

Assessing Student Competencies: Analyzing Readiness for Employment 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
GORLI LAXMI(22481A0563)
CHAPPIDI JASWANTH (22481A0543)
AKUNURI HARSHAVALLI (22481A0507)
GORIPARTHI VENKATA KARTHIK (22481A0561)

Under the Enviabale and Esteemed Guidance of

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

Head of the Department, CSE



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE
(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)
SESHADRI RAO KNOWLEDGE VILLAGE GUDLAVALLERU – 521356
ANDHRA PRADESH

2024-25

SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)
SESHADRI RAO KNOWLEDGE VILLAGE, GUDLAVALLERU
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



CERTIFICATE

This is to Certify that the Project Report *Assessing Student Competencies: Analyzing Readiness for Employment* is a bonafide record of work carried out by **G.LAXMI (22481A0563), CH.JASWANTH(22481A0543), A.HARSHAVALLI(22481A0507), GORIPARTHI VENKATA KARTHIK (22481A0561)**, under the guidance and supervision of Dr.M.BABU RAO ,M.Tech, Ph.D, Head of the Department,CSE, Computer Science and Engineering, in the partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering of Jawaharlal Nehru Technological University Kakinada, Kakinada during the academic year 2024-25.

Project Guide
(Dr. M. BABU RAO)

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

External Examiner

ACKNOWLEDGEMENT

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

We would like to express our deep sense of gratitude and sincere thanks to **Dr.M.BABU RAO ,M.Tech, Ph.D, Head of the Department**, Computer Science and Engineering for 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

GORLI LAXMI(22481A0563)

CHAPPIDI JASWANTH (22481A0543)

AKUNURI HARSHAVALLI (22481A0507)

GORIPARTHI VENKATA KARTHIK (22481A0561)

INDEX

TITLE	PAGE NUMBERS
ABSTRACT	5 - 5
PART-A	6 - 6
CHAPTER 1: INTRODUCTION	6 - 12
1.3 Data Warehousing	8 - 8
CHAPTER 2: KDD PROCESS	13 - 16
Step-1: Collecting & Exploring Dataset	13 - 13
Step-2: Preprocess the Data	15 - 16
CHAPTER 3: OLAP & SCHEMAS	17 - 31
Generate SQL Queries for OLAP Schema Construction	17 - 18
Design & Visualize the Snowflake Schema	23 - 23
Deploy & Load Data into the Snowflake Schema	23 - 24
Create & Execute OLAP Queries	24 - 25
Perform OLAP Operations	26 - 26
CHAPTER 4: STUDENT CLASSIFICATION	31 - 36
Classification of Students Based on Employment Status	31 - 32
Training & Testing the Model	33 - 33
Model Evaluation Metrics	34 - 34
Methodology Overview	34 - 34
Visualization Metrics for Classification	35 - 35
CHAPTER 5: MODEL DEPLOYMENT AND EVALUATION	36 - 37
Data Import and Cleaning	36 - 36
Feature Engineering	36 - 36
PART-B: Classifying DNA Sequence Using Promoters Data	38 - 44
Abstract	38 - 38
Methodology	38 - 39
Model Training	39 - 40
Model Evaluation	40 - 41
Predictions & Results	41 - 42
Visualization	43 - 44
Final Flow Diagram	44 - 44
PART-C: Experiment Analysis	44 - 47
Key Observations from Experimental Analysis	45 - 46
Preprocessing Differences	46 - 46
Impact	46 - 47
Final Conclusion	47 - 48
REFERENCES	49 - 50
PROJECT PROFORMA & CO-PO MAPPING	50 - 51

ABSTRACT

Understanding student career preferences and employment readiness is essential for bridging the gap between academic learning and industry requirements. This project utilizes data-driven analysis to classify students based on their skills, internships, project experience, and job search patterns. The dataset, collected through surveys, was preprocessed to handle missing values, normalize categorical data, and extract relevant features for analysis.

