U3

– Reporting Authoring –

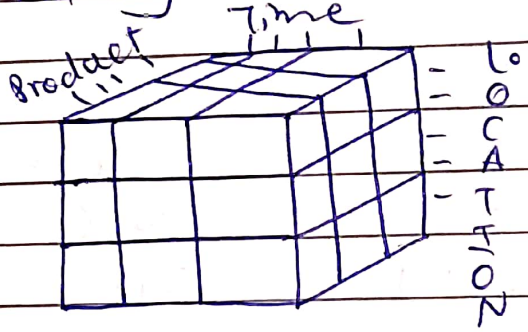
Page No.
Date

# Multidimensional Data Model – (MDM)

MDM is a way of organizing and representing data that focuses on multiple aspects or dimensions of information particularly useful for analyzing large datasets involving various factors and allows for faster retrieval

> OLAP and DW uses MDM databases. It represents data in the form of data cubes. Data cubes allows to ~~more~~ model and view data from many dimensions & perspectives.

> It is defined by dimensions and facts and is represented by fact table




> Once Data cube is ready it can be queried using OLAP such as Slicing, dicing, drill up, drill down, roll up, Pivoting

# Relational Data Model (RDM) –

RDM structured way of organizing data into tables (relations) ~~consists~~ Consisting Rows & Columns. Data is stored and accessed and managed using query language SQL

Key components –

> Tables – table has name & consists of rows (records) and column (attributes).

2) Rows - Individual records in a table, representing a single instance of entity.
3) Columns (Attribute) - Properties of entity
4) Primary key - Unique identifier for each row within table.
5) Foreign key - These are columns that references primary key of another table.
Ex - Employee table.

| ID | Name | Email | Dept ID | Position |
|---|---|---|---|---|
| 1 | XYZ | --@gm | 101 | Manager |
| 2 | ABC | --@gm | 102 | Analyst |
| 3 | SYT | --@m | 103 | Developer. |

Department table.

| Dept ID | Dept Name | Location |
|---|---|---|
| 101 | Sales | NYC |
| 102 | Marketing | IND |
| 103 | IT | BAN |

Employees table has foreign key (DeptID) that establishes a relationship with Department table.

## Types of Reports -

1) List reports - Present detailed info in a tabular format often with rows and colums. Each row represents a data point each column represents an attribute.
Use case
> Displaying customer lists, product catelog

2) Crosstab Reports - Analyze data by summarizing relationships between 2 or more categorical variables in matrix format ~~rows represents variables~~ with categories on the rows and columns and intersection of these categories displays the corresponding value (frequency, count sum, etc.).

Use case
1) Identify trend & patterns betn variable.
2) Analyze the relationship betn variables.

3) Statistical Reports - Summarizes & analyzes quantitative data using statistical measures like mean, median, mode, variance etc. may include tables and charts to present measures.

Use case
1) Providing high quality overviews of central tendency & dispersion
2) comparing different dataset using statistical metrics.

4) Chart Reports - visually represent data patterns and trends using ~~chart~~ charts like bar charts, line charts pie charts or scatter plots. Provides visual pattern & relationships

Use cases
1) Highlighting trends, relations between datapoints.
2) Visual representation.

5) Map Reports - Display data geographically using maps. Can be helpful for analyzing trends distributions or patterns with location.
use case -
1) Identifying geographic areas for specific pattern.
2) Understanding the spatial context of your data

6) Financial Reports - Present financial info about company or organization. ~~They typically~~ includes income statements, balance sheet, Cash flow ~~etc~~.

* Data grouping & sorting -
> Grouping - Involves organizing data points into categories based on shared characteristics. Grouping can make large datasets more manageable by focusing on specific ~~of~~ subset of data. Groping can be used for summarizing data by categories, analyze data pattern & trends in group

> Sorting - ~~Error~~ Sorting arranges data in specific order, either ascending or descending, based on ~~more~~ one or more columns. Sorting is crucial for making it data easier to read & analyze. Sorting used for organizing data in logical sequence, To improve readability, improved data analysis.

