1. Introduction to Big Data

1. What is Big Data

→ ① Big data is a term that is Used to describe data that is high Volume, high Velocity, and high Variety, requires new technologies and techniques to Capture, store and analyze it.

② Big data refers to data that is so large, fast or Complex that it's difficult or impossible to process Using traditional methods.

2. What is Unique about big data?

→ ① Companies have searched for make the best Use of information to improve their business Capabilities.

② Big data is also special because it represents both Significant information which can open new doors to improve business.

3. Enlist tools Used to Big data?

→ 1. NoSQL - Database MangoDB, CouchDB, Cassandra, Bigtable.
2. Mapreduce - Hadoop, Hive, Pig, S4, Kafka.
3. Storage - S3, Hadoop distributed file System.
4. Server - El2, Google app engine, Elastic
5. Processing - R, Yahoo! Pipes, Bigsheets

Above are the Some tools Used for big data.

4. Write Advantages and disadvantage of big data

→

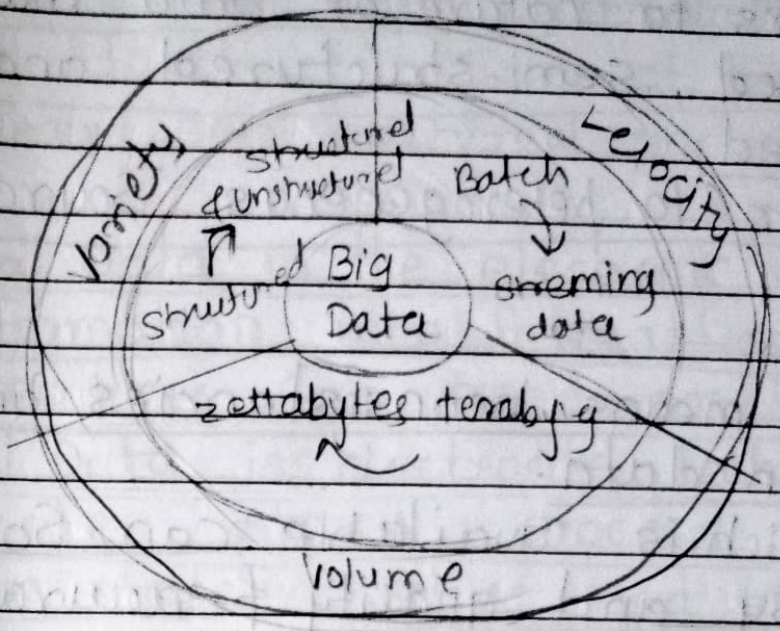
* Advantages:
1. Big data helps in improving Science and research;
2. It improves healthcare
3. Every second additions are made.
4. One platform Carries Unlimited ifommation;
5. It helps in fianicial trading's, Sports, polling etc;
6. It helps in Optimizing business process;

* Disadvantages:
1. lost of Big data is Unstructured
2. it may increase social stratification
3. It can be Used for manipulation of public records;
4. Big data is not Useful in short run.
5. Traditional storage Can Cost lost of money to store big data;

5. Explain 5 'V's of Big Data.

→ I. Valume (Volume)
II. Variety (Variety)
III. Velocity (Velocity)
IV. Veracity (Verosity)
V. Value (Volue)

these are 5 V's of big data



I. Volume

① · The name 'Big Data' itself is related to a size.
② · Volume is huge amount of data.
③ · To determine the Value of data, Size of data plays a very crucial role,
④ · if the Volume of data is very large then it is actually Considerd as a 'Big data'.

II. Velocity

① · velocity refers to the high speed of accumulation of data.
② · In Big data velocity data flows in forms Source like machine, networks, Social media, phones etc.
③ · example- There are more than 3.5 billion Searches per day are made on google.

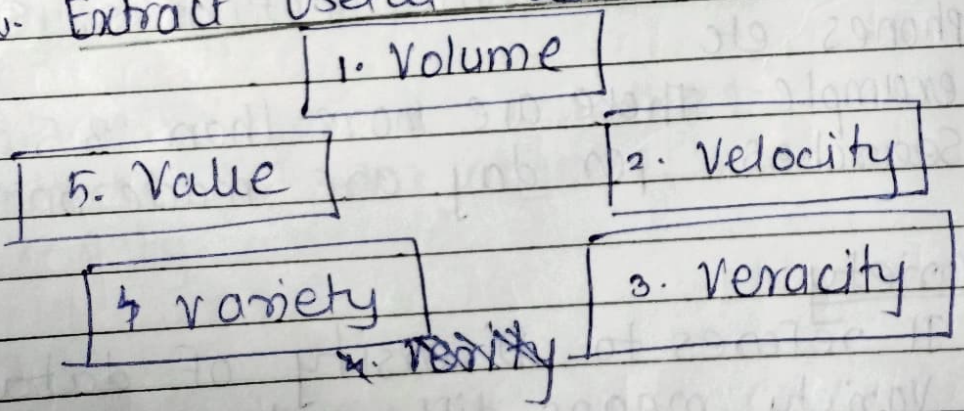
III. Variety

① · It referes to Diversity of data.
② · Variety means different formats of data from various Sources.

② • It refers to nature of data that is structured, semi-structured and Un-structured.
④ • It refers to heterogeneous sources.

IV. Veracity
① • Veracity means inconsistencies and Uncertanity in data.
② • Data which is available can Sometimes get messy and quality & accuracy are difficult to Control.
③ • example- Data in bulk could Create Confusion.

V. Value
① • After having 4 V's into there comes one more V which is Stands for Value.
② • The bulk of data having no value. Unless you turn it into Something Useful.
③ • Data needs to Converted into Something Valuable to extract information.
④ • Value is the most imp V of all 5 V's.
• Extract Useful data.



6. What is Digital Data? OR Explain types of digital data OR. Types of big data OR diff? structured, semi-structured and unstructured

→

> Digital Data is the electronic representation of information in a format or language that machines can read or understand.
> Digital Data is electronic technology that generates, stores, process data in terms of two states: positive and non-positive.

* Types of Digital data and Big data.

I] Structured data -
① . This data is basically an Organized data.
② . It generally has defined the length and format of data.
③ . Structured data has standardized format, has a well defined structure.
④ . Structured data is easily understood by machine language.
⑤ . example - Relational database, name, dates, address, and more.

II] Unstructured data -
① . This data is not-organized data.
② . It refers to data that generally doesn't fit neatly into the traditional row and Coloumn structure of the relational database.
③ . Unstructured data doesn't store in structured database.

④ Word, PDF, text, Media logs, picture etc. are the examples of Unstructured data which can't be stored in the form of rows and Coloumns.

III] Semi-structured data

① · This data basically semi-orgnized data

② · It generally refers to data that has defined the length and format of data. do not Confirm to the formal structure of data.
· log Files, XML data examples of semi-str.

③ · Semi-structured data does not follow the Format of tabular data model because it does not have fixed schema.

7. What is big data analytics & explain types of big analytics.

→

› Big data analysis is the process of extracting Vseful information by analysing different types of big data sets.

› Big Data analytics is Used to discover hidden patterns, market trends, and consumer preferences for the benefits of organizational decision making.

› There are four types of big data analysis.

① Descriptive analysis ② Predictive analysis
③ Exploratory & discovery ④ Prescreptive analysis

I] Descriptive Aalytics

- Descriptive analytics is a stastical method that is Used to search and Summrize historical data in order to identify patterns or meaning.
- The Simplest way to define descriptive analytics is that it answers the question "What has happened?".
- Example- Company reports provide histonic reveiw

II] Predictive analytics.