Exploratory Data Analysis (EDA) was conducted to identify trends and correlations, guiding the selection of appropriate classification models. Various machine learning algorithms, including Decision Trees, Random Forest, and Support Vector Machines (SVM), Navie Bayes, K-Nearest Neighbour were evaluated to determine the most effective approach for categorizing students based on their employment readiness.

The models were assessed using performance metrics such as accuracy, precision, recall, and F1-score, with a confusion matrix providing further insights. The results highlight that RANDOM FOREST achieved the highest accuracy in classifying students into categories such as “Employed” and “Unemployed”.

This study demonstrates the potential of machine learning in career readiness assessment, enabling educational institutions to provide targeted guidance, enhance training programs, and improve student employability. By leveraging data-driven insights, institutions can better support students in transitioning from academics to the workforce.

PART-A

Student Career Pathways: Analyzing Employment Readiness 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

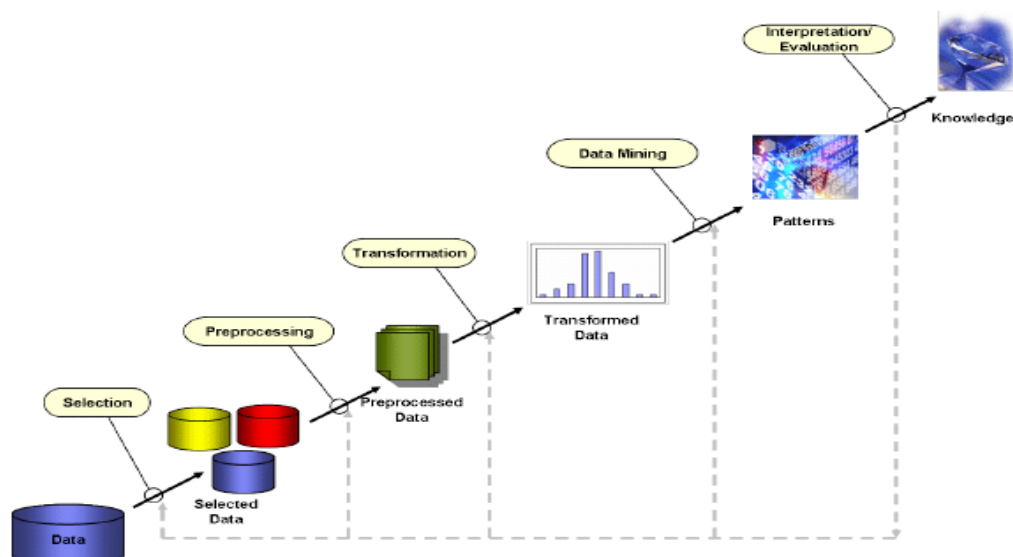


Fig 1: KDD Process

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 that can 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 student career classification project, a data warehouse is used to store and analyze data collected from various sources, including student surveys and institutional records. This centralized data repository enables efficient analysis of career preferences, and employment status trends. By integrating data across different departments and timeframes, the warehouse supports data-driven decision-making and helps in identifying patterns related to student aspirations and outcomes.

1. Data Source Layer (Extracting Data)

- Data is collected from surveys and student records.
- Includes student demographics, academic background, internships, project experience, and job search activities.

2. ETL (Extract, Transform, Load) Process

- **Extraction:** Data is gathered from multiple sources.
- **Transformation:** Data is cleaned, formatted, and standardized.
- **Loading:** The processed data is stored in the warehouse.

3. Data Storage Layer (Fact & Dimension Tables)

- **Fact Table** stores core metrics like total listening time, most-used platform, and user engagement scores.
- **Dimension Tables** include details like user demographics, platform names, and subscription types.

4. OLAP (Online Analytical Processing) for Data Analysis

- Allows multi-dimensional analysis to identify **trends in user behavior**.
- Enables queries like:
 - Which skillsets are most common among students?
 - Do students with internships have higher employment rates?
 - What factors influence job placement success?

5. Data Visualization & Reporting

- Insights are presented using **dashboards, reports, and visual charts**.
- Helps educational institutions and career advisors optimize guidance and training programs based on student career trends..

