

**Group Project**  
on  
**DATA MINING**

**A Project Report/Synopsis submitted in partial fulfillment of the  
requirements for the award of**

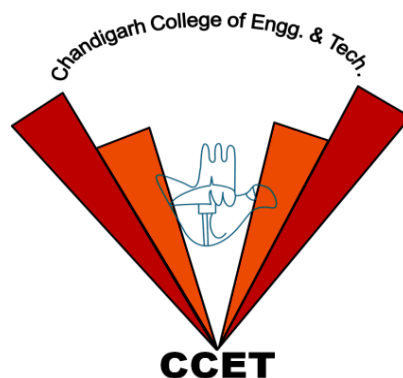
**Bachelor of Engineering  
IN COMPUTER SCIENCE AND ENGINEERING**

**Submitted by**

**MANTASH SINGH**  
(Roll no:CO17335)

**YAMINI**  
(Roll no:LCO17380)

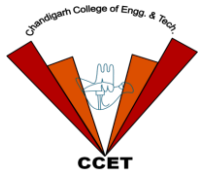
**Under the supervision of (Dr. Ankit Gupta Assistant Professor , CSE Department  
, CCET (Degree Wing))**



**CHANDIGARH COLLEGE OF ENGINEERING AND TECHNOLOGY  
(DEGREE WING)**

Government Institute under Chandigarh (UT) Administration, Affiliated to Panjab University  
, Chandigarh

Sector-26, Chandigarh. PIN -160019  
**APRIL,2020**



## CHANDIGARH COLLEGE OF ENGINEERING AND TECHNOLOGY (DEGREE WING)

Government Institute under Chandigarh (UT) Administration | Affiliated to Panjab University, Chandigarh  
Sector-26, Chandigarh. PIN-160019 | Tel. No. 0172-2750947, 2750943

Website: [www.ccet.ac.in](http://www.ccet.ac.in) | Email: [principal@ccet.ac.in](mailto:principal@ccet.ac.in) | Fax. No. :0172-2750872



---

### Department of Computer Sc. & Engineering

---

#### CANDIDATE'S DECLARATION

WE hereby declare that the work presented in this report entitled “**DATA MINING**”, in fulfillment of the requirement for the award of the degree Bachelor of Engineering in Computer Science & Engineering, submitted in CSE Department, Chandigarh College of Engineering & Technology (Degree wing) affiliated to Punjab University, Chandigarh, is an authentic record of my/our own work carried out during my degree under the guidance of Dr. Ankit Gupta. The work reported in this has not been submitted by me for award of any other degree or diploma.

Date : 14/04/20

MANTASH SINGH (CO17335)

Place : CCET

YAMINI (LCO17380)



**CHANDIGARH COLLEGE OF ENGINEERING AND TECHNOLOGY (DEGREE WING)**

Government Institute under Chandigarh (UT) Administration | Affiliated to Panjab University, Chandigarh  
Sector-26, Chandigarh. PIN-160019 | Tel. No. 0172-2750947, 2750943

Website: [www.ccet.ac.in](http://www.ccet.ac.in) | Email: [principal@ccet.ac.in](mailto:principal@ccet.ac.in) | Fax. No. :0172-275087 2



---

**Department of Computer Sc. & Engineering**

---

**CERTIFICATE**

This is to certify that the Project work entitled “**DATA MINING.**” submitted by

**MANTASH SINGH (Roll no:CO17335)and YAMINI (Roll no:LCO17380)**

fulfillment for the requirements of the award of Bachelor of Engineering Degree in Computer Science & Engineering at Chandigarh College of Engineering and Technology (Degree Wing), Chandigarh is an authentic work carried out by him/her under my supervision and guidance.

To the best of my knowledge, the matter embodied in the project has not been submitted to any other University / Institute for the award of any Degree .

Date : 14/04/20

Place : CCET

Deptt of CSE

CCET(Degree Wing)

Chandigarh



## **CHANDIGARH COLLEGE OF ENGINEERING AND TECHNOLOGY (DEGREE WING)**

Government Institute under Chandigarh (UT) Administration | Affiliated to Panjab University, Chandigarh  
Sector-26, Chandigarh. PIN-160019 | Tel. No. 0172-2750947, 2750943

Website: [www.ccet.ac.in](http://www.ccet.ac.in) | Email: [principal@ccet.ac.in](mailto:principal@ccet.ac.in) | Fax. No. :0172-2750872



---

### **Department of Computer Sc. & Engineering**

---

#### **ACKNOWLEDGEMENT**

We would like to express deep gratitude to Dr. Ankit Gupta Assistant Professor (Computer Science & Engineering), submitted in CSE Department, Chandigarh College of Engineering & Technology(Degree wing) affiliated to Punjab University, Chandigarh without whose permission the training would not be possible. I would also like to thank Dr. Ankit Gupta, Training & Placement Officer, CSE. Department, who recommended me for this training.

