Smart Prediction of Engineering Streams Based on Student Counselling Data :A KDD Approach

Project Report

Submitted to the Faculty of Engineering of

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY KAKINADA, KAKINADA

In partial fulfillment of the requirements for the award of the Degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

Ch. Surya Vamsi (22481A0542)

G. Revathi (22481A0557)

D. Sarika (22481A0548)

A. Veladri (22481A0512)

Under the Enviable and Esteemed Guidance of

Dr. M. Babu Rao, M.Tech ,Ph.D.

Assistant Professor of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)

SESHADRIRAO KNOWLEDGE VILLAGE

GUDLAVALLERU – 521356

ANDHRA PRADESH

2024-25

SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)SESHADRI RAO KNOWLEDGE VILLAGE, GUDLAVALLERU

DEPARTMENT OF

COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the project report entitled **"Smart Prediction of Engineering Streams Based on Student Counselling Data: A KDD APPROACH"** is a bonafide record of work carried out by **Ch. Surya Vamsi (22481A0542) ,G. Revathi (22481A0557) ,D. Sarika (22481A0548) ,A. Veladri (22481A0512)** under the guidance and supervision of **Dr. M. Babu Rao, M.Tech ,Ph.D. ,Assistant Professor,** Computer Science and Engineering, in the partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering of Jawaharlal Nehru Technological University Kakinada, Kakinada during the academic year 2024-25.

Project Guide
(Dr. M. BABU RAO)

Head of the Department
(Dr. M. BABU RAO)

External Examiner

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people who made it possible and whose constantguidance and encouragements crown all the efforts with success.

We would like to express our deep sense of gratitude and sincere thanks to **Dr. M. Babu Rao, M.Tech ,Ph.D., Assistant Professor,** Computer Science and Engineering for his constant guidance, supervision and motivation in completing the project work.

We feel elated to express our floral gratitude and sincere thanks to **Dr. M. Babu Rao**, **M.Tech, Ph.D., Head of the Department**, Computer Science and Engineering for his encouragements all the way during analysis of the project. His annotations, insinuations and criticisms are the key behind the successful completion of the project work.

We would like to thank our beloved principal **Dr. B. KARUNA KUMAR, M.Tech, Ph.D.,** for providing a great support for us in completing our project and giving us the opportunity for doing project.

Our Special thanks to the faculty of our department and programmers of our computer lab. Finally,we thank our family members, non-teaching staff and our friends, who had directly or indirectly helped and supported us in completing our project in time .

Team Members

Ch. Surya Vamsi (22481A0542)

G. Revathi (22481A0557)

D. Sarika (22481A0548)

A. Veladri (22481A0512)

INDEX

CONTENTS	PAGE NO
Abstract	1
PART A: KDD PROCESS	2-26
Chapter 1: Introduction	2-7
1.1 Introduction to KDD	
1.2 Data Warehousing	
1.3 Data Mining	
Chapter 2: Data Mining and Warehousing Process on Collected Dataset	8-22
2.1 Problem Statement	
2.2 Methodology	
Chapter 3: Experimental Analysis	23-26
3.1 Evaluation	
3.2 Conclusion	
PART B: DATA MINING IN DETAIL	27-41
Chapter 1: Introduction on data mining methodology	27-28
1.1 Problem Statement	
1.2 Identification of appropriate methodology	
Chapter 2: Analysis on Dataset	29-30
Chapter 3: Working on Dataset	31-36
Chapter 4: Experimental Analysis	37-41
PART C: FINAL ANALYSIS	42-43
Evaluation of Experimental Analysis	42

Conclusion 43

References 44

List of Program Outcomes and Program Specific Outcomes 45-46

Mapping of Program Outcomes with graduated POs and PSOs 47

ABSTRACT

In the rapidly evolving field of engineering education, the classification of aspirants based on demographic and academic attributes is vital for optimizing branch allocation and refining academic guidance systems. However, traditional manual analysis of student profiles is labor-intensive and often fails to reveal underlying selection patterns. This project seeks to develop a machine learning model to categorize students into engineering branches (e.g., CSE, ECE, ME) using attributes such as rank, category, gender, and locality, aiming to promote equitable allocation, streamline counseling processes, and enhance institutional planning.

The implementation leverages real-world datasets to explore data preprocessing, feature engineering, and model training, utilizing techniques such as decision trees, support vector machines, and k-nearest neighbors. The study focuses on improving model accuracy through feature selection and validation, while employing visualization tools to interpret classification outcomes effectively. Performance metrics like accuracy, precision, and recall are employed to evaluate model efficacy, ensuring reliable predictions.

By integrating data-driven insights with practical application, this project highlights the transformative impact of machine learning on educational decision-making. The findings aim to increase transparency, boost student satisfaction, and support strategic planning in academic institutions, demonstrating the potential of predictive analytics to address challenges in engineering branch allocation and beyond.

PART A: SMART PREDICTION OF ENGINEERING STREAMS BASED ON STUDENT COUNSELLING DATA USING KDD PROCESS

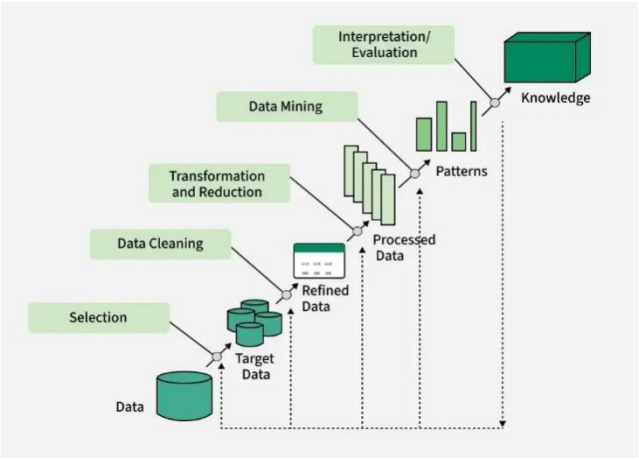
CHAPTER 1: INTRODUCTION

1.1 INTRODUCTION TO KDD

Knowledge Discovery in Databases (KDD) refers to the complete process of uncovering valuable knowledge from large datasets. It starts with the selection of relevant data, followed by preprocessing to clean and organize it, transformation to prepare it for analysis, data mining to uncover patterns and relationships, and concludes with the evaluation and interpretation of results, ultimately producing valuable knowledge or insights. KDD is widely utilized in fields like machine learning, pattern recognition, statistics, artificial intelligence, and data visualization.

The KDD process is iterative, involving repeated refinements to ensure the accuracy and reliability of the knowledge extracted. The whole process consists of the following steps:

1. Data Selection
2. Data Cleaning and Preprocessing
3. Data Transformation and Reduction
4. Data Mining
5. Evaluation and Interpretation of Results



Data Selection

Data Selection is the initial step in the Knowledge Discovery in Databases (KDD) process, where relevant data is identified and chosen for analysis. It involves selecting a dataset or focusing on specific variables, samples, or subsets of data that will be used to extract meaningful insights.

- It ensures that only the most relevant data is used for analysis, improving efficiency and accuracy.
- It involves selecting the entire dataset or narrowing it down to particular features or subsets based on the task's goals.
- Data is selected after thoroughly understanding the application domain.

By carefully selecting data, we ensure that the KDD process delivers accurate, relevant, and actionable insights.

Data Cleaning

In the KDD process, Data Cleaning is essential for ensuring that the dataset is accurate and reliable by correcting errors, handling missing values, removing duplicates, and addressing noisy or outlier data.

- **Missing Values:** Gaps in data are filled with the mean or most probable value to maintain dataset completeness.

- **Noisy Data:** Noise is reduced using techniques like binning, regression, or clustering to smooth or group the data.
- **Removing Duplicates:** Duplicate records are removed to maintain consistency and avoid errors in analysis.

Data cleaning is crucial in KDD to enhance the quality of the data and improve the effectiveness of data mining.

Data Transformation and Reduction

Data Transformation in KDD involves converting data into a format that is more suitable for analysis.

- **Normalization**: Scaling data to a common range for consistency across variables.
- **Discretization**: Converting continuous data into discrete categories for simpler analysis.
- **Data Aggregation**: Summarizing multiple data points (e.g., averages or totals) to simplify analysis.
- **Concept Hierarchy Generation**: Organizing data into hierarchies for a clearer, higher-level view.

Data Reduction helps simplify the dataset while preserving key information.

- **Dimensionality Reduction** (e.g., PCA): Reducing the number of variables while keeping essential data.
- **Numerosity Reduction**: Reducing data points using methods like sampling to maintain critical patterns.
- **Data Compression**: Compacting data for easier storage and processing.

Together, these techniques ensure that the data is ready for deeper analysis and mining.

Data Mining

Data Mining is the process of discovering valuable, previously unknown patterns from large datasets through automatic or semi-automatic means. It involves exploring vast amounts of data to extract useful information that can drive decision-making.

Key characteristics of data mining patterns include:

- **Validity**: Patterns that hold true even with new data.
- **Novelty**: Insights that are non-obvious and surprising.
- **Usefulness**: Information that can be acted upon for practical outcomes.
- **Understandability**: Patterns that are interpretable and meaningful to humans.

In the KDD process, choosing the data mining task is critical. Depending on the objective, the task could involve classification, regression, clustering, or association rule mining. After determining the task, selecting the appropriate data mining algorithms is essential. These algorithms are chosen based on their ability to efficiently and accurately identify patterns that align with the goals of the analysis.

Evaluation and Interpretation of Results

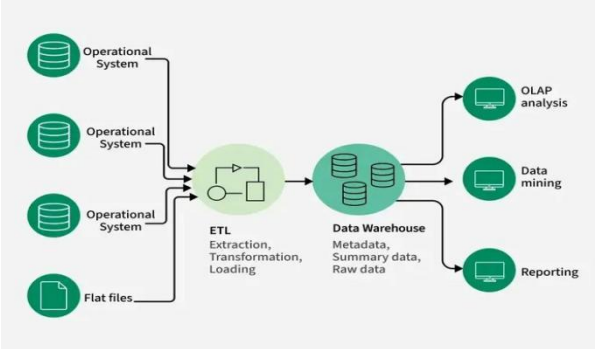
Evaluation in KDD involves assessing the patterns identified during data mining to determine their relevance and usefulness. It includes calculating the "interestingness score" for each pattern, which helps to identify valuable insights. Visualization and summarization techniques are then applied to make the data more understandable and accessible for the user.

Interpretation of Results focuses on presenting these insights in a way that is meaningful and actionable. By effectively communicating the findings, decision-makers can use the results to drive informed actions and strategies.

1.2 DATA WAREHOSUING

A data warehouse is a centralized system used for storing and managing large volumes of data from various sources. It is designed to help businesses analyze historical data and make informed decisions. Data from different operational systems is collected, cleaned, and stored in a structured way, enabling efficient querying and reporting.

- Goal is to produce statistical results that may help in decision-making.
- Ensures fast data retrieval even with the vast datasets.



Need for Data Warehousing

1. Handling Large Volumes of Data: Traditional databases can only store a limited amount of data (MBs to GBs), whereas a data warehouse is designed to handle much larger datasets (TBs), allowing businesses to store and manage massive amounts of historical data**.**

2. Enhanced Analytics: Transactional databases are not optimized for analytical purposes. A data warehouse is built specifically for data analysis, enabling businesses to perform complex queries and gain insights from historical data.

3. Centralized Data Storage: A data warehouse acts as a central repository for all organizational data, helping businesses to integrate data from multiple sources and have a unified view of their operations for better decision-making.

4. Trend Analysis: By storing historical data, a data warehouse allows businesses to analyze trends over time, enabling them to make strategic decisions based on past performance and predict future outcomes.

5. Support for Business Intelligence: Data warehouses support business intelligence tools and reporting systems, providing decision-makers with easy access to critical information, which enhances operational efficiency and supports data-driven strategies.

Components of Data Warehouse

The main components of a data warehouse include:

- Data Sources: These are the various operational systems, databases, and external data feeds that provide raw data to be stored in the warehouse.
- ETL (Extract, Transform, Load) Process: The ETL process is responsible for extracting data from different sources, transforming it into a suitable format, and loading it into the data warehouse.
- Data Warehouse Database: This is the central repository where cleaned and transformed data is stored. It is typically organized in a multidimensional format for efficient querying and reporting.
- Metadata: Metadata describes the structure, source, and usage of data within the warehouse, making it easier for users and systems to understand and work with the data.
- Data Marts: These are smaller, more focused data repositories derived from the data warehouse, designed to meet the needs of specific business departments or functions.
- OLAP (Online Analytical Processing) Tools: OLAP tools allow users to analyze data in multiple dimensions, providing deeper insights and supporting complex analytical queries.
- End-User Access Tools: These are reporting and analysis tools, such as dashboards or Business Intelligence (BI) tools, that enable business users to query the data warehouse and generate reports.

1.3 DATA MINING

Data mining is a process of discovering patterns and knowledge from large amounts of data, utilizing sources such as databases, data warehouses, the internet, and other data repositories. It combines techniques from statistics, artificial intelligence, and machine learning to analyze large datasets and extract meaningful information. This analysis helps identify trends, correlations, and patterns that are not immediately obvious, enabling informed decision-making and predictions.

One of the key breakthroughs in data mining is its ability to handle and analyze big data efficiently. With the increasing volume, velocity, and variety of data, traditional methods are often insufficient. Data mining

techniques like clustering, classification, regression, and association rule learning are essential for extracting valuable insights from complex datasets quickly and accurately.
Data mining is closely related to machine learning and data analytics. While data mining focuses on discovering new patterns within large datasets, machine learning involves developing algorithms that can learn from and make predictions on data. These fields complement each other, enhancing data analysis and predictive modeling capabilities.

Data Mining Block Diagram

The data mining block diagram starts with data understanding, where the data is collected and analyzed to grasp its structure and content. Next, data preparation involves cleaning and transforming the data for better analysis. In the modeling phase, various algorithms are applied to build predictive models. The evaluation phase assesses the models' performance, and finally, deployment integrates the chosenmodel into practical applications for decision-making.

Supervised Learning

Supervised learning is a machine learning technique where models are trained on labeled data. In this project, the model learns to based on user attributes. Common
algorithms used include:

- **K-Nearest Neighbors (KNN)**
- **Decision Trees**
- **Random Forest**
- **Navie Bayes**
- **Support Vector Machine**

Categories of Supervised Learning in This Project:

1. Classification:

- The dataset contains categorical labels (e.g., Name,Gender,Rank .etc).
- Classification algorithms predict a student's career outcome based on factors such as academic background, internships, technical skills, and job search activities..

Algorithm	Description	Type
SVM	Support Vector Machine (SVM) is a supervised learning algorithm used for classification and regression tasks by finding the optimal hyperplane that best separates data points into different categories. It is effective in high-dimensional spaces and is widely used in image recognition, text classification, and bioinformatics.	Classification and Regression

Decision Tree	Highly interpretable classification or regression model that splits data-feature values into branches at decision nodes.	Classification
Naïve Bayes	The Bayesian method is a classification method that makes use of the Bayesian theorem. The theorem updates the prior knowledge of an event with the independent probability of each feature that can affect the event.	Regression and Classification
KNN	K-Nearest Neighbors (KNN) is a supervised learning algorithm that classifies data points based on the labels of their nearest neighbors in the feature space. It assigns the most common label among the closest data points to the new data point.	Regression and Classification
Random Forest	Random Forest is an ensemble learning algorithm that builds multiple decision trees and combines their outputs for robust and accurate classification or regression.	Regression and Classification

➢ **UNSUPERVISED LEARNING**

Unsupervised learning is a type of machine learning where the model is trained on unlabeled data, meaning there are no predefined output labels. The goal is to discover hidden patterns or intrinsic structures within the data. Common techniques include clustering (e.g., K-Means) and association rule learning. This approach is useful for tasks like customer segmentation and anomaly detection.