Page No.
Date

* Filtering Reports - It is the process of refining data to display only the relevant info that meets specific criteria. It helps users focus on subsets of data for analysis, making reports more insightful.
> Benefits
1) focus on specific Trends
2) Identify outliers
3) Simplify complex data.
> Types of filters
1) value filters - based on specific values in one or more column
2) Range filters - based on range of values.
3) Text filters - filtering based on text patterns.
4) Date filters

* Drill Down - (zooming in) -
Drill down allows user to navigate from higher level to lower level of details within the data
It helps users to break down aggregated data into finer details, uncovering the underlying component.
Example - Imagine sales report that shows Sales by region (eg North, east, west..). Drilling down on the North region might reveal sales figures for each state within North region.

* Drill up (zoom out) - opposit to drill down.
It allows you user to navigate from a lower level of details in data to a higher level of aggrega.n
It collapse detailed view for a more concise view.
Ex - Continuing with Sales report drilled down at North region user could drill up to go back to original view showing total sales by region.

Page No.
Date

* Drill through - Allows user to navigate from a summarized view in a report to a completely different but related data. This action is typically initiated by clicking on a specific element within the report it takes user to a new view

Ex. Imagine a report showing total customer purchases by product category. Drilling through on a specific product category (eg. laptops) might take you to seperate report that details individual laptop models sold, their prices, etc.

Nandeesh.

# Data preparation

Page No.
Date

## Incomplete data -

This term refers to datasets that have missing values or lack certain necessary attributes. This can happen due to various reasons such as data entry errors, loss of data.

> Implication -

1) Bias in Analysis - Missing data can lead to biased results if not handled properly.
2) Reduced statistical power - Incomplete datasets can weaken statistical power of an analysis, making it harder to detect true patterns.
3) Inaccurate Model - ML models trained on incomplete data may not generalize well and can produce unreliable predictions.

> Strategies to handle incomplete data -

1) Data manipulaⁿ - Filling in missing values
   1) Mean, median, Mode imputation - Replacing missing values using with the mean median mode of column.
   2) KNN algorithm (KNN) - Using the nearest neighbors to estimate the missing values.
2) Delition method - If missing data is significant or cannot be reliably imputed, removing affected ~~records~~ data from records might be necessary.

## Data Affected by Noice -

Noisy data refers to datasets that contains errors outliers or irrelevant data points that do not represent true values or patterns.

> ML model might learn the noise instedd of the underlying pattern, reducing their ~~the~~ noise instead of the underlying reducing performance on new data. (generalization).

> Strategies -

1) Data Cleaning - Involves identifying and removing noisy data or replacing them using mathematical techniques like Zscores, ~~Standardization~~, Calculating Inter Quartile range

2) Data filtering - Filtering out records with extreme values of those exceeding a certain threshold to reduce influence of outliers.

3) Clustering Algorithms - Using clustering algorithms like DBSCAN, K-Means can be used to find the main groups in the data and treat points far from any cluster center as noise

4) PCA - Reduce dimensionality by transforming data into set of principal components that capture the most significant variance

5) Regression techniques - Use regression techniques that are less sensetive to outliers Suchas ~~RANSAC~~ RANSAC or Huber regression.

* Data ~~reduction~~ Transformaⁿ -

> Standardizaⁿ - Also known as Z-score normalizaⁿ, involves transforming the data so that it has mean of 0 and standard deviaⁿ of 1

$$z = \frac{x - \mu}{sd}$$

Used in algos ~~like~~ that depends on distance metrics (eg KNN, SVM). In optimizaⁿ algos like gradient descent for faster convergence

By putting features on similar scale, z score Normalization ensures all features contribute equally during training.

< z score normalizaⁿ can be sensetive to outliers significantly affecting mean & standard deviaⁿ.

~~2) Min Max Scaling~~

~~> Feature extraction~~

2) Min Max Scaling - Also known as normalization, data transformaⁿ technique ~~used~~ that adjusts the range of the features to scale values to ~~specific~~ specific range, typically [0,1] or [-1,1].

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

> Feature Extraⁿ - key step in data preprocessing pipeline for ML & data analysis focuses on identifying and extracting the most relevant and informative characteristics from raw data. Purpose -

1) Dimensionality reducⁿ - reduces no. of features while retaining important info.
2) Improved Model performance
3) Noise reduction.

Common techniques

1) Principal Component Analysis (PCA) -

PCA is dimensionality reduction technique commonly used in data analysis. PCA aims to simplify data by identifying the underlying structure and reducing no. of features while preserving imp data.

> PCA Process

1) Standardize data. - z score normalizn.

2) Covariance matrix - Compute covariance matrix, which captures relationships betn features.