We have tried my best to keep report simple yet technically correct. I hope I succeed in my attempt.



## **CHANDIGARH COLLEGE OF ENGINEERING AND TECHNOLOGY (DEGREE WING)**

Government Institute under Chandigarh (UT) Administration | Affiliated to Panjab University, Chandigarh

Sector-26, Chandigarh. PIN-160019 | Tel. No. 0172-2750947, 2750943

Website: [www.ccet.ac.in](http://www.ccet.ac.in) | Email: [principal@ccet.ac.in](mailto:principal@ccet.ac.in) | Fax. No. :0172-2750872



---

### **Department of Computer Sc. & Engineering**

---

#### **ABSTRACT**

The objective of a practical training is to learn something about ASSOCIATION RULE MINING and to be familiar with a working style of a technical worker to adjust simply according to industrial environment. This report deals with the equipments their relation and their general operating principle.

## CONTENTS .....

### Contents

CHAPTER 1: CONCEPTS OF DATA MINING.....	2
1.1 ASSOCIATION RULE MINING.....	2
1.2 APRIORI ALGORITHM.....	2
CHAPTER 2: DATA MINING .....	6
There are various steps that are involved in mining data as .....	6
2.2 STEPS .....	6
2.2.1 Data collection .....	6
2.2.2 Data Cleaning.....	7
2.2.3 Data Integration .....	8
2.2.4 Data Transformation .....	8
2.2.5 Data Mining .....	8
2.2.6 Pattern Evaluation and Knowledge Presentation .....	9
CHAPTER 3: SOFTWARE USED .....	10
3.1 WEKA.....	10
3.2 Orange.....	12
CHAPTER 4: PATTERNS FOUND .....	13
4.1 SMOKING AND BEING OVERWEIGHT ATTRACTS COVID19 .....	13

# CHAPTER 1: CONCEPTS OF DATA MINING

## 1.1 ASSOCIATION RULE MINING

Association Rule Mining, as the name suggests, association rules are simple If/Then statements that help discover relationships between seemingly independent relational databases or other data repositories. Most machine learning algorithms work with numeric datasets and hence tend to be mathematical. However, association rule mining is suitable for non-numeric, categorical data and requires just a little bit more than simple counting.

Association rule mining is a procedure which aims to observe frequently occurring patterns, correlations, or associations from datasets found in various kinds of databases such as relational databases, transactional databases, and other forms of repositories. **An association rule has 2 parts:**

- **an antecedent (if) and**
- **a consequent (then)** An antecedent is something that's found in data, and a consequent is an item that is found in combination with the antecedent. Have a look at this rule for instance:

*"If a customer buys bread, he's 70% likely of buying milk."*

In the above association rule, bread is the antecedent and milk is the consequent. Simply put, it can be understood as a retail store's association rule to target their customers better. If the above rule is a result of a thorough analysis of some data sets, it can be used to not only improve customer service but also improve the company's revenue.

Association rules are created by thoroughly analyzing data and looking for frequent if/then patterns. Then, depending on the following two parameters, the important relationships are observed:

1. **Support:** Support indicates how frequently the if/then relationship appears in the database.
2. **Confidence:** Confidence tells about the number of times these relationships have been found to be true. So, in a given transaction with multiple items, Association Rule Mining primarily tries to find the rules that govern how or why such products/items are often bought together. For example, peanut butter and jelly are frequently purchased together because a lot of people like to make PB&J sandwiches.

*Assume you are a retail store and its association*

## 1.2 APRIORI ALGORITHM

**Apriori algorithm** is given by R. Agrawal and R. Srikant in 1994 for finding frequent itemsets in a dataset for boolean association rule. Name of the algorithm is Apriori because it uses prior knowledge of frequent itemset properties. We apply an iterative approach or level-wise search where k-frequent itemsets are used to find k+1 itemsets. To improve the efficiency of level-wise generation of frequent itemsets, an important property is used called *Apriori property* which helps by reducing the search space.

**Apriori Property –**

All non-empty subset of frequent itemset must be frequent. The key concept of Apriori algorithm is its anti-monotonicity of support measure. Apriori assumes that

*All subsets of a frequent itemset must be frequent (Apriori property). If an itemset is infrequent, all its supersets will be infrequent.*

Before we start understanding the algorithm, go through some definitions which are explained in my previous post.