There are two categories of Unsupervised Learning. They are

1.Clustering

2.Association

1.Clustering:

clustering serves as a vital technique in unsupervised learning within data mining. It involves grouping similar data points together into clusters based on their intrinsic characteristics, without predefined labels. Algorithms like K-Means and Hierarchical Clustering help us uncover hidden patterns within our dataset of lens-related attributes. By applying clustering, we aim to identify distinct groups of individuals with similar visual characteristics, facilitating personalized recommendations for lens suitability. This unsupervised approach aids in data exploration and segmentation, providing insights into diverse needs and preferences among individuals. Overall, clustering plays a crucial role in uncovering meaningful patterns and guiding data-driven decision-making in lens recommendation strategies.

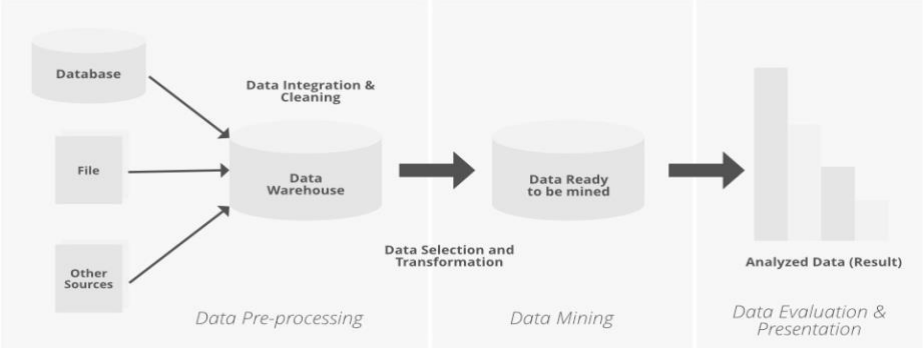
2.Association:

Association analysis is a core technique in unsupervised learning within data mining, aimed at discovering relationships among different attributes or items in a dataset. Algorithms like Apriori andFP- Growth enable us to identify frequent itemsets and association rules within our dataset of lens related attributes. By applying association analysis, we aim to uncover associations between visual characteristics such as age, prescription, tear production rate, and astigmatism status, and the types of

lenses recommended. Additionally, association analysis helps identify relevant features for lens suitability, contributing to the refinement of our predictive models.

How to choose Data Mining Algorithm

Choosing a data mining algorithm depends on the nature of your data and the problem you aim to solve. For labeled data and predictive tasks, supervised learning algorithms like Decision Trees or Logistic Regression are suitable.



Data Mining Basic Diagram

Challenges and Limitations of Data Mining

One significant challenge in data mining is the issue of data quality and preprocessing. Often, real- world datasets are noisy, incomplete, or contain inconsistencies, which can significantly impact the effectiveness of data mining algorithms. Preprocessing tasks such as data cleaning, normalization, and feature selection are crucial for improving the quality of the data and ensuring accurate and reliable results. However, these tasks can be time-consuming and resource-intensive, especially for large and complex datasets. Moreover, even with careful preprocessing, there may still be underlying biases or limitations in the data that can affect the performance and generalization ability of the models. Therefore, addressing data quality and preprocessing challenges remains a critical aspect of successfuldata mining projects.

Applications of Data Mining

1. **Customer Relationship Management (CRM):** Data mining is a cornerstone in Customer Relationship Management (CRM), enabling businesses to delve deep into customer data for actionable insights. By scrutinizing diverse aspects such as demographics, purchase history, and behavioral patterns, companies can discern trends and preferences. This analysis facilitates the identification of high-value customers, prediction of churn rates, and crafting personalized marketing strategies. Leveraging data mining in CRM notonly enhances customer satisfaction but also fosters long-term loyalty and retention. Through targeted campaigns and tailored offerings, businesses can nurture stronger relationships with customers, ultimately driving growth and profitability.
2. **Fraud Detection:** Data mining techniques are instrumental in fraud detection systems, deployed across sectors like banking, insurance, and e-commerce. By scrutinizing transactional data and user behavior, algorithms can swiftly identify irregularities or suspicious trends that may signal fraudulent activity.

Data Mining vs Data Warehousing

Data warehousing and data mining serve distinct but complementary purposes in data management. Data warehousing involves storing and organizing large volumes of data from various sources into a centralized

repository, designed to support efficient querying and reporting for business intelligence. It focuses on the ETL (Extract, Transform, Load) process to ensure data consistency and accessibility.In contrast, data mining analyzes this stored data to discover patterns, trends, and relationships using algorithms and statistical methods. The primary goal of data mining is to transform raw data into actionable insights that inform business strategies and decision-making. While data warehousing emphasizes efficient storage and access, data mining focuses on extracting meaningful knowledge from the data. Together, they enable effective data management and strategic decision-making by leveraging stored data for in-depth analysis and discovery.

SOFTWARE AND HARDWARE REQUIREMENTS:

Software Requirements

Windows 10 or above Operating System

Dataset:

• Clean and structured survey data.

• Handle missing values, duplicates, and inconsistent entries.

Software/Tools:

• Orange Tool

• SQL Server Management Studio

• VS Code

Models:

• Classification Algorithms : Neural Network, Random Forest, KNN, SVM, Naive Bayes, etc.

• Model evaluation metrics: Accuracy, Precision, Recall, F1-score, Confusion Matrix

Hardware Requirements

- **Processor:** Intel Core i5 or higher / AMD equivalent
- **RAM:** Minimum 8 GB (16 GB recommended for faster training)
- **Storage:** At least 100 GB free disk space

CHAPTER 2:DATA MINING AND WARE HOUSING PROCESS ON COLLECTED DATASET

2.1 Problem Statement:

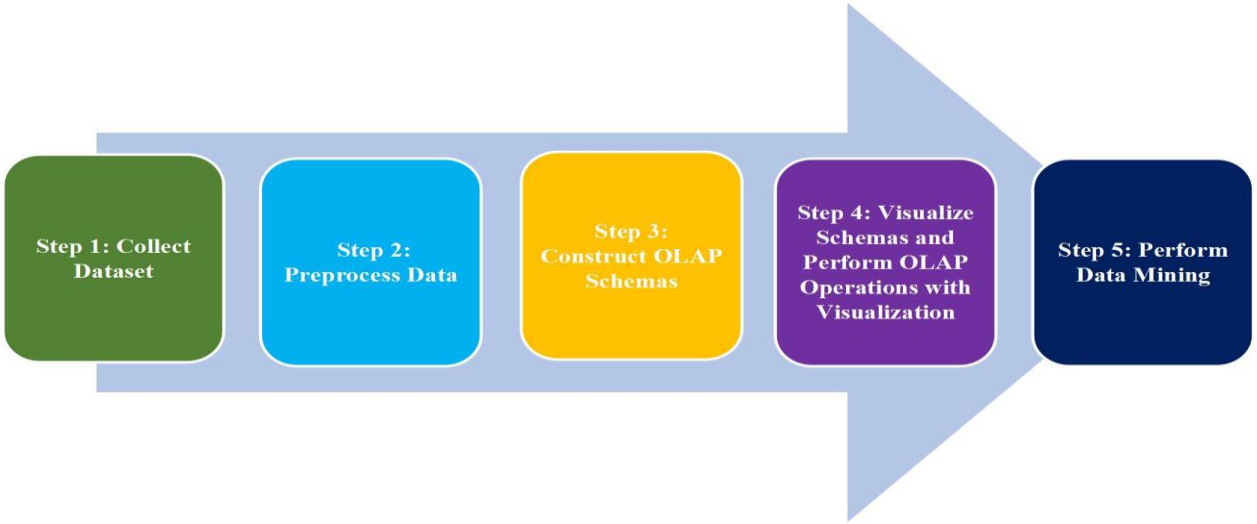
The classification of engineering aspirants based on their demographic and academic attributes is crucial for streamlining branch allocation and enhancing academic guidance systems. However, manual analysis of student profiles is time-consuming and may not accurately capture hidden patterns in selection behavior. This project aims to develop a machine learning model to classify students into different engineering branches (e.g., CSE, ECE, ME, etc.) based on their attributes such as rank, category, gender, and locality, ensuring fair allocation, improved counseling processes, and optimized institutional planning. By leveraging data-driven insights, the system will help academic institutions enhance transparency, student satisfaction, and decision-making accuracy

Objectives

- Automate the classification of engineering aspirants into appropriate branches (e.g., CSE, ECE, ME, etc.) using machine learning techniques based on their academic and demographic profiles.
- Support academic institutions in making data-driven decisions for fair and efficient branch allocation.
- Improve student satisfaction by aligning branch assignments with individual preferences, merit, and eligibility criteria.
- Enhance the transparency and efficiency of counseling and admission processes through predictive analytics.
- Enable better planning and resource allocation within institutions by understanding trends in student branch preferences.

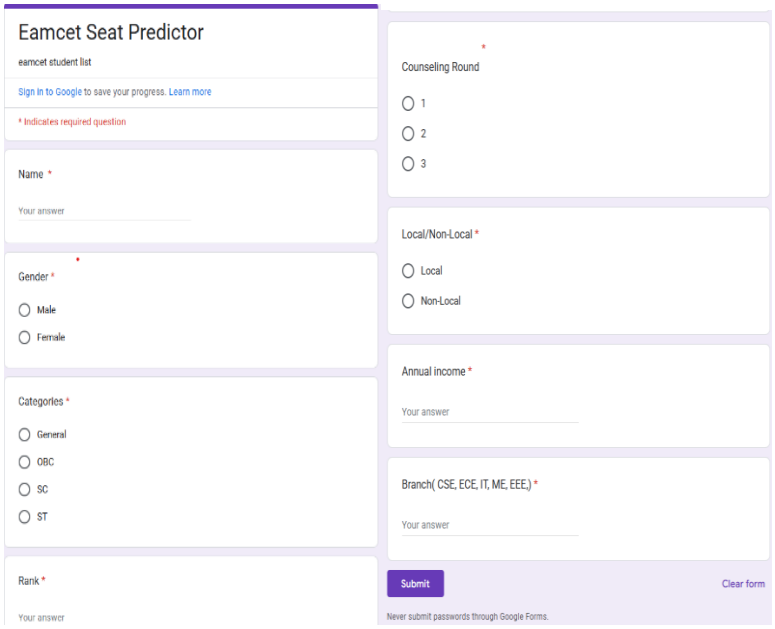
2.2 Methodology:

The KDD process is performed in step by step from collection of data set to the classification and developing the prediction model. There are some intermediatory steps in which we created all three schemas with the help of various tools like SSMS (SQL Server Management Services), Visual Studio and SSAS (SQL Server Analysis Services). The process is explained in step by step below.



STEP-1: COLLECTING & EXPLORING DATASET

- Dataset is collected Using Google Forms
- Designed Google Form – with relevant questions (e.g., name, gender, categories,rank,etc.)



- Shared it with participants through email, social media, or targeted groups.
  Link to the Google form: [Eamcet Seat Predictor]

- Collected Responses – Monitored responses and ensure enough data is gathered.

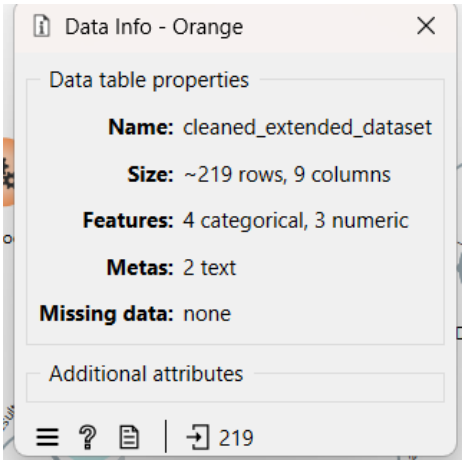
- Exported Data – Downloaded the responses as a CSV file for further processing.

Name	Gender	Categories	Rank	Counseling	Local/Non	Annual inc	Branch(CSE, ECE, IT, ME, EEE,)
Jaswanth	Male	OBC	20000	1	Non-Local	90000	CSE
Dasari Sari	Female	OBC	1	1	Non-Local	500000	CSE
Alla Veladr	Male	General	13000	1	Local	25000	CSE
Lalitha	Female	OBC	13000	2	Local	60000	CSE
Pandu nay	Male	ST	80000	1	Non-Local	80000	CSE
Kurshid be	Female	ST	14000	1	Non-Local	40000	CSE
Vijaya sri	Female	ST	40000	1	Non-Local	85000	ECE
Dustin Har	Male	General	1200	1	Local	100000	IT
Luffy	Male	General	1200	2	Non-Local	150000	ME
Roronoa Z	Male	General	1400	1	Local	200000	CSE
Bhanu sri	Female	General	4836	2	Local	120000	IT
Ulisi.Manil	Male	General	80000	2	Non-Local	100000	CSE
Saranya	Female	General	11706	2	Non-Local	100000	CSE
Revathi	Female	General	10000	1	Local	1000000	CSE
Devaki	Female	General	1000	1	Local	2000000	CSE
Ulisi.Manil	Male	General	80000	2	Non-Local	100000	CSE
Renu	Female	OBC	468967	2	Non-Local	356785	ECE
ulisi.manir	Male	General	80000	2	Non-Local	100000	CSE
Sunny	Male	SC	43678	3	Non-Local	5435799	IT
Uliai.Mani	Male	General	80000	2	Non-Local	100000	CSE
Bunny	Male	ST	5357	2	Non-Local	536805	ME
Ulisi.Manil	Male	General	80000	2	Non-Local	100000	CSE
Chinni	Female	SC	4677	1	Non-Local	467546	ECE
Girija kum	Female	OBC	23456	3	Non-Local	35263	ME
Ganga	Female	General	46787	2	Local	579066	CSE

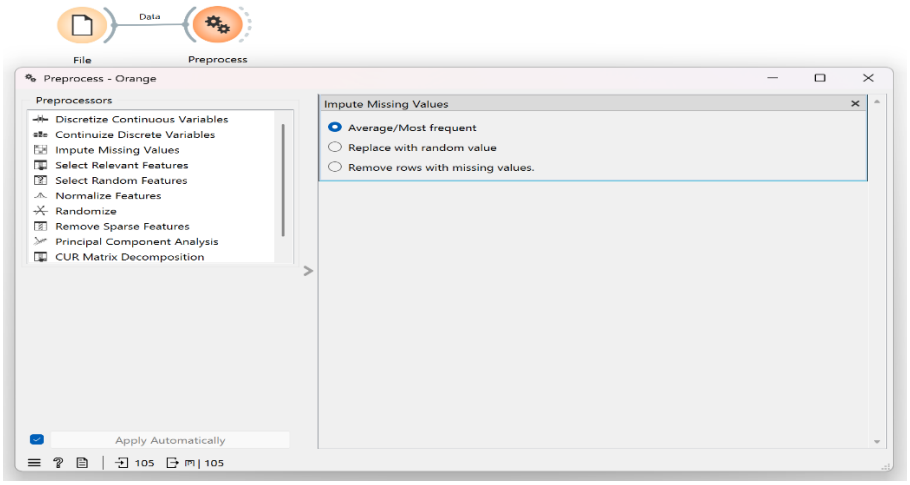
The attributes are:
1.Name
2.gender
3.categories
4.rank
5.counseling round
6.local/non-local
7.annual income
8.branch

Step 2: PREPROCESS THE DATA

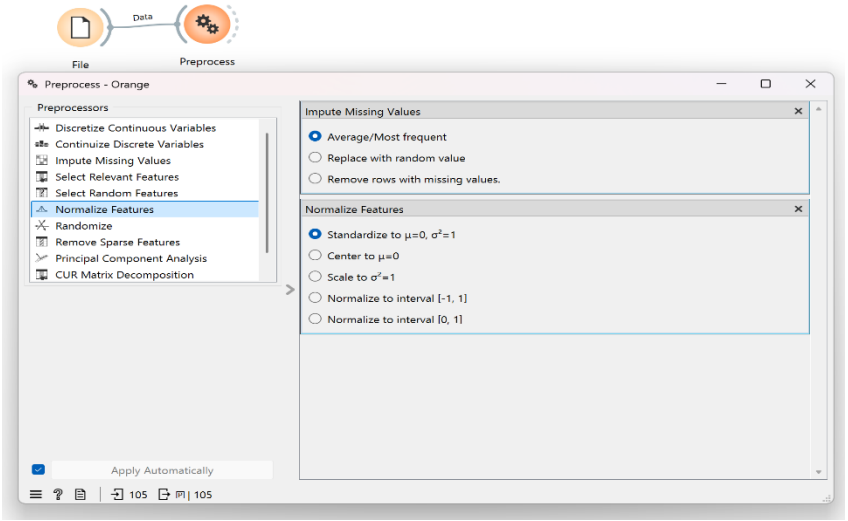
- Preprocess the Dataset Using Orange Tool
  (Select the best preprocessing technique using test and score)
- Load the Dataset – Import the collected CSV file into Orange.



