INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078

(Volume 7, Issue 2 - V7I2-1295)
Available online at: https://www.ijariit.com

Trends and application in data mining

Manish Choudhary
choudharymanish46969@gmail.com
Mukesh Patel School of Technology Management and Engineering, Mumbai, Maharashtra

Sagar Patil
patilsagarp2001@gmail.com
Mukesh Patel School of Technology Management and Engineering, Mumbai, Maharashtra

Mehar Singh
meharsingh0312@gmail.com
Mukesh Patel School of Technology Management and Engineering, Mumbai, Maharashtra

ABSTRACT

In our day to day life, many Terabytes of data are generated by various organizations around the world, as data has been increasing each day in Tremendous amount there is a need for some tools and techniques which may be used and hence help humans to analyze large data repositories and gain some useful information. This need which keeps on increasing on daily basis gives rise to an area known as Data Mining or Knowledge Discovery of Databases. The main aim of Data Mining is to extract some useful unknown patterns or knowledge or information from the vast amount of data repositories. In this automated world, many organizations uses Data Mining as it enhances various fields of human life including education, agriculture, business, etc. using Artificial Intelligence, Machine Learning, Pattern Recognition, and various data visualization techniques. This paper discusses and describes Data Mining and how data mining is performed for Retail Industries, the paper also describes how data mining can be used for Fraud Detection.

Keywords: *Data Mining, Trends in Data Mining, DM, Data Warehouse, Application in DM*

1. INTRODUCTION

With time, information is growing rapidly, huge chunks of data are available with every organization. Data is collected and stored on a daily basis. Data is just data and is of no use until it is transformed into knowledge or useful information. To extract information from them, analysis needs to be performed. This is where data mining comes into role. It helps extract information or transform this raw data into knowledge.

Data mining involves many stages like Data Cleaning, Integration, Data Transformation, techniques and tools for Data Mining, Pattern Evaluation and Data Presentation respectively. This is nothing but the knowledge discovery process. Once the processes are done, we apply it in various domains which we have discussed below. Some applications can be Fraud Detection, Data Mining in Healthcare, Market Basket Analysis, Production Control, Mining for Financial data, Science Exploration, etc.

Applications

Data Mining are widely used in which areas?-

1. Retail Industry
2. Financial Data Analysis
3. Fraud Detection
4. In is widely used in various Telecommunication Industry
5. For Biological Data Analysis
6. Data Mining is also widely used for many Other Scientific Applications
7. It is also used for Intrusion Detection

As seen, Data mining is used for various purposes in various domains, hence we can say that the applications are in diverse fields. Other than applications, we also focus on trends in data mining and the recent evolvements. Data mining concepts are still evolving and some latest trends in this field that we get to see are:

1. Data mining and visualizations.
2. Data mining with complex data types.
3. Biological data mining.
4. Data mining and software engineering.
5. Web mining.
6. Data mining in distributed systems.
7. Data mining with Real-time data.
8. Application Exploration.
9. Scalable and interactive mining methods.
10. Data mining with data warehouses and multi-dimensional modelling, database systems, etc.
11. Multi database data mining.
12. Privacy protection in data mining.

2. APPLICATIONS

Data Mining for the Retail Industry

The retail industry is a major application area for data mining, since it collects huge amounts of data on sales, customer data

which includes their shopping history, their preferences, their consumption, and other services. This case-study focusses on the market-basket analysis which is used by retail industries to know the patterns in which customers buy products together and what type of products are bought frequently. Every enterprise comes across millions of transactional data every day. All the transaction information is nothing but raw data, which needs to be converted into useful information. Hence it has become important for the businesses to learn more about their customers and develop more effective marketing strategies as well as increase sales and decrease costs. This is exactly what data mining does for the business. It tracks who is buying what.

Apriori algorithm is the most widely used and a classic algorithm for learning association rules. It is designed in a way to work on databases containing transactions (transactions can be a list of the items the customer has brought. One transaction is one row where the columns are nothing but the items the customer has purchased when visited the supermarket).

The main purpose for using Apriori Algorithm is to find different rules and association between many different sets of data provided. Apriori Algorithm is also referred as "Market Basket Analysis". Using this algorithm we get the sets of rules which gives us information regarding how often item is there in a given data sets.

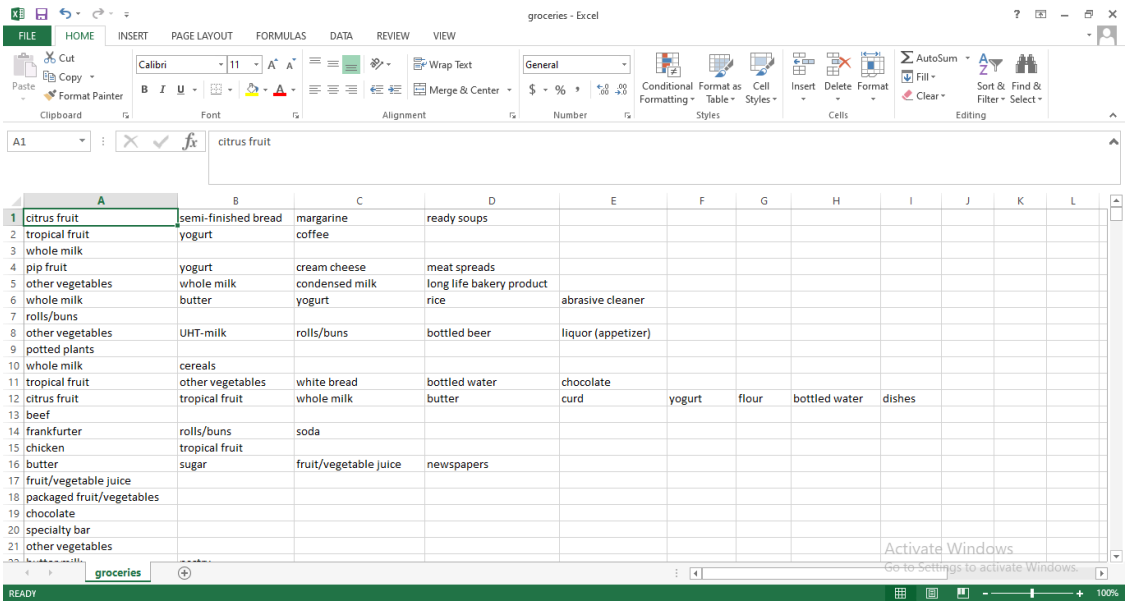
This is retailer's favorite technique for improving sales and customer base. It influences many aspects of a retailer including store layouts, cross selling, upselling, etc.

Eq:
When we visit supermarkets, we notice that some items are always bundled together though we find no co-relation between them. One famous example is beer and diapers.

Or login to your Amazon account and you find the "frequently bought together" section which recommends you some products you should buy. This depends on not just your purchase history but also the purchase history of the customer's who have bought the same item as yours.

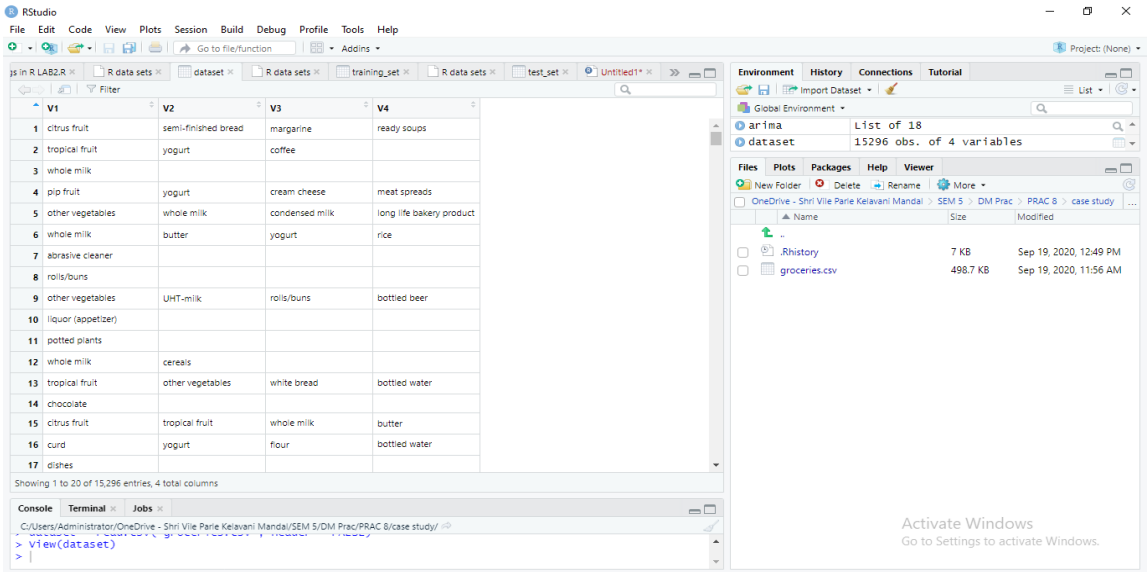
Now for us there is definitely no correlation between beer and diapers but retailers did notice that customers who are buying beer may have already bought diapers or are more likely to buy it. The example listed above is a perfect example of Association Rules in Data Mining. It helps us understand the concept of Apriori algorithms (used majorly for association rules in DM). It helps the customers buy their items with ease, and enhances the sales performance of the departmental store. **This is targeted marketing.**

