Unlocking Patterns in Virtual Learning: A Data Mining Approach to Student Engagement and Platform Optimization 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

JAKKU KUMARSWAMI (22481A0582)

KATRAGADDA NAGA KAVYA (22481A05A0)

KURAKULA TARUN KUMAR (22481A05C6)

JARUGU VENKATA YASWANTH (22481A0586)

Under the Enviable and Esteemed Guidance of

Dr. N. RAJESWARI, M.Tech,Ph.D

Professor & Mentor, 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 Unlocking Patterns in Virtual Learning: A Data Mining Approach to Student Engagement and Platform Optimization is a bonafide record of work carried out by **J.KUMARSWAMI(22481A0582),K.NAGAKAVYA(22481A05A0),K.TARUN KUMAR(22481A05C6),J. VENKATA YASWANTH (22481A0586)**, under the guidance and supervision of **Dr.N.RAJESWARI ,M.Tech, Ph.D,** Professor & Mentor,Computer Science and Engineering, in the partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering of Jawaharlal Nehru Technological University Kakinada, Kakinada during the academic year 2024-25.

Project Guide
(Dr. N.RAJESWARI)

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

External Examiner

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people who made it possible and whose 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.N.RAJESWARI,M.Tech, Ph.D, Professor ,**Computer Science and Engineering for her 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

JAKKU KUMRSWAMI(22481A0563)

KATRAGADDA NAGA KAVYA (22481A05A0)

KURAKULA TARUN KUMAR(22481A05C6)

JARUGU VENKATA YASWANTH (22481A0586)

INDEX

ABSTRACT

The proliferation of online education platforms in recent years has significantly altered the landscape of learning, necessitating a deeper understanding of student engagement, behavioral patterns, and platform efficacy. This project presents a data-driven analysis based on a structured survey aimed at evaluating various dimensions of the online learning experience. The dataset encompasses demographic attributes, platform preferences, course formats, learning styles, feature expectations, satisfaction metrics, and perceived challenges.

Employing data mining methodologies, including exploratory data analysis, correlation mapping, and pattern discovery, the study uncovers latent trends in learner behavior and identifies critical factors influencing engagement and retention. Insights drawn from this analysis reveal key associations between learner profiles and their preferences for content delivery, time investment, and platform usability. Furthermore, the study highlights significant pain points such as lack of personalization, limited interactivity, and cognitive overload.

The results aim to support educators, instructional designers, and e-learning platforms in optimizing content delivery strategies and enhancing learner-centric features. By providing a holistic understanding of student needs in virtual environments, this research contributes to the development of adaptive and data-informed digital education systems

PART-A

Analyzing the Online Learning Experience: Behavioral Trends and Challenges in the Digital Education Era Using the 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

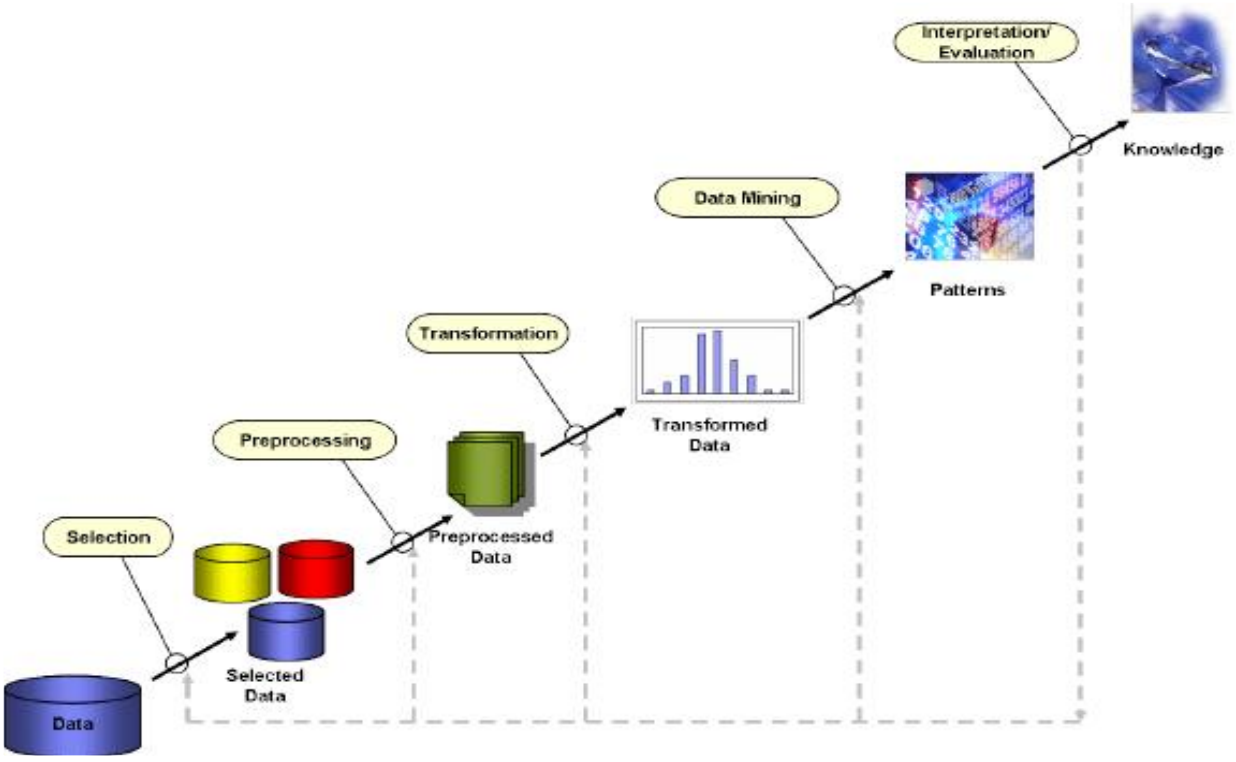


Fig1 : KDD Process

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 online learning survey analysis project, a data warehouse is utilized to store, organize, and analyze responses collected from diverse learners. This centralized data repository enables efficient analysis of learner behavior, platform preferences, satisfaction levels, and challenges associated with digital education. By structuring the data warehouse appropriately, we can extract meaningful insights to support platform improvements and learner-centric recommendations

Data Source Layer (Extracting Data)

- Data is collected directly from student survey responses.
- Includes attributes such as age group, educational level, preferred learning platform, course type, learning style, satisfaction level, and challenges faced.

1. **ETL (Extract, Transform, Load) Process**

   - **Extraction**: Data is gathered from survey forms and digital submissions..
   - **Transformation**: Raw Data is cleaned, missing values are handled , data types are standardized and categorical responses are encoded.
   - **Loading**: The structured and processed data is then loaded into the data warehouse for further analysis.

2. **Data Storage Layer (Fact & Dimension Tables)**

   - **Fact Table** stores core metrics like total hours spent on online learning, Satisfaction score,platform usage frequency,number of courses taken.
   - **Dimension Tables** include details like student demographics (age, education level), platform names(e.g., Coursera, Udemy, YouTube) ,course types(Technical and non Technical) and Learning Styles(video, text based,interactive).

3. **OLAP (Online Analytical Processing) for Data Analysis**

- Enables multi-dimensional querying and slicing of data to explore trends.
- Facilitates insights such as:
- Which platforms are most preferred by different age groups?
- Do learners using interactive courses report higher satisfaction?
- What challenges are more prominent among users of free platforms?
- How does time investment relate to learner satisfaction?

4. **Data Visualization & Reporting**

    - Results are visualized through dashboards, bar graphs, pie charts, and heatmaps.
    - These insights assist educators, platform developers, and policymakers in understanding user needs and enhancing the e-learning experience through evidence-based decisions.

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 data through a structured **online learning survey**. The initial phase involves **data preprocessing**, where raw responses are cleaned, standardized, and normalized to ensure consistency. The refined dataset is then **split into training and testing sets** to enable effective model development and evaluation.

The **training data** is used to build and train a range of **classification and clustering models** aimed at uncovering hidden patterns in learner behavior. These models help analyze and predict various aspects of the online learning experience, such as **platform preference**, **course engagement**, and **satisfaction levels**. Key attributes considered include **age group, education level, learning style, course type, time spent, and challenges faced**.

Ultimately, the data mining process provides actionable insights that support improved content delivery, personalized learning recommendations, and strategic enhancements to online education platforms.

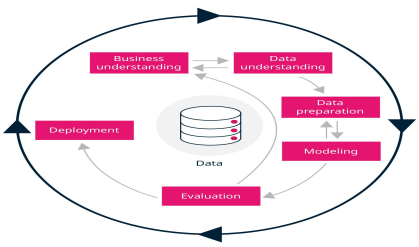


Fig2 :Data Mining Block Diagram

➢ 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 Online learning dataset to grasp its structure and content.
   - Identifying attributes such as age, education level,learning style,course type,time spent etc.
2. **Data Preparation**
   - Cleaning and transforming the dataset by handling missing values, standardizing data, and encoding categorical attributes.
   - Normalizing numerical data for better accuracy in analysis.
3. **Modelling**
   - Applying various classification algorithms like Neural Network, Random Forest, Navie Bayes, KNN,SVM… etc to predict a user's preferred streaming platform.
4. **Evaluation**
   - Assessing model performance using accuracy, precision, recall, and F1-score to ensure reliable predictions.
5. **Deployment**
   - Integrating the best-performing model to provide insights into learner engagement, satisfaction levels, and key factors influencing the effectiveness of online learning platforms

➢ SUPERVISED LEARNING

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

- **K-Nearest Neighbors (KNN)**
- **Neural Network**
- **Random Forest**
- **Navie Bayes**
- **Support Vector Machine**

Categories of Supervised Learning in This Project:

1. **Classification:**
   - The dataset contains categorical labels (e.g., platform name,course type,satisfaction .etc).
   - Classification algorithms are utilized to identify patterns in student engagement and learning preferences based on attributes such as time spent, learning styles, platform features, and usage behavior.

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

Neural Networks	A **neural network** is a machine learning model inspired by the human brain that processes data through interconnected layers of nodes to recognize patterns, make predictions, or classify information.	Classification and regression
Naïve Bayes	The Bayesian method is a classification method that makes use of the Bayesian theorem. The theorem updates the prior knowledge of an event with the independent probability of each feature that can affect the event.	Regression and Classification
KNN	K-Nearest Neighbors (KNN) is a supervised learning algorithm that classifies data points based on the labels of their nearest neighbors in the feature space. It assigns the most common label among the closest data points to the new data point.	Regression and Classification
Random Forest	Random Forest is an ensemble learning algorithm that builds multiple decision trees and combines their outputs for robust and accurate classification or regression	Regression and Classification

➢ UNSUPERVISED LEARNING

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

There are two categories of Unsupervised Learning. They are

1.Clustering

2.Association

1.Clustering:

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

2.Association:

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

How to Choose 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 online learning patterns**, **classification** algorithms are the **best fit.**

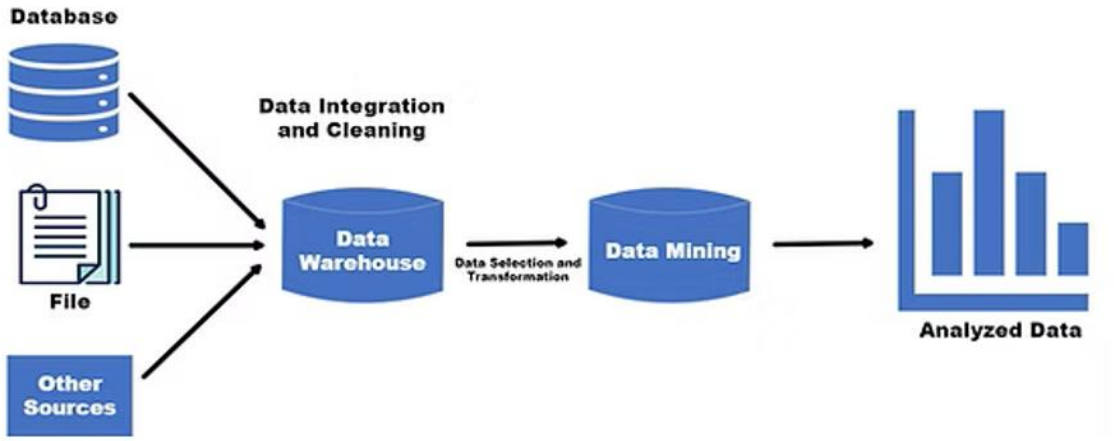


Fig3 : Data Mining Basic Diagram

➢ 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**.**
- **Dimensionality reduction:** 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**.**

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

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.

SOFTWARE AND HARDWARE REQUIREMENTS:

Software Requirements

Dataset:

- Clean and structured survey data.
- Handle missing values, duplicates, and inconsistent entries**.**

Software/Tools:

Windows 10 or above Operating System

- Orange Tool
- SQL Server Management Studio
- VS Code

Models:

- Classification Algorithms : Neural Network, Random Forest, KNN, SVM**,** Naive Bayes, etc.
- Model evaluation metrics: Accuracy, Precision, Recall, F1-score, Confusion Matrix

Hardware Requirements

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

➢ PROBLEM STATEMENT

Understanding learner behavior, preferences, and challenges in online education is vital for improving digital learning platforms and enhancing educational outcomes. However, manually analyzing large volumes of survey responses is time-consuming and may fail to uncover meaningful patterns related to learner engagement, satisfaction, and platform effectiveness.

This project aims to apply **data mining and machine learning techniques** to analyze student survey data and classify learners based on their **preferences, satisfaction levels, learning styles, platform usage, and challenges faced**. The goal is to derive actionable insights that support **personalized learning improvements** and strategic decision-making for educational stakeholders.

Objectives:

- **Identify behavioral patterns** and preferences among different learner segments.
- **Assist educators and platform developers** in tailoring content, features, and teaching approaches.
- **Improve online learning platforms** by addressing key pain points reported by learners.
- **Support data-driven decision-making** for designing inclusive and adaptive e-learning systems.

By leveraging classification, clustering, and trend analysis models, this system enables better understanding and segmentation of learners—ultimately leading to **more engaging, efficient, and learner-centric online education experiences**.

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

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.

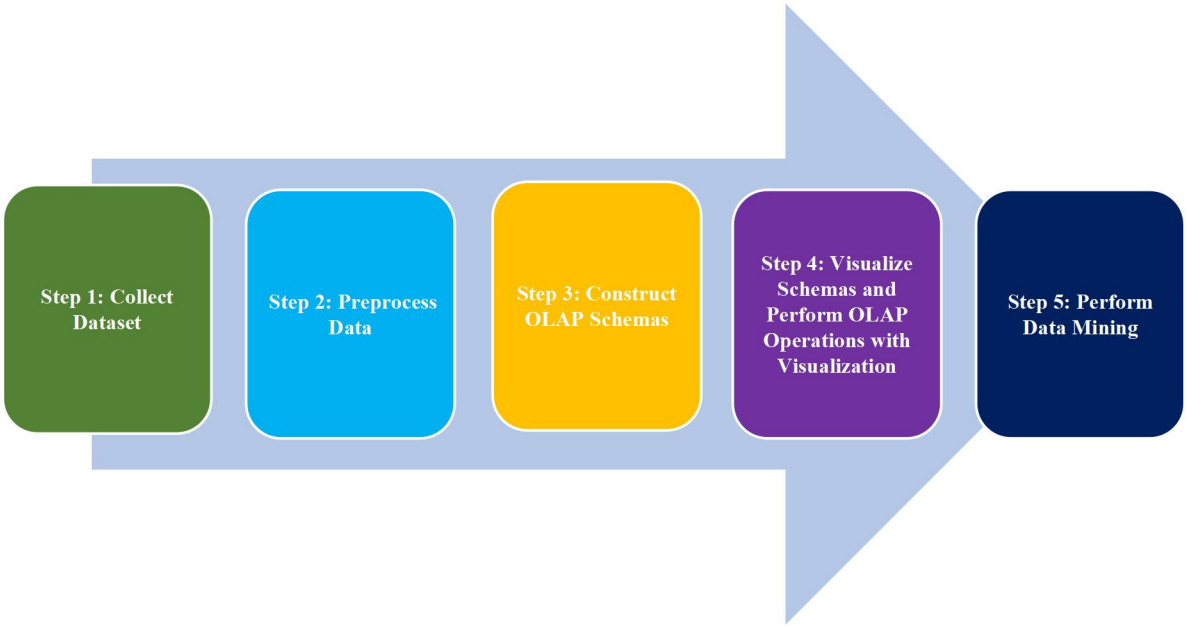
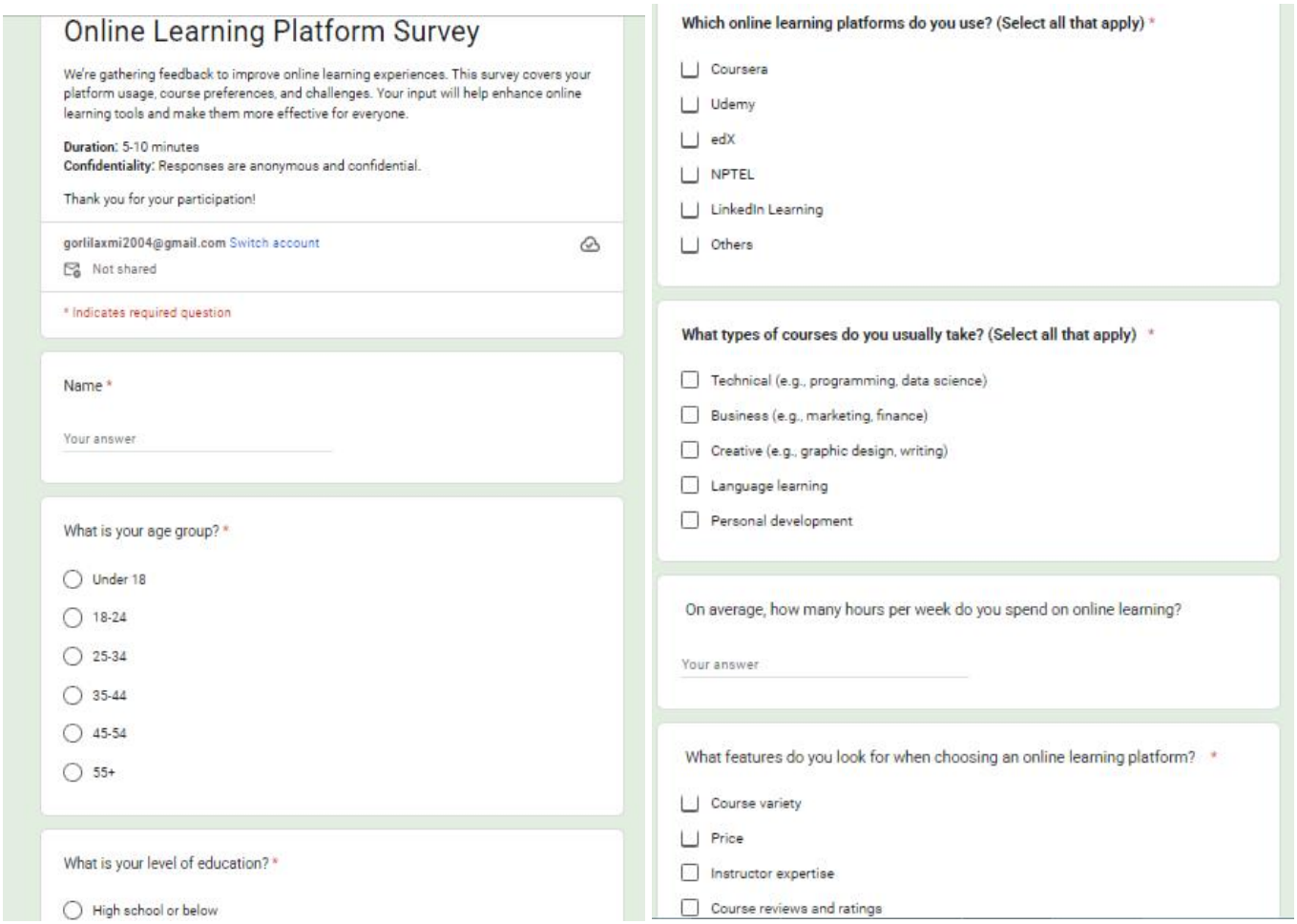


Fig4 : Knowledge Discovery From KDD

STEP-1: COLLECTING & EXPLORING DATASET

1.1. Extracting the Form to Collect Information from Users

The dataset was constructed by collecting responses through a structured Google Form survey aimed at understanding students' experiences with online learning platforms. The survey gathered a wide range of attributes related to the learners' interaction with digital education, including:



Online Learning Platform Survey

We're gathering feedback to improve online learning experiences. This survey covers your platform usage, course preferences, and challenges. Your input will help enhance online learning tools and make them more effective for everyone.

