**Day-1 Internship-2**

1) Data comes from

1. Real world
2. Database
3. Extract
4. Api
5. Market Research/Reachering

After extracting we load data a file structure in DB……..

Datapreprocessing-> extract EDA--🡪 vba🡪feature eng. 🡪feature selection🡪model🡪Hyper parameter tuning🡪deployement

TOOLS

1 sql

2 excel

3 Power Bi

4 Python library

5 Aws

* **SQL**

1. **Drop VS Delete.**

Drop:- delete all data & table

Delete:- delete a specific row / Column

Truncket:- delete all row & auto commit.

1. **Alter Vs Update**
2. **Where vs having**

Where:- normal condition(in,and,or)

Having:- aggregate function (sum,mean,max)

Select \* from table name where group by having order

1. Group by & distinct

Q:- 1) find duplicate & delete duplicate

2) give an example of all cluses

3) create a table using all constraint

(primary key, foreign key, not null, default, check

4) find out age btn 19 to 70

5) find all dob who born 1990 to 1947

**Day-2**

JOIN:-

* **Natural join** (join using logical operator/\*/using prk & fk
  + Select t1.t2 from table(t1)
* **Inner join**
* **Outer Join** 
  + - Left common (left table with common data)
    - Right common (right table with common data)
    - Full join(left +right)
* **Cross**
  + Cartesian products
* **Self** 
  + A single table join with itself with or without using pk & fk.
* **Union** 
  + Union all (duplicate value fetch)
  + Union (duplicate value not phase)
    - Select col-1,col-2 from table(t1) where, group by, having, orderedby

Interview Assignment-2

1. Join & set Operator
   1. select a.team as team\_a,b.team as team\_b from ipl a join ipl b where a.team!=b.team;
2. Union & inner
   1. select a.team as team\_a,b.team as team\_b from ipl a inner join ipl b where a.team!=b.team;
   2. SELECT \* FROM ipl a UNION SELECT \* FROM team\_b;
3. Union all & full
   1. select a.team as team\_a,b.team as team\_b from ipl a Full join ipl b where a.team =b.team;
   2. SELECT \* FROM ipl a UNION ALL SELECT \* FROM team\_b;
4. Difference between cross & full join.
5. Create a IPL table using self join…
6. Take 2 table one is students other is
7. Extract first 3 digit of phone number using Regular expression in NLP
   1. SELECT a.team as team\_a,b.team as team\_b FROM ipl a LEFT JOIN ipl b ON a.team=b.team ;

NLP for ML

Nltk

Token

Stemming

Lemmatization

Q-1) Difference between stopword & tokenization.

Package requirement for code

Tokinize 🡪 punkt package

Stop word 🡪 stopwords

**Day-3**

Function

1)Analytical function/windows functions

Aggregate function

* Sum
* Count
* Min
* Max
* Avg

Value

* Lead()
* Lag()
* Nth\_value
* Fast\_value
* Last\_value()

Rank

* Rank
* Dense\_rank
* Row\_number
* Present

1. Aggregate

* Sum 🡪 after select statement

Select sum col() from table having(cond.)

* MIN🡪minimum
* Max🡪maximum
* Avg🡪average
* Count🡪select count(\*) from table having 🡪 for row & column
* Select count(col) from table having() 🡪for column

1. Value
   * LEAD() 🡪 Forward

(select col, sum(col), lead() over(partition by id //ordered by salary) as LEAD(name);

* + LAG() 🡪 backward (display the value of adjacent row)

(LEAD(name)

* + Nth value()🡪

* + FIRST\_VALUE ()🡪
  + LAST\_VALUE ()🡪

1. Rank

* Rank() 🡪 same data in row skip hear
* Dense Rank() 🡪 same row data not skip
* Row number ()🡪 number of row
* Percent() 🡪 normal find the %
* Select rank() over(position by id orderby col as rank\_name;
* Select avg(practition by id) over(position by id orderby col as rank\_name;

2. Q-1 Difference between SQL & MS SQL?

Q-1) Write a querry to get the average star for each products every month.

Review

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Review-id | User-id | Submit-base | Product-id | Star |
| 6171 | 123 | 6/8/22 | 5001 | 4 |
| 7802 | 265 | 6/10/22 | 69852 | 4 |
| 5293 | 362 | 6/18/22 | 5001 | 3 |
| 6352 | 192 | 7/26/22 | 69852 | 3 |
| 4517 | 981 | 7/5/22 | 69852 | 2 |

Using group by

Approach

For each product\_id for each month note down their stars

ii) Find the averae of stars that are received by eatch product for that particular month

PYTHON

1. Data type
2. Loops
3. Functions
4. Oops
5. Exceptions
6. File handelling
7. Typecasting

**Day-6**

C

Regression

Import numpy,pandas,matplot-lib,

**Day-7**

C

Logistic Regression

Confusion matrix

**Day-8**

C

Decession tree

1. Linear regression
2. Logistic regression

Disadvantages:-

It was overfitting. (because it does have not any endpoint) it formats hierarchical methods.

Some matrices:

1. Information gain:- parent node – child node.
2. Gini index:- information gain is directly proportional to Gini index.
3. Gain ratio :- It was same as information gain
4. Entropy : - how many right answer and how many wrong answer.
5. Chi-square :- difference of every node.
6. Purity :- how many write answer wrote in entropy.

**DT for Regression**

1. Collection data OR retrieve data
2. Data preprocessing
3. EDA

* Numpy
* Pandas

1. VBA

* Matplotlib
* Seaborn 🡪 stastical

1. FeatureExtraction /engineering
2. FeatureExtraction / selection
3. Hyperparameter
4. Deployement

* AWS( s3,
* Azure
* GCP
* GitHub/herirchical

**Day-9**

Random Forest

1. Decession tree

Power Bi

Tenserflow

Piespark

Analysys🡪

Sentiment analysis

Twitter data set

Spam detection

Google search analysis

Web Scrapping selenium