ASSIGNMENT No. 3

Q1) Explain how to build reports with relational vs multidimensional data models?

Ans • Relational data Models:-

1. It is a table-based approach where data is stored in tables with rows and columns.
2. It is ideal for applications that require simple queries and ad hoc reporting.
3. To build reports we follow these steps:-

i) Identify the data for which report needs to be build upon.
ii) Write SQL queries to extract data from relational database.
iii) Choose a reporting tool to report the data.
iv) After choosing the reporting tool we start designing our report.
v) Test & refine the report.

• Multidimensional Data Models:-

1. It is a cube-based approach where data is stored in multiple dimensions, such as time, geography, product and so on.
2. It is ideal for applications that require complex analysis and reporting.
3. To build report we follow these steps:-

i) Determine which cubes in database contain data for the report.
ii) Determine the dimensions to display, conditions / filters to apply for the report.
iii) Select appropriate tool such as Microsoft excel / PowerBI.
iv) After retrieving the data one must format it into a report format.
v) Share the report with the intended audience.

Q2) Explain types of reports :- lists, crosstabs, statistics, chart, map.

Ans • List Report :-

1. It is a simple report that presents data in tabular format.
2. Consists of a table with columns and rows and displays data in a sequential order.
3. It is useful for displaying data in a simple and organized manner.

• Crosstab Report :-

1. It is a type of report that summarizes data by creating a cross-tabulation or pivot table.
2. It shows how data is distributed across multiple dimensions.
3. It is useful in analyzing large dataset and identifying trends or patterns.

• Statistics Report :-

1. A statistics report is a report that presents descriptive statistics such as mean, median, mode, standard deviation, etc.
2. It is useful in summarizing data and identifying key trends.

• Chart Report :-

1. It is a graphical report that makes it easy to understand trends and patterns.
2. Charts can be created in various formats such as bar graphs, pie charts, line graph etc.
3. They make visualization of data easier and provide lucidity of the data.

• Map Report :-

1. It is a type of report that displays data geographically.
2. It is useful in analyzing data related to geographic locations, such as sales data, customer location.

3. Map reports provide insights into regional trends & patterns.

Q3) Define terms : Data Grouping and sorting.

Ans • Data Grouping :-

1. It involves grouping together similar data values based on a common attribute or characteristic.
2. For example we can group sales data by product category or customer region.
3. Grouping data can make it easier to analyze and summarize large dataset.

• Sorting :-

1. It is the process of arranging data values in particular order.
2. Data can be sorted in :-
   - Ascending order or Descending order.
   - Based on attributes.
   - Alpabetical Ordering.
   - Numerical ordering.
3. It helps in anlyzing, comparing & identifying patterns & trends.

Q4) Explain drill-up, drill-down and drill-through capabilities?

Ans Drill-up, drill-down, drill-through are capabilities of data exploration that allow users to navigate through hierarchial data and get more granular insights into data.

• Drill-down :-

1. Drill-down is the process of moving from a higher-level view of data to a more detailed view of data by expanding a hierarchical data structure.
2. For example if we have sales data organized by year, drill down would allow sales data to be broken down by quarters or months within each year.

- Drill-up:-

1. It is the reverse of drill-down where we move from a more detailed view of data to a higher-level view of data.
2. For example; if we have sales data by quarter, drill up will allow us to see sales data for the entire year.

- Drill-thorough:-

1. It is the processing of navigating from summary report to a detailed report by clicking on a data point in the summary report.
2. For example, if we have a report that summarizes sales data by product category, drill-through would allow you to click on a specific product category and see a detailed report that shows the sales data for that category broken down by product, region or other attributes.

Qs) Explain terms :- filtering reports, adding calculations to reports, conditional formatting, adding summary lines to report.

Ans Let us see each term:-

1. Filtering report:

- Process of selecting a subset of data to included in a report based on a specific criteria.

