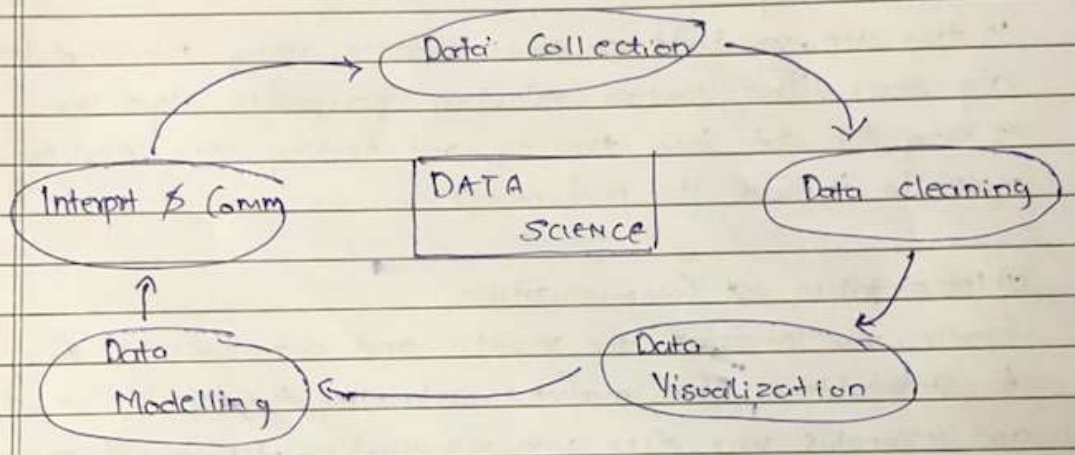
DATA SCIENCE - UNIT 3

Q. State and Explain 5-fundamental steps of Data Science Process

ANS Data science is the art of science of extracting meaningful insights from data. It blends various fields, including statistics, CS and domain expertize to analyze large volume of data and uncover patterns.



(i) Data Collection

This is where you gather raw data from various sources. This can be includes databases, API, web scraping, surveys and more. It's crucial to ensure that data is collected is relevant and of high quality.

(ii) Data Cleaning.

Raw data is often messy and filled with errors, missing values or duplicates. In this step, you clean and preprocess the data to make it ~~outside~~ usable. This might involve filling in missing values, correcting errors and normalizing the data.

(refer log table)

(iii) Data Exploration & Visualization

In this the exploration of data is done to understand it's structure, patterns and relationships. Visualization tools are like chart and graph helps you spot trends and insight that might no be obvious from data alone.

(iv) Data Modelling

In this step, you build an train models using machine learning algorithms. This involves selecting appropriate algorithm, spitting the data into training and testing sets, and tunning the model to improve its performance.

(v) Intepretation & Communication

Finally, you interpret the results and communicate your findings to stakeholders. This involves explaining the insights in a clear and actionable way, often with visualization, reports & presentation.

Ex: Imagine you run a small e-commerce store and want to understand your customer's purchasing behaviour.

(i) Gathering data on customer purchases from your store's database.

(ii) Remove any incomplete or duplicate entries from your dataset

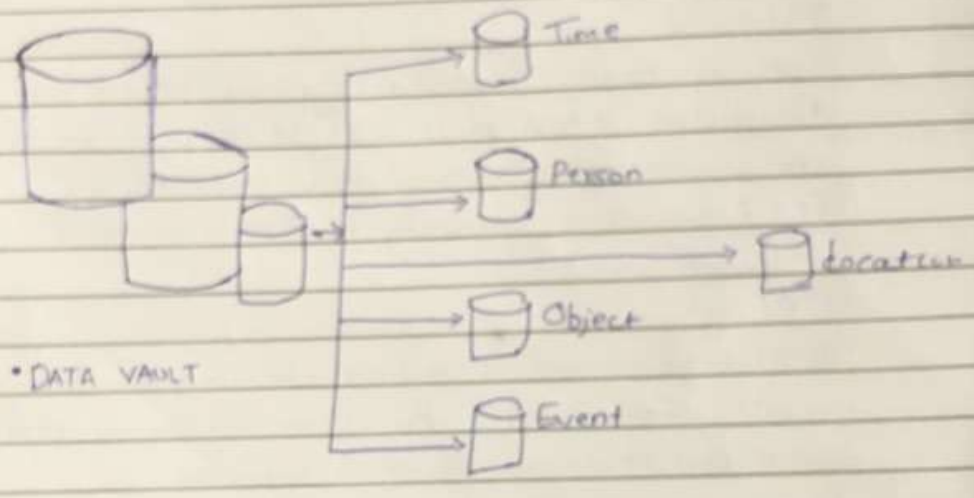
(iii) Visualize the data to see which products are most popular

(iv) Use a ML algo to predict which products a customer is likely to buy next based on their past purchases.