3. DATASET



	A	B	C	D	E	F	G	H	I
1	citrus fruit	semi-finished bread	margarine	ready soups					
2	tropical fruit	yogurt	coffee						
3	whole milk								
4	pip fruit	yogurt	cream cheese	meat spreads					
5	other vegetables	whole milk	condensed milk	long life bakery product					
6	whole milk	butter	yogurt	rice	abrasive cleaner				
7	rolls/buns								
8	other vegetables	UHT-milk	rolls/buns	bottled beer	liquor (appetizer)				
9	potted plants								
10	whole milk	cereals							
11	tropical fruit	other vegetables	white bread	bottled water	chocolate				
12	citrus fruit	tropical fruit	whole milk	butter	curd	yogurt	flour	bottled water	dishes
13	beef								
14	frankfurter	rolls/buns	soda						
15	chicken	tropical fruit							
16	butter	sugar	fruit/vegetable juice	newspapers					
17	fruit/vegetable juice								
18	packaged fruit/vegetables								
19	chocolate								
20	specialty bar								
21	other vegetables								

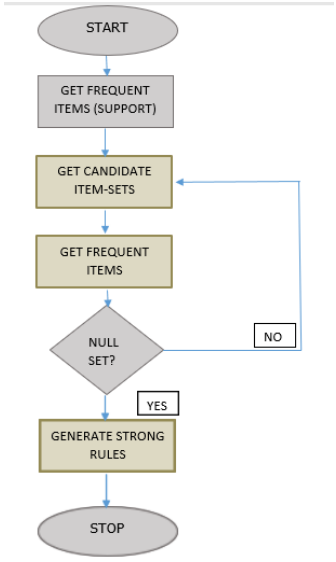
IN R:



Associated Attributes

The attributes for this dataset are the items bought by the customers in the supermarket.

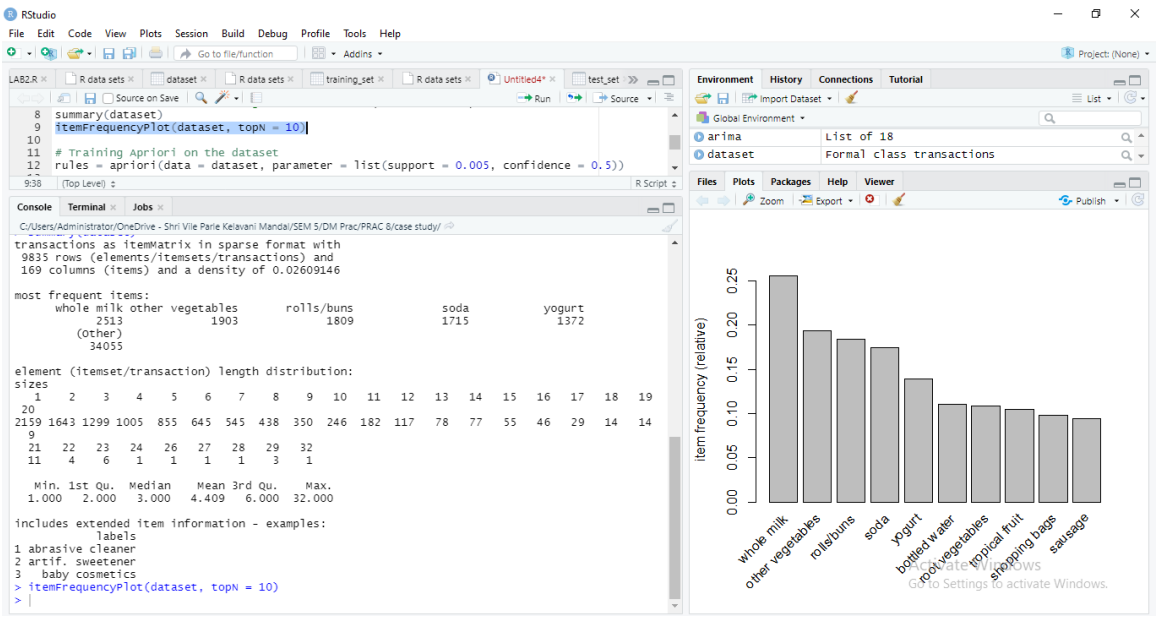
Flow Diagram



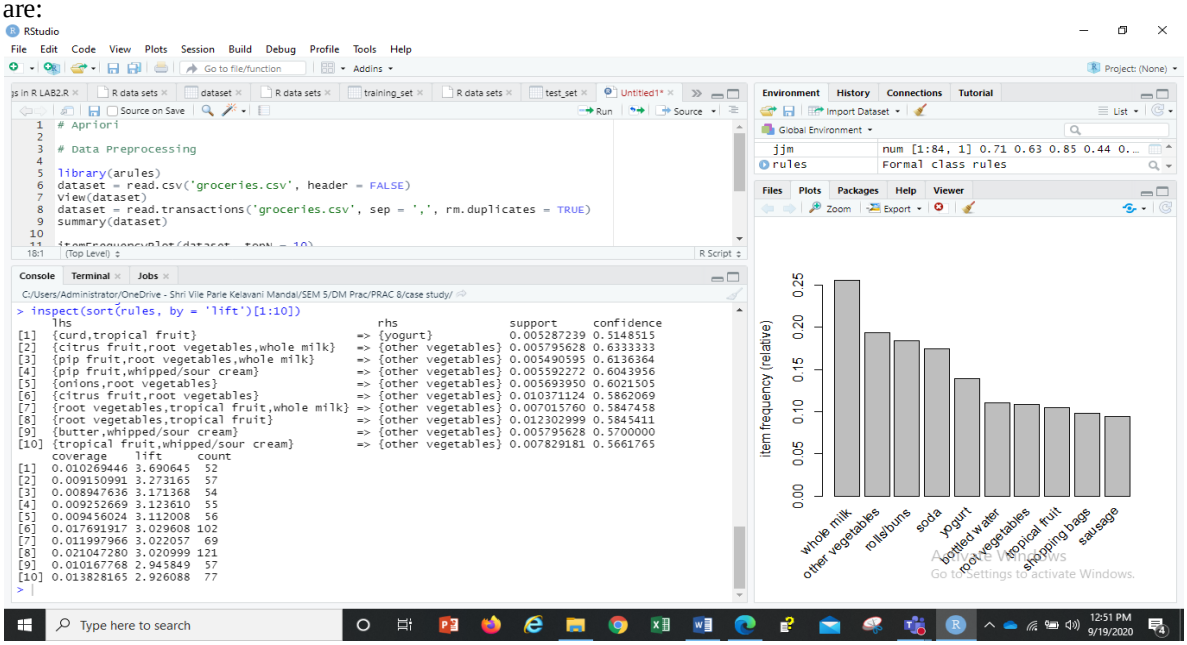
Input and Output

This is the summary of that dataset, with the top 10 products (sa le during pandemic) is as
Shown:

Whole milk, other vegetables, rolls/buns, soda, yogurt, etc. w ith the output
Frequency:



The **support = 0.005**, **confidence = 0.5,** is considered in this code.
The rules are:



Conclusion drawn

As executed,

{Curd, tropical fruit} => {yogurt} is the first set as seen. Mainly customers who bought Curd and Fruits, also bought Yogurt which is obvious in situations like now. Most customers who are going for healthy items, go for these combinations. The super-market can have offers on fruits and dairy products together, or packages of fruits and vegetables (offers on fruits and veggies together), or that of all the three (fruits, veggies, dairy products). These items can be sold together to increase sales.