- Predictive analytics is a stastical method that is Used Utilizes algorithms and machine learning to identify trends in data and Predict future behaviour.
- It has ability to "Predict" what might happen
- This analytics are about Understanding future.
- Example- Training targets, talent management

III] Prespective analysis.

- Prespective analysis is a statistical method Used to generate recommendation and make decisions based on Computational findings of algorithmic model.
- It answer the question "what should we do?".
- It focuses on Advise on possible outcomes.
- example- A training manager Use prespective analysis.

8. Diff' Between structured, Unstructured and Semistructured.

Properties	structured data	Unstructured data	Semi-structured
① organization	well Organized	not organized at all	Partially organized
② Transaction	Matured transaction	No Transaction	Transaction is adapted from DBMS not matured.
③ Flexibility	less Flexible	more Flexible	more Flexible than structured data but less Flexible than Unstructured data.
④ scalability	Very difficult to scale	very Scalable	Simpler to scale than structured data.
⑤ Technology	It is based on relational database.	It is based on character and binary data	It is based on XML/RDF.
⑥ Example	Financial data, Bar codes, Relational database.	Media logs, Vedio, Audio files.	Tweets organized by hastags, folder organized by topics.

classmate Date Page

9. What is ment by Petabyte and zettabyte?

→

- Pettabyte :-
A petabyte is a measure of memory or data storage capacity that is equal to 2 to the 50th power of bytes.

- zettabyte :-
A zettabyte is a Unit of measurement used by technology professionals and the general public to describe a computer or other device's storage capacity.

10. Inlist Big data Applications.

→

1. Healthcare
2. Telecommunication
3. Fianancial Firms
4. Retails
5. Energy & Utilities
6. Education
7. Media & Entertainment.
8. Law Enforcement
9. Marketing
10. New Product development
11. Banking
12. Insurance
13. Agriculture

11. What are the Sources of Big data.

→ • Big data has many Sources. includes.
• Social media Sources Soch as twitter, facebook generate tremendous amounts of Comments and tweets.
• Machines such as Smart phones, meters, generate data.
• Image, Voice and audio data.

2. Introduction to Data Science

1. What is Data Science ? Write down Components of data Science.

→ ①. Data Science analyzes large amount of data
②. Data Science is the study of data to extract meanigful insights for business.
③. Data science means extracting Useful Knowledge for data to Solve business problems.
④. R-language is most important for data science.
⑤. Data Science is the domain of study that deals with vast volume of data Using morden tools and techniques to find meanigful information.

* Components of data science

1. Stastistics - It is most important Component it is way to Collect and analyze the numerical data in large amount.

2. Visualization - It represents data in visual Context so that people can Understand it is easy to access the huge amount of data in visuals.

3. Data enegineering - It is part of data Science which involves acquiring, storing, retrieving and transforming data.

4. Advanced Computing = Heavy lifting of data science is advanced computing. it involves designing, writing, and maintaining the source code of computer programs.

5. Machine learning - It all about to provide training to machine so that it can act as a human brain.

2. Write Down Advantages and disadvantages of data Science.

→

• Advantages :-

1. Data Science helps organizations to reduce Coasts, get into new markets.
2. The domain is in high demand.
3. it offeres abundance of job positions.
4. it helps in securing highly paid Career.
5. it helps in achieve high objectives of business.
6. it is versatile branch of data.

• Disadvantages :-

1. Data Science is blurry term.
2. it requires large amount of knowledge to become data scientist.
3. It has Problems of data privacy.
4. Mastering Data Science is nearly impossible.
5. Tools used for data science are more expensive.

3. Write down applications of Data Science

→

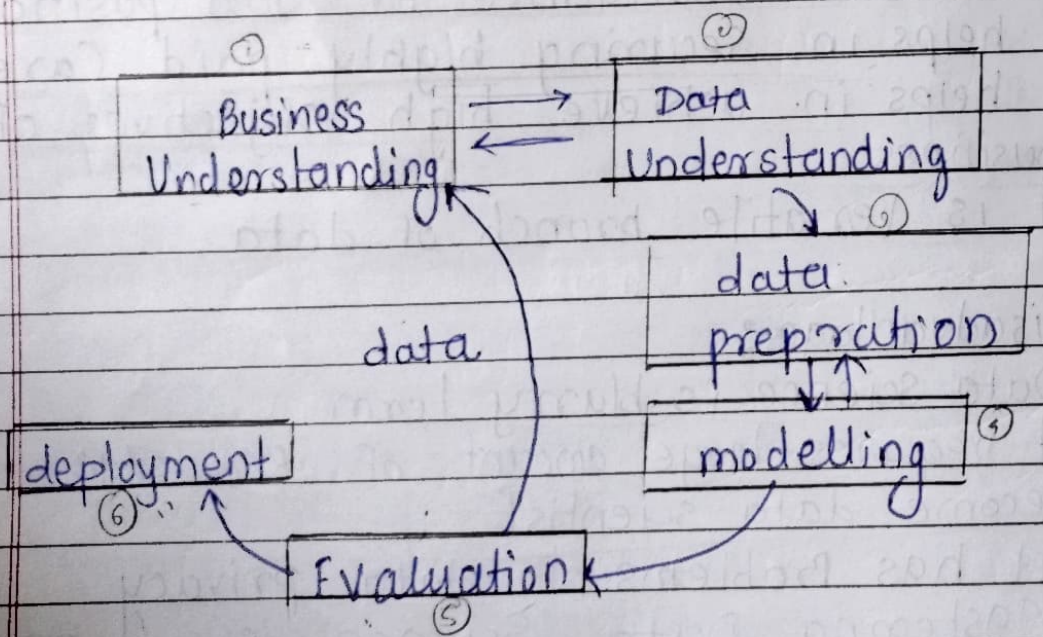
1. Fraud and Risk Detection
2. Healthcare
3. Virtuall assistance for patients and Customer Supporty
4. Internet Search
5. Targeted advertising
6. Gaming
7. Website recommendations.

4. Explain Data Science Process.

→

1. Following are the proces phases and stages involved in data Science process.

2.



Data Science Process stages.

3. Data science process is systematic approach to solving a data problem it provides a structured framework.

1. Business Understanding -
- The First step of this process is Setting a research goal.
- This stage has main purpose that making sure all the users understand the what, how and why of the project.

2. Data Understanding:-
- The Second phase is data Understanding that is data retrieved.
- You want to have data available for analysis.
- This step includes finding Suitable data.

3. Data Prepration-
- This phase or stage involves Data prepration and proceeds with activities in order to get familiar with data.
- In this phase data is designed, Stored, modeling, formating involve.

4. Modeling -
- In this phase various modeling techniques are seleeted and applied.
- This phase Comes after data prepration phase

5. Evaluation-
- At this stage in the project, you have built a model that appears to have high quality.
- It is imporatant to evaluate the model.

6. Deployment:-
- The last step or stage of data Science process model is presenting your results and automating the analysis.
- Deployment is last phase of Data Science process.

5. What is Data Analytics.?

→
① Data analytics is the Science of examining raw data with the purpose of drawing Conclusions about that information.
② Data Analytics is a lifeline for the IT industry.
③ Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering Useful information and Supporting decision making.)
④

6. Need of Big Data analytics.
→ Big data analytics gives future plan for organization & organization to take decision based on Big data Analysis
Thats why Big data analytics is needed and more important.

7. Write down advantages & disadvantages of big data analytics.

→ Advantages.
1) Fraud detection
2) Delivering Relevant Products.
3) optimizing & improving customer Experience.