(v) Discover that customers who buy product A are very likely to also buy Product B. Use this insight to create targeted marketing campaigns & recommendation.

Q: Time - Person - Object - Event - Location Data Vault

Ans: TPOLE Data vault, it is an powerful framework for organizing and analyzing data, especially ~~our~~ when dealing with an complex and interconnected information.



TPOLE

(i) Time

Captures the temporal aspects of events. This includes dates, times and periods. It helps in tracking when event occur, enabling trend analysis over time

Ex: A sales transaction on March 15, 2023, at 3:00 PM

(ii) Person

It involves the individuals, groups or entities associated with the event. Essential for understanding who is involved in various events or transaction

Ex: Abu Sufiyan, a customer making a purchase.

Critical value of ... f(n log table)

(iii) Object

It Represents the items or entities Central to the event. Provides context about what is involved in the event.

Ex: An Vintage Car is been purchased

(iv) Location

Captures spatial Aspects, such as where had the event or the transaction toked place. It is crucial for geographic analysis and understanding where the event happen.

Ex: A transaction occuring at a store in Central Park, Mumbai

(v) Event

Describe occurance or actions, serving as the core of the data model. Encapsulates the activities being recorded, providing a comphrehensive view of what happened.

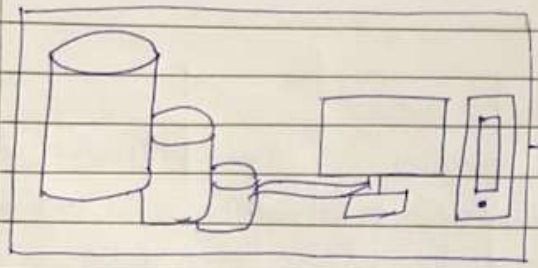
Ex: The actual sales transaction linking the respective Time, Person, Object and Local Entries.

The TPOLE can be written or displayed as:

- Sales Transaction on March 15, 2023 at 300pm by customer AbuSufiyan, purchasing an Vintage Car at a Central Park Store in Mumbai, documenting the purchase event with all valid Papers.

Q. Building an Data Warehouse

Ans. Building an Data Warehouse is like creating an central hub for all of your organization's data.



- To build an Data warehouse all these process and steps are taken in consideration

(i) Req Analysis
(ii) Data Modelling
(iii) ETL Process
(iv) Data Integration
(v) Performance Testing
(vi) Report and analysis

(i) Required Analysis

First you need to understand, what data you need and what you want to achieve with it. This involves talking to stakeholders, identifying data sources and defining key matrics and goals. It sets the foundation for your project.

(ii) Data Modelling

You need to design an data model, which act as an blueprint for how the data will be organized. This involves creating diagrams that represents the structure of data, including facts and dimension tables. This ensures data is organized locally.

Critical value
(In log table)

(iii) ETL Process

Extract, Transform, Load, after data model designing you start the ETL Process. You extract data from various sources, clean and transform it to ensure it's accurate and consistent, and load in into the data warehouse.

(iv) Data Integration

Once the data is loaded, you integrate it from different sources to provide a unified view. This involves mapping and linking related data, ensuring consistency. This step make it easier to analyze the data comprehensively.

(v) Performance Tuning

Optimizing the data warehouse for performance is crucial. This involves indexing and query optimization to ensure that you can retrieve data quickly and efficiently.

(vi) Reporting & Analysis

Finally, you create reports and dashboards to visualize the data and gain insights. This involves developing tools and visualizations that helps stakeholder to understand and interpret the data.

Ex: Imagine you run an online bookstore and want to understand customer buying patterns.

(i) Identify data from sales transactions, cust info and inventory.

(ii) Design schema with tables for sales, Cust - Books - Time

(iii) ETL : E = Extract sales data, cust info and inventory details

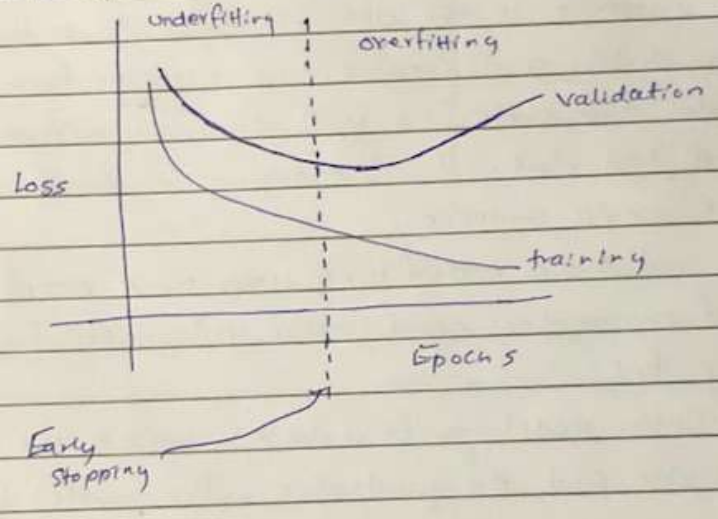
T = Transform data by cleaning and normalizing it

L = load data into warehouse.

