## Data Analytics Tools

* It is a Data Analytics Platform that focused on understanding data and using it potential in business strategy
* It is a free data visualization application that link to many data sources whether that it is a comparate ~~database~~ datawarehouse or mx-excel or web based information.

### Features

→ Not meet any prior ~~prerequisites~~ technical knowledge in order. to ~~use~~ two data source integration

Tableau connecting to

→ The process of data analysis is faster with interactive visual data representation.

highly evole data visualization which provide a number of way in which data can be study

→ It is easy to use, easy to implement with simple ~~design~~ drag and drop interface. which make easy to learn and operate for professional at any level

### R programming

→ R is a open source programming language which

is widely used as statical software and data analytics tool

→ It was designe by Ross Thaka & Robert Gentle Man
→ At the university of auckland, Newzland and it is currently developed by core team.

Feature.

→ R is interpreted programming language.
→ It is a platform independent language
→ with the static packages it also allowed to integrate with other programming language. such as C, C++, java
→ It has a ~~white~~ wide variety of libaries
→ R has CRAN, it is repository of R ~~programming language~~. holding more than 1500 package.

Python

→ It is a scripting language that is simple to use or understand and write as well
→ It is developed by Guido Van Rossum in 1980

Python is a dynamic high level, language
interpreted programming

features.

→ Free and open source
→ Easy to code and easy to read.

→ It support object oriented programming language and it is also integrated language.

→ Python is a dynamically typed language.

Apache spark

- It was created in 2009 by AMP lab at the university of california. berkley.
- Apark spark is large scale data processing engine to perform application 100 time quicker when it come to memory and 10 time faster on disc in Hadoop cluster.
- It is based on data science and it is design popular for developing data pipeline and machine learning models.
- Spark also contain Mplep package which provide a progressive collection of ML algorithm for learning data science processing like classification, collaborative, filtering, regression and clustering.

# Data Analytics life cycle

Phase 1 : Discovery

Phase 2 : Data preparation

Phase 3 : Model Planning

Phase 4 : Model Building

Phase 5 : Communication Result

Phase 6 : operationalization

## Key Role of Data Analytics

Business user

Project sponsored

Project manager

Business Intelligent Agent

Data Base Administrator

Assignment 1.

1. What do you understand by Big Data?
2. Explain different type of Big data platform.

Big data :- The definition of big data is data that contain greater variety, arriving in increasing volumes and with more velocity. This is also known as the three Vs. Put simply, big data is longer, more complex data sets, especially from new data source.

Big data is a term used to describe a collection of data that is huge in size and yet growing exponentially with time

example :- jet engines, social media sites etc.

There are three types of Big data are :-

1. Structured data
2. Semi-structured data
3. Unstructured data

~~Structured data~~ :- ~~It is based on Relational database table~~

It is based on XML / RDF

A big data platform act as a organised storage mediums for large amount of data. Big data platform utilise a combination of data management hardware and software tool to store

aggregated dataset usually onto the cloud.

Google cloud :- Google cloud offers a lots of big data management tools, each with its own specialty. Big Query warehouse petabytes of data in an easily queried format.

Microsoft azure :- Users can analyze data stored on Microsoft's cloud platform, Azure, with a broad spectrum of open-source Apache technologies, including Hadoop and spark.

Amazon web services :- Best known as AWS, Amazon's cloud based platform comes with analytics tool that are designed for everything from data prep and warehousing to SQL queries and data lake design.

Snowflake :- Snowflake is a data warehouse used the for storage, processing and analysis. It runs completely atop the public cloud infrastructures - Amazon web services, Google cloud platform and Microsoft Azure and combines with a new SQL query engine.

2. Write the differentiate between Predictive and prescriptive Analytics and Data analysis and Data Analytics.

| Predictive | prescriptive Analytics. |
| --- | --- |
| Models certain aspects of a business | Models the entire business |
| Forecasts what's likely to happen | Is 100 percent data driven |
| Predicts when it will happen | Recommends specific business decision. |
| Outputs are non-actionable. They only identify the need to take a decision | Considers Interdependencies. Is not bound by static rule |

Data Analysis

Data Analysis is a specialized type of analytics used in business to evaluate data and gain insights.