Duration: 5-10 minutes
Confidentiality: Responses are anonymous and confidential.

Thank you for your participation!

gorlilaxmi2004@gmail.com Switch account

Not shared

* Indicates required question

Name *

Your answer

What is your age group? *

- Under 18
- 18-24
- 25-34
- 35-44
- 45-54
- 55+

What is your level of education? *

- High school or below

Which online learning platforms do you use? (Select all that apply) *

- Coursera
- Udemy
- edX
- NPTEL
- LinkedIn Learning
- Others

What types of courses do you usually take? (Select all that apply) *

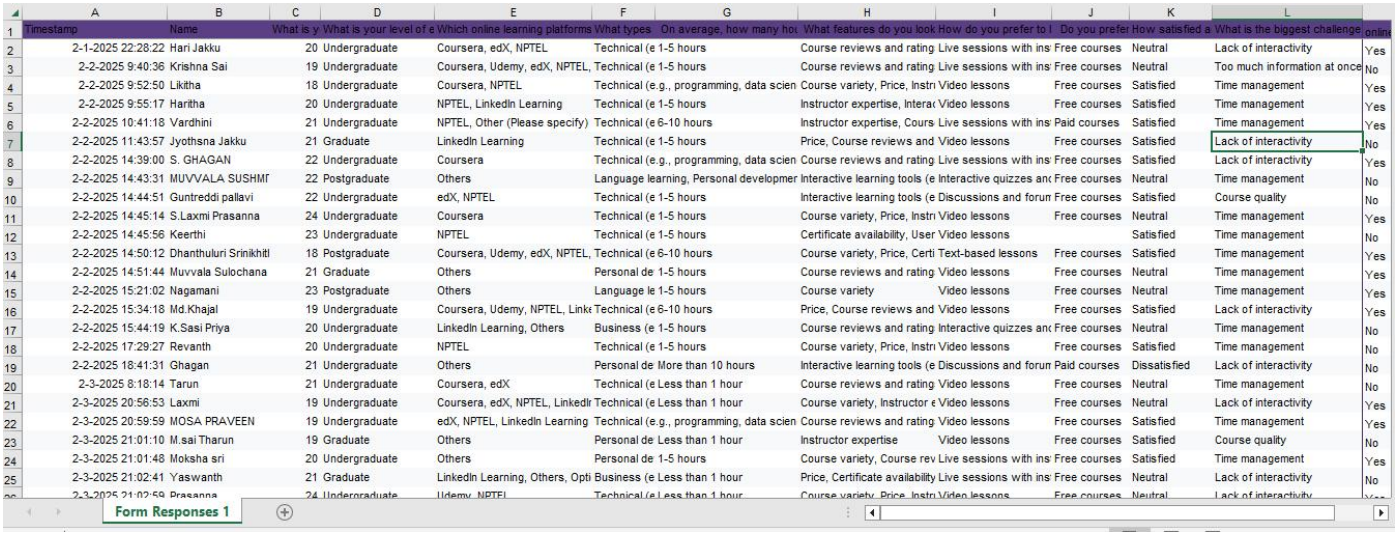
- Technical (e.g., programming, data science)
- Business (e.g., marketing, finance)
- Creative (e.g., graphic design, writing)
- Language learning
- Personal development

On average, how many hours per week do you spend on online learning?

Your answer

What features do you look for when choosing an online learning platform? *

- Course variety
- Price
- Instructor expertise
- Course reviews and ratings



	A	B	C	D	E	F	G	H	I	J	K	L	
1	Timestamp	Name	What is y	What is your level of e	Which online learning platforms	What types	On average, how many ho	What features do you look	How do you prefer to l	Do you prefer	How satisfied a	What is the biggest challenge	onlin
2	2-1-2025 22:28:22	Hari Jakku	20	Undergraduate	Coursera, edX, NPTEL	Technical (e	1-5 hours	Course reviews and rating	Live sessions with ins	Free courses	Neutral	Lack of interactivity	Yes
3	2-2-2025 9:40:36	Krishna Sai	19	Undergraduate	Coursera, Udemy, edX, NPTEL,	Technical (e	1-5 hours	Course reviews and rating	Live sessions with ins	Free courses	Neutral	Too much information at once	No
4	2-2-2025 9:52:50	Likitha	18	Undergraduate	Coursera, NPTEL	Technical (e.g., programming, data scien		Course variety, Price, Instru	Video lessons	Free courses	Satisfied	Time management	Yes
5	2-2-2025 9:55:17	Haritha	20	Undergraduate	NPTEL, LinkedIn Learning	Technical (e	1-5 hours	Instructor expertise, Intera	Video lessons	Free courses	Satisfied	Time management	Yes
6	2-2-2025 10:41:18	Vardhini	21	Undergraduate	NPTEL, Other (Please specify)	Technical (e	6-10 hours	Instructor expertise, Cours	Live sessions with ins	Paid courses	Satisfied	Time management	Yes
7	2-2-2025 11:43:57	Jyothsna Jakku	21	Graduate	LinkedIn Learning	Technical (e	1-5 hours	Price, Course reviews and	Video lessons	Free courses	Satisfied	Lack of interactivity	No
8	2-2-2025 14:39:00	S. GHAGAN	22	Undergraduate	Coursera	Technical (e.g., programming, data scien		Course reviews and rating	Live sessions with ins	Free courses	Satisfied	Lack of interactivity	Yes
9	2-2-2025 14:43:31	MUVVALA SUSHMI	22	Postgraduate	Others	Language learning, Personal developmer		Interactive learning tools (e	Interactive quizzes an	Free courses	Neutral	Time management	No
10	2-2-2025 14:44:51	Guntreddi pallavi	22	Undergraduate	edX, NPTEL	Technical (e	1-5 hours	Interactive learning tools (e	Discussions and forun	Free courses	Satisfied	Course quality	No
11	2-2-2025 14:45:14	S.Laxmi Prasanna	24	Undergraduate	Coursera	Technical (e	1-5 hours	Course variety, Price, Instru	Video lessons	Free courses	Neutral	Time management	Yes
12	2-2-2025 14:45:56	Keerthi	23	Undergraduate	NPTEL	Technical (e	1-5 hours	Certificate availability, User	Video lessons		Satisfied	Time management	No
13	2-2-2025 14:50:12	Dhanthuluri Srinikhitl	18	Postgraduate	Coursera, Udemy, edX, NPTEL,	Technical (e	6-10 hours	Course variety, Price, Certi	Text-based lessons	Free courses	Satisfied	Time management	Yes
14	2-2-2025 14:51:44	Muvvala Sulochana	21	Graduate	Others	Personal de	1-5 hours	Course reviews and rating	Video lessons	Free courses	Neutral	Time management	Yes
15	2-2-2025 15:21:02	Nagamani	23	Postgraduate	Others	Language le	1-5 hours	Course variety	Video lessons	Free courses	Neutral	Time management	Yes
16	2-2-2025 15:34:18	Md.Khajal	19	Undergraduate	Coursera, Udemy, NPTEL, Linkt	Technical (e	6-10 hours	Price, Course reviews and	Video lessons	Free courses	Satisfied	Lack of interactivity	Yes
17	2-2-2025 15:44:19	K.Sasi Priya	20	Undergraduate	LinkedIn Learning, Others	Business (e	1-5 hours	Course reviews and rating	Interactive quizzes an	Free courses	Neutral	Time management	No
18	2-2-2025 17:29:27	Revanth	20	Undergraduate	NPTEL	Technical (e	1-5 hours	Course variety, Price, Instru	Video lessons	Free courses	Satisfied	Time management	No
19	2-2-2025 18:41:31	Ghagan	21	Undergraduate	Others	Personal de	More than 10 hours	Interactive learning tools (e	Discussions and forun	Paid courses	Dissatisfied	Lack of interactivity	No
20	2-3-2025 8:18:14	Tarun	21	Undergraduate	Coursera, edX	Technical (e	Less than 1 hour	Course reviews and rating	Video lessons	Free courses	Neutral	Time management	No
21	2-3-2025 20:56:53	Laxmi	19	Undergraduate	Coursera, edX, NPTEL, LinkedIr	Technical (e	Less than 1 hour	Course variety, Instructor e	Video lessons	Free courses	Neutral	Lack of interactivity	Yes
22	2-3-2025 20:59:59	MOSA PRAVEEN	19	Undergraduate	edX, NPTEL, LinkedIn Learning	Technical (e.g., programming, data scien		Course reviews and rating	Video lessons	Free courses	Satisfied	Time management	Yes
23	2-3-2025 21:01:10	M.sai Tharun	19	Graduate	Others	Personal de	Less than 1 hour	Instructor expertise	Video lessons	Free courses	Satisfied	Course quality	No
24	2-3-2025 21:01:48	Moksha sri	20	Undergraduate	Others	Personal de	1-5 hours	Course variety, Course rev	Live sessions with ins	Free courses	Satisfied	Time management	Yes
25	2-3-2025 21:02:41	Yaswanth	21	Graduate	LinkedIn Learning, Others, Opti	Business (e	Less than 1 hour	Price, Certificate availability	Live sessions with ins	Free courses	Neutral	Lack of interactivity	No
	2-3-2025 21:02:59	Prasanna	24	Undergraduate	Udemy, NPTEL	Technical (e	Less than 1 hour	Course variety, Price, Instru	Video lessons	Free courses	Neutral	Lack of interactivity	

Form Responses 1

The attributs are:

1.Name
2.Age Group
3.Level of Education
4.Platform used for Learning
5.Type of Course (for ex. Techincal ,non-techinical,communication etc..)
6.Time Spent
7.What Features do you look For
8.Challenges Faced
9.Mode of Learning they Prefer(for ex.text-based,video based.)
10.Feedback of Online Learning(For ex.Satisfied,neutral,..)

11.Opinion on Online Learning is it Useful or not
12.what do you prefer either free or paid.
13.is online learning useful?
This rich dataset forms the foundation for conducting exploratory data analysis (EDA) and applying data mining techniques to extract meaningful insights into online learner behavior and preferences.

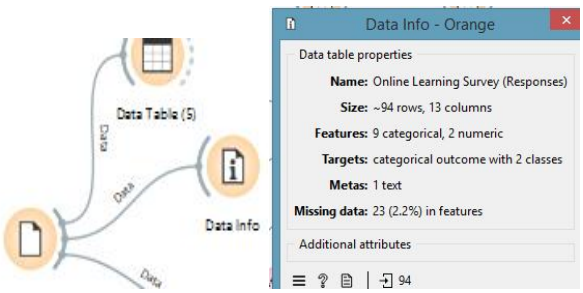
1.2 Defining Survey or Data Collection Methods

- **Online Surveys:** A structured questionnaire was distributed online, including multiple-choice questions to capture online learning preferences, platforms,time spent etc..

  Link: Student Career Preferences and Employment (google.com)

1.3Choosing Attributes for Analysis

- To perform meaningful analysis and derive actionable insights, the following **key attributes** were selected from the collected survey dataset:
- **Learner Demographics**
  Includes **age group**, **educational level**, and **gender (if applicable)** to observe learning trends across different categories of learners.
- **Preferred Learning Platforms**
  Identifies which online platforms (e.g., Coursera, Udemy, YouTube) are most popular and how preferences vary across demographics.
- **Course Type & Learning Style**
  Captures whether students favor **technical or non-technical** courses and their preferred **learning styles** (video-based, interactive, self-paced, etc.).
- **Time Spent on Learning**
  Analyzes the average duration learners dedicate to online learning and its impact on satisfaction or outcomes.
- **Platform Features & Expectations**
  Helps in understanding which features (e.g., certificates, hands-on projects, peer interaction) students value most.
- **Satisfaction Levels**
  Measures how satisfied learners are with their overall online learning experience across different platforms and course types.
- **Challenges Faced**
  Categorizes common issues such as **lack of interactivity**, **distractions**, **technical difficulties**, and **content overload** that hinder learning outcomes.
- These attributes were carefully chosen to ensure comprehensive **segmentation, trend identification, and prediction** of learner behaviors, preferences, and problem areas in online education..



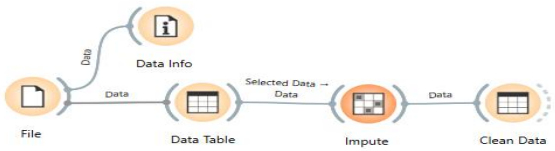
	rning platforms a	Name	Timestamp	st is your age grou	your level of edu	o you usually take	s per week do yo	lo you prefer to le	refer free or paid	lity of courses you	ou face when usi	atforms do you u	r when choosing
1	Yes	Hari Jakku	2025-02-01 22:2...	20	Undergraduate	Technical (e.g., ...	1-5 hours	Live sessions wi...	Free courses	Neutral	Lack of interact...	Coursera, edX, ...	Course reviews ...
2	No	Krishna Sai	2025-02-02 09:4...	19	Undergraduate	Technical (e.g., ...	1-5 hours	Live sessions wi...	Free courses	Neutral	Too much infor...	Coursera, Ude...	Course reviews ...
3	Yes	Likitha	2025-02-02 09:5...	18	Undergraduate	Technical (e.g., ...	?	Video lessons	Free courses	Satisfied	Time manage...	Coursera, NPTEL	Course variety, ...
4	Yes	Haritha	2025-02-02 09:5...	20	Undergraduate	Technical (e.g., ...	1-5 hours	Video lessons	Free courses	Satisfied	Time manage...	NPTEL, LinkedI...	Instructor exper...
5	Yes	Vardhini	2025-02-02 10:4...	21	Undergraduate	Technical (e.g., ...	6-10 hours	Live sessions wi...	Paid courses	Satisfied	Time manage...	NPTEL, Other (...	Instructor exper...
6	No	Jyothsna Jakku	2025-02-02 11:4...	21	Graduate	Technical (e.g., ...	1-5 hours	Video lessons	Free courses	Satisfied	Lack of interact...	LinkedIn Learni...	Price, Course re...
7	Yes	S. GHAGAN	2025-02-02 14:3...	22	Undergraduate	Technical (e.g., ...	?	Live sessions wi...	Free courses	Satisfied	Lack of interact...	Coursera	Course reviews ...
8	No	MUVVALA SUS...	2025-02-02 14:4...	22	Postgraduate	Language learn...	?	Interactive quiz...	Free courses	Neutral	Time manage...	Others	Interactive lear...
9	No	Guntreddi pallavi	2025-02-02 14:4...	22	Undergraduate	Technical (e.g., ...	1-5 hours	Discussions an...	Free courses	Satisfied	Course quality	edX, NPTEL	Interactive lear...
10	Yes	S.Laxmi Prasanna	2025-02-02 14:4...	24	Undergraduate	Technical (e.g., ...	1-5 hours	Video lessons	Free courses	Neutral	Time manage...	Coursera	Course variety, ...
11	No	Keerthi	2025-02-02 14:4...	23	Undergraduate	Technical (e.g., ...	1-5 hours	Video lessons	?	Satisfied	Time manage...	NPTEL	Certificate avail...
12	Yes	Dhanthuluri Sri...	2025-02-02 14:5...	18	Postgraduate	Technical (e.g., ...	6-10 hours	Text-based less...	Free courses	Satisfied	Time manage...	Coursera, Ude...	Course variety, ...
13	Yes	Muvvala Suloc...	2025-02-02 14:5...	21	Graduate	Personal develo...	1-5 hours	Video lessons	Free courses	Neutral	Time manage...	Others	Course reviews ...
14	Yes	Nagamani	2025-02-02 15:2...	23	Postgraduate	Language learn...	1-5 hours	Video lessons	Free courses	Neutral	Time manage...	Others	Course variety
15	Yes	Md.Khajal	2025-02-02 15:3...	19	Undergraduate	Technical (e.g., ...	6-10 hours	Video lessons	Free courses	Satisfied	Lack of interact...	Coursera, Ude...	Price, Course re...
16	No	K.Sasi Priya	2025-02-02 15:4...	20	Undergraduate	Business (e.g., ...	1-5 hours	Interactive quiz...	Free courses	Neutral	Time manage...	LinkedIn Learni...	Course reviews ...
17	No	Revanth	2025-02-02 17:2...	20	Undergraduate	Technical (e.g., ...	1-5 hours	Video lessons	Free courses	Satisfied	Time manage...	NPTEL	Course variety, ...
18	No	Ghagan	2025-02-02 18:4...	21	Undergraduate	Personal develo...	More than 10 h...	Discussions an...	Paid courses	Dissatisfied	Lack of interact...	Others	Interactive lear...
19	No	Tarun	2025-02-03 08:1...	21	Undergraduate	Technical (e.g., ...	Less than 1 hour	Video lessons	Free courses	Neutral	Time manage...	Coursera, edX	Course reviews ...
20	Yes	Laxmi	2025-02-03 20:5...	19	Undergraduate	Technical (e.g., ...	Less than 1 hour	Video lessons	Free courses	Neutral	Lack of interact...	Coursera, edX, ...	Course variety, ...
21	Yes	MOSA PRAVEEN	2025-02-03 20:5...	19	Undergraduate	Technical (e.g., ...	?	Video lessons	Free courses	Satisfied	Time manage...	edX, NPTEL, Lin...	Course reviews ...
22	No	M.sai Tharun	2025-02-03 21:0...	19	Graduate	Personal develo...	Less than 1 hour	Video lessons	Free courses	Satisfied	Course quality	Others	Instructor exper...
23	Yes	Moksha sri	2025-02-03 21:0...	20	Undergraduate	Personal develo...	1-5 hours	Live sessions wi...	Free courses	Satisfied	Time manage...	Others	Course variety, ...
24	No	Yaswanth	2025-02-03 21:0...	21	Graduate	Business (e.g., ...	Less than 1 hour	Live sessions wi...	Free courses	Neutral	Lack of interact...	LinkedIn Learni...	Price, Certificat...
25	Yes	Prasanna	2025-02-03 21:0...	24	Undergraduate	Technical (e.g., ...	Less than 1 hour	Video lessons	Free courses	Neutral	Lack of interact...	Udemy, NPTEL	Course variety, ...
26	Yes	Awad	2025-02-03 21:0...	24	Graduate	Technical (e.g., ...	1-5 hours	Video lessons	Free courses	Neutral	Lack of interact...	Coursera, Udemy	Price
27	No	K. Tejaswini	2025-02-03 21:0...	24	Graduate	Technical (e.g., ...	Less than 1 hour	Interactive quiz...	Free courses	Neutral	Lack of interact...	Coursera, NPTE...	Course variety, ...
28	Yes	Dheerendra	2025-02-03 21:1...	23	Graduate	Technical (e.g., ...	6-10 hours	Video lessons	Free courses	Neutral	Too much infor...	Coursera, Ude...	Course variety, ...
29	No	Maroju roshini	2025-02-03 21:1...	22	Undergraduate	Technical (e.g., ...	1-5 hours	Video lessons	Free courses	Satisfied	Course quality	Others	Course variety
30	Yes	Sai Kiran	2025-02-03 21:1...	21	Graduate	Technical (e.g., ...	1-5 hours	Interactive quiz...	Free courses	Satisfied	Lack of interact...	Coursera, Ude...	Course variety, ...
31	No	Haneesh Bandaru	2025-02-03 21:1...	19	Undergraduate	Technical (e.g., ...	1-5 hours	Video lessons	Free courses	Satisfied	Time manage...	Coursera, Ude...	Course variety, ...
32	No	S Giri Dhar Naidu	2025-02-03 21:1...	19	High school or ...	Language learn...	1-5 hours	Video lessons	?	Neutral	Too much infor...	Others	Interactive lear...
33	Yes	Karagana Bala s...	2025-02-03 21:1...	18	Graduate	Technical (e.g., ...	1-5 hours	Video lessons	Free courses	Satisfied	Course quality	Coursera, Ude...	Course variety, ...