2. Adding calculations to reports:-

- Calculations are used to perform arthimetic or logical operations on data in a report.
- Adding calculations to a report allows you to derive meaningful insights from data and present it in a format easy to understand.

3. Conditional formatting:-

- Conditional formatting helps change appearance of data based on specific conditions or rules.
- It helps draw attention to important data points and

make report visually attractive and appealing.

4. Adding summary lines to reports :-

- Summary lines are used to aggregate data in report.
- Summary lines provide overview of key metrics and make it easy to compete/compare data across different categories or time periods.

ASSIGNMENT 4

Q1) Explain in detail data validation and data transformation?

Ans Let us look at the two in detail :-

- Data validation :

1. It is the process of ensuring data is complete, accurate and consistent.
2. It involves data for errors, inconsistency and missing values.
3. Our primary goal is that the data must be fit to be used in analysing and reporting.
4. Common techniques used :-
   - Data profiling :- Analyze data to identify patterns.
   - Data cleansing :- Involves correcting incomplete data.
   - Data auditing :- Reviewing data.

- Data transformation :-

1. It is the process of converting data from one format or structure to another.
2. It is used to prepare data for analysis or reporting.
3. It involves several techniques such as...
   - Data aggregation :- Combine data from multiple sources.
   - Data normalization :- Organizing data in consistent format.
   - Data enrichment :- Oversampling/undersampling the data.
   - Data summarization :- Reducing data to smaller size.

- Data validation and summarization/transformation are essential steps in preparing data for analysis and reporting.
- Data validation ensures data is complete & accurate while data transformation helps to convert data into a format that is easy to analyze & interpret.

Q2)

Ans

Q3)

Ans 1.

2.

Q2) Explain data reduction with sampling, feature selection, principal component analysis.

Ans Data reduction is the process of reducing the size and complexity of data, while preserving its integrity. Several techniques that help us do this are as follows:-

- Sampling :-
  - It involves selecting a subset of data from a larger dataset.
  - Sampling can be done randomly or systemically, and the sample size must be large enough to be representative of the population to be analysed.
  - Different sampling techniques are simple random sampling, stratified sampling, cluster sampling and systematic sampling.
- Feature selection:-
  - It involves selecting most relevant features from all given features.
  - This done when dataset has many irrelevant features or has features that are highly correlated to each other.
  - Different feature selection such as filter methods, wrapper methods, and embedded methods.
- Principal component analysis (PCA):-
  - PCA is a technique for reducing the dimensionality of a dataset.
  - The principal components represent the maximum variance in original data, and they can be used in subsequent analysis.
  - PCA can help to reduce the complexity of dataset, improve data visualization, and identify patterns and relationship in the data.

Q3) Explain data discretization?

Ans 1. Data discretization is a process of converting continuous data into discrete data by dividing data into intervals or categories.

2. Its main objective is to reduce complexity of continuous

data and transform it into categorical data that can be more easily analyzed or modeled.

3. This process involves grouping data values into intervals or categories based on some criteria.

4. Techniques for data discretization:-
   - Depth binning: Data divided according to range.
   - Clustering: Data values grouped by similarity.
   - Decision tree-based :- Discretization using decision tree.
   - Entropy-based :- Data is grouped by entropy value.

5. Data discretization is useful in :
   - simplying data for analysis.
   - Reducing noise & improving accuracy.
   - Improving interpretability of data.

6. However data discretization can also result in loss of information & precision.

Q4) Differentiate between univatate & bivariate analysis.

Ans

	Univariate analysis	Bivariate analysis
1.	Summarizes single value.	Summarizes two variables.
2.	Does not deal with causes & relationships.	Deals with causes and relationships.
8.	No dependent variable.	Contains dependent variable
4.	Main purpose is to describe.	Main purpose is to explain.
5.	Analysis is not done.	Analysis is done.
6.	Example :- Height.	Example :- Sale of ice-creames and the degree of temperature.

Q6) Explain multivariate analysis with graphical analysis & measures of correlation for numerical attributes.