- Handle Missing Values – Used imputation techniques (mean, median, mode) to fill missing values.



- Normalize & Transform Data – Normalized the numerical values using various techniques (Standardize, z-score, etc..)



- Save Preprocessed Data – Export the cleaned dataset for further use.

Step 3: CREATION OF DATABASE CONSTRUCT OLAP SCHEMAS

The above table was normalized and divided into multiple tables

DIMENSION TABLES:

1.Dim_Branch

	Column Name	Data Type	Key Type
1	branch_id	int	Primary Key
2	branch_name	varchar(50)	

2. Dim_Category

	Column Name	Data Type	Key Type
1	category_id	int	Primary Key
2	category_name	varchar(20)	

3.Dim_Gender

	Column Name	Data Type	Key Type
1	gender_id	int	Primary Key
2	gender_type	varchar(10)	

4. Dim_Location

	Column Name	Data Type	Key Type
1	location_id	int	Primary Key
2	location_type	varchar(20)	

5. Dim_Location_Details

	Column Name	Data Type	Key Type
1	location_detail_id	int	Primary Key
2	location_type	varchar(20)	

6. Dim_Branch_Details

	Column Name	Data Type	Key Type
1	branch_id	int	Primary Key
2	branch_detail_id	int	Foreign Key

7. Dim_Category_Details

	Column Name	Data Type	Key Type
1	category_id	int	Primary Key
2	category_detail_id	int	Foreign Key

FACT TABLES:

8. Fact_Admission

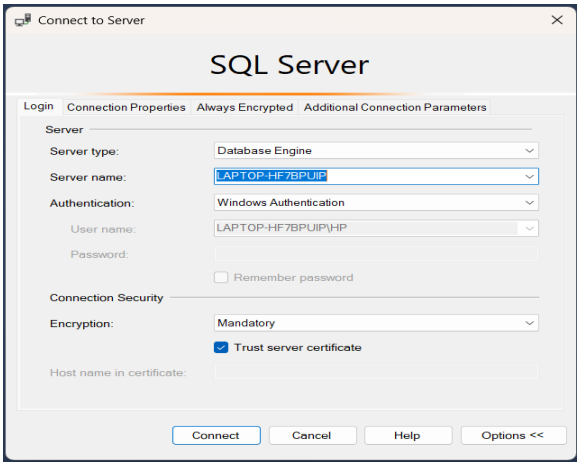
	Column Name	Data Type	Key Type
1	admission_id	int	Primary Key
2	student_name	varchar(100)	
3	branch_id	int	Foreign Key
4	gender_id	int	Foreign Key
5	category_id	int	Foreign Key
6	location_id	int	Foreign Key
7	counseling_round	int	
8	rank	float	
9	annual_income	float	

9. Fact_Student_Financials

	Column Name	Data Type	Key Type
1	financial_id	int	Primary Key
2	student_name	varchar(100)	
3	gender_id	int	Foreign Key
4	category_id	int	Foreign Key
5	location_id	int	Foreign Key
6	annual_income	float	
7	scholarship_amount	float	
8	loan_amount	float	

We have created a database Student and inserted the data into the tables.

generated sql queries to perform OLAP operations.



- Generate SQL Queries for OLAP Schema Construction

3.1 Designing the Schemas:

- ❖ Star Schema
- ❖ Snowflake Schema
- ❖ Fact Constellation Schema.

3.2 SQL Queries :

- To create Fact and Dimension tables
- Inserted data into Tables using Oracle SQL and executed them in SSMS

STEP-4: VISUALIZE SCHEMAS

- Created analysis service multidimensional project to Visual Studio.
- Defined Data Source & Data Source View – Connected the database and define table relationships.
- Generated database diagrams for schemas.
- Verified relationships between tables.
- Created Cubes & Measures – Defined fact tables, measures, and dimensions.

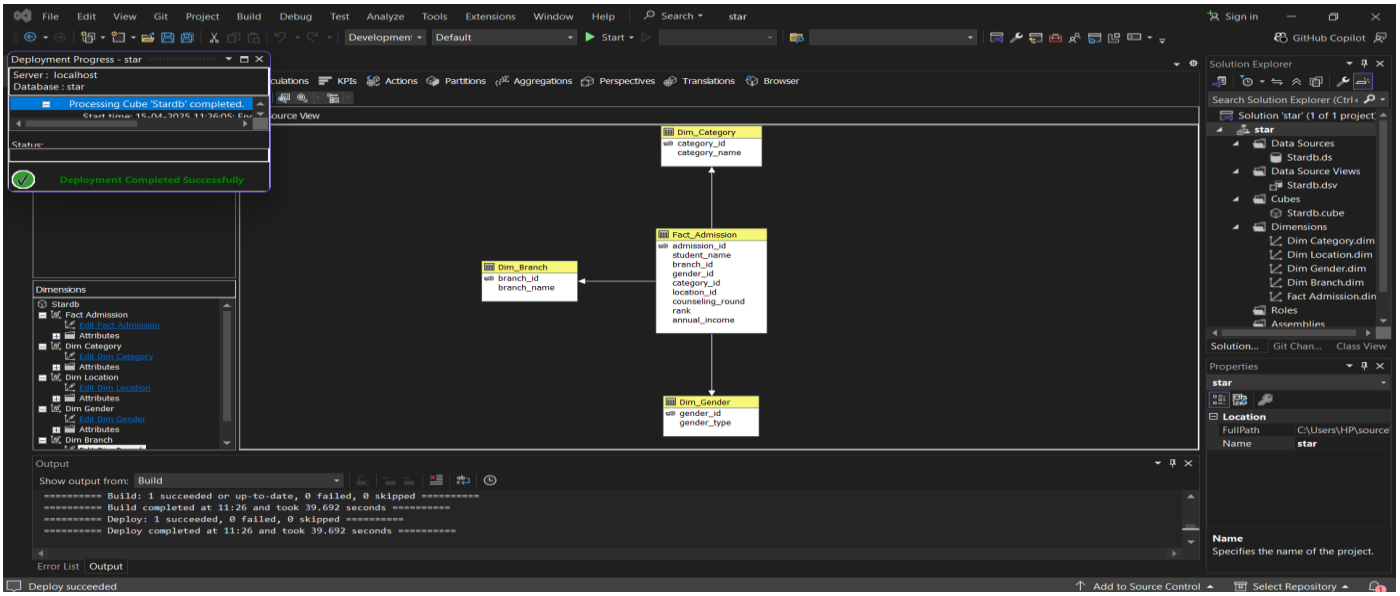
These are the schemas that we have created

4.1Star Schema:

The Star Schema is a denormalized database schema used in OLAP, where a central Fact Table (containing measurable data number of projects) is directly connected to multiple Dimension Tables (such as job search duration etc) in a star-like structure.

4.1.1 Design & Visualize the Schema

- Create the Star Schema with Fact and Dimension tables.
- Define relationships between tables for efficient querying



- Initially Build deploy and process all Multi Dimensional cubes.
- Visualize them in SSAS server
- perform OLAP operations.

4.1.2 Deploy the Data Warehouse & Load Data

- Store structured data into the data warehouse.
- Ensure ETL (Extract, Transform, Load) processes are completed.

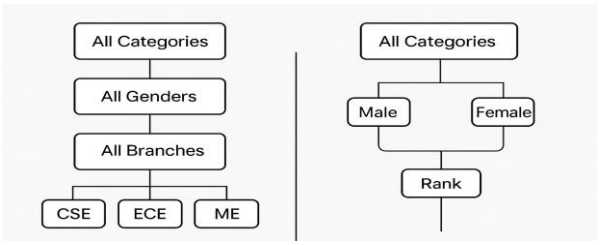
4.1.3 Create & Execute OLAP Queries

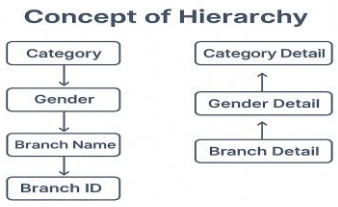
- Write OLAP queries to perform data analysis.
- Use ROLLUP,SLICE, DICE, DRILL-DOWN, and PIVOT operations for multi-dimensional analysis.

4.1.4 Perform OLAP Operations

- Run the queries to process large datasets efficiently.
- Perform aggregations, filtering, and transformations on the stored data.

Concept Hierarchies used :





MDX Queries OLAP operations in STAR SCHEMA:

(A) What is the average rank of students grouped by category and gender?

Roll-Up

```
SELECT
 NON EMPTY
  {
   ([Dim Category].[Category Name].Members) *
   ([Dim Gender].[Gender Type].Members)
  } ON ROWS,
  [Measures].[rank] ON COLUMNS
FROM [Stardb]
```

Output:

Messages Results

		Rank
All	All	262451.3
All	Female	72001
All	Male	190450.3
General	All	43450.3
General	Female	5000
General	Male	38450.3
OBC	All	60001
OBC	Female	13001
OBC	Male	47000
SC	All	25000
SC	Male	25000
ST	All	134000
ST	Female	54000
ST	Male	80000

Execution Time: 1 ms

(B) What is the average rank of students for each combination of category, gender, and branch?
Drill-Down:

```
SELECT
NON EMPTY
  {
    ([Dim Category].[Category Name].Members) *
    ([Dim Gender].[Gender Type].Members) *
    ([Dim Branch].[Branch Name].Members)
  } ON ROWS,
  [Measures].[Rank] ON COLUMNS
FROM [Stardb]
```

Output:

All	All	CSE	165451.3
All	All	ECE	65000
All	All	IT	5000
All	All	ME	27000
All	Female	All	72001
All	Female	CSE	27001
All	Female	ECE	40000
All	Female	IT	5000
All	Male	All	190450.3
All	Male	CSE	138450.3
All	Male	ECE	25000
All	Male	ME	27000
General	All	All	43450.3

Execution Time: 9 ms

(C) The total Annual Income where Gender Type is 'Male'?

Slice

```
SELECT
   {[Measures].[Annual Income]} ON COLUMNS
FROM [Stardb]
WHERE ([Dim Gender].[Gender Type].[Male])
```

Output:

Annual Income
13149500

Execution Time: 4 ms

(D) Rank and Annual Income for students in 'CSE' branch who are 'Local'?
Dice:

```
SELECT
   {[Measures].[Rank], [Measures].[Annual Income]} ON COLUMNS,
```

{[Dim Branch].[Branch Name].[CSE]} ON ROWS
FROM [Stardb]
WHERE ([Dim Location].[Location Type].[Local])

Output:

	Rank	Annual Income
CSE	51450.3	12899500

Execution Time: 3 ms

(E) The sum of Annual Income across Gender Type and Branch Name?
Pivot

SELECT
{[Dim Gender].[Gender Type].Children} ON COLUMNS,
{[Dim Branch].[Branch Name].Children} ON ROWS
FROM [Stardb]
WHERE ([Measures].[Annual Income])

Output:

	Female	Male	Unknown
CSE	600000	13009500	(null)
ECE	85000	75000	(null)
EEE	(null)	(null)	(null)
IT	150000	(null)	(null)
ME	(null)	65000	(null)
Unknown	(null)	(null)	(null)

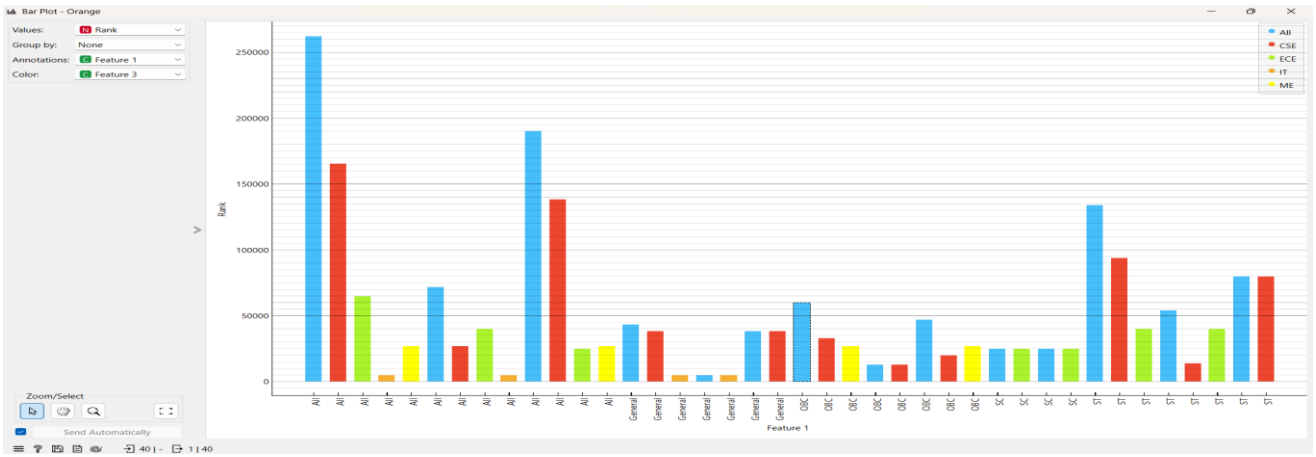
Execution Time: 3 ms

4.1.5 Visualize OLAP Results:

To analyze the distribution of students based on gender and career preferences, a **bar plot** is used. This visualization helps in identifying trends across different career choices..

