PeerLearn Miner – Analyzing the Impact of Peer Learning on Academic Success

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

Bollina Lavanya(22481A0534)

Bhukya Sandeep(22481A0531)

Gorla Simhadri(22481A0562)

Gokapai Lakshmi Prasanna(22481A0559)

Under the Enviable and Esteemed Guidance of

Dr. M. BABU RAO, M. Tech , Ph . D

Head of the Department,CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE

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

SESHADRI RAO 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 PeerLearn Miner – Analyzing the Impact of Peer Learning on Academic Success is a bonafide record of work carried out by **B.Lavanya(22481A0534),B.Sandeep(22481A0531),G.Simhadri(22481A0562),G.Lakshmi Prasanna(22481A0559)** ,under the guidance and supervision of Dr. M. BABU RAO ,M.Tech, Ph.D, Head of the department,CSE , 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 **Head of the Department**

(Dr. M. BABU RAO) **(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 constant guidance 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, Head of the department** , 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, 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** 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

B.LAVANYA(22481A0534)

B.SANDEEP(22481A0531)

G.SIMHADRI(22481A0562)

G.LAKSHMI PRASANNA(22481A0559)

INDEX

PART C- FINAL ANALYSIS

ABSTRACT

Peer learning is a critical factor in academic success, fostering collaboration and knowledge sharing among students. This project employs data mining techniques to analyze the impact of peer learning on student performance by categorizing learners based on their study behaviors and engagement levels. The dataset, collected through surveys, includes attributes such as peer learning participation, preferred learning methods, platforms used, benefits, challenges, and academic outcomes.

To ensure data quality, preprocessing techniques were applied to handle missing values, normalize categorical data, and extract relevant features. Exploratory Data Analysis (EDA) was conducted to uncover patterns and relationships within the dataset, guiding the selection of appropriate data mining models. Various classification and clustering algorithms, including Decision Trees for classification and K-Means Clustering for behavioral grouping, were evaluated to identify distinct learning patterns. Additionally, Association Rule Mining was utilized to discover correlations between different peer learning attributes.

The models were assessed using performance metrics such as accuracy, precision, recall, and F1-score, alongside a confusion matrix for detailed analysis. The results indicate that [mention best-performing model] achieved the highest accuracy in categorizing students based on their peer learning habits.

This study highlights the potential of data mining in education by providing actionable insights into student collaboration and learning effectiveness. The findings can help educational institutions and policymakers develop personalized peer learning strategies, enhance student engagement, and improve academic outcomes through data-driven decision-making.

PART-A:*PEER LEARNING PATTERNS AND ACADEMIC SUCCESS USING KDD PROCESS*

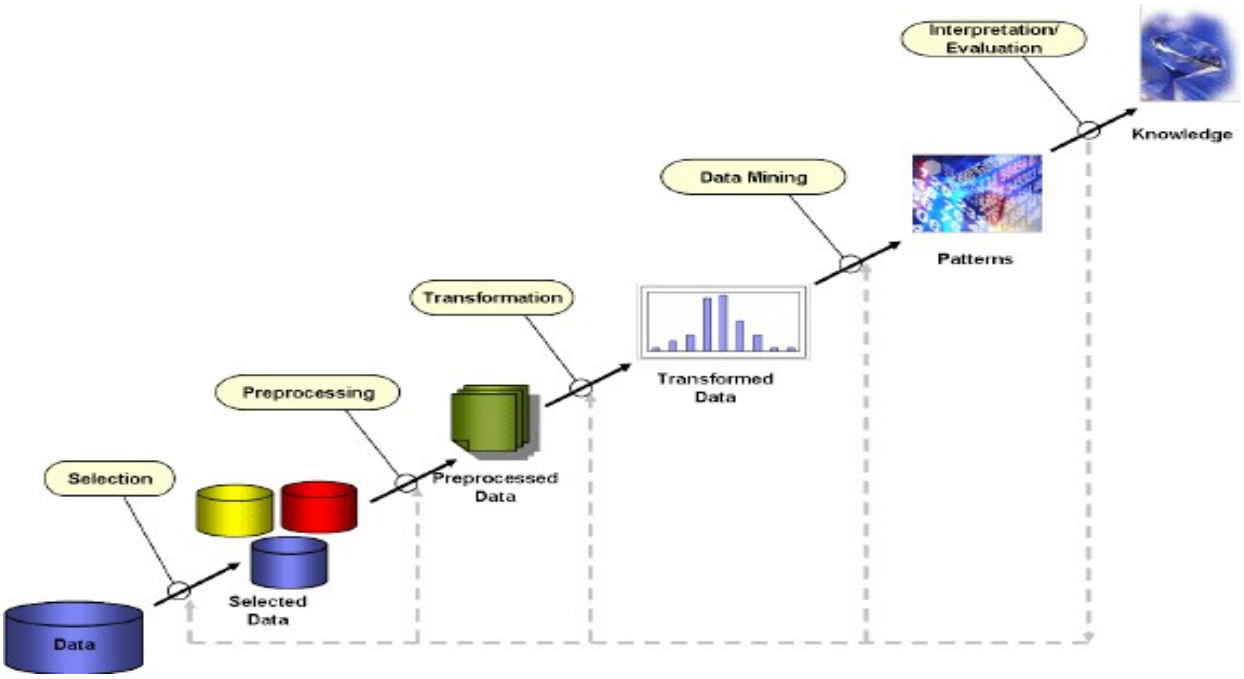
CHAPTER 1: INTRODUCTION

1.1 INTRODUCTION

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



1.2 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 thatcan learn from and make predictions on data. These fields complement each other, enhancing data analysis and predictive modeling capabilities.

1.3 DATA WAREHOUSING

In our peer learning analysis project, a data warehouse is used to store and analyze data collected from various sources, such as student surveys, academic records, and peer learning platforms. This centralized data repository enables efficient trend analysis, student segmentation, and learning behavior evaluation.

1. Data Source Layer (Extracting Data)

- Data is collected from student surveys, academic performance records, and online learning platforms.
- Includes attributes like peer learning participation, study duration, preferred learning methods, and performance metrics.

2. ETL (Extract, Transform, Load) Process

- **Extraction**: Data is gathered from multiple sources (survey forms, institutional databases, and online learning logs).
- **Transformation:** Data is cleaned, formatted, and standardized to maintain consistency.
- **Loading:** The processed data is stored in a structured data warehouse for analysis.

3. Data Storage Layer (Fact & Dimension Tables)

- **Fact Table** stores core metrics like study hours, collaboration frequency, and academic scores.
- **Dimension Tables** include details like student demographics, peer learning methods, and subjects studied.

4. OLAP (Online Analytical Processing) for Data Analysis

- Supports multi-dimensional analysis to identify trends in peer learning behavior and academic performance.
- Enables queries like:
  - Does peer learning improve academic performance?
  - Which peer learning method is most effective?
  - How does participation vary across different student demographics?

5. Data Visualization & Reporting

- Insights are presented using **dashboards, reports, and visual charts.**
- Helps educators and institutions optimize peer learning strategies based on data-driven insights.

➢ 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.

➢ DATA MINING INTRODUCTION

The block diagram for our project begins with collecting **peer learning data** through surveys, followed by **data preprocessing** to clean and standardize the data. The dataset is then split into **training and testing sets**. The training data is used to build and train various **classification and clustering models**. Finally, the models analyze the data to classify students based on **peer learning participation, study duration, and academic performance**.

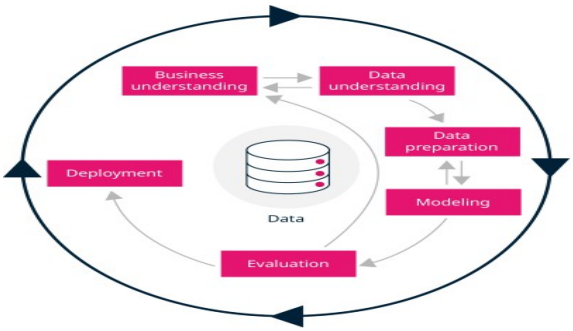


Fig 1.1 Data Mining Block Diagram

DATA MINING BLOCK DIAGRAM EXPLANATION

The data mining process follows structured steps to extract meaningful insights from the dataset:

1. **Data Understanding**
   - Collecting and analyzing the peer learning dataset to grasp its structure and content.
   - Identifying attributes such as peer learning participation, study duration, and academic performance.
2. **Data Preparation**
   - Cleaning and transforming the dataset by handling missing values, standardizing data, and encoding categorical attributes.
3. **Modeling**
   - Applying various classification algorithms like Decision Trees, Random Forest, and Logistic Regression to analyze peer learning behavior and predict academic performance."
4. **Evaluation**
   - Assessing model performance using accuracy, precision, recall, and F1-score.
5. **Deployment**
   - Integrating the best-performing model to provide insights into peer learning impact on academic success.

SUPERVISED LEARNING

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

- K-NearestNeighbors (KNN)
- Decision Trees
- Random Forest
- Logistic Regression

Categories of Supervised Learning in This Project:

1. **Classification:**
   - The dataset contains categorical labels (e.g., High, Medium, Low Academic Performance).
   - Classification algorithms predict a student's academic success based on peer learning participation, study duration, and collaboration frequency.
2. **Regression:**
   - If we analyze study hours as a continuous variable, regression models could predict a student's expected academic performance.
   - However, since our project focuses on performance classification, classification is the primary approach.

Algorithm	Description	Type
Logistic Regression	Extension of linear regression that's used for classification tasks. The output variable is **binary** (e.g., High or Low academic performance).	Classification rather regression
Decision Tree	Highly interpretable classification model that splits **peer learning attributes** into branches at decision nodes to predict academic success.	Classification
Naïve Bayes	The Bayesian method is a classification method that make s use of the Bayesian t orem updates prior knowledge of a student's performance based on independent peer learning fact	Regression and Classification
KNN	K-Nearest Neighbors (KNN) is a supervised learning algorithm that classifies students based on the labels of their nearest academic peers. It assigns the most common performance category among the closest data points to a new student.	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 a Data Mining Algorithm?

Choosing the right data mining algorithm depends on:

- **If the data has labels:**
  Use Supervised Learning (Classification/Regression)**.**
- **If the data has no labels:**
  Use Unsupervised Learning (Clustering/Association)**.**

Since our dataset focuses on **predicting platform preference**, **classification algorithms** are the best fit. However, **clustering and association rule mining** can be used for **user segmentation and behavior analysis**

Fig: Data Mining Basic Diagram

.

➢ CHALLENGES AND LIMITATIONS OF DATA MINING

One of the major challenges in data mining is ensuring **data quality and preprocessing**. In real-world scenarios, datasets often contain **noise, missing values, and inconsistencies**, which can significantly impact the effectiveness of data mining algorithms.

Key Challenges:

- **Data Cleaning & Normalization:** Raw data needs extensive cleaning to remove duplicates, inconsistencies, and errors**.**
- **Feature Selection:** Choosing the most relevant attributes is crucial for improving model accuracy.
- **Resource-Intensive Processing:** Preprocessing large and complex datasets requires significant computational power and time**.**
- **Bias & Data Limitations:** Even after cleaning**,** inherent biases in the data may affect model predictions, leading to skewed insights.

Addressing these challenges is critical for ensuring accurate and reliable predictions in data mining projects.

➢ APPLICATIONS OF DATA MINING

1. Customer Relationship Management (CRM)

Data mining helps businesses analyze customer demographics, purchase history, and behavioral trends to optimize marketing strategies**.**

- Enables personalized recommendations and targeted marketing campaigns.
- Improves customer engagement and retention.
- Identifies high-value customers and predicts churn rates**.**

2. Fraud Detection

Data mining is widely used in banking, insurance, and e-commerce to detect fraudulent transactions.

- Algorithms analyze transactional data to detect anomalies.
- Identifies patterns indicating fraudulent behavior.
- Enhances real-time fraud prevention systems.

PROBLEM STATEMENT

The classification of students based on their peer learning participation and study habits is crucial for understanding its impact on academic success. However, manual analysis is time-consuming and may not accurately capture learning behavior patterns.

This project aims to develop a machine learning model to classify students into different categories (e.g., Highly Engaged, Moderately Engaged, and Less Engaged) based on their peer learning participation, study hours, and collaboration frequency.

Objectives:

- Identify the impact of peer learning on academic performance.
- Provide insights for educators to enhance collaborative learning strategies.
- Help students optimize their study methods based on data-driven findings.

By leveraging **classification models,** the system will enable **better student segmentation**, leading to **improved learning** outcomes **and academic success.**

REQUIREMENTS FOR THIS PROJECT:

Software Requirements	
Software	**Purpose**
Operating System	Windows 10 or above
SQL serever Management Services	For schema creation
Visual Studio	For schema deployment
SQL serever Analysis Services	For OLAP operations
Python (Optional)	Data preprocessing or visualization (if needed)

Hardware Requirements		
Component	**Minimum Specification**	**Recommended Specification**
Processor (CPU)	Intel Core i3 (or equivalent)	Intel Core i5/i7 or AMD Ryzen 5/7
RAM	4 GB	8 GB or more
Storage	250 GB HDD	512 GB SSD or higher
Display	1366 x 768 resolution	Full HD (1920 x 1080)

TOOLS	
TOOL	**Purpose**
Orange Tool	For Data Mining

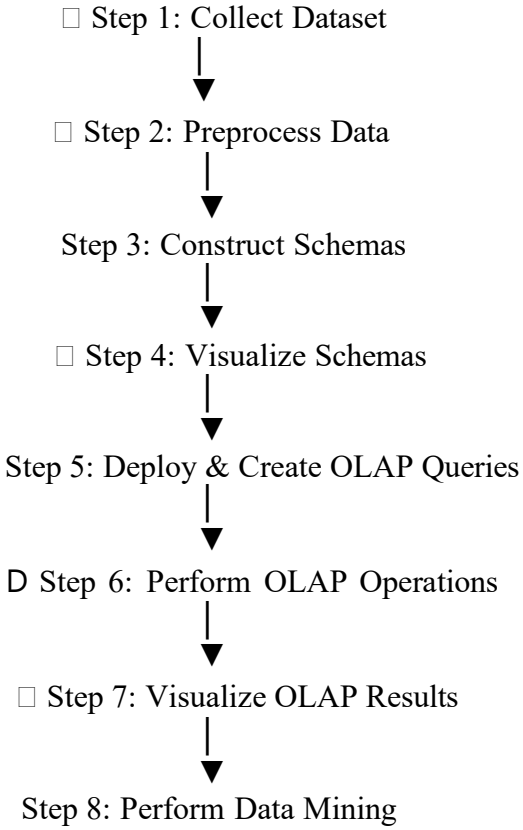
CHAPTER-2: Knowledge Discovery in Databases (KDD) Process

2.1 Problem Statement:

Understanding the impact of peer learning on academic success is essential for enhancing educational strategies and promoting collaborative learning environments. However, manually assessing student participation and its outcomes can be subjective and limited in scope. This project aims to develop a data mining model to analyze and classify students based on their peer learning behaviors and preferences (e.g., Active Collaborators, Occasional Participants, Independent Learners, etc.). By uncovering hidden patterns in student responses, the system will enable educational institutions to promote effective peer learning initiatives, improve student engagement, and support academic performance through data-driven decision-making.

METHODOLOGY:

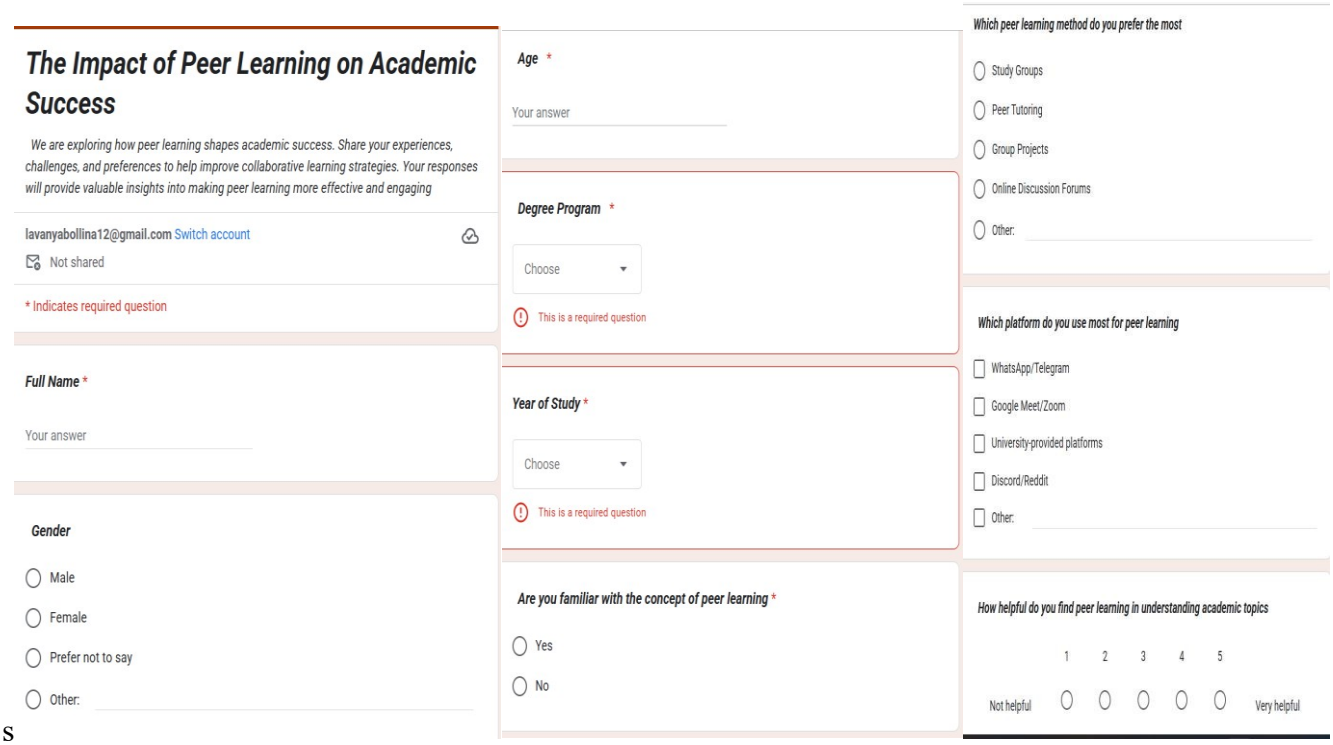
The KDD process is performed in step by step from collection of data set to the classification and developing the prediction model. There are some intermediatory steps in which we created all three schemas with the help of various tools like SSMS(SQL Server Management Services), Visual Studio and SSAS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



STEP-1: COLLECTING & EXPLORING DATASET

1.1. Extracting the Form to Collect Information from Users

The dataset was created by gathering information on peer learning participation and academic performance from various students. Data was collected through surveys and online responses, covering aspects such as study habits, collaboration methods, preferred learning platforms, and challenges faced in peer learning.



s

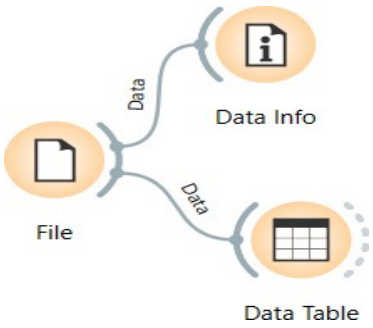
1.2 Defining Survey or Data Collection Methods

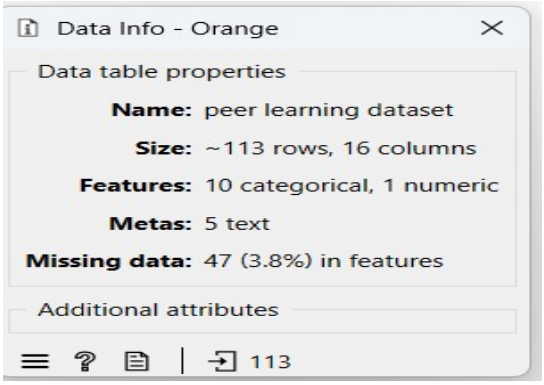
Online Surveys: Conducted through structured questionnaires featuring multiple-choice and rating-scale questions to collect data on user preferences, platform choices, and listening habits.