Ans ➤ Multivariate Analysis :-

1. It is a statistical method that involves analyzing multiple variables simultaneously.
2. The goal here is to identify patterns, relationships and dependencies between variables.
3. It is often used in data science, machine learning and other fields to gain insight into complex datasets.

• Graphical analysis :-

1. It is one of the most common methods of multivariate analysis.
2. Graphical methods involving charts, graphs are used to explore relationships between multiple variables.
3. These representations help identify patterns and trends in the data and provide insights in relationships between variables.

• Measures of coorelation :

1. These are used to quantify strenth & direction of the relationship between two numerical variables.
2. Correlation coefficients are used as a common measure of correlation and one typically calculated using pearson's correlation or Spearman's rank correlation coefficients.
3. They are given as :-

$$\text{Pearson's coefficient} = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma (x_i - \bar{x})^2 \Sigma (y_i - \bar{y})^2}}$$

$$\text{Spearman's rank Coefficient} = 1 - \frac{6 \Sigma d_i^2}{n(n^2 - 1)}$$

ASSIGNMENT No. 05

Q1) Define classification ? explain classification problems.

Ans 1. Classification refers to the process of categorizing or grouping objects or data into distinct classes or categories based on their characteristics or attributes.

2. It is a fundamental task in machine learning and data analysis, where the goal is to develop a model that can automatically assign new, unseen instances to predefined classes.

3. Some classification problem faced are as follows:-

i) Overfitting:
- It happens when model perfoms well on training data but fails to generalize on unseen data.
- It mostly happens when model becomes too complex.

ii) Underfitting:-
- It happens when model is too simple or lacks necessary complexity to generalize well to unseen data.
- It happens is too simple and does not fit well.

iii) Curse of dimensionality:-
- When number of features is too large, the computational complexity and sparsity of data points is quite high.

iv) Noise and outliers:-
- Noisy or erroneous data points & outliers significantly affect the classification process.

v) Feature selection:-
- Choice of relevant features is crucial for accurate classification.

vi) Imbalanced data:-
- Imbalanced data occurs when distribution of classes in the dataset is uneven, with some requiring undersampling or oversampling.

Q2. Explain evaluation of classification models?

Ans 1. Evaluation of classification models involves assessing the performance and effectiveness of a models in predicting the correct class labels for unseen instances.

2. Commonly evaluation methods for classification are as follows

- Accuracy :-
  - Proportion of correctly classified instances out of the total number of instances.
- Confusion Matrix :-
  - A confusion matrix provides a more detailed evaluation by presenting the actual and predicted class labels.
  - It shows true positives, true negatives, false positives, and false negative.
- Precision & Recall :-
  - Precision is the ratio of true positives to the sum of true positives and false positives.
  - Recall is the ratio of true positives to the sum of true positives and false negatives.
- F1 Score :-
  - F1 score combines precision and recall into a single metric.
  - It measures tradeoff between precision & recall.
- ROC Curve :-
  - It is the graphical representation of trade-off between true positive rate (TPR) and the false positive rate (FPR).
  - It helps evaluate model's performance across various thresholds and provides insights into the model's ability to discriminate between classes.
- Validation set :-
  - It is a seperate portion of data set that is used to evaluate the model during the training process.
  - It helps in tunning hyperparameters and detect

overfitting or underfitting.

Q3) Explain Bayesian methods?

Ans 1. Bayesian methods are a class of statistical techniques that are based on the principles of Bayesian inference.

2. They provide a framework for updating and revising beliefs about uncertain quantities or parameters based on new evidence or data.

3. Key Components of Bayesian methods are as follows:-

- Prior Probability :-
- Bayesian methods begin with the specification of prior probability distributions, which represent the intial beliefs or knowledge about unknown quantities.

- Likelihood Function :-
- It represents the probability of observing the data given the values of the unknown parameters.
- Quantifies relationship between observed data and parameters of interest.