3) Eigen decomposition - Perform eigen decompositn on covariance matrix to find eigenvalue & eigenvector eigenvector represents direction of max variance & eigen value represents magnitude of variance.

4) Selecting Principal Component - Selecting top k eigenvector corresponding to the rank by arranging eigenvalues in descending order.

5) Dimensionality reduct - by selecting top k Principal component dimensionality reducn is achieved.

> Benefits

1) Reduced complexity

2) Improved ML performance.

3) Feature extraction.

Data Reduction - Important process in data preprocessing that helps to manage large datasets by reducing their size while retaining ~~original~~ essential info

> Techniques

i) Sampling - Involves selecting a subset of data points from the larger dataset. This subset should represent the overall dataset adequately, allowing for efficient analysis

Types of Sampling -

i) Random Sampling - Selecting data points randomly from the entire dataset. Each data point has an equal probability of being seleted. These can be chosen in various way by selecting ~~eah~~ each data point randomly or dividing dataset into distinct subgroups based on certain characteristic.

ii) Stratified Sampling - Data is split into partition and samples are drawn from each partition

iii) Systematic Sampling - Data points are selected at a fixed interval from an ordered list. This method is efficient but requires data to be well structured.

~~iv) clustering sample -~~

iv) Cluster Sampling - Population divided into groups (clusters) based on shared characteristics Then, random samples are chosen from each cluster.

2) Feature Selection -
Involves selecting the most relevant features (variables) from the dataset while discarding less important or redundant features.
Helps in improving model performance, reducing overfitting & decreasing computational cost.

> Types -
1) Filter method - These technique evaluate individual features based on statistical measures like corelation with target variable. ~~feat~~ Feature with highest scores are selected.
These methods are generally computationally efficient but might not capture complex relations.
Ex - Chi-square test, Corelation analysis, ~~F-value~~, F-value Calculation For ANOVA.

2) Wrapper method - Evaluate feature subsets by training and testing a machine learning model on different combinaⁿ features. Aims to find the subset that produces the best model ~~pers~~ performance.
Techniques
i) Forward - starts with an empty set of features and adds features one by one.
ii) Backward Eliminaⁿ - Starts with all features and removes them one by one evaluating model performance at each step.

3) Embedded methods - This method performs feature selection as part of model training. Some algorithms like LASSO regression or tree based methods performs feature selection during training by assigning weights or importance scores to features.

Data discretization - Also known as binning it is the process of grouping continuous values of variables into contiguous intervals.
It is a data reduction mechanism because it diminishes data from large domain of numeric values to a subset of categorical values.

> Steps -
1) Sorting continuous values of the features
2) Evaluating a cut point for splitting intervals for merging
3) Splitting or merging intervals of continuous value

> Typical Methods of Data Discritizaⁿ -
1) Binning - Techniques in Binning
i) Equal width Binning - Simplest method, where the entire range of the continuous features is divided into specific no. of ~~bins~~ bins of equal ~~length~~ width.
Ex. if the range of features is 0 to 100 and we choose 5 bins would cover an interval of 20 un
[0-20] [20-40] [40-60] [60-80] [80-100]

ii) Equal-Frequency binning - Data is divided into bins containing approximately the same number of data points. However, the bin width might not be equal, especially if the data is not evenly distributed.

2) Histogram Analysis - It is an unsupervised discretization technique because it does not use class info like binning. There are various partition rules used to define histograms. In equal width histogram, values are partitioned in equal size bins. histograms are effective for data with multiple attributes to capture depencies between them.

3) Cluster Analysis - Popular data discritization method. A clustering algo can be applied to descritize a numeric, Partitioning the values into clusters or groups based on similarity scores Partitioning data into clusters There are many choices of clustering definitions & clustering algos like K-means or K-medoids.

4) Corelation analysis - Supervised discritization method. Also known as chi merge algorithm It is performed recursively by finding best neighbouring intervals that have similar distribution of classes and merge them.

# Data Explorat$^n$

1) Univariate analysis - It is simplest form of data analysis that involves examining each variable in a dataset independently. Helps to understanding the basic structure and characteristic of individual variable.

> Graphical analysis of categorical Attributes - Purpose to visualize the distribution & frequency of categorical variable

Techniques

1) Bar chart - Display freq or propor$^n$ of categories. Each Bar shows category & heigh represent freq.
2) Pie chart - Shows relative proportions or different categories. Each slice represents a category's proportion to the whole.
3) Line graph - Used to track changes over period of time

> Graphical analysis of Numerical attributes - To understand distribution, central tendency & Spread of numerical value.

