Data Mining & Predictive Analytics (DMPA)

6 COs :-

CO1- Comprehand fundamental concepts of data mining, predictive analytics, CRISP-DM process and explain their applications.

CO2- Apply appropriate data pre-processing, EDA and dimension reduction methods to prepare dataset for effective analysis.

Attribute - feature - field - column
record - row

fill the missing values - Imputation.

CO3- Apply univariate & multivariate statistical analytics on the data in order to access underlying patterns and relationships.

univariate - using single parameter or feature. ex- avg. of age.
multivariate - using multiple parameter or more than one feature.

CO4- Describe and apply key data preparation techniques including cross validation, bias variance trade off, overfitting control to enhance model training and validation.

CO5- Analyse predictive modelling technique such as simple linear regression and multiple regression to model relationship between variables.

Yes/no. - classification
Prediction continuous - Regression

CO6- Explain and demonstrate the k-nearest neighbour algorithm for classification & prediction task with their applicability on different problem domain.

Book - Data mining & Predictive Analysis
— Daniel T. Larose

Data

Data is a raw fact, information, statistics which can be in various form like no.s, text, sound, images or any other form.

Why data mining?

The explosive growth of data that means production of data is too much in various areas, due to increase in size of database, increase in computerised growth in the society automatic analysis overcome on the manual analysis which

shows the need of data mining.

Evolution -

In 1960 the term data is evolved.

1960 - data collected & database creation.

1970 - Database Management System (SQL)

1980 - Advanced DBMS (RDBMS)

1990-2000 - Data mining or KDD.

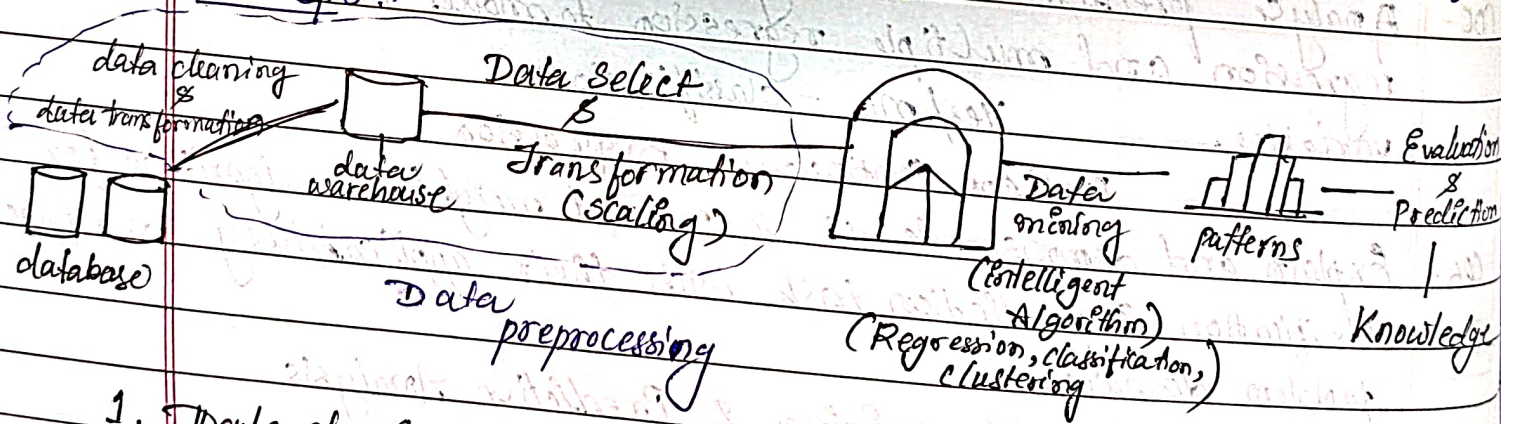
What is data mining?

Data mining refers to extracting or mining knowledge from large amount of data.

- Another term for data mining is Knowledge discovery from Data (KDD)
- Finding hidden info. or pattern from database.

KDD - Extraction of interesting information or patterns from the data in large database.

7 - steps :-



1. Data cleaning - To remove the noise and to consistent data.
2. Data integration - where multiple data sources may be combined and stored to warehouse.
3. Data Select - where data relevent to the analysis tasks are retrived from the database.
4. Data transformation - where data are transformed or consolidated into appropriate forms for mining by performing summary or aggregation operators that is scalling.
5. Data mining - An essential process where intelligent methods

are applied in order to extract data patterns.

6. Pattern Evaluation - To identify the truely interesting patterns representing knowledge based on measures.
(Precision, Accuracy, f1 score, kapa, recall, sensitivity, specificity

7. Knowledge Presentation - where visualization & knowledge representation techniques (pie chart, bar, curve etc) are used to present the mined knowledge to the users.

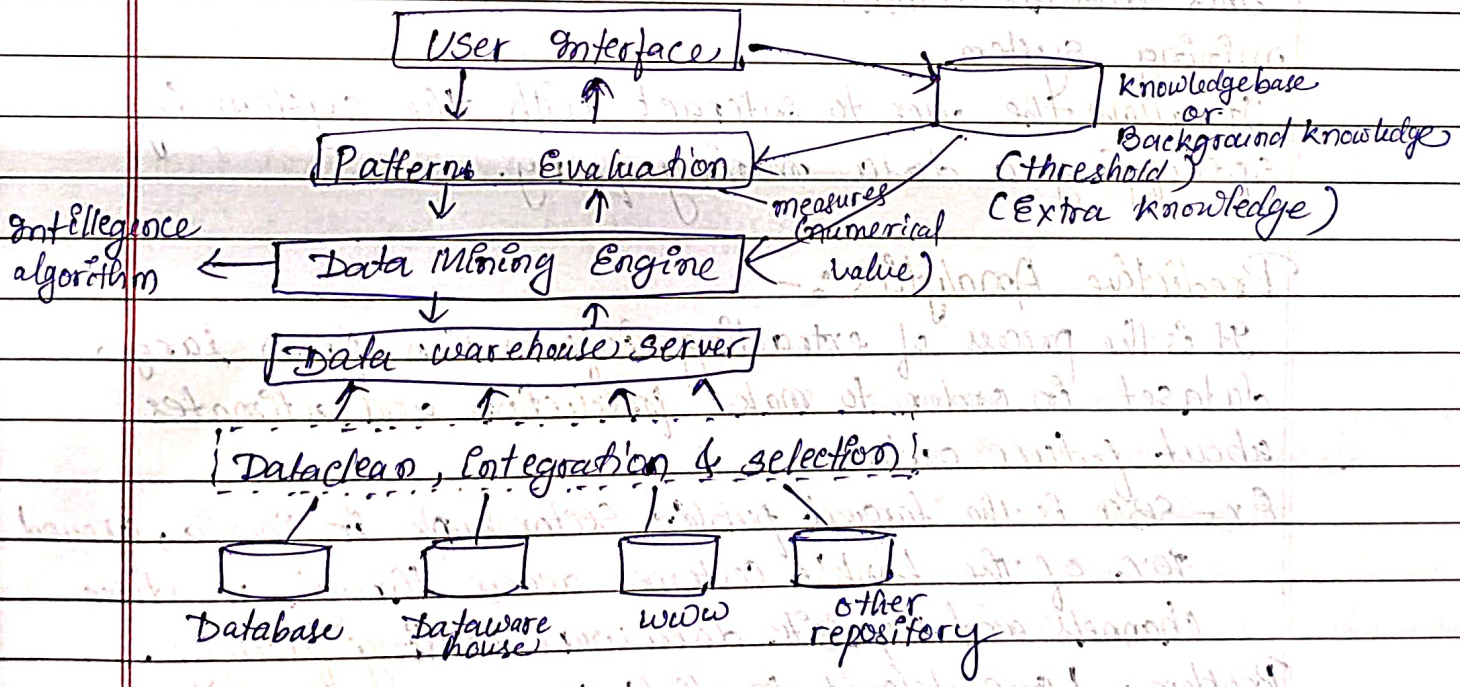
AUC - Area Under Curve

ROC - Receivers operating curve.

plot - scatter plot, box plot etc.

5/9/25

Architecture of Data Mining :-



Database, Data warehouse, www, other repository :-
This is a set of database datawarehouse, spreadsheet or other kind of information repositories. Data cleaning & data integration techniques may be performed on the data. The database or the datawarehouse server is responsible for fetching the relevant data based on users request.

Knowledge Base :-
This is the domain knowledge or background knowledge used to guide the search or evaluate the resulting patterns.