It is described as a particularized form of analytics

It ~~support decision making by analyzing~~ It analyzes the data by focusing on insights into business data.

It support inferential analysis

One cannot find anonymous relation with the help of this

A Descriptive analysis can be performed on this

Data Analytics

Data analytics is a traditional or generic type of analytics used in enterprises to make data-driven decision

It is described as a traditional form or generic form of analytics

It support decision making by analyzing enterprise data

It does not deal with inferential analysis

One can find anonymous relation with the help of this

Descriptive analysis cannot be performed on this

## Characteristics of Big Data.

Volume:-

It refers to the huge amount of data. the name big data itself is related to an size

Big data is vast volume of the data generated from many sources daily such as business process, social media platform, human interaction

ex:- facebook can generate approximation per day.

Variety:- Big Data can be structured, unstructured, and semi-structured that are be collected from different sources.

It data will be only collected by from data base and sheet in the vast best these days. data will come in many form that are emails, video's, photos, audio etc.

Veracity:- It refers to the inconsisting and uncertainty in data it means how much the data is Relevable.

It is many way to filter and translate the data.

Velocity is the process of been able and manage the data.

Value.

It refers to expect useful data.

value is not a data that is we process or least is reliable on and valuable in nature that is

The value of big data usually come from inside discovery and pattern recognition that leads to more effective operation. It is stronger customer relationship.

Velocity:- It refers the speed in which companies receive ~~store~~ store the manage data.

It refers to speed by which the data is created in real time

It contain the linking of incoming data set speed, rate of change and activity burst.

The primary aspects of big data is to provide demanding data rapidly.

Importance of big data.

Understanding marketing condition:

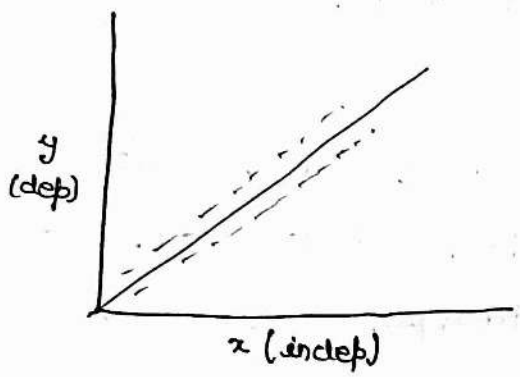
Time stressing

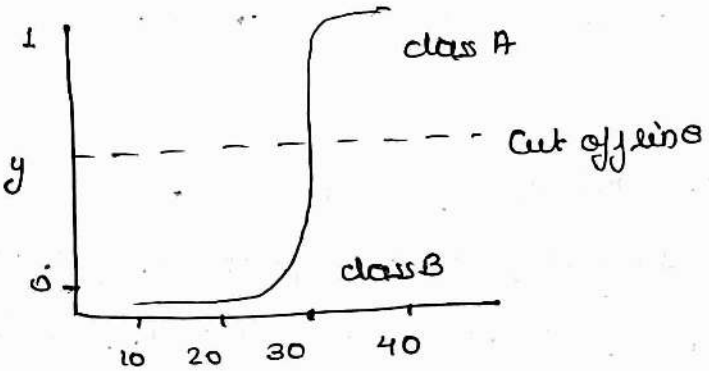
Offering marketing inside.

Unit:-2

## Regression Analysis

1. Dependent & Independent Variable
2. No outlier
3. No multi collinearity
4. Underfitting & overfitting




$y = \alpha + \alpha_1 x_1$

Linear Regression

$y = \frac{1}{1+e^{-x}}$

Logistic Regression

Regression Analysis :- It is a statistical technique of measuring the relationship between variables. it provide a value of the dependent variable from the value of a independent variable

The main use of regression analysis is to determine the strength of predictores. forecast an effect an identifying a towards.

example :- A gym supplement company is used RA technique to determine how prizes and advertising can effect the sales of the supplement

Logistic Regression:- In Logistic Regression dependent variable is in binary form and independent variable can be continous or binary.

