Theory for Viva

Practical 1:-

Data wrangling ?
→ Process of cleaning, transforming & preparing raw data in structure suitable for analysis.

Big data → large collection of dataset & cannot process using traditional techniques.

numpy, Pandas
Matplotlib, Seaborn,
Scikit learn

label encoding:-
It converts the label in numeric form as most of library are compatible with the numerical data.

i) Numpy:- It is the fundament library used in python. You can Perform basic mathematical operation and also operation of array [add, mutiply slicie, index] and basic operation like floating & indexing

ii) Pandas:-
Pandas is an open source library providing high performance easy-to-use data structure and analysis tools.
It can used for manipulating the data frame.
- (update, delete, add colum to data frame)
- Handling & Identifying missing values
- Plotting.

iii) Scikit learn :- open source library featuring algorithms like random forest, k-mean clustering SVM etc.
It is use for classification : SVM etc
Regression (linear & logistic)
model selection [for finding accuracy & score].

i) matplot lib :- matplot lib is an Data visuali' library use to plot different visuliz' technique such as charts, graph, maps & plots

ii) Seaborn :- Seaborn is advance version of matplotlib it has better design and specialized visuliz' techique. It is high level because provide the simplified functi'..

Techique to handle missing data
fill in using fillna & interpolation
dropna
replacena

Practical 2 :-

Data transformation :- refers to ETL ie extract transform load.

Smoothing -> noises are removed
Aggregation -> presenting data in suitable format.
Generali' -> low level -> high level
Normaliz^n -> converting data variable in given range to normalize data
z- score
min-max
Decimal scaling.

Method to detect outlier

boxplot → visual

Z-score :-

scatter plot

inter quartile method

x = [1, 2, 2, 3, 4, 4, 4, 5]

mode ⇒ 4

unimodal → one mode

bimodal - 2 mods. (two values appear most with same high freq)

multimode, nomode

· describe() → mean, median, max, ming, count, 25%, 75%.

Practical 4:- Concept of linear regression

→ It is algorithm used for the prediction of the target variable based on independent variable or feature variable, used in supervised learning task.

It has linear relationship betw feature variable & target variable

It Gives the mean square error Betw actual value and predicted value which tells accuracy of model

It eq: $y = mx + b + e$

a. Gradient descent → method of (linear regression use to find optimal coefficient to a regression model by minimizing difference b/w predicte & actual value.

b. Least square method

Supervised and Unsupervised learning :-

(1) In Supervised learning → Algorithm learns from labeled data or known data whose each training set consist features & corresponding variable. based on this, it make on new pattern & insight from unseen data base.

(2) Unsupervised learning :- Algorithm learn from a Unlabel data which has no target variable to predict. It clusters similar data pts together goal is to find pattern and insights in data without any guidance.

Interpolate :- Interpolate is method or function used to fill the missing value in dataset by the estimations. method

1) linear :- estimate missing value bas on adjacent non-miss value

2) Polynomial :- use polynomial for on non-missing value to estimate value

Practical 5

What is logistic Regression :-

logistic regression is algorithm used for binary classification, which has only two possible outcome true/false Yes/no 1/0. It predicts the probability of occurance of data point of logic one particular class

method of logistic regression
① Binary logistic
② multinomial logistic regression :-

- linear regression give continous output
- logistic regression give constant output → sigmoid
  - Stock price, house price
  - Predicting where patient has cancer or not

Confusion matrix :- It's a matrix of table used to give performance of model
true +ve, true -ve, false +ve, false -ve.

Standard scalar → Remove mean & unit the SD
* as mean can be zero centred

Precision :- measure of accuracy of the instance
Recall :- measure of ability to correctly identify the +ve instance.

Practical 6:- Naive Bayes base of Bayes probability which ass all feature are independent of each other and calculate probability of feature occur in class.

Practical 8 → matplotlib.

Practical 9 → 8 10

Pyplot - 1 dimensional representation.

(11) →

what is Impla:-

It's an open source parallely processing SQL query engine data store in space-hadoop cluster.

All to run sql queries on big data providing fast process.

(12) Scala → Scalable programming language which has syntax similar to the java.

It's Object oriented & func° lang

Everything in scala is object & operation are method call.

Hadoop :- Open source framework for distributed storage which store a documents in chunk on different computer system & process large data set.

It' is scalabel

It is used for Big data analysis & Data noisy.

mapreduce → It is a model use in Hadoop which process and analyzes large data set in parallel clusters - break task in small - task & distribut across node in cluster.