Data Mining Engine :-
This is essential to the data mining system & consists of a set of modules for task prediction and evaluation analysis.

Pattern Evaluation :-
This component typically employ interestingness measures and interact with the data mining modules based on the search.

User Interface :-
This module communicates between user and the data mining system.
It allow the user to interact with the system by specifying a data mining query.

Predictive Analytics :-
It is the process of extracting information from large dataset in order to make prediction and estimates about future outcomes.
Ex- SBI is the largest public sector bank in India, around 70% of the bank's customer access through multiple channels and requisite data ware housing facility.
Problem - Loan default prediction
Datamining knowledge - Analysing past records the bank finds customers with high credit utilization & irregular payment history are likely to default.
Predictive Analytics - when a new customer apply for a loan the bank uses this pattern to predict the likelyhood of default.

Problem 2 -

Health Care -

Data mining knowledge :- hospitals study patience histories and discover that people with high bp, obesity and smoking habits are more prone to heart diseases.

Predictive Analytics - A doctor uses this pattern in a predictive model to identify patients who are at risk of developing heart disease in the future.

Data Mining on what kind of Data :-

- Here we will discuss the number of different data repositories on which mining can be performed.

Repositories are :-

(i) Relational Database
(ii) Data warehouse
(iii) Transactional Database
(iv) Advanced Database System

* Advance Database System
  - (a) ⟶ Object Relational Database (ORD)
  - (b) ⟶ Spatial Database & Spatiotemporal database (geographical)
  - (c) ⟶ Temporal database, Sequence database
  - (d) ⟶ Text database, Multimedia database
  - (e) ⟶ Heterogeneous Database (Combination of heterogeneous database) & Legacy Database (Centuries) (Communication)
  - (f) ⟶ Time series Database (stock mkt)

(v) World Wide Web (www)

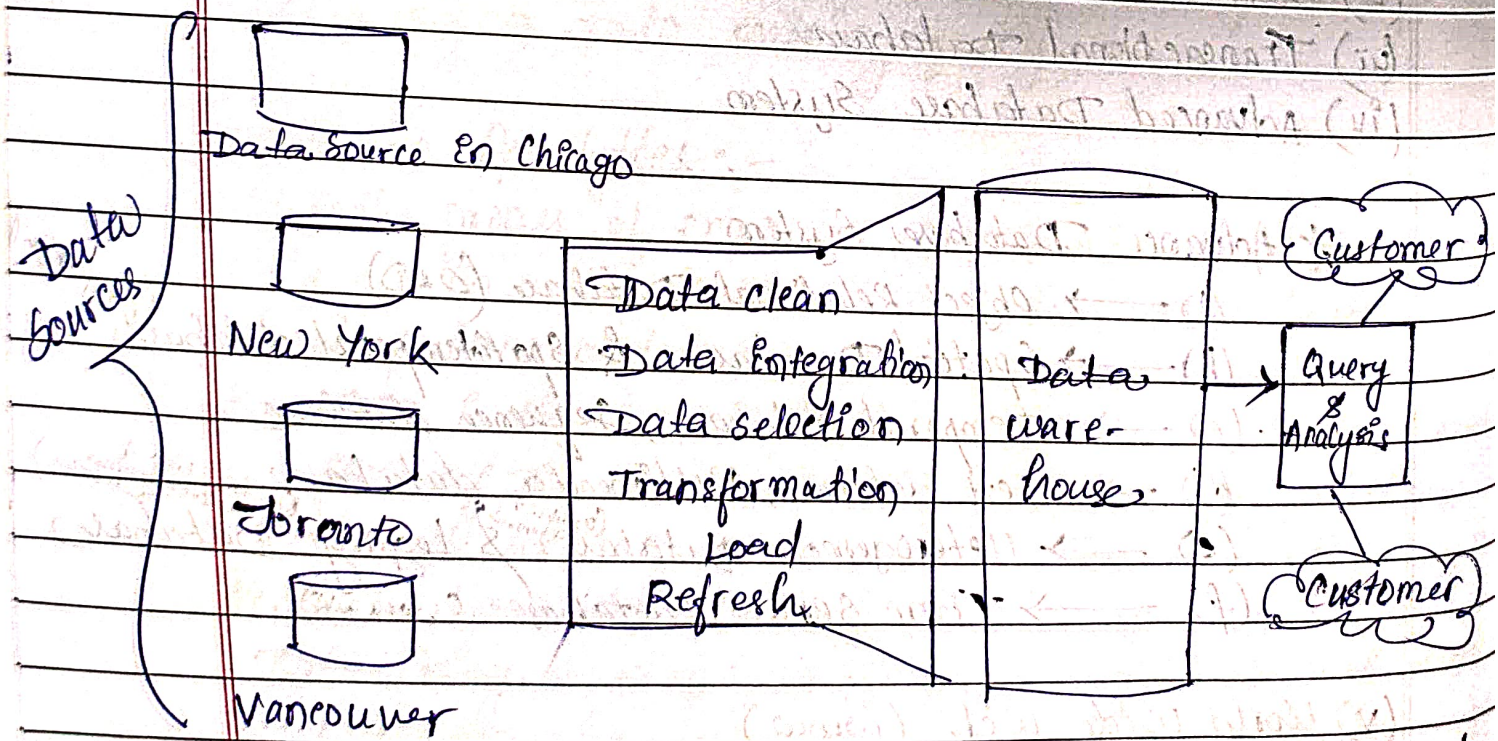
(i) Relational Database

→ DBMS consist of database and set of software program.

DBMS → Database, Software Program

→ Database consist of a collection of inter-related data.
→ Software programs are used to manage and access the data, store in the database.
→ The software program involve mechanisms for database structures, data storage, data sharing or distributing data access and for ensuring the consistency and security of the information stored.
→ The Relational Database is a collection of tables, each table is assigned a unique name.
→ Each table is consist of a set of attributes and a large set of tuples. (records)
→ Each tuple in a relational table is represented by a unique key and described by a set of attribute values

Data Warehouse



→ A data warehouse is a repository of information collected from multiple sources stored under unify scheme and usually reside at a single site.
→ Data warehouse are constructed through a process of Data cleaning, Data integration, Data transformation,

Data loading and Periodic Data Refreshing.

→ In computing data warehouse, also known as enterprise data warehouse (EDW) it is a system used for reporting and data analysis.

10/09/25

Q. Differentiate bet^n data mining & predictive Analytics.

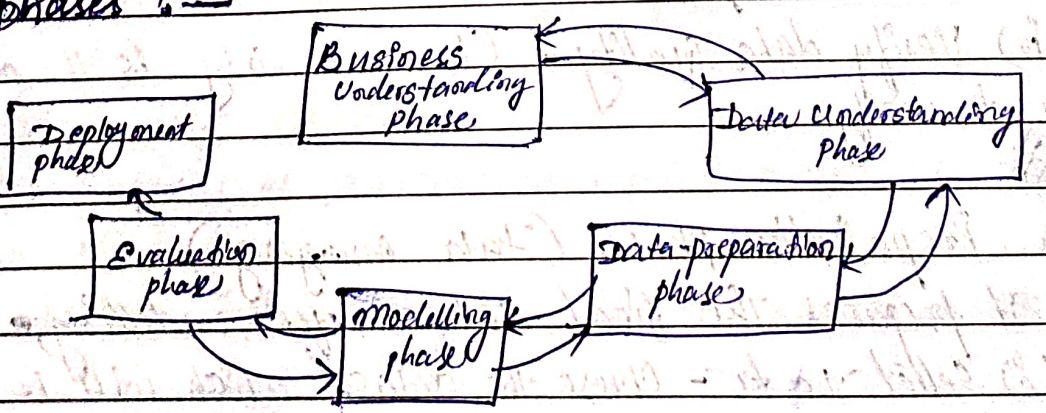
	Data Mining	Predictive Analytics
	- It is the process of extracting knowledge from large dataset.	- Apply the data mining knowledge, or patterns to predicts future outcomes.
Objective:	- to explore & understand data.	- to predict & forecast future values or probabilities.
Approach:	- descriptive	- predictive
Techniques:	- Regression, classification, clustering, association etc.	- eg Linear regression, decision tree,
Output:	- hidden patterns	- risk factors, probabilities
Type of Data	- past data & present data	- using only past data.
Dependency	- Independent	- Dependent

CRISP-DM

(Cross Industry Standard Process for Data Mining)

CRISP provide a standardize data mining process across industry for fitting data mining into the general problem solving strategy of a business or a research unit.