This also helps for identify patterns within customers.

E.g. here:

We can see customers who are more health conscious, buy items like curd, tropical fruits, and vegetables together. We can also know how frequent these items are bought by the respective customers. To increase the sales two things can be done:

- Either put these products together (same or nearby shelves), so the customer doesn't have to go searching for them.
- Or these items can be put in faraway shelves so the customers eventually take a look around the shop and maybe they will come across some other products which they wouldn't buy otherwise.

Sentiment analysis –Natural Language Processing

Natural Language Processing (or NLP) is nothing but using machine learning for text. We humans can process text or interpret language. When someone says "Hi" to us, we understand what they mean and give our reply "Hello". This is how conversations get going. But you never say "Hi" to a machine right? Because machines cannot interpret language or text. They don't understand "Hi" or "Hello". This is when Natural Language Processing comes into picture. This trains the machine to process text and gain insights from it. Whenever you say something to your smart phone, it is converted into text. This is possible only with an NLP algorithm. The perfect example for NLP is your Siri.

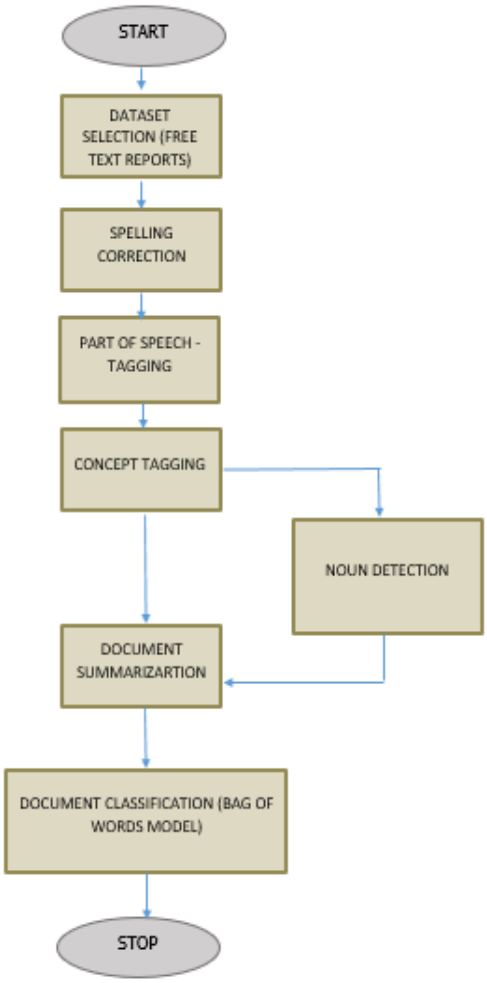
NLP has many applications as text processing nowadays has become necessary with the huge information companies are gathering.

This is used by businesses for sentiment analysis of their customers. How will the business know that they are preferred by their customers? How many customers are liking your products? How many reviews are negative for your product? All these can be answered on the basis of the feedback provided by your customers. Every business comes across millions of customer feedback or reviews which cannot be read and analyzed by employees one by one, and this is why we use NLP. We feed the customer reviews to the model and it gives you an analysis on the number of positive reviews and the number of negative reviews.

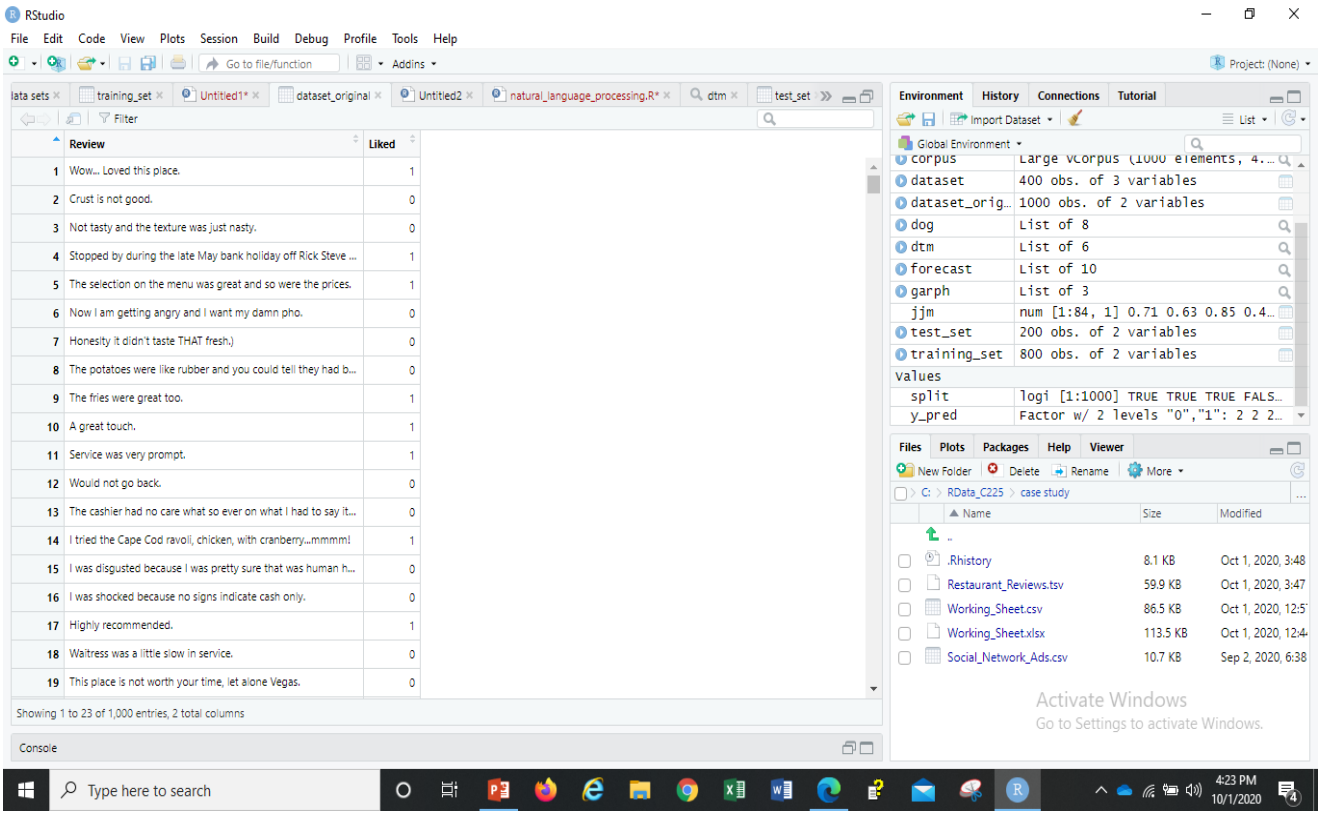
NLP can also be applied on books and helps in genre prediction of the book. Machine translator, Text recognition, Speech recognition systems, Classification (classify genre or languages) are some famous NLP applications.

Some most widely used with NLP algorithms for classification are Logistic Regression, Decision trees CART model, Naive Bayes, Hidden Markov Models, etc. A very well-known model in NLP is the Bag of Words model. It is a model used to preprocess the texts to classify before fitting the classification algorithms on the observations containing the texts.