Techniques.

1) Histogram - Displays freq distribution of numerical variable
2) Box plot - Summarizes data using minimum, 1$^{st}$ quartile, median, 3$^{rd}$ quartile & maximum value
3) Density plot - Shows data distribu$^n$ in a smooth continuous curve, useful for identifying data's shape.

> Measures of central tendency - To identify the center point of a numerical dataset.

Measures -

1) Mean = $\frac{\Sigma f}{n} \frac{\Sigma x}{}$

2) Median - middle value
3) Mode - Most frequent value

> Measure of dispersion of numerical Attributes -
To understand spread or variability in dataset
Measures -
1) Range - Range = Max - Min
2) Variance - Avg of Squared difference from mean
Variance ($\sigma^2$) = $\frac{\Sigma (x-\bar{x})^2}{n}$
3) Standard deviaⁿ - Square root of variance indicating data deviaⁿ.
Standar deviaⁿ ($\sigma$) = $\sqrt{\frac{\Sigma (x-\bar{x})^2}{n}}$
4) Interquartile Range (IQR) - Difference bt$^n$ 3rd quartile & 1st quartile.
IQR = $Q_3 - Q_1$

> Identificaⁿ of Outliers -
To detect data points that significantly differ from rest of dataset.
Technique
1) Box-Plot - Oultliers as points beyond whiskers
2) Z-Score - $z = \frac{x_e - \bar{x}_e}{\sigma}$ typically
$|z| > 3$ are considered outliers.
3) IQR method.

2) Bivariate - Involves examining the relationship between two variables to understand how one variable influences or is associated with the other.

> Graphical Analysis -

To visually inspect relation betn 2 variables.

Technique.

1) Scatter plot - used 2 numerical variables, showing data points plotted on 2D graph. The pattern indicates relations (positive, negative or no relan).

2) Line plot - useful for time series data showing trends over time.

3) Box plot comparison - comparing numerical variable across different categories of variables.

4) Heatmap - Represents relations betn 2 variables using colors to show co-relation.

> Measures of Correlation for Numerical Attribute

To quantify the strength & direcn of relationship betn 2 variables.

Technique

1) Pearson Correlation Coefficient (r) - Measures linear relationship betn two continuous variable

Value range :- -1 to +1

Positive value - Positive linear relation

Negative value - Negative linear relation

2) Spearman Rank Correlan - Measures monotonic relationship betn two ~~var~~ variables based on their rank

useful when dat is not normally distributed.

3) Contengency table - To summarize relationship between 2 variables. A matrix that displays the freq. distribun of variables. Each cell represent the count of occurrences

Chi-square test used to determine significant associa$^{n}$ betn 2 variables.

3) Multivariate - Involves examining more than 2 variables to understand their relationships. This type reveal complex patterns in data.

→ Graphical Analysis -

1) Scatter plot Matrix.
2) Heatmap
3) 3D scatter plot
4) Parallel Co-ordinate plot - Line represent each data point and each axis corresponds to different variable.

2) Measure of correlation.

1) Correla$^{n}$ matrix - A table of Correlation betn multiple variables.
2) Pearson Correlation
3) Partial Correla$^{n}$ - extends pearson's correla$^{n}$ it helps to isolate the correlation betn 2 variables while controlling effect of another variable.

| | Univariate | Bivariate | Multivariate |
|---|---|---|---|
| 1) | single variable at a time | Relationship between 2 variables | between multiple variables |
| 2) | understand distribution (center, Mean, Median mode) | Identify pattern correlations | Explore complex relations identify clusters. |
| 3) | Freq tables. Mesures of central tendency (mean, median Mode) variance, Stndard deviation | Scatter plot, correlation coefficients (Pearson Correlation, Spearman's rank) | Scatter plot matrix pair plots, 3D Scatter plot, Partial Correlation. |
| 4) | To understand basic characteristic of single variable | Measure & describe Streyth and direction of relationship | To understand interactions and combined effects of multiple variables. |
| 5) | Graphical techniq Histogram Bar charts Box plot | Scatter plot Box plot Heatmap | Scatter plot matrix Pair plots, 3D Scatter plots ~~pr~~ |

Scanned with OKEN Scanner

Q5

Impact of ML on BI Process.