6 phases :—



(CRISP-DM process)

Business Understanding Phase - This phase focus on understanding the objectives & requirements of the project.

This phase having 4 tasks -

(i) Determine business objectives -

Throughly understand from a business perspective what the costemer really want and then define business success criteria

(ii) Access Situation -

Determine resources availability project requirement, acass contigency and conduct a cost benifit analysis.

(iii) Determine data mining Goals -

In addition to defining the business objective you should define technical data mining perspective

(iv) Product - Project - plan -

Select technologies & tools & define detailed plan for each project phase

Data Understanding Phase -

It drives the focus to identify, collect and analyse the dataset, this phase also has 4 tasks -

(i) Collect initial data - Acquire the necessary data & load it into your analysis tool.

(ii) Describe data - Examine the data & document like data format, no. of records or fields.

(iii) Explore data - You have to visualize & identify relations among the data.

(iv) Verify data Quality - Clean or dirty is the datas, must have to verify.

Data Preparation phase (Data Munging) -

It prepares the final dataset for modelling.

(i) Select data - Choose the dataset which will be used.

(ii) Clean data - To correct, impute or remove erroneous data

(iii) Construct data - Derive new attribute that will be helpful.
(iv) Integrate data - Create a new dataset by combining data from multiple sources.
(v) Format data - Reformat data as necessary.

Modelling Phase -
In this phase build and access various phase based on alternate modelling techniques
(i) Select modelling techniques - Determine which algo you want to ~~track~~.
(ii) ~~Set~~ generate test design - Need to split the data into training testing & validation.
(iii) Build Model - Execute few lines of code
(iv) Access Model - Multiple models are competing against each other to interpret model result based on domain knowledge.

Evaluation Phase -
It focus on technical model accessment.
(i) Evaluate Result - which one should approve for the business, review process
(ii) Review Process - Properly executed or not summarize it and correct anything if needed.
(iii) Determine next Phase - based on the previous task determine wether to proceed for deployment or not.

Deployment Phase -
A model is not particularly useful unless the customer can access its result.
(i) Plan deployment - develop and document a plan for deploying a model.
(ii) Plan monitoring & maintenance - To avoid the issues during the operational phase of a model.

(iii) produce final report - the project to document; a summary of the project.
(iv) Review Project — what went well, and what could have been better, how to improve.

Fallacies of DM (Misconception)

1. Data mining tools are automated tools that can be deployed on data repository to find answers to our problems.
Reality :- There are no automated data mining tool which will automatically solve your problem.
There are methodology available such as crisp-DM which streamline the data mining process into the overall business plan.

2. Data mining process is autonomous, requiring no human intervention.
Reality :- Without skilled human intervention (expert), blind use of data mining software will only provide wrong answer to wrong question; thus human intervention is required to update the model

3. Data mining phase for itself quickly.
Reality :- Benifit take time; cost exclude starter cost data collection, data warehouse, preparation cost, software & skilled analysis. 12/09/25

4. Data mining software packages are intuitive & easy to use
Reality :- No ease of use varies, it may require statistical and mathematical knowledge of particular application domain.

5. Data mining will identify the causes of ~~the~~ our business or research problem.
Reality :- KDD process will help to show hidden patterns

Types of Learning
Supervised Unsupervised Re-inforcement

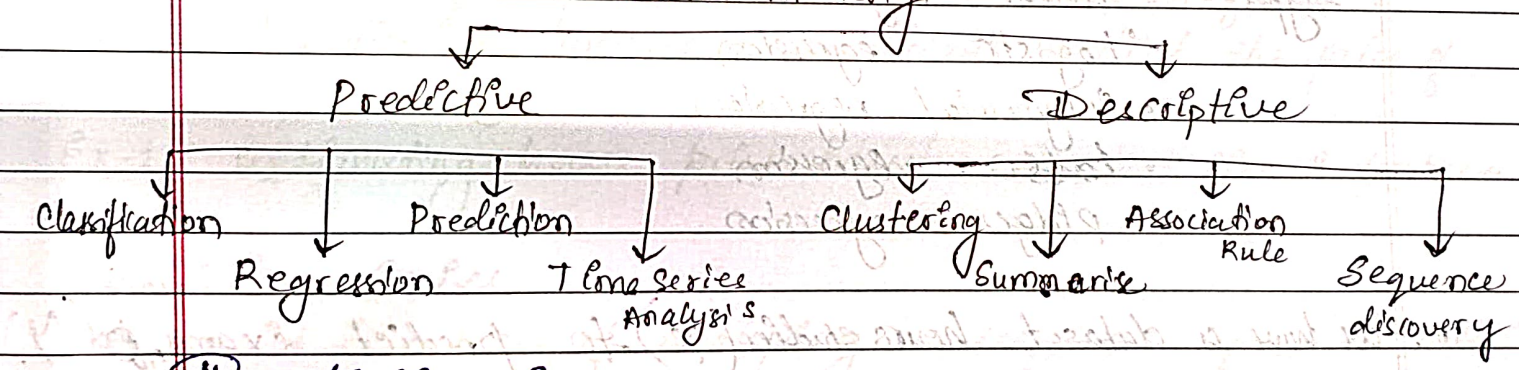
but human judgement and domain expertise needed.

6. Data mining will automatically clean a machine database
Reality :- still needs to handle missing values, outlayers and inconsistency manually.

7. Data mining always provide a positive result.
Reality :- Not gaurenteed, poor data or assumption can leads to misleading outcome.

Task of Data Mining :—

Data mining task involves finding patterns & useful info. from large dataset.



Predictive :—

Classification (supervised learning)

Classification is a supervised learning technique in data mining that involve categorizing or classifying data into ~~3 differe~~ pre-defined classes, categorizing or groups based on their features or attributes.

In supervised learning label data can be used to build a model that can predic the class of unseen data. It is of 2 types — Binary classification
Multinary classification

Steps to build classification
- data preparation
- feature selection
- split for train and test

- model selection
- model training
- model evaluation
- model tuning
- Ensemble learning
- Model Deployment

Regression (Estimation)

It is a supervised learning technique used to predict a continuous numerical value by analysing the relationship between dependent variable & independent variable using past data.

Types -
- Linear regression
- Logistic regression
- Polynomial regression
- Lasso regression
- Ridge regression

Q. We have a dataset hours studies (x) to predict exams score (y)

x	1	2	3
y	2	4	5

$\hat{y} = a + bx$

$a = ?$

$b = ?$

Step-1 $\bar{x} = 2 \qquad \bar{y} = \frac{11}{3} = 3.67$

Step-2 $b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{4}{3} = 1.5$

x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	-1	-1.67	1.67	2.78
2	4	0	1.67	0	0
3	5	1	2.67	2.67	7.128
				4.34	9.91

$a = \bar{y} - b\bar{x}$

$= 0.67$

$\boxed{y = 0.67 + 1.5x}$ - predictive eq^n

$x = 4$

Exam score = 6.67 $\{0.67 + 1.5 \times 4$

Prediction

It is similar to classification & estimation except for prediction the result lies in future.

Ex-1. Predicting wether a candidate predict the price of a stock 3 month into the future.

Ex-1. Estimating house price based on location size & features.

Time Series

It is a way of analysing a sequence of data point collected over an interval of time.

It composed of 4 main element -

→ trend - A long term movement of data

→ Seasonality - A predictable recurring pattern of repetation over a fixed regular interval like days, week month

→ Cycle - A long term movement of fluctuation.

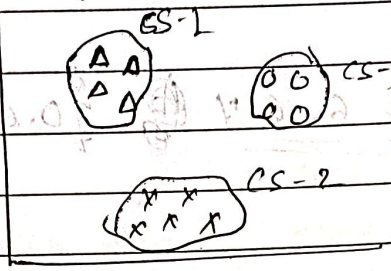
→ Irregularity - Random or unpredictable fluctuation in the data that are not explained by trend, season or cycle.

Descriptive :-

Clustering

It refers to grouping of records observations into a similar object. This is mainly used in unsupervised learning.

The data point having less distance belong to same cluster & more distance belongs to different cluster. Clustering segment the whole dataset into relatively homogeneous sub-groups.



Summarization

It is the process of reducing large dataset into shorter, more understandable format that highlight key patterns, trends and relationship.

Types -
- Descriptive - uses statistical measures to describe main features of numerical data.
  Ex - mean, median, mode.
- Aggregation - combining data into simpler summary into sum average and group count.
- Sampling - select a representative subset of the data.