Flow Diagram



Dataset



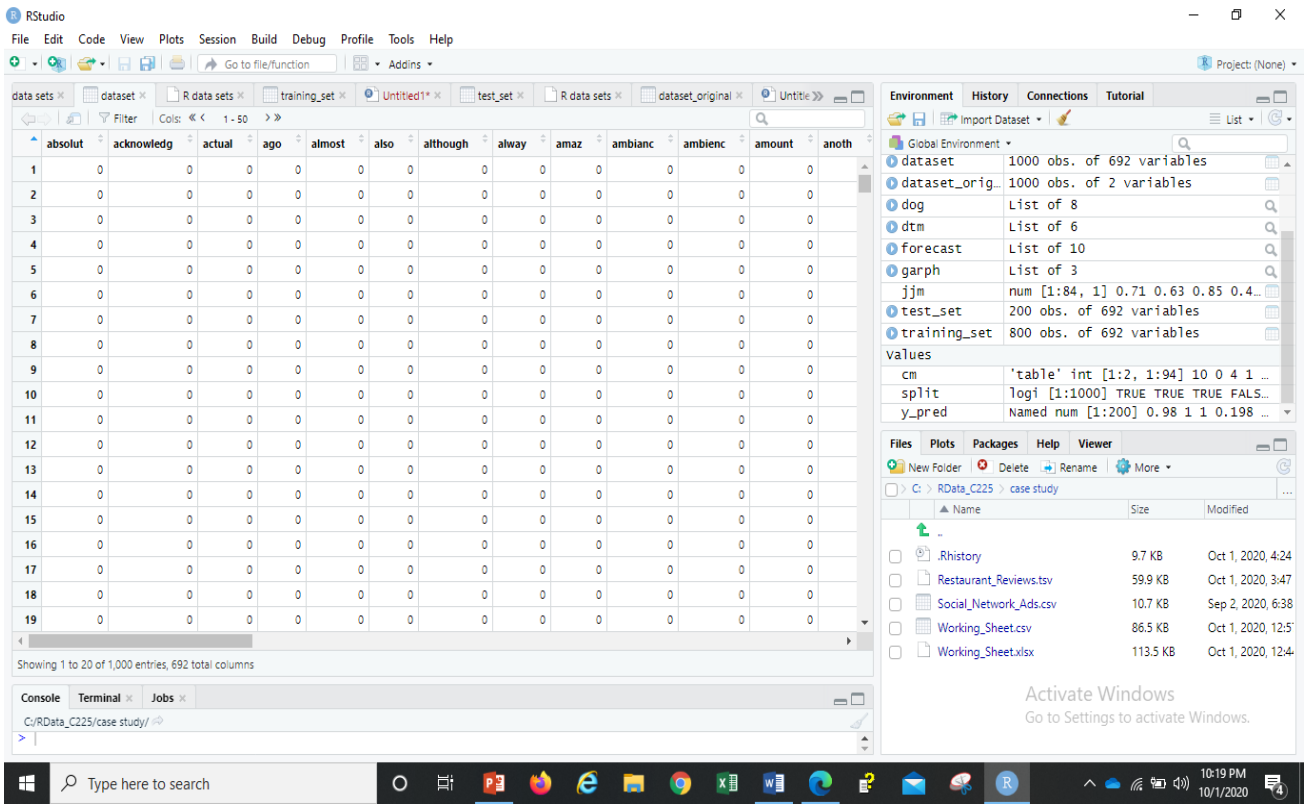
Bag of Words In Matrix Form

This is just a simplified representation. We use it to represent text data for our NLP algorithm. This tells us about the words in the inputted document/ text. When our NLP model is inputted with the dataset, it simplifies it and every word with its frequency is inserted in the bag of words model. This helps extract features from the text.

Each sentence is taken one at a time as treated as a document. All words in the given document will be then separated (there are rules for this too, like punctuation is ignored). Finally we get a "Bag of words" which contains all the words present in the document.

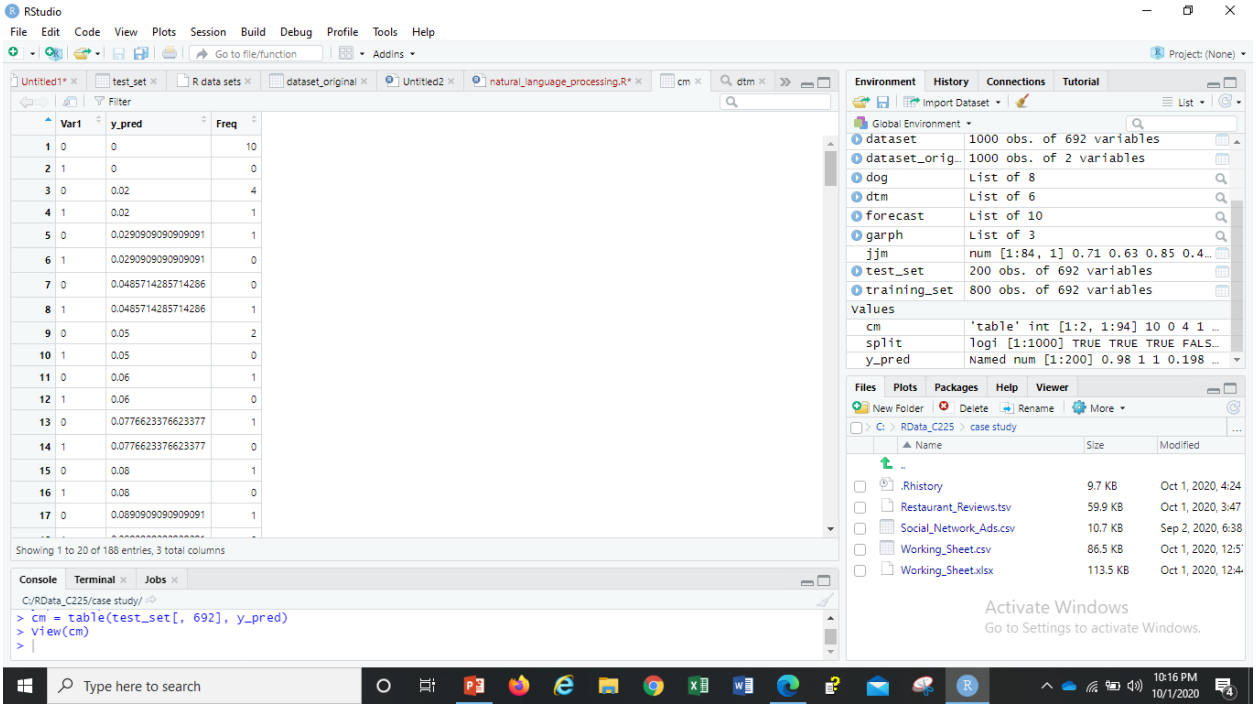
As seen in the image, the bag of words contains all the words starting with "absolute" and each entry tells whether the word is present in that transaction (or sentence) or not. If the word is present in the sentence, the entry is 1 and 0 otherwise.

This process of converting NLP text into numbers (which tell the frequency of the occurrence of words) is called **vectorization**.



4. RESULTS

Confusion Matrix



Fraud Detection

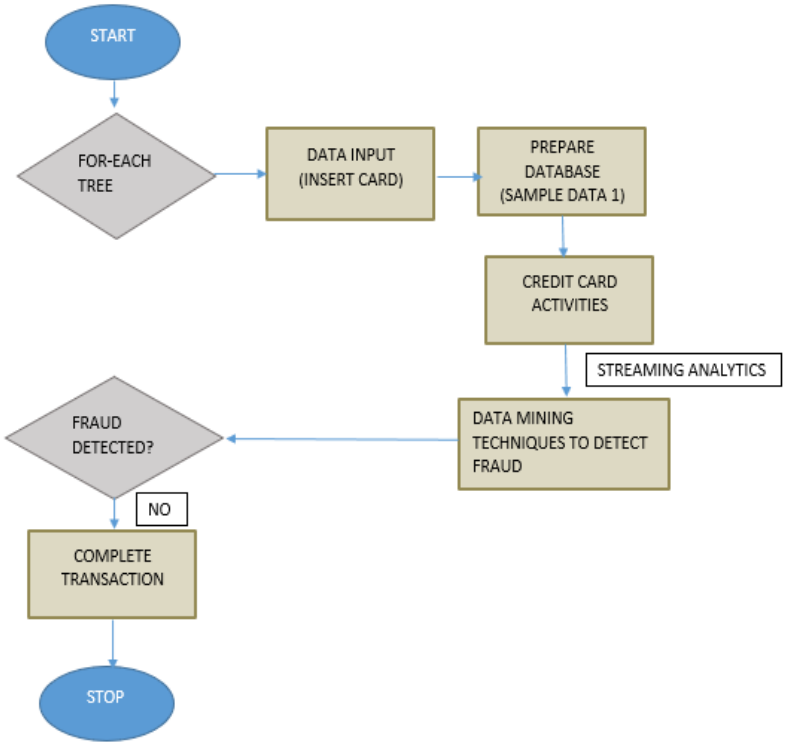
We live in a world where large numbers of transaction occur daily. So we use the technology to handel tha data. But the criminal also become more smarter. So artificial intelligence is used to reduce fraud. Fraud detection is a activity where we can prevent data from the criminals.The Fraud can be detected in many ways like asking for identification, number like OTP ,photos, etc.

