Data Scienc

YOUNG INDIAN

Date :
Page No. :

data Science

A feild that gives insights from structured and unstructured data, using diffrent Scientific methods and Algorithms. Consequently help in generating insights, making Predictions, driving data driven Solutions.

Need

we need data Science to abundant Complex data into insights enabling informed decisions, innovations and Solutions to challenges across various feilds like healthCare, finance and enviornmental Conservation.

Life Cycle of data Science.

- The things go into the formation of data Science.

Buisness Understanding ⇒ Data Collection ⇒ Data Preparation ⇒

Model deployment ⇐ Model building ⇐ Exploratory data Analysis and Maintanence.

- Buisness understanding - first step - defining the buisness Problem because a well defined buisness statement define a specific goal and is the key to Success of the Project.

- Data Collection - Second step - Data acquisition - Essential step in data Science because data needs to be relevant to solve the buisness problem Correctly.

- Data Preparation - Third step - A Crucial step in DS Project as it helps clean and bring the data into shape which is required for further Analysis and Modeling. Also refered as data cleaning.

TEACHER SIGNATURE ....

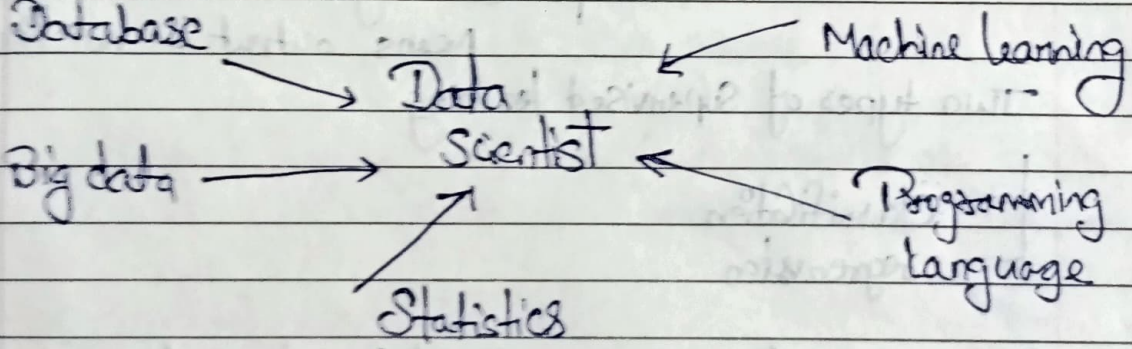
we treat issues like missing values, outliers and also transform the data into required format.

Exploratory data Analysis - Fourth step - Data is Analysed using Summary Statistics and graphically to understand key Patterns. This simplest step but highly effective in creating some useful patterns.

Model Building - Fifth Step - Two types -

1) Discriptive Analytics
- Involves insights based on historical data
2) Predictive Modeling
- Involves future predictions.

Model deployment and Maintenance - model is ready to deploy in real world. The deployment can occur offline on the web, on the cloud, any Android or ios app.



Python

- Pandas - Data Analysis and Manipulation
  - Provide fast, flexible data structure, Numerious functions and methods
- Numpy - Scientific Computations
  - Contains multidimensional array and matrix datastructures
- Seaborn / matplotlib - Data visualisation
- PowerBI / Tableau - Visualization tools

Machine learning

- ML Algorithms
  - Supervised - when we worked with labelled data means output is known
  - unsupervised - Algorithm has to make guess means output is not known

Two types of Supervised learning

- Classification
- Regression

- HyperParameter Tuning - Model optimization
  - they Control the behaviour of Algorithm.
  - Done to improve results

TEACHER SIGNATURE ....

Scikit Learn – Machine learning library supports both supervised and unsupervised learning
- Provides various tools for model fitting, data preprocessing, Model selection, Model evaluation

In R – Mlr Package / Caret used

for Big data Traditional machine learning Algorithms are not used we use deep learning.

Deep learning:

- Neural networks: working with diffent kind of layers and neural networks then deep learning is preffered
  - There are set of Algorithms which are modeled you know loosely after human brain to recognize patterns
- Algorithms
  - CNN → Convolution Neural network
  - RNN → Recurrent Neural network
  - LSTMs → long short term memory Networks
- Natural language Processing – A subfeild of AI Conferned with interaction b/w Machine & humans

- Libraries – → Tensorflow
  → Pytorch

TEACHER SIGNATURE ....