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 the **student career preferences dataset**, followed by **data preprocessing** to clean and normalize the data. The dataset is then **split into training and testing sets**. The training data is used to build and train **various classification models**. Finally, the **models classify** the data to predict a student's employment readiness based on **attributes** such as **academic background, internships, project experience, and job search activities**.

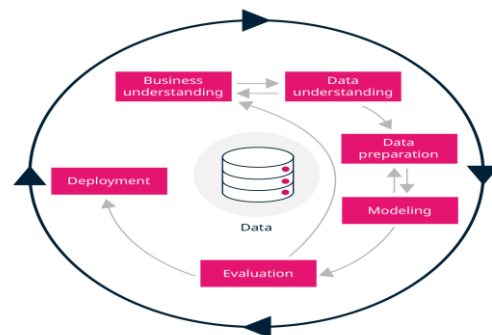


Fig2 :Data Mining Block Diagram

❓ 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 student career dataset to grasp its structure and content.
- Identifying attributes such as academic background, internships, project experience, and job search activities.

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 Decision Trees, 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 student career readiness and factors influencing employment success

❓ 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)**
- **Decision Trees**
- **Random Forest**
- **Navie Bayes**
- **Support Vector Machine**

Categories of Supervised Learning in This Project:

1. Classification:

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

Algorithm	Description	Type
SVM	Support Vector Machine (SVM) is a supervised learning algorithm used for classification and regression tasks by finding the optimal hyperplane that best separates data points into different categories. It is effective in high-dimensional spaces and is widely used in image recognition, text classification, and bioinformatics.	Classification and regression
Decision Tree	Highly interpretable classification or regression model that splits data-feature values into branches at decision nodes.	Classification
Naïve Bayes	The Bayesian method is a classification method that makes use of the Bayesian theorem. The theorem updates the prior knowledge of an event with the independent probability of each feature that can affect the event.	Regression and Classification
KNN	K-Nearest Neighbors (KNN) is a supervised learning algorithm that classifies data points based on the labels of their nearest neighbors in the feature space. It assigns the most common label among the closest data points to the new data point.	Regression and Classification

Random Forest	Random Forest is an ensemble learning algorithm that builds multiple decision trees and combines their outputs for robust and accurate classification or regression	Regression and Classification
---------------	---	-------------------------------

❑ UNSUPERVISED LEARNING

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

There are two categories of Unsupervised Learning. They are

- 1.Clustering
- 2.Association

1.Clustering:

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

2.Association:

Association analysis is a core technique in unsupervised learning within data mining, aimed at discovering relationships among different attributes or items in a dataset. Algorithms like Apriori and FP-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 student career outcomes**, **classification** algorithms are the **best fit**. However, clustering and association rule mining can be used for student segmentation and career trend analysis.



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

Windows 10 or above Operating System

Dataset:

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

Software/Tools:

- Orange Tool
- SQL Server Management Studio
- VS Code

Models:

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

Hardware Requirements

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

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

❑ PROBLEM STATEMENT

The classification of students based on their academic background, skills, and career preferences is crucial for personalized career guidance and job recommendations. However, manual categorization is time-consuming and may not accurately capture student career readiness patterns.

This project aims to develop a machine learning model to classify students into different categories (e.g., Employed, Seeking Job, Pursuing Higher Studies, Entrepreneurship) based on their academic performance, internships, technical skills, and job search activities

Objectives:

- Provide personalized career guidance based on student academic and skill profiles.
- Assist educational institutions in optimizing training programs by analyzing student career preferences.
- Help recruiters and career counselors make data-driven decisions to improve student employability.

By leveraging classification models, the system will enable **better student segmentation**, leading to **improved career guidance** and **enhanced employability support**.

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 intermediary 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.