Observation:

- ST and General categories show significantly higher rank values, especially in the CSE branch.
- CSE branch consistently appears with the highest ranks across all categories, suggesting it is the most in-demand or competitive.
- IT and ECE branches have moderate rank values, indicating a balanced preference or accessibility.
- SC and OBC categories display comparatively lower rank values across branches, suggesting better ranks or reservation benefits.

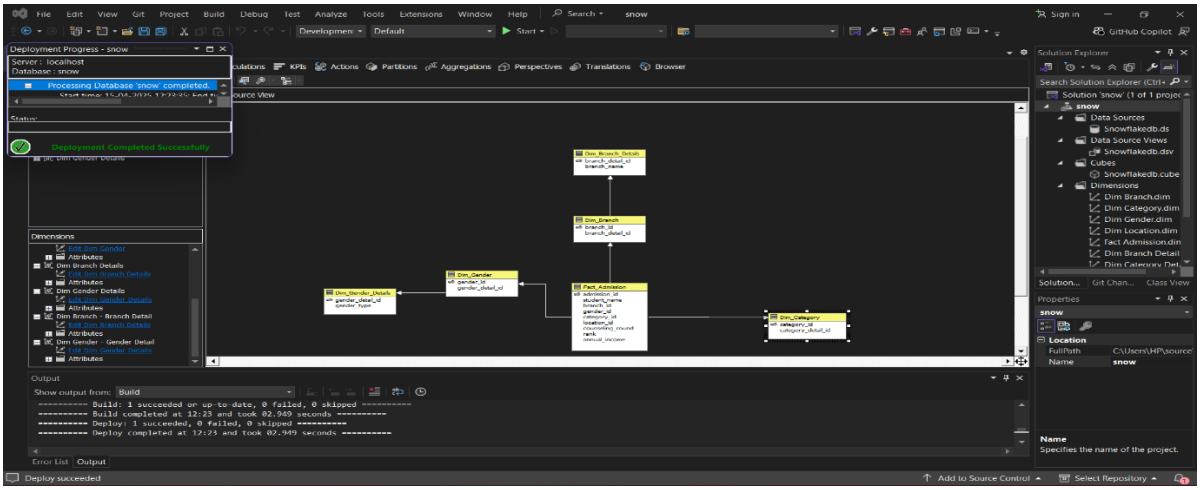


4.2SNOWFLAKE SCHEMA:

The Snowflake Schema is a normalized version of the Star Schema, where dimension tables are further divided into sub-dimensions, reducing redundancy. Below are the steps to implement it in OLAP:

4.2.1 Design & Visualize the Snowflake Schema

- Identify Fact Tables
- Identify Dimension Tables
- Normalize dimension tables by breaking them into sub-dimensions
- Ensure foreign key relationships between tables.



4.2.2 Deploy & Load Data into the Snowflake Schema

- Implement the schema in a Data Warehouse SnowFlake
- Load Fact and Dimension Tables into the database.
- Ensure proper data integrity and indexing for performance.
- Deployed the schema to the Data Warehouse.
- Configured SQL Server Analysis Services (SSAS) for OLAP processing and reporting.

MDX Queries for OLAP operations in Snow Flake Schema:

(A) What is the average rank when Gender Details are rolled up to Gender Type, considering different Categories and Branches?

Roll-up

```
SELECT
  NON EMPTY
  {
    ([Dim Category].[Category Id].[Category Id].Members) *
    ([Dim Gender].[Gender Id].[Gender Id].Members) *
    ([Dim Branch].[Branch Id].[Branch Id].Members)
  } ON ROWS,
  [Measures].[Rank] ON COLUMNS
FROM [Snowflakedb]
```

Output:

			Rank
1	1	1	38450.3
1	2	3	5000
2	1	1	20000
2	1	4	27000
2	2	1	13001
3	1	2	25000
4	1	1	80000
4	2	1	14000
4	2	2	40000

Execution Time: 3 ms

(B) How does the average rank vary when we drill down from Branch Name to Branch ID across different Categories and Genders?

Drill Down:

```
SELECT
  NON EMPTY
  {
    ([Dim Category].[Category Detail Id].[Category Detail Id].Members) *
    ([Dim Gender Details].[Gender Type].[Gender Type].Members) *
    ([Dim Branch Details].[Branch Name].[Branch Name].Members)
  } ON ROWS,
  [Measures].[Rank] ON COLUMNS
FROM [Snowflakedb]
```

Output:

			Rank
1	Female	IT	5000
1	Male	CSE	38450.3
2	Female	CSE	13001
2	Male	CSE	20000
2	Male	ME	27000
3	Male	ECE	25000
4	Female	CSE	14000
4	Female	ECE	40000
4	Male	CSE	80000

Execution Time: 5 ms

(C) The total rank for students in the "ST" category only?

Slice: Slice for Category = "ST"

Query:

```
SELECT
    [Measures].[Rank] ON COLUMNS
FROM [Snowflakedb]
WHERE ([Dim Category Details].[Category Name].[ST])
```

Output:

Execution Time: 3 ms

(D) The total rank of students from "Local" and "Non-Local" locations, filtered for "General" and "OBC" categories?

Dice: Dice for (Location Type = Local, Non-Local and Category = General, OBC and Measure = Rank)
Query:

```
SELECT
   [Measures].[Rank] ON COLUMNS,
   [Dim Location Details].[Location Type].Children ON ROWS
FROM [Snowflakedb]
WHERE ({[Dim Category Details].[Category Name].[General],
      [Dim Category Details].[Category Name].[OBC]})
```

Output:

Messages | Results

	Rank
Local	56450.3
Non-Local	47001
Unknown	(null)

Execution Time: 3 ms

(E) The total annual income with branches as rows and locations as columns?

Pivot: Interchanging rows and columns for slice operation
Query:

```
SELECT
   [Dim Location Details].[Location Type].Children ON COLUMNS,
   [Dim Branch Details].[Branch Name].Children ON ROWS
FROM [Snowflakedb]
WHERE ([Measures].[Annual Income])
```

Output:

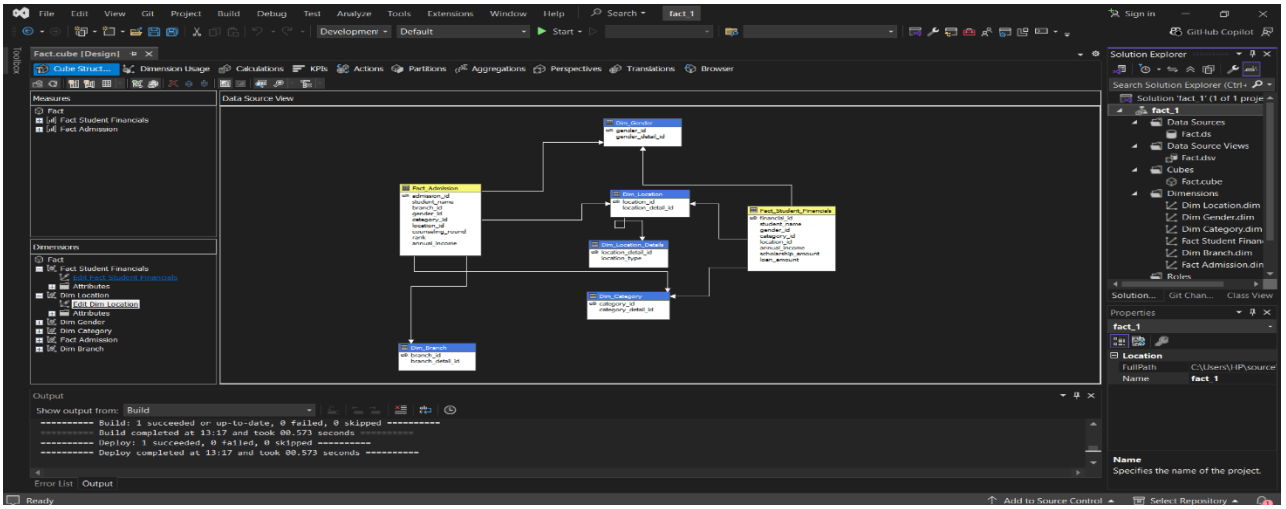
Messages | Results

	Local	Non-Local	Unknown
CSE	12899500	710000	(null)
ECE	75000	85000	(null)
EEE	(null)	(null)	(null)
IT	150000	(null)	(null)
ME	(null)	65000	(null)
Unknown	(null)	(null)	(null)

Execution Time: 3 ms

4.3FACT CONSTELLATION:

A **Fact Constellation Schema** is a complex OLAP schema where multiple fact tables share common dimension tables, allowing for more flexible analysis across different business processes. It combines multiple star schemas into a single structure, with each fact table connecting to shared dimensions.



4.3.1 Design & Visualize the FACT CONSTELLATION SCHEMA

- Multiple fact tables represent different business processes.
- Dimension tables are shared across multiple fact tables.
- Fact tables have foreign key references to common dimension tables.
- Normalized dimension tables reduce redundancy p

- Time dimension is often shared across fact tables.
- Flexible schema for handling various business processes.
- No direct relationship between fact tables; they connect through shared dimensions.

4.3.2 Deploy & Load Data into the Snowflake Schema:

1. Implement Snowflake Schema in a data warehouse.
2. Load Fact and Dimension Tables into the database.
3. Ensure data integrity and optimize performance with proper indexing.
4. Deploy schema to the data warehouse.
5. Configure SQL Server Analysis Services (SSAS) for OLAP processing and reporting

MDX Queries for OLAP operations in Fact Constellation:

(A) What is the average annual income grouped by gender?

Roll-up:

```
SELECT
  [Measures].[Annual Income] ON COLUMNS,
  [Dim Gender].[Gender Type].MEMBERS ON ROWS
FROM [Fact]
```

Output:

Messages | Results

	Annual Income
All	13984500
Female	835000
Male	13149500
Unknown	(null)

Execution Time: 22 ms

(B) Show rank of all students in the CSE branch?

Drill Down:

```
SELECT
  [Measures].[Rank] ON COLUMNS,
  [Fact Admission].[Student Name].MEMBERS ON ROWS
FROM [Fact]
WHERE ([Dim Branch].[Branch Name].[CSE])
```

Output:

	Rank
All	165451.3
Alla Veladri	38450.3
Dasari Sarika	1
Jaswanth	20000
Kurshid Begam	14000
Lalitha	13000

Execution Time: 5 ms

(C) What is the total rank of male students from non-local areas?

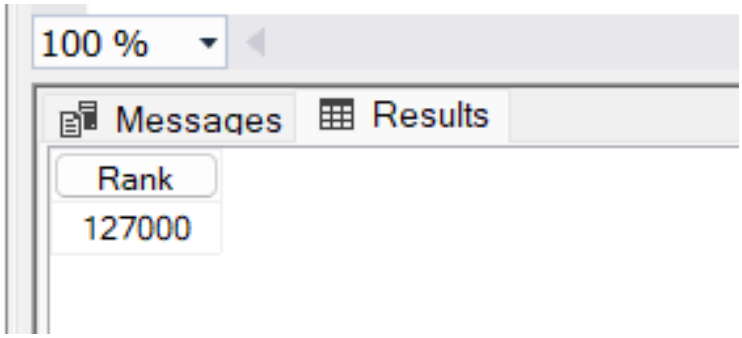
Slice: Filters the cube on one or more dimension members to analyze a single slice.

Query:

```
SELECT
  [Measures].[Rank] ON COLUMNS
FROM [Fact]
WHERE (
```

[Dim Gender].[Gender Type].[Male],
[Dim Location].[Location Type].[Non-Local]
)

Output:



Execution Time: 4 ms

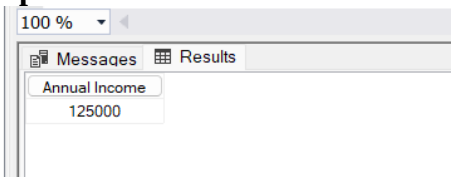
(D) Find the total annual income of ST female students from EEE branch?

Dice: Dice for Branch → Branch Name; Gender → Gender Type; Category → Category Name

Query:

```
SELECT
  [Measures].[Annual Income] ON COLUMNS
FROM [Fact]
WHERE (
  [Dim Gender].[Gender Type].[Female],
  [Dim Category].[Category Name].[ST],
  [Dim Branch].[Branch Name].[EEE]
)
```

Output:



Execution Time: 4 ms

(E) Compare rank of students across branch (columns) and gender (rows)?

Pivot: Interchanging rows and columns for slice operation

Query:

```
SELECT
  [Dim Branch].[Branch Name].MEMBERS ON COLUMNS,
  [Dim Gender].[Gender Type].MEMBERS ON ROWS
FROM [Fact]
WHERE ([Measures].[Rank])
```

Output:

Messages | Results

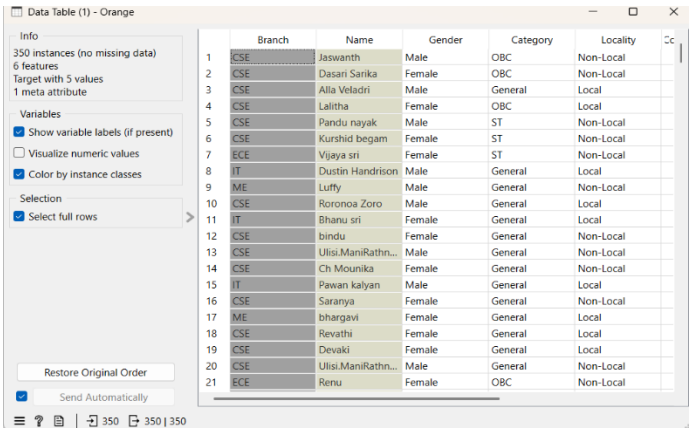
	All	CSE	ECE	EEE	IT	ME	Unknown
All	262451.3	165451.3	65000	(null)	5000	27000	(null)
Female	72001	27001	40000	(null)	5000	(null)	(null)
Male	190450.3	138450.3	25000	(null)	(null)	27000	(null)
Unknown	(null)	(null)	(null)	(null)	(null)	(null)	(null)

Execution Time: 3 ms

Step 5: Perform Data Mining on the Dataset

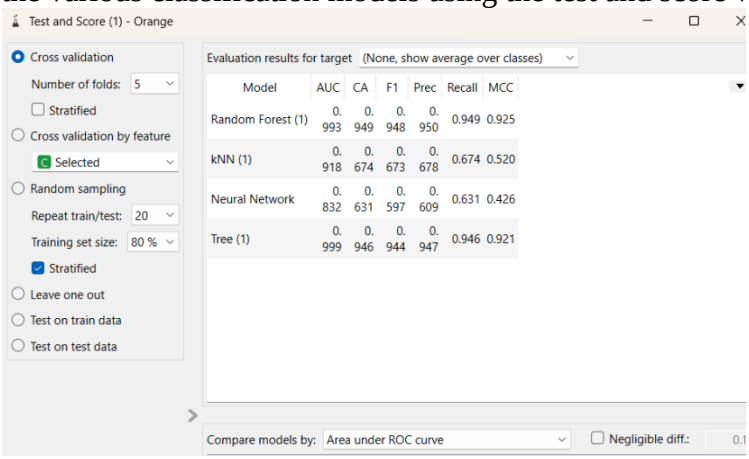
One can perform any data mining technique like Classification, Regression, Clustering and Association rule Mining. According to my data set we choose to perform Classification on data set we want to Classify the type . We mentioned this also in our problem Statement.

- The Most suitable model for my preprocessed dataset is Classification. The target class has three values in my classification model.