Disadvantage.
1) lack of Commitement & Patience
2) low quality of data
3) More Complex.

8. Explain types of Data analytics

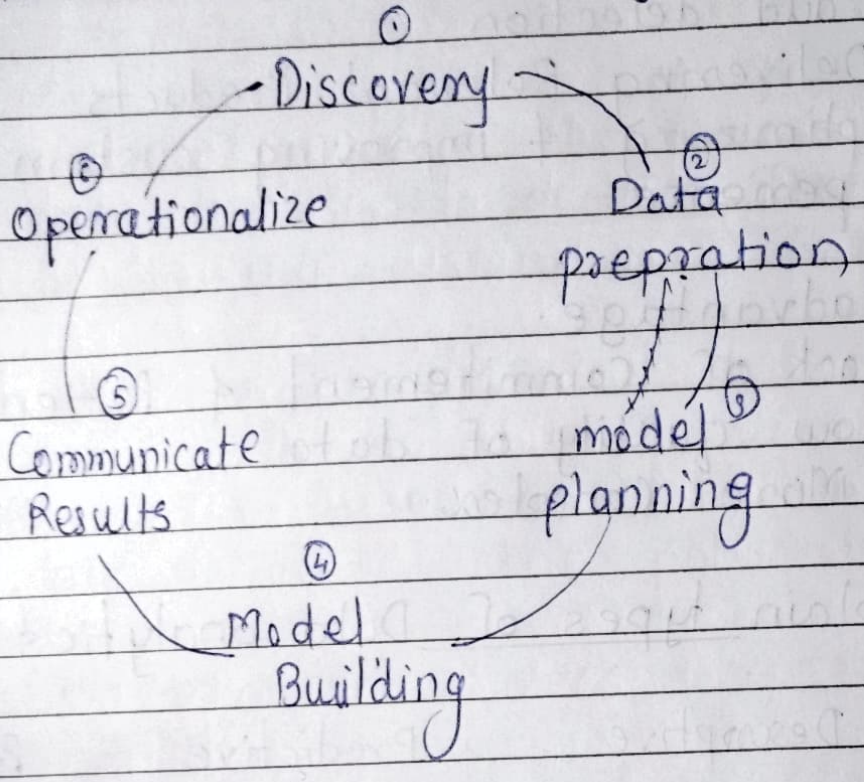
diff^n bet^n Predictive & Prespective descriptive analys

→

Properties	Descriptive	Predictive	Prespective
Dep^n.	Summrizes historical data to identify hidden patterns.	It is the Use of data to Predict the future events	analyzes data and provides possible outcome
Focus on	Insights into the past	Understanding the Future	Advise on possible outcomes
Answers the question	what has happened?	what might happen?	What should we do?
Tools Used	Data aggregation, Data mining	stastics, modeling simulation	Business rule machine learning optimization
Examples	Demand trends Survey results	Training targets, talent management.	Training manager.

9. Explain Phases of Data analytics lifecycle.

→



① Phase1 - Discovery
- Team learns business domain.
- data Science team is trained and researches the issue
- The team comes with an initial hypothesis, which can be later Confirmed with evidence.

② Phase-2 Data Prepration
- This phase involve analysing, preparing data and modeling
- It required to have an analytic Sandbox
- Data prepration tasks can be repeated and not in predefined sequence
- Some tools used Commonly for this process include - Hadoop, Apline miner, Open Refine etc.

④ Phase-3 Model Planning.

In This phase, team determines the methods, techniques, and workflow of data.

- In this phase data Science teams create data sets that can be used for training for testing, production and training goals.
- Some tools used in this stage are MATLAB and STASTICA.

⑤ Phase-4 Model building -

- In this phase, team creates datasets for training, testing as well as production use.
- In this phase the team builds and excutes model based on work done in the model planning phase.
- Tools used are free and open-source octave, Rand PL/R. etc.

⑥ Phase-5 Communication Results-

- In this phase, stakeholders, determines if the results of project are success or failare based on the Certeria.

⑦ Phase-6 Operationlize

- In this phase team ddivers final reports, briefing, code and technical doaments
- Uses free tools Such as WEKA, SQL MADlib and octave.

10. What is stastical Inference ? explain with example ?

→

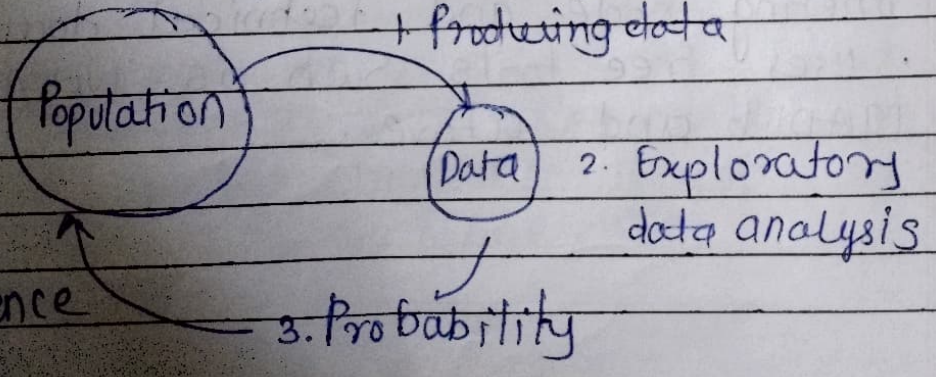
① Statistical Inference is the process of Using data analysis to deduce properties of an Underlying distribution of probability

② Statistical inference is method for drawing and measuring the reliability of Conclusion about population based on info Obtained from a Sample of the population.

③ Statistical Inference involves four-step process that is.
i. Data production =
ii Exploratory Data analysis =
iii. Probability =
iv. Inference =

④ Stastical Inference includes two broad areas.
i. Stastical Estimation
Point Estimation — interval estimation
ii. Stastical Hypothesis Testing.

⑤



11. What is Population Explain

→

① Population refers to the collection of all elements that have possessing common characteristics that Comprises Universe.

② Population is defined as a total of the items Under Consideration.

③ Population focuses on Identifying characteristics

④ Example - automobiles with 4 wheels, people who Consume olive oil.

⑤ Types of population :-

i. Finite Population
ii. Infinite Population
iii. Existent population
iv. Hypothetical population

12. What is Sample.

→

① Sample is a part of population chosen at random

② Sample is that part of the population from which information is Collected.

③ Sample is defined as proportion of population selected.

④ Sample focus on Making inference about population.

⑤ Example - 20 Cars from each Make.

13. Explain Stastical sta Statistical Modeling

→

① stastical modeling is an approach to stastical data analysis that helps researches discovers Something about a

Steps of Stastical modeling:-
④ ① model selection ② Model fitting ③ model visulization.

Phenomenon that is assumed to exits
② This Stastical modeling approach helps explain the Variability found in the dataset
③ Example - Public Health data, Social media data.

14. Define Probability

→

① Probability theory developed form the study of games of chances like dice and cards.
② Process like flipping a coin, rolling a die or drawing a card from a deck are called Probability experiments.
③ Probability theory is a the foundation for stastical inference.
④ Probability = The no. of ways of achieving success / The total no. of possible outcomes

15. Probability distribution

→

① The Probability distribution of a random Variable x is the system of numbers.
② There are two types of probability distribution
a. discrete Probability distribution
b. Continous Probability distribution

16. What is Correlation & types of Correlation

→

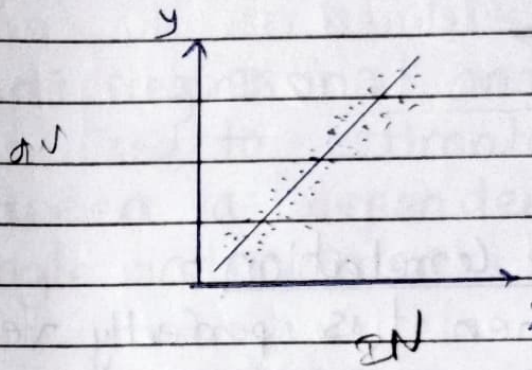
① Correl Correlation is a statistical measure which determines ~~for~~ association of two Variables.