Association Rule

Association task for data mining is the job of finding which attributes go together.

Ex - Market basket analysis, where items are purchased together.

This rule shows how frequently an item set occurs in a transaction.

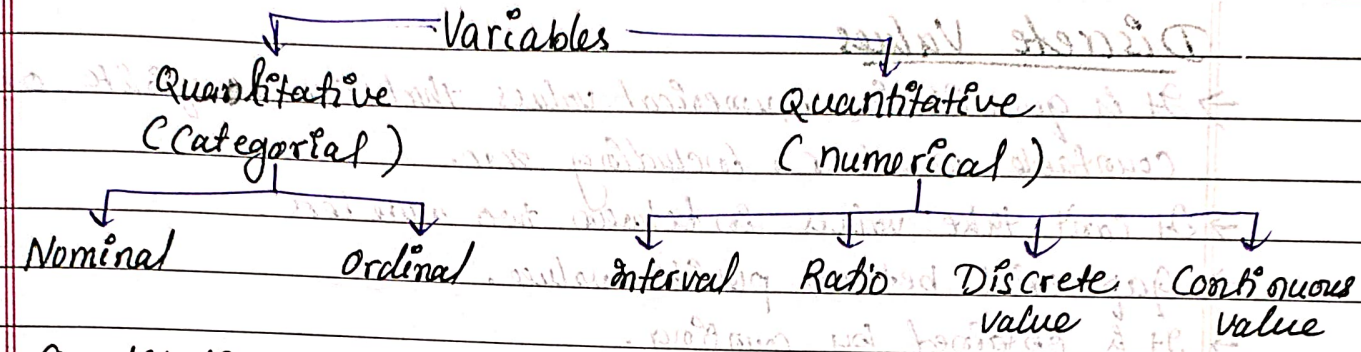
Sequence discovery -

It is a data mining technique used to identify frequently occurring ordered patterns in sequential data such as customer's purchase history.

Appendix —

Data summarization & visualization :-

Applicant	Marital status	Mortgage	Income	Rank	Year	Risk
1.	Single	Y	38K	2	2009	Good
2.	married	Y	32K	7	2010	Good
3.	other	N	25K	9	2011	Good
4.	other	N	36K	3	2009	Good
5.	other	Y	33K	4	2010	Good
6.	other	N	24K	10	2007	Bad
7.	married	Y	25K	8	2009	Good
8.	married	Y	48K	1	2010	Good
9.	married	Y	32K	6	2011	Bad
10.	married	Y	32K	5	2009	Good



Qualitative Data

→ Describes qualities or categories of Data.
→ Also called categorial variables.
→ Eg :- Marital status, mortage, rank and risk.

Quantitative Data

→ It takes numerical values and allows arithmetic operation.
→ Also called numerical variables.
→ Eg :- Income, year, etc.

Nominal

→ It is used for names, labels or categories without order.
→ Ex :- Marital status, Mortage & risk.

Ordinal

→ It maintains a particular order
→ Eg :- Ranks, Grades.

Interval & Ratio

* Interval
→ It is a quantitative data defined within an Interval without natural zero.
Eg :- Year.

* Ratio
→ It is a quantitative data for which mathematical operations can be performed.
Eg :- Income

Discrete Values

→ It is quantitative numerical values that can be finite or countable number, including zero.
→ It can't take values in between two numbers.
→ Gap exists between possible values.
→ It is obtained by counting.
→ Graphical representation :- Bar chart
Eg :- Year

Continuous Values

→ It is a Quantitative value that can be any value within a given range.
→ No gaps between the points.
→ Eg :- Income
→ Get from measurables.
→ Graphical representation :- Histogram

Predictor Variable

→ It is a variable whose value is used to help predict the value of response variable.
→ The predictor variable in table are all variables except risk.
→ Response variable, also called, Dependent variable or output variable or target variable which try to predict, explain or measure the effect.
Eg :- Effect of study hours on exam score.
prediction → study hours
response → exam score

Measurement of Center, variability & position :-

→ Measures of Central Tendency :-
• It measures of the location of middle or center of data distribution.
Eg :- Mean, Median, Mode & Mid Range.

Mean

→ It is an Arithmetic average.
→ The sum of all data values / No. of values

Data = [2, 4, 6, 8, 10]
Mean = $\frac{30}{5} = 6$

Advantage
Easy to compute, including all points.
Disadvantage
Sensitive to outliers.

Median

→ The middle value when the data is arranged in ascending or descending order.

Data = [3, 5, 7, 9, 11] } odd dataset
Median = 7

Data = [3, 5, 7, 9] } Even dataset
Median = $\frac{5+7}{2} = 6$

→ If no. of observation is odd, median is the middle value.
→ If no. of observation is even, median is average of two middle value

Advantages
→ Not affected by outliers.

Mode

→ The value that appear most frequently in a dataset.
→ It is of three types
(i) Unimodal (ii) Bimodal (iii) Multimodal
(iv) No modal

Data :- [2, 4, 4, 5, 6]
mode = 4

Advantage
Used in categorial data

Limitations
May not exist or may not be unique.

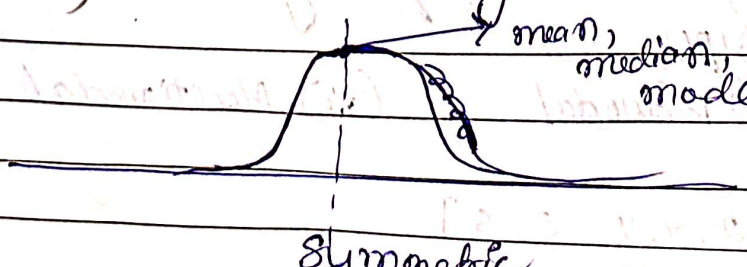
Mid Range

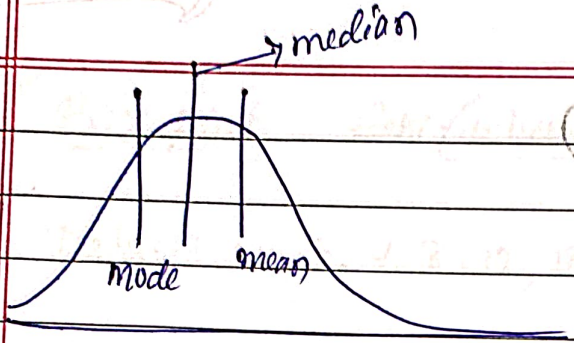
→ It is the avg of largest and smallest value in the dataset.
→ It is applicable for numerical data.

Measure	Best used	Sensitive to outliers	Datatype
Mean	Numerical & Symmetric data	Yes	Interval / Ratio
Median	Skewed data / with outliers	No	Ordinal & Interval / Ratio
Mode	Most frequent item needed	No	Nominal / Ordinal / Interval / Ratio

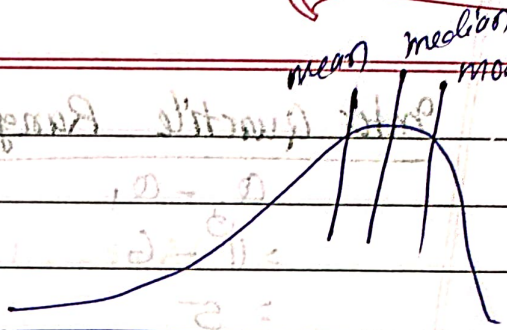
Measuring Dispersion of Data

(i) Symmetry
→ In a unimodel frequency call with perfect symmetric data distribution, mean, median & mode are all at the same center value.
→ In real data is not symmetric.





Positive skewed
(mode < median)



negative skewed
(mode > median)

Measurement of dispersion (variability)

1 → Range 2 → Quartiles (Q_1, Q_2, Q_3)
3 → Inter Quartile Range (IQR) ($Q_3 - Q_1$)
4 → Five number summary (min, max, Q_1, Q_2, Q_3)
5 → Box Plot
6 → Z-score 7 → variance
8 → standard deviation

Range

→ It is the difference between max & min value in the dataset.

Eg :- Dataset is [5, 7, 9, 10, 12],

Range = 12 - 5 = 7

Quartile

→ Divides ordered data into 4 equal parts.

→ Q_1 → First Quartile which represents 25 percentile of data below it

→ Q_2 → Second Quartile known as median throughout represent 50 percentile of data below it.

→ Q_3 → third Quartile which represents 75 percentile of data below it.