Consider the following dataset and we will find frequent itemsets and generate association rules for them.

TID	items
T1	I1, I2 , I5
T2	I2,I4
T3	I2,I3
T4	I1,I2,I4
T5	I1,I3
T6	I2,I3
T7	I1,I3
T8	I1,I2,I3,I5
T9	I1,I2,I3

minimum support count is 2 minimum confidence is 60%

#### Step-1: K=1

(I) Create a table containing support count of each item present in dataset – Called **C1(candidate set)**

Itemset	sup_count
I1	6
I2	7
I3	6
I4	2
I5	2

(II) compare candidate set item's support count with minimum support count(here min\_support=2 if support\_count of candidate set items is less than min\_support then remove those items). This gives us itemset L1.

Itemset	sup_count
I1	6
I2	7
I3	6
I4	2
I5	2

#### Step-2: K=2

- Generate candidate set C2 using L1 (this is called join step). Condition of joining  $L_{k-1}$  and  $L_{k-1}$  is that it should have (K-2) elements in common.
- Check all subsets of an itemset are frequent or not and if not frequent remove that itemset.(Example subset of {I1, I2} are {I1}, {I2} they are frequent.Check for each itemset)
- Now find support count of these itemsets by searching in dataset.



Itemset	sup_count
I1,I2	4
I1,I3	4
I1,I4	1
I1,I5	2
I2,I3	4
I2,I4	2
I2,I5	2
I3,I4	0
I3,I5	1
I4,I5	0

(II) compare candidate (C2) support count with minimum support count(here min\_support=2 if support\_count of candidate set item is less than min\_support then remove those items) this gives us itemset L2.

Itemset	sup_count
I1,I2	4
I1,I3	4
I1,I5	2
I2,I3	4
I2,I4	2
I2,I5	2
I2,I5	2

### Step-3:

Generate candidate set C3 using L2 (join step). Condition of joining  $L_{k-1}$  and  $L_{k-1}$  is that it should have (K-2) elements in common. So here, for L2, first element should match.

So itemset generated by joining L2 is {I1, I2, I3} {I1, I2, I5} {I1, I3, I5} {I2, I3, I4} {I2, I4, I5} {I2, I3, I5}

Check if all subsets of these itemsets are frequent or not and if not, then remove that itemset.(Here subset of {I1, I2, I3} are {I1, I2}, {I2, I3}, {I1, I3} which are frequent. For {I2, I3, I4}, subset {I3, I4} is not frequent so remove it. Similarly check for every itemset) • find support count of these remaining itemset by searching in dataset.

Itemset	sup_count
I1,I2,I3	2
I1,I2,I5	2

(II) Compare candidate (C3) support count with minimum support count(here min\_support=2 if support\_count of candidate set item is less than min\_support then remove those items) this gives us itemset L3.

Itemset	sup_count
I1,I2,I3	2
I1,I2,I5	2

**Step-4:**

- Generate candidate set C4 using L3 (join step). Condition of joining  $L_{k-1}$  and  $L_{k-1}$  ( $K=4$ ) is that, they should have ( $K-2$ ) elements in common. So here, for L3, first 2 elements (items) should match.
- Check all subsets of these itemsets are frequent or not (Here itemset formed by joining L3 is {I1, I2, I3, I5} so its subset contains {I1, I3, I5}, which is not frequent). So no itemset in C4
- We stop here because no frequent itemsets are found further

Thus, we have discovered all the frequent item-sets. Now generation of strong association rule comes into picture. For that we need to calculate confidence of each rule.

**Confidence –**

A confidence of 60% means that 60% of the customers, who purchased milk and bread also bought butter.

$$\text{Confidence}(A \rightarrow B) = \frac{\text{Support\_count}(A \cup B)}{\text{Support\_count}(A)}$$

So here, by taking an example of any frequent itemset, we will show the rule generation. Itemset {I1, I2, I3} //from L3