Here goal is to find the best fitting model for dependent and independent variable relationship.

# Dependent variable

It deals with the propability to measure the relationship between the both dependent and independent variable.

It is used to find the possibility and propability of occuring an event.

Sigmaid function

Sigmaid function simply convert independent variable into propability expression within range of 0 and 1 with respect to dependent variable.

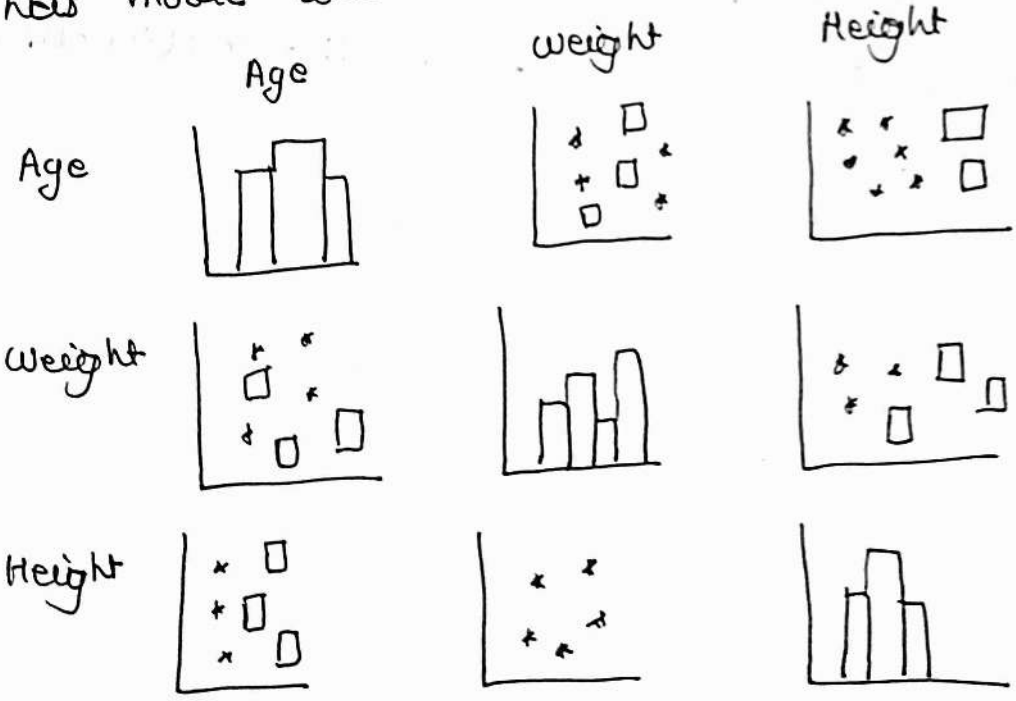
If there are various classes in the dependent variable then it is known as multi - collinearity logistic regression.

## Multivariate analysis

Multivariate analysis is involves evaluating multiple variable to identify any possible association among them. It is define as a process involving multiple dependent variable resulting one outcome

example:- We can not predict the whether condition of any year based on the season their are multiple factors like pollution, humidity, precipitation etc. precipation in which whether depend.