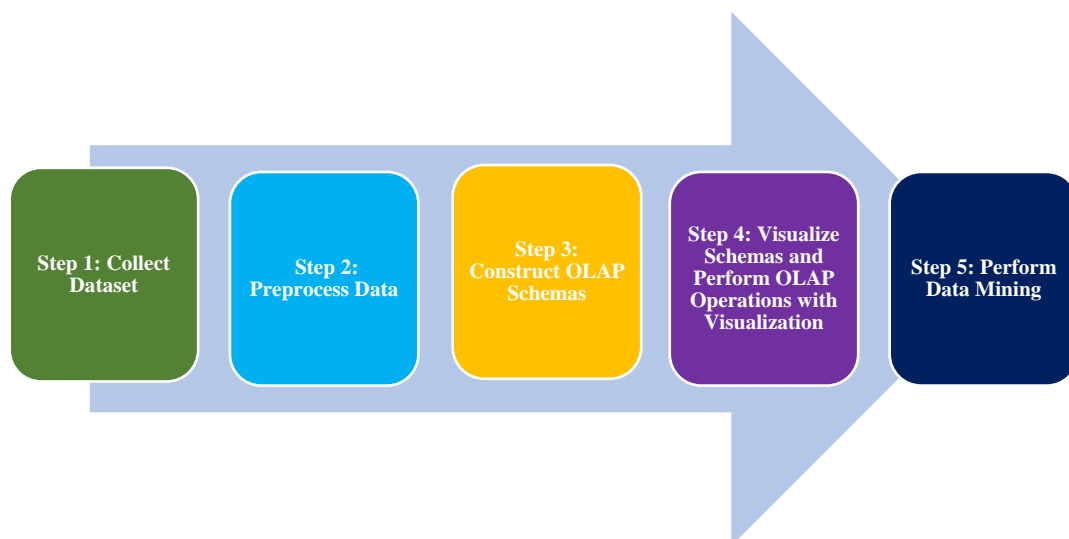


Fig4 : Knowledge Discovery From KDD

STEP-1: COLLECTING & EXPLORING DATASET

1.1. Extracting the Form to Collect Information from Users

The dataset was created by gathering information on student career preferences and employment status. Data was collected through **surveys and academic records**, covering aspects such as **education background, internships, technical skills, job search activities, and career aspirations**.

Student Career Preferences and Employment

gorilaxmi2004@gmail.com [Switch account](#)

Not shared

* Indicates required question

Student Name *

Your answer

Gender

☐ Male

☐ Female

☐ Prefer not to say

Age

Your answer

Degree Program

Choose

Year of Study

Choose

Desired Career Path *

☐ Software Development

☐ Data Science

☐ Web Development

☐ Machine Learning

☐ Cybersecurity

☐ AI

☐ Other

Skills Acquired *

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Time	Student Name	Gender	Age	Degree Program	Year of Study	Desired Career	Skills Acquired	Internship	Number Of Projects	Employment	Job Search	Job Search Platforms Used			
2	####	Prasanna	Female	19	B.Tech		3	Web Develop C/Python/Java	Yes		0	Unemployed				
3	####	Lavanya	Female	18	B.Tech		3	Web Develop C/Python/Java	Yes		1	Unemployed				
4	####	PECHETTI ROHIT SRI S	Male	19	B.Tech		3	Software Dev C/Python/Java	No		0	Unemployed				
5	####	CH JASWANTH	Male	20	B.Tech		3	Software Dev C/Python/Java	No		1	Unemployed				
6	####	Bukkuru Syam	Male	21	B.Tech		3	Web Develop C/Python/Java	Yes		2	Unemployed	2 months	LinkedIn		
7	####	Haritha	Female		B.Tech		3	Software Dev C/Python/Java	Yes		1	Unemployed		LinkedIn		
8	####	Sivanagaraju	Male	20	B.Tech		3	Software Dev C/Python/Java	No		0	Unemployed				
9	####	Hari	Male	21	B.Tech		3	Machine Learn C/Python/Java	Yes		2	Employed		LinkedIn, Naukri		
10	####	Goriparthi Venkata Ka	Male	19	B.Tech		3	Web Develop HTML/CSS/Java	No		2	Unemployed	2	LinkedIn		
11	####	Manichandra Dudduki	Male	20	B.Tech		3	Software Dev C/Python/Java	Yes		1	Unemployed	1	LinkedIn, Campus Placements		
12	####	Keerthi	Female	20	B.Tech		3	Web Develop C/Python/Java	Yes		1	Unemployed	0	LinkedIn		
13	####	Mounika	Female	19	Other		3	Web Develop HTML/CSS/Java	Yes	more than 3	Unemployed	2 months	LinkedIn, Naukri			
14	####	GOTTIPATI SAI SRAVA	Female	20	B.Tech		3	Web Develop C/Python/Java	Yes		0	Unemployed		LinkedIn		
15	####	Likitha	Female	20	B.Tech		3	Web Develop C/Python/Java	Yes		2	Unemployed	1	LinkedIn, Campus Placements		
16	####	Dollu santhosh kumar	Male	24	B.Sc			Machine Learn HTML/CSS/Java	Yes		3	Employed	2022	Indeed		
17	####	Aggala karthik	Male	20	B.Sc		4	Other	No		0	Unemployed		Other		
18	####	Gullipalli chanti	Male	21	B.Sc		3	Web Develop	Others	No	0	Unemployed	1	Indeed, Naukri		
19	####	Yamini	Female	20	B.Tech		3	Web Develop C/Python/Java	Yes		0	Unemployed	0	LinkedIn, Campus Placements		
20	####	Gorle dharani	Female	18	B.Sc		2	Other	C/Python/Java	No	0	Unemployed	0	Other		
21	####	Jaya Shankar	Male	20	B.Tech		3	Other	Others	No	0	Unemployed		LinkedIn		
22	####	Kimidi susmitha	Female	18	B.Sc		2	Data Science, C/Python/Java	No		0	Unemployed		Campus Placements, Other		
23	####	Paroju Harika	Female	18	B.Sc		2	Other	C/Python/Java	No	0	Unemployed	0	Other		