Dataset → [5, 7, 9, 10, 12]

$Q_2 \rightarrow 9$

As, 5, 7 are even,

$Q_1 \rightarrow \frac{5+7}{2} = 6$

Similarly $Q_3 \rightarrow \frac{10+12}{2} = 11$

Inter Quartile Range (IQR)

$$Q_3 - Q_1$$
$$= 11 - 6$$
$$= 5$$

★ **Capping**

For identifying suspected outliers, the common rule is to single out values falling atleast → 1.5 X IQR

★ **Above Q_3 or Below Q_1**

Five Number Summary

→ It is a concise statistical summary of data consisting of minimum, Q_1, Q_2, Q_3 & maximum.

Data Set = [5, 7, 8, 12, 13, 14, 18, 21, 23, 25]

$$Q_2 = \frac{13+14}{2} = \frac{27}{2} = 13.5$$

$Q_1 = 8$

$Q_3 = 21$

min = 5

max = 25

Box plot ~~Cat~~ (whisker Plot)

Q. Dataset 5, 7, 8, 12, 13, 14, 18, 21, 23, 25

Q_2 = 13.5

10 → Even → $\frac{13+14}{2} = 13.5$ (Q_2)

min = 5

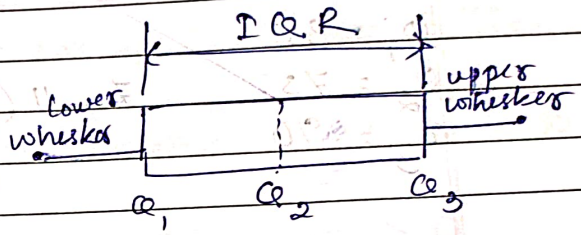
max = 25

Q_2 = 13.5

Q_1 = 8

Q_3 = 21

5, 7, 8, 12, 13 → Q_1



Lower = $Q_1 - 1.5 \times IQR = 11.5$

upper = $Q_3 + 1.5 \times IQR = 40.5$

$IQR = Q_3 - Q_1 = 21 - 8 = 13$

Box plot is a graphical representation of 5 number summary of a dataset.

Features - (i) Box - Extends from Q_1 to Q_3 that is IQR

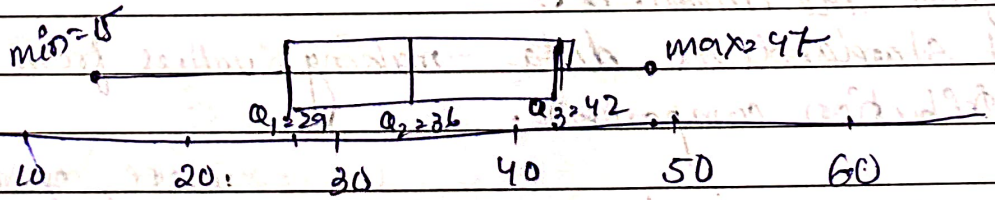
(ii) Median line - A line inside the box shows median.

(iii) Whisker - A line extending from Q_1 to min & Q_3 to the max

(iv) Outliers - Data points that lie beyon lower fench

$= Q_1 - 1.5 \times IQR$

upper fench = $Q_3 + 1.5 \times IQR$



Q. Collect the math mark out of 100 from 10 student of a class. Records are 11, 7, 35, 55, 64, 90, 86, 88, 95, and 97.

7, 11, 35, 55, 64 | 86, 88, 90, 95, 97

min = 7
max = 97
$Q_1 = 35$
$Q_2 = 75$
$Q_3 = 90$

$Q_2 = \frac{64+86}{2} = 75$

7, 11, (35) 55, 64 — Q_1

86, 88, 90, 95, 97 — Q_3

lower fench $= Q_1 - 1.5 \times IQR$
$= 35 - 1.5 \times 55$
$= -47.5$

$IQR = Q_3 - Q_1$
$= 90 - 35$
$= 55$

upper fench $= Q_3 + 1.5 \times IQR$
$= 90 + 1.5 \times 55$
$= 172.5$

All values come under this, so no outliers.

Z-Score

It measures how many standard deviations a datapoint is from the mean.
It standardizes data making values from different distribution comparable.

$$Z = \frac{x - \mu}{\sigma}$$

where, x = observed value
μ = mean of dataset
σ = standard deviation

Note

if $Z = 0$ mean position
$Z > 0$ above mean
$Z < 0$ below mean

$|Z| > 3$ outlier.

note: if Z is 0, value is at mean.
Z greater than 0, value is above mean.
Z smaller than 0, value is below mean.
outlier = datapoint with magnitude of $Z > 3$ are outliers.

Q. In an exam average score is 70, standard deviation is 10. Student score is 85. Calculate Z.

$$Z = \frac{85 - 70}{10} = \frac{15}{10} = 1.5 > 0$$ above mean.

if $x = 65$

$$\frac{65 - 70}{10} = -0.5 < 0$$ below mean.

Variance

It measures the avg. square deviation from the mean.

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

Q. 2, 4, 6 $\quad \bar{x} = \frac{12}{3} = 4$

	$x - \bar{x}$	$(x - \bar{x})^2$
2	-2	4
4	0	0
6	2	4
		8

$$\sigma^2 = \frac{1}{3} \times 8 = \frac{8}{3} = 2.6$$

Standard Deviation

It is the square root of variance.

$$\sqrt{\sigma^2} = \sqrt{\frac{8}{3}} = 2\sqrt{\frac{2}{3}}$$

Ch-2

Data Preprocessing

Raw data contained in database unpreprocessed, incomplete and noisy.

Data preprocessing is the process of preparing raw data for analysis by cleaning & transforming it into a usable format.

Types of data where we use data preprocessing :-

→ Real dataset often having missing or incomplete entries.
→ Data may contain type error, duplicates or outliers.
→ Different features may have different scale not suitable for data mining models.
→ Some features are irrelevant, redundant or highly co-related.
→ Values not consistency with the policy.

Steps of data pre-processing :-

- data cleaning
- data integration
- data transformation
- data reduction

Data cleaning :-

It is a process of identifying & correcting errors and inconsistency in the dataset.
- It involve handling missing values, removing duplicate and correcting incorrect or outlier data to ensure the dataset is accurate.
- Clean data is essential for effective analysis as it improves the quality of results and enhances the performance of data model.

Missing values

- When a data is absent in a dataset, it is called missing values.

(a) Ignore the tuple or record :-
This is usually done when the class level is missing, this method is not very effective unless the tuple contains several attributes with missing value.

(b) Fill in the missing value manually.
This approach is time consuming may not be feasible for a large dataset.

(c) Global Value
Use a global const. to fill in a missing value.

(d) Use the attribute mean to fill in the missing value
Use the attribute mean for all sample belonging to the same class as the given tuple.

(e) Use the most preferable probable value to fill in the missing value.
Decision tree, regression, bayesian formula

Noisy Data

Noise is a random error or variance in a measured variable.

Incorrect attribute values may due to
- faulty data.
- data entry problem
- data transmission problem
- technology limitation
- inconsistency in naming convention

Other data problem like duplicate records incomplete data.

— Binning Method
first sort the data and the sorted values are

distributed into a no. of buckets or bins, then, these bins can smooth by mean, median & boundaries.

(a) By bin means:-
Each value in a mean is replaced by mean value of the mean.

(b) By bin median:-
Each bin value is replaced by bin median.

(c) By bin boundaries:-
Minimum & maximum values in a given mean are identified as the bin boundary.
Each bin value is then replaced by closest boundary value.

Equal width partitioning (Distance)

It divides the range into N-intervals of equal sizes.

If A & B are lowest & highest value of the attribute, the width of the interval will be

$$W = \frac{B-A}{N}$$

The most straight forward method but the outliers may dominant the presentation

Equal-depth Partitioning (OR Equal-Frequency Partitioning)

It divides the range into N intervals each containing approximately same no. of sample.

- Good data scaling.
- Managing categorical attribute.

Q- Sorted data for price is 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34.

Sol:

bin 1 - 4, 8, 9, 15
bin 2 - 21, 21, 24, 25
bin 3 - 26, 28, 29, 34

~~Equal Depth Partitioning~~ :-

Smoothing by bin means,

mean 1 = $\frac{4+8+9+15}{4} = 9$

bin 1 = 9, 9, 9, 9

mean 2 = $\frac{21+21+24+25}{4} = 22.75$

bin 2 = 23, 23, 23, 23.

mean 3 = $\frac{26+28+29+34}{4} = 29.25$

bin 3 = 29, 29, 29, 29

Smoothing by bin medians.

