EXPERIMENT NO.2

AIM: Experiment to identify Business aspects for a identified domain and perform analysis for the same.

PROBLEM STATEMENT:
Develop a comprehensive database schema optimized for financial management and analysis. The schema should facilitate the detailed tracking and analysis of financial transactions, budgets, and categories, providing insights for various financial activities.

DATA-SET DESCRIPTION:
https://www.kaggle.com/datasets/census/business-and-industry-reports

This dataset contains information about the country, age, salary, and purchased status of customers who were surveyed by a company. The country column indicates the nationality of the customer, the age column shows the age of the customer in years, the salary column shows the annual income of the customer in US dollars, and the purchased column shows whether the customer bought the product offered by the company or not. The purchased column has binary values: Yes or No.
This dataset is useful for understanding the demographic and behavioral characteristics of the customers, and how they relate to their purchasing decisions. The dataset can be used for exploratory data analysis, data visualization, data cleaning, and machine learning tasks such as classification or clustering. The dataset can help the company to identify the target market, segment customers, and optimize marketing strategies.

Attributes(Dependent variable / Independent variable / Control variable) :

Age: Describes age of user.
Salary: Describes the salary.
Country: Describes the country for where the data is used and based on location of companies.

IEEE PAPERS:
https://ieeexplore.ieee.org/document/7807072
https://ieeexplore.ieee.org/abstract/document/8858722

● Dependent variable (DV) : A dependent variable is the outcome or response being studied in a research project. It is the variable that researchers are interested in understanding or predicting based on changes in other variables.
● Independent variable (IV) : An independent variable is a factor that researchers manipulate or control in an experiment or study. It is the variable that is hypothesized to have an effect on the dependent variable.
● Control variable (CV) : A control variable is a factor that is kept constant or held consistent in an experiment or study. It is used to minimize the potential influence of extraneous factors on the relationship between the independent and dependent variables. Control variables are not the focus of the study but are included to ensure that any observed effects can be attributed to the independent variable being investigated.

The set of attributes and dependent variables for the dataset about country, age, salary, and purchased status of customers:

Country: IV
Age: IV
Salary: IV
Purchased: DV

The dependent variable in this dataset is purchased, which indicates whether the customer bought the product or not. The independent variables are country, age, and salary, which are the factors that may influence the purchasing decision of the customer. There are no control variables in this dataset, as all the variables are either dependent or independent

Types of attributes (Nominal/ Discrete/ Ordinal/ Continuous):
● **Nominal:**
Description: Nominal attributes are categorical variables representing different categories or groups without any inherent order or ranking.
● **Ordinal:**
Description: Ordinal attributes are categorical variables with a natural order or ranking, but the differences between categories may not be uniformly quantifiable.
● **Continuous:**
Description: Continuous attributes are numerical variables that can take any real value within a certain range. They are measured and can have an infinite number of possible values within an interval.

- **Discrete:**

Description: Discrete attributes are numerical variables that can only take
on specific, distinct values, typically whole numbers. They represent
counts or quantities that cannot be subdivided further.

FOR OUR DATASET:

The types of attribute distribution for your dataset from
(normal/discrete/ordinal/continuous) are as follows:

- Country: This is a discrete attribute, as it has a finite number of possible values
  that are not ordered. The distribution of this attribute can be described using the
  mode, which is the most frequent category, or the relative frequency, which is the
  proportion of each category in the dataset.
- Age: This is a continuous attribute, as it has an infinite number of possible values
  that are ordered. The distribution of this attribute can be described using the mean,
  which is the average age of the customers, the median, which is the middle age of
  the customers, the standard deviation, which measures the variability of the ages,
  or the skewness, which measures the symmetry of the distribution. The
  distribution of this attribute can be visualized using a histogram or a box plot,
  which show the shape and spread of the data.
- Salary: This is also a continuous attribute, as it has an infinite number of possible
  values that are ordered. The distribution of this attribute can be described and
  visualized using the same methods as the age attribute.
- Purchased: This is a discrete attribute, as it has a finite number of possible values
  that are not ordered. The distribution of this attribute can be described and
  visualized using the same methods as the country attribute.