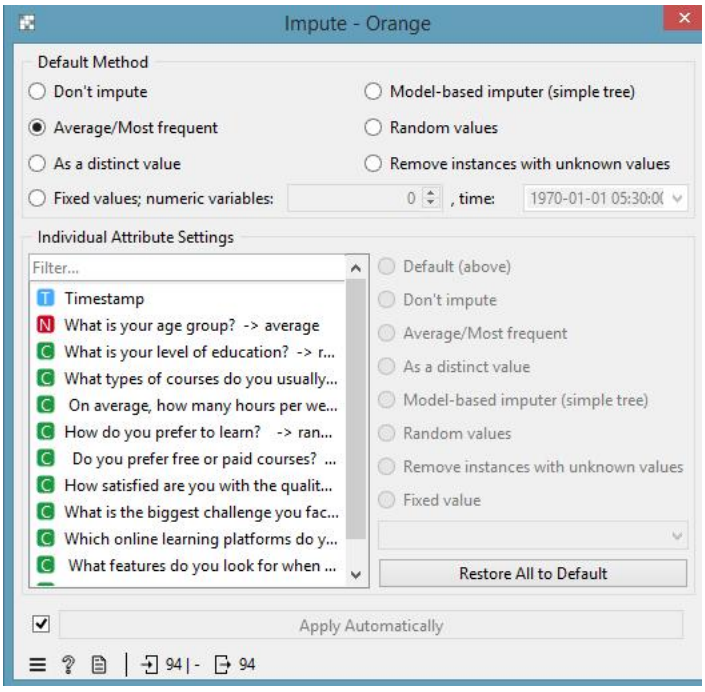
Step-2: PREPROCESS THE DATA

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 **rondom/most frequent values**.





2.2 Data Cleaning & Transformation

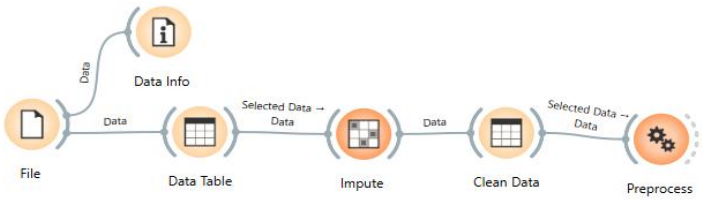
- Standardized text-based attributes**.**
- Converted categorical values into numerical form for analysis.

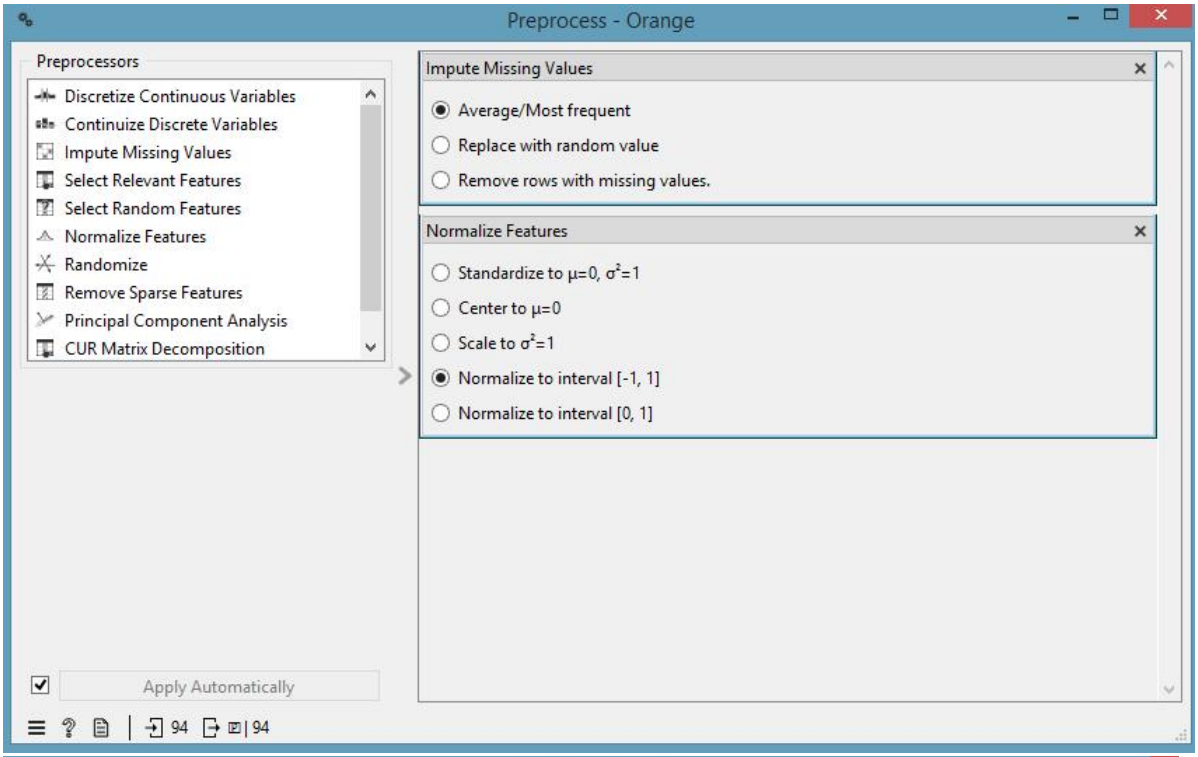
2.3 Removing Duplicates & Inconsistencies

- Removed duplicate survey responses.
- Ensured data consistency and integrity..
- Normalization data for better processing.

2.4 Normalization

- **Normalization** was applied to standardize numerical values.





preprocessed data - Orange

	rning platforms a	Name	Timestamp	at is your age grou	your level of edu	o you usually take	s per week do yo	lo you prefer to le	refer free or paid	lity of courses yo	ou face when usi	atforms do you u	r when choosing
1	Yes	Hari Jakku	2025-02-01 22:2...	-0.6190	Undergraduate	Technical (e.g., ...	1-5 hours	Live sessions wi...	Free courses	Neutral	Lack of interact...	Coursera, edX, ...	Course reviews ...
2	No	Krishna Sai	2025-02-02 09:4...	-0.7143	Undergraduate	Technical (e.g., ...	1-5 hours	Live sessions wi...	Free courses	Neutral	Too much infor...	Coursera, Ude...	Course reviews ...
3	Yes	Likitha	2025-02-02 09:5...	-0.8095	Undergraduate	Technical (e.g., ...	1-5 hours	Video lessons	Free courses	Satisfied	Time manage...	Coursera, NPTEL	Course variety, ...
4	Yes	Haritha	2025-02-02 09:5...	-0.6190	Undergraduate	Technical (e.g., ...	1-5 hours	Video lessons	Free courses	Satisfied	Time manage...	NPTEL, LinkedI...	Instructor exper...
5	Yes	Vardhini	2025-02-02 10:4...	-0.5238	Undergraduate	Technical (e.g., ...	6-10 hours	Live sessions wi...	Paid courses	Satisfied	Time manage...	NPTEL, Other (...	Instructor exper...
6	No	Jyothsna Jakku	2025-02-02 11:4...	-0.5238	Graduate	Technical (e.g., ...	1-5 hours	Video lessons	Free courses	Satisfied	Lack of interact...	LinkedIn Learni...	Price, Course re...
7	Yes	S. GHAGAN	2025-02-02 14:3...	-0.4286	Undergraduate	Technical (e.g., ...	1-5 hours	Live sessions wi...	Free courses	Satisfied	Lack of interact...	Coursera	Course reviews ...
8	No	MUVVALA SUS...	2025-02-02 14:4...	-0.4286	Postgraduate	Language learn...	1-5 hours	Interactive quiz...	Free courses	Neutral	Time manage...	Others	Interactive lear...
9	No	Guntreddi pallavi	2025-02-02 14:4...	-0.4286	Undergraduate	Technical (e.g., ...	1-5 hours	Discussions an...	Free courses	Satisfied	Course quality	edX, NPTEL	Interactive lear...
10	Yes	S.Laxmi Prasanna	2025-02-02 14:4...	-0.2381	Undergraduate	Technical (e.g., ...	1-5 hours	Video lessons	Free courses	Neutral	Time manage...	Coursera	Course variety, ...
11	No	Keerthi	2025-02-02 14:4...	-0.3333	Undergraduate	Technical (e.g., ...	1-5 hours	Video lessons	Free courses	Satisfied	Time manage...	NPTEL	Certificate avail...
12	Yes	Dhanthuluri Sri...	2025-02-02 14:5...	-0.8095	Postgraduate	Technical (e.g., ...	6-10 hours	Text-based less...	Free courses	Satisfied	Time manage...	Coursera, Ude...	Course variety, ...
13	Yes	Muvvala Suloc...	2025-02-02 14:5...	-0.5238	Graduate	Personal develo...	1-5 hours	Video lessons	Free courses	Neutral	Time manage...	Others	Course reviews ...
14	Yes	Nagamani	2025-02-02 15:2...	-0.3333	Postgraduate	Language learn...	1-5 hours	Video lessons	Free courses	Neutral	Time manage...	Others	Course variety
15	Yes	Md.Khajal	2025-02-02 15:3...	-0.7143	Undergraduate	Technical (e.g., ...	6-10 hours	Video lessons	Free courses	Satisfied	Lack of interact...	Coursera, Ude...	Price, Course re...
16	No	K.Sasi Priya	2025-02-02 15:4...	-0.6190	Undergraduate	Business (e.g., ...	1-5 hours	Interactive quiz...	Free courses	Neutral	Time manage...	LinkedIn Learni...	Course reviews ...
17	No	Revanth	2025-02-02 17:2...	-0.6190	Undergraduate	Technical (e.g., ...	1-5 hours	Video lessons	Free courses	Satisfied	Time manage...	NPTEL	Course variety, ...
18	No	Ghagan	2025-02-02 18:4...	-0.5238	Undergraduate	Personal develo...	More than 10 h...	Discussions an...	Paid courses	Dissatisfied	Lack of interact...	Others	Interactive lear...
19	No	Tarun	2025-02-03 08:1...	-0.5238	Undergraduate	Technical (e.g., ...	Less than 1 hour	Video lessons	Free courses	Neutral	Time manage...	Coursera, edX	Course reviews ...
20	Yes	Laxmi	2025-02-03 20:5...	-0.7143	Undergraduate	Technical (e.g., ...	Less than 1 hour	Video lessons	Free courses	Neutral	Lack of interact...	Coursera, edX, ...	Course variety, ...
21	Yes	MOSA PRAVEEN	2025-02-03 20:5...	-0.7143	Undergraduate	Technical (e.g., ...	1-5 hours	Video lessons	Free courses	Satisfied	Time manage...	edX, NPTEL, Lin...	Course reviews ...
22	No	M.sai Tharun	2025-02-03 21:0...	-0.7143	Graduate	Personal develo...	Less than 1 hour	Video lessons	Free courses	Satisfied	Course quality	Others	Instructor exper...
23	Yes	Moksha sri	2025-02-03 21:0...	-0.6190	Undergraduate	Personal develo...	1-5 hours	Live sessions wi...	Free courses	Satisfied	Time manage...	Others	Course variety, ...
24	No	Yaswanth	2025-02-03 21:0...	-0.5238	Graduate	Business (e.g., ...	Less than 1 hour	Live sessions wi...	Free courses	Neutral	Lack of interact...	LinkedIn Learni...	Price, Certificat...
25	Yes	Prasanna	2025-02-03 21:0...	-0.2381	Undergraduate	Technical (e.g., ...	Less than 1 hour	Video lessons	Free courses	Neutral	Lack of interact...	Udemy, NPTEL	Course variety, ...
26	Yes	Awad	2025-02-03 21:0...	-0.2381	Graduate	Technical (e.g., ...	1-5 hours	Video lessons	Free courses	Neutral	Lack of interact...	Coursera, Udemy	Price
27	No	K. Tejaswini	2025-02-03 21:0...	-0.2381	Graduate	Technical (e.g., ...	Less than 1 hour	Interactive quiz...	Free courses	Neutral	Lack of interact...	Coursera, NPTE...	Course variety, ...
28	Yes	Dheerendra	2025-02-03 21:1...	-0.3333	Graduate	Technical (e.g., ...	6-10 hours	Video lessons	Free courses	Neutral	Too much infor...	Coursera, Ude...	Course variety, ...
29	No	Maroju roshini	2025-02-03 21:1...	-0.4286	Undergraduate	Technical (e.g., ...	1-5 hours	Video lessons	Free courses	Satisfied	Course quality	Others	Course variety
30	Yes	Sai Kiran	2025-02-03 21:1...	-0.5238	Graduate	Technical (e.g., ...	1-5 hours	Interactive quiz...	Free courses	Satisfied	Lack of interact...	Coursera, Ude...	Course variety, ...
31	No	Haneesh Bandaru	2025-02-03 21:1...	-0.7143	Undergraduate	Technical (e.g., ...	1-5 hours	Video lessons	Free courses	Satisfied	Time manage...	Coursera, Ude...	Course variety, ...
32	No	S Giri Dhar Naidu	2025-02-03 21:1...	-0.7143	High school or ...	Language learn...	1-5 hours	Video lessons	Free courses	Neutral	Too much infor...	Others	Interactive lear...
33	Yes	Karagana Bala s...	2025-02-03 21:1...	-0.8095	Graduate	Technical (e.g., ...	1-5 hours	Video lessons	Free courses	Satisfied	Course quality	Coursera, Ude...	Course variety, ...

STEP-3: CREATION OF DATABASE CONSTRUCT OLAP SCHEMAS

The above table was normalized and divided into multiple tables:

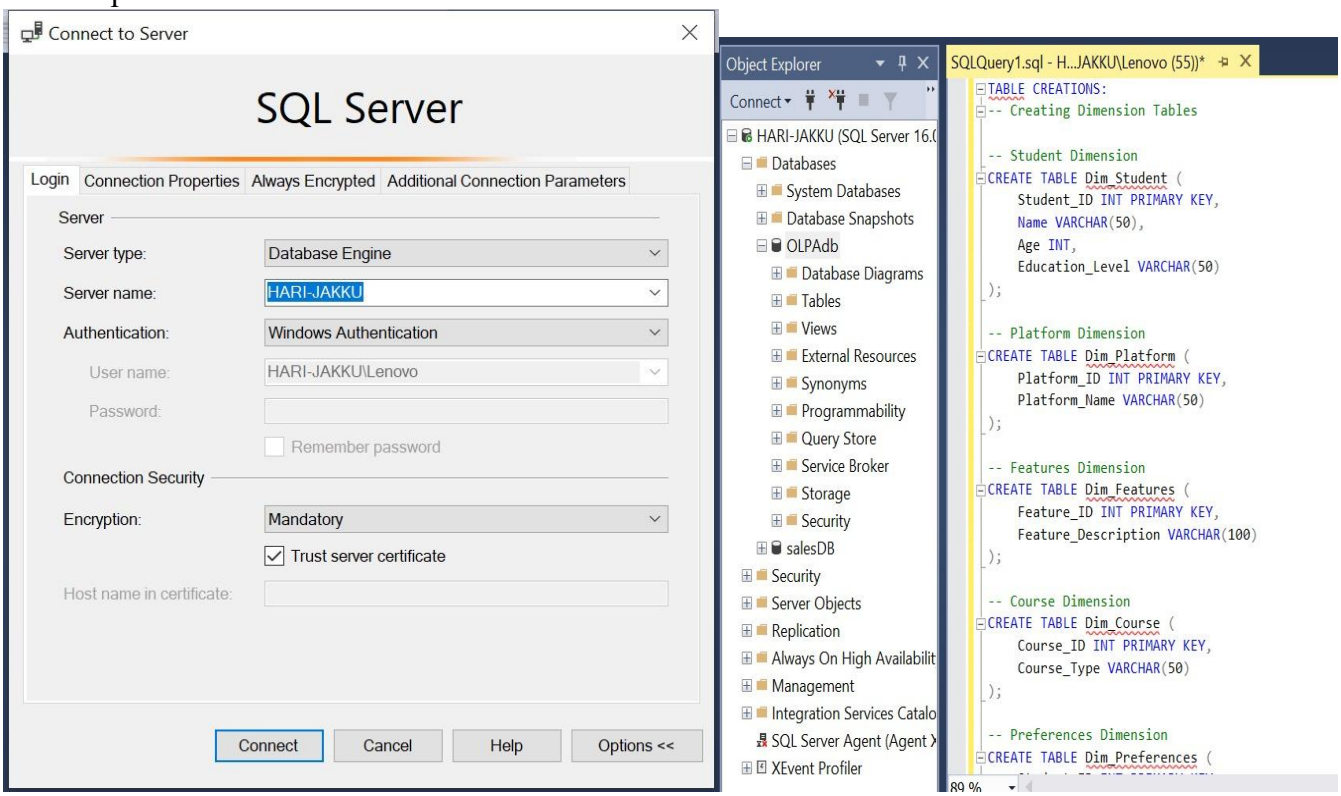
Dimension Tables

Table Name	Attribute	Data Type	Key Type
Dim_Student	Student_ID	INT	Primary Key
	Name	VARCHAR(50)	
	Age	INT	
	Education_Level	VARCHAR(50)	
Dim_Platform	Platform_ID	INT	Primary Key
	Platform_Name	VARCHAR(50)	
Dim_Features	Feature_ID	INT	Primary Key
	Feature_Description	VARCHAR(100)	
Dim_Course	Course_ID	INT	Primary Key
	Course_Type	VARCHAR(50)	
Dim_Preferences	Student_ID	INT	Primary Key, Foreign Key → `Dim_Student(Student_ID)`
	Learning_Style	VARCHAR(50)	
	Course_Type	VARCHAR(50)	
	Learning_Challenge	VARCHAR(100)	

Fact Tables:

Fact_Course_Engagement	Student_ID	INT	Primary Key (composite), Foreign Key → `Dim_Student(Student_ID)`
	Platform_ID	INT	Primary Key (composite), Foreign Key → `Dim_Platform(Platform_ID)`
	Course_ID	INT	Primary Key (composite), Foreign Key → `Dim_Course(Course_ID)`
	Feature_ID	INT	Foreign Key → `Dim_Features(Feature_ID)`
	Learning_Method	VARCHAR(50)	
Fact_Online_Learning	Student_ID	INT	Primary Key (composite), Foreign Key → `Dim_Student(Student_ID)`
	Course_ID	INT	Primary Key (composite), Foreign Key → `Dim_Course(Course_ID)`
	Time_Spent	INT	
	Satisfaction_Level	INT	

We have created a database Student and inserted the data into the tables. generated sql queries to perform OLAP operations.



Generate SQL Queries for OLAP Schema Construction

3.1 Designing the Schemas:

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

3.2 SQL Queries :

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

STEP-4: VISUALIZE SCHEMAS

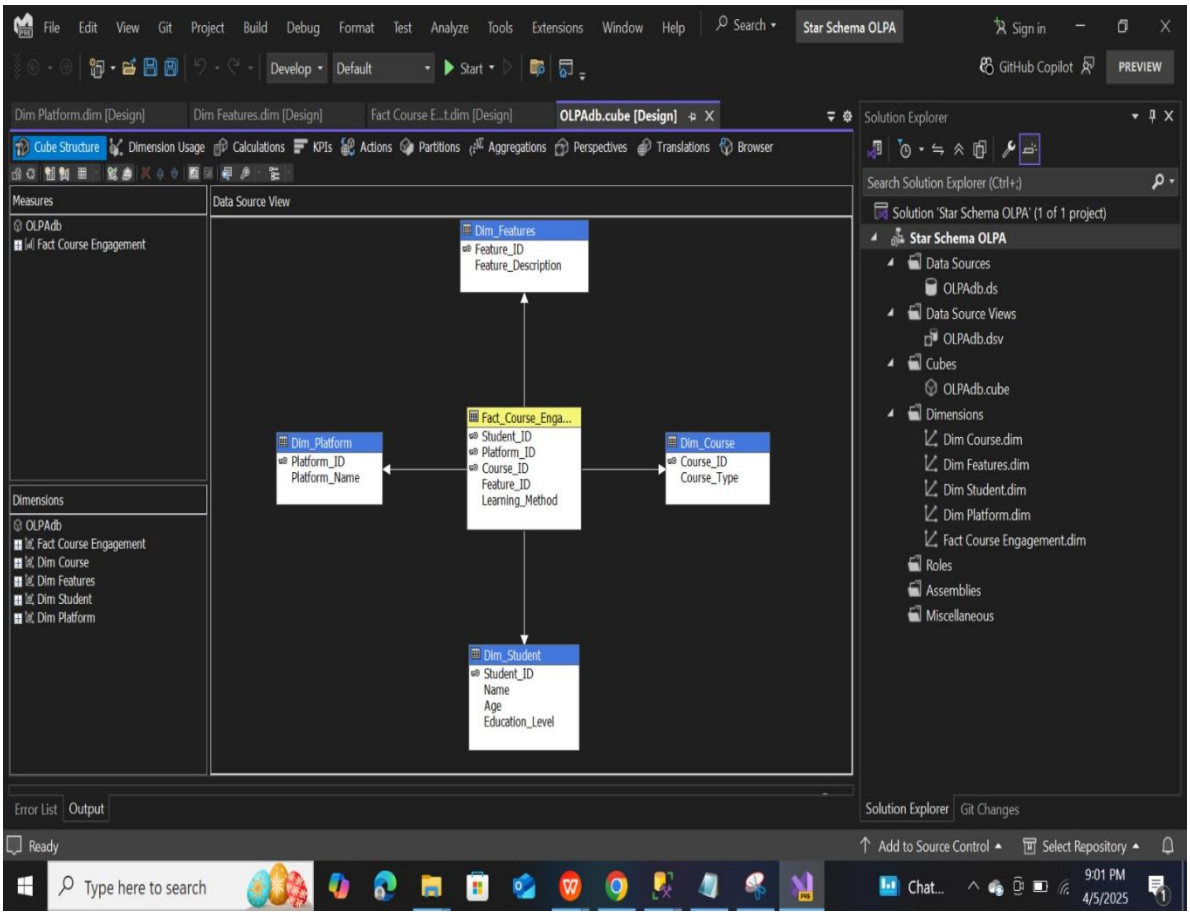
- Developed a multidimensional analysis project using Visual Studio.
- Configured Data Source & Data Source View, establishing connections to the database and defining table relationships.
- Designed database schema diagrams to visualize data structure.
- Validated table relationships to ensure data integrity.
- Created Cubes & Measures, defining fact tables, dimensions, and key performance measures for analysis.
- Initially Build deploy and process all Multi Dimensional cubes.
- Visualize them in SSAS server
- perform OLAP operations.