② Correlation is used to represent linear relationship between two variable.

③ Types of Correlation

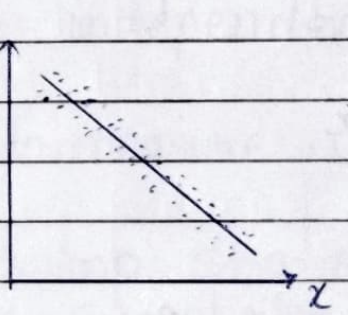
a) Positive Correlation -

- positive correlation indicates positive association bet^n the variables.



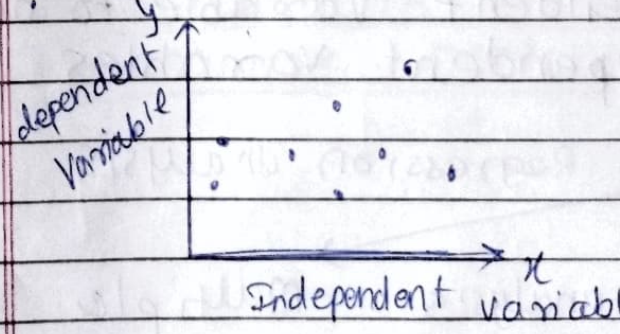
b) Negative Correlation -

- It indicates -ve association bet^n the variables.



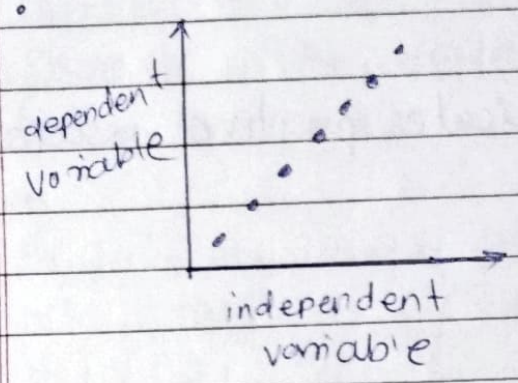
c) No Correlation

- In the no relation bet^n variables.



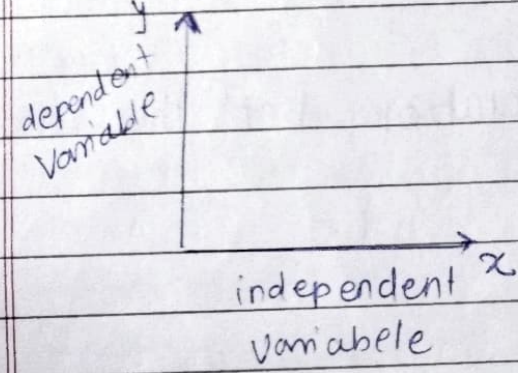
d) Perfectly positive Correlation-

• when $r=1$, then it is perfectly +ve correlation



e) Perfectly Negative Correlation-

• When $r=-1$, then it is perfectly -ve correlation

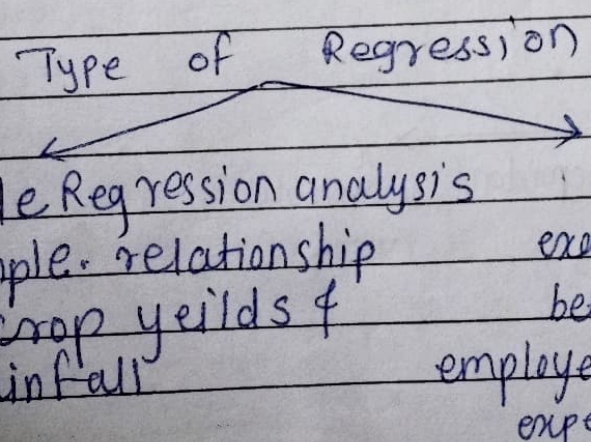


17. What is Regression & Regression analysis.

→

① Regression is a stastical technique that relates a dependent variable to one or more independent variables.

②



③ The stastical technique that expresses a functional relationship between two or more variables in the form of an equation to estimate the value of a variables, based on the given value of another variable is called regression analysis".

④ Simple regression analysis -
→ it is used to estimate the relationship between a dependent variable and a single independent variable.

⑤ Multiple regression analysis
→ it is used to estimate the relationship between a dependent variable and one or more independent variable.

* Techniques to measure Correlation.
→
. There are 3 imp statistical tool used to measure correlation.

a) Scatters Diagrams
b) Karl Pearson's Coefficient of Correlation
c) Spearman's rank correlations.

3. Introduction to Machine learning

1. What is machine learning & write down need of machine learning.

→

1. Machine learning is defined as the capability of a machine to imitate intelligent human behavior.
2. Machine learning is subset of artificial intelligence, which focuses on using stastical techiques to build intelligent computer system to learn from available database.
3. Machine learning is a scientific disipline that is concerned with the design and development of algorithms that allows computers to evolve behaviors based on available data from database
4. Need of Machine learning -

- Diversity of data - data is bieng generated from diff^n channels and its natures are diff^n
- Capacity and dimensions - There is large number of data available lead to growth of data
- Speed - As data volume increases, speed at which data is captured and transformed.
- Complexity - With increase in complexity of data, high data quality and security is required.
- Applicability - It means applicability of data to business and performence improvement.

Remaining ques.
Understood machine learning

2. Write down advantages and disadvantages of machine learning.

→

- Advantages-

1. ML is Used in Variety of applications Such as banking and financial Sector, healthcare etc.
2. ML has ~~Cabilit~~ Capabilities to handle ~~multi-dimensional~~ and multi-variety data.
3. ~~ML allows time~~ cycles reduction and efficient Utilization of resources.
4. The process of automation of tasks is easily possible.
5. Due to ML, there are tools available to provide quality improvement in Complex data.

- Disadvantages-

1. Acquisition of relevant data is the major challenge.
2. It is impossible to make Immediate accurate predications with a machine learning System
3. Machine learning needs a lots of training data for future prediction.
4. interpetation of results is also a major challenge to determine effectiveness of machine learning algorithms.
5. Possibility of High Error

- Advantage- ML is Used for data analysis and Knowledge discovery in data base, data mining, etc.

Ross Dhaka & Robber Gentleman

3. Introduction to R language & Write features of R language

① R is a programming language and software enviornment for stastical analysis, graphic representation and reporting.

② R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand and Currently developed by the R Development Core Team.

③ Features of R :-
- R is well developed, simple and effective programming language
- R has an effective data handling and storage faeility
- R provide graphical faeilities for data analysis

④ R languange includes Varias Components Such as:

1. Attributes
2. Data types
   a. Veetors
   b. List
   c. Matrices
3. Data frame
4. Control Structures i.e if else, for, while etc.

Ross Ihaka & Robber Gentelen

3. Introduction to R language & Write features of R language

① R is a programming language and software used enviornment for stastical analysis, graphic representation and reporting.

② R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand and Currently developed by the R Development Core Team.

③ Features of R :-
- R is well developed, simple and effective programming language.
- R has an effective data handling and Storage faeility.
- R provide graphical faeilities for data analysis.

