

Data mining is a very powerful tool that should be used with utmost care for increasing customer satisfaction, providing best, safe and useful products at reasonable and economical prices. This should be used for making the business more competitive and profitable. Data mining should be used in any way that affects the privacy of common man, so that the confidentiality and individuality of human being is preserved. It should not be used in any way that may cause undue hardship, financial loss or emotional setback.

## References

1. Nick Wreden, *Communications Week Interactive*, February 17, 1997.
2. Karen Watterson, *Datamining poised to go mainstream*, October 1999.
3. Information and Privacy Commissioner/Ontario, *Data Mining: Staking a Claim on Your Privacy*, January 1998.
4. Author unknown *Data Mining: What is Data Mining?*
5. Pilot Software, *Data Mining White Paper — Profitabl Applications* found at, [www.pilotsw.com/dmpaper/dmindex.htm#dmapp](http://www.pilotsw.com/dmpaper/dmindex.htm#dmapp).
6. Jim Gray, *Data Management: Past, Present, and Future*, at [www.research.microsoft.com/~gray/DB\\_History.htm](http://www.research.microsoft.com/~gray/DB_History.htm).
7. Jonathan Wu, *Business Intelligence: The Value in Mining Data*, *DM Review online*, February, 2002
8. Richard A. Spinello, *Case Studies in Information and Computer Ethics*. New Jersey: Prentice Hall, 1997, p. 73
9. Anne Wells Branscomb, *Who Owns Information: From Privacy to Public Access*. New York: Basic Books, 1994, p.
10. Zak Pines, *Data Mining – A universal Tool*, from [www.hpcwire.com/dsstar](http://www.hpcwire.com/dsstar).