4.1STAR SCHEMA:

The **Star Schema** is a denormalized database schema used in OLAP, where a central **Fact Table** (containing measurable data) is directly connected to multiple **Dimension Tables** (such as platform,course etc) in a star-like structure.

4.1.1 Design & Visualize the Schema

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



4.1.2 Deploy the Data Warehouse & Load Data

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

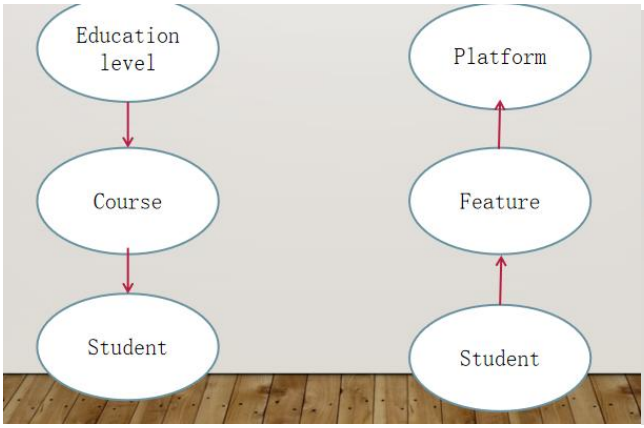
4.1.3 Create & Execute OLAP Queries

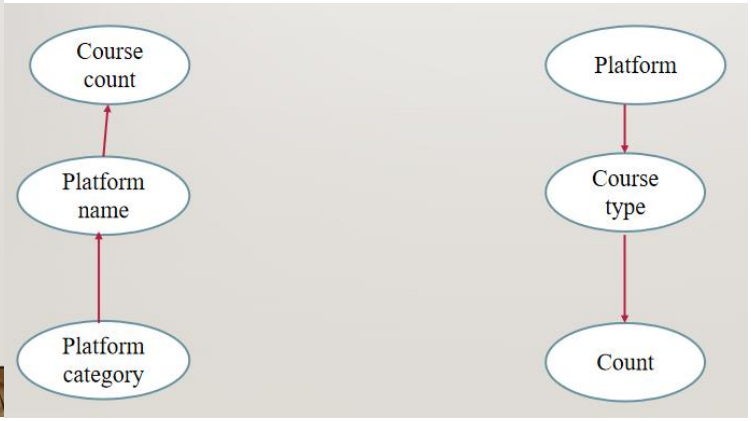
- Write OLAP queries to perform data analysis.
- Use ROLLUP,SLICE, DICE, DRILL-DOWN, and PIVOT operations for multi-dimensional analysis.

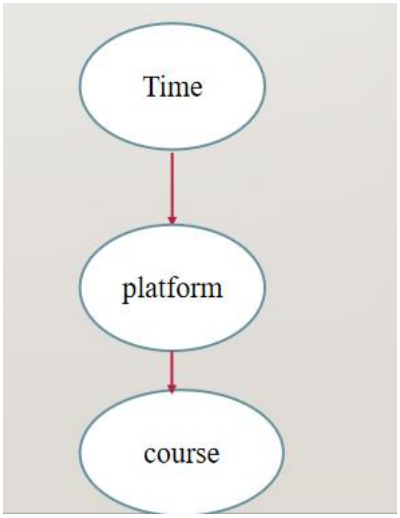
4.1.4 Perform OLAP Operations

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

Concept Hierarchies used :







MDX Queries OLAP operations in STAR SCHEMA:

A) Summarize the data by Student, Course Type,Education level:
Roll-Up (Aggregation of Course):

```
SELECT
   [Dim Student].[Education Level].MEMBERS ON COLUMNS,
   [Dim Course].[Course Type].MEMBERS ON ROWS
FROM [OLPAdb]
```

Output:

Messages | Results

	All	Diploma	Postgraduate	Undergraduate
All	10	2	3	5
AI & ML	1	(null)	1	(null)
Business Management	1	(null)	1	(null)
Cloud Computing	1	1	(null)	(null)
Cybersecurity	1	(null)	1	(null)
Data Science	1	(null)	(null)	1
Finance	1	(null)	(null)	1
Graphic Design	1	1	(null)	(null)
Marketing	1	(null)	(null)	1
Software Engineering	1	(null)	(null)	1
Web Development	1	(null)	(null)	1

B) Drill down from platforms to individual students:?

Drill-Down:

```
SELECT
   [Dim Platform].[Platform Name].MEMBERS ON COLUMNS,
   NONEMPTY(
      CROSSJOIN(
         [Dim Features].[Feature Description].MEMBERS,
         [Dim Student].[Student ID].MEMBERS
      )
   ) ON ROWS
FROM [OLPAdb]
```

OUTPUT:

Messages | Results

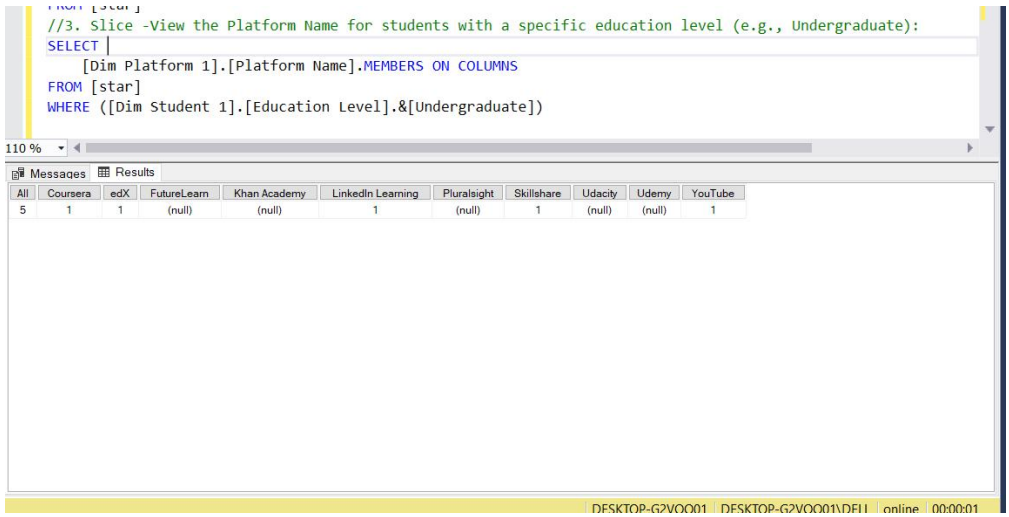
	All	1	2	3	4	5	6	7	8	9	10
All	10	1	1	1	1	1	1	1	1	1	1
Coursera	1	1	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)
edX	1	(null)	(null)	1	(null)	(null)	(null)	(null)	(null)	(null)	(null)
FutureLearn	1	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	1	(null)
Khan Academy	1	(null)	(null)	(null)	1	(null)	(null)	(null)	(null)	(null)	(null)
LinkedIn Learning	1	(null)	(null)	(null)	(null)	1	(null)	(null)	(null)	(null)	(null)
Pluralsight	1	(null)	(null)	(null)	(null)	(null)	1	(null)	(null)	(null)	(null)
Skillshare	1	(null)	(null)	(null)	(null)	(null)	(null)	1	(null)	(null)	(null)
Udacity	1	(null)	(null)	(null)	(null)	(null)	(null)	(null)	1	(null)	(null)
Udemy	1	(null)	1	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)
YouTube	1	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	1

C) View the Platform Name for students with a specific education level (e.g., Undergraduate)

Slice :

```
SELECT
   [Dim Platform 1].[Platform Name].MEMBERS ON COLUMNS
FROM [star]
WHERE ([Dim Student 1].[Education Level].&[Undergraduate])
```

OUTPUT:

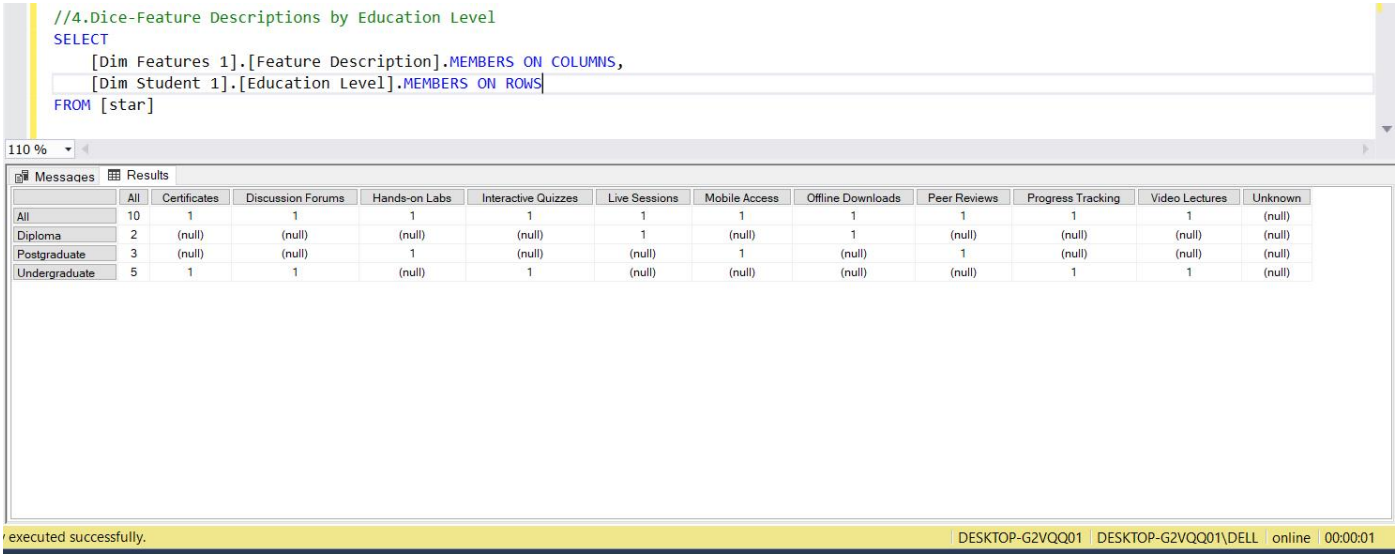


D) Feature Descriptions by Education Level

Dice:

```
SELECT
    [Dim Features 1].[Feature Description].MEMBERS ON COLUMNS,
    [Dim Student 1].[Education Level].MEMBERS ON ROWS
FROM [star]
```

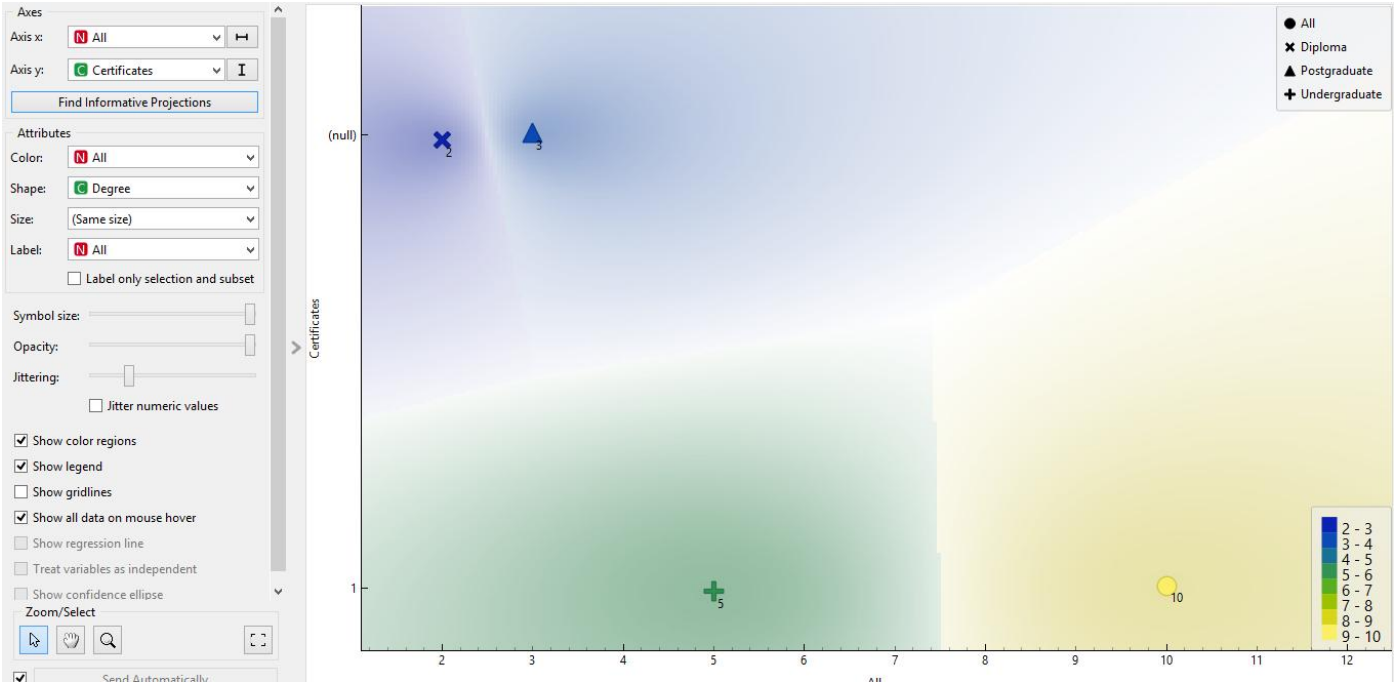
OUTPUT:



	All	Certificates	Discussion Forums	Hands-on Labs	Interactive Quizzes	Live Sessions	Mobile Access	Offline Downloads	Peer Reviews	Progress Tracking	Video Lectures	Unknown
All	10	1	1	1	1	1	1	1	1	1	1	(null)
Diploma	2	(null)	(null)	(null)	(null)	1	(null)	1	(null)	(null)	(null)	(null)
Postgraduate	3	(null)	(null)	1	(null)	(null)	1	(null)	1	(null)	(null)	(null)
Undergraduate	5	1	1	(null)	1	(null)	(null)	(null)	(null)	1	1	(null)

4.1.5 Visualize OLAP Results:

This visualization explores the relationship between the **number of certificates earned** and an aggregated metric labeled "All" (potentially representing overall user activity, score, or performance). The plot categorizes users by **education level** using different shapes and colors.



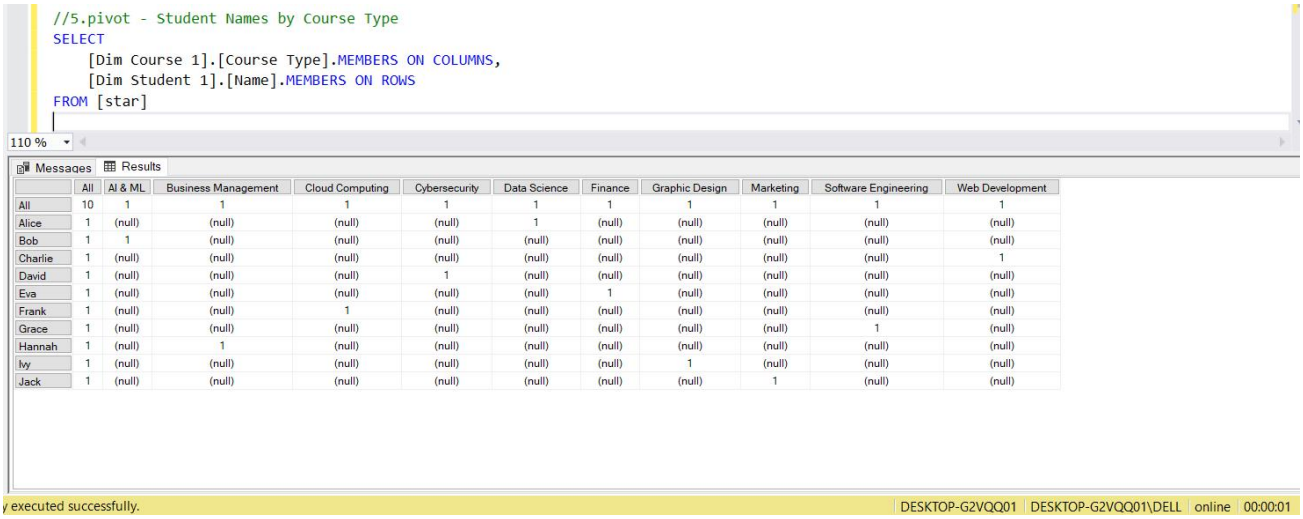
Scatter Plot Configuration:

- **X-Axis**: All
  (Represents an aggregated metric — could be engagement, score, or time spent)
- **Y-Axis**: Certificates
  (Number of certificates achieved by users)
- **Color**: All
  (Color intensity changes based on value — from blue [low] to yellow [high])
- **Shape**: Degree
  (Represents user's educational level:)
  - ✖ Diploma
  - ▲ Postgraduate
  - ✚ Undergraduate
  - ● All

E) Student Names by Course Type
Pivot (Rearranging Dimensions):

```
SELECT
  [Dim Course 1].[Course Type].MEMBERS ON COLUMNS,
  [Dim Student 1].[Name].MEMBERS ON ROWS
FROM [star]
```

OUTPUT:



```
//5.pivot - Student Names by Course Type
SELECT
    [Dim Course 1].[Course Type].MEMBERS ON COLUMNS,
    [Dim Student 1].[Name].MEMBERS ON ROWS
FROM [star]
```

110 %

Messages | Results

	All	AI & ML	Business Management	Cloud Computing	Cybersecurity	Data Science	Finance	Graphic Design	Marketing	Software Engineering	Web Development
All	10	1	1	1	1	1	1	1	1	1	1
Alice	1	(null)	(null)	(null)	(null)	1	(null)	(null)	(null)	(null)	(null)
Bob	1	1	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)
Charlie	1	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	1
David	1	(null)	(null)	(null)	1	(null)	(null)	(null)	(null)	(null)	(null)
Eva	1	(null)	(null)	(null)	(null)	(null)	1	(null)	(null)	(null)	(null)
Frank	1	(null)	(null)	1	(null)	(null)	(null)	(null)	(null)	(null)	(null)
Grace	1	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	1	(null)
Hannah	1	(null)	1	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)
Ivy	1	(null)	(null)	(null)	(null)	(null)	(null)	1	(null)	(null)	(null)
Jack	1	(null)	(null)	(null)	(null)	(null)	(null)	(null)	1	(null)	(null)

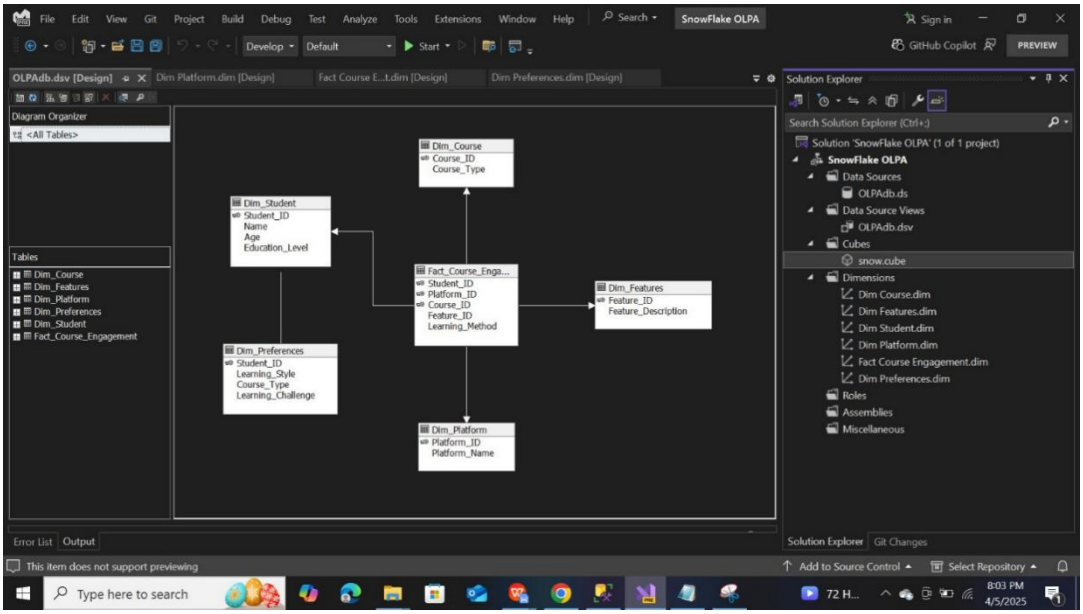
y executed successfully. DESKTOP-G2VQQ01 | DESKTOP-G2VQQ01\DELL | online | 00:00:01

4.2<u>SNOWFLAKE SCHEMA:</u>

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, Course Engagement)
- Identify **Dimension Tables** (e.g., student,preference,course etc).
- Normalize dimension tables by breaking them into **sub-dimensions** (e.g., student→preference).
- Ensure **foreign key relationships** between tables.