④ R languange includes various Components Such as:

1. Attributes
2. Data types
   a. Veetors
   b. List
   c. Matrices
3. Data frame
4. Control Shuctures i.e if .. else, For, While etc.

4. Explain two main type of Machine learning. OR

* I. Supervised Machine Learning

② Also known as Predictive analysis.

① In Supervised Machine learning, the training is Controllerd by external Superior.

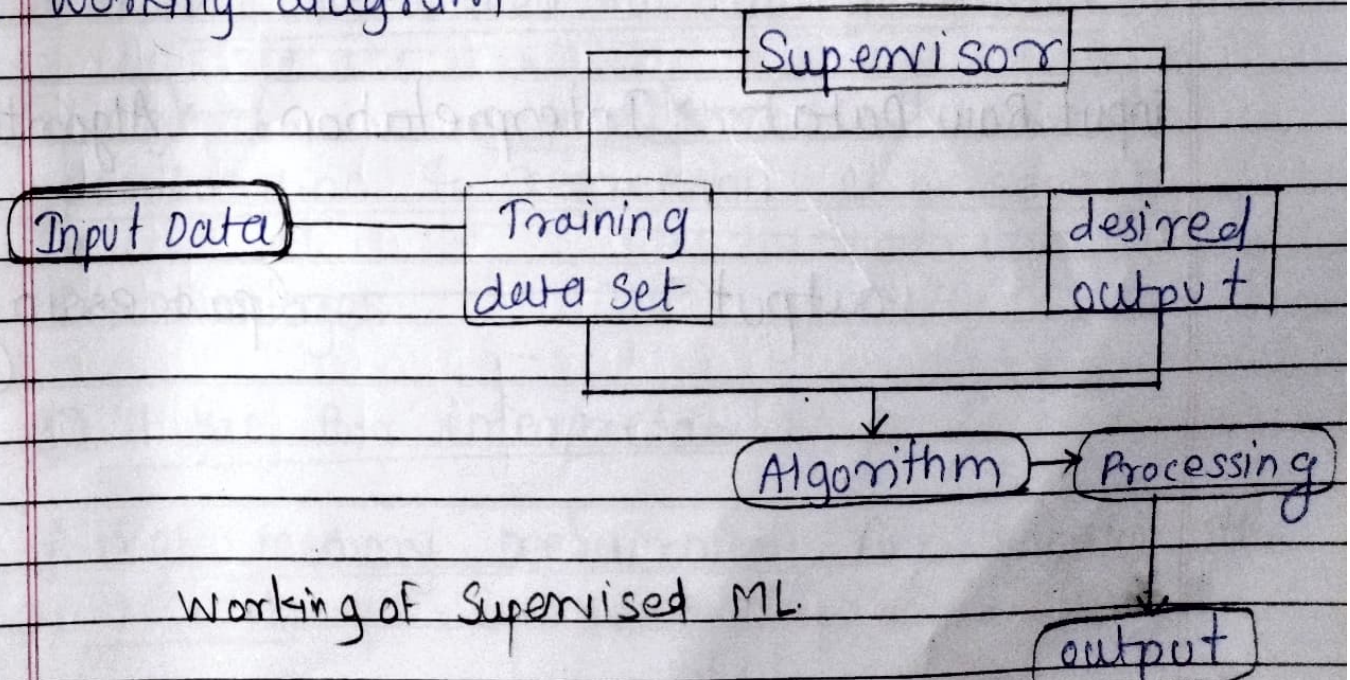
② Common use of Supervised Machine learning is to use historical data to predict future events.

③ In Supervised learning, Computer is provided with input that are labeled with their desired output.

④ Supervised learning is the type of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output.

⑤ Supervised learning used to train machines so Can develop predictive data model.

⑥ Working diagram



Working of Supervised ML.

⑦ SML described as "learn from the past and predict the future".

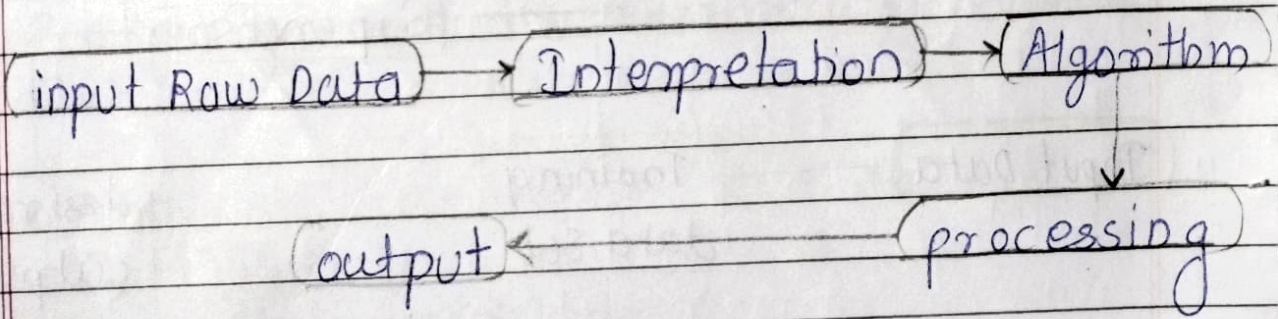
* II. Unsupervised Machine learning

① In unsupervised learning there is no Supervisor to observe the process.
② In Unsupervised learning, the network is provided with inputs but not with desired output.
③ In unsupervised learning data is Unlabeled.
④ Unsupervised learning is Used to create recommedation System.
⑤ following are the Some examples of Unsupervised machine learning applications
a) Social network analysis.

b) Image Segmentation

c) Google News.

⑥ Working diagram of Unsupervised learning :



5. Short Note on K-Nearest Neighbors (KNN).

→

① K-NN stands for K-Nearest Neighbors.
② KNN is one of the Simplest ML algorithm.
③ KNN based on supervised learning algorithm
④ KNN does not learn anything in training period hence KNN is called as lazy learner.
⑤ KNN model is used for classification as well as regression.
a. In case of classification, new data points get classified in a particular class.

b. In Case of regression, new data gets labeled based on the avg. value of KNN.

⑥ Advantages
① Algorithm is simple and easy to implement.
② There is no need to build a model.
③ algorithm is versatile, that is it used for classification & regression also search.

⑦ Disadvantages.
① It has Poor interpretability
② High memory requirment for storing the results.

6. Naive Bayes Algorithm (short note)

→

① Naive Bayes algorithm is a Supervised learning algorithm.

② Naive Bayes is a classification technique based on Baye's Theorm.

③ Naive Bayes algorith model is easy to build & Particularly Useful for all very large data set.

④ The principle behind Naive Bayes theorm is By Bayes theorm or Bayes Rule.

⑤ Application of Naive Bayes Algorithm
a. it is Used for credit Scoring
b. it is Used in medical data classification
c. it is Used in text classification Such as Spam filtering etc.

⑥ Baye's Theorem
a. also known as Bayes' Rule or Bayes' law.
b. Bayes' law is Used to Calculate the Conditional probability.

c. Bayes theorm represented as :

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \qquad P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

⑦ Advantages
i. it is easy and fast to predict class of test data set.