Link: https://forms.gle/ES8XxS9GF7FQA6oo7

1.3 Choosing Attributes for Analysis

- The key attributes selected for analysis include:
  Peer Learning Participation – Indicates whether a student engages in peer learning.
  Preferred Methods – Identifies if students prefer discussions, group studies, or online collaboration.
  Academic Performance – Measures academic success based on grades or self-assessment.
  Challenges Faced – Highlights difficulties like scheduling conflicts or resource limitations.
  Platforms Used – Specifies whether peer learning occurs on campus, online forums, or study groups.





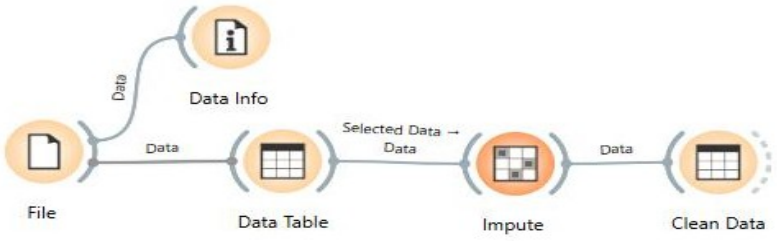
	Timestamp	Full Name	Age	Year of Study	s to improve peer	Gender	Degree Program	with the concept	participate in pee	ng method do yo	do you use most l	er learning in und	do you g
1	2025/02/04 9:0...	laya	18	3	no	Female	B.Tech	Yes	Weekly	Study Groups	WhatsApp/Tele...	5	Better un
2	2025/02/04 9:3...	Likitha	20	3	no	Female	B.Tech	Yes	Monthly	Study Groups	WhatsApp/Tele...	4	Better un
3	2025/02/04 9:4...	Hari Jakku	21	3	nothing	Male	B.Tech	Yes	Monthly	Group Projects	Google Meet/Z...	4	Better un
4	2025/02/04 9:4...	Akunuri Harsha...	18	3	No	Prefer not to say	B.Tech	No	Never	No	I don't know	1	?
5	2025/02/04 9:4...	Tunuguntla Go...	19	3	nothing	Female	B.Tech	Yes	Rarely	Study Groups	WhatsApp/Tele...	4	Better un
6	2025/02/04 9:5...	Revathi	20	3	nothing	Female	B.Tech	Yes	Rarely	Group Projects	WhatsApp/Tele...	4	Better un
7	2025/02/04 9:5...	D.Pujitha	20	3	None	Female	B.Tech	Yes	Weekly	Peer Tutoring	WhatsApp/Tele...	4	Better un
8	2025/02/04 9:5...	G SaiBabu	21	3	nothing	Male	B.Tech	Yes	Rarely	Study Groups	WhatsApp/Tele...	4	Better un
9	2025/02/04 10:...	Banavathu Vasu	19	3	implementing s...	Male	B.Tech	Yes	Daily	Study Groups	WhatsApp/Tele...	5	Improvec
10	2025/02/04 10:...	Pullamma	40	1	creatuve rpojec...	Prefer not to say	B.Tech	No	Never	I don't know ab...	I don't know ab...	1	?
11	2025/02/04 10:...	Nallamolu Nag...	20	3	great iddea	Female	B.Tech	Yes	Weekly	Group Projects	WhatsApp/Tele...	4	Improvec
12	2025/02/04 10:...	G.Lakshmi pras...	20	3	do more pear a...	Female	B.Tech	Yes	Weekly	Peer Tutoring	WhatsApp/Tele...	4	Better un
13	2025/02/04 10:...	B.Rasi	20	2	No	Female	B.Sc	Yes	Weekly	?	WhatsApp/Tele...	2	Better un
14	2025/02/04 10:...	Santi. Murali ...	17	2	nothing more	Male	Other	Yes	Daily	Study Groups	WhatsApp/Tele...	4	Better un
15	2025/02/04 10:...	Lahari	20	3	no	Female	B.Tech	Yes	Monthly	Study Groups	?	4	Better un
16	2025/02/04 10:...	HARITHA SAI	20	3	increasse peer l...	Female	B.Tech	Yes	Weekly	Peer Tutoring	WhatsApp/Tele...	4	Better un
17	2025/02/04 10:...	Shakina	19	3	nothing	Female	B.Tech	Yes	Rarely	Online Discussi...	WhatsApp/Tele...	3	Improvec
18	2025/02/04 11:...	Anand babu	20	3	yes improve act...	Male	B.Tech	No	Rarely	Online Discussi...	?	1	Better un
19	2025/02/05 7:1...	laya	18	3	if you improve i...	Female	B.Tech	Yes	Weekly	Study Groups	WhatsApp/Tele...	5	Better un
20	2025/02/05 7:2...	BHUKYA SAND...	22	3	creative project...	Male	B.Tech	Yes	Weekly	Group Projects	Google Meet/Z...	5	Improvec
21	2025/02/05 7:2...	Layabonam	18	3	Great idea 💡	Female	B.Tech	Yes	Weekly	Group Projects	Discord/Reddit	3	Improvec
22	2025/02/05 7:2...	Sandy	19	4	No	Female	B.Tech	Yes	Rarely	?	University-provi...	3	Reduced
23	2025/02/05 7:2...	Sivanagaraju	21	4	No	Male	B.Com	Yes	Monthly	Group Projects	?	3	?
24	2025/02/05 7:3...	Gajjala	50	4	nothing special ...	Male	Other	Yes	Monthly	?	WhatsApp/Tele...	3	Better un
25	2025/02/05 7:4...	Sandeep S	21	3	yes improve act...	Male	B.Tech	Yes	Rarely	Online Discussi...	WhatsApp/Tele...	1	Better un
26	2025/02/05 7:5...	Chandika Purna...	20	3	no	Male	B.Tech	No	Rarely	Online Discussi...	WhatsApp/Tele...	3	Better un
27	2025/02/05 7:5...	Bskn	55	3	Great idea 💡	Male	B.Tech	Yes	Rarely	Online Discussi...	Discord/Reddit	?	Better un
28	2025/02/05 8:3...	Bindu	19	3	no	Female	B.Tech	Yes	Daily	Study Groups	WhatsApp/Tele...	1	Better un
29	2025/02/05 10:...	B bhavana	21	4	yes improve act...	Female	B.Tech	Yes	Weekly	Group Projects	WhatsApp/Tele...	4	Better un
30	2025/02/05 10:...	Nikhil	19	2	nothing	Male	B.Tech	Yes	Monthly	Study Groups	WhatsApp/Tele...	3	Increasec

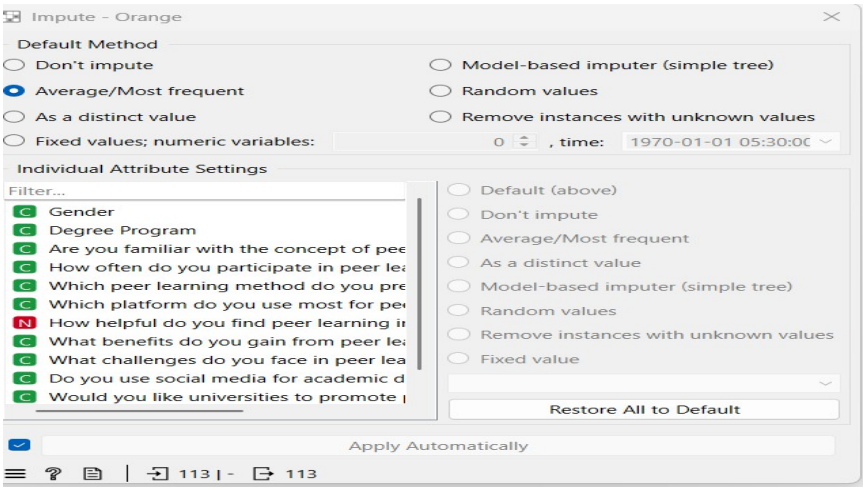
Step-2: PREPROCESS THE DATA

Preprocess the Dataset Using ORANGE TOOL

2.1 Handling Missing Values

- **Numerical values** were filled using the **Average /Most frequent** method.
- **Categorical values** (e.g., subscription type) were filled using the **mode**.
- Records with excessive missing values were removed.





2.2 Data Cleaning & Transformation

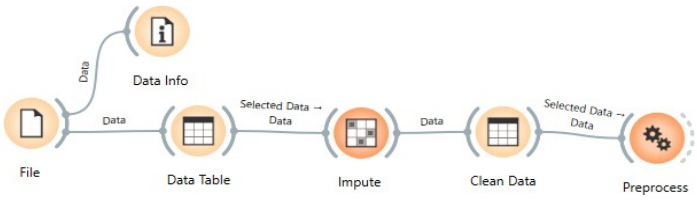
- Standardized text-based attributes (e.g., ensuring "Spotify" is uniformly formatted).
- Converted categorical values into numerical form for analysis.

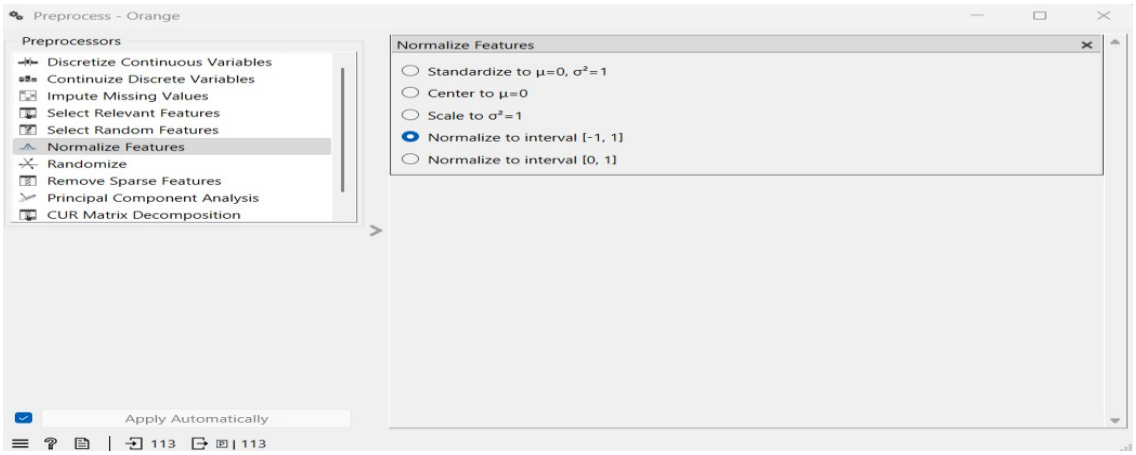
2.3 Removing Duplicates & Inconsistencies

- Removed duplicate survey responses.
- Ensured data consistency and integrity.
- Applied feature selection and dimensionality reduction to optimize the dataset.
- Normalization and encoding of categorical data for better processing.

2.4 Normalization & Encoding

- **Categorical attributes** were encoded (e.g., Free = 0, Premium = 1).
- **Normalization** was applied to standardize numerical values





	Study	s to improve peer	Gender	Degree Program	with the concept	participate in pee	ng method do yo	do you use most	er learning in und	do you gain from	ges do you face in	l media for acade	es to promote pe
1		no	Female	B.Tech	Yes	Weekly	Study Groups	WhatsApp/Tele...	1.0	Better understa...	Time managem...	No	Yes
2		no	Female	B.Tech	Yes	Monthly	Study Groups	WhatsApp/Tele...	0.5	Better understa...	Misinformation ...	Yes	Yes
3		nothing	Male	B.Tech	Yes	Monthly	Group Projects	Google Meet/Z...	0.5	Better understa...	Difficulty findin...	Yes	Yes
4		No	Prefer not to say	B.Tech	No	Never	No	I don't know	-1.0	Better understa...	Time managem...	Yes	No
5		nothing	Female	B.Tech	Yes	Rarely	Study Groups	WhatsApp/Tele...	0.5	Better understa...	Time managem...	Yes	Yes
6		nothing	Female	B.Tech	Yes	Rarely	Group Projects	WhatsApp/Tele...	0.5	Better understa...	Difficulty findin...	Yes	Yes
7		None	Female	B.Tech	Yes	Weekly	Peer Tutoring	WhatsApp/Tele...	0.5	Better understa...	Time managem...	Yes	Yes
8		nothing	Male	B.Tech	Yes	Rarely	Study Groups	WhatsApp/Tele...	0.5	Better understa...	Time managem...	Yes	Yes
9		implementing s...	Male	B.Tech	Yes	Daily	Study Groups	WhatsApp/Tele...	1.0	Improved probl...	Difficulty findin...	Yes	Yes
10		creatuve rpojec...	Prefer not to say	B.Tech	No	Never	I don't know ab...	I don't know ab...	-1.0	Better understa...	I don't know ab...	Yes	No
11		great iddea	Female	B.Tech	Yes	Weekly	Group Projects	WhatsApp/Tele...	0.5	Improved probl...	Lack of commit...	Yes	Yes
12		do more pear a...	Female	B.Tech	Yes	Weekly	Peer Tutoring	WhatsApp/Tele...	0.5	Better understa...	Time managem...	Yes	Yes
13		No	Female	B.Sc	Yes	Weekly	Study Groups	WhatsApp/Tele...	-0.5	Better understa...	Time managem...	Yes	Yes
14		nothing more	Male	Other	Yes	Daily	Study Groups	WhatsApp/Tele...	0.5	Better understa...	Lack of commit...	Yes	Yes
15		no	Female	B.Tech	Yes	Monthly	Study Groups	WhatsApp/Tele...	0.5	Better understa...	Time managem...	No	Yes
16		increasse peer l...	Female	B.Tech	Yes	Weekly	Peer Tutoring	WhatsApp/Tele...	0.5	Better understa...	Difficulty findin...	Yes	Yes
17		nothing	Female	B.Tech	Yes	Rarely	Online Discussi...	WhatsApp/Tele...	0.0	Improved probl...	Lack of commit...	Yes	No
18		yes improve act...	Male	B.Tech	No	Rarely	Online Discussi...	WhatsApp/Tele...	-1.0	Better understa...	Difficulty findin...	Yes	Yes
19		if you improve i...	Female	B.Tech	Yes	Weekly	Study Groups	WhatsApp/Tele...	1.0	Better understa...	Lack of commit...	Yes	Yes
20		creative project...	Male	B.Tech	Yes	Weekly	Group Projects	Google Meet/Z...	1.0	Improved probl...	Time managem...	Yes	Yes
21		Great idea 💡	Female	B.Tech	Yes	Weekly	Group Projects	Discord/Reddit	0.0	Improved probl...	Misinformation ...	Yes	Yes
22		No	Female	B.Tech	Yes	Rarely	Study Groups	University-provi...	0.0	Reduced stress ...	Time managem...	Yes	Yes
23		No	Male	B.Com	Yes	Monthly	Group Projects	WhatsApp/Tele...	0.0	Better understa...	Misinformation ...	No	Yes
24		nothing special ...	Male	Other	Yes	Monthly	Study Groups	WhatsApp/Tele...	0.0	Better understa...	Difficulty findin...	Yes	Yes
25		yes improve act...	Male	B.Tech	Yes	Rarely	Online Discussi...	WhatsApp/Tele...	-1.0	Better understa...	Time managem...	Yes	No
26		no	Male	B.Tech	No	Rarely	Online Discussi...	WhatsApp/Tele...	0.0	Better understa...	Difficulty findin...	No	Yes
27		Great idea 💡	Male	B.Tech	Yes	Rarely	Online Discussi...	Discord/Reddit	0.360	Better understa...	Misinformation ...	Yes	Yes
28		no	Female	B.Tech	Yes	Daily	Study Groups	WhatsApp/Tele...	-1.0	Better understa...	Difficulty findin...	Yes	No
29		yes improve act...	Female	B.Tech	Yes	Weekly	Group Projects	WhatsApp/Tele...	0.5	Better understa...	Difficulty findin...	Yes	Yes
30		nothing	Male	B.Tech	Yes	Monthly	Study Groups	WhatsApp/Tele...	0.0	Increased enga...	Difficulty findin...	Yes	Yes

Step 3:

Fields or attributes for my dataset are

- ***Full Name***
- ***Gender***
- ***Age***
- ***Degree Program***
- ***Year of Study***
- ***Are you familiar with the concept of peer learning***
- ***How often do you participate in peer learning activities***
- ***Which peer learning method do you prefer the most***
- ***Which platform do you use most for peer learning***
- ***How helpful do you find peer learning in understanding academic topics***
- ***What benefits do you gain from peer learning***
- ***What challenges do you face in peer learning***
- ***Do you use social media for academic discussions***
- ***Would you like universities to promote peer learning more actively***
- ***Do you have any suggestions to improve peer learning at your institution***

Schema Construction by Normalizing the Dataset (Using Database Engine)

- After Normalizing the following tables are identified:
- **Dimension Tables:**

  1.student dim

`Student_Dim`

Attribute	Data Type	Key
student_id	INT	Primary Key
full_name	VARCHAR(50)	
gender	VARCHAR(10)	
degree_program_id	INT	Foreign Key
year_of_study	INT	

2. Method_dim

Attribute	Data Type	Key
method_id	INT	Primary Key
method_name	VARCHAR(50)	

3. Time_dim

Attribute	Data Type	Key
time_id	INT	Primary Key
full_date	DATE	
year	INT	
month	VARCHAR(10)	
day_of_week	VARCHAR(10)	

4. platform_dim:

Attribute	Data Type	Key
platform_id	INT	Primary Key
platform_name	VARCHAR(50)	

5. Benefits_dim

Attribute	Data Type	Key
benefit_id	INT	Primary Key
benefit_description	VARCHAR(100)	

challenges_dim

Attribute	Data Type	Key
challenge_id	INT	Primary Key
challenge_description	VARCHAR(100)	

- **Sub-Dimension tables/Extended tables:**

1.Degree_program_dim(from student_dim)

Attribute	Data Type	Key
degree_program_id	INT	Primary Key
degree_program	VARCHAR(50)	

- **Fact tables:**

1. Peer_learning_fact

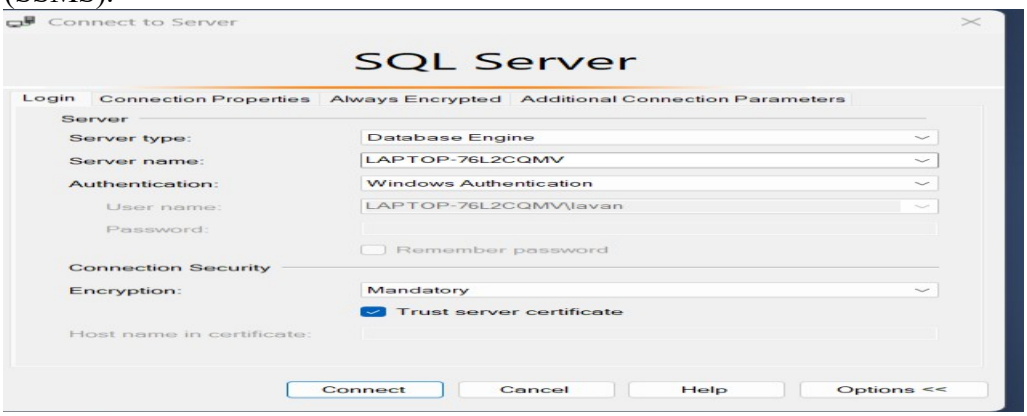
Attribute	Data Type	Key
learning_id	INT	Primary Key
student_id	INT	Foreign Key
time_id	INT	Foreign Key
method_id	INT	Foreign Key
total_sessions	INT	
study_hours	FLOAT	