SO rules can be

$$[I1 \wedge I2] \Rightarrow [I3] \text{ //confidence} = \frac{\text{sup}(I1 \wedge I2 \wedge I3)}{\text{sup}(I1 \wedge I2)} = \frac{2}{4} * 100 = 50\%$$

$$[I1 \wedge I3] \Rightarrow [I2] \text{ //confidence} = \frac{\text{sup}(I1 \wedge I2 \wedge I3)}{\text{sup}(I1 \wedge I3)} = \frac{2}{4} * 100 = 50\%$$

$$[I2 \wedge I3] \Rightarrow [I1] \text{ //confidence} = \frac{\text{sup}(I1 \wedge I2 \wedge I3)}{\text{sup}(I2 \wedge I3)} = \frac{2}{4} * 100 = 50\%$$

$$[I1] \Rightarrow [I2 \wedge I3] \text{ //confidence} = \frac{\text{sup}(I1 \wedge I2 \wedge I3)}{\text{sup}(I1)} = \frac{2}{6} * 100 = 33\%$$

$$[I2] \Rightarrow [I1 \wedge I3] \text{ //confidence} = \frac{\text{sup}(I1 \wedge I2 \wedge I3)}{\text{sup}(I2)} = \frac{2}{7} * 100 = 28\%$$

$$[I3] \Rightarrow [I1 \wedge I2] \text{ //confidence} = \frac{\text{sup}(I1 \wedge I2 \wedge I3)}{\text{sup}(I3)} = \frac{2}{6} * 100 = 33\%$$

So if minimum confidence is 50%, then first 3 rules can be considered as strong association rules.

# CHAPTER 2: DATA MINING

## 2.1 STEPS OF DATA MINING:

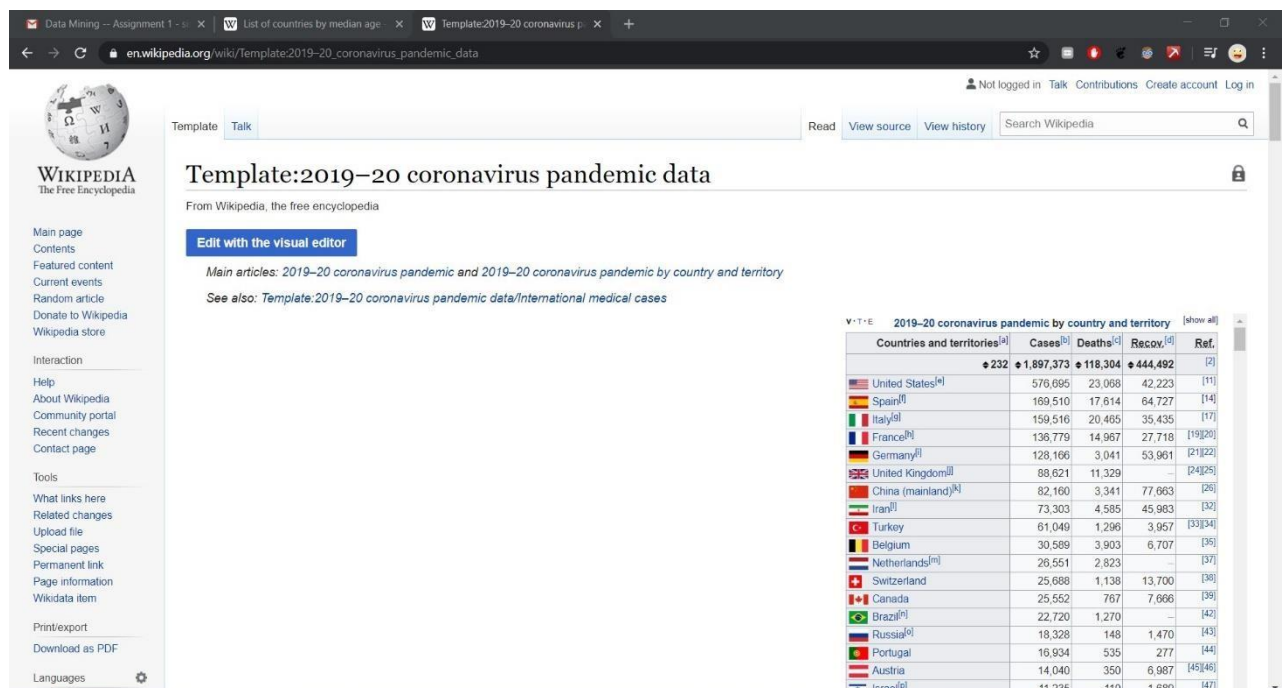
There are various steps that are involved in mining data as

1. Data collection
2. Data Cleaning
3. Data Integration
4. Data transformation
5. Data mining
6. Pattern Evaluation and Knowledge Presentation

## 2.2 STEPS

### 2.2.1 Data collection

All the COVID-19 related data was collected from a single Wikipedia page using a web scraper automatically.