Here the main reason of using data mining techniques is to detect and tackle fraud. The data analytics techniques like AL, data mining efficiently test our system to detect fraud.

Prevention from fraud:

1. Be ensure that a single system is used by single person.
2. Recheck your transaction regular.
3. Don't use private information in public places.

Flow-Diagram



Data Mining for the Infrastructure Industry – Soil-CBR Prediction

California Bearing Ratio or CBR test is the ratio of the force per unit area required to penetrate a soil mass with a standard circular piston of 50 mm diameter, at the rate of 1.25 mm/min to that of force required to penetrate sample of compacted stone having CBR of 100%.

This depends on many factors, especially the sand type and the liquid limit of it.

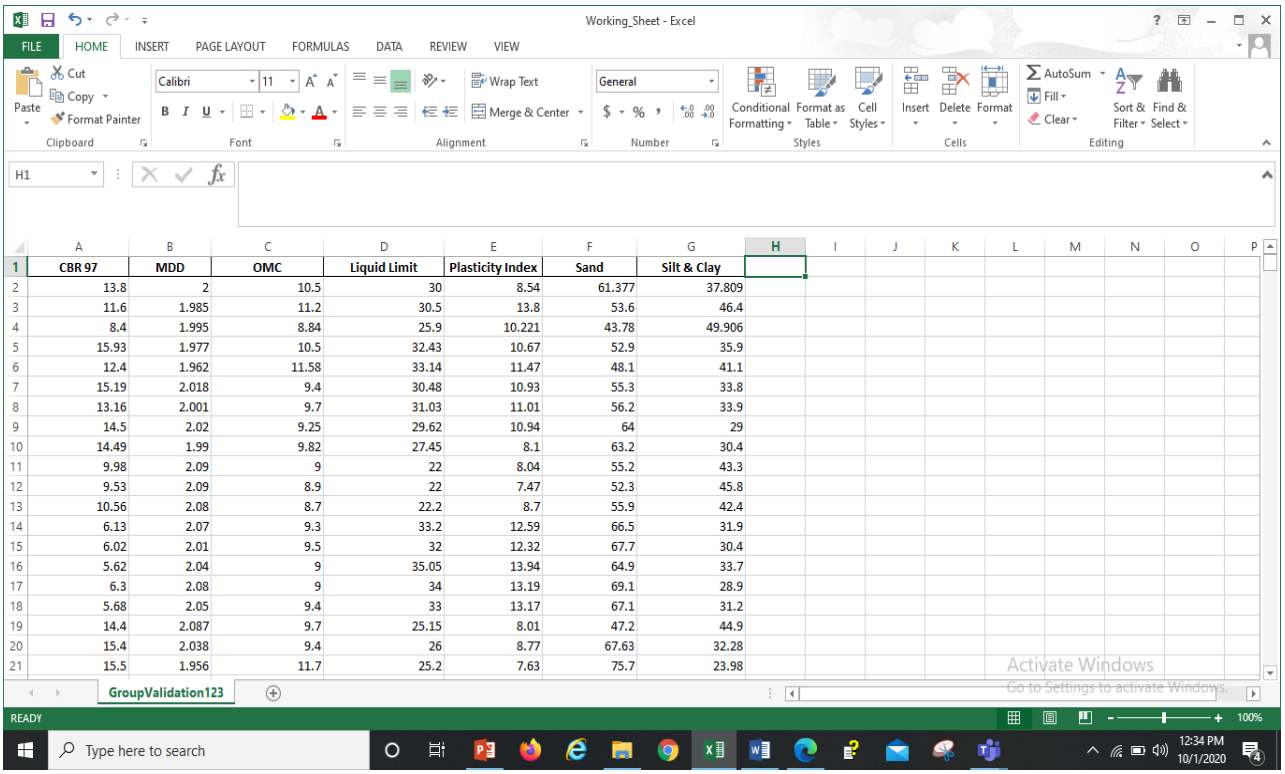
This technique is actually used to classify the soil-grade or to evaluate it. This helps determine the materials used in flexible pavements (coarse materials).

This is important when it comes infrastructural projects taken by companies like railways, highways, ports (sea), etc. Evaluation of the soil characterization is important before building the structures because this affects the process as well as the structure. Before construction the soil properties will help determine the problems which can occur during construction as well as the time the project will take.

Associated Attributes

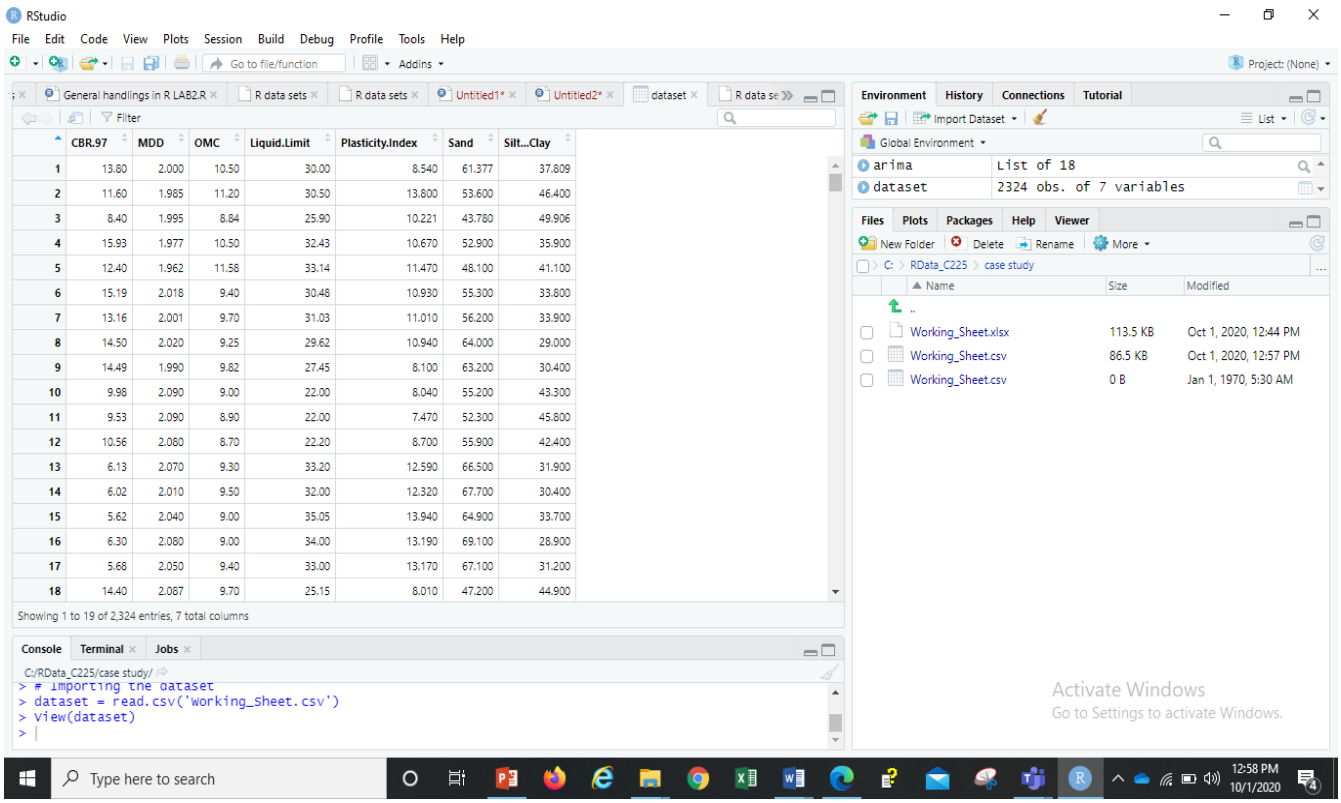
The attributes here are MDD, OMC, Liquid Limit, Plasticity Index, Sand, Silt and Clay, etc.

Here the attribute we are predicting is CBR (The **dependent variable**) The **independent variables** are MDD, OMC, Liquid Limit, Plasticity Index, Sand, Silt and Clay.