2. student_performance_fact

Attribute	Data Type	Key
performance_id	INT	Primary Key
student_id	INT	Foreign Key
time_id	INT	Foreign Key
platform_id	INT	Foreign Key
total_study_hours	FLOAT	
exam_score	FLOAT	
grade	VARCHAR(5)	
benefit_id	INT	Foreign Key
challenge_id	INT	Foreign Key

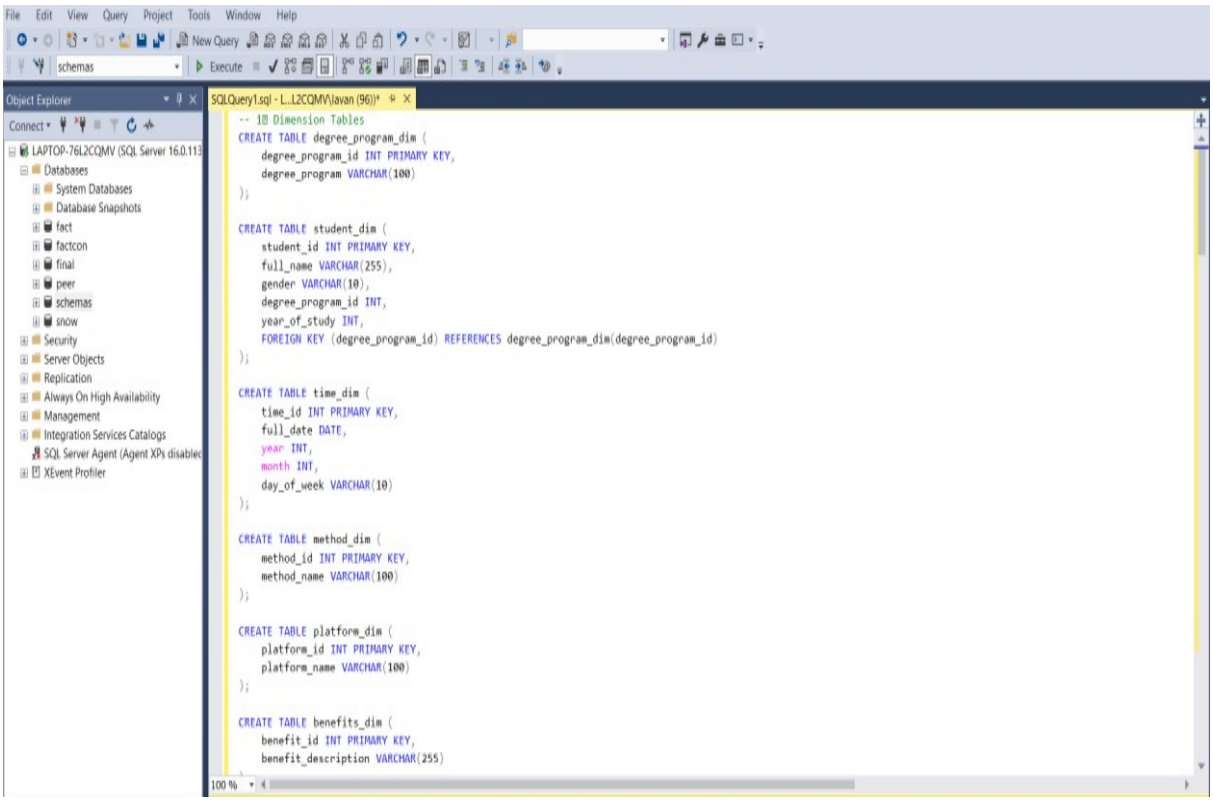
- Now Create a database “peer” to insert all these tables.

Generate SQL queries to create and insert data into all the tables in SQL Server Database Engine (SSMS).



Database Engine

SQL queries inserted for schema creation

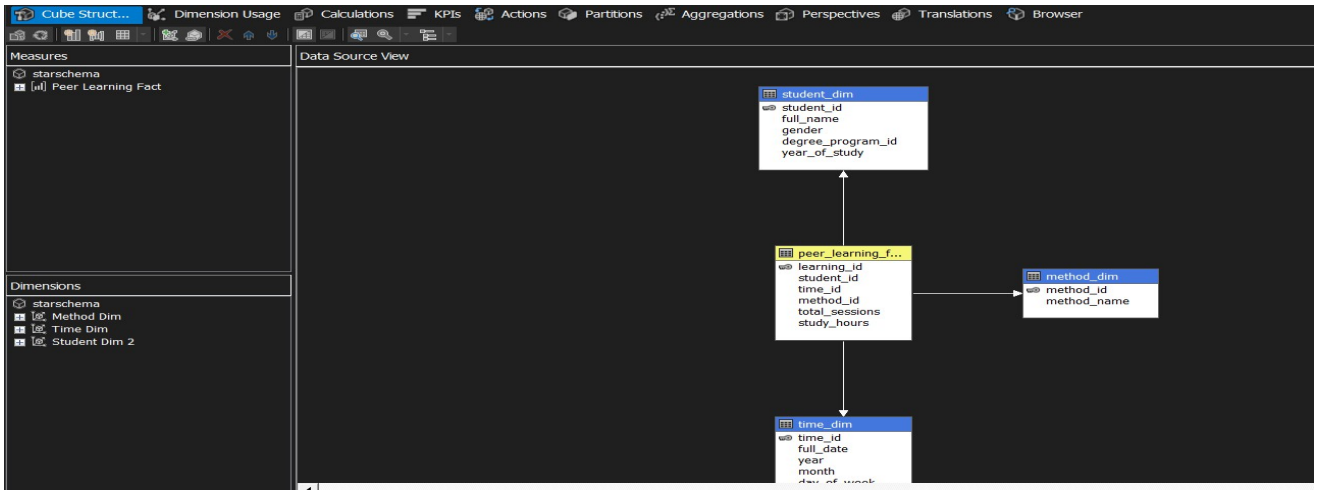


Step 4: Schema Visualization in Visual Studio

- Create analysis service multidimensional project in Visual Studio.
- Create Data Sources & Data Source Views by Connected the database.
- View database diagrams for schemas and verify relationships between tables.
- Create Multidimensional cubes for schema.

These are the schemas

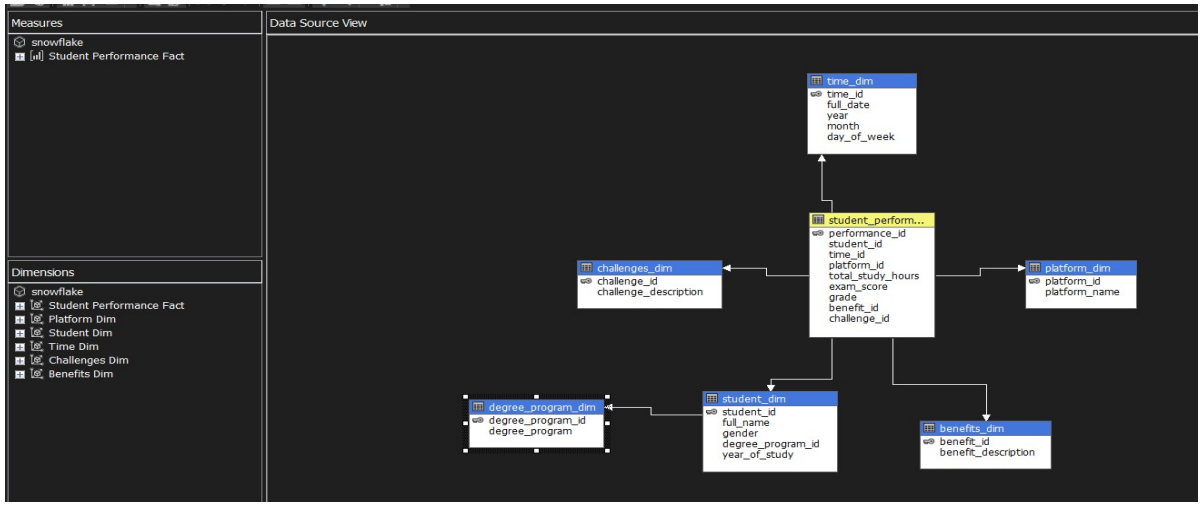
Star Schema:



Star schema has one Fact table and three Dimension tables.

- Fact table is Streaming_Fact.
- Dimension tables are student _Dim, method_Dim and time_Dim.

Snowflake Schema:



Snowflake schema has one Fact table, five Dimension tables and one Sub-Dimension tables.
student_dim,time_dim,platform_dim,Challenges_dim, Benefits_dim

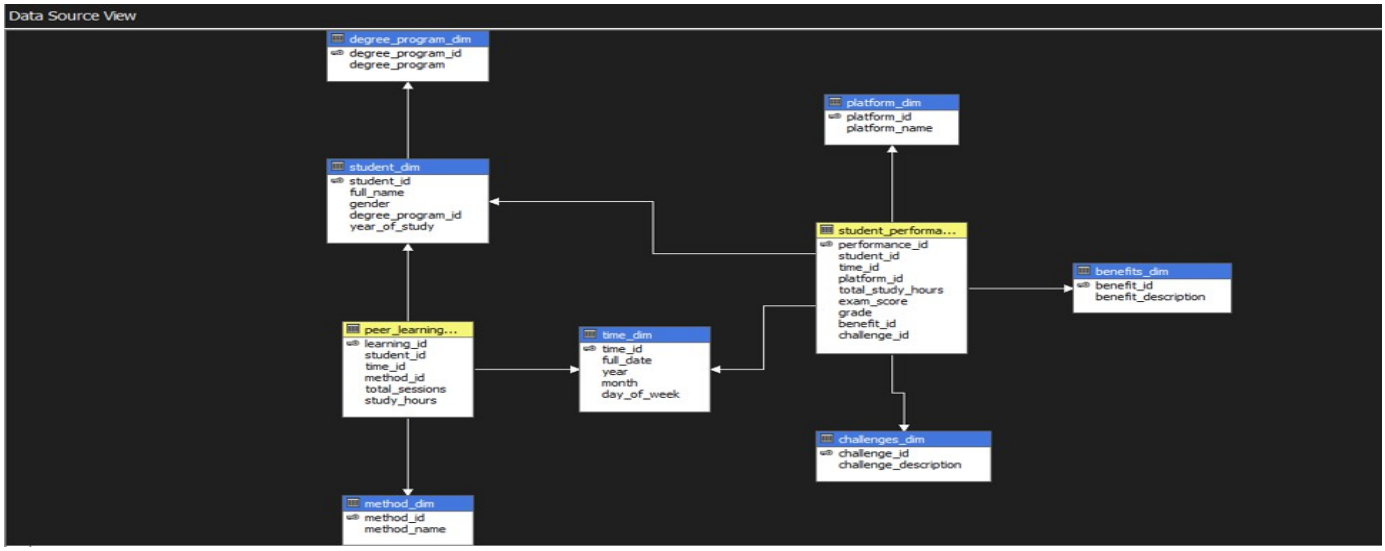
❖ **Extended tables :**

1)Degree_Program_dim

Fact table :

1)student_Performance_fact

Fact Constellation:

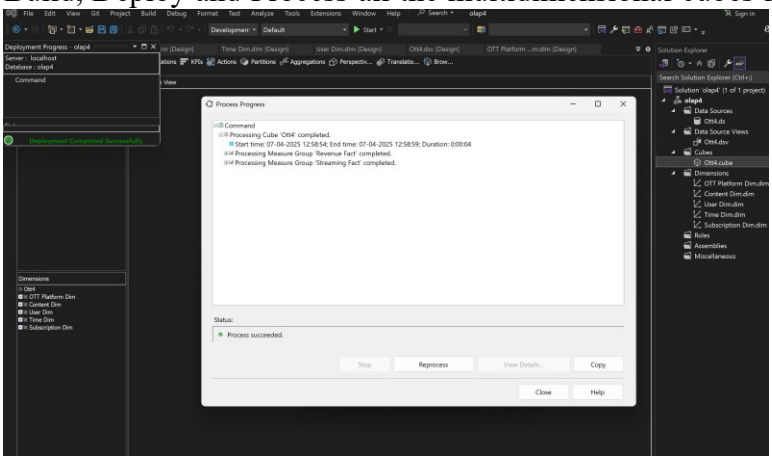


Fact Constellation has two Fact tables, six Dimension tables and one Sub-Dimension tables.

- Fact tables are Peer_Learning_Fact and Student_Performance_Fact
- Dimension tables are student_dim,Time_dim, Method_dim,Platform_dim,Challenges_dim, Benefits_dim, Degree_Program_dim

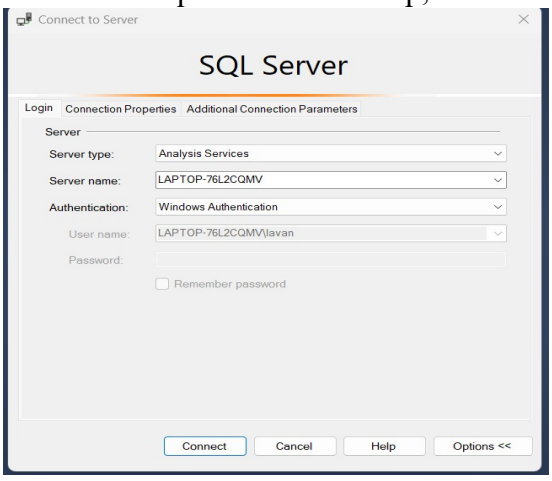
Step 5: Build, Deploy and Process Multidimensional Cubes & Create OLAP Operations.

- Build, Deploy and Process all the multidimensional cubes in visual studio.



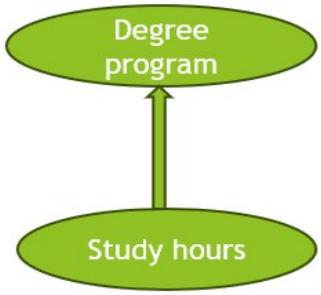
Build, Deploy and Processing Multidimensional Cubes

- Write MDX queries for Roll-Up, Drill-Down, Slice, Dice and Pivot operations in SSAS.



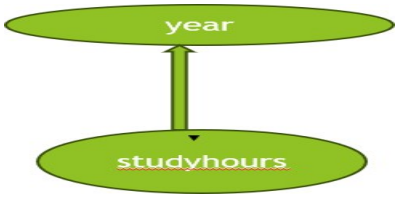
Analysis Services

- A **concept hierarchy** defines levels of abstraction in a dimension. It allows **attributes to be organized from low-level to high-level**, enabling data to be viewed at different levels of granularity.
- Concept hierarchies are **essential in data warehousing schemas** (like **star**, **snowflake**, and **fact constellation**) to support **OLAP operations** such as **roll-up**, **drill-down**, **slice**, and **dice** effectively.
- These are the concept hierarchies used to perform OLAP operations.



Degree_program hierarchy

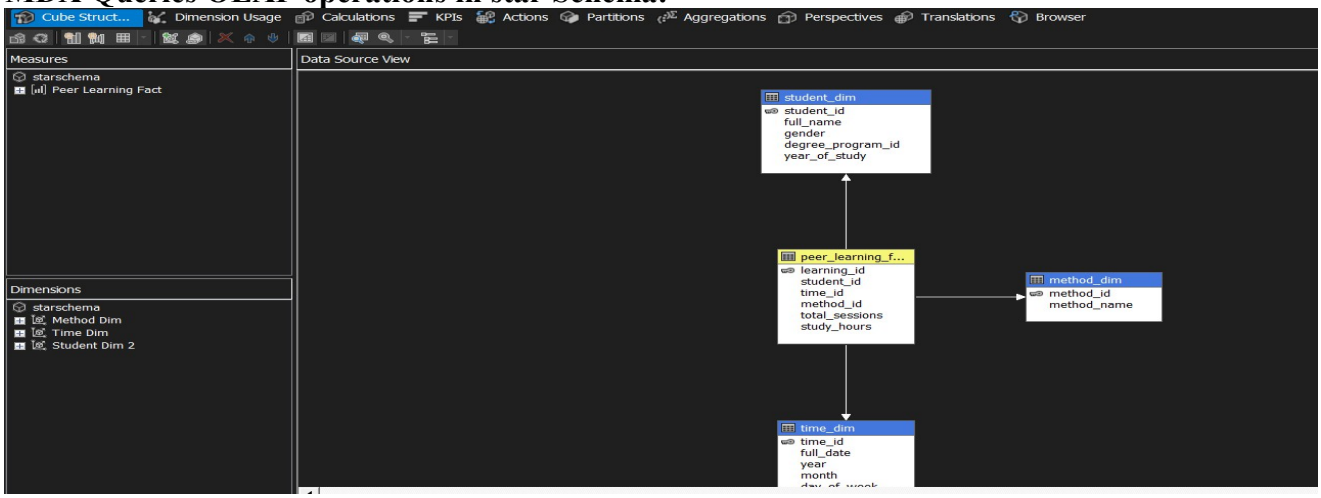
Time hierarchy



Year hierarchy

Step 6: Perform OLAP Operations using MDX (Multidimensional Expressions) Queries.

MDX Queries OLAP operations in star Schema:



(A) How can we analyze the total study hours by aggregating data across different years and years of study?

Concept Hierarchy Used: 🎓 Student Dimension → {Year of Study → Semester → Course

ROLLUP:

Query:

```
SELECT
  [Measures].[Study Hours] ON COLUMNS,
  NONEMPTY(
    CROSSJOIN(
```

```
    EXCEPT([Time Dim].[Year].MEMBERS, {[Time Dim].[Year].[All], [Time Dim].[Year].[Unknown]}),
    CROSSJOIN(
      EXCEPT([Student Dim 2].[Year Of Study].MEMBERS, {[Student Dim 2].[Year Of Study].[All],
[Student Dim 2].[Year Of Study].[Unknown]}),
      EXCEPT([Method Dim].[Method Name].MEMBERS, {[Method Dim].[Method Name].[All], [Method
Dim].[Method Name].[Unknown]})
    )
  )
 ) ON ROWS
FROM [starschema]
```

Output:

			Study Hours
2024	1	Coding Competitions	13
2024	1	Study Groups	14.2
2024	1	Webinars	15.6
2024	2	Group Discussions	12.5
2024	2	One-on-One Tutoring	9.5
2024	2	Online Courses	18.2
2024	3	Hackathons	11.3
2024	3	Video Tutorials	12
2024	3	Workshops	8
2024	4	Interactive Quizzes	13.9
2024	4	Online Forums	16.8
2024	4	Seminars	7.8
2025	1	Hands-on Labs	8.6
2025	2	Research Collaboration	9.7
2025	3	Case Studies	11.5

Execution Time: 5 ms

(B) How can we break down study hours further by showing detailed data for each month within a given year?