The screenshot shows the Wikipedia page for the template '2019–20 coronavirus pandemic data'. It includes a sidebar with navigation links and a main content area with a table of COVID-19 statistics by country and territory. The table has columns for Countries and territories, Cases, Deaths, Recov., and Ref. The data is sorted by total cases in descending order.

Countries and territories <sup>[a]</sup>	Cases <sup>[b]</sup>	Deaths <sup>[c]</sup>	Recov. <sup>[d]</sup>	Ref.
<b>232</b>	<b>1,897,373</b>	<b>118,304</b>	<b>444,492</b>	<b>[2]</b>
<span><span></span></span> United States <sup>[a]</sup>	579,695	23,068	42,223	<sup>[11]</sup>
<span><span></span></span> Spain <sup>[a]</sup>	169,510	17,614	64,727	<sup>[14]</sup>
<span><span></span></span> Italy <sup>[a]</sup>	159,516	20,465	35,435	<sup>[17]</sup>
<span><span></span></span> France <sup>[a]</sup>	136,779	14,967	27,718	<sup>[19][20]</sup>
<span><span></span></span> Germany <sup>[a]</sup>	128,166	3,041	53,961	<sup>[21][22]</sup>
<span><span></span></span> United Kingdom <sup>[a]</sup>	88,621	11,329	—	<sup>[24][25]</sup>
<span><span></span></span> China (mainland) <sup>[a]</sup>	82,160	3,341	77,663	<sup>[26]</sup>
<span><span></span></span> Iran <sup>[a]</sup>	73,303	4,585	45,983	<sup>[32]</sup>
<span><span></span></span> Turkey	61,049	1,296	3,957	<sup>[33][34]</sup>
<span><span></span></span> Belgium	30,589	3,903	6,707	<sup>[35]</sup>
<span><span></span></span> Netherlands <sup>[a]</sup>	26,551	2,823	—	<sup>[37]</sup>
<span><span></span></span> Switzerland	25,688	1,138	13,700	<sup>[38]</sup>
<span><span></span></span> Canada	25,552	767	7,666	<sup>[39]</sup>
<span><span></span></span> Brazil <sup>[a]</sup>	22,720	1,270	—	<sup>[42]</sup>
<span><span></span></span> Russia <sup>[a]</sup>	18,328	148	1,470	<sup>[43]</sup>
<span><span></span></span> Portugal	16,934	535	277	<sup>[44]</sup>
<span><span></span></span> Austria	14,040	350	6,987	<sup>[45][46]</sup>
<span><span></span></span> Israel <sup>[a]</sup>	11,235	110	1,689	<sup>[47]</sup>

We also used the same site to collect the median ages of different countries i.e. Wikipedia

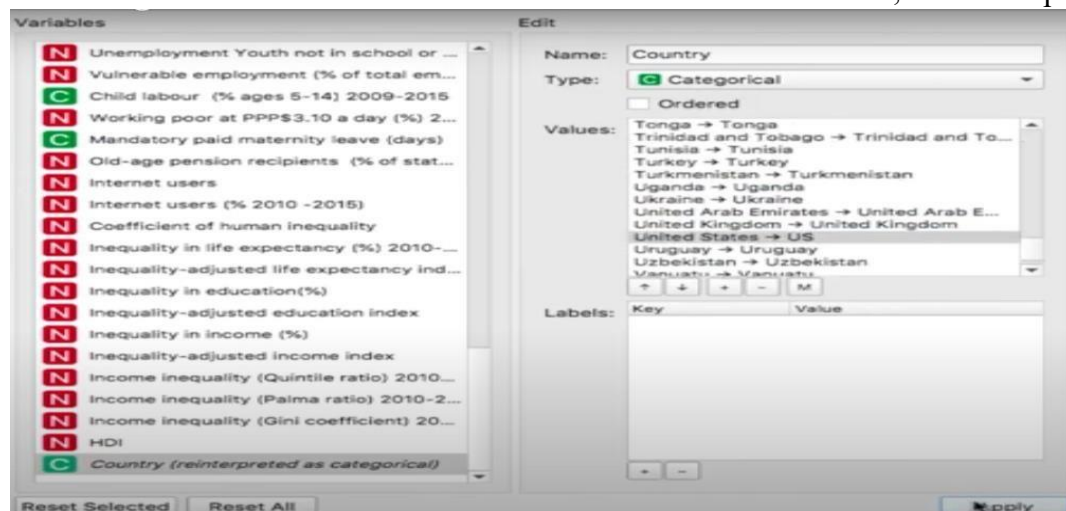
The screenshot shows the Wikipedia page titled "List of countries by median age". The page includes a sidebar with navigation links, a main content area with a table of countries, and a world map showing median age by country. The table lists countries with their rank, median age, and male/female median ages.