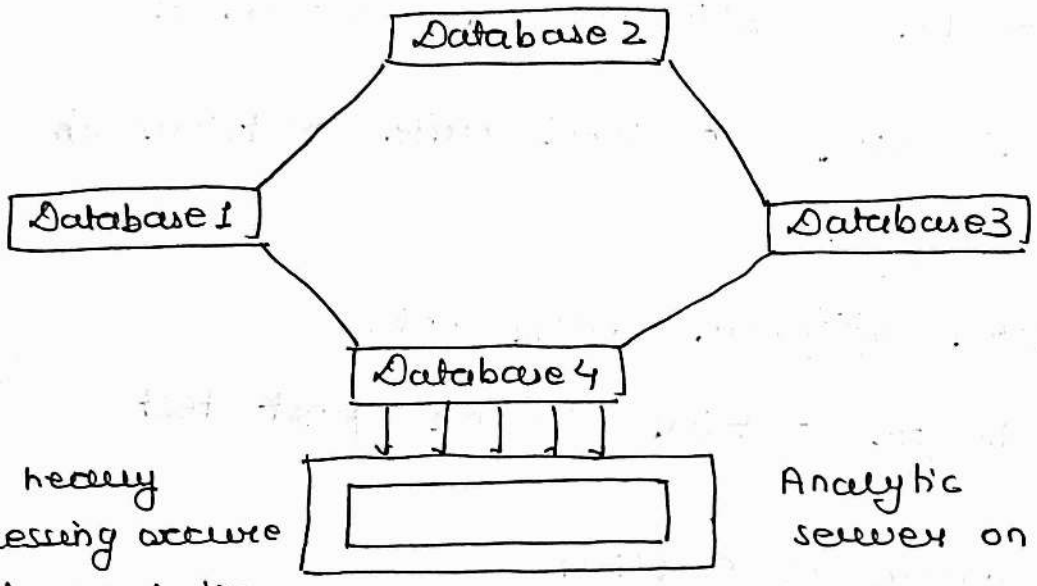
As if consider more than one factor of independent variable that influences the variability of dependent variable hence the conclusion bond has more accurate.



## Assignment No:-2

Write a short note on analytics scalability?

In analytic scalability we have to pull the data together in a separate analytic environment and then start performing analysis.




Analysts also perform data preparation, which is made up of joins, aggregations, derivations, and transformation. In this process, they pull data from various sources and merge it all together to create the variables required for analysis.

What is Data cleaning? How do we handle outlier in data set?

Data cleaning is the process of fixing or removing incorrect corrupted, incorrectly formatted, duplicate or incompleted data within a dataset. When combining multiple data sources, there are many opportunity for data to be duplicated or mislabeled

There are 5 ways to deal with outliers in data sets :.

- Set up a filter in your testing tool.
- Remove or change outlier during post test analysis.
- Change the value of outlier.
- Consider the underlying distribution
- Consider the value of mild outlier.

Differentiate the role and responsibilities of data analyst and data scientist.

Data ~~analyst~~ analysts typically work with structured data to solve tangible business problem using tool like SQL, R or python, data visualization software and statistical analysis

Common tasks for a data analysts include:

Collaborating with organisational leaders to identify informational needs.

Requiring data from primary and secondary sources.

Cleaning and reorganising data for analysis

Analysing data set to spot trends and patterns that can be translated into actionable insights.

Data Scientist after deal with the unknown by using more advanced data technique to make prediction about the future. They might automate their own machine learning algorithm or can handle both structured and unstructured data. This role is generally considered a more advance version of data analyst some of their task include

Gathering, cleaning and processing raw data.

Designing predictive models and machine learning algorithm to mine big data set.

Building data visualisation tools, dashboard and reports.

Writing programs to automate data collection and processing.

Explain Data Modelling and Data building phases.

Phases of Data Modelling :-

Conceptual Data Models :- Focus on the general state of the system, the entities to be included business requirement of the database to be built, and the type of data to be stored.

Logical Models: Are about building the database Primary, secondary and foreign keys are worked out, along with restrictions and the actual tables and columns to be used.

The phases of Data building are :-

- Data Discovery
- Data preparation
- Planning of data models
- Building of data models
- Communication of result
- Operationlisation

Differentiate between Discrete data and Continous data.

| Discrete Data | Continous Data. |
| --- | --- |
| Takes specific countable values | Takes any measured value within a specific range. |
| Ordinal data value and integer values represent discrete value. | Decimal numbers and fraction represent continous data. |
| Easily counted on something as simple as a number line | Requires more in-depth measurement took and method like curves and be skewed |
| Discrete data remain constant over a specific time interval | Continous data varies over time and can have separate values at any given point. |

Explain various phases of Analytics life cycle.