Concept hierarchy: Year → Month→studyhours

DRILL DOWN

Query:

```
SELECT
  [Measures].[Study Hours] ON COLUMNS,
  NONEMPTY(
    CROSSJOIN(
      CROSSJOIN(
        EXCEPT([Time Dim].[Year].MEMBERS, {[Time Dim].[Year].[All], [Time
Dim].[Year].[Unknown]}),
        EXCEPT([Time Dim].[Month].MEMBERS, {[Time Dim].[Month].[All], [Time
Dim].[Month].[Unknown]})
      ),
      CROSSJOIN(
      CROSSJOIN(
          EXCEPT([Student Dim 2].[Year Of Study].MEMBERS, {[Student Dim 2].[Year Of
Study].[All], [Student Dim 2].[Year Of Study].[Unknown]}),
          EXCEPT([Student Dim 2].[Degree Program Id].MEMBERS, {[Student Dim 2].[Degree
Program Id].[All], [Student Dim 2].[Degree Program Id].[Unknown]})
        ),
        EXCEPT([Method Dim].[Method Name].MEMBERS, {[Method Dim].[Method Name].[All],
[Method Dim].[Method Name].[Unknown]})
      )
    )
  ) ON ROWS
FROM [starschema]
```

OUTPUT:

					Study Hours
2024	1	2	1	Group Discussions	12.5
2024	10	3	10	Video Tutorials	12
2024	11	1	11	Webinars	15.6
2024	12	4	12	Interactive Quizzes	13.9
2024	2	3	2	Workshops	8
2024	3	1	3	Study Groups	14.2
2024	4	4	4	Online Forums	16.8
2024	5	2	5	One-on-One Tutoring	9.5
2024	6	3	6	Hackathons	11.3
2024	7	1	7	Coding Competitions	13
2024	8	4	8	Seminars	7.8
2024	9	2	9	Online Courses	18.2
2025	1	2	13	Research Collaboration	9.7
2025	2	3	14	Case Studies	11.5
2025	3	1	15	Hands-on Labs	8.6

Execution time:4ms

C) How can we retrieve the total study hours for students in their third year while focusing on different study methods?

SLICE:

Query:

```
SELECT
  [Measures].[Study Hours] ON COLUMNS,
  NONEMPTY(
    EXCEPT([Method Dim].[Method Name].MEMBERS, {[Method Dim].[Method Name].[All], [Method Dim].[Method Name].[Unknown]})
  ) ON ROWS
FROM [starschema]
WHERE ([Student Dim 2].[Year Of Study].[3])
```

OUTPUT:

	Study Hours
Case Studies	11.5
Hackathons	11.3
Video Tutorials	12
Workshops	8

Execution time:6ms

D) How can we analyze study hours by different study methods and years while considering only third-year students?

DICE:

Query:

```
SELECT
  [Measures].[Study Hours] ON COLUMNS,
  NONEMPTY(
    CROSSJOIN(
      EXCEPT([Method Dim].[Method Name].MEMBERS, {[Method Dim].[Method Name].[All], [Method Dim].[Method Name].[Unknown]}),
      EXCEPT([Time Dim].[Year].MEMBERS, {[Time Dim].[Year].[All], [Time Dim].[Year].[Unknown]})
    )
  ) ON ROWS
FROM [starschema]
WHERE ([Student Dim 2].[Year Of Study].[3])
```

		Study Hours
Case Studies	2025	11.5
Coding Competitions	2024	13
Group Discussions	2024	12.5
Hackathons	2024	11.3
Hands-on Labs	2025	8.6
Interactive Quizzes	2024	13.9
One-on-One Tutoring	2024	9.5
Online Courses	2024	18.2
Online Forums	2024	16.8
Research Collaboration	2025	9.7
Seminars	2024	7.8
Study Groups	2024	14.2
Video Tutorials	2024	12
Webinars	2024	15.6
Workshops	2024	8

Execution Time:8ms

E) How can we compare the total study sessions across different days of the week and study methods?

PIVOT:

Query:

```
SELECT
  NONEMPTY(
    EXCEPT([Time Dim].[Day Of Week].MEMBERS, {[Time Dim].[Day Of Week].[All]}),
    [Measures].[Total Sessions]
  ) ON COLUMNS,
  NONEMPTY(
    EXCEPT([Method Dim].[Method Name].MEMBERS, {[Method Dim].[Method Name].[All]}),
    [Measures].[Total Sessions]
  ) ON ROWS
FROM [starschema]
```

OUTPUT:

	Friday	Monday	Saturday	Sunday	Thursday	Tuesday	Wednesday
Case Studies	(null)	(null)	(null)	(null)	5	(null)	(null)
Coding Competitions	6	(null)	(null)	(null)	(null)	(null)	(null)
Group Discussions	5	(null)	(null)	(null)	(null)	(null)	(null)
Hackathons	(null)	(null)	(null)	5	(null)	(null)	(null)
Hands-on Labs	3	(null)	(null)	(null)	(null)	(null)	(null)
Interactive Quizzes	(null)	(null)	(null)	(null)	(null)	6	(null)
One-on-One Tutoring	(null)	(null)	4	(null)	(null)	(null)	(null)
Online Courses	(null)	(null)	(null)	8	(null)	(null)	(null)
Online Forums	(null)	(null)	7	(null)	(null)	(null)	(null)
Research Collaboration	(null)	(null)	(null)	(null)	(null)	(null)	4
Seminars	(null)	(null)	3	(null)	(null)	(null)	(null)
Study Groups	6	(null)	(null)	(null)	(null)	(null)	(null)
Video Tutorials	(null)	(null)	(null)	5	(null)	(null)	(null)
Webinars	(null)	7	(null)	(null)	(null)	(null)	(null)
Workshops	(null)	(null)	3	(null)	(null)	(null)	(null)

Execution Time:7 ms

Step 7:

4.1.5 Visualize OLAP Results

Prepare OLAP Output for Visualization

- Select key OLAP operation results related to study hours.
- Export the selected data as an Excel sheet for further visualization.
- Ensure the dataset includes relevant attributes such as Year, Month, Study Hours,

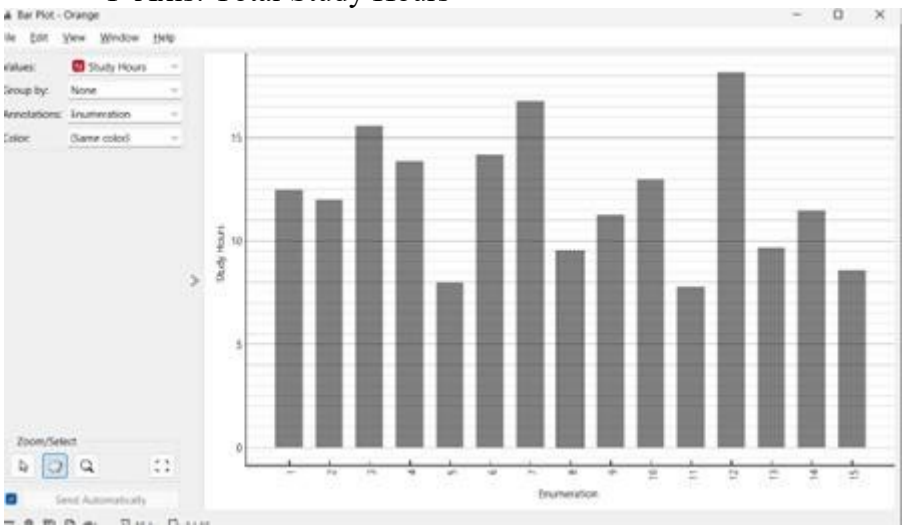
Monthly Study Hours Breakdown: (**study_hours_drill.csv**)

- The data is structured to show study hours aggregated by year and month.

		Study Hours
2024	1	12.5
2024	10	12
2024	11	15.6
2024	12	13.9
2024	2	8
2024	3	14.2
2024	4	16.8
2024	5	9.5
2024	6	11.3
2024	7	13
2024	8	7.8
2024	9	18.2
2025	1	9.7
2025	2	11.5
2025	3	8.6

Bar Chart Configuration:

- X-Axis: Year and Month
- Y-Axis: Total Study Hours

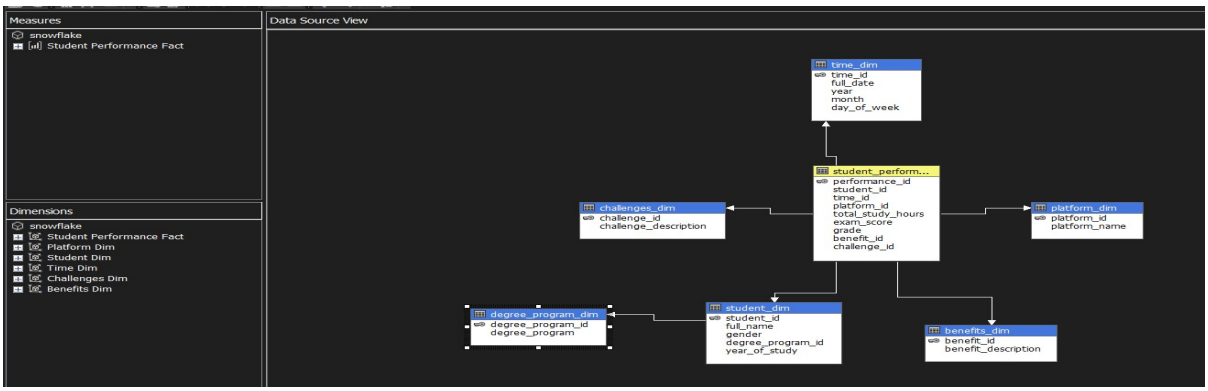


4.1 SNOWFLAKE SCHEMA:

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** (e.g.,exam_score,grade)
- Identify **Dimension Tables** (e.g.,student,benefits,platfrom..etc).
- Normalize dimension tables by breaking them into **sub-dimensions** (e.g., student → degree_program).
- Ensure **foreign key relationships** between tables.



MDX Queries OLAP operations in SNOWFLAKE SCHEMA:

A) How can we analyze the total study hours by aggregating data across different challenges and degree programs?

ROLL UP:

Query:

```
SELECT
  [Measures].[Total Study Hours] ON COLUMNS,
  NONEMPTY(
    CROSSJOIN(
      EXCEPT([Time Dim].[Year].MEMBERS, {[Time Dim].[Year].[All]}),
      CROSSJOIN(
        EXCEPT([Student Dim].[Year Of Study].MEMBERS, {[Student Dim].[Year Of Study].[All]}),
        CROSSJOIN(
          EXCEPT([Platform Dim].[Platform Name].MEMBERS, {[Platform Dim].[Platform Name].[All]}),
          CROSSJOIN(
            EXCEPT([Challenges Dim].[Challenge Description].MEMBERS, {[Challenges Dim].[Challenge
Description].[All]}),
            EXCEPT([Benefits Dim].[Benefit Description].MEMBERS, {[Benefits Dim].[Benefit
Description].[All]})
          )
        )
      )
    )
  ) ON ROWS
FROM [snowflake]
```

OUTPUT:

					Total Study Hours
2024	1	Khan Academy	Low Engagement	Increased Motivation	13
2024	1	Pluralsight	Technical Difficulties	Career Opportunities	15.6
2024	1	Slack	Scheduling Conflicts	Higher Grades	14.2
2024	2	Coursera	Distractions	Enhanced Problem-Solving	9.5
2024	2	Google Classroom	Internet Connectivity Issues	Improved Understanding	12.5
2024	2	YouTube	Different Learning Paces	Confidence Boost	18.2
2024	3	edX	Lack of Time	Better Communication	11.3
2024	3	LinkedIn Learning	Overloaded Curriculum	Skill Development	12
2024	3	Microsoft Teams	Difficulty in Understanding	Better Retention	8
2024	4	Skillshare	Lack of Study Materials	Time Management	13.9
2024	4	Udemy	Communication Barriers	Exposure to New Ideas	7.8
2024	4	Zoom	Lack of Motivation	Networking Opportunities	16.8
2025	1	Reddit Study Groups	Difficulty in Applying Concepts	Critical Thinking	8.6
2025	2	Peer-to-Peer Networks	Exam Anxiety	Collaboration Skills	9.7
2025	3	GitHub Discussions	Limited Peer Support	Self-Learning Ability	11.5

Excecution Time:9 ms

B) How can we examine exam scores in more detail by breaking them down across years, months, and learning platforms?

Concept hierarchy: Year ⟶ Month ⟶ platfromname→examscore

DRILLDOWN:

Query:

```
SELECT
  [Measures].[Exam Score] ON COLUMNS,
  NONEMPTY(
    CROSSJOIN(
      CROSSJOIN(
        EXCEPT([Time Dim].[Year].MEMBERS, {[Time Dim].[Year].[All]}),
        EXCEPT([Time Dim].[Month].MEMBERS, {[Time Dim].[Month].[All]})
      ),
```

```
    CROSSJOIN(
      CROSSJOIN(
        EXCEPT([Student Dim].[Year Of Study].MEMBERS, {[Student Dim].[Year Of Study].[All]}),
        EXCEPT([Student Dim].[Degree Program Id].MEMBERS, {[Student Dim].[Degree Program
Id].[All]})
      ),
      CROSSJOIN(
        EXCEPT([Platform Dim].[Platform Name].MEMBERS, {[Platform Dim].[Platform Name].[All]}),
        CROSSJOIN(
          EXCEPT([Challenges Dim].[Challenge Description].MEMBERS, {[Challenges Dim].[Challenge
Description].[All]}),
          EXCEPT([Benefits Dim].[Benefit Description].MEMBERS, {[Benefits Dim].[Benefit
Description].[All]})
        )
      )
    )
  )
) ON ROWS
FROM [snowflake]
```

OUTPUT:

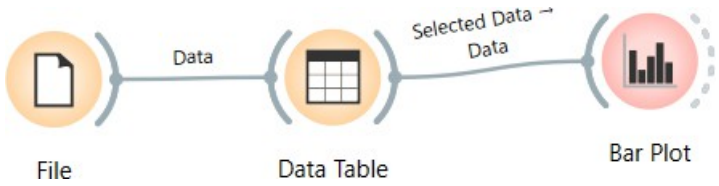
							Exam Score
2024	1	2	1	Google Classroom	Internet Connectivity Issues	Improved Understanding	85.5
2024	10	3	10	LinkedIn Learning	Overloaded Curriculum	Skill Development	81.7
2024	11	1	11	Pluralsight	Technical Difficulties	Career Opportunities	89.2
2024	12	4	12	Skillshare	Lack of Study Materials	Time Management	86
2024	2	3	2	Microsoft Teams	Difficulty in Understanding	Better Retention	78
2024	3	1	3	Slack	Scheduling Conflicts	Higher Grades	92
2024	4	4	4	Zoom	Lack of Motivation	Networking Opportunities	88.5
2024	5	2	5	Coursera	Distractions	Enhanced Problem-Solving	74
2024	6	3	6	edX	Lack of Time	Better Communication	80.2
2024	7	1	7	Khan Academy	Low Engagement	Increased Motivation	85
2024	8	4	8	Udemy	Communication Barriers	Exposure to New Ideas	70.5
2024	9	2	9	YouTube	Different Learning Paces	Confidence Boost	95.3
2025	1	2	13	Peer-to-Peer Networks	Exam Anxiety	Collaboration Skills	76.3
2025	2	3	14	GitHub Discussions	Limited Peer Support	Self-Learning Ability	82.8
2025	3	1	15	Reddit Study Groups	Difficulty in Applying Concepts	Critical Thinking	72.1

Execution Time:5 ms

Step 7:

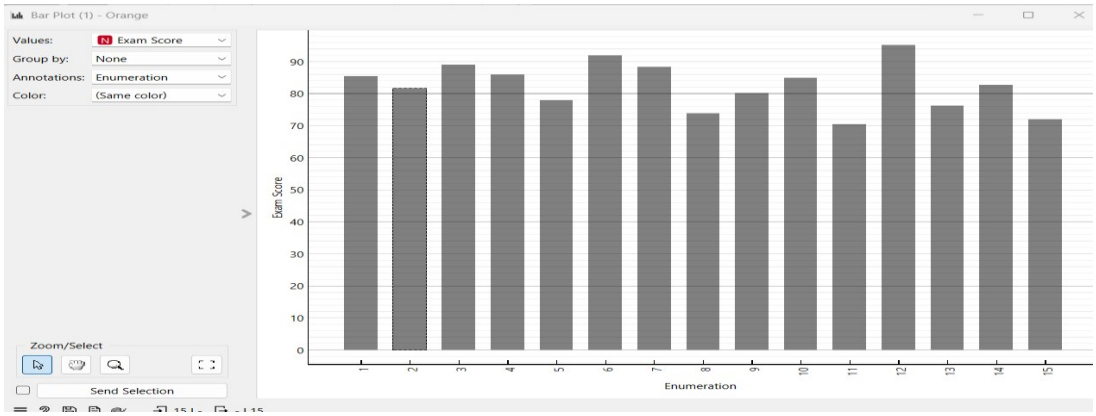
Visualize OLAP Results:

Visualizing exam score Using Orange Tool



Bar Chart Configuration:

- **X-Axis:** Year and Month
- **Y-Axis:** exam scoring



C) How can we retrieve the total study hours for different learning platforms while focusing only on the year 2024?

SLICE:

Query:

```
SELECT
  [Measures].[Total Study Hours] ON COLUMNS,
[Platform Dim].[Platform Name].MEMBERS ON ROWS
FROM [snowflake]
WHERE ([Time Dim].[Year].[2024])
```

OUTPUT:

Messages Results

	Total Study Hours
All	152.8
Coursera	9.5
edX	11.3
GitHub Discussions	(null)
Google Classroom	12.5
Khan Academy	13
LinkedIn Learning	12
Microsoft Teams	8
Peer-to-Peer Networks	(null)
Pluralsight	15.6
Reddit Study Groups	(null)
Skillshare	13.9
Slack	14.2
Udemy	7.8
YouTube	18.2
Zoom	16.8

Execution Time:5 ms

D) How can we analyze exam scores based on different learning platforms and the benefits they provide, while considering only third-year students?

DICE:

Query:

```
SELECT
  [Measures].[Exam Score] ON COLUMNS,
  NONEMPTY(
    CROSSJOIN(
      EXCEPT([Platform Dim].[Platform Name].MEMBERS, {[Platform Dim].[Platform Name].[All]}),
      EXCEPT([Benefits Dim].[Benefit Description].MEMBERS, {[Benefits Dim].[Benefit
Description].[All]})
    ), [Measures].[Exam Score]  -- Ensures only meaningful data
  ) ON ROWS
FROM [snowflake]
WHERE ([Student Dim].[Year Of Study].[3])
```

OUTPUT:

Messages Results

		Exam Score
edX	Better Communication	80.2
GitHub Discussions	Self-Learning Ability	82.8
LinkedIn Learning	Skill Development	81.7
Microsoft Teams	Better Retention	78

Execution time:6 ms

E) How can we compare challenges based on different days of the week?