Country/Territory	Rank	Median (Years)	Male (Years)	Female (Years)
Afghanistan	208	18.9	18.8	18.9
Albania	95	32.9	31.6	34.3
Algeria	136	28.1	27.8	28.4
American Samoa	122	25.5	25.1	26.0
Andorra	10	44.3	44.4	44.1

## 2.2.2 Data Cleaning

The data we have collected may contain errors, missing values, noisy or inconsistent data. So, we get rid of such anomalies. This step was very time consuming, due to all the manual work required.

In this step, we used orange to rename the files to represent the dates when the data was captured and to remove inconsistencies such as the use of abbreviations like “UK”, “US” to represent



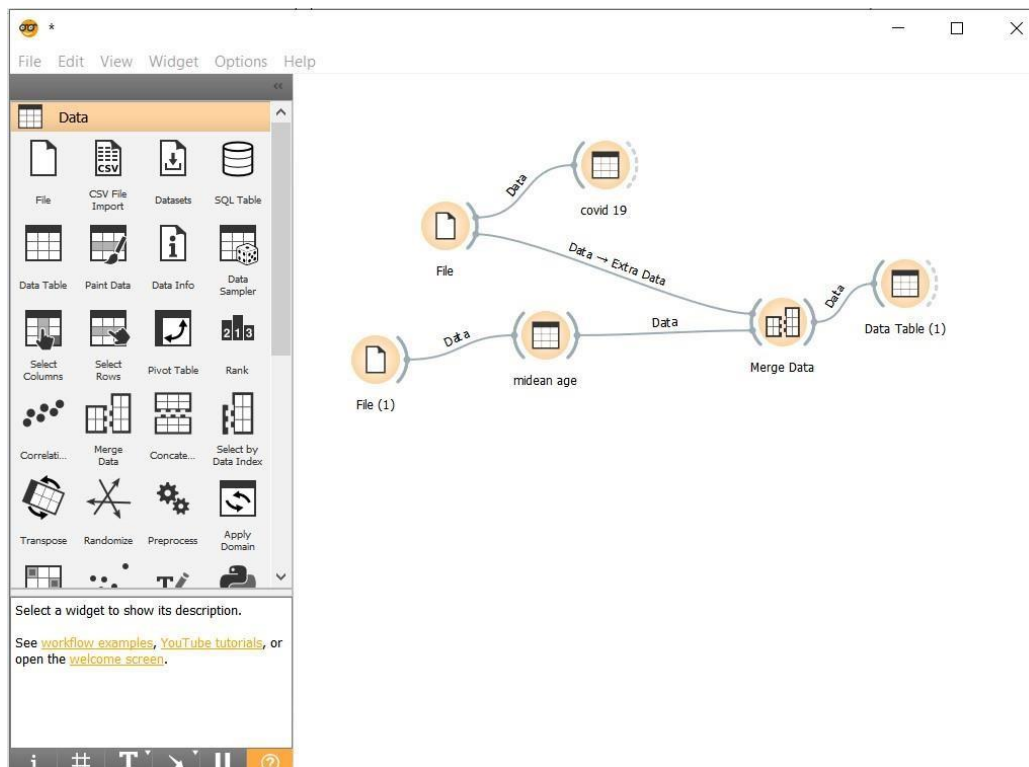
countries, and disambiguation of names such as “China” and “China (Mainland)” to represent the same things.

### 2.2.3 Data Integration

Data cleaning in data mining is the process of detecting and removing corrupt or inaccurate records from a record set, table or database. 1 You can ignore the tuple. This is done when class label is missing. This method is not very effective , unless the tuple contains several attributes with missing values.