$med_1 = 8.5$
bin 1 = 8.5, 8.5, 8.5, 8.5
$med_2 = 22.5$
bin 2 = 22.5, 22.5, 22.5, 22.5
$med_3 = 28.5$
bin 3 = 28.5, 28.5, 28.5, 28.5

Smoothing by bin boundaries.

bin 1 = 4, 4, 4, 15
bin 2 = 21, 21, 21, 25
bin 3 = 26, 26, 26, 34

with Arduino

Bins

May not have equal no. of data
Need not to be arranged

Equal-width

S-1
min = 4
max = 34
Range = 34 - 4 = 30

S-2
width = interval = Range / no. of Bin = 30/3 = 10 (3)

S-3
Bin 1 [4, 14] : 4, 8, 9
Bin 2 [14, 24] : 15, 21, 21, 24
Bin 3 [24, 34] : 25, 26, 28, 29, 34
(Range)

Equal-depth or frequency

S-1 Data must be sorted.
S-2 Take 3 bins, total value in the data set is 12. So each bin should have 4 values.
S-3
bin 1 - 4, 8, 9, 15
bin 2 - 21, 21, 24, 25
bin 3 - 26, 28, 29, 34

Regression

Data can be smooth by fitting the data to a function such as regression.
(i) Linear Regression
(ii) Multiple Linear Regression
(iii) Clustering.

Misclassification :-

Misclassification means a model predict the wrong class level compared to the actual or true class level in your dataset.

Brand	Frequency
USA	1
France	1
US	156
Europe	46
Japan	51

Misclassification

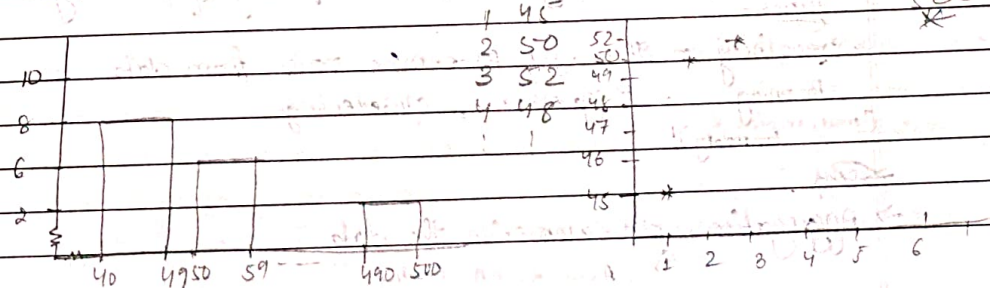
Brand	Frequency
US	1
Europe	1
US	156
Europe	46
Japan	51

In frequency distribution dataset having 5 class, however two of the class USA & France have count only one each which is clearly happening that two of the records have been inconsistently classified (Misclassification) with respect to original data.
To maintain consistency the record have been labelled US and Europe instead of USA and France.

Graphical Method for Identifying Outliers :-

Outliers are extreme values that go against the trend of remaining data.
Graphical Methods for identifying outliers for numerical variables is histogram & scatter plot.

Q. 45, 50, 52, 48, 47, 49, 51, 46, 500, 50, 48, 49, 52, 47, 51
45, 46, 47, 47, 48, 48, 49, 49, 50, 50, 51, 51, 52, 52, 500



Data Integration :-

It is the merging of data from multiple data sources.
It help to reduce and avoid redundancy and inconsistency in the resulting data set which improve the accuracy and speed of the subsequence data-mining process.

Entity Identification Problem :-

Ex

Employee Details

Customer Name	Mobile no.	Email Id.
Rina	93855 6507	Rina@gmail.com

Customer Details

Customer Name	Contact no.	Email
Rina Mallokid	9338556507	Rina@gmail.com

It occurs when we try to determine wether two or more records from different data sources refer to the same real world entity.

Solution -
- → Rule based Matching
- → Probabilistic Matching (similarity)
- → Machine Learning Algorithm

Data Transformation :-

It is the process of converting data from one format or structure to another format or structure to ensure compatibility, consistency and better quality of data for analysis.

Types -
- → Smoothing - It is used to remove noise from data
  Binning, Regression, clustering
  (mean, median & boundary)
- → Aggregation - It summerises the data.
  Ex - Avg. mean, sum.
- → Generalization - Replacing detailed data with higher-level concep
  Eg - Age - 23 → 'youth Adult'
  Age of a student 23 converted to higher-level 'Youth-Adult'
- → Normalization - Rescaling a value to a standard range, and the range must be between (0-1).
- → Attribute construction - Create a new attribute from a existing data.
  Ex - use DOB to calculate age.

Min-Max Normalization (Scaling)

(0-1)

$$x'_{mm} = \frac{x - min(x)}{range} = \frac{x - min(x)}{max(x) - min(x)}$$

Min-max normalization is on the original data where,
x = original data
min(x) = min value of attribute
max(x) = max value of attribute
x'_{mm} = Normalized values

Q. 50, 60, 70, 80, 90, 100. Calculate min-max normalization for 80.

$$x'_{mm} = \frac{80 - 50}{50} = \frac{30}{50} = 0.6$$

Q. Suppose the min & max value for the attribute income are 12000 and 98000 respectively. We like to map income to the range 0-1 by min-max normalization. a value of 73600. What is the transformed value.

$$x'_{mm} = \frac{73600 - 12000}{98000 - 12000} = \frac{61600}{86000} = 0.716$$

In terms of mid-range, what would be the transformed value.

$$\frac{12000+98000}{2}$$

$$\bar{X} = 55000$$

$$x'_{min} = \frac{55000-12000}{86000}$$

$$= \frac{43000}{86000} = 0.5$$

Z-score standardization :-

$$\text{Z-score} = \frac{X - \text{mean of } (x)}{SD(x)}$$

$$= \frac{X-\mu}{\sigma}$$

Decimal Scaling :-

It is a normalise technique where the normalized value lies between -1 to 1.

$$X_{decimal} = \frac{X}{10^d}$$

where,

d represents the no. of digits in the data value with the largest absolute value.

Q. -987, -120, 56, 300, 850.

$$X_{decimal} = \frac{X}{10^d} = \frac{-987}{10^3} = -0.987$$

$$120_{decimal} = \frac{-120}{10^3} = -0.120$$

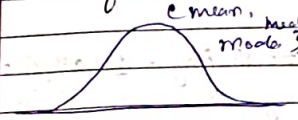
$$56_{decimal} = \frac{56}{10^2} = 0.56$$

$$300_{decimal} = \frac{300}{10^3} = 0.3$$

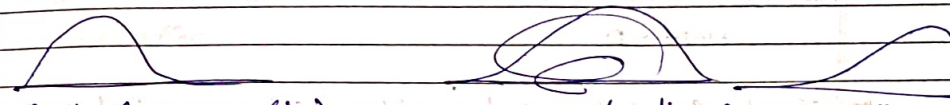
$$850_{decimal} = 0.850$$

That is the normalized value lies between -1 to +1

Transformation to Achieve Normality :-



The normal distribution is a continuous probability distribution commonly known as bell curve which is a symmetric curve.



$$\text{Skewness} = \frac{3(\text{mean} - \text{median})}{SD}$$

Remove Skewness -

To eliminate the skewness apply transformation to the data.

(i) Log transformation

$$\bar{x} = \log(x+1)$$

(ii) Square root transformation

$$\bar{x} = \sqrt{x}$$

(iii) Inverse Squarerent transformation $\bar{x} = \frac{1}{\sqrt{x}}$

Flag variable

These are dummy variables or indicator variables which is converting categorial variables into only 2 variables (0/1).

Ex: Lets take a categorical variable, gender having 2 variables female or male.

If Gender = female then Gender_Flag = 0

If Gender = Male then Gender_Flag = 1

Ex-2: Categorical predictor region having 4 variable. North, South, East, west.

north_flag : If region = north then north_flag = 1 otherwise north_flag = 0

east_flag : If region = east then east_flag = 1 otherwise east_flag = 0

South_flag : If region = south then south_flag = 1 otherwise south_flag = 0

Otherwise.

west = 0

Transforming categorical variables into numerical variables.

In data mining categorical variables must be converted to numerical format for machine learning algorithms.

(i) Level Encoding

Ex -

size — large (0)
size — Medium (1)
size — small (2)

region — East (0)
region — west (1)
region — north (2)
region — south (3)

(ii) One-hot-Encoding

color — Red
color — Green
color — Blue

Red	Green	Blue
1	0	0
0	1	0
0	0	1

(iii) Binary-Encoding

city — Bhubaneswar (1) → 001
city — Cuttack (2) → 010
city — Rourkela (3) → 011

10/9/25

Q. Suppose that a hospital tested the age & body fat data for some randomly selected adults.

age	23	23	27	27	39	41	47	49	50
% fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2

(a) Calculate mean, median, mode & SD of age & % fat.