PIVOT

Query:

```
SELECT
  NONEMPTY(
    EXCEPT([Time Dim].[Day Of Week].MEMBERS, {[Time Dim].[Day Of Week].[All]})
  ) ON COLUMNS,
  NONEMPTY(
    EXCEPT([Challenges Dim].[Challenge Description].MEMBERS, {[Challenges Dim].[Challenge
Description].[All]})
  ) ON ROWS
FROM [snowflake]
```

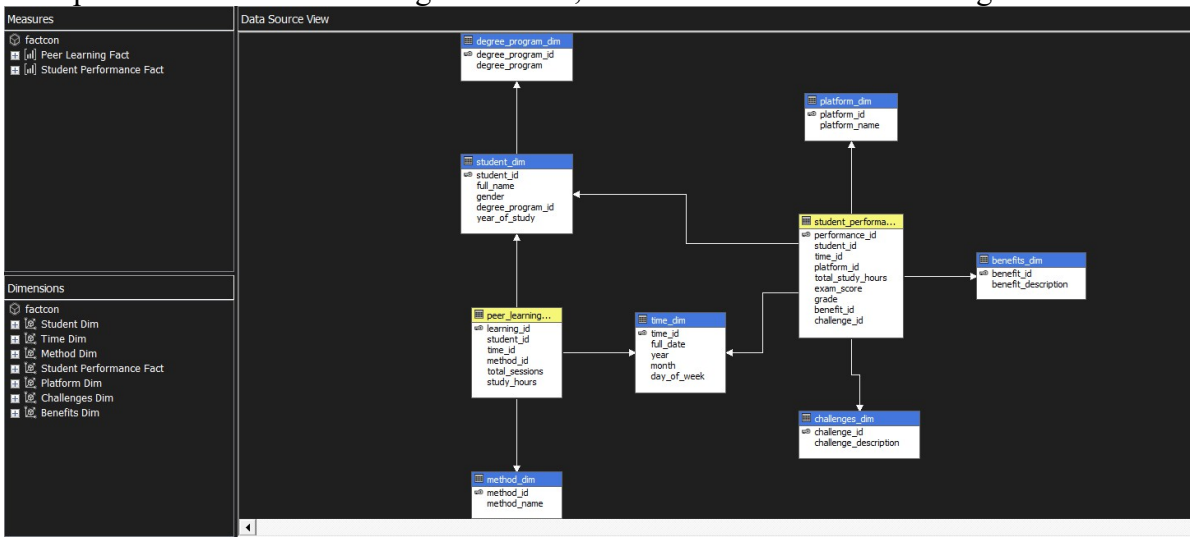
OUTPUT:

	Friday	Monday	Saturday	Sunday	Thursday	Tuesday	Wednesday
Communication Barriers	(null)	(null)	7.8	(null)	(null)	(null)	(null)
Different Learning Paces	(null)	(null)	(null)	18.2	(null)	(null)	(null)
Difficulty in Applying Concepts	8.6	(null)	(null)	(null)	(null)	(null)	(null)
Difficulty in Understanding	(null)	(null)	8	(null)	(null)	(null)	(null)
Distractions	(null)	(null)	9.5	(null)	(null)	(null)	(null)
Exam Anxiety	(null)	(null)	(null)	(null)	(null)	(null)	9.7
Internet Connectivity Issues	12.5	(null)	(null)	(null)	(null)	(null)	(null)
Lack of Motivation	(null)	(null)	16.8	(null)	(null)	(null)	(null)
Lack of Study Materials	(null)	(null)	(null)	(null)	(null)	13.9	(null)
Lack of Time	(null)	(null)	(null)	11.3	(null)	(null)	(null)
Limited Peer Support	(null)	(null)	(null)	(null)	11.5	(null)	(null)
Low Engagement	13	(null)	(null)	(null)	(null)	(null)	(null)
Overloaded Curriculum	(null)	(null)	(null)	12	(null)	(null)	(null)
Scheduling Conflicts	14.2	(null)	(null)	(null)	(null)	(null)	(null)
Technical Difficulties	(null)	15.6	(null)	(null)	(null)	(null)	(null)

Execution time:11 ms

4.2 FACT 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.



MDX Queries OLAP operations in FACTCONSTELLATION SCHEMA:

(A) How can we analyze overall study performance by aggregating total study hours and exam scores across different years and degree programs?

Concept Hierarchy: Year → Degree Program → examscore→studyhours

ROLL UP:

Query:

```
SELECT
  {[Measures].[Exam Score], [Measures].[Total Study Hours]} ON COLUMNS,
  NONEMPTY(
    CROSSJOIN(
      EXCEPT([Time Dim].[Year].MEMBERS, {[Time Dim].[Year].[All]}),
      EXCEPT([Student Dim].[Degree Program].MEMBERS, {[Student Dim].[Degree Program].[All]})
    )
  ) ON ROWS
FROM [factcon]
```

OUTPUT:

		Exam Score	Total Study Hours
2024	Artificial Intelligence	92	14.2
2024	Bioinformatics	81.7	12
2024	Business Analytics	89.2	15.6
2024	Computer Science	85.5	12.5
2024	Cybersecurity	74	9.5
2024	Data Science	78	8
2024	Information Technology	86	13.9
2024	Mathematics	85	13
2024	Physics	70.5	7.8
2024	Robotics	88.5	16.8
2024	Software Engineering	80.2	11.3
2024	Statistics	95.3	18.2
2025	Blockchain	72.1	8.6
2025	Cloud Computing	76.3	9.7
2025	Machine Learning	82.8	11.5

MDX Execution Time: 4 ms

(B) How can we retrieve a detailed breakdown of exam scores and study hours by drilling down from years to months to full dates while also including student degree programs and individual student names?

Concept Hierarchy:

Time Dim: Year → Month → Full Date

Student Dim: Degree Program → Full Name

DRILL DOWN:

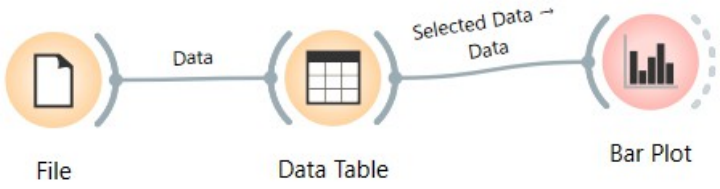
Query

```
SELECT
  {[Measures].[Exam Score], [Measures].[Study Hours]} ON COLUMNS,
  NONEMPTY(
    CROSSJOIN(
      EXCEPT([Time Dim].[Year].MEMBERS, {[Time Dim].[Year].[All]}),
      EXCEPT([Time Dim].[Month].MEMBERS, {[Time Dim].[Month].[All]}),
      EXCEPT([Time Dim].[Full  Date].MEMBERS, {[Time Dim].[Full  Date].[All]}),
      EXCEPT([Student Dim].[Degree Program].MEMBERS, {[Student Dim].[Degree Program].[All]}),
      EXCEPT([Student Dim].[Full Name].MEMBERS, {[Student Dim].[Full Name].[All]})
    ),
    [Measures].[Exam Score]  -- Ensuring only meaningful data is retrieved
  ) ON ROWS
FROM [factcon]
```

OUTPUT:

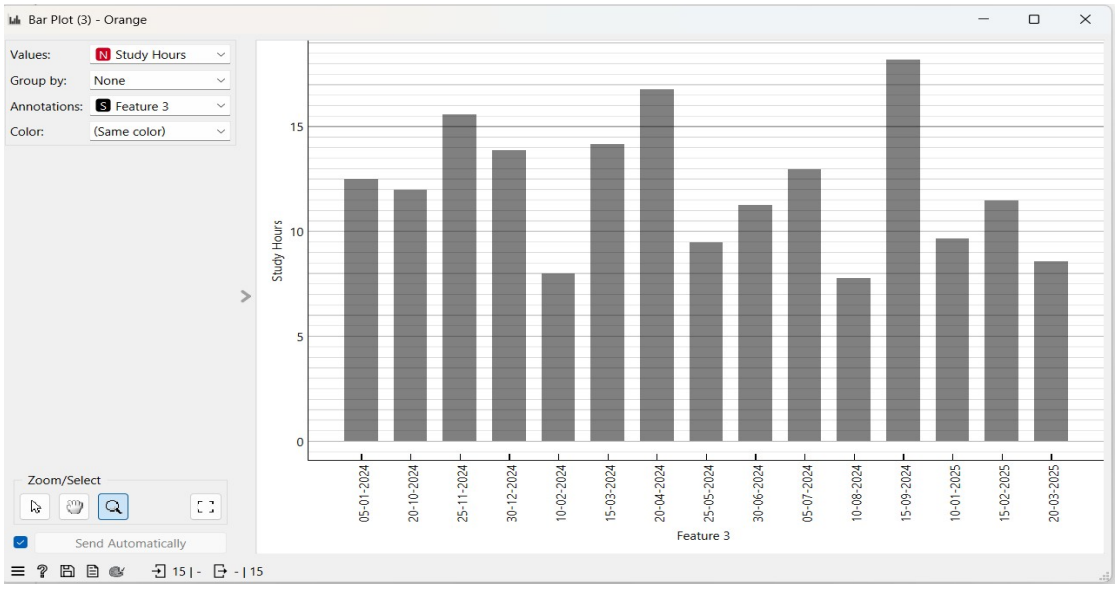
					Exam Score	Study Hours
2024	1	2024-01-05	Computer Science	Alice Johnson	85.5	12.5
2024	10	2024-10-20	Bioinformatics	Jack Taylor	81.7	12
2024	11	2024-11-25	Business Analytics	Kelly Adams	89.2	15.6
2024	12	2024-12-30	Information Technology	Liam Wright	86	13.9
2024	2	2024-02-10	Data Science	Bob Smith	78	8
2024	3	2024-03-15	Artificial Intelligence	Charlie Davis	92	14.2
2024	4	2024-04-20	Robotics	David Brown	88.5	16.8
2024	5	2024-05-25	Cybersecurity	Emma Wilson	74	9.5
2024	6	2024-06-30	Software Engineering	Frank White	80.2	11.3
2024	7	2024-07-05	Mathematics	Grace Miller	85	13
2024	8	2024-08-10	Physics	Hannah Moore	70.5	7.8
2024	9	2024-09-15	Statistics	Ian Clark	95.3	18.2
2025	1	2025-01-10	Cloud Computing	Mia Lewis	76.3	9.7
2025	2	2025-02-15	Machine Learning	Noah Scott	82.8	11.5
2025	3	2025-03-20	Blockchain	Olivia Martin	72.1	8.6

Step 7:Visualize OLAP Results:
Visualizing exam score Using Orange Tool



Bar Chart Configuration:

- **X-Axis:** Year and Month
- **Y-Axis:** study hours



MDX Execution Time:9 ms

(C) How can we analyze total study hours while focusing only on male students in the year 2024, broken down by different study methods?
Slice :
Query:
SELECT
[Measures].[Total Study Hours] ON COLUMNS,

```
  EXCEPT([Method Dim].[Method Name].MEMBERS, {[Method Dim].[Method Name].[All]}) ON
ROWS
FROM [factcon]
WHERE ([Time Dim].[Year].[2024], [Student Dim].[Gender].[Male])
```

OUTPUT:

	Total Study Hours
Case Studies	94.4
Coding Competitions	94.4
Group Discussions	94.4
Hackathons	94.4
Hands-on Labs	94.4
Interactive Quizzes	94.4
One-on-One Tutoring	94.4
Online Courses	94.4
Online Forums	94.4
Research Collaboration	94.4
Seminars	94.4
Study Groups	94.4
Video Tutorials	94.4
Webinars	94.4
Workshops	94.4
Unknown	94.4

MDX Execution Time: 3 ms

(D How can we analyze exam scores for third-year students while filtering for specific learning platforms and benefits received?

Dice :

Query:

```
SELECT
  [Measures].[Exam Score] ON COLUMNS,
  NONEMPTY(
    CROSSJOIN(
      EXCEPT([Platform Dim].[Platform Name].MEMBERS, {[Platform Dim].[Platform Name].[All]}),
      EXCEPT([Benefits Dim].[Benefit Description].MEMBERS, {[Benefits Dim].[Benefit
Description].[All]})
    )
  ) ON ROWS
FROM [factcon]
WHERE ([Student Dim].[Year Of Study].[3])
```

OUTPUT:

		Exam Score
Coursera	Increased Motivation	(null)
Coursera	Networking Opportunities	(null)
Coursera	Self-Learning Ability	(null)
Coursera	Skill Development	(null)
Coursera	Time Management	(null)
Coursera	Unknown	(null)
edX	Better Communication	80.2
edX	Better Retention	(null)
edX	Career Opportunities	(null)
edX	Collaboration Skills	(null)
edX	Confidence Boost	(null)
edX	Critical Thinking	(null)
edX	Enhanced Problem-Solving	(null)
edX	Exposure to New Ideas	(null)
edX	Higher Grades	(null)

MDX Execution Time: 6ms

(E) **How can we analyze student challenges across different days of the week to identify patterns in study difficulties?**

Pivot:

Query:

```
SELECT
  NONEMPTY(
    EXCEPT([Time Dim].[Day Of Week].MEMBERS, {[Time Dim].[Day Of Week].[All]})
  ) ON COLUMNS,
```

```
  NONEMPTY(
    EXCEPT([Challenges Dim].[Challenge Description].MEMBERS, {[Challenges Dim].[Challenge
Description].[All]})
  ) ON ROWS
FROM [factcon]
```

OUTPUT:

	Friday	Monday	Saturday	Sunday	Thursday	Tuesday	Wednesday
Communication Barriers	20	7	17	18	5	6	4
Different Learning Paces	20	7	17	18	5	6	4
Difficulty in Applying Concepts	20	7	17	18	5	6	4
Difficulty in Understanding	20	7	17	18	5	6	4
Distractions	20	7	17	18	5	6	4
Exam Anxiety	20	7	17	18	5	6	4
Internet Connectivity Issues	20	7	17	18	5	6	4
Lack of Motivation	20	7	17	18	5	6	4
Lack of Study Materials	20	7	17	18	5	6	4
Lack of Time	20	7	17	18	5	6	4

MDX Execution Time: 6 ms

Step 8:perform data mining

Classification of Peer Learning Preferences

Objective:

We aim to classify users based on their preference for universities actively promoting peer learning (Yes/No) using Supervised Machine Learning techniques.

5.1 DATA PREPARATION FOR CLASSIFICATION

Dataset Features:

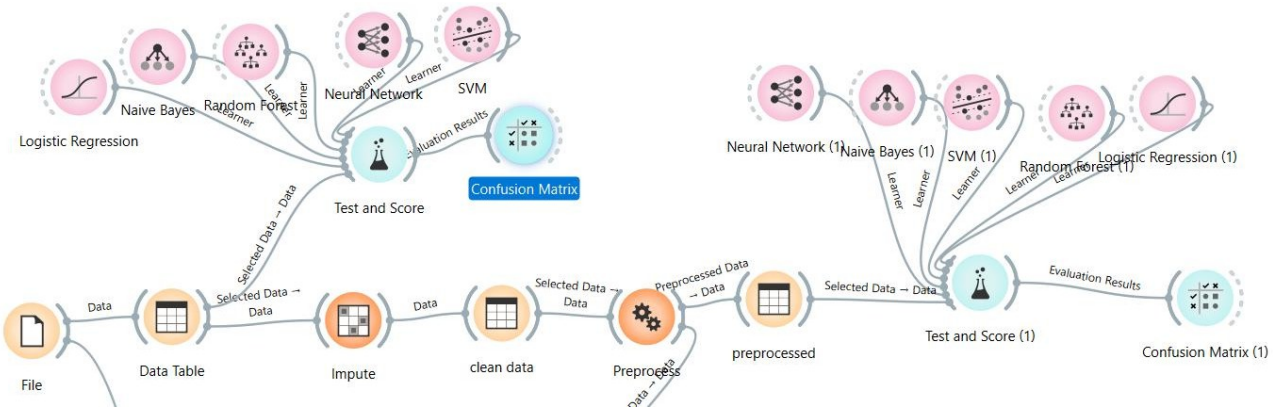
- **Student Attributes:** Age, Gender, Degree Program, Year of Study
- **Peer Learning Participation:** Preferred Methods, Frequency of Peer Learning Sessions, Platforms Used
- **Engagement & Benefits:** Study Improvement, Challenges Faced, Effectiveness Rating
- **Target Variable:** preference for universities actively promoting peer learning

Data Preprocessing:

1. Handle missing values.
2. Normalize numerical data
3. Encode categorical data

5.2 SELECTING CLASSIFICATION ALGORITHMS

- We use Supervised ML models to predict whether students think universities should promote peer learning more actively (Yes/No).
- **Random Forest**
- **Neural Networks**
- **Logistic Regression**
- **naïve Bayes**
- **Support Vector Machine (SVM)**
-



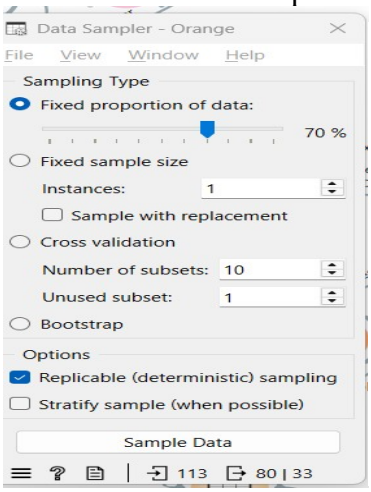
- Test the Accuracy of the Individual Models by the TEST&SCORE Evaluation
- The Highest Accuracy in Test& Score is Regarded as a Best MODEL Approach To Classify the Dataset

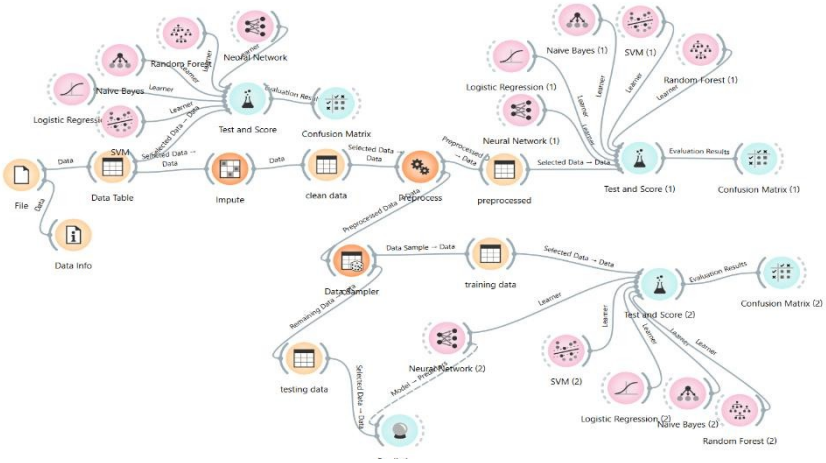
5.3 TRAINING & TESTING THE MODEL

- **Dataset Split:**

    1. **Training Set (80%)** – Used to train the model.
    2. **Testing Set (20%)** – Used to evaluate the model.

    Train models to learn patterns in **peer learning behavior** and predict whether universities yes/no





5.4 MODEL EVALUATION METRICS

To determine the best classification model, we evaluate using:

- **Accuracy:** How often the model correctly predicts the streaming platform.
- **Precision:** How many predicted platforms were correct.
- **Recall:** How well the model identifies actual platform users.

- **F1-Score:** Balances precision and recall.
- **Confusion Matrix:** Compares predicted vs. actual platform classification.

5.5 METHODOLOGY OVERVIEW

Step 1: Data Collection & Preprocessing

Gather data on student peer learning behavior.

● Label data based on whether universities should promote peer learning more actively (Yes/No).

Step 2: Model Training & Classification

Train classification models to predict if a student believes universities should promote peer learning more actively.

Step 3: Evaluate & Compare Models

- Use metrics like accuracy, precision, recall, and F1-score to select the best model.