### 2.2.4 Data Transformation

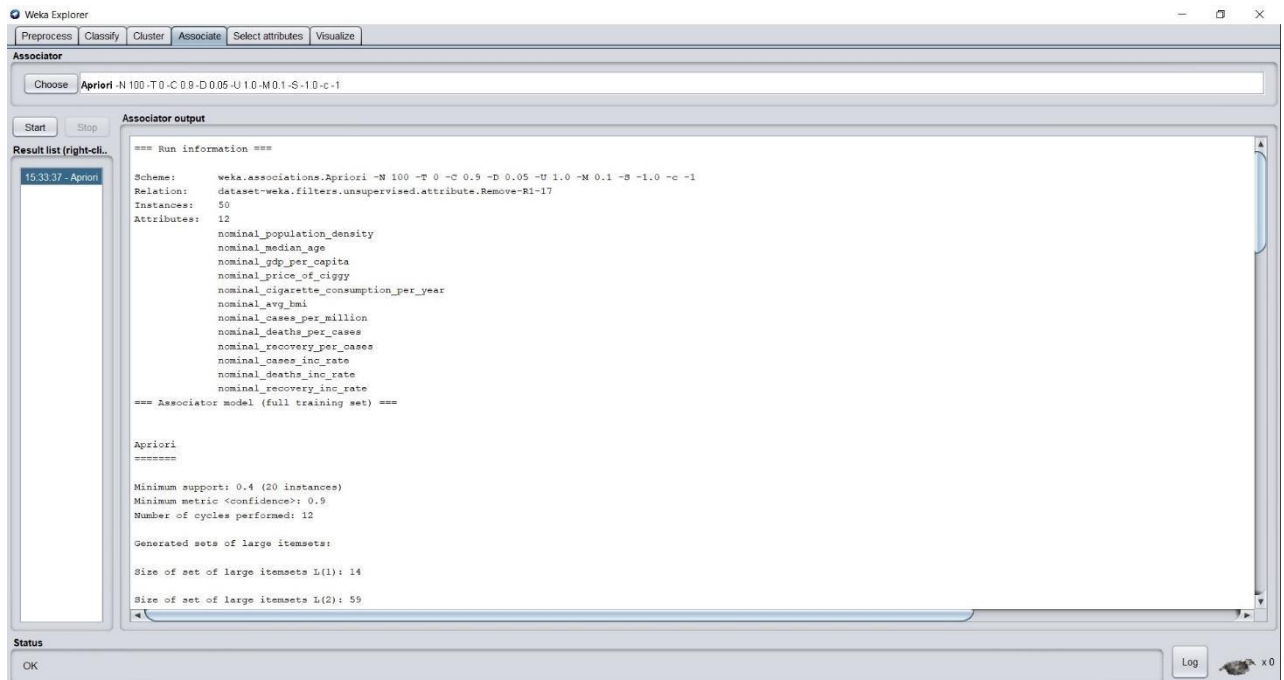
In computing, Data transformation is the process of converting data from one format or structure into another format or structure. It is a fundamental aspect of most data integration and data management tasks such as data wrangling, data warehousing, data integration and application integration.



### 2.2.5 Data Mining

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.[1] Data mining is an

interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use



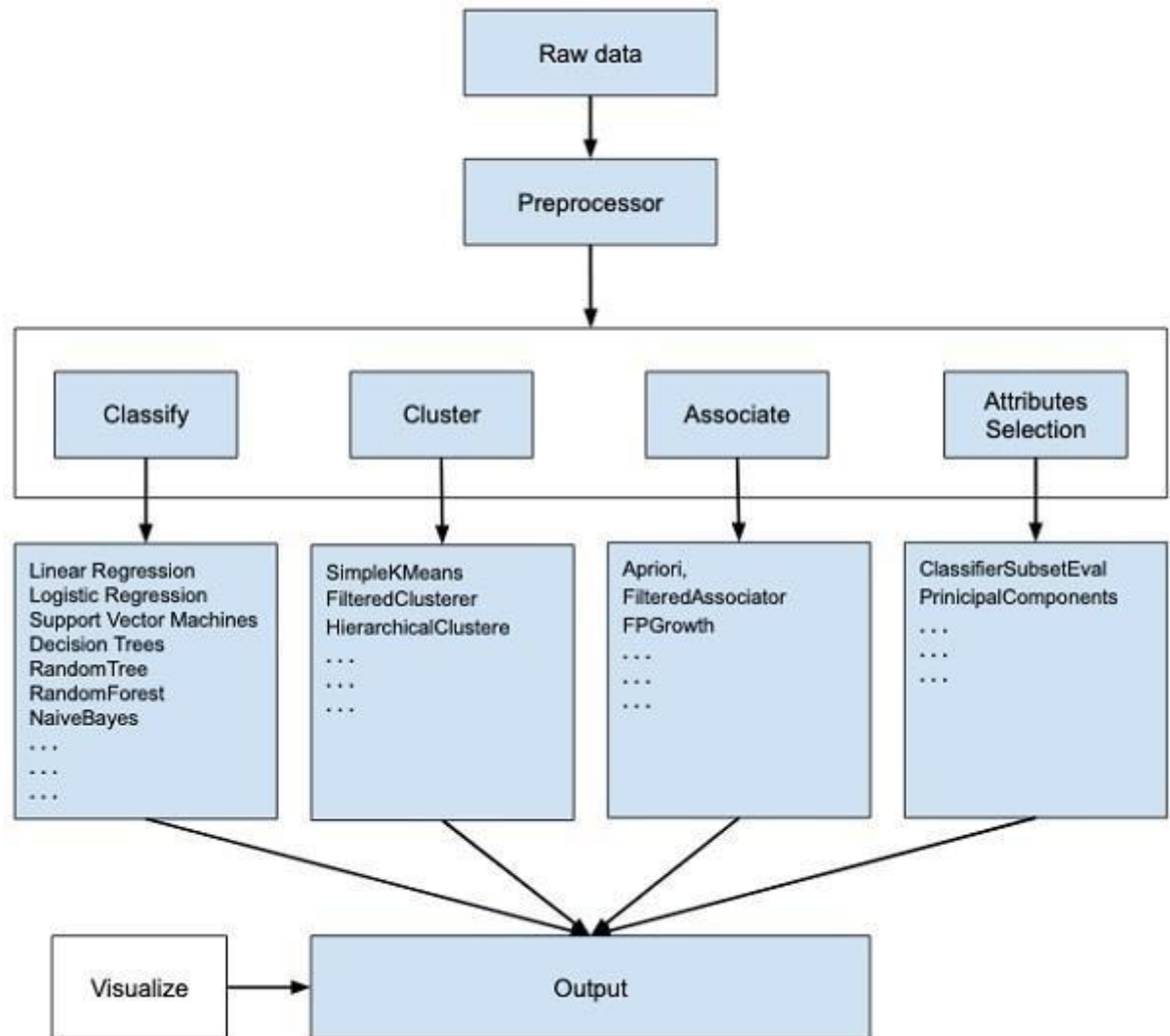
## **2.2.6 Pattern Evaluation and Knowledge Presentation**