	A	B	C	D	E	F	G	H
1	CBR 97	MDD	OMC	Liquid Limit	Plasticity Index	Sand	Silt & Clay	
2	13.8	2	10.5	30	8.54	61.377	37.809	
3	11.6	1.985	11.2	30.5	13.8	53.6	46.4	
4	8.4	1.995	8.84	25.9	10.221	43.78	49.906	
5	15.93	1.977	10.5	32.43	10.67	52.9	35.9	
6	12.4	1.962	11.58	33.14	11.47	48.1	41.1	
7	15.19	2.018	9.4	30.48	10.93	55.3	33.8	
8	13.16	2.001	9.7	31.03	11.01	56.2	33.9	
9	14.5	2.02	9.25	29.62	10.94	64	29	
10	14.49	1.99	9.82	27.45	8.1	63.2	30.4	
11	9.98	2.09	9	22	8.04	55.2	43.3	
12	9.53	2.09	8.9	22	7.47	52.3	45.8	
13	10.56	2.08	8.7	22.2	8.7	55.9	42.4	
14	6.13	2.07	9.3	33.2	12.59	66.5	31.9	
15	6.02	2.01	9.5	32	12.32	67.7	30.4	
16	5.62	2.04	9	35.05	13.94	64.9	33.7	
17	6.3	2.08	9	34	13.19	69.1	28.9	
18	5.68	2.05	9.4	33	13.17	67.1	31.2	
19	14.4	2.087	9.7	25.15	8.01	47.2	44.9	
20	15.4	2.038	9.4	26	8.77	67.63	32.28	
21	15.5	1.956	11.7	25.2	7.63	75.7	23.98	

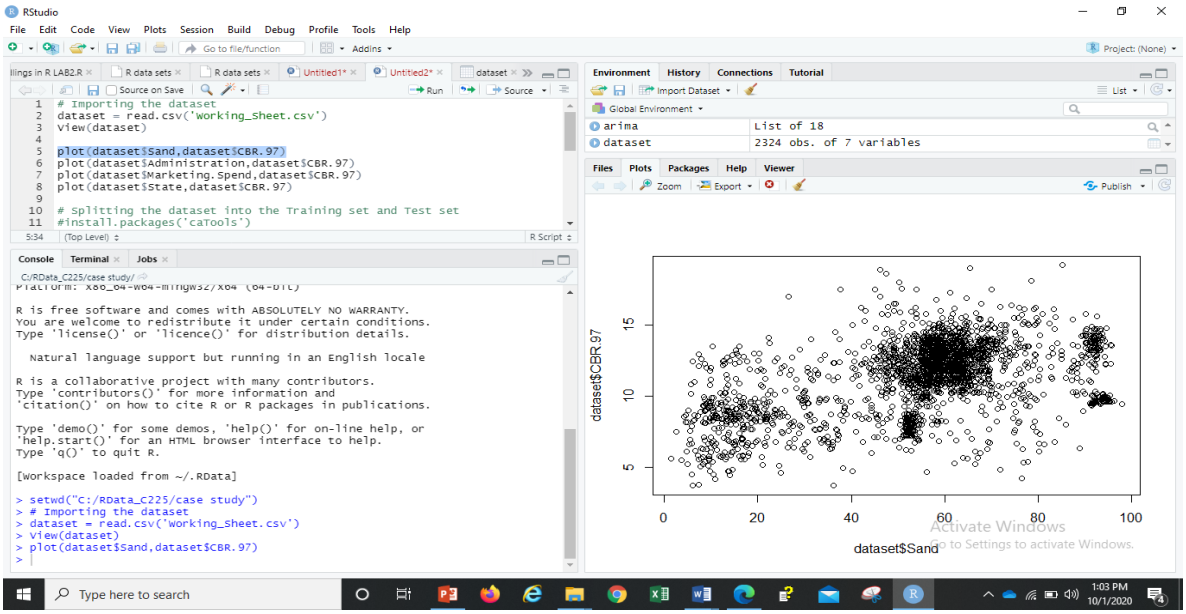
Regression analysis is used here. We use multiple regression because the independent variables/attributes are more than one.



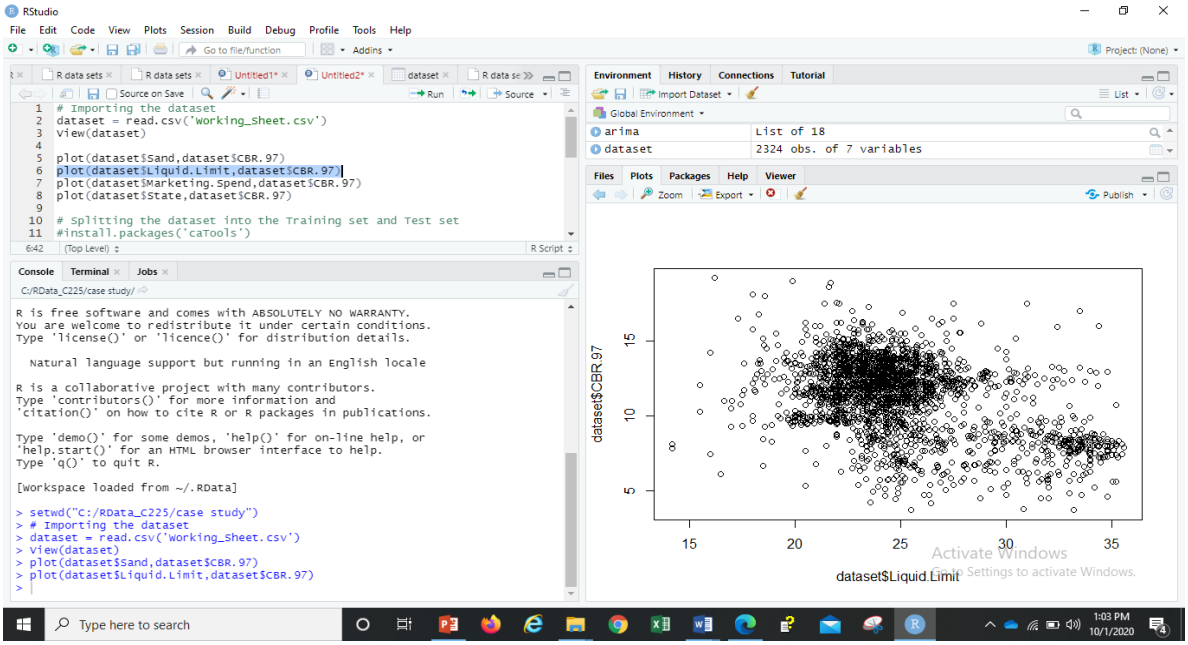
	CBR.97	MDD	OMC	Liquid.Limit	Plasticity.Index	Sand	Silt...Clay
1	13.80	2.000	10.50	30.00	8.540	61.377	37.809
2	11.60	1.985	11.20	30.50	13.800	53.600	46.400
3	8.40	1.995	8.84	25.90	10.221	43.780	49.906
4	15.93	1.977	10.50	32.43	10.670	52.900	35.900
5	12.40	1.962	11.58	33.14	11.470	48.100	41.100
6	15.19	2.018	9.40	30.48	10.930	55.300	33.800
7	13.16	2.001	9.70	31.03	11.010	56.200	33.900
8	14.50	2.020	9.25	29.62	10.940	64.000	29.000
9	14.49	1.990	9.82	27.45	8.100	63.200	30.400
10	9.98	2.090	9.00	22.00	8.040	55.200	43.300
11	9.53	2.090	8.90	22.00	7.470	52.300	45.800
12	10.56	2.080	8.70	22.20	8.700	55.900	42.400
13	6.13	2.070	9.30	33.20	12.590	66.500	31.900
14	6.02	2.010	9.50	32.00	12.320	67.700	30.400
15	5.62	2.040	9.00	35.05	13.940	64.900	33.700
16	6.30	2.080	9.00	34.00	13.190	69.100	28.900
17	5.68	2.050	9.40	33.00	13.170	67.100	31.200
18	14.40	2.087	9.70	25.15	8.010	47.200	44.900