The data Analytics life cycle is as follows:.

Phase 1:

Discovery:. Develop content and understand come to know about data sources needed available for the project.

Phase 2:

Data preparation: Step to explore preprocess, and condition data prior to modelling and analysis.

Phase 3:

Model Planning:- Data science team develop data sets for training, testing and production purposes.

Phase 4:

Model Building:- Team develops dataset for testing, training, and production purpose.

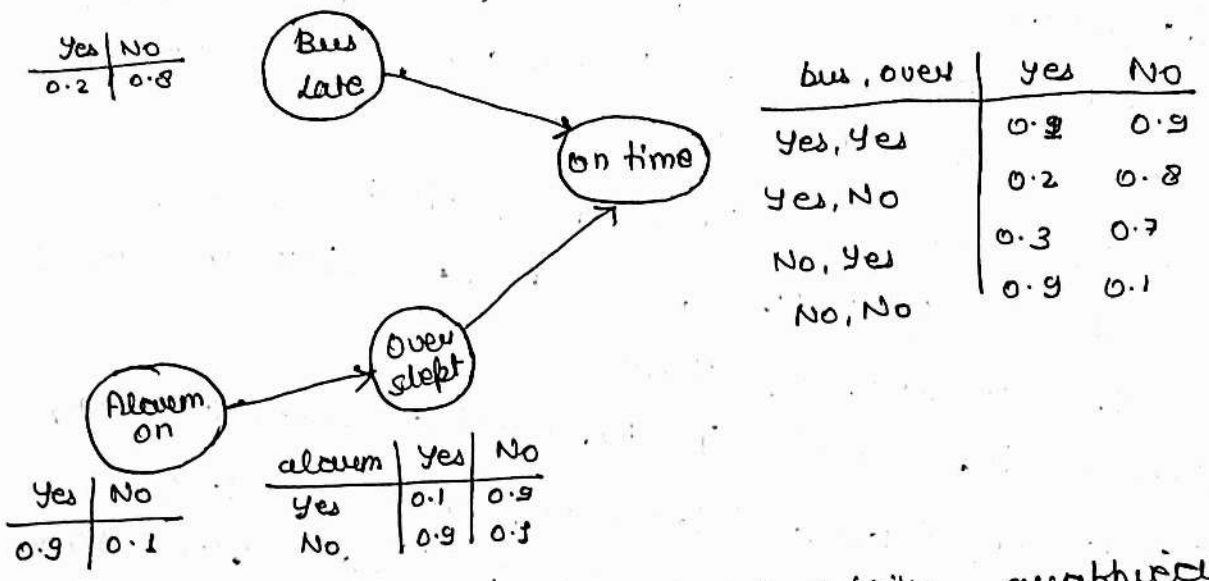
Phase 5:

Communication Result:- After executing model team need to compare outcomes of modeling to criteria established for whether a success or failure.

Phase 6:

Operationaling:- The team delivers final reports, codes.

Bayesian Network.




| Yes | No |
|---|---|
| 0.2 | 0.8 |

| bus, over | Yes | No |
|---|---|---|
| Yes, Yes | 0.9 | 0.9 |
| Yes, No | 0.2 | 0.8 |
| No, Yes | 0.3 | 0.7 |
| No, No | 0.9 | 0.1 |

| Yes | No |
|---|---|
| 0.9 | 0.1 |

| alarm | Yes | No |
|---|---|---|
| Yes | 0.1 | 0.9 |
| No | 0.9 | 0.1 |

It is widely used class of probabilistic graphical models they consists of two part a structure and parameter.

The structure is a directly acyclic graph that express conditional independence and dependency among random variables associated with each node.

The parameter consists of conditional probability distribution associated with each node.