Page No.
Date

~~Classification - It is a Supervised learning tasks that involves predicting the category or class of a given data points. Cat Good is~~

Fundamental task

Classificaⁿ - It is a fundamental task in ML that involves assigning data points to specific categories such as classifying reviews to positive or negative.

Classification Problems - These problems involves dependent & independent variable dependet variables used for training model to predict dependant variable.
eg - Spam ham classificaⁿ on emails.
Types of classificaⁿ problems -
1) Binary classificaⁿ - Predicting one of two possible categories.
2) Multiclass - Predicting one of many possible categories.
eg - handwritten digit recognition.

Evaluaⁿ of classificaⁿ model -
1) Accuracy - $\frac{\text{No. of Correct predicⁿ}}{\text{Total no. of predicⁿ}}$

2) Precision - Ratio of correctly predicted positive observaⁿ to the ~~toata~~ total predicted positives.

$$\frac{TP}{TP + FP}.$$

3) Recall = $\frac{TP}{TP + FN}$

4) F1 Score = $2 \frac{Precision \times Recall}{Precision + Recall}$

5) Confusion matrix - Table describing performance of classification model

| | 0 | 1 |
|---|---|---|
| 0 | TP | FP |
| 1 | FN | TN |

0 1

6) AUC-ROC curve - graphical representation of TP rate versus FP rate at various threshold settings. AUC represents Area Under the Curve.

~~A~~ Baysian methods -

> Bayes Theorem - Statistical and probability theory is a mathematical formula used to determine the conditional probability of events. It describes the probability of an event based on prior knowledge of the conditions that might be relevant to the event.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

P(A|B) (Posterior) - ~~Probab~~ P(A) after observing B
P(B|A) (likelihood) - P(B) given A is true
P(A) - P(A) before observing B
P(B) - total probability of observing B.

It is based on 3 components.
1) Background knowledge - model's parameter
2) Data - observed evidences
3) Posterior inference - combinaⁿ of first 2 component

> Naive Bayes classifier - It is a classificaⁿ based on Bayes' Theorem. ~~with an independent~~ The 'naive' assumption is that all features are conditionally independent. Commonly used in text classificaⁿ tasks like spam ~~ham~~ detecⁿ sentiment analysis.

> Baysian inference - used in fields like medical ~~dig~~ diagnosis, where the likelihood of a disease is updated based on test results.
ex - ~~P(D~~

$$P(\text{Decease}|\text{PositiveTest}) = \frac{P(\text{PositiveTest}|\text{Disease}) \cdot P(\text{Disease})}{P(\text{PositiveTest})}$$

> Baysian networks - Graphical model that represents probablistic relationship among ~~of~~ a set of variables

> Markov Chain Monte Carlo - Probability distribation based on constructing a Markov chain.

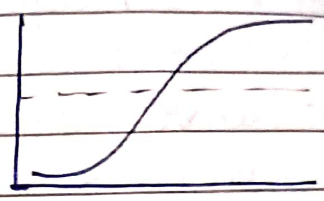
> Benefits of Baysian methods -
1) incorporating prior knowledge
2) Handling uncertainity
3) Flexibility - They can applied to problems with limited data or complex models.

(LR)

Logistic Regression — Fundamental ML task. For classificaⁿ tasks in ML, Logistic regression deals with categorical dependant variables.

> Logistic regression employs a Sigmoid function to map the linear combination of its features to probability value betⁿ 0 & 1.

$$LR = \frac{1}{1 + e^{-z}}$$




> LR is trained using an iterative algo that aims to minimize cost function.

> By adjusting the weights and bias term through optimization process.

> Applicaⁿ -
Spam filtering
Fraud detecⁿ
customer churn.

Clustering — Grouping data points together based on their inherent similarities without predefined labels. It's like sorting a pile of fruit based on their characteristics (color, shape, etc)

> Clustering methods -

> Partitioning method - dividing data points into a fixed no. of clusters (K) in one go. iteratively refine the cluster.
~~model~~

Page No.
Date

Partitioning models -

1) K means clustering - widely used partition method. randomly assigns data points to k initial clusters and then iteratively perform steps -
> Calculate centroid - mean value of each feature is calculated for each cluster creating centroid.
> Reassign Data points - Each data point is assigned to the cluster with nearest centroid.
> Update centroid & reassign.