4.2.2 Deploy & Load Data into the Snowflake Schema

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

4.2.3 Create & Execute OLAP Queries

- Write **SQL queries** for analytical processing:
  1) **ROLLUP** – Aggregate data across different levels.
  2) **CUBE** – Compute multi-dimensional aggregates.
  3) **DRILL-DOWN** – View data at finer granularity.
  4) **SLICE & DICE** – Filter and analyze subsets of data.

4.2.4 Perform OLAP Operations

- Use OLAP processing to retrieve and manipulate large datasets efficiently.
- Run complex queries on multi-dimensional data using **MDX (Multi-Dimensional Expressions)** or SQL-based OLAP tools.

<u>MDX Queries OLAP operations in SNOWFLAKE SCHEMA:</u>

(A) Course engagement count per platform?

Roll-Up (Aggregation):

```
SELECT
   {[Measures].[Fact Course Engagement Count]} ON COLUMNS,
   {[Dim Platform].[Platform Name].Children} ON ROWS
FROM [snow]
```

OUTPUT:

```
--SNOW FLAKE  SCHEMA---
--1.ROLL UP

SELECT
    {[Measures].[Fact Course Engagement Count]} ON COLUMNS,
    {[Dim Platform].[Platform Name].Children} ON ROWS
FROM [snow]
```

89 %

Messages | Results

	Fact Course Engagement Count
Coursera	1
edX	1
FutureLearn	1
Khan Academy	1
LinkedIn Learning	1
Pluralsight	1
Skillshare	1
Udacity	1
Udemy	1
YouTube	1

(B) View All Degree Program

Drill-Down

```
SELECT
   {[Measures].[Fact Course Engagement Count]} ON COLUMNS,
   CROSSJOIN(
      [Dim Platform].[Platform Name].Children,
      [Dim Course].[Course Type].Children
   ) ON ROWS
FROM [snow]
```

OUTPUT:

```
--2. DRILL DOWN---

SELECT
    {[Measures].[Fact Course Engagement Count]} ON COLUMNS,
    CROSSJOIN(
        [Dim Platform].[Platform Name].Children,
        [Dim Course].[Course Type].Children
    ) ON ROWS
FROM [snow]
```

89 %

Messages | Results

		Fact Course Engagement Count
Coursera	AI & ML	(null)
Coursera	Business Management	(null)
Coursera	Cloud Computing	(null)
Coursera	Cybersecurity	(null)
Coursera	Data Science	1
Coursera	Finance	(null)
Coursera	Graphic Design	(null)
Coursera	Marketing	(null)
Coursera	Software Engineering	(null)
Coursera	Web Development	(null)
edX	AI & ML	(null)
edX	Business Management	(null)
edX	Cloud Computing	(null)
edX	Cybersecurity	(null)
edX	Data Science	(null)
edX	Finance	(null)

:

(C) Display Degree Programs and Project Completion Status
Slice

```
SELECT
   {[Measures].[Fact Course Engagement Count]} ON COLUMNS
FROM [snow]
```

WHERE ([Dim Platform].[Platform Name].[Coursera])

OUTPUT:

```
--3,SLICE--
SELECT
    {[Measures].[Fact Course Engagement Count]} ON COLUMNS
FROM [snow]
WHERE ([Dim Platform].[Platform Name].[Coursera])
```

39 %

Messages | Results

Fact Course Engagement Count
1

(D) View Project Completion Status

Dice

```
SELECT
   {[Measures].[Fact Course Engagement Count]} ON COLUMNS,
   [Dim Platform].[Platform Name].Members ON ROWS
FROM [snow]
```

OUTPUT:

```
--4.DICE--
SELECT
    {[Measures].[Fact Course Engagement Count]} ON COLUMNS,
    [Dim Platform].[Platform Name].Members ON ROWS
FROM [snow]
```

9 %

Messages | Results

	Fact Course Engagement Count
All	10
Coursera	1
edX	1
FutureLearn	1
Khan Academy	1
LinkedIn Learning	1
Pluralsight	1
Skillshare	1
Udacity	1
Udemy	1
YouTube	1

(E) Display Degree Programs with project completion status 'Completed' and Degree Program names less than 'M' alphabetically.

Pivot (Rearranging Dimensions):

SELECT

{[Dim Student].[Age].Children} ON COLUMNS,

{[Dim Platform].[Platform Name].Children} ON ROWS

FROM [snow]

OUTPUT:

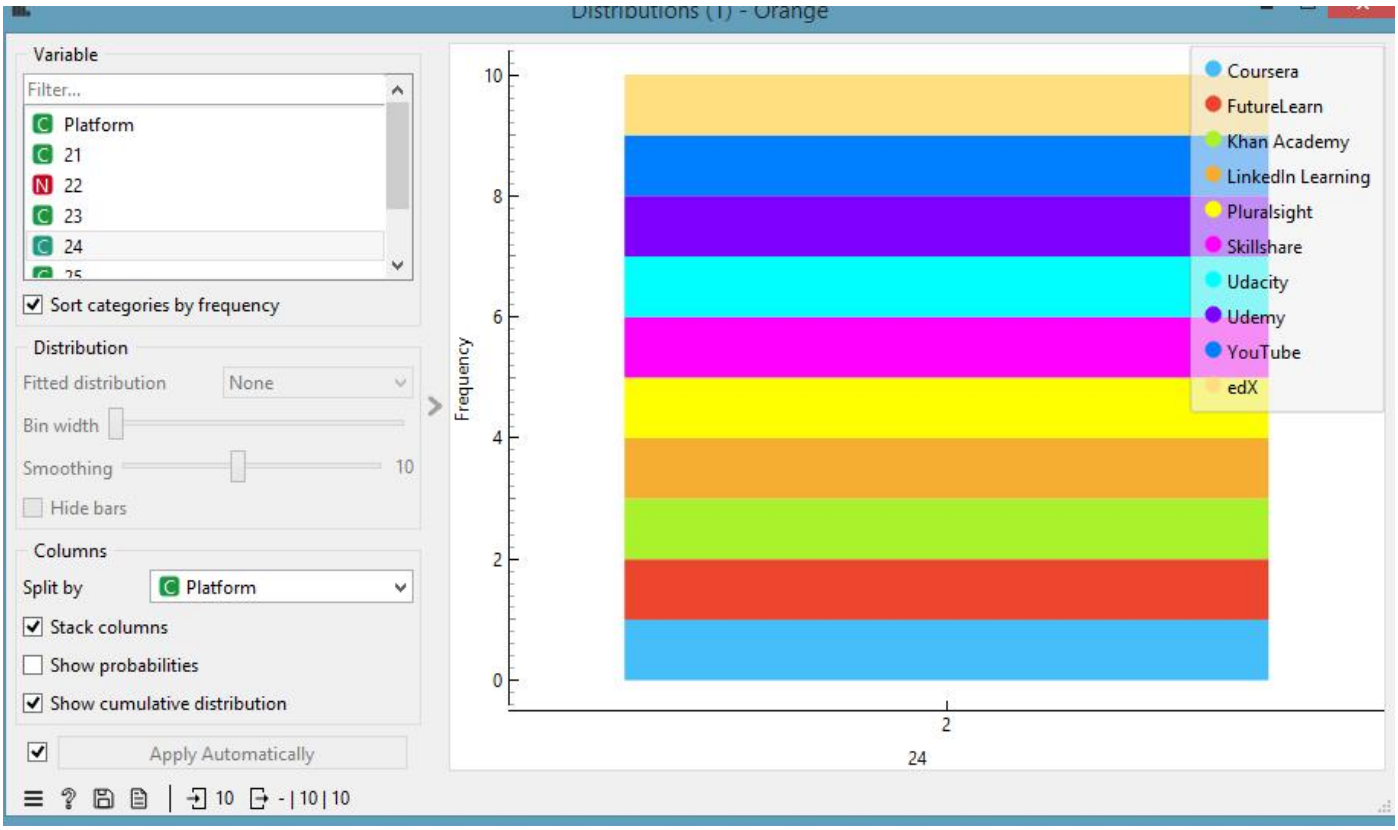
```
--5.pivot--
SELECT
    {[Dim Student].[Age].Children} ON COLUMNS,
    {[Dim Platform].[Platform Name].Children} ON ROWS
FROM [snow]
```

9 %

Messages | Results

	21	22	23	24	25	26
Coursera	1	3	2	2	1	1
edX	1	3	2	2	1	1
FutureLearn	1	3	2	2	1	1
Khan Academy	1	3	2	2	1	1
LinkedIn Learning	1	3	2	2	1	1
Pluralsight	1	3	2	2	1	1
Skillshare	1	3	2	2	1	1
Udacity	1	3	2	2	1	1
Udemy	1	3	2	2	1	1
YouTube	1	3	2	2	1	1

Visualize OLAP Results:



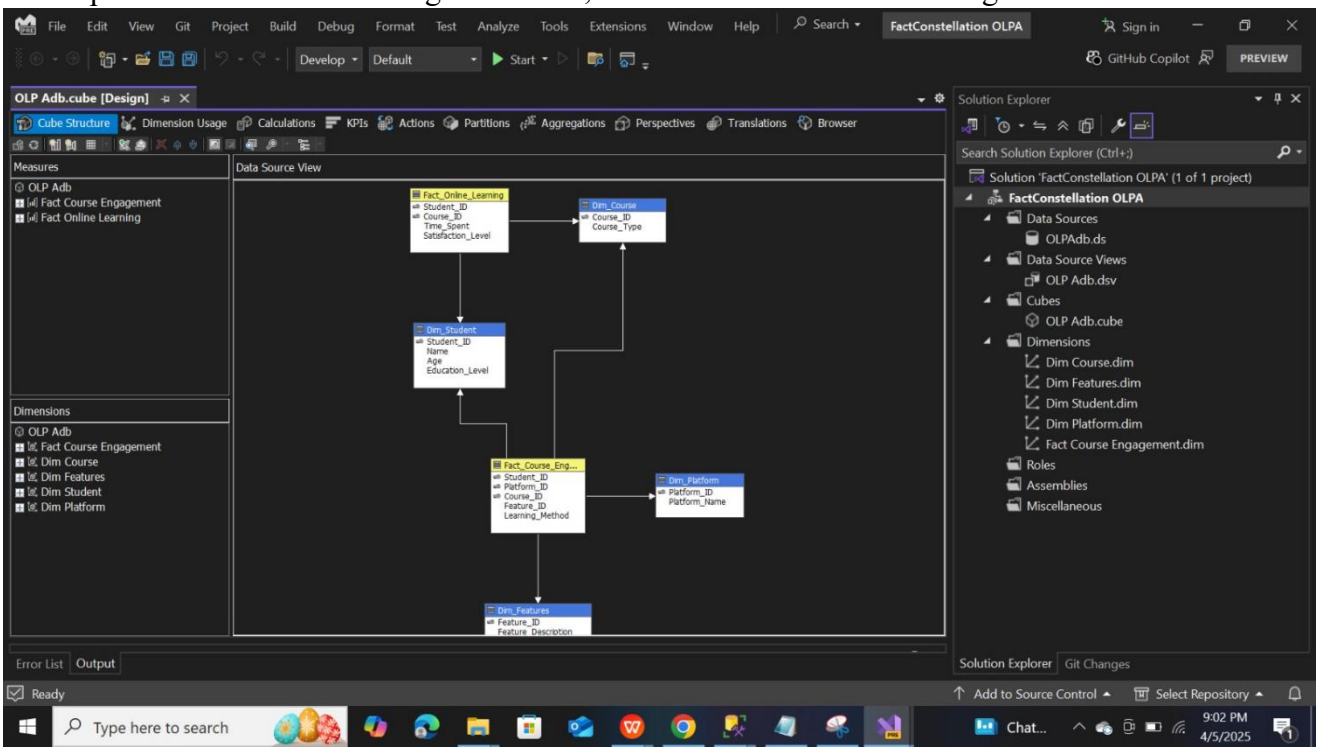
To analyze the popularity and usage of various online learning platforms among users, a **horizontal bar plot** is used. This visualization offers insights into which platforms are most frequently accessed or preferred, helping in understanding user behavior and trends in e-learning.

Bar Plot Configuration:

- **X-Axis**: Frequency (Number of users)
- **Y-Axis**: Platform ID (Representing different online platforms)
- **Split By**: Platform (Each bar colored uniquely)
- **Observation**:
  - Each colored horizontal bar represents a specific online learning platform.
  - The frequency value for each platform is uniformly **1**, indicating equal user representation for each platform in the sample.
  - Some Platforms included:
  - This uniform distribution suggests either a sample designed for balance or equal interest across various platforms.

4.3 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.



4.3.1 Design & Visualize the FACT CONSTELLATION SCHEMA

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

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

4.3.2 Deploy & Load Data into the Snowflake Schema:

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

4.3.3 Create & Execute OLAP Queries:

1. **ROLLUP**: Aggregate data at different levels.
2. **CUBE**: Compute multi-dimensional aggregates.
3. **DRILL-DOWN**: View data with more detail.
4. **SLICE & DICE**: Filter and analyze data subsets.

4.3.4 Perform OLAP Operations:

1. Use OLAP tools to retrieve and manipulate large datasets.
2. Run complex queries using **MDX** or **SQL-based OLAP** tools.

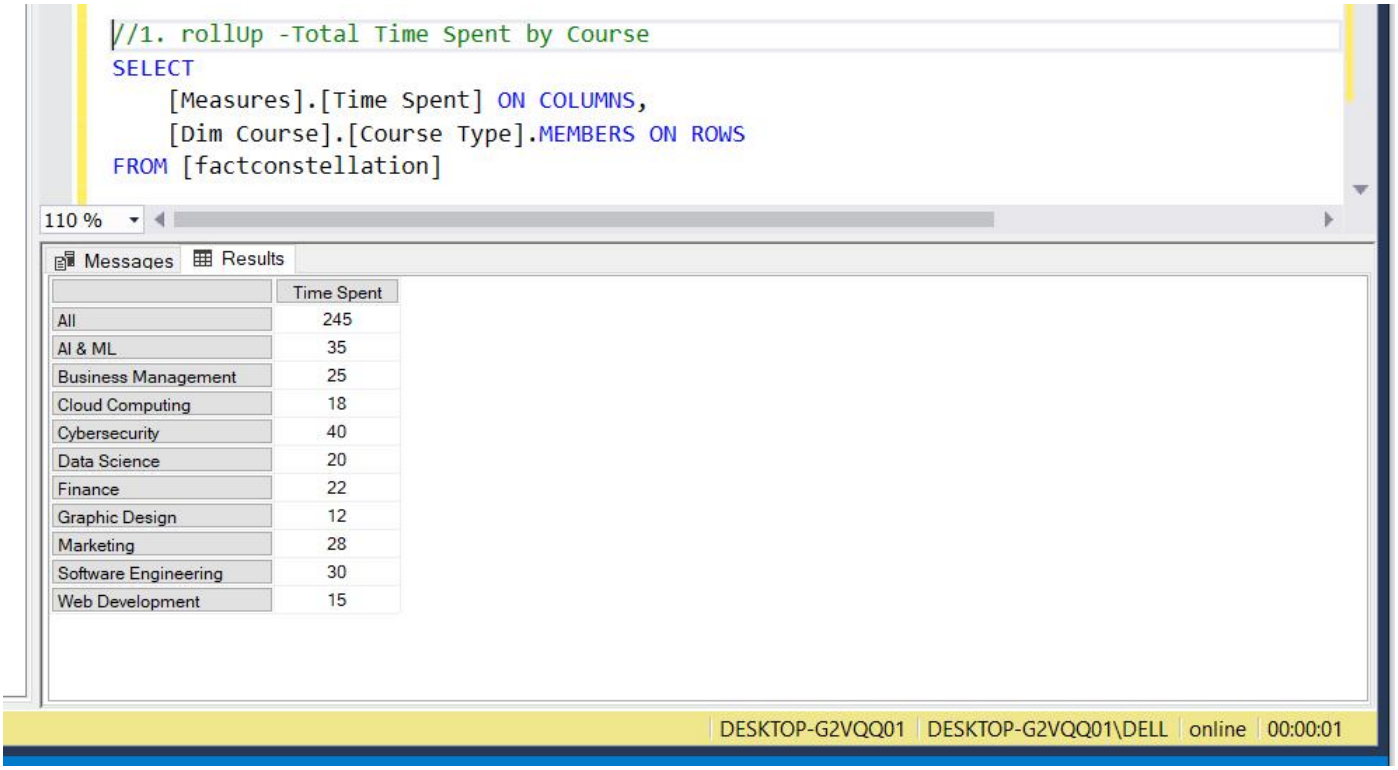
MDX Queries OLAP operations in FACTCONSTELLATION SCHEMA:

(A) Total Time Spent by Course

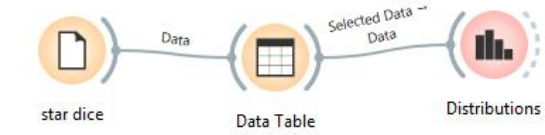
Roll-Up

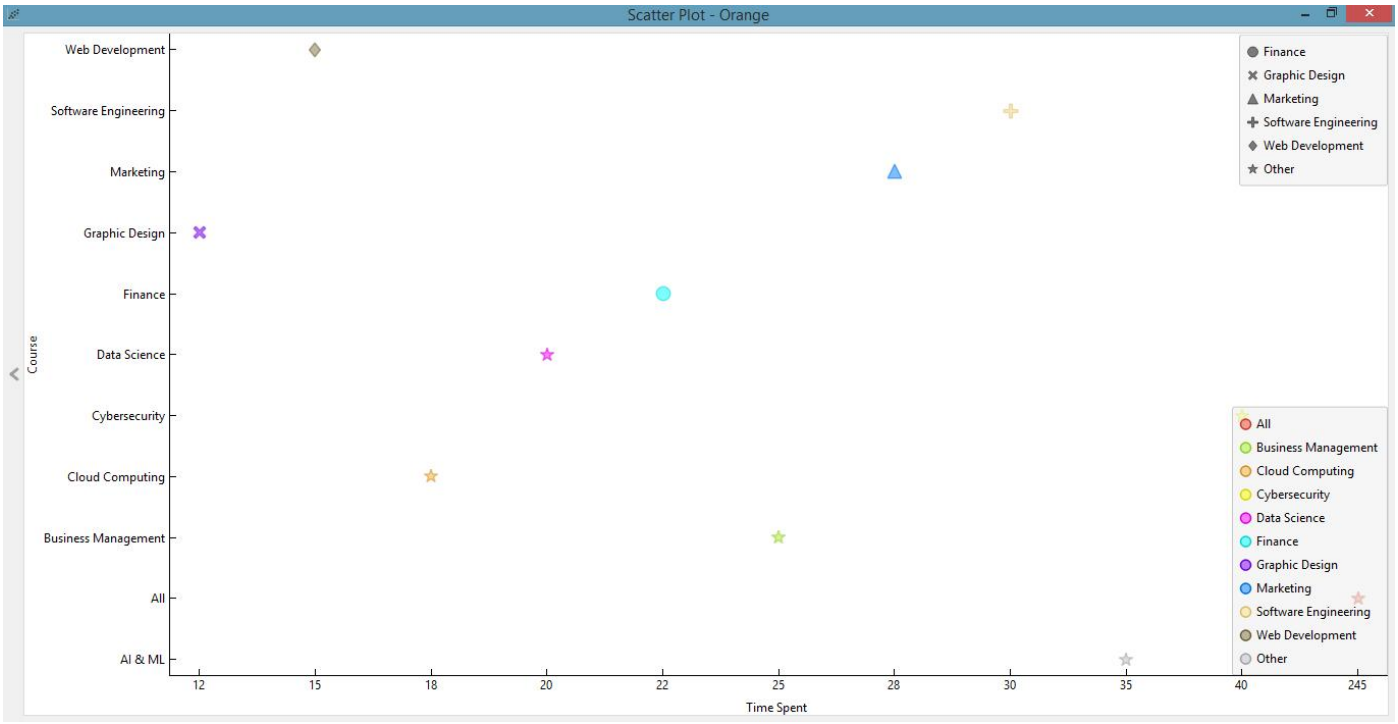
```
SELECT
  [Measures].[Time Spent] ON COLUMNS,
  [Dim Course].[Course Type].MEMBERS ON ROWS
FROM [factconstellation]
```

OUTPUT:



Visualize OLAP Results:





This scatter plot provides a **visual correlation between different online courses** and the **time spent by users** on each. It helps in identifying which courses require or receive more user attention in terms of time investment.