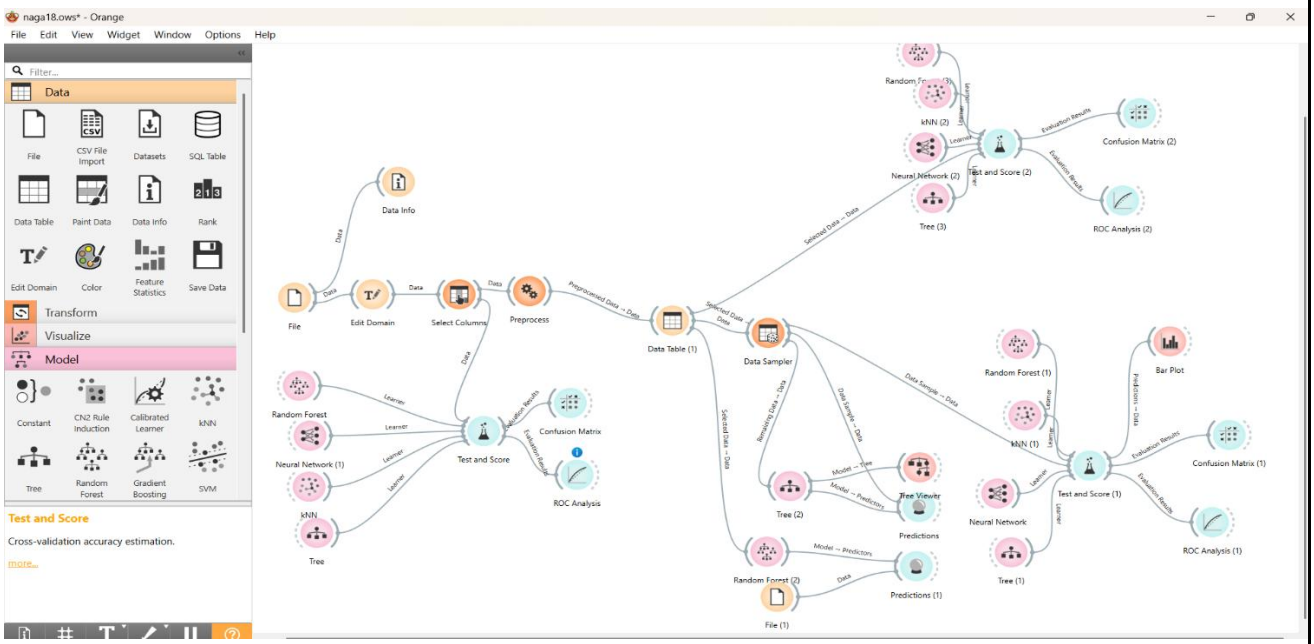
	Branch	Name	Gender	Category	Locality
1	CSE	Jaswanth	Male	OBC	Non-Local
2	CSE	Dasari Sarika	Female	OBC	Non-Local
3	CSE	Alla Veladri	Male	General	Local
4	CSE	Lalitha	Female	OBC	Local
5	CSE	Pandu nayak	Male	ST	Non-Local
6	CSE	Kurshid begam	Female	ST	Non-Local
7	ECE	Vijaya sri	Female	ST	Non-Local
8	IT	Dustin Handrison	Male	General	Local
9	ME	Luffy	Male	General	Non-Local
10	CSE	Roronoa Zoro	Male	General	Local
11	IT	Bhanu sri	Female	General	Local
12	CSE	bindu	Female	General	Non-Local
13	CSE	Ulisi.ManiRathn...	Male	General	Non-Local
14	CSE	Ch Mounika	Female	General	Non-Local
15	IT	Pawan kalyan	Male	General	Local
16	CSE	Saranya	Female	General	Non-Local
17	ME	bhargavi	Female	General	Local
18	CSE	Revathi	Female	General	Local
19	CSE	Devaki	Female	General	Local
20	CSE	Ulisi.ManiRathn...	Male	General	Non-Local
21	ECE	Renu	Female	OBC	Non-Local

- Check the Classification (Classification Methodology is explained further in Part-B in detail) accuracy for the various classification models using the test and score widget.



Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest (1)	0.993	0.949	0.948	0.950	0.949	0.925
kNN (1)	0.918	0.674	0.673	0.678	0.674	0.520
Neural Network	0.832	0.631	0.597	0.609	0.631	0.426
Tree (1)	0.999	0.946	0.944	0.947	0.946	0.921

- Here the Workflow Model in Orange Tool for my classification.



CHAPTER 3: EXPERIMENTAL ANALYSIS

- Study Classifier Accuracy
  Use Test & Score widget to view the classifier output, including accuracy, precision, recall, F-measure, and other metrics.

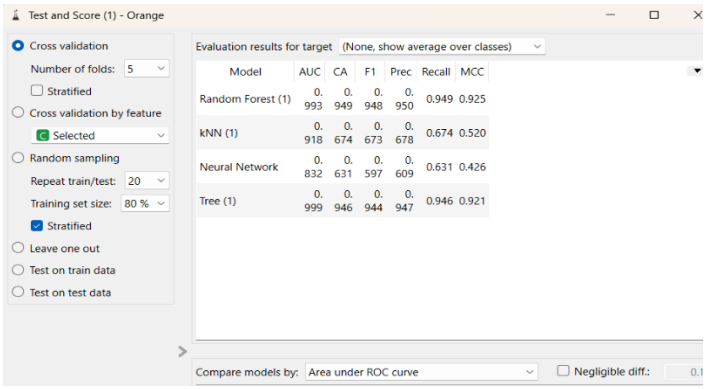


Figure 3: Test & Score measurements

Figure 3 shows readings of various evaluation metrics like AUC, CA, F1, Frequency, etc…

- Evaluate Model Performance
  Observe the confusion matrix and derive metrics such as Accuracy, F- measure, True Positive Rate (TPR), False Positive Rate (FPR), Precision, and Recall.
  Apply cross-validation strategy with various fold levels in the Test & Score widget to compare accuracy results.

- This is the confusion matrix for the best classification model (Here in our case best model based on CA is Classification tree)

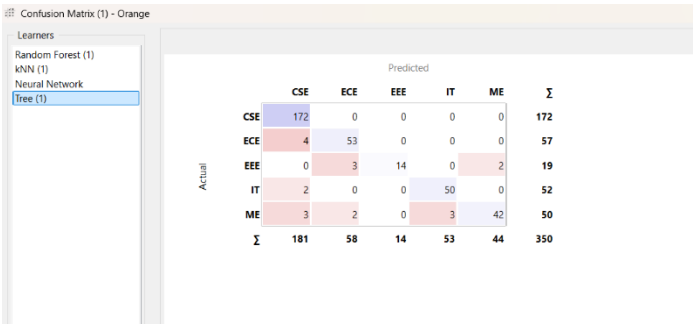


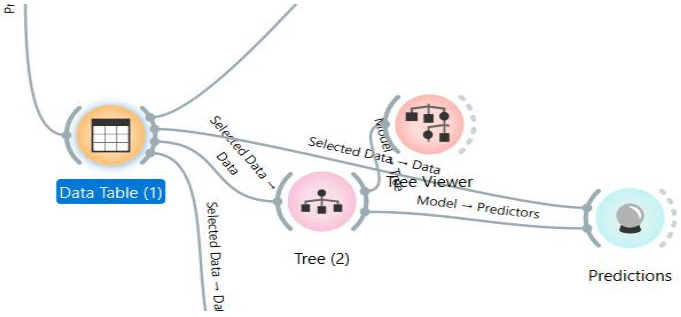
Figure 4: Confusion matrix of tree

Figure 4 shows that the classification performance of a decision tree model for Branches It also accurately classifies most "CSE" (172/177) and "ME" (42/50) instances. However, it struggles with "ECE" (5/13 correct) and "IT" (0/14 correct), indicating significant misclassifications.

Prediction model:

Now based on the classifier accuracy we developed a predicion model by spliting the dataset into training dataset and testing dataset.

The training dataset is directly given from the preprocessor and test data is given externally to the prediction model.



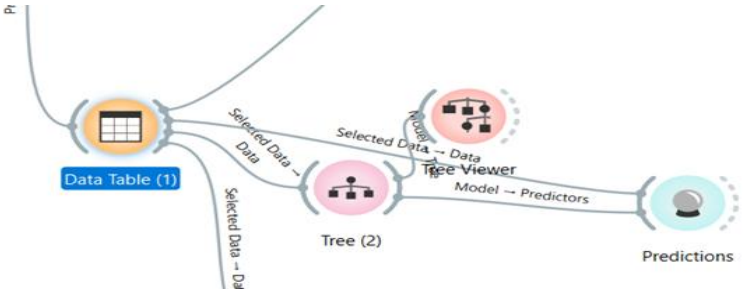
- The predictions that are given by our classification tree prediction model is
  The prediction model also gives error rate and accuracy.

Predictions - Orange

Show probabilities for: Classes in data — Show classification errors

	Tree (2)	error	Branch	Name	Gender	Category	Locality	Counseling Rounc	Rank	Annual Income
1	1.00 : 0.00 : 0.00 : 0.00 : 0.00 → CSE	0.000	CSE	Jaswanth	Male	OBC	Non-Local	1	20000.0	90000
2	1.00 : 0.00 : 0.00 : 0.00 : 0.00 → CSE	0.000	CSE	Dasari Sarika	Female	OBC	Non-Local	1	1.0	500000
3	1.00 : 0.00 : 0.00 : 0.00 : 0.00 → CSE	0.000	CSE	Alla Veladri	Male	General	Local	1	38450.3	12839500
4	1.00 : 0.00 : 0.00 : 0.00 : 0.00 → CSE	0.000	CSE	Lalitha	Female	OBC	Local	2	13000.0	60000
5	1.00 : 0.00 : 0.00 : 0.00 : 0.00 → CSE	0.000	CSE	Pandu nayak	Male	ST	Non-Local	1	80000.0	80000
6	1.00 : 0.00 : 0.00 : 0.00 : 0.00 → CSE	0.000	CSE	Kurshid begam	Female	ST	Non-Local	1	14000.0	40000
7	0.00 : 1.00 : 0.00 : 0.00 : 0.00 → ECE	0.000	ECE	Vijaya sri	Female	ST	Non-Local	1	40000.0	85000
8	0.00 : 0.00 : 0.00 : 1.00 : 0.00 → IT	0.000	IT	Dustin Handrison	Male	General	Local	1	1200.0	12839500
9	0.00 : 0.00 : 0.00 : 0.00 : 1.00 → ME	0.000	ME	Luffy	Male	General	Non-Local	2	38450.3	12839500
10	1.00 : 0.00 : 0.00 : 0.00 : 0.00 → CSE	0.000	CSE	Roronoa Zoro	Male	General	Local	1	1400.0	12839500
11	0.00 : 0.00 : 0.00 : 1.00 : 0.00 → IT	0.000	IT	Bhanu sri	Female	General	Local	2	4836.0	120000
12	1.00 : 0.00 : 0.00 : 0.00 : 0.00 → CSE	0.000	CSE	bindu	Female	General	Non-Local	1	38450.3	90000

- Now coming to the Visualization of the model we selected a tree viewer for our classification tree model.



- The classification tree for the Classification model is as follows

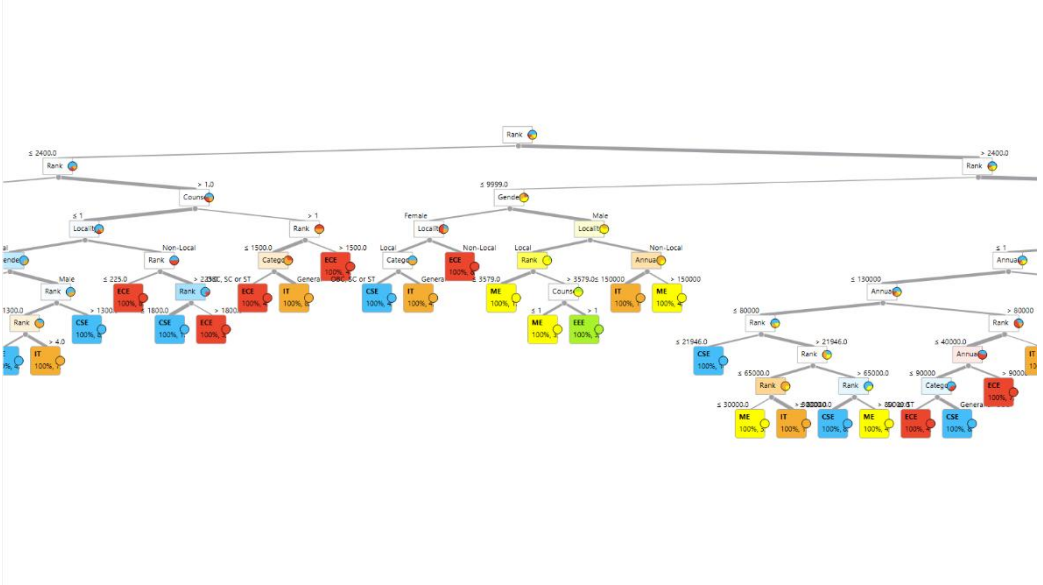


Figure 10: Classification tree

CONCLUSION:

In this project of our project, we successfully implemented an end-to-end data-driven approach to classify Branch for student based on their rank,gender,cast and annual income. I began by collecting and preprocessing the dataset, ensuring data quality through cleaning, normalization, and feature engineering. Following this, I designed and implemented OLAP schemas (Star, Snowflake, and Fact Constellation), inserted data into fact and dimension tables using Oracle SQL in SSMS, and visualized these schemas in Visual Studio.

After deploying the OLAP schemas on the SSAS server, we performed OLAP operations such as Drill-Down, Roll-Up, Slice, and Dice to extract meaningful insights from user data. The OLAP results were then visualized using Orange Tool, providing a clear representation of user behavior trends. Finally, we conducted a classification analysis to categorize branches like CSE,ME,ECE,IT,EEE based on their Rank and gender. Multiple machine learning models were evaluated, and the best-performing model was identified using accuracy, precision, recall, F1-score, and ROC analysis. This part of the project demonstrates the effectiveness of OLAP and machine learning in understanding predicting branch to student. The insights gained can be leveraged for personalized recommendations, targeted marketing, and content optimization, ultimately enhancing the user experience**.**

In this project, we successfully classified students based on their **student eamcet details** using **Supervised Machine Learning techniques**. Among the various models tested, the **Decision Tree** model demonstrated the highest accuracy and outperformed other classifiers. This classification model can help identify employability trends and assist institutions in tailoring training programs based on student details and rank preserved.

PART-B: TITANIC SURVIVAL
CHAPTER 1: INTRODUCTION ON DATA MINING METHODOLOGY

1.1 Problem Statement :

Each team is assigned with a dataset randomly to perform data mining using various techniques like Classification, Regression. Clustering and Association Rule Mining.
We need to identify the appropriate technique to perform data mining on assigned dataset.
We are assigned with "A10_titanic.tab" dataset.