METHODOLOGY AND ALGORITHMS FOR ANALYSIS:

 the three data mining approaches (DT, clustering, and NN). The algorithms are
implemented and tested using SSAS package. SSAS provides a part to perform different
data mining approaches. All data mining approaches help the decision-maker in
supporting decision on the basis of the obtained results. The sales manager can use this
basis to predict the sales rate, sales items, and quantities by using the analytical graphs
and results obtained from the DT, clustering, and NN approaches. The results of the
graphs and figures from data mining approaches are combined with the reports and KPIs
in one dashboard

1. Decision Tree (DT) processing in financial data analysis helps in:
Classification: Sorting financial data into different groups.
Prediction: Forecasting future trends or outcomes based on historical data.
Feature Importance: Identifying the most influential factors in financial outcomes.
Risk Assessment: Evaluating and managing financial risks.
Portfolio Management: Analyzing investment options and optimizing portfolios.
Fraud Detection: Identifying abnormal patterns in financial transactions.
Customer Segmentation: Grouping customers based on behavior or preferences for targeted strategies.

2. **the K-Nearest Neighbors (KNN)** algorithm in financial data analysis using the SSAS:
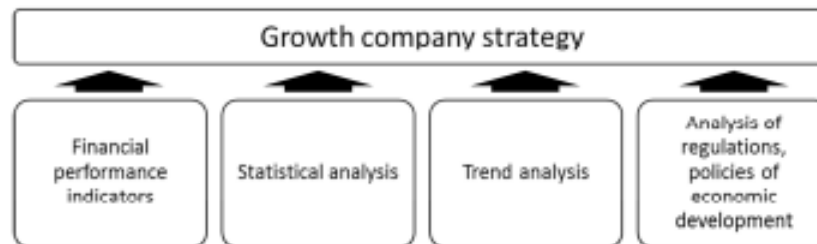**Pattern Recognization:**Recognizes patterns and classifies data based on similarity.
Anomaly Detection: Detects anomalies, aiding in fraud detection.
Regression: Predicts continuous variables like stock prices.
Portfolio Management :Helps in portfolio management and customer segmentation.
Market Analysis: Analyzes markets for informed decision-making..

ANALYSIS:
The analysis of data involves various stages and methods. The data preprocessing stage includes importing sales data from an operational system, performing data cleansing, and transforming the data by deriving new columns such as date dimensions and address dimensions . The data is then stored in a staging table for further transformation processes as part of the ETL process .

The model-driven decision support system (DSS) framework methodology consists of five stages: data pre-processing, ETL, building sales cube and KPIs, building reports, and deploying the sales DSS dashboard . The DSS utilizes a knowledge-driven approach based on three data mining algorithms: Decision Trees, Neural Networks, and Clustering .

The OLAP operations in the DSS allow analysts and managers to view sales information from different perspectives, while predefined KPIs automatically compare and indicate critical situations . The system is configured to perform online ETL, data mining, and KPI analysis using fresh data, with scheduled tasks to handle extraction, transformation, and loading processes .

Overall, the data analysis here, involves a comprehensive approach that integrates data preprocessing, ETL processes, data mining algorithms, OLAP operations, and KPI analysis to support strategic decision-making in sales management

INFERENCES AND EFFECTS:

Analyzing the provided business survey data reveals valuable insights for optimizing marketing strategies and enhancing customer targeting. By examining demographics such as age, salary, and purchase decisions across different countries, businesses can tailor their approaches to consumer segments more effectively. Machine learning techniques can be applied to predict future purchasing behavior, allowing for better sales forecasting and inventory management. Understanding geographical trends and customer profiles enables businesses to personalize marketing campaigns and product offerings, ultimately improving customer satisfaction and loyalty. Moreover, insights gained from the analysis can inform pricing strategies, product development, and overall business decision-making, leading to increased sales revenue and market competitiveness.

REFERENCES:

1.Narayan, R.; and Meht, G. (2020). Design of customer information management system. Data Communication and Networks, 1049, 195-216.

2. Tria, F.D.; Lefons, E.; and Tangorra, F. (2017). Evaluation of data warehouse design methodologies in the context of big data. Proceedings of the International Conference on Big Data Analytics and Knowledge Discovery. Lyon, France.

3. Kasie, F.M.; Bright, G.; and Walker, A. (2017). Decision support systems in manufacturing: a survey and future trends. Journal of Modelling in Management, 12(3), 432-454.

4. Wagner, W.P. (2017). Trends in expert system development: A longitudinal content analysis of over thirty years of expert system case studies. Expert Systems with Applications, 76, 85-96