	MODELS	AUC	CA	F1	PREC	RECALL	MCC
WITHOUT PREPROCESSING	SVM	0.682	0.823	0.810	0.803	0.823	0.318
	NAÏVE BAYES	0.751	0.726	0.747	0.780	0.726	0.235
	RANDOM FOREST	0.815	0.805	0.769	0.756	0.805	0.143
	LOGISTIC REGRESSION	0.741	0.805	0.785	0.775	0.805	0.216
	NEURAL NETWORKS	0.718	0.823	0.810	0.803	0.823	0.318
WITH PREPROCESSING (WITHOUT SAMPLER)	SVM	0.707	0.823	0.758	0. 771	0.823	0.114
	NAÏVE BAYES	0.744	0.690	0.724	0. 799	0.690	0.277
	RANDOM FOREST	0. 796	0.823	0.810	0.803	0.823	0.318
	LOGISTIC REGRESSION	0.746	0.814	0.791	0.782	0.814	0.239
	NEURAL NETWORKS	0.744	0.841	0.834	0.829	0.841	0.410
WITH PREPROCESSING (WITHSAMPLER)	SVM	0.636	0.812	0.728	0. 660	0.812	0.000
	NAÏVE BAYES	0.763	0.662	0.699	0. 820	0.662	0.338
	RANDOM FOREST	0.757	0.825	0.798	0.797	0.825	0.305
	LOGISTIC REGRESSION	0.726	0.800	0.769	0.759	0.800	0. 191
	NEURAL NETWORKS	0.713	0.825	0.814	0.809	00.825	0.366

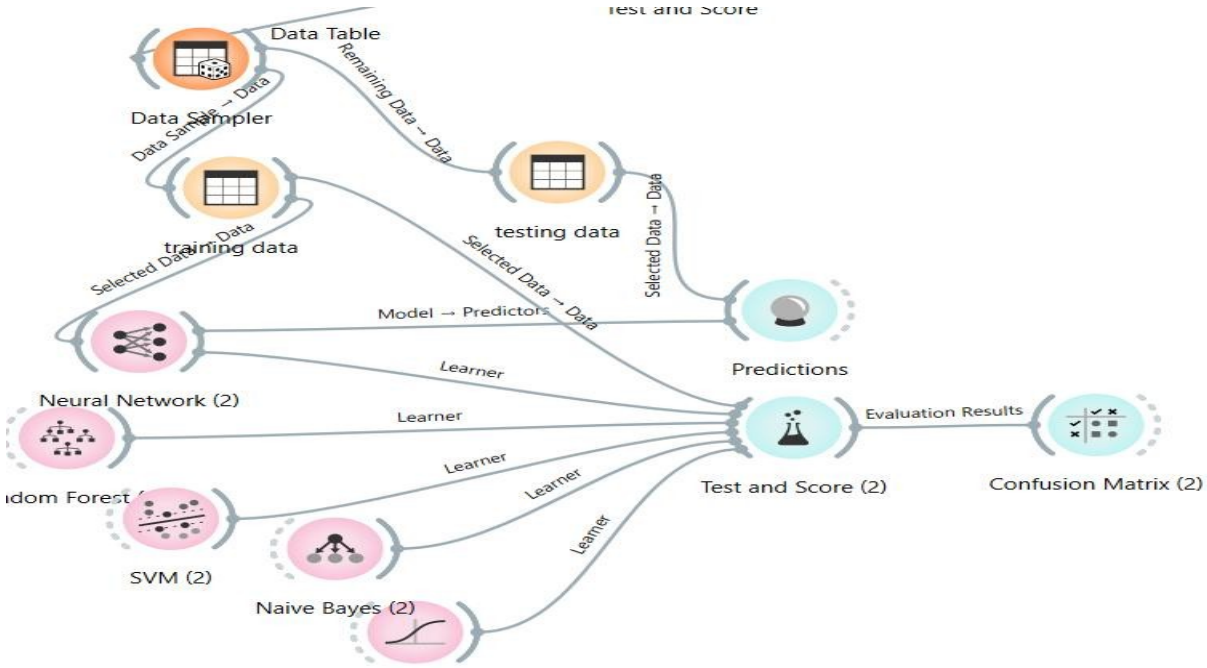
“ Neural Network Model Achieved the Highest Accuracy “

After testing various models,**Neural Network** demonstrated the highest accuracy in classifying students based on whether universities should promote peer learning more actively

VISUALIZATION & PREDICTION ANALYSIS:

- **Data Preprocessing & Sampling:**

- The 30 % testing data is used for prediction



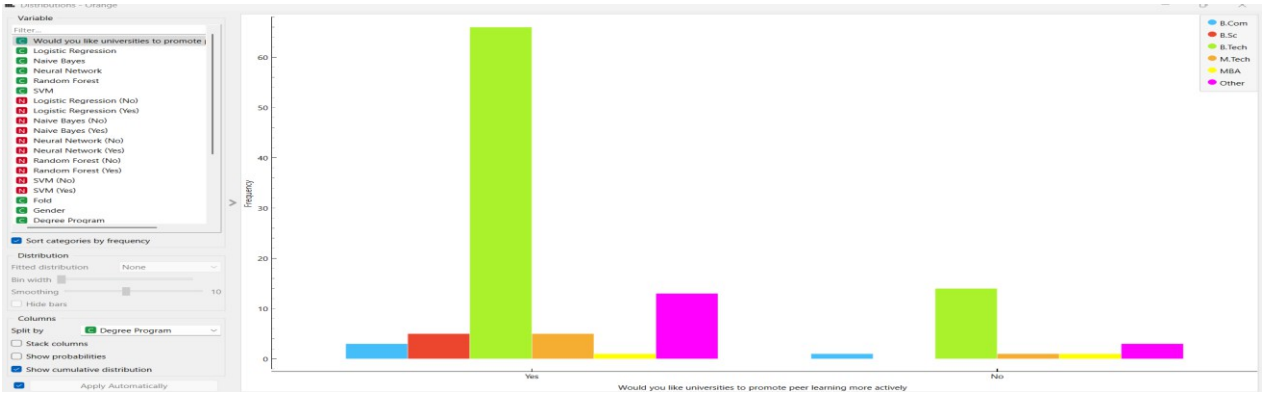
	Neural Network (2)	error	es to promote pe	Timestamp	Full Name	Age	Year of Study	s to improve peer	Gender	Degree Program	with the concept	participate in pee	ng method do yo	do you use most	er learning
1	0.00 : 1.00 → Yes	0.003	Yes	2025/02/05 11:...	Ch.Sunitha	19	1	imrpove activities	Female	Other	Yes	Weekly	Study Groups	WhatsApp/Tele...	0.5
2	0.01 : 0.99 → Yes	0.009	Yes	2025/02/05 4:2...	Samantha	23	3	Improved probl...	Female	B.Tech	Yes	Weekly	Online Discussi...	Google Meet/Z...	0.5
3	0.38 : 0.62 → Yes	0.383	Yes	2025/02/05 1:1...	M.Swayamprak...	19	2	nothing	Prefer not to say	B.Tech	Yes	Weekly	Study Groups	WhatsApp/Tele...	0.0
4	0.01 : 0.99 → Yes	0.009	Yes	2025/02/06 5:4...	Nagarjuna	23	3	Improved probl...	Male	B.Tech	Yes	Weekly	Peer Tutoring	Discord/Reddit	0.5
5	0.00 : 1.00 → Yes	0.001	Yes	2025/02/05 4:2...	Virat	21	3	Improved probl...	Male	B.Tech	Yes	Weekly	Group Projects	Google Meet/Z...	1.0
6	0.00 : 1.00 → Yes	0.004	Yes	2025/02/05 8:4...	Kollati Ribka	19	3	Conduct classes	Female	B.Tech	No	Daily	Group Projects	Google Meet/Z...	1.0
7	0.88 : 0.12 → No	0.118	No	2025/02/05 10:...	Harsha	19	1	yes improve act...	Male	B.Tech	No	Never	Study Groups	WhatsApp/Tele...	0.360
8	0.08 : 0.92 → Yes	0.082	Yes	2025/02/10 3:3...	Aditya	24	3	yes improve act...	Male	B.Tech	Yes	Daily	Study Groups	WhatsApp/Tele...	0.0
9	0.02 : 0.98 → Yes	0.016	Yes	2025/02/05 4:2...	Sampath	21	3	Improved probl...	Male	B.Tech	Yes	Daily	Group Projects	Discord/Reddit	0.0
10	0.04 : 0.96 → Yes	0.039	Yes	2025/02/05 5:0...	Yaswanth Kond...	19	2	Good	Male	B.Tech	Yes	Daily	Study Groups	WhatsApp/Tele...	1.0
11	0.00 : 1.00 → Yes	0.000	Yes	2025/02/06 5:4...	Venkatesh	26	4	Improved probl...	Male	B.Tech	Yes	Daily	Peer Tutoring	Google Meet/Z...	0.5
12	0.55 : 0.45 → No	0.551	Yes	2025/02/05 11:...	rajesh	23	1	yes improve act...	Male	B.Com	Yes	Rarely	Group Projects	University-provi...	0.5
13	0.99 : 0.01 → No	0.988	Yes	2025/02/05 10:...	Nikhil	19	2	nothing	Male	B.Tech	Yes	Monthly	Study Groups	WhatsApp/Tele...	0.0
14	0.11 : 0.89 → Yes	0.112	Yes	2025/02/06 7:1...	Anjali	20	3	No	Female	B.Sc	No	Rarely	Online Discussi...	University-provi...	1.0
15	0.05 : 0.95 → Yes	0.054	Yes	2025/02/04 9:3...	Likitha	20	3	no	Female	B.Tech	Yes	Monthly	Study Groups	WhatsApp/Tele...	0.5
16	0.41 : 0.59 → Yes	0.412	Yes	2025/02/05 4:0...	Sivanagaraju	21	2	No	Male	M.Tech	Yes	Monthly	Study Groups	Discord/Reddit	0.5
17	0.01 : 0.99 → Yes	0.014	Yes	2025/02/05 7:2...	Sandy	19	4	No	Female	B.Tech	Yes	Rarely	Study Groups	University-provi...	0.0
18	0.02 : 0.98 → Yes	0.021	Yes	2025/02/04 9:4...	Hari Jakku	21	3	nothing	Male	B.Tech	Yes	Monthly	Group Projects	Google Meet/Z...	0.5
19	0.03 : 0.97 → Yes	0.030	Yes	2025/02/05 7:3...	Gajjala	50	4	nothing special ...	Male	Other	Yes	Monthly	Study Groups	WhatsApp/Tele...	0.0
20	0.18 : 0.82 → Yes	0.177	Yes	2025/02/06 5:2...	Sandeep reddy ...	25	3	Improved probl...	Male	B.Com	Yes	Weekly	Peer Tutoring	University-provi...	0.5
21	0.02 : 0.98 → Yes	0.978	No	2025/02/10 3:3...	Purnima	25	2	no suggestions	Female	MBA	Yes	Monthly	Group Projects	WhatsApp/Tele...	0.0
22	0.14 : 0.86 → Yes	0.143	Yes	2025/02/05 8:1...	Gangireddy Dh...	20	3	Nothing	Male	B.Tech	Yes	Rarely	Study Groups	Google Meet/Z...	0.0
23	0.04 : 0.96 → Yes	0.035	Yes	2025/02/06 5:2...	Hardik pandya	21	2	Increased enga...	Male	M.Tech	Yes	Weekly	Peer Tutoring	Google Meet/Z...	0.5
24	0.00 : 1.00 → Yes	0.001	Yes	2025/02/06 5:2...	Lokesh	22	3	Increased enga...	Male	B.Tech	Yes	Weekly	Peer Tutoring	University-provi...	1.0
25	0.02 : 0.98 → Yes	0.017	Yes	2025/02/12 10:...	varshitha katuri	17	1	fest fot college	Female	B.Tech	Yes	Weekly	Study Groups	University-provi...	1.0
26	0.01 : 0.99 → Yes	0.012	Yes	2025/02/05 7:2...	Layabonam	18	3	Great idea	Female	B.Tech	Yes	Weekly	Group Projects	Discord/Reddit	0.0

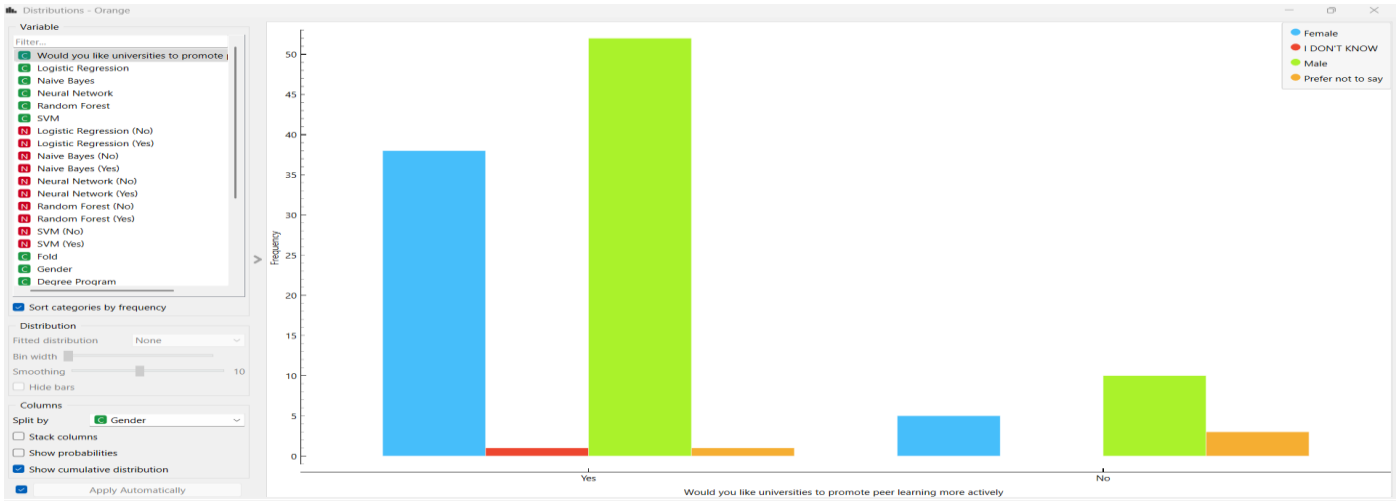
Show perfomance scores Target class: Yes

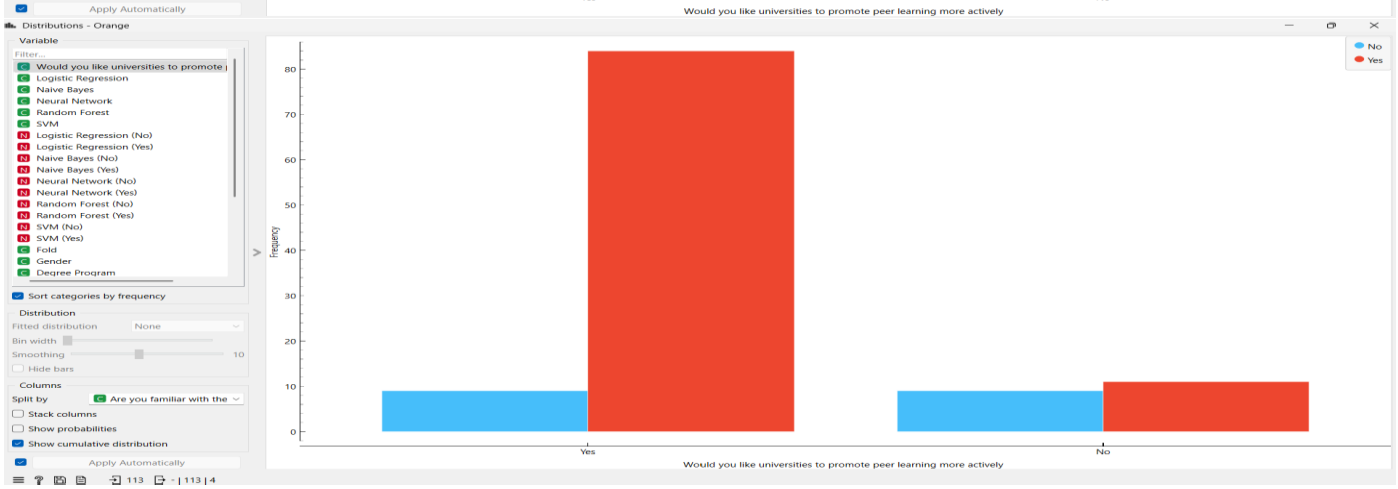
Model	AUC	CA	F1	Prec	Recall	MCC
Neural Network (2)	0.836	0.879	0.929	0.929	0.929	0.529

5.6 VISUALIZATION METRICS FOR CLASSIFICATION:

To analyze and validate the classification results, we utilize **Distributions** for visualizing Classisfication







5.7 EVALUATION METRICS:

- **Confusion Matrix** was used to analyze correct and incorrect classifications.

WITHOUT PREPROCESSING

Actual \ Predicted	No	Yes	Σ
No	7	13	20
Yes	7	86	93
Σ	14	99	113

WITH PREPROCESSING

Actual \ Predicted	No	Yes	Σ
No	9	11	20
Yes	7	86	93
Σ	16	97	113

Through these evaluations, we successfully classified students based on their preference for universities promoting peer learning more actively. This insight can help educational institutions design better collaborative learning strategies and enhance student engagement.

5.8 EXPERIMENT ANALYSIS:

The experiment aimed to classify students based on their preference for universities actively promoting peer learning using supervised machine learning models. Various models, including SVM, Random Forest, naïve bayes, and Neural Networks, were trained on student data consisting of attributes like peer learning participation, preferred methods, platforms used, benefits, and challenges faced. After preprocessing the data (including handling missing values, normalization, and encoding), the models were evaluated using metrics such as accuracy, precision, recall, and F1-score. The Neural Network model achieved the highest accuracy and was identified as the best model for classifying student responses. Additional techniques, like SMOTE, were applied to balance the dataset. Visualization methods, such as confusion matrices and distributions, were used to validate the classification results.

CONCLUSION:

In this project, we successfully classified students based on their preference for universities actively promoting peer learning using Supervised Machine Learning techniques. The Neural Network model demonstrated the highest accuracy among all tested classifiers, highlighting the key factors that influence students' perspectives on peer learning. This study provides valuable insights for educational institutions to refine and enhance their collaborative learning strategies.

PART-B: Space Shuttle Landing Decision Data using Data Mining

1.1 Problem Statement:

The objective of this study is to analyze and classify space shuttle landing decisions using machine learning techniques based on key control parameters. Using the "shuttle-landing-control.tab" dataset, which includes attributes such as wind conditions, stability metrics, and flight dynamics, this project aims to develop an effective classification model to predict the appropriate landing site decision. Various supervised learning algorithms will be evaluated to identify the most accurate and robust model. Through comprehensive data preprocessing, feature selection, and performance evaluation using metrics like accuracy, precision, recall, and F1-score, this study seeks to enhance decision-making in aerospace landing systems and contribute to the safety and reliability of space missions.

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.

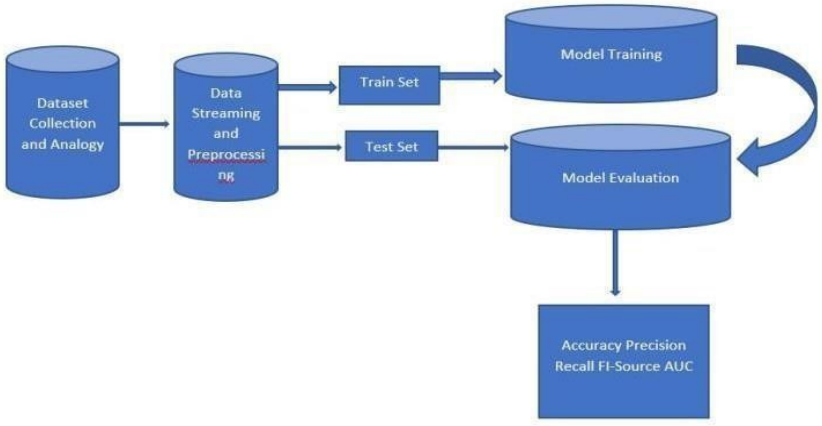
1.2.1 Dataset Overview

The **shuttle-landing-control.tab** dataset contains information related to shuttle landing scenarios, including various attributes that influence the landing process. These attributes may include **weather conditions, wind speed, landing site characteristics, shuttle type, decision outcomes, and control parameters**.

The goal is to **perform classification** with the target variable **"decision"**, which determines the appropriate landing control action based on the given conditions. By applying classification techniques, we aim to predict the best landing control strategy for different shuttle landing scenarios.