1.2 Identification of appropriate Methodology:

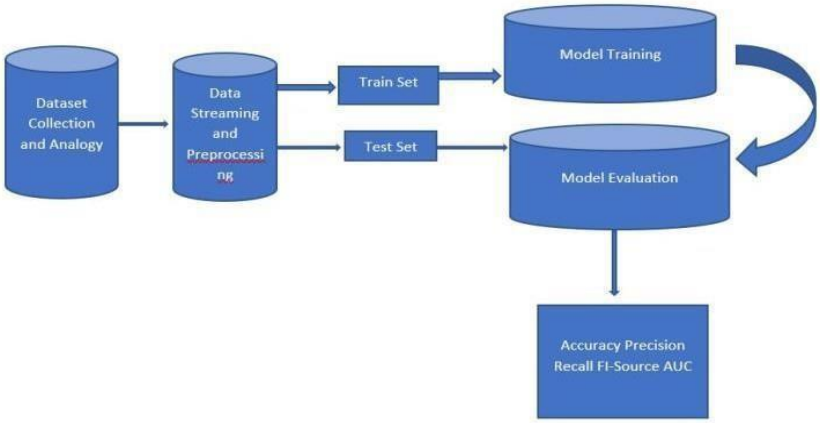
First the dataset assigned to us is loaded into orange tool to known about the dataset. Orange tool identified our dataset to be multi target dataset. We decided that the dataset would be better for classification.

1.2.1 Dataset Overview

The dataset A10_titanic.tab contains information about passager, including their characteristics such as status,age,sex,survived. The goal is to perform classification with the target variable "survived".

1.2.2 Methodology

We need the processes the dataset and make sure there are no redundancies test various classification algorithms and then develop the prediction model by training with training dataset and testing it with the testing dataset.

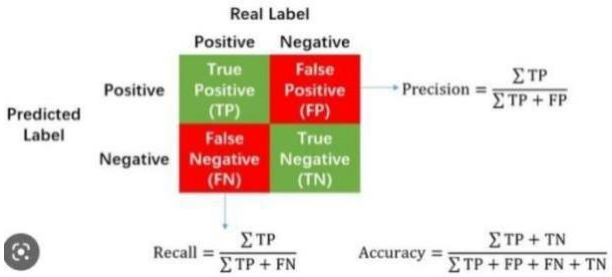


1.2.3 Machine Learning Models

We intend to use Supervised Machine Learning models to classify passengers by selecting target variable as survived. They are Naïve Bayes, Logistic Regression, Random Forest, Decision Tree, KNN, SVM.

1.2.4 Evaluation Metrics

Since this is a classification problem, we can use:
Accuracy: Accuracy is Calculated and Compared and best one should be noticed.
Precision: It counts the number of predictions from the positive class that are actually in that class.
Recall: It calculates how many positive class predictions were made using all of the dataset's positive examples.
F-Measure: It offers a single score that evenly weighs issues of precision and recall.
Confusion Matrix: It is used to determine the classification models performance for a set of test data.

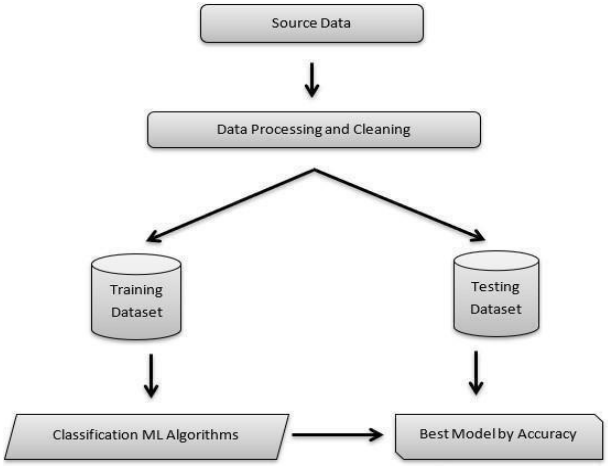


Confusion matrix

Block Diagram

The diagram illustrates the machine learning workflow for classification:

1. Source Data undergoes Data Processing and Cleaning to remove inconsistencies and prepare it for analysis.
2. The dataset is split into Training and Testing sets, ensuring proper evaluation.
3. Classification algorithms are applied to the training set, and the best model is selected based on accuracy from the testing set.

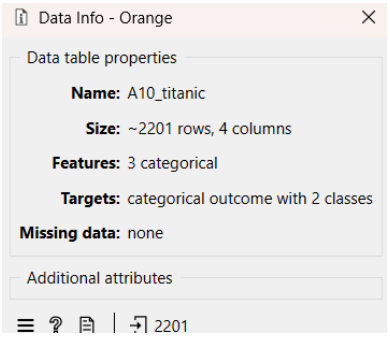


Block Diagram

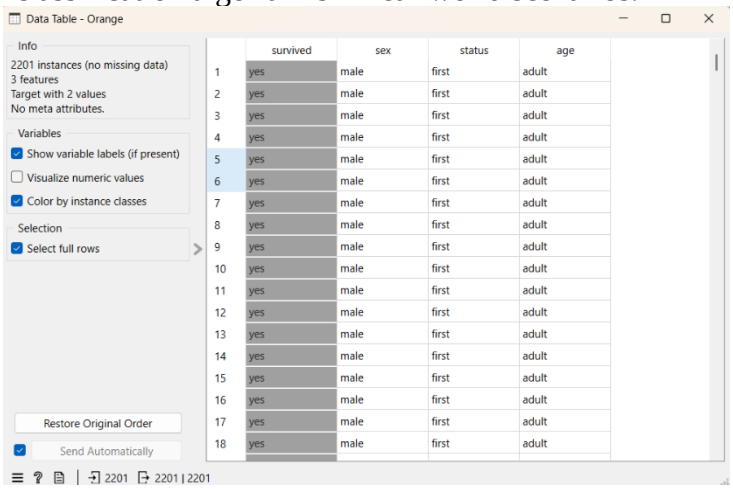
CHAPTER 2: ANALYSIS ON THE DATASET

Data Set Description:

This dataset contains demographic and survival information of passengers aboard the RMS Titanic. It is commonly used for binary classification tasks in machine learning. The goal is to predict whether a passenger **survived** or **did not survive**, based on a few basic attributes.
These features include:



Each entry in the dataset offers valuable insights into survival patterns and demographic factors aboard the Titanic, supporting the classification of passenger outcomes based on key attributes such as class, age, and gender. This dataset serves as a foundational resource for analyzing survival dependencies, evaluating the impact of socio-economic status on emergency outcomes, and developing predictive models to understand human behavior in crisis situations. It is instrumental in exploring historical data-driven decision-making and improving the accuracy of classification algorithms in real-world scenarios.



Dataset Splitting:

Given the limited dataset, we carefully split the data into **training and test sets** to ensure a robust model evaluation. The dataset was divided as follows:

- **Training Set:** 70% of the data
- **Test Set:** 30% of the data

This split was designed to maintain a balanced representation of survival outcomes in both training and testing sets, enabling effective development and evaluation of machine learning models despite the limited number of categorical attributes. This approach ensures that the model generalizes well and captures meaningful patterns across different passenger groups.

Data Visualization:

For data visualization, we utilized the Orange Data Mining tool to explore relationships and correlations among key features such as **passenger class, age group, gender, and survival status**. This analysis helped uncover dependencies between attributes and identify influential factors contributing to survival prediction. Prior to modeling, we conducted data cleaning to remove non-informative rows and ensure the dataset contained no missing or duplicate entries. We then analyzed the distribution of survival outcomes across different passenger groups, which simplified the model and enhanced both interpretability and predictive performance.

Data Validation, Cleaning, and Preparation Process

We meticulously assessed the Titanic dataset to ensure its accuracy and readiness for predictive analysis. The process began by identifying key variables such as passenger class, age group, gender, and survival status (the target variable). Through careful inspection, we removed irrelevant metadata and addressed structural inconsistencies to maintain the dataset's integrity for classification tasks. To enhance model reliability, we applied data preprocessing techniques using the Orange Data Mining tool. Our approach included:

• **Handling Missing Values:** Although the dataset contained minimal missing information, any inconsistencies were resolved using appropriate imputation methods based on model-driven accuracy comparisons.

• **Target Variable Verification:** The **survived column**, representing the target outcome, was reviewed and cleaned to ensure consistency across all entries.

• **Normalization of Categorical Inputs:** Categorical features such as **status, age, and sex** were encoded and preprocessed using Orange's Preprocessing widget to improve model interpretability and performance.

In real-world scenarios, datasets often suffer from imbalances or structural noise, making validation a critical step. We validated the Titanic dataset by examining variable types (categorical vs. numerical), verifying class distributions, and confirming a balanced representation of survival outcomes. To evaluate model performance and fine-tune hyperparameters, the dataset was partitioned into training and validation sets. This ensured unbiased model evaluation and helped optimize classification accuracy across unseen data.

Machine Learning Techniques and Model Selection

We implemented and evaluated multiple **machine learning algorithms** to classify passengers survived effectively. The following five models were tested using the **Test & Score widget** in Orange:

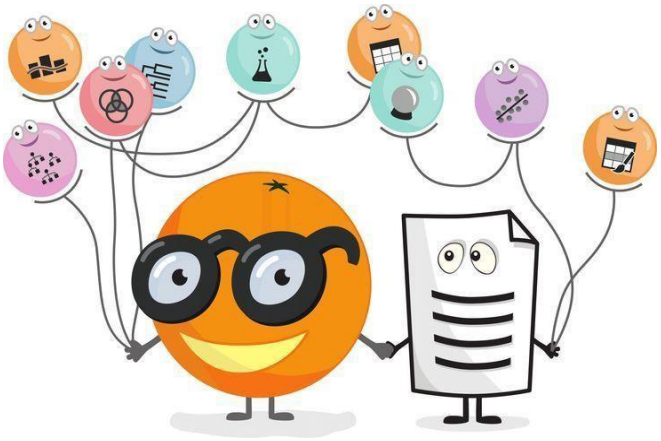
1. **Decision Tree**
2. **Random Forest**
3. **Neural Networks**
4. **KNN**

The **best-performing model** was selected based on **the highest accuracy** during testing. This approach allowed us to refine the dataset and optimize model performance for **Titanic survival prediction.**

CHAPTER 3: WORKING ON THE DATASET (DEVELOPING PREDICTION MODEL)

Orange Data Mining tool description:

The Orange tool is an open-source data visualization and analysis tool that offers a user-friendly interface for performing various machine learning and data mining tasks. It provides a visual programming interface where users can create work flows by connecting different components, suchas data loaders, preprocessing tools, and machine learning algorithms.



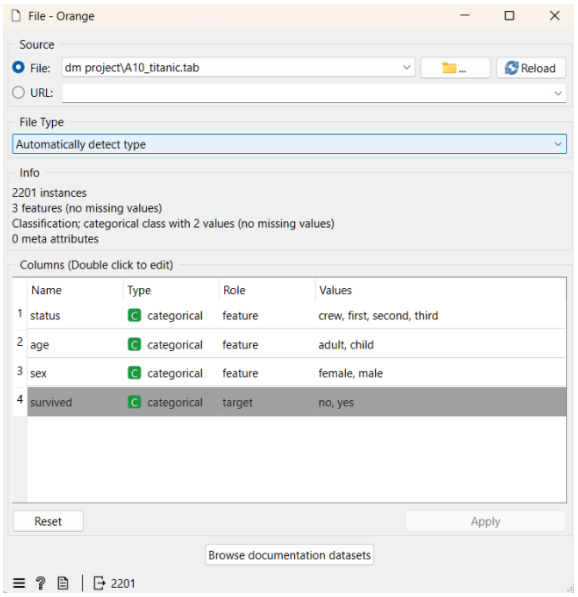
Step-by-Step Guide for Classification Using Orange

Step 1: Open Orange Canvas

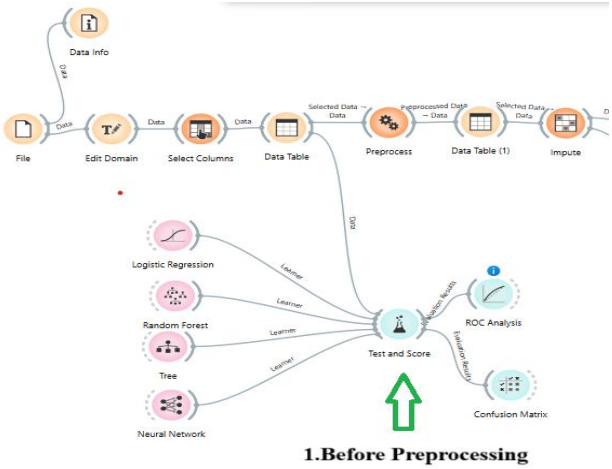
- Launch the Orange tool.
- Open the Orange Canvas to start creating your workflow.

Step 2: Load Dataset

- Drag and drop the "File" widget onto the canvas.
- Click on the "File" widget and then click on the "Browse" button.
- Choose your dataset(e.g.” A10_titanic.tab”) and open it.



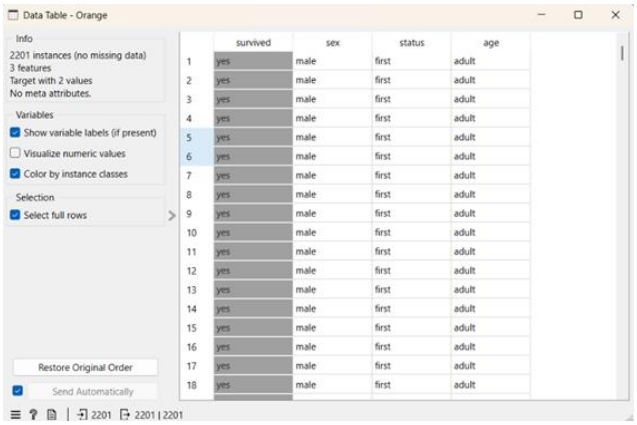
Step 3: Test the accuracies for various classification algorithms before preprocessing & choose the top four according to their accuracies.



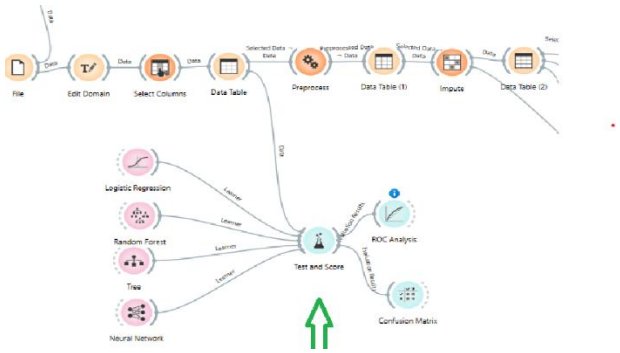
1.Before Preprocessing

Step 4: Preprocessing dataset

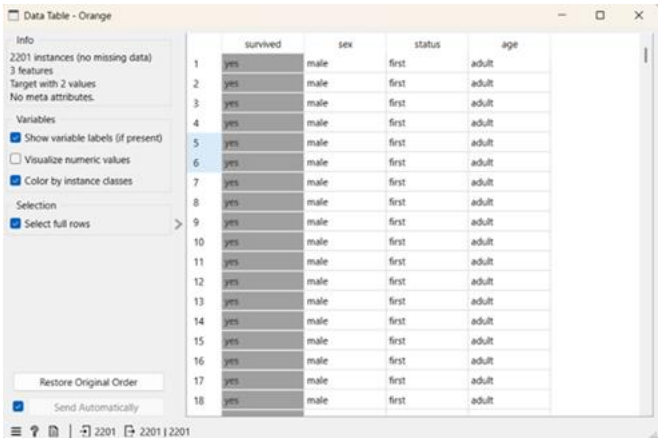
- Drag and drop the "Preprocess" widget onto the canvas.
- Connect the "File" widget to the "Preprocess" widget.
- Select the preprocess technique to remove missing values and normalize the numeric values.
- To check whether the missing values are replaced or not connect it to the “Data table widget”. Data table shows the information related to dataset.