5. RELATIONSHIP BETWEEN SOIL CBR AND OTHER FACTORS

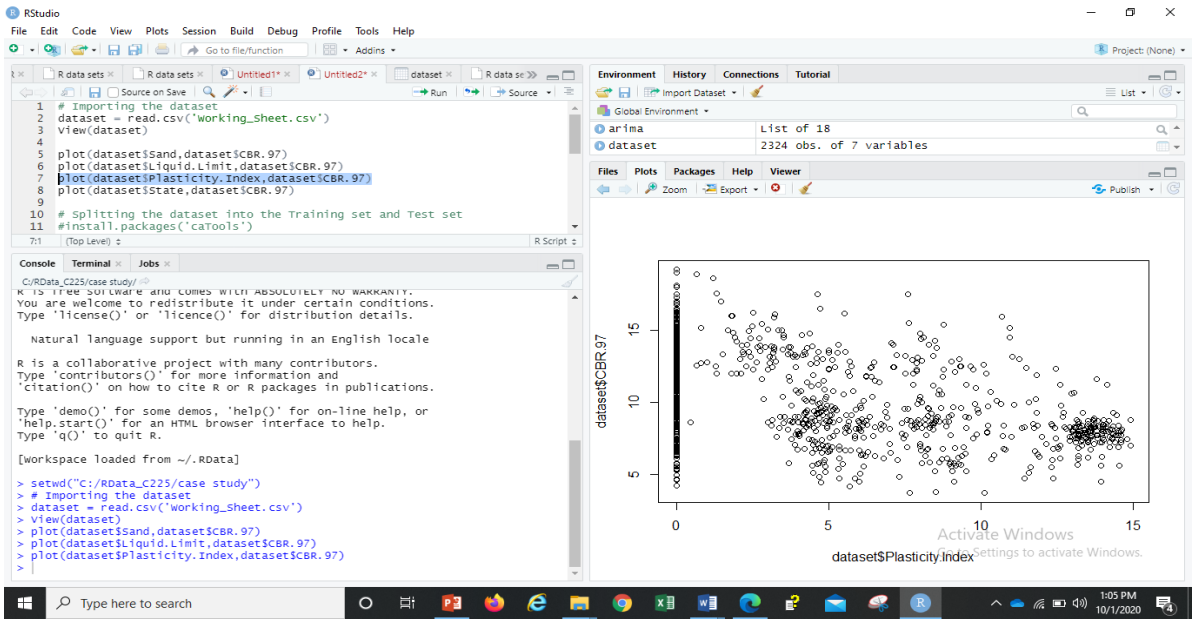
Sand and **CBR**



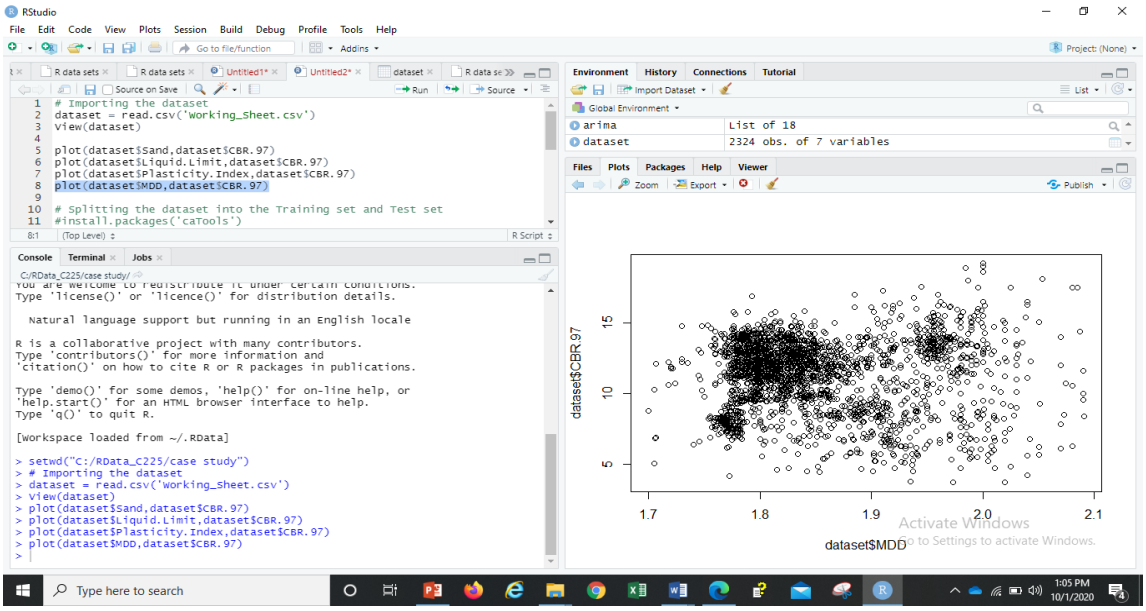
Liquid Limit and **CBR**



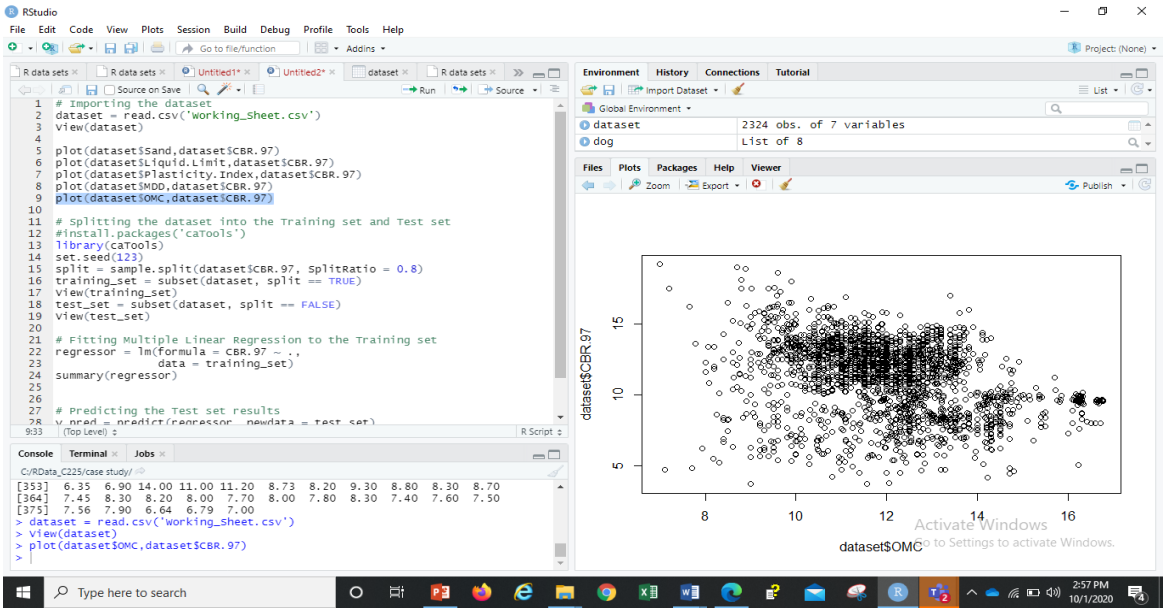
Plasticity index and **CBR**



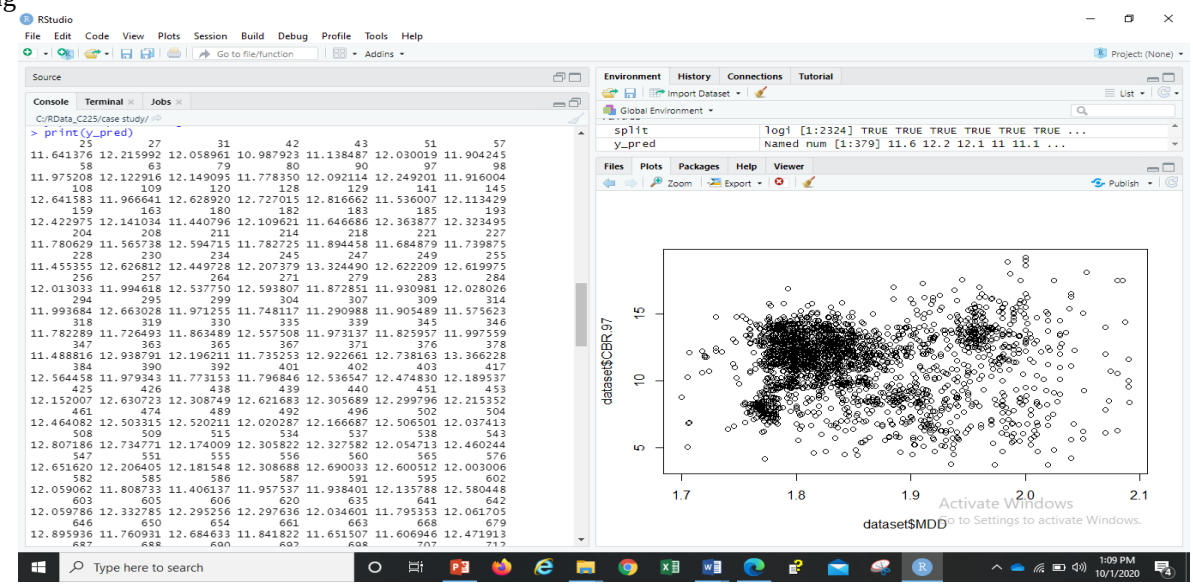
MDD and **CBR**



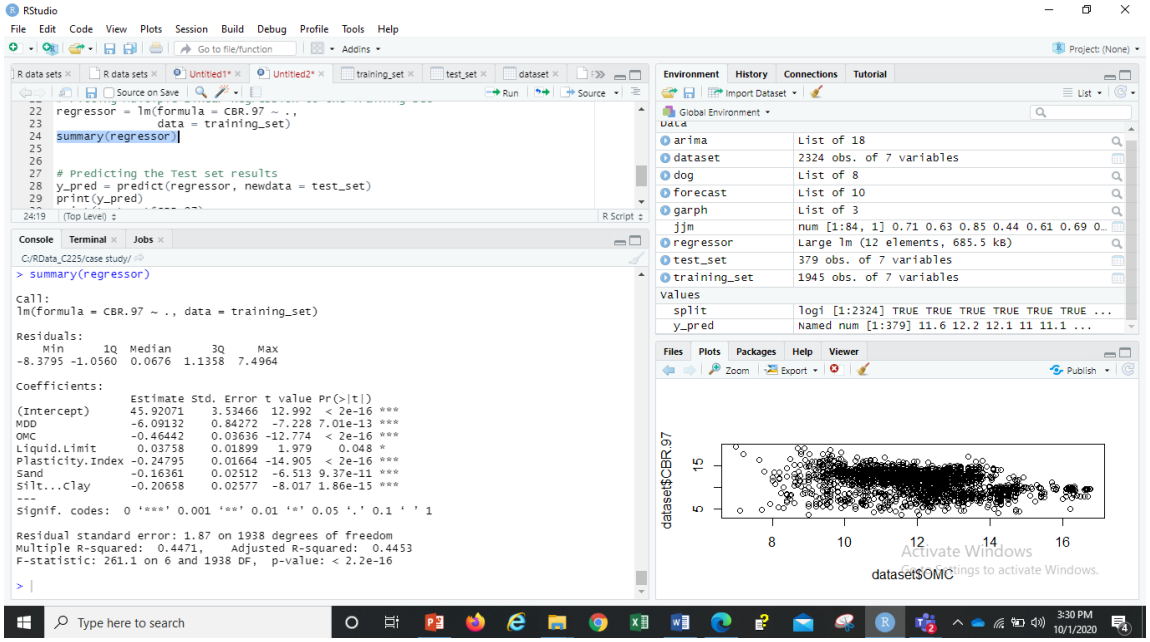
OMC and **CBR**

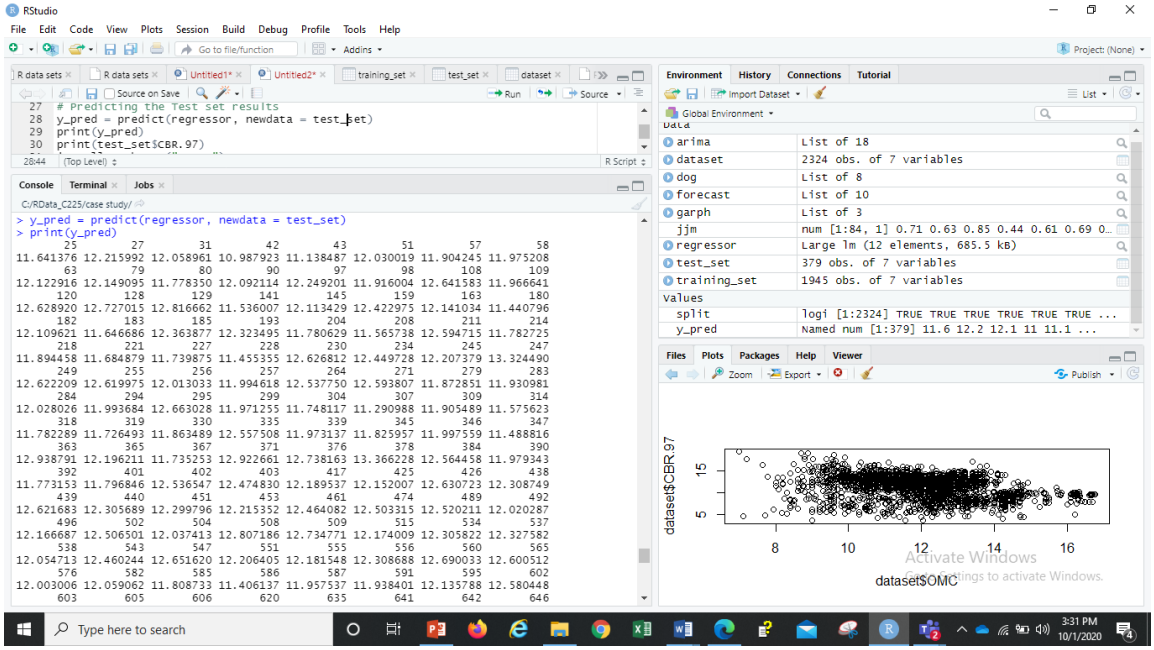


Training Set and Test Set Splitting
Split Ratio = 0.8
Test set contains: 379 rows
Training set contains: 1945 rows
Predicting



6. RESULTS





Financial Data Analysis

Financial data collected in the banking and financial industries are often relatively complete, reliable, and of high quality, which facilitates systematic data analysis and data mining. The Financial sector generates a large amount of data like customer data, logs from their financial products, transaction data that can be used in order to support decision making, together with external data, like social media data and data from websites. The financial data can be collected by many sectors like bank, customer and distributor. So the management of data is very complex, so we use data mining to manage the data and take appropriate decision.

Where can data mining be used:

1. Loan Prediction
   - Selection of Feature and attribute relevance ranking
   - For Loan payment performances
   - For Predicting Consumer credit rating
2. Clustering techniques used to group customers for targeted marketing
3. Designing of Data Warehouses for data mining and multi-dimensional analysis
4. Detection of crimes such as money concealment and similarly other financial related crimes.

7. CONCLUSION