Data Science

YOUNG INDIAN

Date :
Page No. :

Data Science

A feild that gives insights from structured and unstructured data, using different Scientific methods and Algorithms. Consequently help in generating insights, making Predictions, driving data driven Solutions.

Need

we need data Science to turn abundant Complex data into insights enabling informed decisions, innovations and Solutions to challenges across various feilds like healthCare, finance and environmental Conservation.

Life Cycle of data Science.

- The things go into the formation of data Science.

Buisness Understanding ⇒ Data Collection ⇒ Data Preparation ⇒

Model deployment ⇐ Model building ⇐ Exploratory data Analysis and Maintenance.

- Buisness understanding - first step - defining the buisness Problem because a well defined buisness statement define a specific goal and is the key to Succes of the Project.

- Data Collection - Second step - Data acquisition - Essential step in data Science because data needs to be relevant to solve the buisness problem Correctly.

- Data Preparation - Third step - A Crucial step in DS Project as it helps clean and bring the data into shape which is required for further Analysis and Modeling. Also refered as data cleaning.

TEACHER SIGNATURE ....

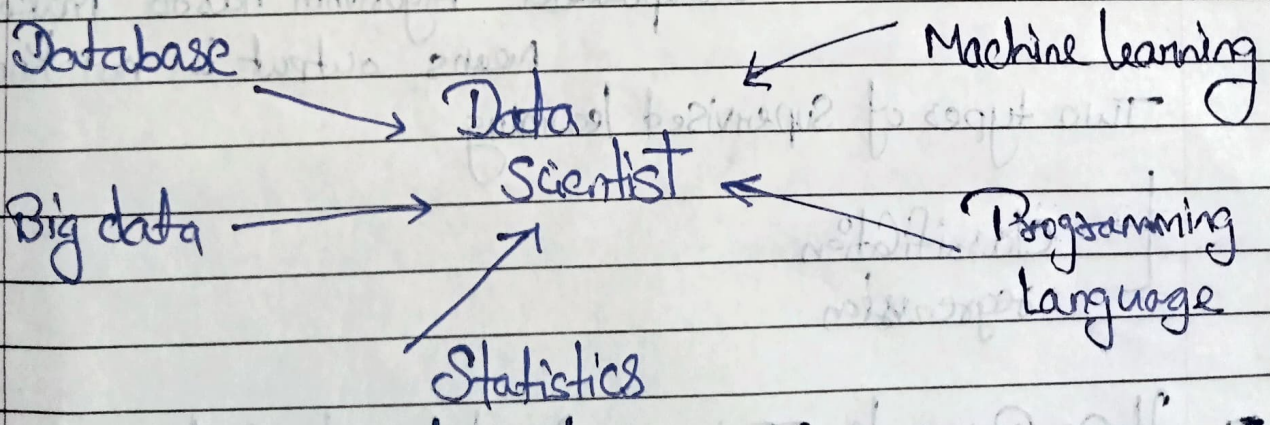
we treat issues like missing values, outliers and also transform the data into required format.

Exploratory data Analysis - Fourth step - Data is Analysed using Summary Statistics and graphically to understand Key Patterns. This simplest step but highly effective in earthing some useful patterns.

Model Building - Fifth step - Two types -

1) Descriptive Analytics
- Involves insights based on historical data
2) Predictive Modeling
- Involves future predictions.

Model deployment and Maintenance - model is ready to deploy in real world. The deployment can occur offline on the web, on the cloud, any Android or ios app.



Statistics

Y YOUNG INDIA

Date :
Page No. :

- Identify ⇒ Predict ⇒ Design ⇒ Estimate

Descriptive Statistics:

↳ It is a Summary Statistics that quantitatively Summarises features

- Variance
- Mean
- Mode
- Median
- Skewness

Inferential Statistics

↳ Inferring Something from small data assuming that this will also apply to large data.