(iii) Combines sales, customer and book data to provide a unified view of customer buying patterns
(iv) Optimize for queries, such as indexing frequently queried column
(v) Create Db to visualize top selling-books, cust demographic and sales trend.

Q. OverFitting and UnderFitting

Ans: Overfitting and underfitting are key concepts in machine learning, representing two extremes of model performance that should be avoided for optimal truth.



• OVERFITTING:

Overfitting occurs when a model learns not just the underlying patterns in the training data but also the noise and random fluctuation. A model that overfits will show high accuracy on training data but low accuracy on test data. overfitting happens when the model is too complex, such as having too many parameters or layers, which allows it to fit too closely to training data.

Critical value
(ln log table)

To prevent and avoid overfitting, use simpler models, prune complex models, employ regulation techniques like L1 and L2 regulation and validate the model using techniques like cross-validation.

Ex: If you train a decision tree to classify whether an email is spam, and the tree is too deep, it might learn specific, irrelevant details unique to the training emails rather than general spam indicators. As a result, it performs poorly on new emails

• UNDERFITTING:

Underfitting happens when a model is too simple to capture the underlying patterns in the data. It fails to perform well even on the training data, and consequently also performs poorly on new data. A model that underfits will show low accuracy on both training and test data. It's like trying to summarize a complex book with an single sentence.

To prevent and avoid underfitting, use more complex models, increase the number of parameters, and ensure adequate feature engineering and training time.

Ex: Using A linear model to fit a dataset where the relationship between variable and ~~its~~ is quadratic. The model fails to capture the curvature and thus performs poorly.

Balancing Two: The goal in machine learning is to find a balance between overfitting and underfitting.

Q. CROSS VALIDATION TEST.

Ans Cross validation test is a technique used in ML and statistics to evaluate the performance of a model. It helps to ensure that the model generalizes well to unseen data, rather than just performing well on the data it was trained on.

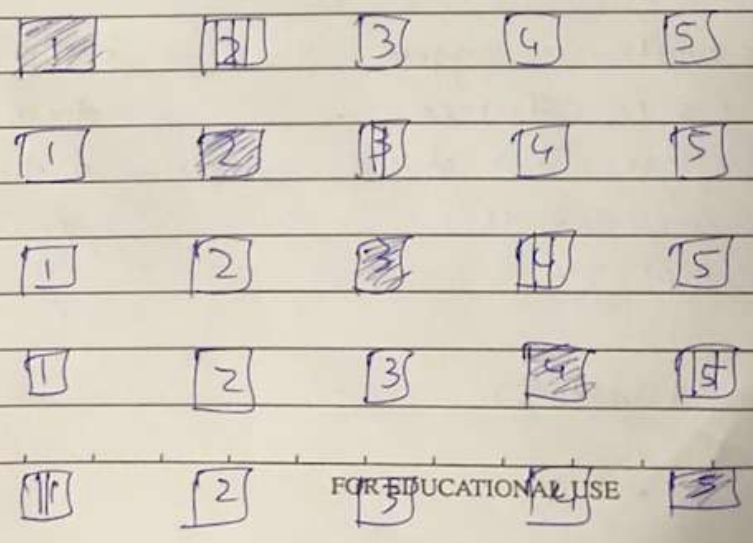
WORKING:

(i) Data Spitting: The original dataset is divided into smaller subsets. One common method is K-fold cross validation, where the data is split into k-equally sized folds.

(ii) Training & Testing:
The model is ~~tester~~ trained on k-1 folds and tested on remaining fold. This process is repeated k-times, each time with a different fold as the test set.

(iii) Performance Averaging:
The performance of model is evaluated in each iteration. The results are averaged to produce a single performance test.

Ex: For a 5-fold cross Validation.



Critical value
(ln log table)

- Test set
- Training set
- Validation set

(i) Split the data into 5 parts
(ii) Train the model on 4 parts and test on 5th
(iii) Repeat this process 5 times, each time with a different parth as the test set
(iv) Average te results to get final performance matrix

Q Hypothesis.

Ans: A Hypothesis is an statement or an assumption about population parameter that you want to test. It's often an educated guess based on prior evidence

Hypothesis

NULL HYPO

- This hypothesis suggets there's no difference no effect. Think of it as an "default" assumption.

ALT HYPO

- This hypothesis suggests there is an effect or a difference. It's what you want to prove

- Test in Hypothesis.

