③

Mean :-
arithmetic average of a dataset
• adding no. in dataset divided by obser

$\bar{x}$ (mean) $= \frac{\Sigma x}{N}$ → observation
N → no. of observation

② Mode :-
Most frequently occuring observation or value.

$$\text{mode} = l + \left[\frac{f_m - f_1}{2f_m - f_1 - f_2}\right] \times h$$

f_m = freq. possessed by modal class
f_1 = class before
f_2 = class after modal class
h = width

③ Median
Middle number in a dataset when no listed either assending and desending order.

for odd obs.
Media = $\left(\frac{n+1}{2}\right)^{th}$ observation

for even
Median = $\frac{(n/2)^{th}\text{ obs} + (n/2+1)^{th}\text{ obs}}{2}$

Scanned with CamScanner

Smoothing :→
Process that is used to remove noise from the dataset using some Algorithms
It helps in prediching the patterns

Aggregation :→
is a nothing but data collection is the Method of storing and presenting data in a summary format

Generalization :-
It converts low level data attributes to high-level attribute using concept hierarchy.
eg Age in Numerical form (20, 35) converted into categorical value (young, old)

Normalization :-
Data Normalization involves converting all data variables into a given range.

Min-Max normalization
Transfer original data linearly.

Z-score normalization :-
In Z score Normalization the value of an attribute (A), are normalized based on the Mean of A and its SD.

Scanned with CamScanner

Standard deviation: sqre of variance

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n-1}} \qquad \sigma^2$$

x - each value
$\bar{x}$ - Mean
n = no. of value

(4) Linear Regression:-
Machine learning algorithm based on supervised learning
It targets prediction values on the basis of independent variables.

Multivariate Regression:-
it concerns the study of two or more predictor variable

$$y = a + bx + cx_2$$

Mean square Error is
represent the error of the estimator or predictive model created based on the given set of observation in the sample

$$MSE = \frac{1}{n} \Sigma (y - \bar{y})^2$$

Sq of diff betn
~~actual & predicted~~

⑥
⑤ ④

Page No.
Date

⑤ Logistic regression

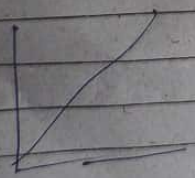
Logistic Regression :-

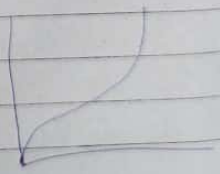
classification techniques are an essential part of Machine learning and data Mining application.

- is one of the most simply & commonly used machine learning algorithm for two class classification

statistical method for predicting binary class

linear	logistic
continuous o/p	custom o/p
e.g. House price	e.g. patient has a cancer or not
estimated using Ordinary Least Square (OLS)	estimated using Maximum Likelihood Estimation (MLE)





④

Scanned with CamScanner

Accuracy:- calculated as the number of correctly instance divided by total number of instance

$$acc = \frac{Tp + TN}{Tp + fp + TN + fN} = \frac{Tp + TN}{Pos + Neg}$$

Error Rate:- calculated as the no of incorrectly classified instance divided by the total number of instance

$$error = 1 - acc.$$

precision:-
calculated as the no. of correctly classified tve instance divided by the total no of instance which are predicted positive. it is also called confidence

$$precision = \frac{Tp}{Tp + fp}$$

Recall:-

$$= \frac{Tp}{Tp + fn}$$

Scanned with CamScanner

Page No.
Date

Types of Logistic Regression

Binary Logistic Regression:-
The target variable has only two possible outcomes such as spam or Not spam, Cancer or No Cancer.

Multinomial Logistic Regression:-
The target variable has three or more nominal categories such as predicting the type of Wine.

Ordinal logistic Regression:-
The target variable has three or more ordinal categories such as restaurant or product rating from 1to5

Confusion Matrix
It contains information about actual and predicted classification done by classification system.

actual \ predicted		
	TP	FN
	FP	TN

Confusion Matrix

Scanned with CamScanner

(6)

Naive Bayes:- Used for classification of categorical data

prior probabilities:-
Calculate for some event based on no other information

Bayes Rule :- $P\left(\frac{A}{B}\right) = \frac{P(B/A)\,P(A)}{P(B)}$

(7)

Text Analysis
is a deriving & extending high quality information from text

Pre-processing Method in Text analysis
1. Tokenization.
The process of breaking text paragraph into small chunk or sentence

Scanned with CamScanner