Missing values before preprocessing



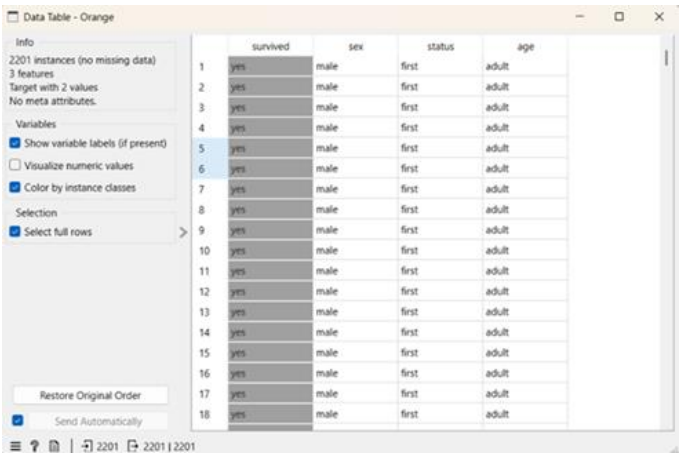
Preprocessing and imputing



Missing values after preprocessing

Step 5: Imputing target variables

- After the preprocessing since there are missing values in the target variable connect the “Preprocess” widget to “impute” widget to particularly impute a specific also select the imputing technique that shows highest accuracy among chosen models.



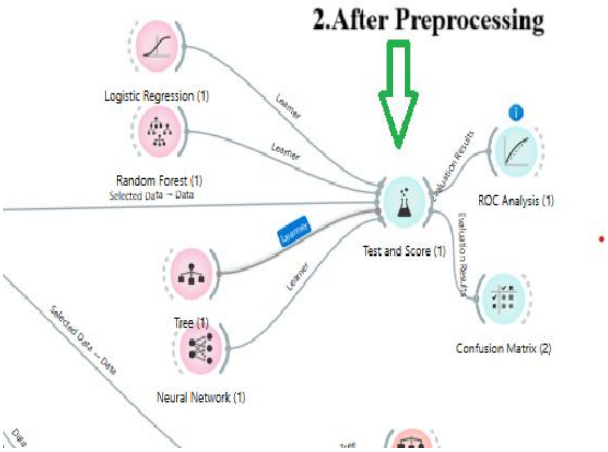
Dataset now has no missing values after imputing the target variable

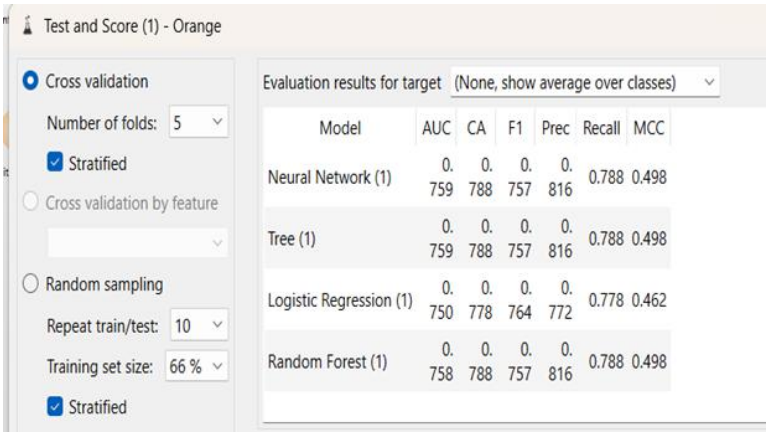
Step 6: Testing accuracy of various classification algorithms

- Drag and drop the "Test & Score" widget

5. Connect the " Neural Networks", ”Random Forest”, Decision Tree”, “KNN” widgets to the "Test & Score" widget.

- Click on the "Test & Score" widget to view the classifier output, including accuracy, precision, recall, F-measure, and other metrics.

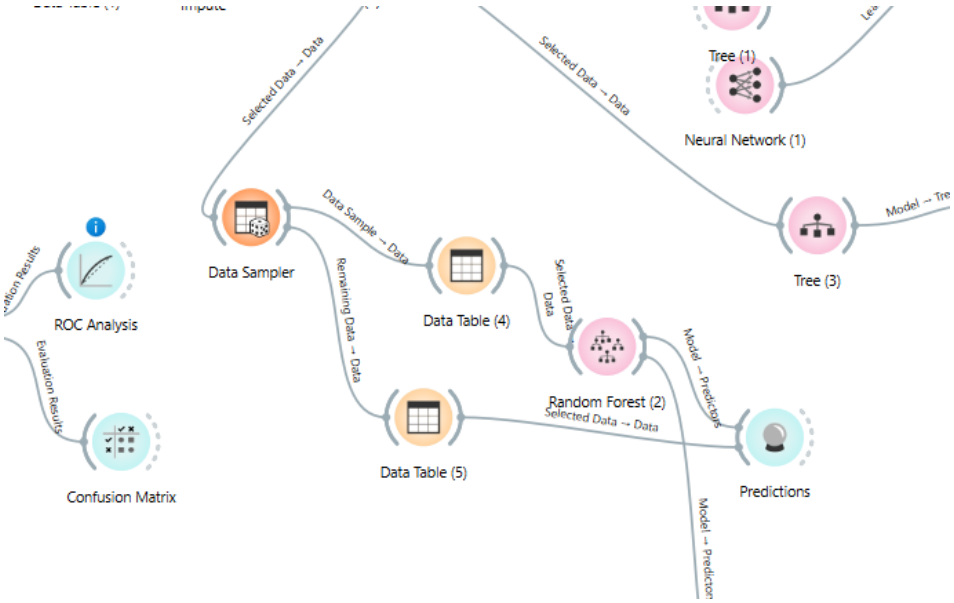




- Make note of classifier accuracies CA to compare various algorithms before and after preprocessing.
- Apply cross-validation strategy with various fold levels in the "Test & Score" widget to compare accuracy results.

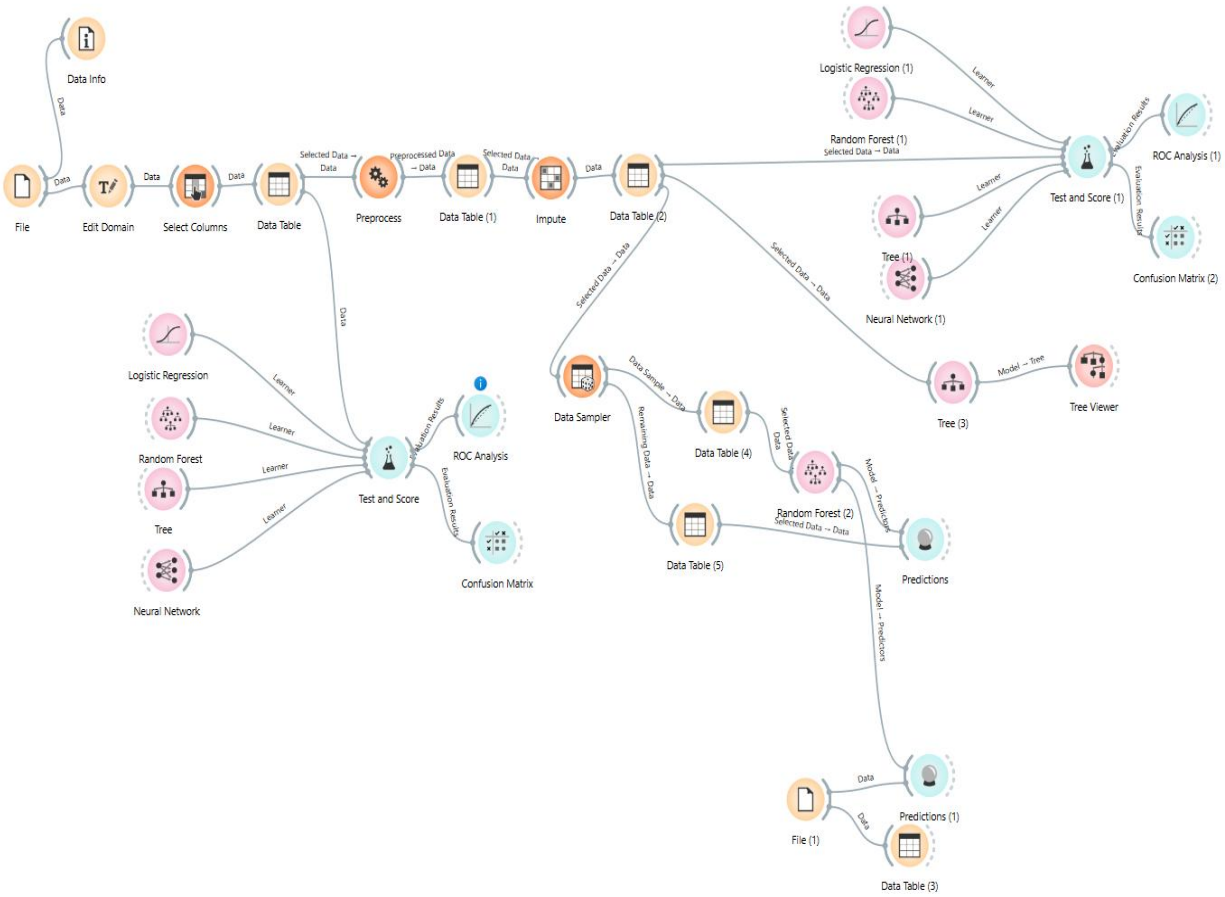
Step 7: Developing prediction model for the learning algorithm with best accuracy.

- The prediction model needs both training and test data. Based on the training and test data the prediction model can be developed in two ways-
  First way is by splitting the dataset into training and test datasets using the data sampler
  This is clearly explained the figure below:



Prediction model by data sampler

Entire Workflow:



Step 8: Perform Visualization for the algorithms. Here We choose classification tree to visualize the output in the orange tool. (Since other techniques failed to visualize our dataset properly we preferred classification tree)

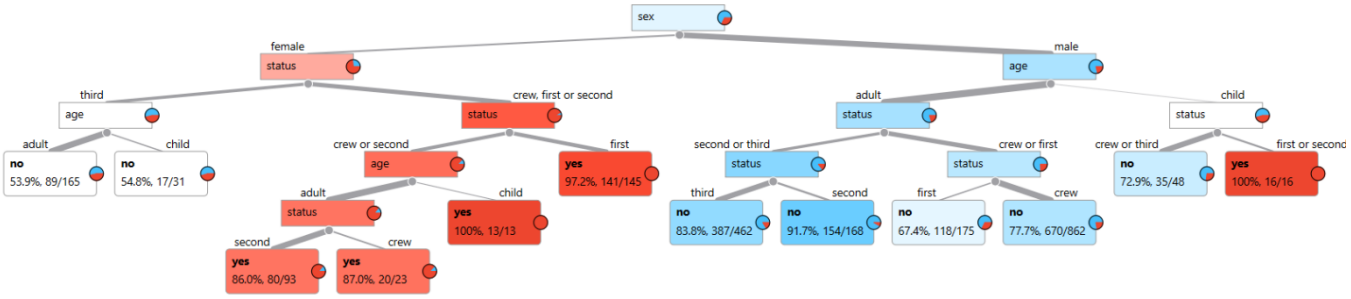


Figure 11: Classification tree for Span attribute

Figure 11 depicts the decision tree visually represents classification based on multiple attributes like **SEX,AGE,STATUS**. Nodes split based on attribute values, with leaf nodes showing classification percentages. Green nodes indicate higher confidence in classification, while red nodes represent lower confidence, suggesting possible misclassifications or mixed results in those branches.

CHAPTER 4: EXPERIMENTAL ANALYSIS

- Based on the Classifier accuracy that is shown in the Test & Score widget we choose to evaluate KNN and Logistic Regression algorithms using various metrics like confusion matrix and Roc Analysis.

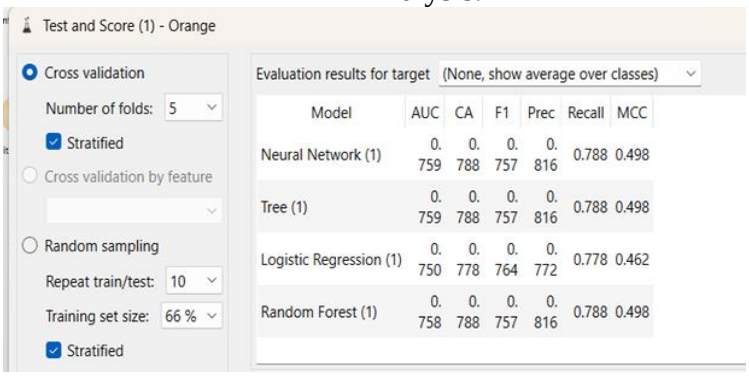


Figure 12: CA before preprocessing

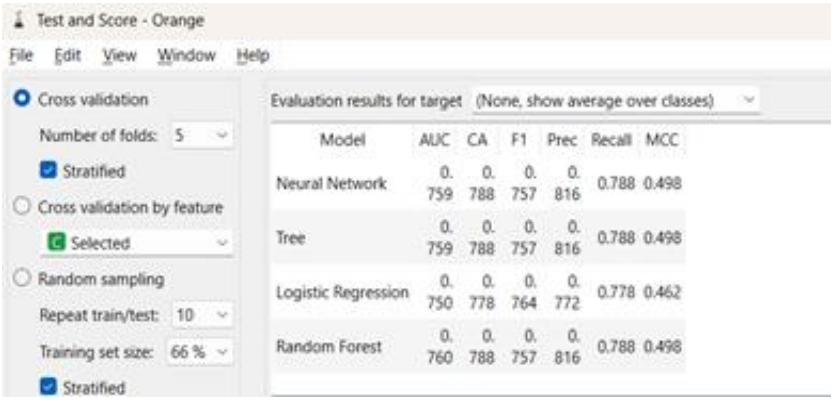


Figure 13: CA after preprocessing

Figure 12 shows Test & Score before preprocessing (5-fold Cross Validation, 66% Training Size)

- Neural Network: 0.759
- Decision Tree: 0.759
- Logistic Regression: 0.750
- Random Forest: 0.758

Figure 13 shows Test & Score after preprocessing (5-fold Cross Validation, 66% Training Size)

- Neural Network: 0.759
- Decision Tree: 0.759
- Logistic Regression: 0.750
- Random Forest: 0.760

Comparison Observations:
All models improved their CA values after preprocessing while maintaining a 66% training size and 5-fold cross-validation.
Random Forest observed a slight improvement from 0.758 to 0.760 after preprocessing.
Logistic Regression, Decision Tree, and Neural Network maintained consistent performance before and after preprocessing, with scores of 0.750, 0.759, and 0.759 respectively.
Analysis on Confusion matrices:
These are the confusion matrices for the two best classification algorithms.