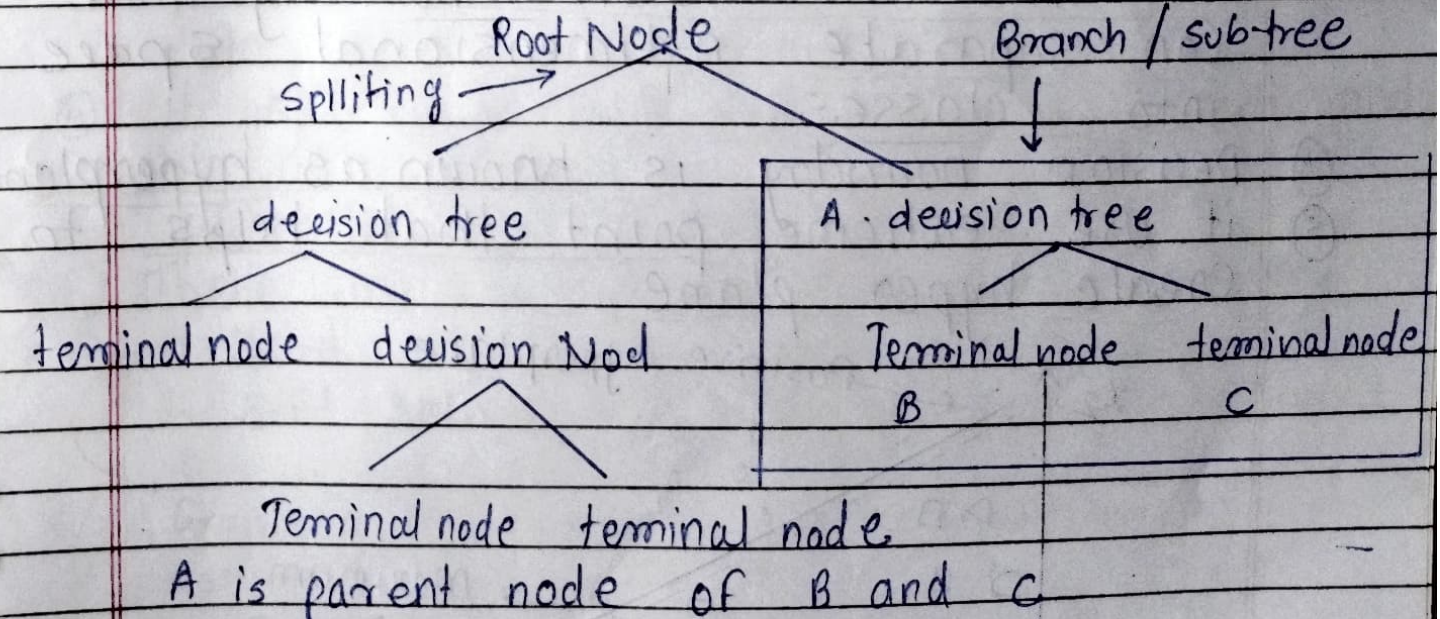
2. it also Performs well in multi-class prediction
3. it is effortlessly trainable, even with a small available data set.

⑧ Disadvantages
1. Parametically dependent
1. assumes independence of features.

7. short note on decision tree.

→

① A decision tree also called as prediction tree.
② A decision tree uses a tree structure to specify sequences of decisions (conditions)
③ It is called decision tree because, structure is similar as tree.
④ decision tree starts with root node & then expand its branches.
⑤ decision tree simply asks a question based on answer is (yes / no).
⑥ structure of decision tree :-



A is parent node of B and C

6. Naive Bayes Algorithm (short note)

→

① Naive Bayes algorithm is a Supervised learning algorithm.

② Naive Bayes is a classification technique based on Baye's Theorm.

③ Naive Bayes algorith model is easy to build & Particularly Useful for all very large data Set.

④ The principle behind Naive Bayes theorm is By Bayes theorm or Baxyes Rule.

⑤ Application of Naive Bayes Algorithm
a. it is Used for credit Scoring
b. it is Used in medical data classification
c. it is Used in text classification Such as Spam filtering etc.

⑥ Baye's Theorem.
a. also known as Bayes' Rule or Bayes' law.
b. Bayes' law is Used to Calculate the Conditional probability.
c. Bayes theorm represented as :-

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \qquad P(A|B) = \frac{P(B/A)\,P(A)}{P(B)}$$

⑦ Advantages
1. it is easy and fast to predict class of test data set.

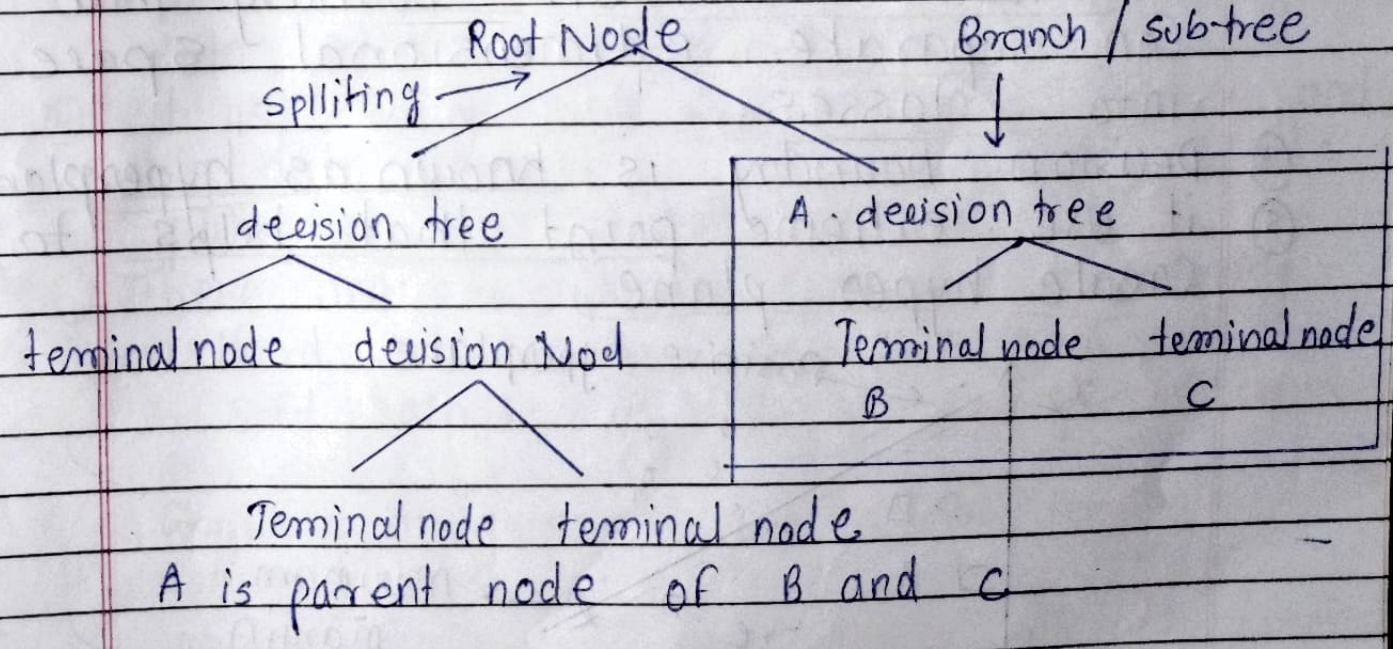
2. it also Performs well in multi-class prediction
3. it is effortlessly trainable, even with a small available data set.

⑧ Disadvantages
1. Paritically dependen
1. assumes in dependence of features.

7. short note on decision tree.