(b) Transform the above numeric attribute using z-score normalization.

age	% fat
23	9.5
23	26.5
27	7.8
27	17.8
39	31.4
41	25.9
47	27.4
49	27.2
50	31.2

mean of age = 36.2

mean of % fat = $\frac{204.7}{9}$ = 22.74

If age,

median = 39

median of % fat = 26.5

7.8, 9.5, 17.8, 25.9, 26.5, 27.2, 27.4, 31.2, 31.4

mode of age = 23, 27

mode of % fat = no mode

Flag variable

These are dummy variables or indicator variables which is converting categorial variables into only 2 variables (0/1).

Ex: Let take a categorical variable, gender having 2 variables, female or male.

If Gender = female then Gender_Flag = 0

If Gender = Male then Gender_Flag = 1

Ex-2: Categorical predictor region having 4 variable. North, South, East, west.

north_flag : If region = north then north_flag = 1 otherwise north_flag = 0

east_flag : If region = east then east_flag = 1 otherwise east_flag = 0

South_flag : If region = south then South_flag = 1 otherwise south_flag = 0

Otherwise,

west = 0

Transforming categorical variables into numerical variables.

In data mining categorical variables must be converted to numerical format for machine learning algorithms.

(i) Level Encoding

Ex -

size ← large (0), medium (1), small (2)

region ← East (0), west (1), north (2), south (3)

(ii) One-hot-Encoding

color ← Red, Green, Blue

Red	Green	Blue
1	0	0
0	1	0
0	0	1

(iii) Binary Encoding

city ← Bhubaneswar (1) → 001, Cuttack (2) → 010, Rourkela (3) → 011

10/9/25

Q. Suppose that a hospital tested the age & body fat data for some randomly selected adults.

age	23	23	27	27	39	41	47	49	50
% fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2

(a) Calculate mean, median, mode & SD of age & % fat.

(b) Transform the above numeric attribute using z-score normalization.

age	% fat
23	9.5
23	26.5
27	7.8
27	17.8
39	31.4
41	25.9
47	27.4
49	27.2
50	31.2

mean of age = 36.2

mean of % fat = $\frac{204.7}{9}$ = 22.74

for age,

median = 39

median of % fat = ~~28.9~~ 26.5

7.8, 9.5, 17.8, 25.9, 26.5, 27.2, 27.4, 31.2, 31.4

mode of age = 23, 27

mode of % fat = no mode

SD of age,

$$\sqrt{\sigma^2} = \sqrt{\frac{1}{9}\sum_{n=1}^{N}(x_i - \bar{x})^2}$$

$\bar{x} = 36.2$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
23	-13.2	174.24
23	-13.2	174.24
27	-9.2	84.64
27	-9.2	84.64
39	2.8	7.84
41	4.8	23.04
47	10.8	116.64
49	12.8	163.84
50	13.8	190.44
		1019.56

$$SD = \sqrt{\frac{1019.56}{9}}$$

$$= \sqrt{113.284}$$

$$= 10.64$$

SD of % fat,

$$\sqrt{\sigma^2} = \sqrt{\frac{1}{9}\sum_{n=1}^{9}(x_i - \bar{x})^2}$$

$\bar{x} = 22.74$

% fat x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
9.5	-13.24	175.2976
26.5	3.76	14.1376
7.8	-14.94	223.2036
17.8	-4.94	24.4036
31.4	8.66	74.9956
25.9	3.16	9.9856
27.4	4.66	21.7156
27.2	4.46	19.8916
31.2	8.46	71.5716
		635.2037

$$SD = \sqrt{\frac{635.2037}{9}}$$

$$= \sqrt{70.578}$$

$$= 8.401$$

b) $$Z = \frac{x_i - \mu}{\sigma}$$

for age, $\mu = 36.2$ $\sigma = 10.64$

x_i	$x - \mu$	$\frac{x-\mu}{\sigma} = Z$
23	-13.2	-1.240
23	-13.2	-1.240
27	-9.2	-0.864
27	-9.2	-0.864
39	2.8	0.263
41	4.8	0.451
47	10.8	1.015
49	12.8	1.203
50	13.8	1.297

for % fat

x_i	$x_i - \mu$	Z
9.5	-13.24	-1.576
26.5	3.76	0.448
7.8	-14.94	-1.778
17.8	-4.94	-0.588
31.4	8.66	1.03
25.9	3.16	0.376
27.4	4.66	0.554
27.2	4.46	0.530
31.2	8.46	1.007

(c) Calculate correlation coefficient $r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$

(c) $$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

x	y	xy	x^2	y^2
22	9.5	218.5	529	90.25
23	26.5	609.5	529	702.25
27	7.8	210.6	729	60.84
27	17.8	480.6	729	316.84
39	31.4	1224.6	1521	985.96
41	25.9	1061.9	1681	670.81
47	27.4	1287.8	2209	750.76
49	27.2	1332.8	2401	739.84
50	31.2	1560	2500	973.44
326	204.7	7986.3	12828	5290.99

$$r = \frac{9 \times 7986.3 - (326 \times 204.7)}{\sqrt{(9 \times 12828 - (326)^2)(9 \times 5290.99 - (204.7)^2)}}$$

$$= \frac{71876.7 - 66,732.2}{\sqrt{(115452 - 106276)(47618.91 - 41902.09)}}$$

$$= \frac{5144.5}{\sqrt{9176 \times 5716.82}}$$

$$= \frac{5144.5}{\sqrt{52457540.32}} \qquad \frac{5144.5}{7242.758} = 0.71029$$

a. Binning

Sale Price records

T = { 50, 55, 65, 72, 11, 13, 15, 35, 92, 108, 150, 5, 8, 10, 187, 204, 210, 215 }

Partition them into 4 bins & smoothing using bin-boundar

(a) Equal-width

(b) Equal-depth.

T = { 5, 8, 10, 11, 13, 15, 35, 50, 55, 65, 72, 92, 108, 150, 187, 204, 210, 215 }

(b) Equal-depth

bin 1 : 5, 8, 10, 11

bin 2 : 13, 15, 35, 50

bin 3 : 55, 65, 72, 92, 108

bin 4 : 150, 187, 204, 210, 215

(a) Equal width min = 5 max = 215

Range = 210

width = $\frac{210}{4}$ = 52.5

Bin 1 [5, 57.5] : 5, 8, 10, 11, 13, 15, 35, 50, 55

Bin 2 [57.5, 110] : 65, 72, 92, 108.

Bin 3 [110, 167.5] : 150.

Bin 4 [167.5, 220] : 187, 204, 210, 215

~~Chapter 2~~

Transforming categorical variable into Numerical variables:

Ex- Region (East, west, North, South)

Relation

1<2<3<4

East < West < North < South

Categorical variable	Numeric variable
East	1
west	2
North	3
South	4

This is simply transformed categorical variable 'region' into a single numerical variable rather than using several flag variables.

Reclassifying categorical variable :-

Reclassifying categorical variables means combining re-grouping or transforming categorical variables into fewer and more meaningful groups to improve data analysis or model performance. when a categorical variable has too many categories or irrelivant, we merge or rename them to make analysis easier.

Reason -

(i) Too many categories exists.
(ii) Some categories have very few observations.
(iii) several categories have similar meaning.
(iv) To create more useful or predictive grouping for data mining algorithms.

Ex- 50 states can be re-classified as economic-level as Richest, middle-Range and Poor state.

Adding an index field :-

An index field is a new column or variable added to the dataset to uniquely identify each record. It is especially useful when the dataset doesn't contain unique key or id,

Reason -

- To uniquely identify each records
- data tracking
- debugging
- joining dataset

Ex-

Customer

Index	Name	Age	Salary
1	Rima	32	20 K
2	Aditya	55	50 K
3	Raj	15	70 K
4	Malala	22	55 K

Removing variables that are not useful :-

Some variables in the dataset may not contain useful information for data-mining task, these variables should be identified and removed to simplify the dataset and improve perfermance of the model.

Reason -

- To reduce the noise (Irrelevant data)
- Improve computation
- Simplified Interpretation
- Avoid redundancy

Ex- Unary variables, Near unary variables.

Unary variables :-

It is a variable or column that has only one value for all the records in a dataset.