Student Career Preferences and

The attributes are:

1. Student Name
2. Gender
3. Age
4. Degree Program
5. Year of Study
6. Desired Career path
7. Skills Acquired
8. Internship Experience
9. Number Of Projects Done
10. Employment Status
11. Job Search Duration
12. Job Search Platforms Used

1.2 Defining Survey or Data Collection Methods

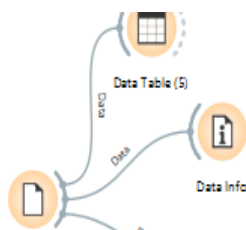
- **Online Surveys:** A structured questionnaire was distributed online, including multiple-choice questions to capture student preferences, academic background, career choices, and employment status..

Link: [Student Career Preferences and Employment \(google.com\)](https://www.google.com)

1.3 Choosing Attributes for Analysis

The key attributes selected for analysis include:

- **Career Status** – Identifies whether the student is employed, seeking a job, pursuing higher studies, or interested in entrepreneurship.
- **Student Demographics** – Includes age, gender, and location to analyze career trends.
- **Academic Background** – Tracks degree, specialization, and academic performance.
- **Internship Experience** – Differentiates between students with and without internship experience.
- **Technical Skills & Certifications** – Identifies key skills and certifications relevant to career readiness.



Data table properties	
Name:	Student Career Preferences and Employment (Responses) - Form responses 1
Size:	~83 rows, 13 columns
Features:	11 categorical
Targets:	categorical outcome with 2 classes
Metas:	1 text
Missing data:	39 (4.3%) in features
Additional attributes	
83	