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 used Supervised Machine Learning models to classify shuttle landing decisions using "decision" as the target variable. The models include Naïve Bayes, Logistic Regression, Random Forest, Decision Tree, KNN, and SVM.

1.2.4 Evaluation Metrics

Since this is a classification problem, we can use:

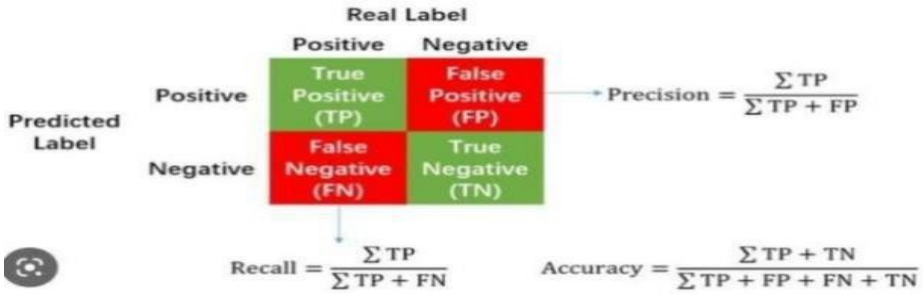
Accuracy: Accuracy is Calculated and Compared and best one should be noticed.

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.

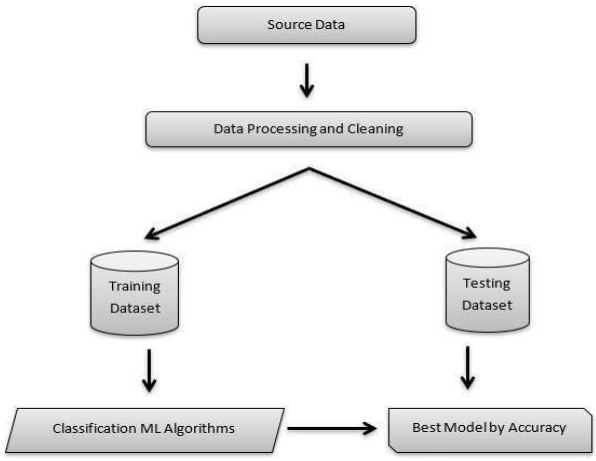
F-Measure: It offers a single score that evenly weighs issues of precision and recall. **Confusion Matrix:** It is used to determine the classification models performance for a set of test data.



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.

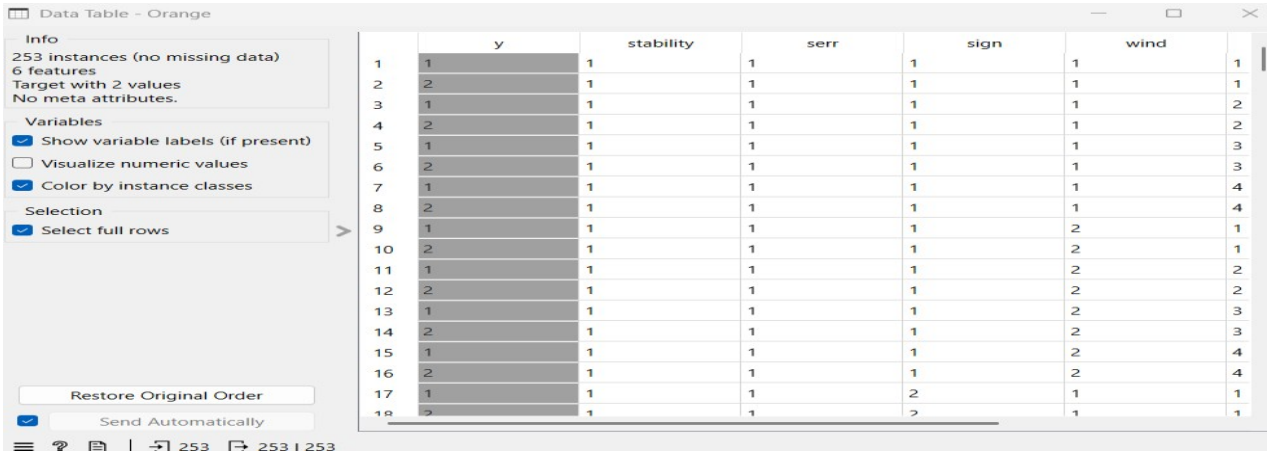


CHAPTER 2: ANALYSIS ON THE DATASET

Data Set Description:

The dataset comprises information related to shuttle landing control scenarios. It consists of records capturing multiple features that influence shuttle landing decisions under varying conditions. These features include: **wind, weather, visibility, cloud, airspeed, altitude, control inputs**, and other operational parameters affecting the landing outcome.

Each entry in the dataset provides valuable insights into shuttle landing control decisions, aiding in the classification of landing outcomes based on key attributes. This dataset is instrumental in analyzing landing conditions, optimizing control strategies, and improving future shuttle landing procedures.



Data Validation, Cleaning, and Preparation Process:

We meticulously assessed the dataset to ensure its accuracy and readiness for analysis. We began by identifying relevant variables such as **landing decision (target variable), wind speed, visibility, weather conditions, airspeed, altitude, and control inputs**. Through careful examination, we addressed any missing or duplicate values, ensuring the dataset's integrity for classification.

To enhance the reliability of our model, we applied **data preprocessing techniques** using the **Orange Data Mining** tool. Our approach involved:

- **Handling Missing Values:** Missing data in features were imputed using the best-suited imputation technique based on accuracy results from different models.
- **Target Variable Imputation:** Missing values in the **landing decision** column (target variable) were handled using the **Impute** widget to maintain data consistency.
- **Normalization of Numerical Features:** Numerical attributes such as **windspeed, airspeed, altitude, and cloud cover** were normalized in the **Preprocessing** widget to ensure uniform scaling and improve model performance.

In real-world scenarios, datasets may not always be a true representation of the population, making **data validation** essential. We validated the dataset by examining **data types (categorical or numerical)** and ensuring a balanced distribution of target classes.

To evaluate **model performance** and tune hyperparameters, we used a **sample dataset split into training and validation sets**. This ensured an unbiased evaluation of model fit during testing.

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 span types** in both sets, allowing us to effectively train and test machine learning models despite the limited number of records.

Data Visualization:

For data visualization, we utilized the **Orange Data Mining** tool to analyze relationships and correlations among key features such as **wind conditions, weather, visibility, airspeed, altitude, and control inputs**. This helped us detect dependencies between features and identify influential attributes for classification. Before proceeding with modeling, we performed **data cleaning** to ensure no duplicate or missing values were present. Next, we analyzed the distribution of **landing decisions (target variable)**. This simplified the model, improving both interpretability and performance.

Machine Learning Techniques and Model Selection

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

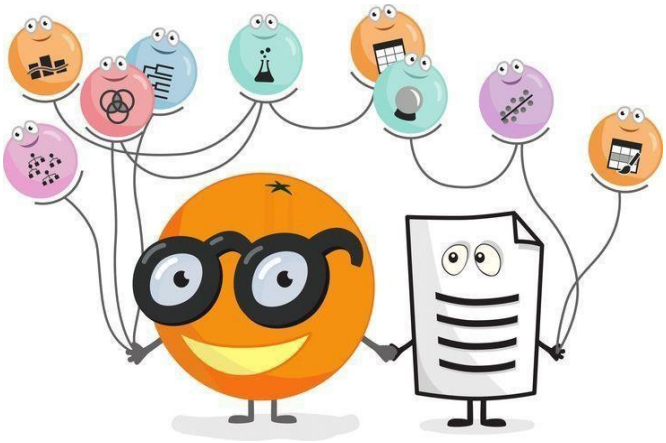
1. Random Forest
2. Naïve Bayes
3. Neural Network
4. SVM

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 **shuttle landing control classification**.

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 interfacewhere users can creatework flows by connecting different components, suchas data loaders, preprocessing tools, and machine learning algorithms.



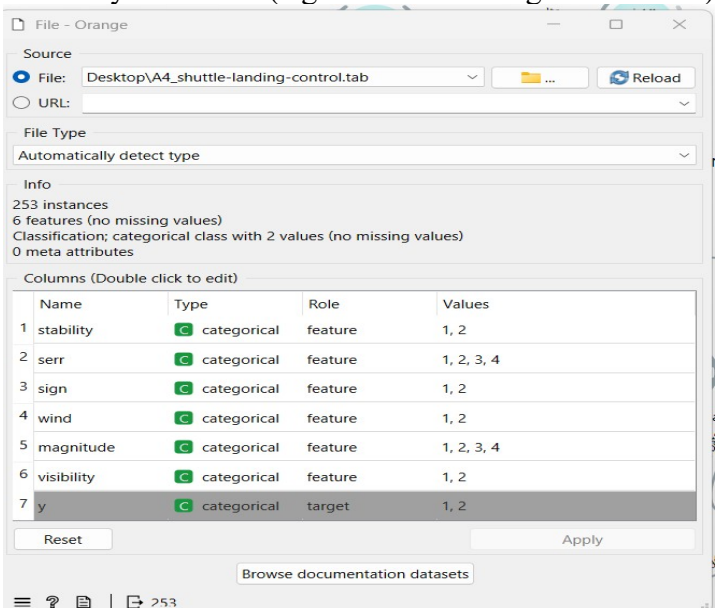
Step-by-Step Guide for Classification Using Orange

Step 1: Open Orange Canvas

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

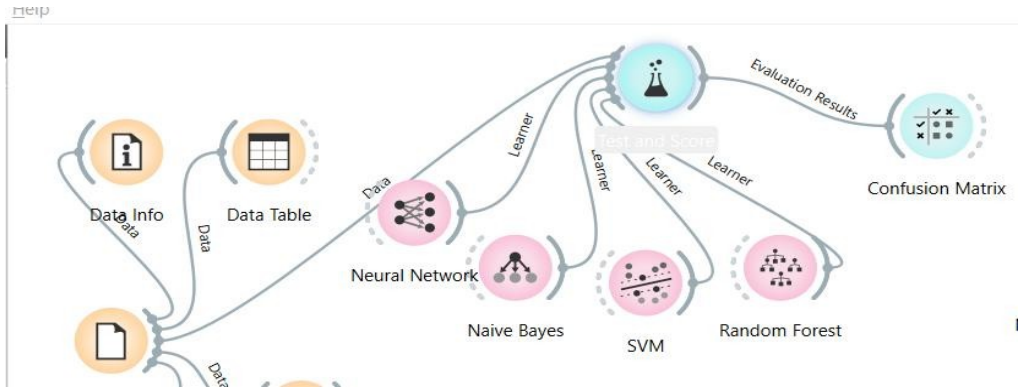
Step 2: Load Dataset

- 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.” shuttle-landing-control.tab”) and open it.



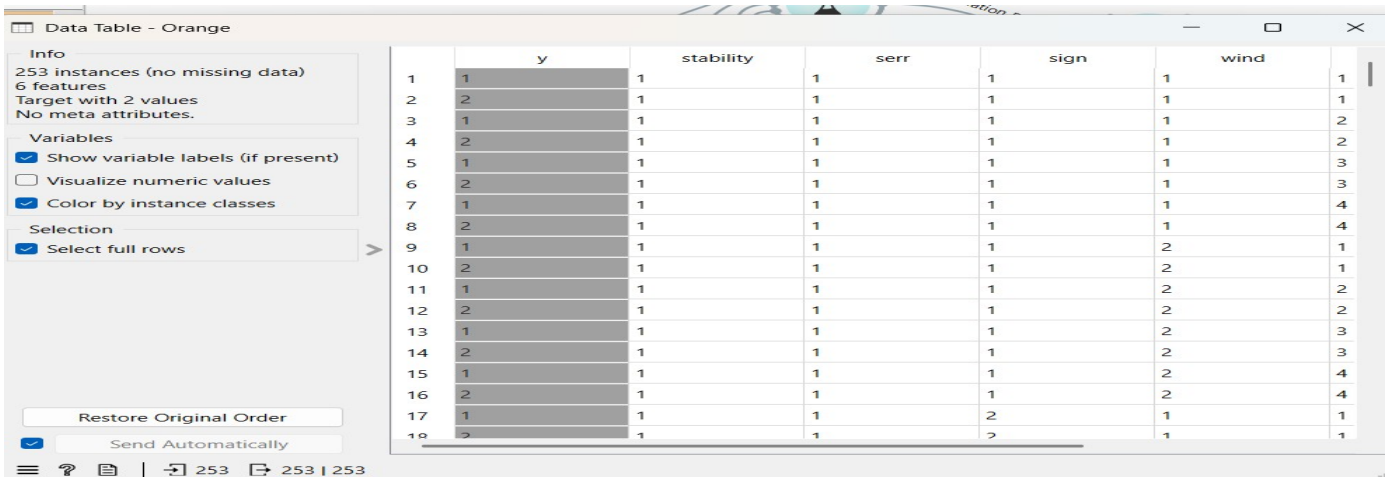
Step 3: Test the accuracies for various classification algorithms before preprocessing & choose the top five according to their accuracies.

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.

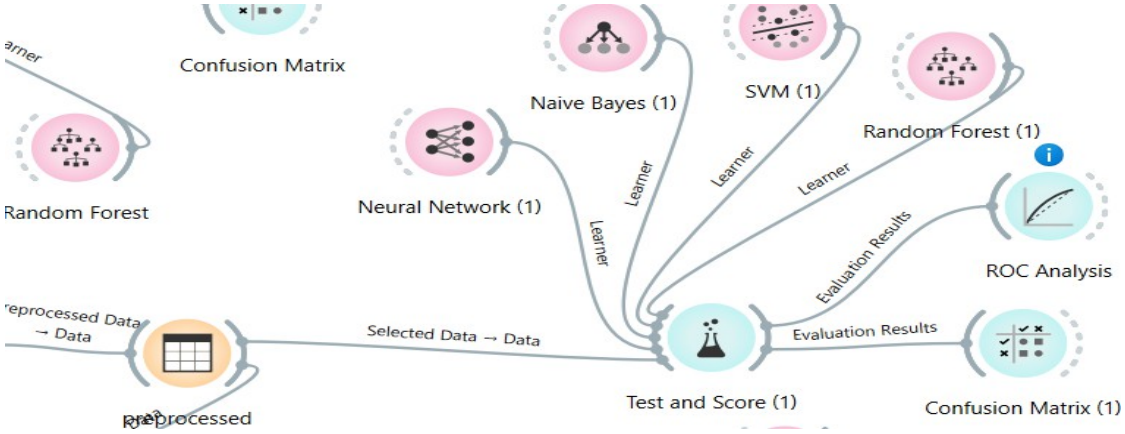


	y	stability	serr	sign	wind	
1	1	1	1	1	1	1
2	2	1	1	1	1	1
3	1	1	1	1	1	2
4	2	1	1	1	1	2
5	1	1	1	1	1	3
6	2	1	1	1	1	3
7	1	1	1	1	1	4
8	2	1	1	1	1	4
9	1	1	1	1	2	1
10	2	1	1	1	2	1
11	1	1	1	1	2	2
12	2	1	1	1	2	2
13	1	1	1	1	2	3
14	2	1	1	1	2	3
15	1	1	1	1	2	4
16	2	1	1	1	2	4
17	1	1	1	2	1	1
18	2	1	1	2	1	1

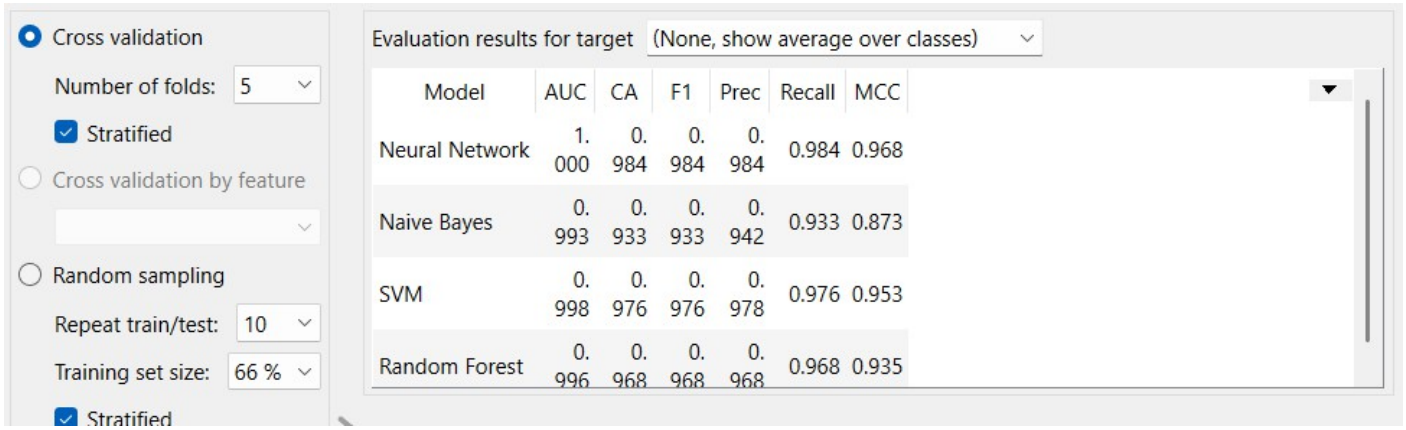
No missing values in the given dataset

Step-5:Testing accuracy of various classification algorithms

- Drag and drop the "Test & Score" widget
- Connect the "SVM", ”Random Forest”, “Neural Network”, “naïve Bayes” 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.

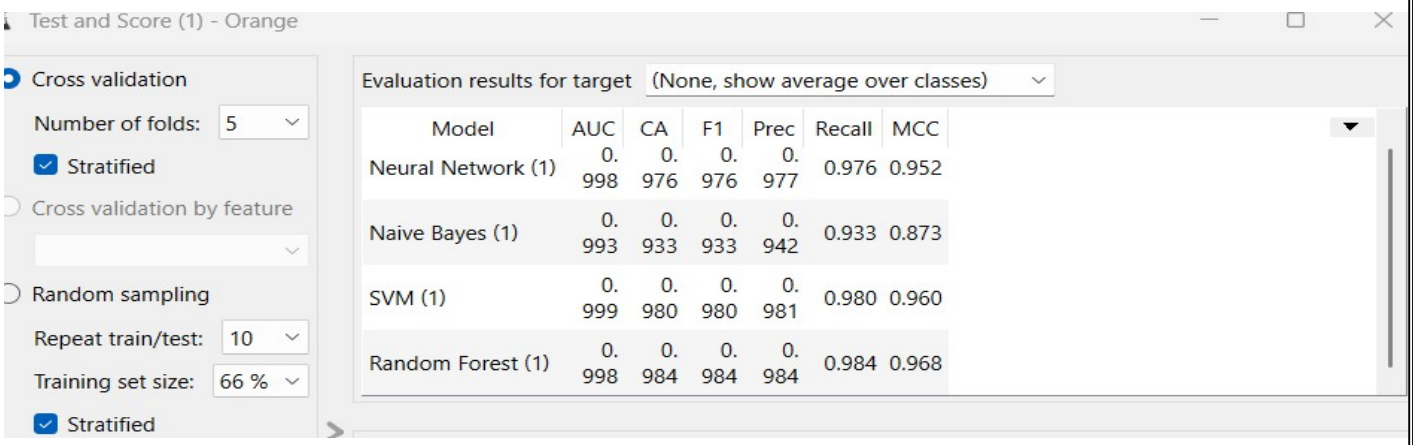


After preprocessing the data



Model	AUC	CA	F1	Prec	Recall	MCC
Neural Network	1.000	0.984	0.984	0.984	0.984	0.968
Naive Bayes	0.993	0.933	0.933	0.942	0.933	0.873
SVM	0.998	0.976	0.976	0.978	0.976	0.953
Random Forest	0.996	0.968	0.968	0.968	0.968	0.935

- 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.