- Hypothesis
- Probability
- Regression Analysis
- chi-Square
- Anova

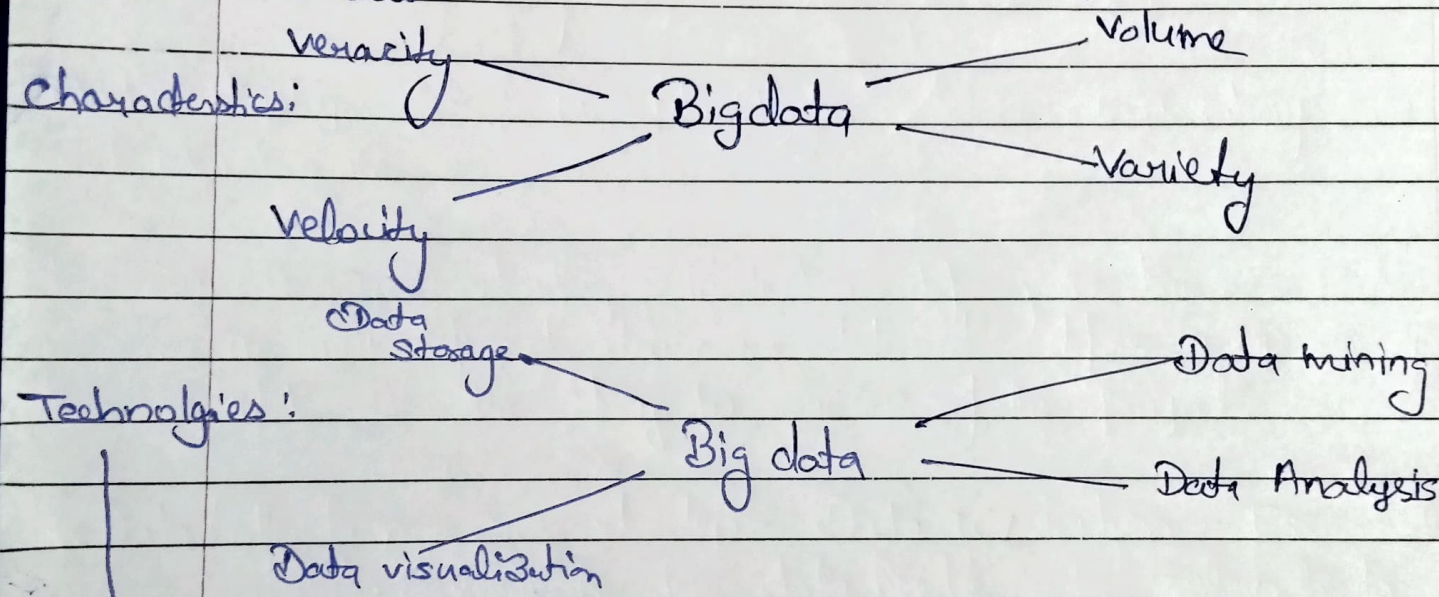
Databases

- Non-Relational DBS
- follows no pre-set Schema
- MongoDB, Apache Cassandra

- Relational DBS
- Data is related, in tabular format, easy.
- MySQL, Oracle

Big data

- Large and Complex data and Companies are dealing with huge volume of data especially from new data sources.



for data Storage — hadoop - process data in distributed enviornment. it can store & Analyze data, high speed, low cost

→ Excel, Web Scrapping, Linux, Git, Cloud

TEACHER SIGNATURE ....

Topic _____ Date ___/___/___ Page _____

Data

Data refers to facts and Statistics Collected together for reference and Analysis.

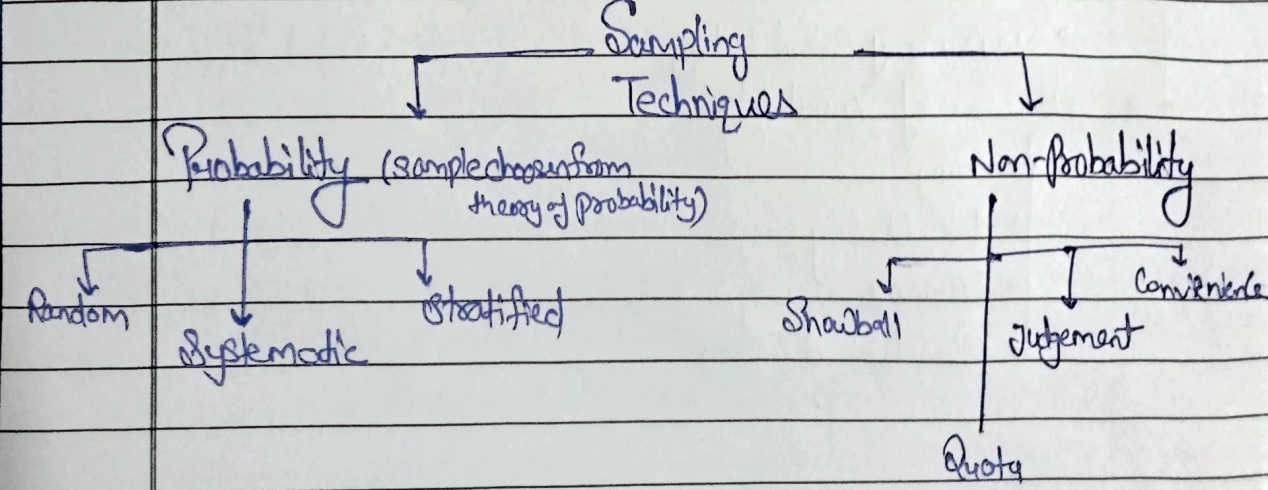
- Qualitative data → deals with characteristics & descriptors
  - Nominal - no inherent order
  - Ordinal - ordered series
- Quantitative data
  - Discrete - hold finite no. of possible values
  - Continuous - hold infinite no. of possible values