→

① A decision tree also called as prediction tree.
② A decision tree uses a tree structure to specify sequences of decisions (conditions)
③ It is called decision tree because, structure is similar as tree.
④ decision tree starts with root node & then expand its branches.
⑤ decision tree simply asks a question based on answer is (yes / No).
⑥ structure of decision tree :-



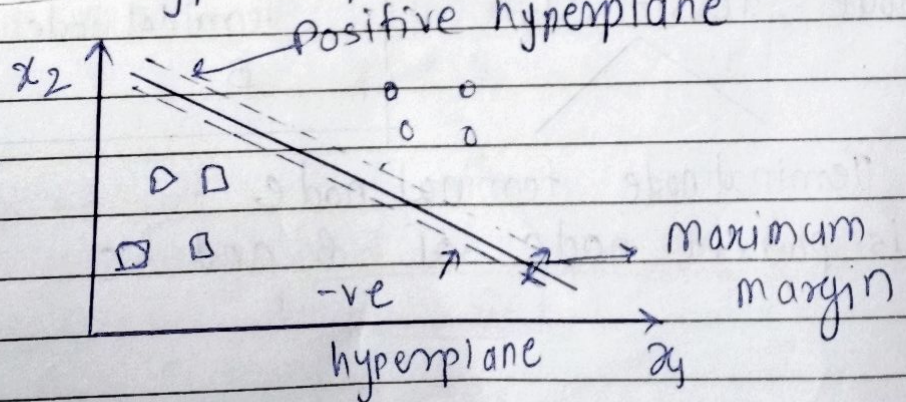
⑦ Advantages.
a. Simple to Understand and interpet
b. Able to handle both numerical and Categorical data.
c. requires little data prepration.
d. Performs well with large data set
e. fast and accurate.

⑧ Disadvantages
a. Trees Can be very non-robust
b. Poor Performance
c. Expensive

8. Write short Note on SVM.

→

① SVM stands for Support Vector Machine
② SVM is one of the most popular Supervised learning algorithm.
③ The goal of SVM algorith is to Create best line or decision boundry that Can separate n-dimensional space into classes.
④ Decision boundry is known as hyperplane
⑤ it use extreme point that helps to create hyper plane.



classmate
Date
Page

⑥ Advantages:
1. it works well with exact clear margin Sepration.
2. it is effective in high dimensional Spaces.
3. it uses subset of training points;

⑦ disadvantages
1. SVM does not directly provide probability
2. When data set has more noise it does not Perform very well.
3. When we have large data set then SVM have training time higher;

g. Cluster Analysis

→

① Cluster analysis is a data one of the best data analysis technique.
② cluster analysis is a method of Unsupervised machine learning;
③ cluster analysis is a task of dividing data points into a number of groups.
④ it is very useful for data mining and big data because it automatically finds patterns in data.
⑤ There are two main type of clustering
a. Hard clustering eg K-means.
b. Soft clustering. eg. fuzzy C-Means

Grouping the data items such that each item only assigned one cluster

Method of Grouping data items Such data item can exist in multiple cluster

10. Write short note on K-Means.

→

① K-means clustering is one of the Simplest and popular Unsupervised machine learning algorithm

② This algorithm involves telling the algorithms how many possible cluster there are in the dataset.

③ K-means is a partitional clustering algorithm.

④ Advantage

1. Simple easy to understand and to implement
2. it provide easily interpretable clustering results.
3. Fast and efficient to use

⑤ disadvantage

1. algorithm is only applicable if the mean is defined
2. do not work properly for large amount of data points.

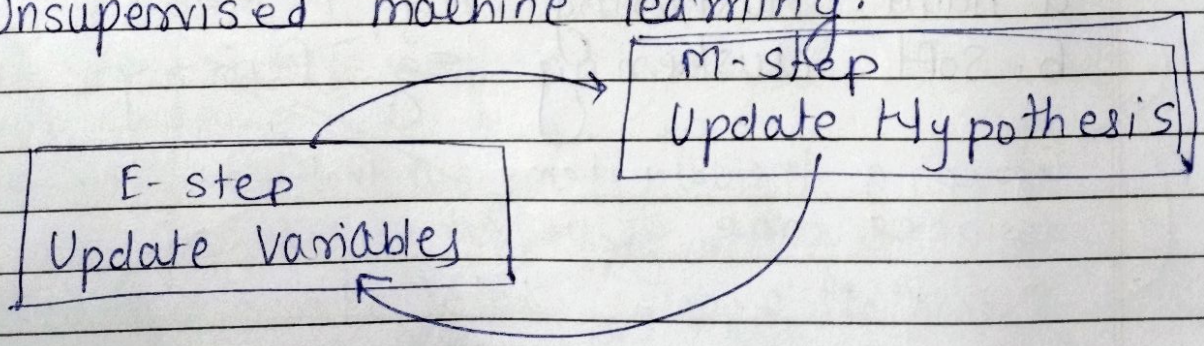
11. short note on EM-Algorithm.

→

① EM stands for Expectation-Maximization algorithm

② it is defined as the combination of various Unsupervised machine learning.

③



12. Association Rule Mining.

→

① Association Rule mining is a procedure which is meant to find frequent patterns, correlations, associations from data sets

② Association rule mining discovers strong association or correlation relationship among data

③ 1. Basket Data analysis
2. Cross marketing
3. Catlog design are the main applications of Association rule mining.

13. Apriori Algorithm.

→

① Apriori algorithm was proposed by Agrawal and srikant in 1994.

② Apriori is designed to operate on database containing transactions

③ Apriori algorithm uses "bottom up" aproch.

④ Advantages.
1. Efficient and effective with large item sets.
2. Easily implementable
3. provides real insight.

⑤ Disadvantage.
1. Assumes transitional database in memory
2. Requires many database scans
3. Slow
4. Higher runtime for execution.

14. What is Regression Analysis:

→

① Regression analysis is a set of statistical processes of ~~estimating~~ defining the relationship among variables.

② Regression analysis it is very useful and powerful tool technique.

③ Regression analysis has two main types

i] Linear Regression analysis:-

① linear regression analysis is one of the type of regression analysis.

② linear regression is used to predict the value of variable based on the value of another variable.

③ The variable you want to predict is called the dependent variable

④ The variable you are using to predict the other variables value is called independent variable.

⑤ General mathematical eqⁿ linear regression

$$y = ax + b$$

ii] Non-linear Regression analysis:-

① it is ~~form~~ regression analysis in which observational data are modeled by a function which is a non-linear combination of model parameters and depends on one or more independent variables.

4. Data analytics with R/WEKA machine learnin

1] What is WEKA?

→

① WEKA is an open source software provides tools for data Preprosseeing, implementation of machine learning algorithms.

② A WEKA is a Collection of machine learning algorithms.

③ A WEAA is Used for data mining task

④ WEKA tool contain data pre-processeing, classification, regression, clustering, visulization etc.

⑤ it is Suited for developing new machine learning schemes.

2] What is Data manipulation.

→

① Data manipulation is the modification of information to make it easier to read and to be more organized.

② When data Collection process is done by machine so many error are in the data but data manipulation remove that error & arrange data in Structured order.

③ example.
Data log is Sorted in alphabetical order to easy findout.

3] What is data Visulization?

→

① Data visulization is a technique Used for the graphical representation of data.
② Data Visualization Uses charts, histograms, graphs.
③ Data Visualization make it easy to Understand data or info.
④ Data visualization make it easy to recognize patterns, trends and exceptions in our data.

⑤ Advantage of data visualization in R:-
① R offers broad Collection of visulization libaries.
② R also offer data visualization in the form of 3d model.

⑥ Disadvantage of data visualization in R:-
① R is only perffered for data visuulization
② R Data visulization Using R is slow for large amount of data.

4] What is Pipe Operator?