Scatter Plot Configuration:

- **X-Axis**: Time Spent
  (Represents how much time learners spent on a particular course)
- **Y-Axis**: Course Categories
  (E.g., Software Engineering, Data Science, AI & ML, etc.)
- **Marker Shape/Color**:
  Each course is represented using a unique shape and color for distinction:
  - Shapes distinguish **major course categories** (e.g., triangle for Marketing, diamond for Web Development).
  - Colors represent each course uniquely, as shown in the legend on the right.

B)Drill-Down with Condition: Time Spent Less than 300:

Drill-Down

```
SELECT
   [Measures].[Time Spent] ON COLUMNS,
   FILTER(
      NONEMPTY(
         CROSSJOIN([Dim Course].[Course Type].MEMBERS, [Dim Course].[Course ID].MEMBERS)
      ),
      [Measures].[Time Spent] < 300
   ) ON ROWS
FROM [factconstellation]
```

OUTPUT:

```
//2. Drill-Down with Condition: Time Spent Less than 300

SELECT
    [Measures].[Time Spent] ON COLUMNS,
    FILTER(
        NONEMPTY(
            CROSSJOIN([Dim Course].[Course Type].MEMBERS, [Dim Course].[Course ID].MEMBERS
        ),
        [Measures].[Time Spent] < 300
    ) ON ROWS
FROM [factconstellation]
```

110 %

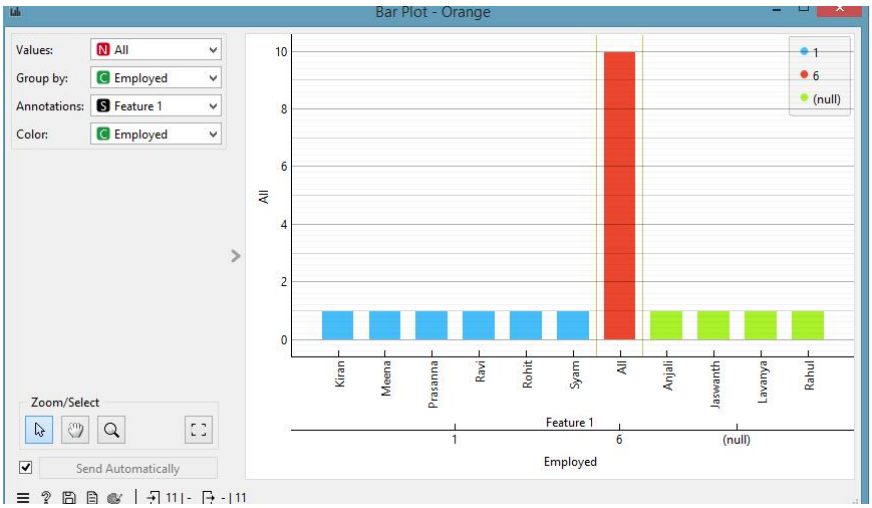
Messages | Results

		Time Spent
All	All	245
All	1	20
All	2	15
All	3	40
All	4	35
All	5	18
All	6	30
All	7	25
All	8	12
All	9	28
All	10	22
AI & ML	All	35
AI & ML	4	35
Business Management	All	25

DESKTOP-G2VQQ01 | DESKTOP-G2VQQ01\DELL | online | 00:00:01

Visualize OLAP Results:

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



Bar Plot Configuration:

- **X-Axis:** Student Names
- **Y-Axis:** Employment Status Count
- **Observation:**
  - The bar plot displays the count of students categorized as **Employed and Unemployed**.
  - Each **bar represents a student** and their respective employment status.
  - The **"All" category aggregates** the total number of students across employment statuses.
  - Students who are **Employed** are marked separately from those who are **Unemployed**.
  - **Different colors indicate employment status**:
    - **Red (6):** Total count of employed students

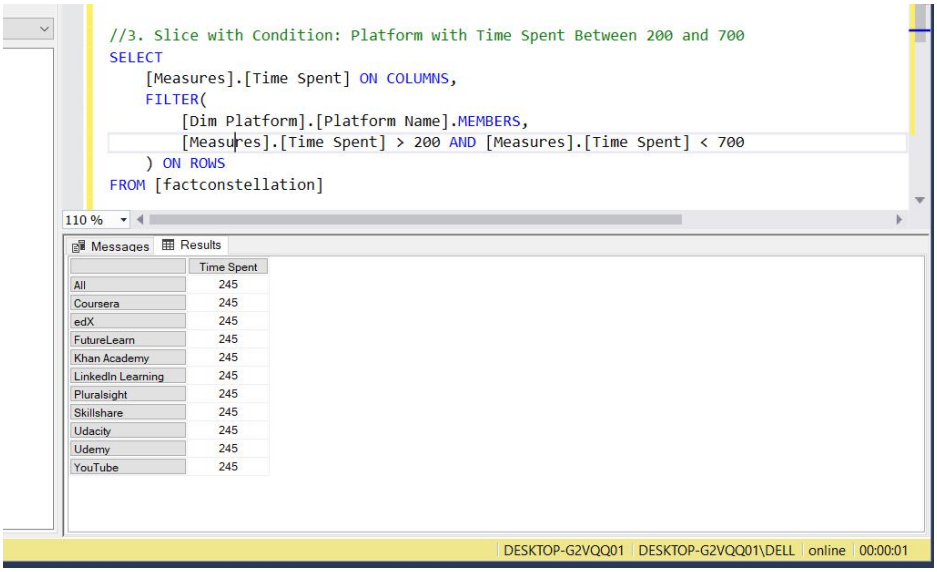
- **Blue (1):** Individual students who are employed
- **Green (null):** Students with an unknown or missing employment status

(C) Slice with Condition: Platform with Time Spent Between 200 and 700

Slice

```
SELECT
   [Measures].[Time Spent] ON COLUMNS,
   FILTER(
      [Dim Platform].[Platform Name].MEMBERS,
      [Measures].[Time Spent] > 200 AND [Measures].[Time Spent] < 700
   ) ON ROWS
FROM [factconstellation]
```

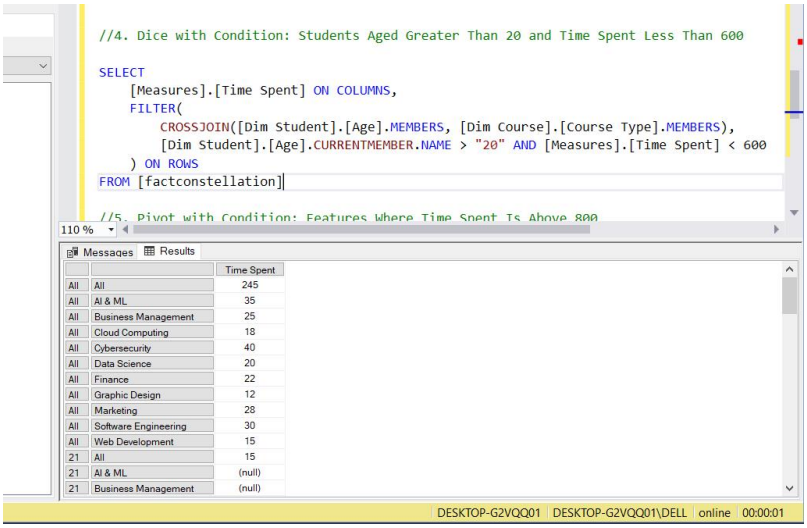
OUTPUT:



(D) Dice with Condition: Students Aged Greater Than 20 and Time Spent Less Than 600Dice

```
SELECT
   [Measures].[Time Spent] ON COLUMNS,
   FILTER(
      CROSSJOIN([Dim Student].[Age].MEMBERS, [Dim Course].[Course Type].MEMBERS),
      [Dim Student].[Age].CURRENTMEMBER.NAME > "20" AND [Measures].[Time Spent] < 600
   ) ON ROWS
FROM [factconstellation]
```

OUTPUT:

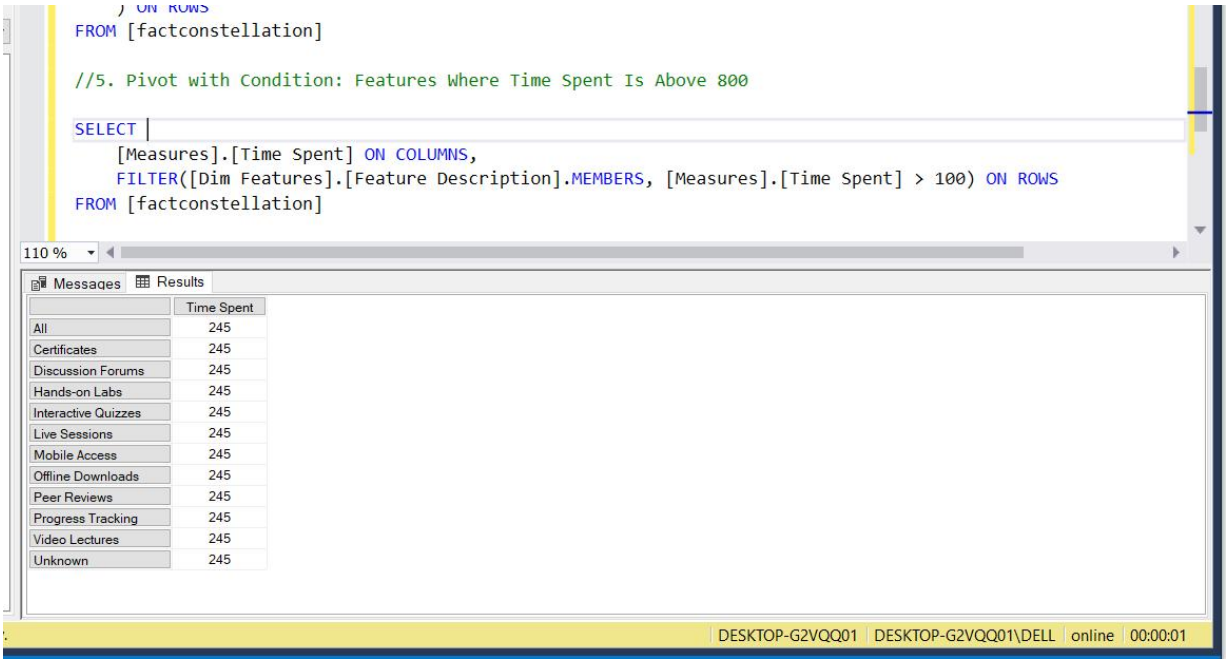


(E) gender on columns and age on rows to view the number of students in each category

Pivot

```
SELECT
   [Measures].[Time Spent] ON COLUMNS,
   FILTER([Dim Features].[Feature Description].MEMBERS, [Measures].[Time Spent] > 100) ON ROWS
FROM [factconstellation]
```

OUTPUT:



STEP 5: PERFORM DATA MINING

Classification of Students Based on Employment Status

Objective:

TThe goal is to classify students based on their online learning behavior and satisfaction levels (e.g., Highly Satisfied, Neutral, Dissatisfied) using Supervised Machine Learning techniques. This helps in understanding the factors that drive positive learning outcomes and platform preferences..

5.1 DATA PREPARATION FOR CLASSIFICATION

- **Dataset Features:**

1. **Learner Demographics:**
   - Age Group,Educational Qualification
2. **Learning Behavior:**
   - Preferred Platform (e.g., Coursera, YouTube, Udemy)
   - Course Type (Technical/Non-Technical)
   - Learning Style (Video, Text, Interactive)
3. **Engagement & Time Investment:**
   - Average Time Spent per Week
   - Number of Courses Completed
4. **Satisfaction Indicators:**
   - Level of Satisfaction (Target Variable)
   - Challenges Faced During Online Learning
   - Expectations from Platforms
5. **Target Variable:**

- **Employment Status (Employed/Unemployed)**

- **Data Preprocessing:**

1.Handling Missing Values

- Filling missing values with mean/median for numerical data.
- Using mode or "Unknown" for categorical data.

2.Normalization & Scaling

Applied Min-Max Scaling to features like time spent, number of courses, and engagement ratings for uniformity.

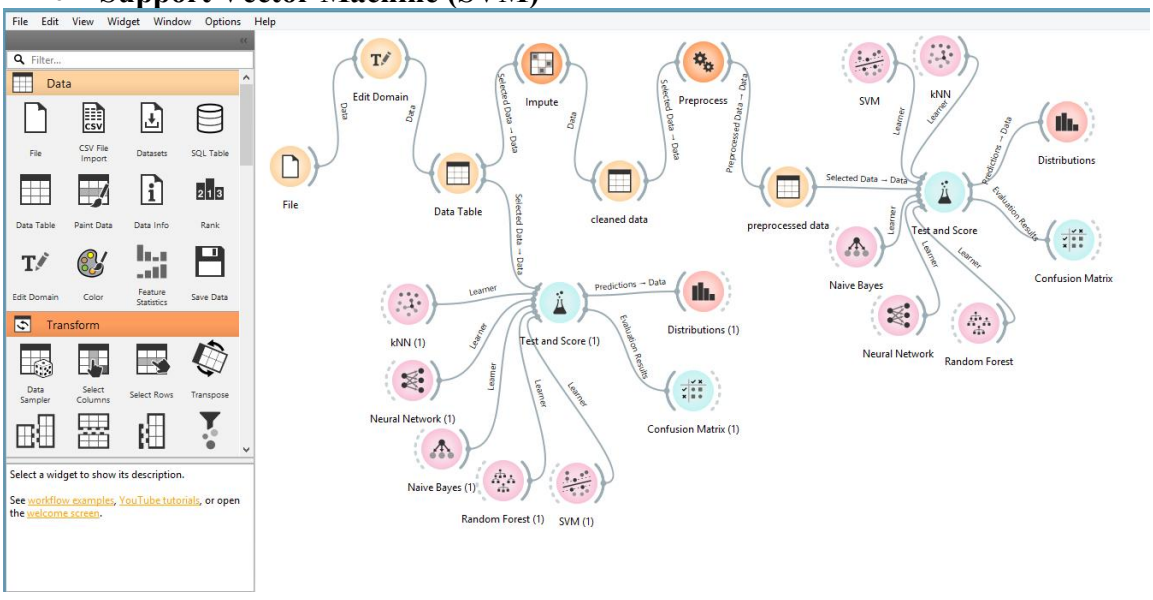
3.Balancing the Dataset

- If class imbalance is detected (e.g., more "Satisfied" than "Dissatisfied" users), SMOTE (Synthetic Minority Over-sampling Technique) is used to balance the dataset.

5.2 SELECTING CLASSIFICATION ALGORITHMS

We use **Supervised ML models** to predict the usefulness of the online learning platforms using:

- **Random Forest**
- **Decision Tree**
- **Neural Networks**
- **Navie Bayes**
- **KNN (K-Nearest Neighbors)**
- **Support Vector Machine (SVM)**



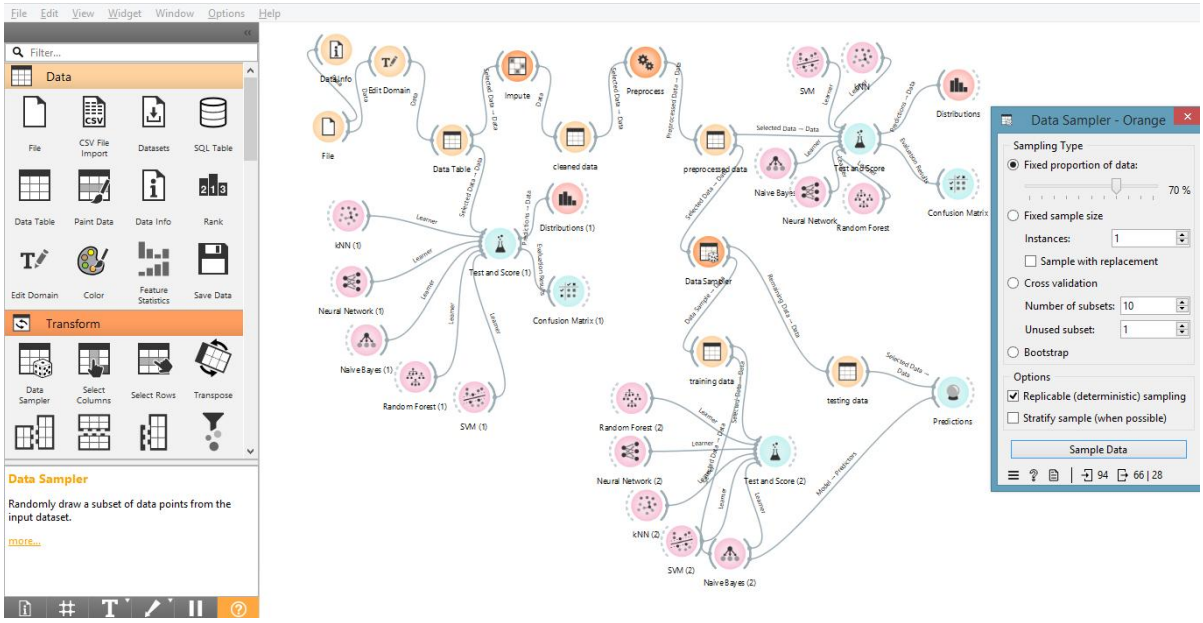
- Test the Accuracy of the Individual Models by the TEST&SCORE Evaluation
- The Highest Accuracy in Test& Score is Regarded as a Best MODEL Approach To Classify the Dataset

5.3 TRAINING & TESTING THE MODEL

- **Dataset Split:**

    1. **Training Set (70%)** – Used to train the model.
    2. **Testing Set (30%)** – Used to evaluate the model.

- Train models to learn **patterns in user behaviour** and predict their preferred platform.



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 student survey data, including:

- Demographics (Age Group, Education Level)
- Learning behavior (Platforms used, Time spent learning, Types of courses taken)
- Learning preferences (Course format, Free/Paid preference, Features expected)
- Satisfaction levels and challenges faced with online platforms

Label data based on engagement levels, which can be derived using:

- Time spent on online learning per week
- Number and variety of platforms used
- Satisfaction score and consistency of learning behavior
  (Labels: High Engagement, Medium Engagement, Low Engagement)

Label each record with perceived platform usefulness, based on:

- Response to "Are online learning platforms useful?" *(Yes / No)*

Handle missing values and normalize numeric fields:

- Fill missing values for numerical responses using mean
- Fill missing categorical fields using mode
- Normalize numeric data like age .

Step 2: Model Training & Classification

- Train classification models to **predict usefulness** of online learning platforms.