- Baye's Theorem :-
- It is the fundamental principle of Bayesian inference.
- It mathematically combines the prior probability distribution and likelihood function to obtain the posterior probability distribution.
- It is given as :-

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

- Markov Chain Monte Carlo (MCMC) :-
- Bayesian methods rely on MCMC to sample from the posterior distribution when analytical solutions are not feasible.
- It generates a sequence of samples from the posterior

distribution, allowing for the estimation of various summary statistics or perform complex analysis.

Q4) Explain clustering methods partition methods and hierarchial methods?

Ans 1. Clustering Methods are unsupervised learning methods used to group similar objects or data points together based on their inherent patterns or similarities.

2. Two commonly used types of clustering methods are partioning methods & hierarchial methods.

- Partion-based methods:-
- Partition-based clustering methods divide the dataset into non-overlapping subsets or partitions, where each partition represents a distinct cluster. Some algorithms are:-

* K-Means :-
- It aims to partition data into k clusters, where k is pre-specified.
- It iteratively assigns data points to nearest cluster.

* K-Mediods:-
- K-medoids is similar to K-means, but uses actual data points as cluster representatives instead of centroids.

- Hierarchial methods:-
- These methods create a hierarchy of clusters by recursively merging or splitting clusters based on thier similarities.
- They form a tree structure called as dendograms.
- Two main types of hierarchial methods are:-

* Agglomerative:
- It iteratively merges most similar clusters until a single clusters is formed.

* Divisive:-
- It starts with entire dataset as a single cluster and recursively splits clusters into smaller subclusters.

Q5 Define terms of Association Rule: Structure of Association Rule and Apriori Algorithm.

Ans. Association Rule :-

1. It is a pattern that describes a relationship between items in a dataset.
2. It typically takes the form of "if X, then Y," where X and Y represent sets of items.
3. Association rules are used in market basket analysis and other applications to uncover interesting relationship or associations among items.

- Structure :-

Antecedent (X) ⟶ Consequent (Y)

The antecedent (X) represents the set of items that precede the arrow and are used to predict the consequent (Y).

- Apriori algorithm :-

1. It is a classic algorithm for mining association rules from transactional datasets.
2. It works in two step process :-

a. Frequent Itemset Generation :-

- Algorithm scans transactional dataset multiple times to discover frequent individual items.
- It uses them to generate candidate itemsets of length two.

b. Rule Generation :-

- Once frequent itemsets are obtained, the algorithm generates associations rules by considering different combination of antecedents and consequents.

3. The Apriori algorithm is efficient because it prunes the search space by exploiting the apriori property.

ASSIGNMENT 6

Q.1) Explain tools for business intelligence and role in analytical tools in BI?

Ans Tools for business intelligence are software applications or platforms that help organizations collect, analyze, and visualize data to support decision making and gain valuable insights. Some common tools are :-

1. Datawarehousing tools :-
- These tools facilitate collection, integration, and storage of data from various sources into a centralized repository known as a data warehouse, ensuring consistency, quality and accessibility for analysis.

2. Reporting & Querying Tools :-
- Reporting tools enable users to create and generate reports based on data from the data warehouse. Querying tools allow users to retrieve specific information from data warehouse using queries.

3. Online Analytical Processing (OLAP) Tools :-
- OLAP tools provide multidimensional analysis capabilities to explore data in different dimensions.

4. Data Visualization Tools :-
- The tools focus on presenting data in visually appealing & interactive formats such as charts, graphs, map, and infographics.

5. Data Integration & ETL :-
- ETL tools help extract, transform and load data into data warehouse.

6. Role of Analytics in BI :-
1) Gaining Insights in data.
2) Support Decision Making

3. Improve operational efficiency & productivity.

Q2) Explain Case study of analytical tools :- WEKA, KNIME, Rapid Miner, R.

Ans. WEKA:

- WEKA stands for Waikato Environment for Knowledge Analysis.
- It is open-source data mining tool also used for ML.
- Provides comphrehensive collection of algorithms & tools.
- It is used for preprocessing, classification, regression, clustering, association rule etc.
- It has been used in healthcare, finance & marketing.