2) K medoid clustering - similar to K mean but uses medoid instead of means. Medoid is centrally located data in cluster.

2) Heirarchical method - Creates tree like structure representing relation between data points.
Models -
1) Agglomerative - starts by considering each data point as a seperate cluster. Then iteratively merges the 2 closest clusters based on distance measures until single cluster remains.
2) Divisive - Approach starts with all data points in a single cluster and then iteratively splits cluster that maximizes a seperation criterion until stopping criterion is reached.

3) Density-Based clustering method - type of unsupervised ML Technique, density based method focus on areas with high concentration of data points

Page No.
Date

Models -

1) DBSCAN - Identifies high density areas by focusing on core points - data points surrounded by minimum number of neighbors within a specified radius.
Clusters are formed ~~based on~~ by connecting core points that are density reachable. ie reaching from one core to other core point through a chain of neighbors with radius.

2) OPTICS. - Similar to DBSCAN, OPTICS focuses on density but takes different approach. It orders data based on reachability distance. No need to predefine the radius ~~or~~ or minimum neighbors -

* Evaluaⁿ of clustering models.

~~Clustering algos are~~

> Internal Evaluaⁿ Metrics.

1) Inertia (Within cluster sum of Squares) - It calculates sum squared distances between ~~Each~~ each data point and centroid of cluster. Lower inertia implies that clusters are compact and well seperated.

2) Silhoutee Coeff - Evaluates how well data points are assigned to their clusters. It considers both the distance to the assigned cluster's centroid and the distance to nearest ~~the~~ neighbouring cluster's centroid.

3) Davies-Bouldin Index - Evaluate the average similarity ratio of each cluster with the cluster that is most similar to it. Lower values indicate bete better clusters.

3) Random Index - Measures the similarity betn clustering results and the groud truth labels. considers both true postives & true negatives.

4) Normalized Mutual Info (NMI) - Measures the amount of info shared between the clustering results and groud truth labels.

5) Calinsk-Harabez Index - Compares average distance between clusters to the variance within clusters. Higher value suggests better seperation

6) Davies Bouldin Index - It measures the ratio of the within-cluster scatter to the between cluster seperan; lower value better clustering

A Association Rule - It is a technique in data mining that helps discover interesting relations between items in a large dataset. Often applied in market basket anadysis, where goal is to find assoeian between items purchased together.

Structure -

1) Anteeedent (x) - set of item that appear in rule's premise (before the arrow (=>))
2) Consequent (Y) - set of items that appear in rule's conclusion (After the arrow (=>)).
3) Support - Frequency of occurence of both X & Y
4) Confidonce - Probability of Y appearing in transac given x is present

5) Lift - ratio of observed support to expected support if X & Y were independent.

> Apriori Algo - It is one of the algorithm used for transaction data in Associa" rule learning. Allows us to mine the frequent itemse in order to generate associa" rule beh them.

1) define min support & min Threshold confiden
2) find frequent item sets
3) Generate Associa" rules

Ex -

| TID | Items |
|---|---|
| $T_1$ | I1, I2, I5 |
| $T_2$ | I2, I4 |
| $T_3$ | I2, I3 |
| $T_4$ | I1, I2, I4 |
| $T_5$ | I1, I3 |
| $T_6$ | I2, I3 |
| $T_7$ | I1, I3 |
| $T_8$ | I1, I2, I3, I5 |
| $T_9$ | I1, I2, I3 |

min support count is 2 ~~min coeff~~
min confidence is 60% Frequency Table -

| ~~TID~~ | |
|---|---|
| I1 | 6 |
| I2 | 7 |
| I3 | 6 |
| I4 | 2 |
| I5 | 2 |

minimum support count is 2 so every item is kept in table no one is removed.

Page No.
Date

Frequency for itemsize 2

| | | |
|---|---|---|
| (11,12) | - 4 | (11,12) |
| (11,13) | - 4 | (11,13) |
| (11,14) | - 1 | (11,15) |
| (11,15) | - 2 | (12,13) |
| (12,13) | - 4 | (12,14) |
| (12,14) | - 2 | (12,15) |
| (12,15) | - 2 | |
| (13,14) | - 0 | |
| (13,15) | - 1 | |
| (14,15) | - 0 | |

min support is 2 →

Frequency for itemsize 3

| | | |
|---|---|---|
| (11,12,13) | - 2 | (11,12,13) |
| (11,12,14) | - 1 | (11,12,15) |
| (11,12,15) | - 2 | |

min. support is 2 →

Associa rule for (11,12,13)
min support 2 min confidence - 60%.