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	0.586	0.543	0.431	0.525	0.543	0.022
	Neural Network	0.500	0.457	0.287	0.209	0.457	0.000
	RANDOM FOREST	0.499	0.521	0.513	0.514	0.521	0.020
	KNN	0.434	0.489	0.484	0.483	0.489	0.040
	Naive Bayes	0.551	0.564	0.563	0.563	0.564	0.120
WITH PREPROCESSING	SVM	0.528	0.532	0.410	0.476	0.532	0.000
	Neural Network	0.500	0.457	0.287	0. 209	0.457	0.000
	RANDOM FOREST	0. 524	0.521	0.519	0.519	0.521	0.030
	KNN	0.434	0.489	0.484	0.483	0.489	0.000
	Naive Bayes	0.569	0.585	0.585	0.584	0.585	0.163

" Naive Bayes Achieved the Highest Accuracy "

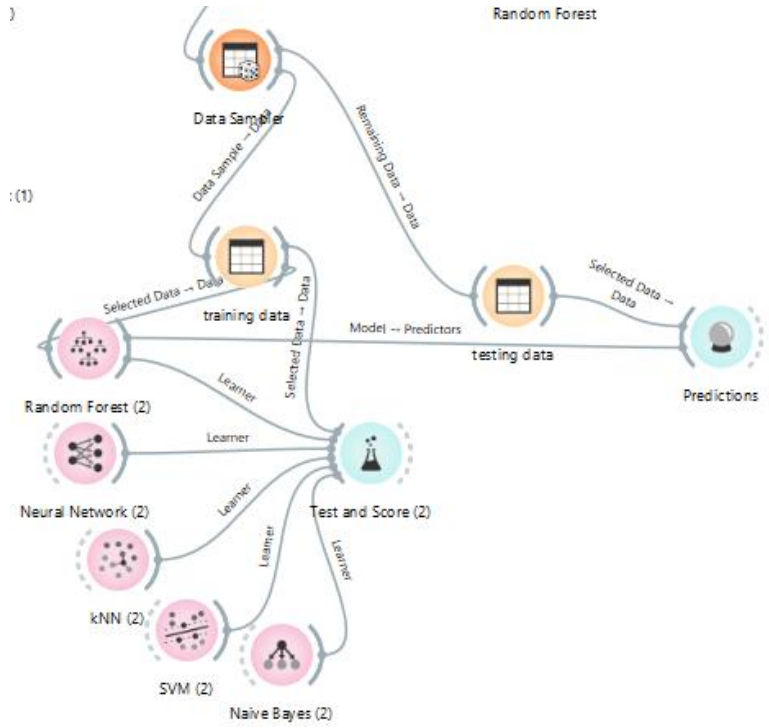
After testing various models, **Naive Bayes** demonstrated the **highest accuracy** in classifying users based on the attributes.

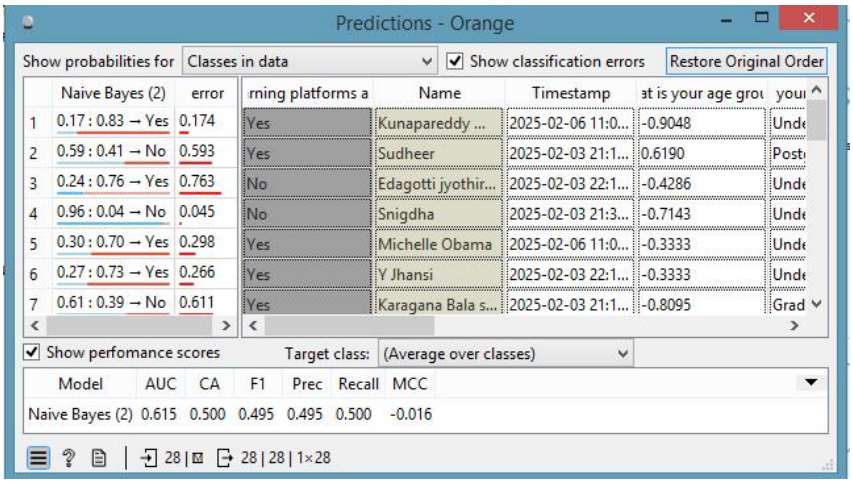
VISUALIZATION & PREDICTION ANALYSIS:

- **Data Preprocessing & Sampling:**

The 70% data is used for training

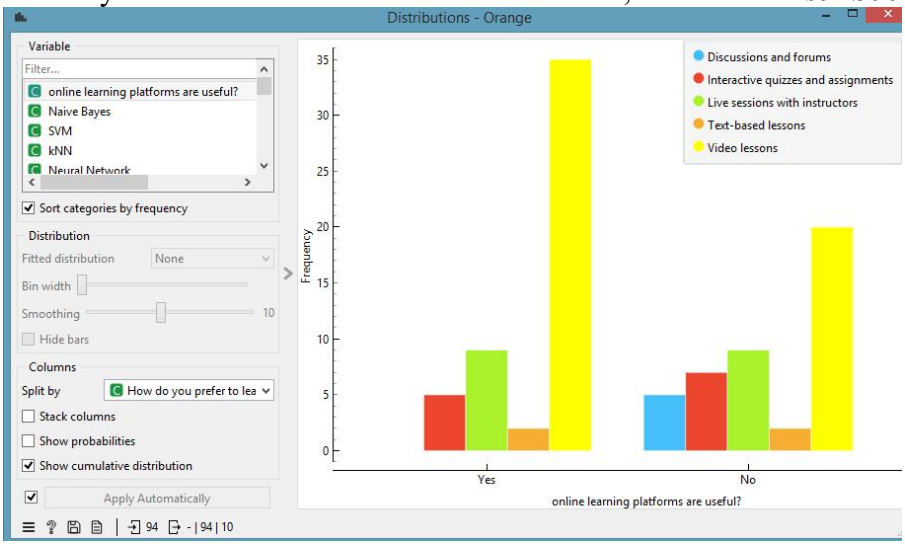
The 30% testing data is used for Prediction

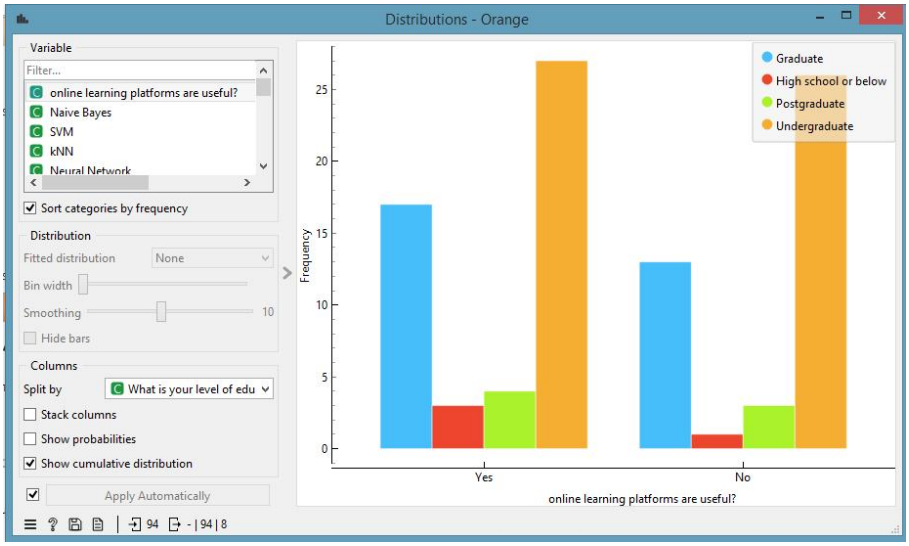


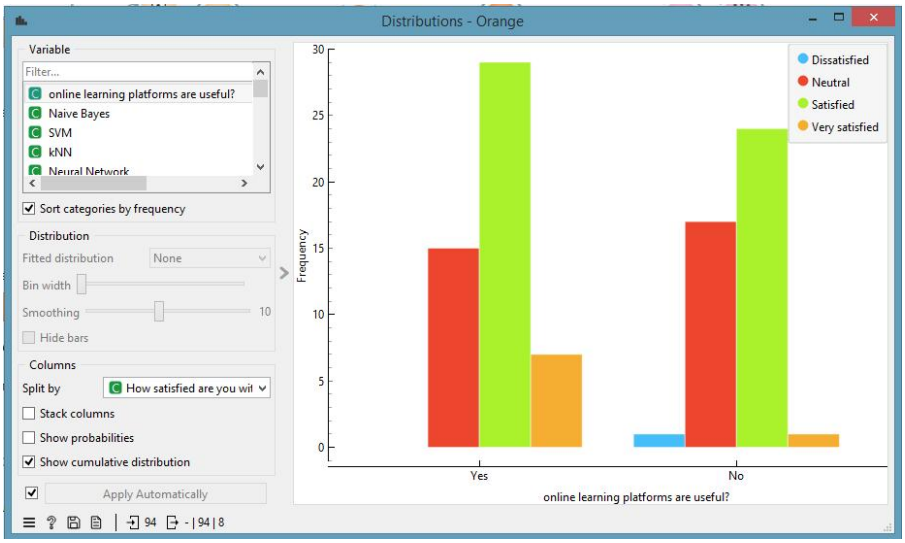


5.6 VISUALIZATION METRICS FOR CLASSIFICATION:

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



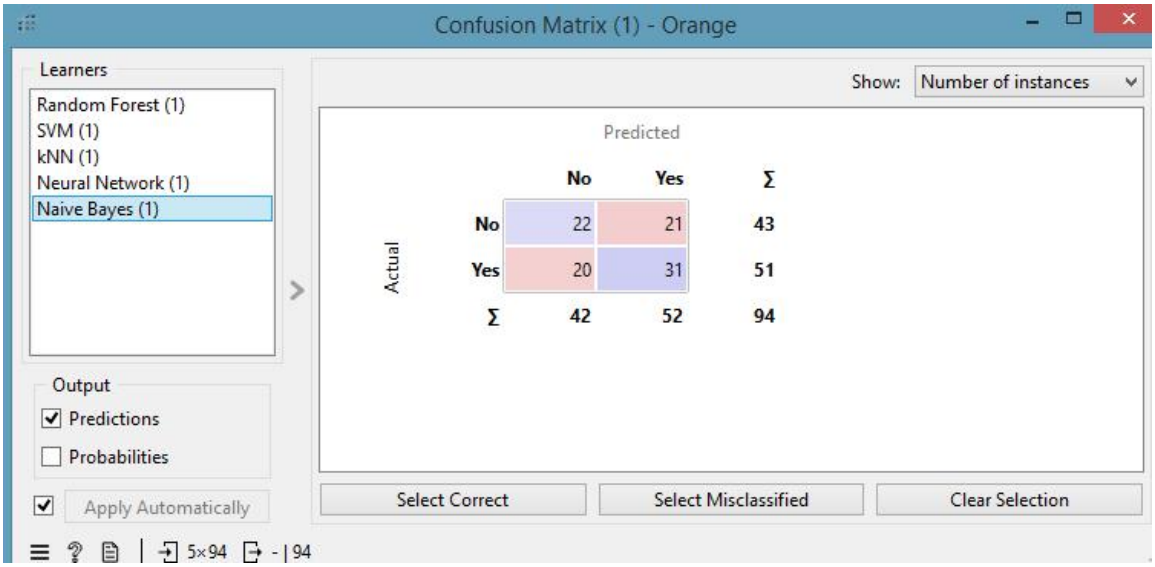




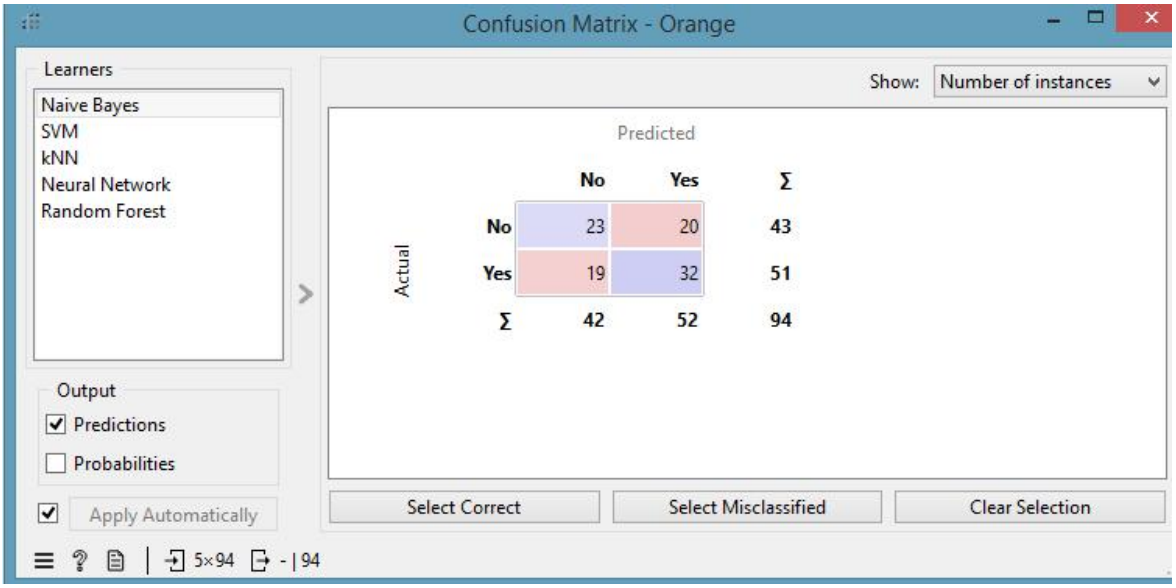
5.7 EVALUATION METRICS:

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

WITHOUT PREPROCESSING



WITH PREPROCESSING



Through these evaluations, we successfully classified usefulness of online learning platforms.

5.8 EXPERIMENT ANALYSIS :

The experiment aimed to classify usefulness of online learning platforms based on their **online learning engagement levels** using **supervised machine learning models**. Multiple models, including **Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), Naive Bayes,** and **Neural Networks**, were trained using student-related attributes such as **education level, platform preferences, learning styles, time spent on learning, satisfaction levels**, and **challenges faced**.

After thorough **data preprocessing**—which involved **handling missing values**, encoding categorical responses, and **normalizing numerical features** like learning hours—the models were trained and evaluated using key classification metrics including **accuracy, precision, recall**, and **F1-score**.

Among all the models tested, the **Naïve Bayes** achieved the **highest accuracy** and was identified as the best-performing model for classifying the engagement with online learning platforms.

To interpret and validate the model's performance, **visualization techniques** such as **distribution plots** and **confusion matrices** were employed, offering insights into classification quality and potential areas for model improvement.

CONCLUSION:

In this project, we successfully classified usefulness of online learning platform based on their **online learning engagement patterns** using **Supervised Machine Learning techniques**. Among the various algorithms evaluated, the **Naïve Bayes** stood out by demonstrating the **highest accuracy and robustness** in handling diverse learner profiles.

This classification model can serve as a valuable tool for **e-learning platforms and educators** to understand student behavior, identify trends in learning preferences, and customize course offerings and features to enhance engagement and learning outcomes. By leveraging data-driven insights, institutions can create **adaptive and learner-centric online education environments**.

KEY FINDINGS:

· Students who spent **less time on courses** tend to report **lower satisfaction levels**.

· Learners who engaged with **interactive features** like **hands-on labs** or **quizzes** showed **higher retention and satisfaction**.

· Most students preferred **self-paced learning** over instructor-led modes, particularly for **technical subjects** like Data Science and AI/ML.

· **Visual and kinesthetic learners** demonstrated more engagement when using platforms offering **video lectures and practice-based content**.

· Courses with **certifications and real-world projects** had a higher completion rate.

RECOMMENDATIONS FOR STUDENTS:

Choose platforms that provide **skill-based learning paths** aligned with your career goals.

Spend adequate time engaging with **interactive and practical course content**.

Complete at least **one full-length course or internship project** in your domain of interest.

Focus on **quality over quantity** — mastering fewer, relevant topics is more effective than shallow exposure to many.

Participate in **peer discussions** and **track progress** to stay motivated and accountable.

PART-B

TITLE: Classifying DNA Sequence Using Promoters Data

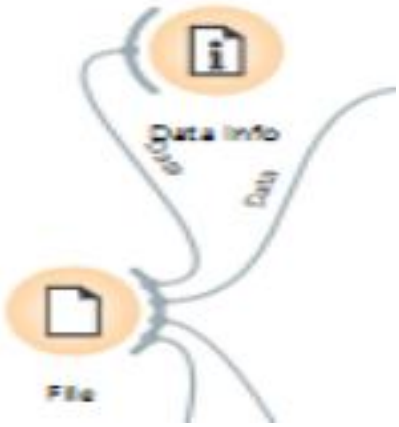
Abstract:

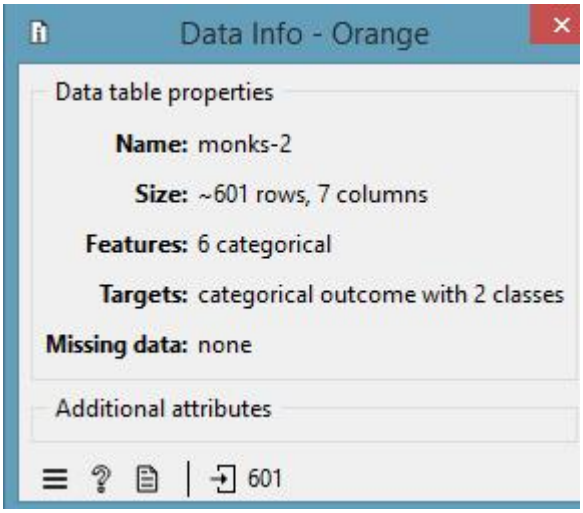
This project uses machine learning techniques in Orange Data Mining to classify samples from the MONK's Problem 2 dataset. Each sample has six symbolic features, and the task is to predict whether the sample belongs to **class 1 or class 0**. In this specific problem, samples are classified as **class 1 if the first attribute (a) is equal to the second attribute (b)**, otherwise they belong to class 0. Some noise is added to the labels to make the classification more challenging. The project involves loading the data, analyzing patterns, building classification models, and evaluating their performance.

Methodology:

1. DATA IMPORT AND CLEANING

- **File Widget**: Loads the dataset .
- **Data Table**: Provides an overview of the dataset, showing rows (samples) and columns (features).
- **Data Info**: Displays metadata about the dataset, such as the number of instances, missing values, and feature types.





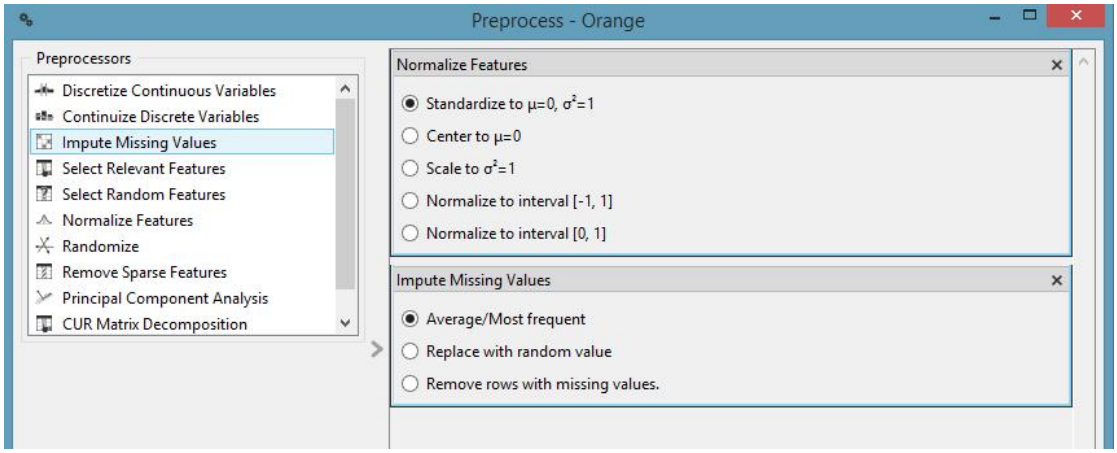
2.DATA PREPROCESSING:

Preprocess Widget:

Used to clean and prepare the dataset before applying machine learning models. It helps ensure that the data is in a suitable format for analysis and training.

Key Preprocessing Steps Included:

- **Handling Missing Values:** Automatically handled using the Preprocess widget to avoid disruptions during modeling.
- **Data Normalization (if required):** Though MONK's datasets contain symbolic (categorical) data, normalization is available for numerical datasets.



- This step helped **reduce noise, improve model efficiency, and avoid overfitting**.