This step involves visualization, transformation, removing redundant patterns etc from the patterns we generated.

# CHAPTER 3: SOFTWARE USED

## 3.1 WEKA

WEKA - an open source software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems. What WEKA offers is summarized in the following diagram –



If you observe the beginning of the flow of the image, you will understand that there are many stages in dealing with Big Data to make it suitable for machine learning –

First, you will start with the raw data collected from the field. This data may contain several null values and irrelevant fields. You use the data preprocessing tools provided in WEKA to cleanse the data.

Then, you would save the preprocessed data in your local storage for applying ML algorithms.

Next, depending on the kind of ML model that you are trying to develop you would select one of the options such as **Classify**, **Cluster**, or **Associate**. The **Attributes Selection** allows the automatic selection of features to create a reduced dataset.

Note that under each category, WEKA provides the implementation of several algorithms. You would select an algorithm of your choice, set the desired parameters and run it on the dataset.

Then, WEKA would give you the statistical output of the model processing. It provides you a visualization tool to inspect the data.

The various models can be applied on the same dataset. You can then compare the outputs of different models and select the best that meets your purpose.

Thus, the use of WEKA results in a quicker development of machine learning models on the whole.

Now that we have seen what WEKA is and what it does, in the next chapter let us learn how to install WEKA on your local computer.





## **3.2 Orange**

Orange is a component-based visual programming software package for data visualization, machine learning, data mining, and data analysis.

Orange components are called widgets and they range from simple data visualization, subset selection, and preprocessing, to empirical evaluation of learning algorithms and predictive modeling.

Visual programming is implemented through an interface in which workflows are created by linking predefined or user-designed widgets, while advanced users can use Orange as a Python library for data manipulation and widget alteration



## CHAPTER 4: PATTERNS FOUND

### **4.1 SMOKING AND BEING OVERWEIGHT ATTRACTS COVID19**

---

nominal\_gdp\_per\_capita=RICH nominal\_cigarette\_consumption\_per\_year=HIGH nominal\_avg\_bmi=FAT 21  
==> nominal\_median\_age=OLD 20 <conf:(0.95)> lift:(1.19) lev:(0.06) [3] conv:(2.1)

---

As per the above research done , we found an intrusting pattern .According to which if a country has HIGH GDP per capita i.e. rich , median age is old is combined with high cigarette consumption per year and average BMI i.e. body mass index is FAT .

Then there are high chances of COVID-19 to spread in that area and infect more people .

As Covid19 is a respiratory disease, it was obvious that Covid19 will be more harmful to smokers than non-smokers. What is surprising, though that smoking is only been a key factor where population density is high.

Also , Obesity can weaken the body's immune system and reduce its ability to fight off infections.