Statistics for data Science

Statistics is an area of Applied mathematics Concerned with the data Collection, Analysis, interpretation, Presentation,

Terminologies:

- Population: A Collection or set of individuals or objects or events whose properties are to be Analyzed.
- Sample: A Subset of population is Called 'Sample'. A well Choosen Sample Contain most of information about a Particular population Parameter.



VATS

Teacher's Signature .....

Topic _____ Date ___/___/___ Page _____

Random Sampling

- Each member of Population has equal chance of being selected in the Sample

Systematic Sampling

- Every nth record is chosen from Population to be a Part of Sample.

Stratified Sampling

A Stratum is a Subset of Population that shares at least one Common characteristics.
Random Sampling is used to select a Sufficient no. of subjects from each stratum.

Types of Statistics

Descriptive Statistics
- uses the data to provide description of the Population either through numerical calc, graphs.

→ Measure of Central tendency
Measure of Variability
- range
- Inter Quality range
- Variance
- Standard deviation

Inferential Statistic
- Makes inferences and Predictions about a Population based on Sample of data taken from Population.

Teacher's Signature .....

Mean:

Measure of Average of all the values in a sample is called Mean.

Median:

Center of value, arranged in Asc order.

Mode:

The most occurrent in sample set.

Range:

- The given measure of how spread apart the values in dataset are.

$$\text{Range} = \text{Max}(x_i) - \text{Min}(x_i)$$

Quantiles:

- Tell us about the spread of dataset by breaking dataset into quarters.

Variance:

- how much a random variable differs from its expected value.

$$s^2 = \frac{\sum_{i}^{n+1} (x_i - \bar{x})^2}{n}$$

Standard deviation:

- Difference btw each element from the mean

Deviation $= (x_i - \mu)$

Topic _____ Date ____/____/____ Page _____

Example.

Data:- 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4

- Mean:- $\frac{9+2+5+4+12+7+8+11+9+3+7+4+12+5+4+10+9+6+9+4}{20}$

$$\boxed{\mu = 7}$$

$(x_i - \mu)^2 \Rightarrow (9-7)^2, (2-7)^2, (5-7)^2$ ------ So on

we get,

Sum = $\{4, 25, 4, 9, 25, 0, 1, 16, 4, 16, 0, 9, 25, 4, 9, 9, 4, 1, 4, 9\}$

Standard deviation,

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$$

$$= \sqrt{\frac{1}{20}(\text{Sum})}$$

$$\sigma^2 = 8.9$$

$$\boxed{\sigma = 2.983}$$

Teacher's Signature .....

Baye's Theorem:

- Shows the relation between one Conditional probability and its inverse.

- $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$

  - $P(A|B)$ → Posterior
  - $P(B|A)$ → likelihood matrix
  - $P(A)$ → Prior
  - $P(B)$ → Evidence

- Conditional Probability, $P(A|x) = \frac{P(A \cap x)}{P(x)}$

- Interval Estimate:

  An interval or range of values used to Estimate a Population Parameter is called interval Estimate.

- Margin of Error:

  The greatest possible distance between the Point Estimate and value of Parameter it is estimating.

$$E = Z_c \frac{\sigma}{\sqrt{n}}$$

Hypothesis testing

- used to formally check whether hypothesis is accepted or rejected.

Topic _____ Date ____/____/____ Page _____

State hypotheses: Stating a null or alternative hypotheses
formulate Analysis plan: Construction of Analysis plan.
Analyses of Sample data: Calculation and interpretation of
Test statistics as described in Analysis plan.
Interpret Result: - Application of desicion rule described
in Analysis plan.