A Bayesian network is compact, flexible and interpretable representation of a joint probability distribution

Kernel: There are the type of algorithm that are used for pattern analysis. these method involve using linear classifier to solve non-linear problem SVM uses Kernel methods to solve classification and regression issues it use Kernel method to take data as input and transform it into the required

from of processing the data.

Assent Essentially Kernel method are algorithm, that make it possible to implicitly project the data is a high dimension state.

Time series analysis :- It is a specific way of analysing a sequence of data po point collected over an interval of time.

In time series analysis Analysist seperate data point at consistent interval for a set of period of time rather than just recording the data points intere

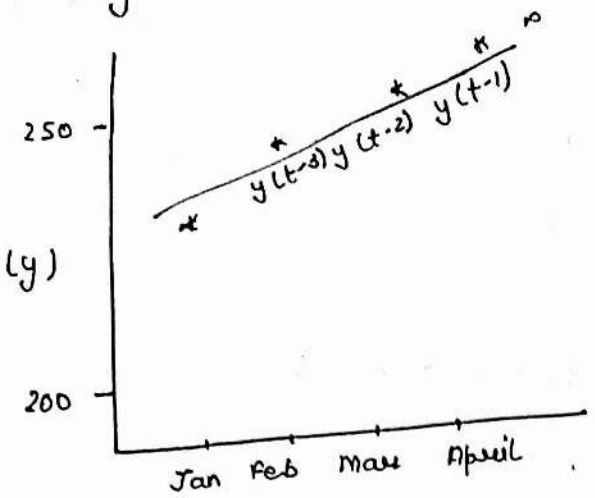
It require a large number of data point to insure consistence and reliability.

Time series analysis help organisation to understand the underlink process of trends, or systematic pattern over time.

Organisation uses time series for forecasting to produce the likelihood of future events.

Example:- Whether data, Rainfall measurement, stop crisis, interest rate

Auto Regression :-




$$y(t) = \beta_0 + \beta_1 y(t-1) + \beta_2 y(t-2) + \ldots \text{error}(t)$$

- It is a statistical model used for time series analysis
- An auto Regression model describe How a particular variables passed values influences its current value.
- AR model use passed values to predict feature value

In AR model response variable by dependence upon on the previous value of y at a pre determine constant time log

Here time log on daily basis or weekly

The most oper ese case for this type of models is stake with stake market prices.

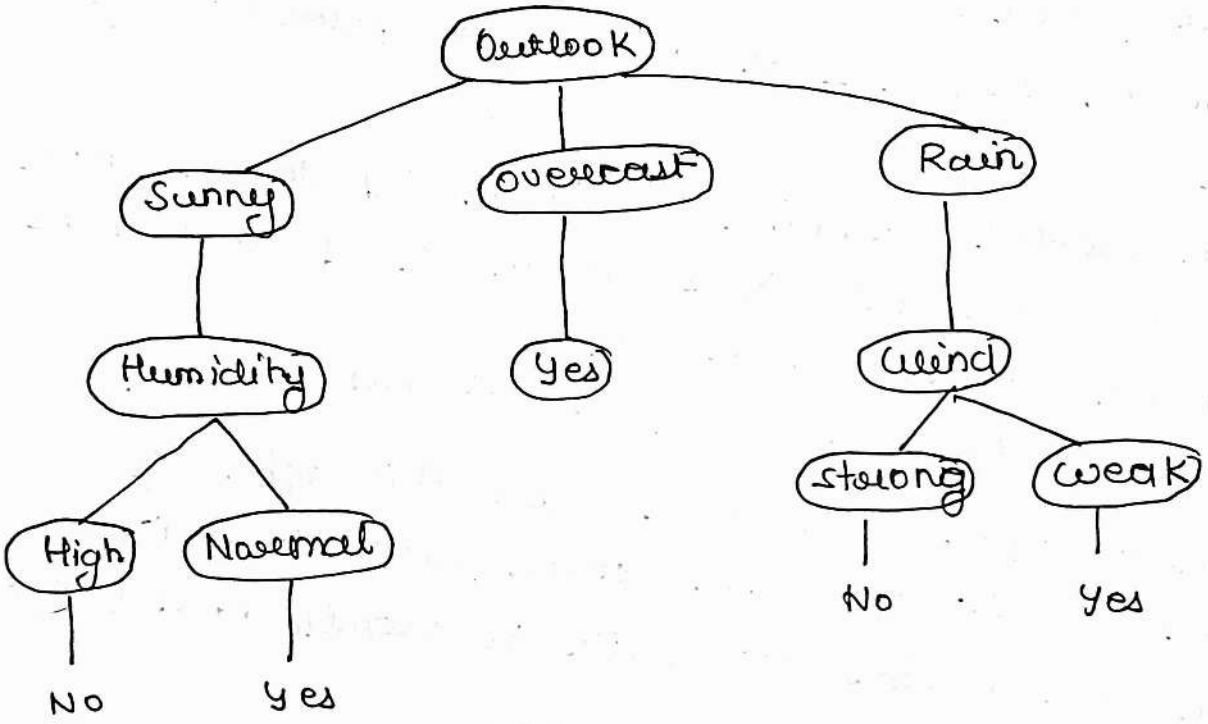
Where the prices to take is highly correlated with the price one-day ago.