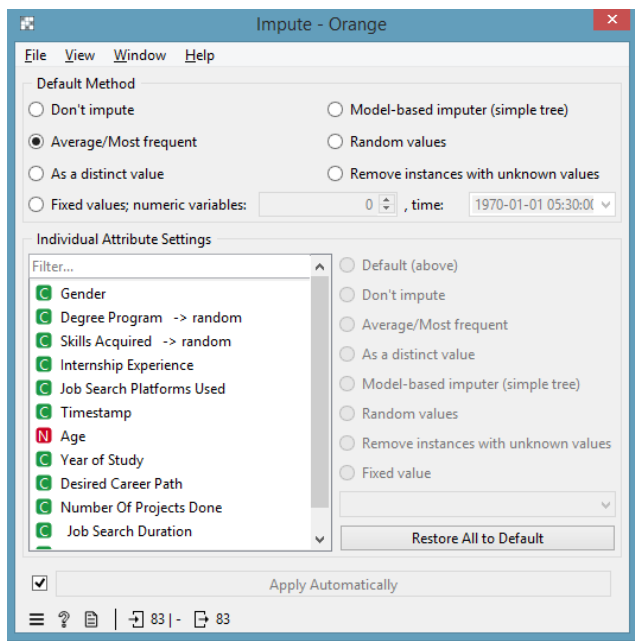
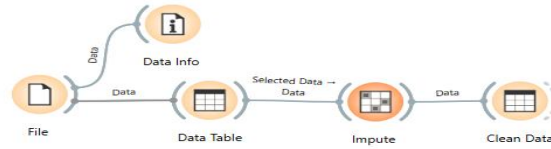
employment Status	Student Name	Gender	Degree Program	Skills Acquired	Internship Experience	Job Search Platforms Used	Timestamp	Age	Year of Study	Desired Career Path	Number Of Projects Done	Job Search Duration
Unemployed	Prasanna	Female	B.Tech	C/Python/Java...	Yes	Other	02/02/2025 12:1...	19	3	Web Developm...	0	?
Unemployed	Lavanya	Female	B.Tech	C/Python/Java...	Yes	?	02/02/2025 12:2...	18	3	Web Developm...	1	?
Unemployed	PECHETTI ROHL...	Male	B.Tech	C/Python/Java...	No	?	02/02/2025 12:2...	19	3	Software Devel...	0	?
Unemployed	CH JASWANTH	Male	B.Tech	C/Python/Java...	No	LinkedIn, Camp...	02/02/2025 12:2...	20	3	Software Devel...	1	0
Unemployed	Bukkuru Syam	Male	B.Tech	C/Python/Java...	Yes	LinkedIn	02/02/2025 12:3...	21	3	Web Developm...	2	2 months
Unemployed	Haritha	Female	B.Tech	C/Python/Java...	Yes	LinkedIn, Camp...	02/02/2025 12:3...	?	3	Software Devel...	1	?
Unemployed	Sivanagaraju	Male	B.Tech	C/Python/Java...	No	?	02/02/2025 12:3...	20	3	Software Devel...	0	?
Employed	Hari	Male	B.Tech	C/Python/Java...	Yes	LinkedIn, Naukri	02/02/2025 12:4...	21	3	Machine Learni...	2	?
Unemployed	Goriparthi Venk...	Male	B.Tech	HTML/CSS/Java...	No	LinkedIn	02/02/2025 12:4...	19	3	Web Developm...	2	2
Unemployed	Manichandra D...	Male	B.Tech	C/Python/Java...	Yes	LinkedIn, Camp...	02/02/2025 13:0...	20	3	Software Devel...	1	1
Unemployed	Keerthi	Female	B.Tech	C/Python/Java...	Yes	LinkedIn	02/02/2025 13:0...	20	3	Web Developm...	1	0
Unemployed	Mounika	Female	Other	HTML/CSS/Java...	Yes	LinkedIn, Naukri	02/02/2025 13:0...	19	3	Web Developm...	more than 3	2 months
Unemployed	GOTTIPATI SAI ...	Female	B.Tech	C/Python/Java...	Yes	LinkedIn	02/02/2025 13:0...	20	3	Web Developm...	0	?
Unemployed	Lakitha	Female	B.Tech	C/Python/Java...	Yes	LinkedIn, Camp...	02/02/2025 13:0...	20	3	Web Developm...	2	1
Employed	Dollu santhosh ...	Male	B.Sc	HTML/CSS/Java...	Yes	Indeed	02/02/2025 13:0...	24	?	Machine Learni...	3	2022
Unemployed	Aggala karthik	Male	B.Sc	Others	No	Other	02/02/2025 13:1...	20	4	Other	0	?
Unemployed	Gullipalli chanti	Male	B.Sc	Others	No	Indeed, Naukri	02/02/2025 13:2...	21	3	Web Developm...	0	01
Unemployed	Yamini	Female	B.Tech	C/Python/Java...	Yes	LinkedIn, Camp...	02/02/2025 13:2...	20	3	Web Developm...	0	0
Unemployed	Gorle dharani	Female	B.Sc	C/Python/Java...	No	Other	02/02/2025 13:2...	18	2	Other	0	0
Unemployed	Jaya Shankar	Male	B.Tech	Others	No	LinkedIn	02/02/2025 13:2...	20	3	Other	0	?
Unemployed	Kimidi susmitha	Female	B.Sc	C/Python/Java...	No	Campus Place...	02/02/2025 13:3...	18	2	Data Science, O...	0	?
Unemployed	Parju Harika	Female	B.Sc	C/Python/Java...	No	Other	02/02/2025 13:3...	18	2	Other	0	0
Unemployed	Aswini	Female	Other	Others	No	Other	02/02/2025 13:4...	45	4	Other	0	?
Unemployed	B Navya Sai sree	Female	B.Tech	Others	Yes	LinkedIn, Naukri...	02/02/2025 13:4...	19	3	Other	1	?
Unemployed	Nohiya	Female	B.Sc	C/Python/Java...	No	Other	02/02/2025 13:5...	19	2	Other	0	?
Unemployed	MEGHANA YA...	Female	B.Tech	C/Python/Java...	Yes	LinkedIn	02/02/2025 13:5...	19	3	Software Devel...	0	1
Unemployed	Balasingi avanthi	Female	B.Sc	Others	No	Other	02/02/2025 14:0...	31-06-2006	2	Other	2	2