| Rule | Support | Confidence |
|---|---|---|
| (11^12) → 13 | 2 | 2/4 = 50% |
| (11^13) → 12 | 2 | 2/4 = 50% |
| (12^13) → 11 | 2 | 2/4 = 50% |
| 13 → (11^12) | 2 | 2/6 ≈ 33.3% |
| 12 → (11^13) | 2 | 2/7 ≈ 28.57% |
| 11 → (12^13) | 2 | 2/6 ≈ 33.3% |

Confidence = Support (A ∪ B) / support (A).

eg (11^12) → 13 = Sup((11^12)∪13) / S(11^12) = 2 ← count of triplet formed / 4 ← count support of 2(11,12) together from freq table.

= 1/2 ≈ 50%

∴ (11,12,13) does not hold any associa rule.

For (I1, I2, I5)

| Rule | Support | Confidence. |
|---|---|---|
| (I1 ^ I2) → I5) | 2 | 2/4 = 50%. X |
| (I1 ^ I5) → I2 | 2 | 2/2 = 100%. |
| (I2 ^ I5) → I1 | 2 | 2/2 = 100%. |
| I1 → (I2 ^ I5) | 2 | 2/6 ≈ 33.3%. X |
| I2 → (I1 ^ I5) | 2 | 2/7 ≈ 28%. X |
| I5 → (I1 ^ I2) | 2 | 2/2 = 100%. |

Consider those rules having confidence more than min confidence level ie 60%.
ie rules (I1 ^ I5) → I2, (I2 ^ I5) → I1 & I5 → (I1 ^ I2)

So the Association rule For (I1, I2, I5) is.
(I1 ^ I5) → I2, (I2 ^ I5) → I1 & I5 (I1 ^ I2)

Scanned with OKEN Scanner

# Unit 6
## BI Applications

Page No.
Date

* Role of Analytical tools in BI -
Analytical tools are the backbone of BI. They act as powerful instrument for collecting, transforming analyzing & visualizing data to extract valuable insight for decision making.
> BI tools can connect to various data sources to gather info
> Data might need cleaning, filtering & transformation to ensure consistency and usability for analysis.
> BI tools provides descriptive statistics and visualization to summarize data. By drilling down into data, analytical tools help identify the cause of observed outcome.
> Advanced Analytica tools help to predict on new data using ml algos. This includes predicting Sales forecast, customer behaviour etc.
> Analytical tools allows creation of interactive dashboards. Tools enable the generation of detailed report that can be customized.
> After all analysis these tools helps in further business decision making.

* Analytical tools -
> WEKA (Waikato Environment for Knowledge Analysis) - Designed for data mining tasks provides collection of visualizing tools & algos.
> It offers wide range of ml algos, including classification, regression, clustering & association rule mining
> Weka provides visualization tools to help understand the data and the learning algo.
> It has user friendly GUI

2) KNIME - free open source platform for data mining & ml.
> KNIME easy to use grap GUI to create workflows for data analysis. offers drag & drop workflow for building data analysis pipelines
> Set of powerful extensions and integrations make KNIME a versatile & scalable.
> Tools like R, Python, and WEKA can be integrated with KNIME.
> KNIME includes tools for data preprocessing transformation, modeling & visualization.

3) RapidMiner — Data Science platform that provides an integrated environment for data preparation, ml, dl, text mining.
> Offers a free open source edition and a paid commercial ~~version~~ version.
> Similar to KNIME it uses drag and drop approach for building workflows.
> Includes data preprocessing, ml, text mining and data visualization.

4) R - Freely available with vast collection of packages for various statistical & graphical techniques.
> CRAN (Comprehensive R Archive Network) hosts thousands of packages that extends R's capabilities.
> R excels at data visualization, offering packages like ggplot2.

(BA)

Page No.
Date

Business Analytics - Science of extracting meaningful info from data to improve business decision making. BA focuses on transforming raw data into from various sources into actionable insights that can guide business strategies.

> key stages
1) Data acquisition - Data collecⁿ from various sources
2) Data analysis & Exploraⁿ - Apply statistical methods data mining & ml algos. to uncover patterns.
3) Data visualizaⁿ.

> Benefits
1) Data-driven decision making
2) Operational Efficiency - identify areas for improvement in processes.
3) Risk management.
4) Improved customer understanding.