• KNIME.

- KNIME stands for Konstanz Information Miner.
- It is open source data analytics & integration platform.
- It seamlessly integrates with other programming tools, and is widely used in pharmaceuticals & manufacturing.

• Rapidminer:

- It is a powerful and user-friendly data science platform.
- It provides visual workflow environment by connecting pre-built components called operators.
- It is widely used for churn prediction, demand forecasting and sentiment analysis.

• R :-

- Open-source programming language and environment for statistical computing and graphics.
- R has a strong community support and is widely used in academia and industry for tasks like statistical modelling, data visualization, advance analytics.
- Particularly used by statisticians and data scientists due to its versatility and extensive statistical capabilities.

Q3) Explain Data analytics & Business Analytics?

Ans • Data Analytics:-

- It focuses on extracting meaingful insights from data.
- It helps in decision making and improve processes.
- It involves the following steps:-

1) Data collection from varibus sources.
2) Data cleaning & preparation.
3) Exploratory Data Analysis.
4) Statistical Analysis on data.
5) Data Visualization & Reporting.

• Business Analytics:-

- It encompasses a broder set of techniques & methodologies that leverage data tools to drive decision making.
- It focuses on using data insights to solve complex business problems & generate value.
- It involves following components:-

1) Descriptive Analysis:-
- Understanding historical & providing insights.

2) Predictive Analysis:-
- It uses statistical models and machine learnig to forecast outcomes.

3) Prescriptive Analytics:-
- It recommends actions based on predictive insights.
- Prescriptive analysis leverages optimization techniques and decision models to provide guidance on best course.

Q4) Describe BI and human resource management?

Ans Business Intelligence and Human Resource Management (HRM) are two distinct areas that can be integrated to enhance HRM processes & decision-making within an organization.

Here's how BI can be applied to HRM:-

1) HR Data Analysis:-
- BI tools can be used to analyze HR data, such as employee demographics, performance evaluations, training records.
- By examining this data, HR professionals can identify trends, patterns and correlations.

2) HR Metrics & KPIs:-
- BI allows the measurement and tracking of HR metrics and key performance indicators.
- Metrics like employee turnover rates, time-to-hire, absenteeism, and training effectiveness can be monitored through dashboards and reports.

3) Workforce Planning and Talent Management:-
- BI enables HR professionals to forecast workforce demand and plan for future needs.
- It helps in predicting talent gaps and deterring recruitment.
- For example:- Sir Alex Ferguson used Power BI on portugal's FC to get Cristiano Ronaldo AKA Siuuuu.

4) Employee Engagement & Satisfaction:-
- BI can help measure and track employee engagement and satisfaction levels.
- This helps HR professional identify areas of improvement

Qs) Explain BI Applications in CRM, Logistics & Production?

Ans) BI Application in CRM:-
- CRM system focus on three things:-
  - Managing Customer Interaction.
  - Sales processes.
  - Customer data.

- BI enhances CRM by providing valuable insights into customer behaviour, preferences & sales performance.

2) BI Applications in Logistics :-

- Logistics involves the management of the flow of goods, transportation, and supply chain processes.
- BI can provide valuable insights to optimize logistics operations.
- Applications of BI in logistics are as follows:-
  - Supply Chain Visibility.
  - Performance Monitoring.
  - Warehouse Optimization.
  - Risk Management.

3) BI Applications in Production :-

- Production involves the manufacturing and production processes within an organization.
- BI can provide insights to optimize production efficiency, quality control and resource utilization.
- Applications in productions are as follows:-
  - Production performance Monitoring.
  - Quality Control.
  - Supply & Demand Planning.
  - Cost Analysis.

By leveraging BI applications in CRM, logistics, and production, organizations can make data-driven decisions, optimize operations, improve customer satisfaction and drive overall business performance.