Data Mining has achieved colossal success in various application areas, Hence due to increasing in use of Data Mining applications Data Mining has become an important concept of Computer Science and has also shown high potential for future developments.

The Increasing Technology and new application areas in future also tends to provide new challenges and opportunities for implementation of data Mining.

In the Conclusion statement , It can be encouragingly said that Data Mining is The future and is a focuses of the world. There are many issue yet to be resolved in the Data Mining field but improvement are continuously made in the Data Mining field. The capability to continually change and provide new thoughtful is the principle benefit of DM, and will be at the core of DM bright and promising future. For making a right decision by an Organization they must be having right information at the right time which used to be major concern for many organizations –

we can say it has been almost resolved due to Data Mining Concept. In the future world will be competing to generate information from large amount of data rather then collecting these data.

8. REFERENCES

[1] Data Mining Techniques and Trends – A Review Shital H bhojani
[2] www.google.com
[3] Tutorials point
[4] www.flatworldsolutions.com
[5] https://link.springer.com/content/pdf/10.1007/s10618-007-0067-9.pdf
[6] http://ijiet.com/wp-content/uploads/2016/05/89.pdf
[7] http://www.ijmer.com/papers/Vol2_Issue6/ES2646574663.pdf
[8] https://commons.erau.edu/cgi/viewcontent.cgi?article=1040&context=jdfsl
[9] https://www.sciencedirect.com/science/article/pii/S0167923610001302
[10] https://scialert.net/fulltext/?doi=itj.2011.710.716
[11] https://www.softwaretestinghelp.com/apriori-algorithm/