> Application
1) Marketing
2) Sales
3) Finance
4) Human Resource.

ERP & BI - ERP & BI are two critical systems that work together to streamline operations, improve decision making and gain insights.

> ERP System Acts as a central ~~ba~~ hub for integrating and managing core business process like accounting, inventory, hr.
> A single database that ensures consistency and ~~accurate~~ accuracy of data across the organizaⁿ
> Provides real time access to data, improving visibility & decision making.

> BI refers to technologies, applications and practices for the collection, integration, analysis and presentation of info.
> Integrating ERP systems with BI tools enhances the value of both by combining operational data
> Benefits.
1) Enhanced decision making
2) Improved ~~do~~ performance
3) Comprehensive insight
4) Forecasting & planning

* BI applicat in CRM -
Customer Relationship Management (CRM) systems Focus on managing all aspects ~~Foa~~ of a Customer's relationship with a business, while BI provides tools to extract valuable insights
Application -
1) Customer Segmentat - BI analyzes customer data to segment customers into distinct groups based on demographics, ~~Rich~~ purchase behaviour, interests, etc.
2) Customer churn - Analyzes customer data and interact pattern. helps to prioritize customer and personalize marketing strategies.
3) Optimizat customer service -
4) Decision making.

# BI applicaⁿ in logistics & production

In the logistics & production sector, a significant volume of data is generated, derived from multiple operations that are carried out everyday

> BI application -

1) Inventory management - Moniter & manage inventory levels, ensuring optimal stock levels are maintained.
   Tools - dashboards, Predictive analytics.
2) Demand Forecasting - To predict future product demand based on historical data.
   BI Tools - Time series analysis, ml algos, predictive modeling.
3) Production planning & Scheduling - BI helps analyze hitorical data on production times, machine performance & material requirements to optimize producⁿ schedules.
4) Quality Control - Analyze data on product defect and quality control process
5) Transportaⁿ management - BI can analyze data on transportaⁿ cost, carrier performance, delivery routes. This analysis allows us to find cost saving opportunities, optimizing routes & improved delivery
6) ~~Enhanced~~
6) Enhance customer ~~servt~~ service.

## Role of BI in Finance

1) Financial reporting
2) Budgeting & forecasting
3) Risk management.
4) Performance management.
5) Investment analysis
6) Fraud detection & prevention.

## Role of BI in Marketing

1) Customer segmentation -
2) Customer behaviour analysis
3) Sales performance management
4) Market trend analysis
5) Social media analytics.
6) Customer churn.

## Role of BI in Telecommunication -

1) Customer churn prediction
2) Network optimization
3) Fraud detection.
4) Service usage
5) Predictive maintainance
6) Network investment & capacity planning

## BI in Salesforce management.

1) Sales performance analysis
2) Pipeline & forecast management
manage Sales pipeline & forecast future sale.
3) CRM integration
4) Sales activity monitoring
5) Churn prediction -
6) Customer segmentation.

Page No
Date

## BI in HR management

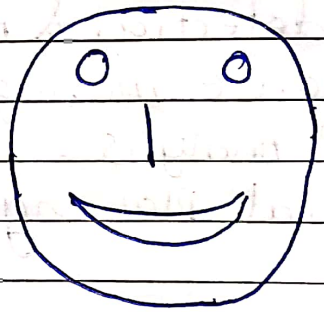
1) Recruitment & Talent Acquisi^n - To streamline hiring process & identify best condidates.
2) Employee performance management.
3) Workforce Analytics.
4) Employee retention
5) Turnover Analysis
6) Training & Development
7) Workforce planning & optimiza^n

## BI applica^n in Banking

1) CRM
2) Risk management
3) Credit scoring & Loan management.
4) Fraud detection & Preventions.
5) Performance management.
6) Transac^n Analysis
7) Financial Planning & Analysis.

So This ends here last subject & last exam of my engineering student life, It was an incridible journey.
I wish you All the best for your future :)

Ae
Engineering ke ctrl+c aur ctrl+v ke dino ko alvida kehete hue ab 'Save as' karke naye chapters Shuru karte hain! Farewell coding buddies aur meri cyber sweetheart (Crush), ab 'Run' karte hain apna 'life.exe' All the best aur yaad rakhna, life ke bugs ka fix toh hum sab kar he lenge.



Surender