→

① Pipe operator is a special operational function.
② Pipe operator wrap multiple functions together.
③ Pipe operator it is denoted as %>%.
④ Pipe operator make it easy when we need to perform Various operations on a dataset to derive the results.

5) What is data analysis.

→

① Data analysis is defined as a process of cleaning transforming and modeling data to discover Useful information for business decision making.

② The main purpose of data analysis is to extract useful information from data and making decision of business.

③ following are the Some time types of data analysis.

① Text analysis ② stastical analysis { descriptive, Inferential }

③ Diagonostic analysis ④ Predictive & Prespective.

④ Proces of data analysis / steps of data analysis proces

Identify business Problems ①
└ Data quisition ②
└ Process/clean data ③
└ Exploratry Analysis's ④
└ Generate model ⑤
└ Validate model ⑥
└ Visualize model ⑦

Text
stastical
Diagonistic
Predictive
descriptiv

Step
① Identity business problem
② data acquisition
③ Process / clean data
④ Exploratory analysis
⑤ model Generation & Validation.

* Remaining Important Questions *

* Remaining important Questions.

Q. What is statistical Modeling?

→

① stastical modeling is the process of applying a statistical alalysis to a dataset.

② stastical modeling is a mathematical representation of observed data.

③ example.
1) Public health data
2) Social media data.

④ steps of stastical modeling process
1) model selection
2) model fitting
3) model validations.

Q. What is Correlation & Write techniques for measuring Correlation.

→

1. correlation analysis is method of statistical evolution used to study the strength of a relationship bet^n two numerically measured, continuous variables.

2. Correlation analysis when we use we want to estabilish possible Connection bet^n variables.

3. For techniques for measuring correlation see chapter no. 2 Notes back.

Q. Diff^n Bet^n Population and Sample

→

Population	Sample
1. Population refers to the collection of all elements possecing common characteristics.	1. Sample means a Subgroup of members of population chosen for participation in the study.
2. it includes each and every Unit of group.	2. it includes only a handful of units of population.
3. Parameter is characteristic	3. Stastic.
4. it focuses on identifying characteristics	4. it focuses on making inferences about Population.
5. it is defined as total of the items Under Considerations.	5. it is defined as a proportion of the population selected
6. example - Automobiles with four wheels, people who Consume olive oil	6. example - 20 Cars form each make.

steps of Regression analysis
dif^{n} bet^{n} Correlation & regression

Q. What is Regression.

→

1. A regression is a stastical technique that relates a dependent variable to one or more independent variable.
2. Regression Significance is used to determine whatever wheather the relationship exits or not
3. example.
we can say age & height is described using linear regression model.

4. Type of Regression model

Regression ~~analysis~~

Simple — Multiple

5. I. Simple Regression:-

- It is used to estimate/identify a relationship bet^{n} dependent variable to a single independent variable.
- example - A relationship bet^{n}.
Crop & rainfall.

6. II. Multiple Regression

- It is used to identify a relationship bet^{n} dependent variable to a one or more independent variable
- example- A relationship bet^{n} salary & employee & their eprience eduction.

7. Regression analysis steps

1. Identify the statement of problem
2. choice of Relavant Variables
3. Collection of data on Relavant Variables
4. Specification of model
5. choice of method for fitting the data
6. fitting of model
7. Validation of model
8. Using the chosen model.

8. Diff^n bet^n Correlation and Regression

→

Points	Correlation	Regression
Meaning	Correlation is a corelationship bet^n two variables	Regression is a relationship bet^n dependent to independent variable
Usage	To represent linear relationship bet^n two variable	to fit a best line bet^n dependent to independent Variable.
Dependent & independent var	No difference	Both are different variables.
Objective	To find the numerical value bet^n variables	To estimate the value of fixed Variables.

Q. describe Hadoop?

→

① Hadoop is one of the best data analysis tool used in Big data.

② Apache Hadoop is an open source frame-work.

③ Hadoop is used to store process and analyze data which are very huge in volume.

④ Hadoop is used to efficiently store and process large datasets ~~ranging in size~~ from gigabytes to petabytes of data.

⑤ Hadoop allow clustering multiple computers to analyze massive datasets in parallel more quickly.

⑥ Hadoop is written in Java.

Q. Write down tools used for big data

→

① NoSQL - Database, MangoDB, Cassandra.

② MapReduce - Hadoop, Hive, Pig, Kafka.

③ Storage - Hadoop distributed file system, S3

④ Server - EC2, Google app engine

⑤ Processing - R, Yahoo! pipes etc.

Above all tools used for big data.

Q. Why hadoop is used for big data analysis?

→

① Hadoop makes it easier to use all the storage and processing capacity in cluster servers, and to execute distributed process agai-nts huge amount of data.
② Hadoop provides building blocks on which other services and applications can be build.
③ Hence Hadoop is used in big data analysis.

* What are the features of hadoop.

→

① Hadoop is open source
② Hadoop is easy to use.
③ Hadoop is very cost-effective.
④ Hadoop is faster in Data processing.
⑤ Hadoop provides feasibility.

Q. Explain need for morden Comporate Sector to go for Big data Strategy.

→

① Companies use big data strategy in their systems to improve operations, provide better customer service, and take other more actions that ultimately can increase revenue and profits.
② Big data processes large amount of data and help to grow organisation and organisation wealth. hence morde Corporate Sector

need to go for Big data strategy.

Q: What is training. testing and cross validation of machine learning models.

→

① Training a model means learning good values for all the weight and bias from labeld.
② The testing stastical based test for machine learning model.
③ The team would train, test and validate their model on real world data.
④ Cross-validation is a technique for validating model efficiency by training it on the subset of input data.

Q. Explain tools used in Big data.

→

I] Apache Hadoop-
it is open source framework from apache and runs commodiaty hardware. it is used to store process and analyze Big data.

II] Apache Spark -
it is a considered as a successor of Hadoop as it overcomes the drawbacks of its Spark unlike Hadoop supports both real time as well as batch processing.

III] Apache strom-

1. Apache storm is big data analysis to tool.
2. Apache storm is a free and open Source distributed realtime Compulation System.
3. Apache storm is efficiently process Unbounded storm of data.
4. Storm is designed to process vast amount of data in a fault-tolerant.

Q. What is Overfitting and Underfitting.

→

① Overfitting means that your model makes not accurate predictions. in case of overfitting train error is very small and valltest error is large.
Overfitting can happen due to low bias to high variance.

② Underfitting means that your model makes acurate but intially incorrect prediction. in case of Underfitting, train error is large and val/test error is large too.
Underfitting can happen due to high bias to high Variance.

Q. Explain Deep learning.

→

① Deep learning is a machine learning techique that teaches Computers to do what comes naturally to humans.

② Deep learning is a key technology behind driverless cars.
③ Deep learning is Subset of machine learning.
④ Applications
1. Self driving Car
2. Natural lang processing
3. Virtual assistant.
4. Visual recognisation
5. Fraud detection.
6. Healthcare.

Q. Explain Role of machine learning algorithm.
→
1. Machine learning algorithms uses historical data as input to predict new output values.
2. A machine learning is a System by which AI System Conduct its task.
3. The two main process of machine learning algorithm are classified classification and regression.

Q. Short note on Random forest.
→ ① it is Supervised machine learning algorithm
② it used widely in classification and regression problems.
③ it builds decision tree on different Samples and make decision tree.
④ it provides better results for classification problems.

Q. Understood maching learning

1. The Broad - ML Predeeting things based on what they done in the past.

2. The Practical - ML finds the relationship bet^n in your data.

3. The Technical - ML use stastical methods to predict value.

4. The mathematical - ML attempts to predict the value y given an input.