Model	AUC	CA	F1	Prec	Recall	MCC
Neural Network (1)	0.998	0.976	0.976	0.977	0.976	0.952
Naive Bayes (1)	0.993	0.933	0.933	0.942	0.933	0.873
SVM (1)	0.999	0.980	0.980	0.981	0.980	0.960
Random Forest (1)	0.998	0.984	0.984	0.984	0.984	0.968

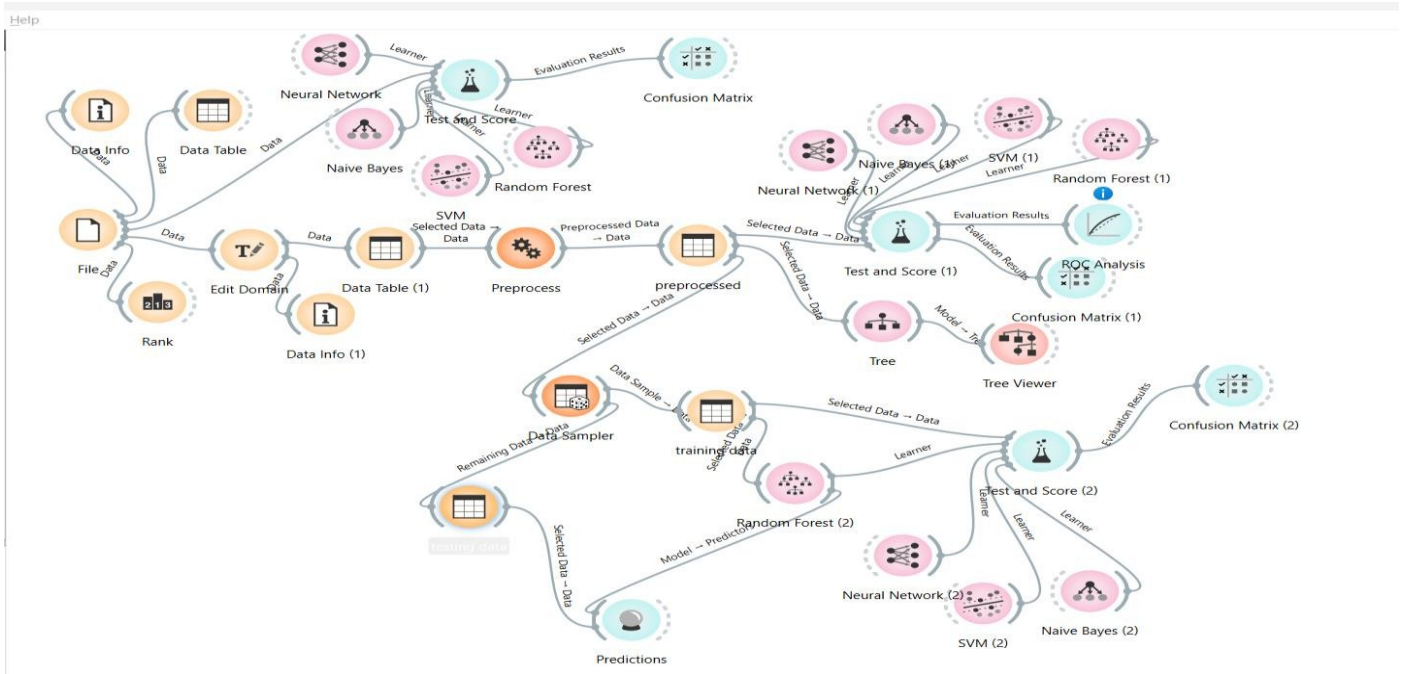
- **Neural networks** showed best CA before preprocessing and **Random Forest** showed best CA after applying preprocessing techniques.

Step 6: 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:
  Second way is by creating a separate test data with the help of available dataset and giving available dataset as training data.

First way workflow:

Entire Workflow:



Step 7: 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)

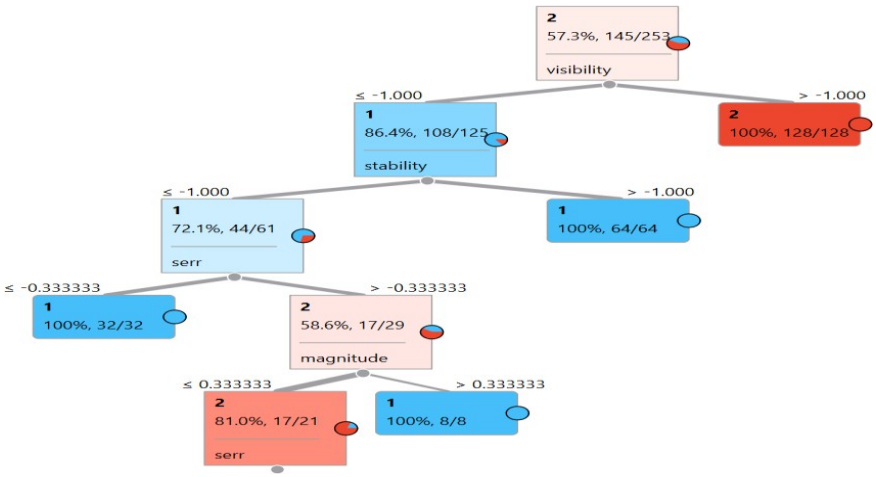
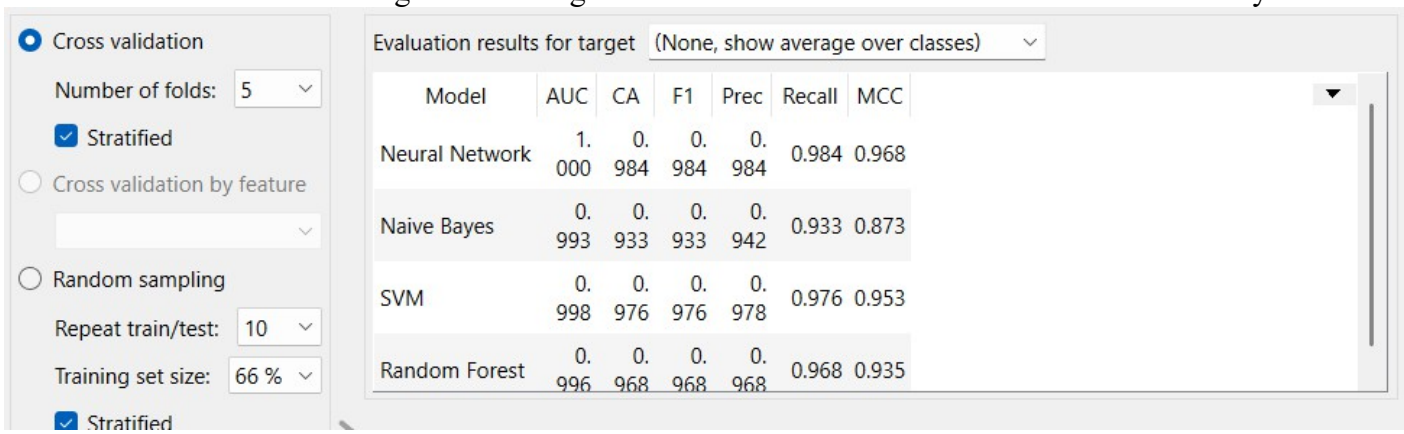


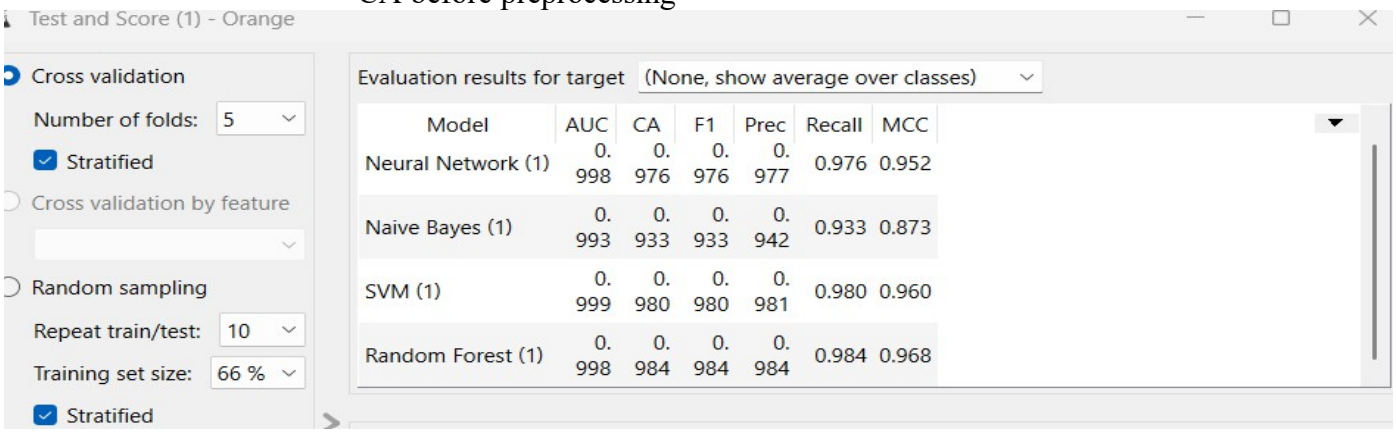
Figure depicts the decision tree, visually representing the classification of shuttle landing control decisions based on multiple attributes like **visibility, stability, serr, and magnitude**. Nodes split based on attribute values, with leaf nodes showing classification percentages. **Blue nodes indicate 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 Neural networks and random forest algorithms using various metrics like confusion matrix and Roc analysis



CA before preprocessing



CA after preprocessing

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

- **Neural Network:** 0.984
- **Naive Bayes:** 0.933
- **SVM:** 0.976
- **Random Forest:** 0.968

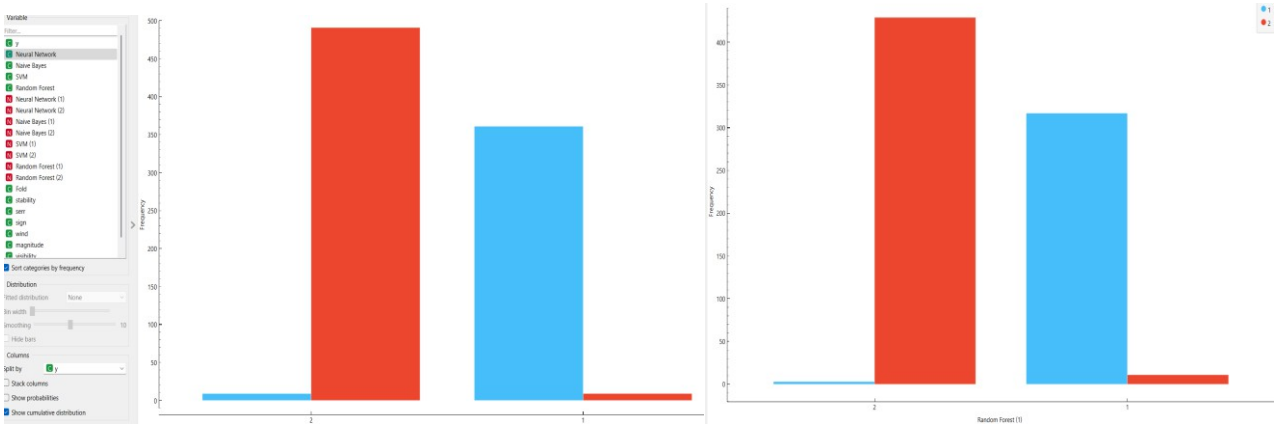
Figure 13 shows Test & Score **after preprocessing** (5-fold Cross Validation, 66% Training Size)

- **Neural Network:** 0.976 (Slight decrease)
- **Naive Bayes:** 0.933 (No change)
- **SVM:** 0.980 (Increased)
- **Random Forest:** 0.984 (Increased)

Comparison Observations:
All models except **Neural Network** improved their **CA** values after preprocessing.
Random Forest saw the highest improvement, from **0.968 to 0.984**.

SVM also showed a positive shift from **0.976 to 0.980**.
Naive Bayes remained constant at **0.933**, while **Neural Network** experienced a minor drop from **0.984 to 0.976**, but still performed at a high accuracy level.
Preprocessing contributed positively to most models, particularly enhancing the performance of **Random Forest** and **SVM** in classifying shuttle landing decisions.

Without preprocessing with preprocessing



Analysis on Confusion matrices:

These are the confusion matrices for the two best classification algorithms.

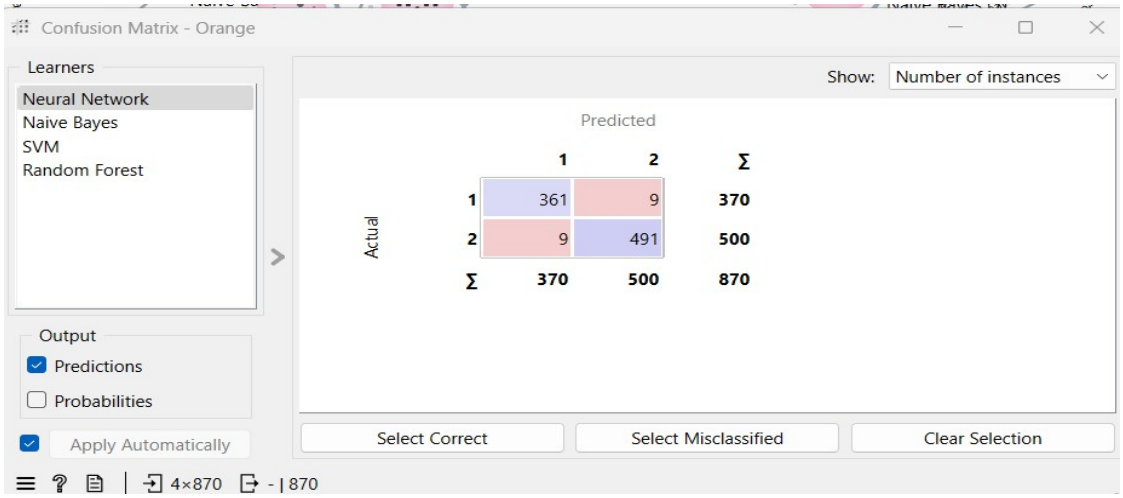


Figure 14: Confusion matrix of Neural Network

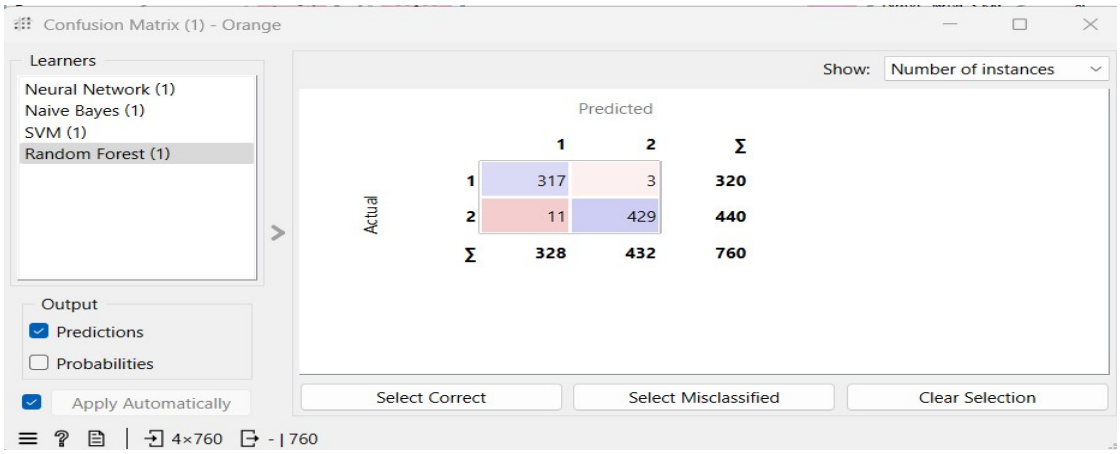


Figure 15: Confusion matrix of Random Forest

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

Overall Accuracy:

- **Neural Network:** Correct classifications = (361 + 491) = **852 out of 870**
- **Random Forest:** Correct classifications = (317 + 429) = **746 out of 760**
- **Neural Network** shows better overall accuracy than Random Forest.

Class-wise Performance:

- **Class 1:**
  - **Neural Network:** 361 correctly classified, 9 misclassified as class 2.
  - **Random Forest:** 317 correctly classified, 3 misclassified as class 2.
  - **Neural Network** performs better for Class 1.
- **Class 2:**
  - **Neural Network:** 491 correctly classified, 9 misclassified as class 1.
  - **Random Forest:** 429 correctly classified, 11 misclassified as class 1.
  - **Neural Network** performs better for Class 2.

Misclassification Trends:

- Neural Network has fewer misclassifications in both classes.
- Random Forest has slightly fewer misclassifications of Class 1 into Class 2 but higher errors in Class 2.
- Neural Network provides more consistent performance across both classes.

Conclusion:

The **Neural Network** model clearly outperforms Random Forest in both class-wise and overall accuracy (852/870 vs. 746/760). If the focus is on achieving high precision and minimizing misclassification across all classes, **Neural Network is the better choice**. However, **Random Forest** still shows competitive performance with relatively fewer parameters and might be considered if model simplicity is prioritized.

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 used a **real-world dataset** collected from institutional sources and surveys.

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

- The **generated dataset (Part B)** was well-structured with no missing values. Preprocessing /Dimensionality reductionwas straightforward, and the data was ready for model training with little effort.

- The **real-world dataset (Part A)** required extensive preprocessing, such as handling missing values, normalization, and dimensionality reduction due to inconsistencies.

- Although more challenging, Part B reflects real-world scenarios and helps in building more robust and generalized models.

2.2. Model Performance Analysis

- Several machine learning classifiers such as **K-Nearest Neighbors (KNN), Random Forest, Support Vector Machines (SVM)**,**k-Nearest Neighbors (k-NN)**,**Decision Tree**,**Neural Networks** were tested
- Their performance was evaluated using metrics like **classification accuracy (CA), confusion matrices**, and **ROC curves**.

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 in part B, highlighting the importance of data quality.
- Random Forest and neural network 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 balanced and pre-structured.
- No missing values.
- Feature engineering was already aligned with model requirements.

Part B: Extensive Preprocessing Required

- Missing values were handled using imputation.
- Normalization and feature scaling were applied.
- Redundant attributes were removed.

- Class imbalance issues were addressed using resampling techniques.
-

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**.

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 B achieved high accuracy quickly due to clean and pre-processed data.
  - Part A required extensive cleaning and transformation but provided insights closer to real-world applications.
  - **Neural Network was the top-performing classifier overall for the peer learning dataset.**
    This study successfully applied machine learning techniques to analyze student data focused on peer learning and its impact on academic success. The classification aimed to predict whether students support the idea that universities should actively promote peer learning. Among the various models tested, Neural Network achieved the highest accuracy, demonstrating its strength in capturing complex patterns within student feedback and learning behavior.
    The results highlight the potential of AI and data mining in educational environments to uncover insights into collaborative learning trends, student preferences, and strategies that can enhance academic support through peer-based initiatives.
  - **Random Forest was the top-performing classifier overall for the Shuttle Landing Control dataset.**
    This study successfully applied machine learning techniques to the Shuttle Landing Control dataset to classify control decisions as either "Yes" or "No" based on a variety of flight conditions. Among the tested models, Random Forest achieved the highest accuracy, confirming its robustness and adaptability in handling high-dimensional control-related data. The results support the application of AI in aerospace and control systems for enhancing decision-making, increasing flight safety, and developing intelligent guidance systems for complex operational environments.

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.

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 algorithmsto 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