Index	Name	Age	Salary	Country
1	Rima	32	20 K	India
2	Aditya	55	50 K	India
3	Raj	15	70 K	India
4	Malala	22	55 K	India

→ unary variab

Nearly Unary:-

Ex-

Index	Name	Age	Salary	Country
1	Rima	32	20K	India
2	Aditya	55	50K	India
3	Raj	15	70K	India
4	Malala	22	55K	USA

Variables that should probably not be removed:-
Some variables that appear unimportant at first but later turned out to be quite useful

Types of variable -
- Categorical identifiers that groups records meaningful.
- Variables with potential interaction effect.
- variables with missing or rare values.

CHAPTER - 3

Exploratory Data Analysis (EDA)

→ Hypothesis testing (z-test, ANNOVA, chi square test)
→ Exploratory Data Analysis (Regression, clustering, Box plot, histogram, scatter plot)

Churn Dataset (UCI - University of California Institute)

There are two distinct approaches to data analysis.
- Hypothesis testing: A conformatory formal procedure that test a pre-specified idea or assumption.
- Exploratory Data Analysis:
  It is an open ended discovery oriented process where the goal is to learn what the data suggest without a fixed hypothesis.

Both approaches play a complimentary role in data mining statistic and machine learning.

Hypothesis Testing :-
H_0 → null hypothesis
H_a or H_1 → Alternative hypothesis
- It is a statistical method used to make decision or draw conclusion about a population based on sample data.
- It helps to determine wether there is enough evidence to support or reject a particular belief.

Key features:-
- starts with an a prior hypothesis or assumption (before examining the data in detail.)
- involves two competing statement.
  (a) Null hypothesis (H_0)
  (b) Alternative hypothesis (H_1)

Ex- A mobile phone operator may hypothesis
H_0 - Mkt share has not decreased after fee hike.
H_1 - Mkt share has decreased after fee hike.

Exploratory Data Analysis :-

It is an approach to analyse datasets that emphasises visual exploration and descriptive statistics to uncover patterns, anomalies and relationships without pre-determining assumptions.

Objectives :-

- understand the structures of data. (variable, range, columns)
- Examine the distribution of categorical variables.
- Look at the histograms of numeric variables to understand their spread and shape.
- Explore the relationship among set of variables. (Predictor and target variables)
- Detect outliers, missing values and data quality issue.
- Develop initial hypothesis and guide subsequent modelling.

Common EDA techniques

→ Graphical: Histogram, Scatter plot, box plot, heat maps.
→ Numerical: Summary statistics (Mean, median, mode, variance, skewness), correlation coefficient.
→ Subset / Group Analysis
Identifying the clusters, trends, or interesting subsets.

★ EDA acts as the foundation of Data Analysis, shaping the direction of further investigation & hypothesis testing.

Complementary Roles :-

→ EDA comes first (discovery stage)
Helps analyst understand the dataset, distribution and uncover important relationships and patterns that could indicate important areas for further investigation.
→ Hypothesis testing follows (confirmation stage)
validates the patterns or suspicious suggested by EDA with statistical rigor, i.e. testing assumptions with formal procedures.

★ Together they form a powerful cycle of discovery and confirmation in Data mining & statistical analysis.

Hypothesis Testing Vs Exploratory Data Analysis (EDA)

	Hypothesis Testing	EDA
Purpose →	To perform or reject a pre-specified idea.	→ To discover patterns, understood distributions, and generate new ideas.
Approach →	Deductive, confirmatory	→ Inductive, discovery oriented.
When used →	when ~~clear~~ clear, theory driven questions exist.	→ when data are unfamiliar, large or complex.
Focus →	Formal decision making	→ Investigation of variables, distributions and relationships.
Tools →	Statistical tests (t-test, chi-square, ANOVA, regression)	→ Graphical (Histogram, Scatter plot, box plot) & descriptive statistics.
Outcome →	Binary decision (reject/fail to reject H_0)	→ Insights, hypothesis, directions for further study
Flexibility →	Rigid, structured	→ Flexible, iterative

EDA on the Churn Dataset (A case study)

→ The churn dataset (UCI ML repository) is used to demonstrate EDA methods applied in a real world business scenario.
→ EDA helps in :
- Detecting anomalies or missing data
- Identifying patterns & relationships among variables.
- Suggesting potential predictors for the target variable.
- Gaining domain insights through visualization & summary statistics before formal modeling.

Churn Example - Getting to know the Dataset

Overview of Dataset

- Number of observations (Rows): 3333 customers.
- Number of Predictors (Features): 20
- Target variable: Churn - indicates whether a customer has left the company (True or false).
- The dataset contains a mix of categorical, integer valued, and continuous features describing customer demographics, account information, service usage, & interactions with customer service.

Variables in Dataset

(a) Customer Identification

1) State - categorical; 50 US states & the district of Columbia.
2) Account length - Integer; duration (in days) the account has been active.
3) Area code - categorical; geographical area code.
4) Phone number - Unique identifier (effectively a surrogate for customer ID)

(b) Service plans

5) International plan → Dichotomous Categorical (Yes/No)
6) voice mail plan → Dichotomous categorical (Yes/No)
7) Number of voice mail messages → Integer, count of saved messages

(c) Usage Metrics

Total minutes, calls, charges, ... (8 to 19)
(Continuous/Int)

(d) Customer Service Interaction

No. of calls to customer service. (Int) (20)

(e) Target Variable

21) Churn - Boolean (True/False)
Indicates if the customer left the company.

24/10/25

Type	Variables	Description
Categorical	State, Area Code	Indicate geographic origin.
Identification	Phone number	Serves as a customer ID surrogate
Flag variables	International Plan, Voice Mail Plan	Dichotomous variable Yes/No
Numerical (Continuous/Integer)	Total day	Capture usage statistics
Target	Churn	T/F

Univariate Vs Multivariate Analysis

→ Univariate analysis explores a single variable in isolation to understand its distribution, central tendency, spread and shape.
→ It doesn't deal with relationships or dependencies.

Purpose

→ Understand data range, outliers and overall pattern.
→ Identify missing or extreme values.
→ Decide on data transformation (normalisation, log based transformation)
→ Check assumptions for future modeling.

Types of variable	Common Techniques	Visualisation
Categorical	Frequency counts, proportions, mode	Bar chart, pie chart
Numerical	Mean, median, σ, skewness	Histogram, boxplot, density plot

→ Multivariate Analysis investigates two or more variables simultaneously to detect patterns, relationships, correlation, and interactions between them.

Purpose

→ Find dependencies and interaction effects between variables.
→ Identify predictors for a target variable.
→ Support feature selection and hypothesis formulation.

Relationship Type	Typical Analysis	Visualization
Two categorical	Contingency table, chi square test	Clustered bar chart
One categorical + one numeric	Group means, boxplot	Side by side boxplot
Two numeric	Correlation, regression line	Scatter plot
Many numeric	PCA, heatmap	Matrix plot

Contingency Table

→ Also called cross Table.
→ It is a type of table using statistics to show the frequency distribution of the variables i.e how two or more categorical variables are related to each other.

29 | 10 | 25

~~Relation type~~ ~~Visualization~~

Exploring Categorical Variables

International plan vs Churn

A comparison of the proportion of churners and non-churns with international Plan (yes, 9.69% of customers) or without (no, % of customers).

	International Plan (No)	International Plan (Yes)	
Churn = False	2664	186	= 2850
Churn = True	346	137	= 483
	3010	323	3333

Column wise: 3010 + 323 = 3333

2664 346 2664

Churn rate for International plan (No) = $\frac{346}{3010}$ = 11.49

Churn rate for International plan (Yes) = $\frac{137}{323}$ = 42.41

∴ Customers with international plans are over 3x more likely to leave.

for voice mail.

	VMail Plan = No	VMail Plan = Yes	
Churn = False	2411	576	= 2987
Churn = True	483	55	= 538
	2894	631	3525

Column total = 2894 + 631 = 3525 ≠ 3333

	Vmail Plan = No	Vmail Plan = Yes	
Churn = False	2008	842	= 2850
Churn = True	403	80	= 483
	2411	922	3333

Colum total = 2411 + 922 = 3333

Churn rate for Vmail Plan (No) = $\frac{403}{2411}$ = 16.71%.

Churn rate for Vmail Plan (Yes) = $\frac{80}{922}$ = 8.67%.

∴ Those without the plan are twice likely to Churn.

Suggestion: Make voice mail plans more accessible or attractive to increase retention.