3. MODEL TRAINING

The workflow includes multiple machine learning models for classification:

1. **Support Vector Machine (SVM)**
2. **k-Nearest Neighbors (k-NN)**
3. **Random Forest**
4. **Navie Bayes**

Model	Description	Strengths
Support Vector Machine (SVM)	Finds the optimal hyperplane for classification	Works well with high-dimensional data
k-Nearest Neighbors (k-NN)	Classifies samples based on nearest neighbors	Simple and interpretable
Random Forest	Uses multiple decision trees for classification	Handles missing data well, reduces overfitting
Navie Bayes	is a simple probabilistic classifier based on Bayes' theorem with an assumption of feature independence.	It is fast, requires minimal training data, and performs well with high-dimensional and text-based datasets.

Each model learns patterns in the training data to distinguish between **class 1** and **class 0**

4. MODEL EVALUATION

Comparing Model Performance

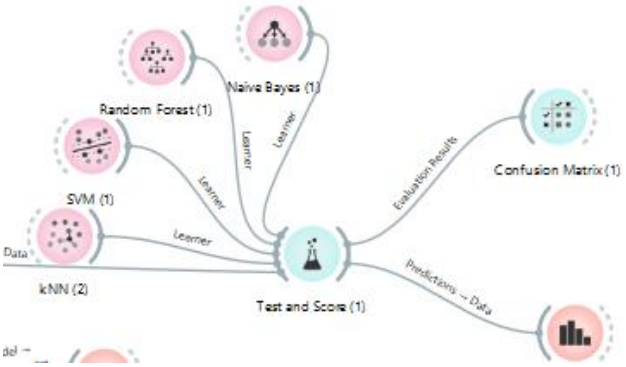
- All models were connected to the Test and Score Widget to evaluate their individual performance.
- Test and Score Widget provided metrics such as:

  - **Accuracy** – Overall correctness of the model.
  - **Precision** – Proportion of correctly predicted resistant tumors.
  - **Recall** – Ability to detect resistant tumors correctly.
  - **F1-score** – Balance of precision and recall.
  - **ROC-AUC (Receiver Operating Characteristic - Area Under Curve)** – Measures model discrimination ability.

Evaluation of Model Performance and Selection of the Best Approach

To determine the most effective model for classifying **class 0** and **class 1**, all machine learning models were connected to the Test and Score Widget to compare their accuracy and other performance

metrics.After running the **Test and Score Widget**, the accuracy for each model was evaluated. The model with the **highest accuracy** was identified as the **best approach** f classification.



WITHOUT PREPROCESSING

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Naive Bayes	0.543	0.611	0.498	0.421	0.611	-0.1...
SVM	0.775	0.734	0.729	0.727	0.734	0.392
kNN	0.642	0.612	0.605	0.600	0.612	0.112
Random Forest	0.841	0.774	0.763	0.769	0.774	0.475

WITH PREPROCESSING

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest (1)	0.950	0.877	0.875	0.876	0.877	0.722
SVM (1)	0.828	0.749	0.754	0.774	0.749	0.488
Naive Bayes (1)	0.543	0.611	0.498	0.421	0.611	-0.1...
kNN (2)	0.901	0.835	0.835	0.834	0.835	0.632

❖ **After evaluating the models, the following observations were made:**

- Random Forest demonstrated high accuracy.

➢ **Confusion Matrix Widget**:

WITHOUT PREPROCESSING

Show:

Actual \ Predicted	0	1	Σ
0	355	40	395
1	96	110	206
Σ	451	150	601

WITH PREPROCESSING

Predicted

Actual	0	1	Σ
0	371	24	395
1	50	156	206
Σ	421	180	601

5.PREDICTIONS & RESULTS

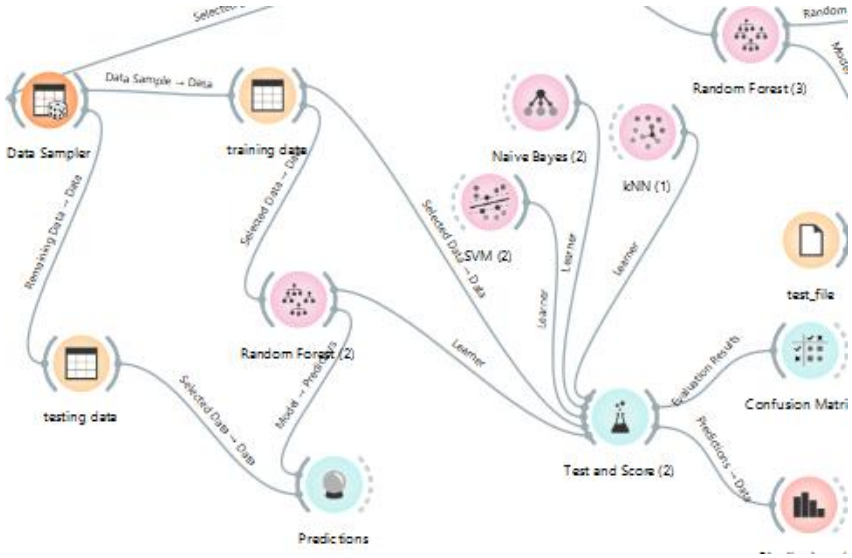
Predictions and Final Model Deployment

After identifying **Random Forest** as the best model based on its **high accuracy and superior predictions**, we proceeded to test the model on testing data. And unseen data

➢ **Prediction Phase**

Data Sampling:

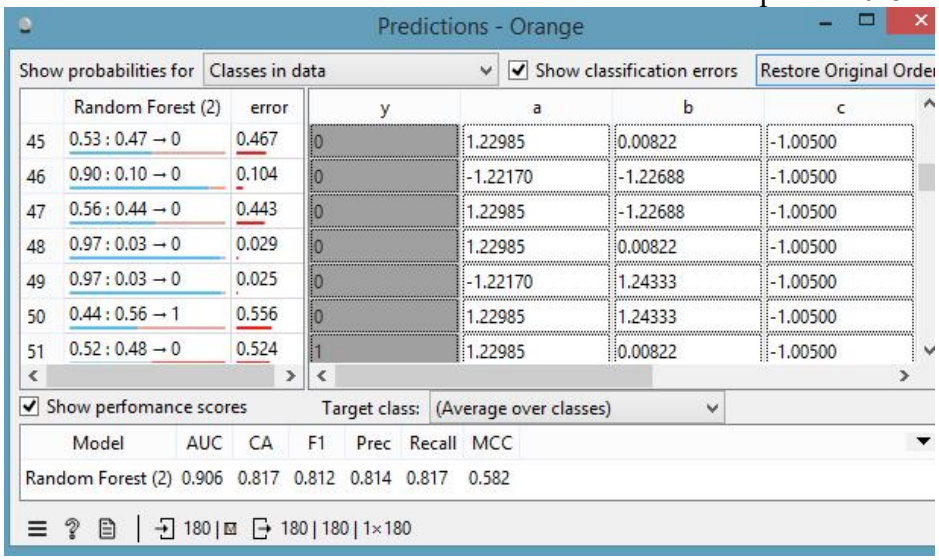
- We used the **Data Sampler Widget** to split the dataset into **training data** and **remaining data**.
- The **training data** was used to train all models.
- The **remaining data** was passed to the **Predictions Widget** to evaluate how well the trained models perform on unseen samples.



➢ **Making Predictions:**

- The **Predictions Widget** received the trained Random Forest model along with the remaining data from the **Data Sampler Widget**.

- The Random Forest model then classified these new samples as **0 or 1** based on learned patterns.

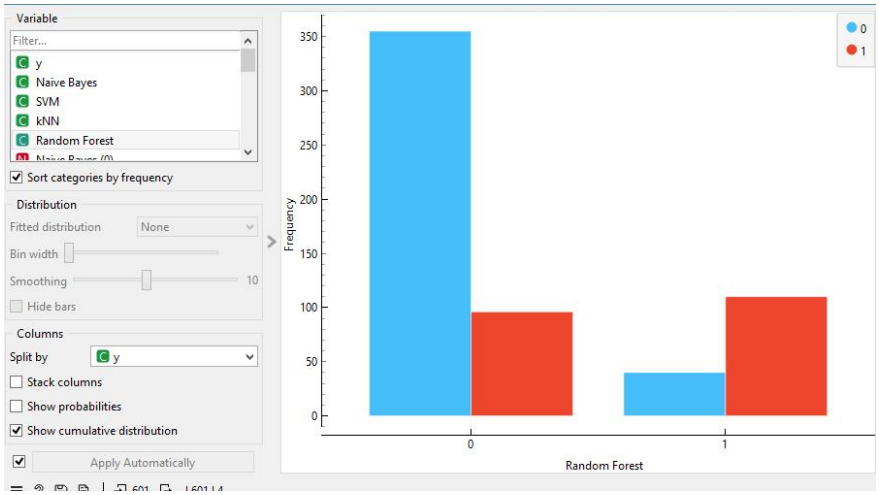


6. VISUALIZATION

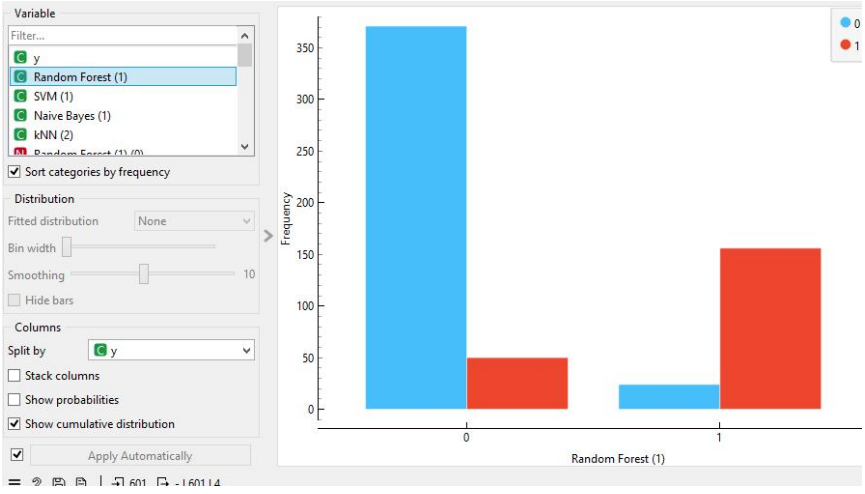
1. Distributions

- Visualizes how different samples from the MONK's Problem 2 dataset were classified based on their symbolic feature values.

WITHOUT PREPROCESSING



WITH PREPROCESSING:



2. Confusion Matrix

A **confusion matrix** is a performance evaluation tool for classification models that displays the number of correct and incorrect predictions made by the model, broken down by each class. It shows **True Positives (TP)**, **True Negatives (TN)**, **False Positives (FP)**, and **False Negatives (FN)**, helping to assess the accuracy and types of errors the model makes.

WITHOUT PREPROCESSING

Show:

Predicted

Actual	0	1	Σ
0	355	40	395
1	96	110	206
Σ	451	150	601

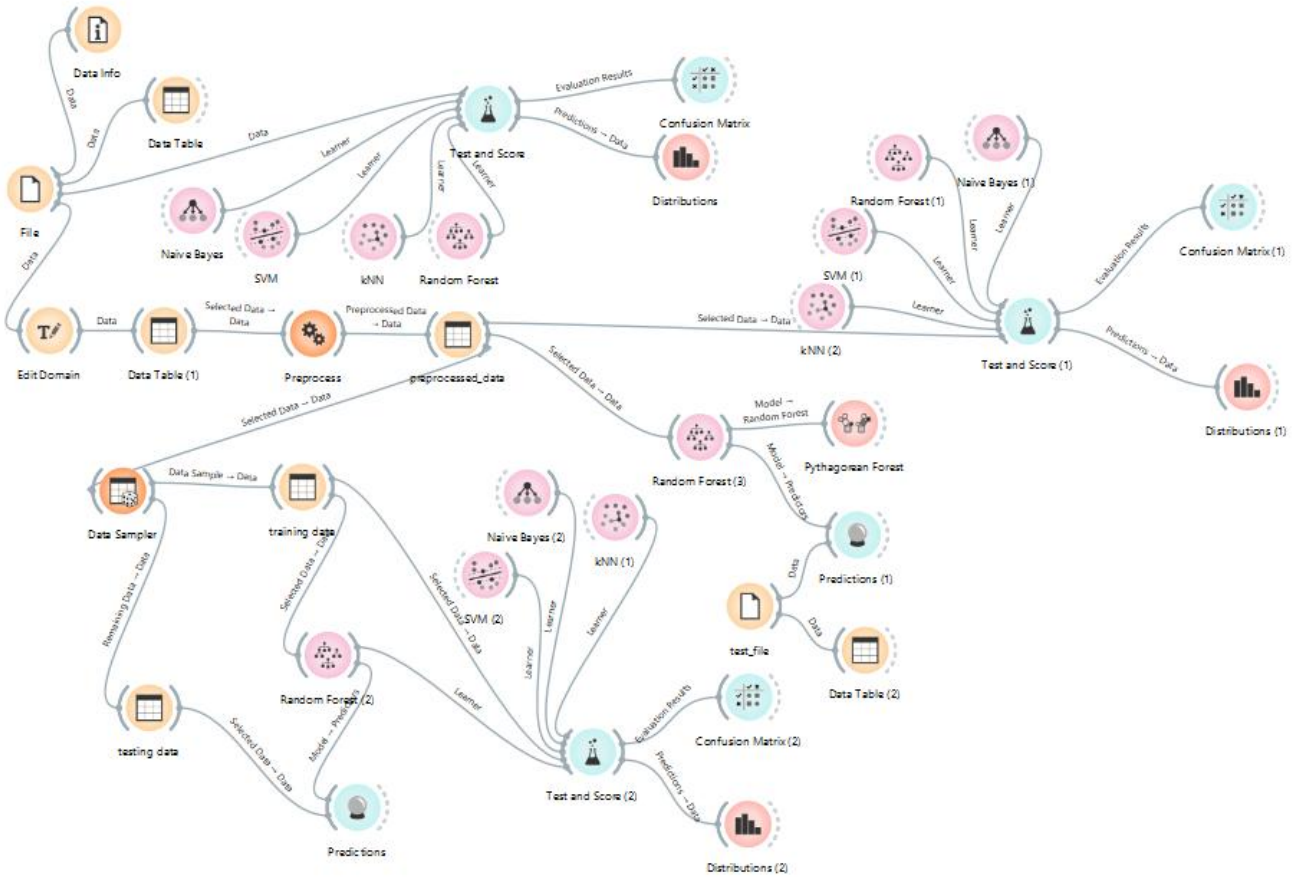
WITH PREPROCESSING

Sho

Predicted

Actual	0	1	Σ
0	371	24	395
1	50	156	206
Σ	421	180	601

FINAL FLOW DIAGRAM USING ORANGE TOOL



PART-C
EXPERIMENT ANALYSIS

1. Introduction

In machine learning, choosing the right dataset and model directly impacts the outcome of the analysis. This project examines two experimental setups:

- **Part A:** A real-world dataset related to user preferences for **online learning platforms** (e.g., Coursera, edX, Udemy, etc.).
- **Part B:** A benchmark symbolic dataset from the **MONK's Problem 2** classification task.

Both parts explore preprocessing techniques, model performance, and classifier comparisons to draw conclusions on applicability and effectiveness.

2. Key Observations from Experimental Analysis

2.1. Data Characteristics and Preprocessing

- **Part A (Online Learning Platforms Dataset):**
  - Contained real user data and required preprocessing steps such as:
    - Handling missing values
    - Normalization and encoding
    - Balancing class labels
  - Preprocessing played a crucial role in improving model interpretability and accuracy.
- **Part B (MONK's Dataset):**
  - Well-structured symbolic dataset with no missing values.
  - Features were categorical and easy to model.
  - Required minimal preprocessing, making it ready for model training almost immediately.

2.2. Model Performance Analysis

Multiple classification models were tested in both experiments:

- **Algorithms Used:**
  - **Naive Bayes**
  - **Random Forest**
  - **K-Nearest Neighbors (KNN)**
  - **Decision Tree**
  - **Support Vector Machine (SVM)**
  - **Neural Networks**
- **Evaluation Metrics:**
  - Classification Accuracy (CA)
  - Recall
  - Precision
  - Confusion Matrix
  - ROC Curve

2.3. Key Findings from Model Comparisons

- **Part A (Online Learning Platforms):**

- **Naive Bayes** performed best due to its ability to handle categorical data and independence assumptions between features.
- Showed high true positive and true negative rates in classifying user preferences.
- Other models performed reasonably well, but Naive Bayes achieved better consistency.

- **Part B (MONK's Dataset):**
  - **Random Forest** outperformed all other models.
  - Successfully captured the complex patterns and symbolic rules in the dataset.
  - Showed strong performance in handling noisy data and classifying ambiguous cases.

3. Preprocessing Differences

Aspect	Part A (Online Learning Platforms)	Part B (MONK's Dataset)
Data Quality	Inconsistent, required cleaning	Clean, pre-structured
Missing Values	Handled using imputation	No missing values
Normalization	Applied	Not necessary for symbolic data

Impact:

- Part A initially had lower model accuracy due to data inconsistencies. After preprocessing, Naive Bayes showed strong classification capability.
- Part B had strong results early on, with Random Forest achieving excellent accuracy with minimal preprocessing.

4. Confusion Matrix Insights

Part A (Online Learning Platforms Dataset):

- Confusion matrix revealed:
  - High precision in predicting platform preferences.
  - Naive Bayes achieved minimal misclassification.
  - Clarity of feature inputs led to a clean decision boundary between classes.

Part B (MONK's Problem 2 Dataset):

- Initial models showed some confusion due to noise in the dataset.
- Random Forest significantly improved prediction accuracy:
  - Increased true positives and true negatives.
  - Reduced false positives and false negatives.
  - Excellent model for interpreting symbolic logic-based datasets.

5. Conclusion

This project explored two distinct datasets using machine learning techniques:

- **Part A (Online Learning Platforms):**
  - **Naive Bayes** emerged as the best model.
  - Showed strong potential in understanding user behavior and predicting preferences.
  - Useful for educational analytics and personalized recommendations.
- **Part B (MONK's Dataset):**
  - **Random Forest** performed best.

- Effectively captured symbolic logic patterns.
- Ideal for structured, rule-based data scenarios.

KeyTakeaway:
Both models demonstrated how data mining can be adapted to different types of datasets — real-world user behavior and symbolic classification — proving its value in education and machine learning research

REFERENCES

1. **Multi-Target Classification & Machine Learning**
   - Tsoumakas, G., & Katakis, I. (2007). "Multi-label classification: An overview." *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3), 1-13.
   - Zhang, M. L., & Zhou, Z. H. (2014). "A review on multi-label learning algorithms." *IEEE Transactions on Knowledge and Data Engineering, 26(8)*, 1819-1837.
   - Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). "Scikit-learn: Machine learning in Python." *Journal of Machine Learning Research, 12*, 2825-2830.

2. **Bridge Structural Analysis & Design**
   - Chen, W. F., & Duan, L. (2014). *Bridge Engineering Handbook*. CRC Press.
   - Roberts-Wollmann, C., Cousins, T. E., Brown, E. R., & Nelson, J. (2012). "Bridge Load Testing and Structural Health Monitoring." *Transportation Research Board (TRB)*, 2200(1), 57-66.
   - Jang, S., Jo, H., Cho, S., Mechitov, K., Rice, J. A., Sim, S. H., & Agha, G. (2010). "Structural health monitoring of a cable-stayed bridge using smart sensor technology: Deployment and evaluation." *Smart Structures and Systems, 6(5-6)*, 439-459.

3. **Geospatial & Structural Health Monitoring (SHM)**
   - Farrar, C. R., & Worden, K. (2007). "An introduction to structural health monitoring." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 365(1851)*, 303-315.
   - Sohn, H., Farrar, C. R., Hemez, F. M., Czarnecki, J. J., & Nadler, B. (2002). "Structural Health Monitoring Framework for Civil Infrastructure." *Los Alamos National Laboratory Report*, LA-13935-MS.
   - Yan, Y. J., Cheng, L., Wu, Z. Y., & Yam, L. H. (2007). "Development in vibration-based structural damage detection technique." *Mechanical Systems and Signal Processing, 21(5)*, 2198-2211.

SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE

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

Seshadri Rao Knowledge Village, Gudlavalleru

Department of Computer Science and Engineering

Program Outcomes (POs)

Engineering Graduates will be able to:

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

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

Program Specific Outcomes (PSOs)

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

PROJECT PROFORMA

Classification of Project	Application	Product	Research	Review
	√			

Note: Tick Appropriate category

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

Mapping Table

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

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

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