CHI SQUARE TEST.

A chi square test is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if a difference between observe data and expected data is due to chance of if its is due to relatiship btwn variables.

Ex:

Gender	Action	Comedy	Total
Male	40	10	50 — V2 (value 2)
Female	20	30	50
Total	60 — V1 (value 1)	40 — V1 (value 1)	100

Null HYP = No Associ Btwn G / M

Alt : Asso Betwn G & M

Soln:

$$\text{Form} = \frac{\text{Total 1} \times \text{Total 2}}{\text{Grand Total}}$$

$$= \frac{50 \times 60}{100} = 30 \qquad \frac{50 \times 40}{100} = 20$$

$$= \frac{50 \times 60}{100} = 30 \qquad \frac{50 \times 40}{100} = 20$$

Expected Value =

Gendre	~~Male~~ A	~~Female~~ B
male	30	20
Female	30	20

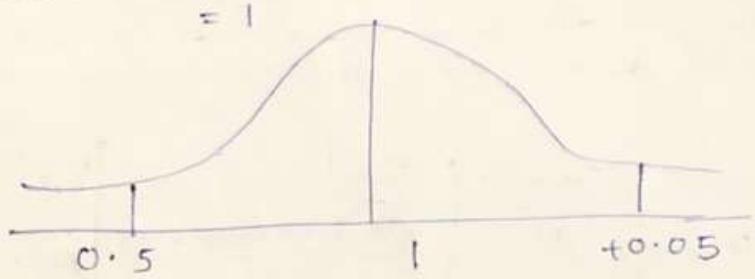
Chi square Test Statistics

$$\chi^2 = \frac{\Sigma (O - E)^2}{E}$$

Observed v	Expected E	O - E	(O - E)²	χ^2 - chi
40	30	10	100	0.3
20	30	-10	100	0.3
10	20	-10	100	0.2
30	20	10	100	0.2
				1

Critical Value of 1 = 2.706
(in log table)

Degree of Freedom = Col - 1 × row - 1
= 2 - 1 × 2 - 1
= 1



∵ x^2 calculated value (1) > x^2 tabular value critical (2.706). ~~Null Hypo Rejected~~

∴ There is no Association btwn G & M

If calc value is less than tab value we will accept Null hypo

Q T-test

Ans: It is an statistical tool used to compare the means of two or more group. It's ratio that measure significance of difference btwn means of groups while taking their variance etc..

Ex: Group 1 [Tradi Method] : [55, 60, 65, 70, 62]
Group 2 [New Method] : [68, 75, 80, 85, 78]

S1: Null : No relation btwn G1 & G2
Alt : Relation btwn G1 & G2

S2: G1 : $\bar{x}$ = 62.4
G2 : $\bar{x}$ = 77.2

S3: variance :
G1

$$= \frac{(\text{value} - \bar{x})^2 + (\text{value} - \bar{x})^2 \ldots (\text{value} - x_n)^n}{n-1}$$

$$= \frac{(55-62.4)^2 + (60-62.4)^2 + (65-62.4)^2 + (70-62.4)^2 - (62-62.4)^2}{5-1}$$

$$= \frac{(-7.4)^2 + (-2.4)^2 + (2.6)^2 + (7.6)^2 + (-0.4)^2}{4}$$

$$= \frac{54.76 + 5.76 + 6.76 + 57.76 + 0.16}{4}$$

$= \frac{125.2}{4}$

$= 31.34$

Variance σ_2:

$$= \frac{(value - \bar{x})^2 + (value - \bar{x})^2 \ldots (value - x_n)^n}{n-1}$$

$$= \frac{(68-77.2)^2 + (75-77.2)^2 + (80-77.2)^2 + (85-77.2)^2 + (78-77.2)^2}{5-1}$$

$$= \frac{(-9.2)^2 + (-2.2)^2 + (2.8)^2 + (7.8)^2 + (0.8)^2}{4}$$

$$= \frac{84.64 + 4.84 + 7.84 + 60.84 + 0.64}{4}$$

$$= \frac{158.8}{4}$$

$= 39.7$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{S_1}{n_1}\right) + \left(\frac{S_2}{n_2}\right)}}$$

$$= \frac{62.4 - 77.2}{\sqrt{\left(\frac{31.3}{5}\right)^2 + \left(\frac{39.7}{5}\right)^2}}$$

$$= \frac{62.4 - 77.2}{\sqrt{\frac{31.3 + 39.75}{5}}}$$

$$= \frac{-14.8}{14.2} = -3.92$$

Degree of Freedom $= n_1 - 1 + n_2 + 1$

$= 5 - 1 + 5 - 1$

$= 8$

By Def Consi 8 - 0.05 (in log table)

$= 2.306$

Cal > Tabular

∴ Rejected