Induction Rule :- Rule induction is one of the most imposetant techniques of machine Learning since regulay hiden in data are frequently expressed in terms of Rules. ~~induction~~

| ID | Age | Income | Student |
|---|---|---|---|
| 1 | young | 25000 | Yes |
| 2 | old | 35000 | No |
| 3 | Medium | 16,000 | Yes |

R1: IF age = youth AND student = YES
THEN Buy Computer = YES

R2: IF age = 'old' AND Income > 25000
THEN buy computer = NO.




{ IF OutHook = Sunny AND
Humidity = High THEN
PLAY = NO }

- Rule induction is one of the fundamental tool. used for the same.
- Rule induction is data mining process of reducing if ~~their or rule~~. IF-then from a data set
- These symbolic decision rule explain inherit relationship between the attributes and class label in the data set.

IF
Defining the pre-condition or coverage of the rule
Then
Then present studing classification or prediction or decision expression.

Characteristics of Rule.
Mutually exclusive rule
Exhaustive Rule.

Assignment :- 1 (Unit 2).

Perform Linear Regression of demand on price from the following data :. also predict the value of y that correspond to a value of $x = 7.5$.

$y = x+$

| Price | Demand. |
|---|---|
| 2.0 | 75 |
| 2.5 | 60 |
| 3.0 | 55 |
| 3.5 | 50 |
| 4.0 | 45 |
| 4.5 | 40 |

$y = a + bx$

$a = \frac{\Sigma y + b(\Sigma x)}{n}$ ✓

$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$ ✓

| $x$ | $y$ | $\Sigma x$ | $\Sigma y$ | $\Sigma xy$ | $\Sigma x^2$ |
|---|---|---|---|---|---|
| 2.0 | 75 | | | 150 | 4 |
| 2.5 | 60 | | | 150 | 6.25 |
| 3.0 | 55 | | | 165 | 9 |
| 3.5 | 50 | | | 175 | 12.25 |
| 4.0 | 45 | | | 180 | 16 |
| 4.5 | 40 | | | 180 | 20.25 |
| 19.5 | 325 | | | 1000 | 67.75 |

$n = 6$

$a =$ ~~$325+$~~

$b = \frac{6(1000) - (19.5)(325)}{6(67.75) - (19.5)^2}$

$b = \frac{-337.5}{406.5 - 380.25}$

$b = \frac{-337.5}{26.25}$

$b = -12.857$

$$a = \frac{325 + 12.057(19.5)}{6}$$

$$a = \frac{325 + 250.711}{6}$$

$$a = \frac{74.209}{6} \quad \frac{575.711}{6}$$

$$a = 12.3015 \quad 95.95$$

$$y = 95.95 + (-12.057)(7.5)$$

$$= 95.95 - 96.4275$$

$$= -0.4775$$