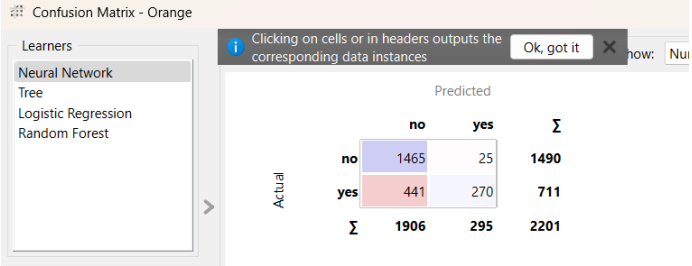


Figure 14: Confusion matrix of Neural Network

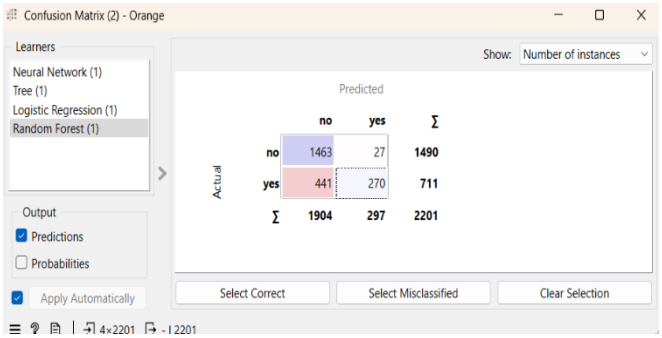


Figure 15: Confusion matrix of Random Forest

The observations that can be made by Figure 14 & Figure 15 are as follows

Overall Accuracy:

- **Neural Network:**
  - Correct classifications = 1465 (no) + 270 (yes) = **1735 out of 2201**
  - Accuracy = **78.83%**
- **Random Forest:**
  - Correct classifications = 1463 (no) + 270 (yes) = **1733 out of 2201**
  - Accuracy = **78.74%**

◆ **Observation:** Neural Network has a **slightly higher overall accuracy** than Random Forest (by 0.09%).

Class-wise Performance:

Class: No

- **Neural Network:**
  - 1465 correctly predicted as "no"
  - 25 misclassified as "yes"
- **Random Forest:**
  - 1463 correctly predicted as "no"
  - 27 misclassified as "yes"
- **Result:** Neural Network performs marginally better on "No" class (fewer misclassifications)

Class: Yes

- **Neural Network:**
  - 270 correctly predicted as "yes"
  - 441 misclassified as "no"
- **Random Forest:**
  - 270 correctly predicted as "yes"
  - 441 misclassified as "no"
- **Result:** Both models perform identically for "Yes" class

Misclassification Trends:

- **Neural Network:**
  - 25 instances of “No” misclassified as “Yes”
  - 441 instances of “Yes” misclassified as “No”

- **Random Forest:**
    - 27 instances of “No” misclassified as “Yes”
    - 441 instances of “Yes” misclassified as “No”

In conclusion, The Neural Network model demonstrated slightly better overall accuracy (1735/2201) compared to the Random Forest model (1733/2201). Both models performed equally well for the "Yes" class, correctly classifying 270 instances, but the Neural Network had fewer misclassifications in the "No" class (25 vs. 27). This suggests that the Neural Network is marginally more effective in reducing false positives. Given the very close performance, either model could be chosen depending on other factors like interpretability or computational efficiency; however, if the primary goal is maximizing accuracy and minimizing false alarms, the Neural Network is the preferable choice.

Analysis on ROC (Receiver Operating Characteristic) analysis:
These are the ROC analysis for the two best classification algorithms.

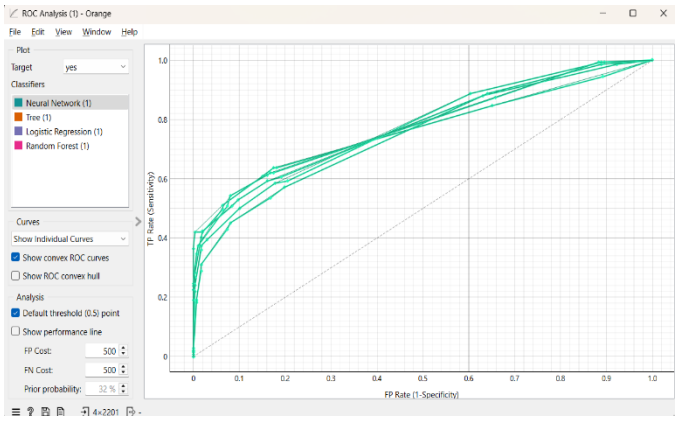


Figure 16: ROC Analysis of Neural Network

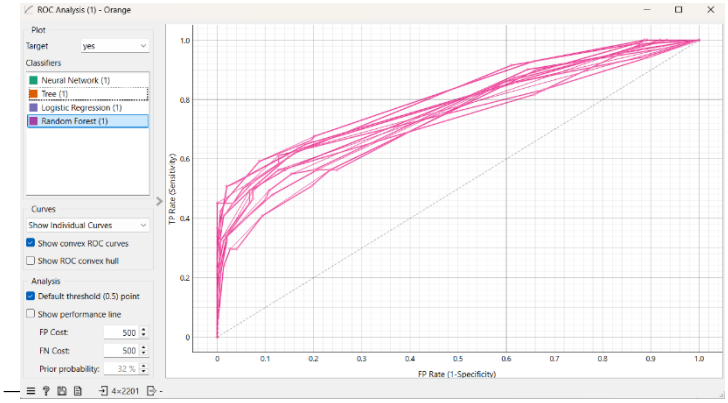


Figure 17: ROC Analysis of Random Forest

Figure 16 & 17 depict ROC analysis for different classifiers, with the following important metrics highlighted:

1. **Default Threshold**:
    - The evaluation is performed at a threshold of **0.5**.
2. **False Positive Cost & Prior Probability**:
    - **False Positive Cost** is set to **500**.
    - **False Negative Cost** is also set to **500**.
    - **Prior Probability** is noted at **32%**.

These parameters influence the decision-making costs and guide the threshold optimization process for the models in the analysis.

CONCLUSION:

In this classification analysis using Orange Data Mining, we evaluated multiple machine learning algorithms before and after preprocessing. The key findings are as follows:

1. **Preprocessing Impact**
    - Data preprocessing improved model performance by handling missing values and normalizing numeric features.
    - Random Forest showed the best classification accuracy (CA) after preprocessing, while Neural Networks performed best before preprocessing.
2. **Overall Classification Accuracy**
    - Neural Networks achieved a slightly better overall accuracy (1735/2201) compared to Random Forest (1733 / 2201)
    - Neural Networks performed better for Class No, while both are better for Class Yes.
3. **Confusion Matrix Insights**
    - Neural Network and Random Forest achieved identical classification results (1735 correct predictions out of 2201), showing robust prediction capabilities.
    - Detailed analysis showed Neural Network slightly edged out others in stable accuracy across classes, while Random Forest offered comparable performance..
4. **ROC Analysis**
    - Both Neural Networks and Random Forest demonstrated strong classification capabilities.
    - The ROC curves provide insights into threshold optimization, with default thresholds set at 0.5 and false positive/negative costs at 500.

Final Decision

- For balanced performance across classes, both models perform similarly and are suitable depending on the specific application requirements
- If minimizing false positives is a priority (i.e., not wrongly predicting someone survived when they didn't), the Neural Network is the better choice.

Future Scope for the Bridges MT2 Dataset with Multi-Target Classification

The **Bridges MT2 dataset**, which contains detailed attributes of bridge structures, presents several avenues for future research and development. Given its multi-target nature, various improvements and applications can be explored:

1. Model Optimization & Ensemble Learning

- **Explore ensemble techniques like Voting Classifiers or Boosting to enhance survival prediction.**
- **Introduce deep learning models through Python scripting or TensorFlow integration with Orange.**

2. Advanced Feature Engineering

- **Derive features like "family size," "deck," or "ticket group size" to improve class separation.**
- **Use NLP techniques on text fields like "Name" and "Cabin" to extract hidden patterns.**

3. Class Imbalance Handling

- **Implement SMOTE or under-sampling techniques to handle the slight imbalance in survival vs. non-survival classes.**

4. Explainability & Fairness Analysis

- **Use tools like SHAP values or LIME to explain model decisions.**
- **Analyze gender or socio-economic bias in predictions, ensuring fairness in survival modeling.**

5. Deployment & Visualization

- **Create a simple web app for real-time survival prediction using pre-trained models.**

- **Integrate Orange workflows with Power BI or Dashboards for intuitive stakeholder communication.**

By expanding these research directions, the **Titanic survival dataset** can be leveraged for developing more accurate and insightful **passenger survival prediction systems**, contributing to improved understanding of human behavior during disasters. Advanced machine learning techniques, feature engineering, and real-time analytics can transform this dataset into a foundation for training safety response models, designing optimized evacuation strategies, and enhancing maritime safety protocols. Furthermore, incorporating historical context, demographic profiling, and behavioral analysis can support broader applications in transportation safety research, policy development, and educational simulations aimed at minimizing casualties in real-world emergencies.

PART C: FINAL ANALYSIS

1. Introduction

In data mining and machine learning, dataset selection plays a crucial role in determining the effectiveness of the models applied. This study analyzed two different experimental setups:
Part A, which used a generated dataset, and
Part B, which used an online dataset collected from external sources.
This section aims to integrate insights from both experiments and provide final conclusions regarding their performance, applicability, and limitations.

2. Key Observations from Experimental Analysis

2.1. Data Characteristics and Preprocessing

One of the fundamental differences between the two parts was the nature of the dataset used.

- Generated dataset (Part A) was pre-structured, leading to minimal preprocessing efforts. There were some missing values or inconsistencies, making it easier for models to achieve high accuracy with basic tuning.
- Online dataset (Part B) required extensive data cleaning, imputing and normalization due to missing or inconsistent values. This extra preprocessing influenced the model performance significantly.
- Despite the challenges of Part B, working with real-world datasets helps build more generalized models that can perform well on unseen data.

2.2. Model Performance Analysis

- Across both experiments, various classifiers (KNN, Logistic Regression, Gradient Boosting, SGD, and Random Forest, etc…) were tested, and their performances were evaluated using classification accuracy (CA), ROC curves, and confusion matrices.

2.3. Key Findings from Model Comparisons

- Neural Network consistently performed the best across both datasets, benefiting from its ability to classify instances effectively when properly tuned.
- Random Forest showed improved performance after preprocessing, suggesting that real-world data benefits significantly from preprocessing techniques.
- Decision Tree had a lower initial accuracy but improved post-preprocessing, highlighting their dependence on high-quality input data.

3. Preprocessing Differences

Part A: Minimal Preprocessing

- Data was structured and balanced.
- No missing values, meaning models performed well **without extensive preprocessing**.
- Features were designed to align with model input requirements.

Part B: Extensive Preprocessing Required

- Missing values had to be handled using imputation techniques.
- Normalization and scaling were applied to align feature ranges.
- Feature selection was needed to remove redundant attributes.
- Data imbalance was addressed to improve model performance.

Impact:

- Models in **Part B performed worse initially** but improved **significantly after preprocessing**.
- **Preprocessing had a major impact on classification accuracy (CA) in Part B**.

4. Findings from ROC Analysis

- Before preprocessing, the models struggled with inconsistent and incomplete data, leading to lower accuracy.42
- After preprocessing, all models improved significantly, showing that clean and well-prepared data

leads to better predictions.

- Neural Networks, SVM, and Random Forest benefited the most from data cleaning and transformation.
- Preprocessing played a crucial role in making the models more accurate, stable, and useful for real world applications**.**

5. Conclusion

This study compared the experimental analysis of a generated dataset (Part A) and an online dataset (Part B) for classifying. Key takeaways include:

Part A achieved higher accuracy faster due to well-structured data.

Part B required significant preprocessing but provided a more realistic assessment of classifier performance.

Neural Networks emerged as the best-performing model, with Random Forest and Decision tree also showing promising results.

REFERENCES

1. **Multi-Target Classification & Machine Learning**
    - Tsoumakas, G., & Katakis, I. (2007). "Multi-label classification: An overview." *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3), 1-13.

    - Zhang, M. L., & Zhou, Z. H. (2014). "A review on multi-label learning algorithms." *IEEE Transactions on Knowledge and Data Engineering, 26(8)*, 1819-1837.

    - Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). "Scikit-learn: Machine learning in Python." *Journal of Machine Learning Research, 12*, 2825-2830.

2. **Bridge Structural Analysis & Design**
    - Chen, W. F., & Duan, L. (2014). *Bridge Engineering Handbook*. CRC Press.

    - Roberts-Wollmann, C., Cousins, T. E., Brown, E. R., & Nelson, J. (2012). "Bridge Load Testing and Structural Health Monitoring." *Transportation Research Board (TRB)*, 2200(1), 57-66.

    - Jang, S., Jo, H., Cho, S., Mechitov, K., Rice, J. A., Sim, S. H., & Agha, G. (2010). "Structural health monitoring of a cable-stayed bridge using smart sensor technology: Deployment and evaluation." *Smart Structures and Systems, 6(5-6)*, 439-459.

3. **Geospatial & Structural Health Monitoring (SHM)**
    - Farrar, C. R., & Worden, K. (2007). "An introduction to structural health monitoring." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 365(1851)*, 303-315.

    - Sohn, H., Farrar, C. R., Hemez, F. M., Czarnecki, J. J., & Nadler, B. (2002). "Structural Health Monitoring Framework for Civil Infrastructure." *Los Alamos National Laboratory Report*, LA-13935-MS.

    - Yan, Y. J., Cheng, L., Wu, Z. Y., & Yam, L. H. (2007). "Development in vibration-based structural damage detection technique." *Mechanical Systems and Signal Processing, 21(5)*, 2198-2211.

SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)
Seshadri Rao Knowledge Village, Gudlavalleru

Department of Computer Science and Engineering

Program Outcomes (POs)

Engineering Graduates will be able to:

1. **Engineering knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
2. **Problem analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
3. **Design/development of solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
4. **Conduct investigations of complex problems**: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of theinformation to provide valid conclusions to meet the desired needs.
5. **Modern tool usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
6. **The engineer and society**: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
7. **Environment and sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need forsustainable development.
8. **Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and normsof the engineering practice.
9. **Individual and team work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

10. **Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
11. **Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and rite
12. **Life-long learning**: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

Program Specific Outcomes (PSOs)

PSO1 : Design, develop, test and maintain reliable software systems and intelligent systems.PSO2 : Design and develop web sites, web apps and mobile apps.

PROJECT PROFORMA

Classification of Project	Application	Product	Research	Review
	√			

Note: Tick Appropriate category

Data Mining Outcomes	
Course Outcome (CO1)	Describe fundamentals, and functionalities of data mining system and data preprocessing techniques.
Course Outcome (CO2)	Illustrate the major concepts and operations of multi dimensional data models.
Course Outcome (CO3)	Analyze the performance of association rule mining algorithms for finding frequent item sets from the large databases.
Course Outcome (CO4)	Apply classification algorithms to solve classification problems.
Course Outcome (CO5)	Use clustering methods to create clusters for the given data set.

Mapping Table

CS3509 : DATA MINING														
Course Outcomes	Program Outcomes and Program Specific Outcome													
	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12		PSO 1	PSO 2
CO1	1	1										1			
CO2	1											1			
CO3	2	3	2									2		1	
CO4	2	2	3	2								2		2	
CO5	1	2	3	1								2		1	

Note: Map each Data Mining outcomes with POs and PSOs with either 1 or 2 or 3 based onlevel of mapping as follows:

1-Slightly (Low) mapped 2-Moderately (Medium) mapped 3-Substantially (High) mapped