Step-2: PREPROCESS THE DATA

Preprocess the Dataset Using ORANGE TOOL

2.1 Handling Missing Values

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



2.2 Data Cleaning & Transformation

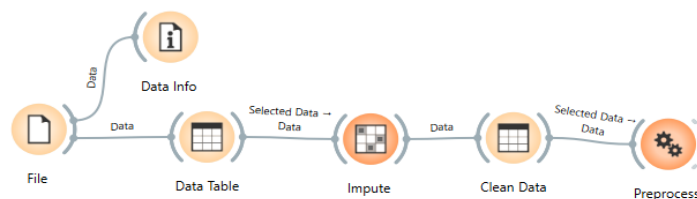
- Standardized text-based attributes.
- Converted categorical values into numerical form for analysis.

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.



STEP-3: CREATION OF DATABASE CONSTRUCT OLAP SCHEMAS

The above table was normalized and divided into multiple tables

DIMENSION TABLES:

1. Dim_Gender

Attribute	Data Type	Key Type
GenderID	INT	Primary Key
Gender	NVARCHAR(10)	

2. Dim_Degree

Attribute	Data Type	Key Type
DegreeID	INT	Primary Key
Degree_Program	NVARCHAR(50)	

3. Dim_Skills

Attribute	Data Type	Key Type
SkillID	INT	Primary Key
Skills_Acquired	NVARCHAR(255)	

4. Dim_Technologies

Attribute	Data Type	Key Type
TechID	INT	Primary Key
Technologies_Used	NVARCHAR(255)	

5. Dim_Student

Attribute	Data Type	Key Type
StudentID	INT	Primary Key
Name	NVARCHAR(100)	
GenderID	INT	Foreign Key → Dim_Gender(GenderID)
Age	INT	
DegreeID	INT	Foreign Key → Dim_Degree(DegreeID)
Year_of_Study	INT	

6. Dim_Career

Attribute	Data Type	Key Type
CareerID	INT	Primary Key
Desired_Career_Path	NVARCHAR(100)	
SkillID	INT	Foreign Key → Dim_Skills(SkillID)

7. Dim_Project

Attribute	Data Type	Key Type
ProjectID	INT	Primary Key
Number_Of_Projects_Done	INT	
TechID	INT	Foreign Key → Dim_Technologies(TechID)

8. Dim_Internship

Attribute	Data Type	Key Type
InternshipID	INT	Primary Key
Internship_Experience	NVARCHAR(3)	
Duration	NVARCHAR(20)	

FACT TABLES:

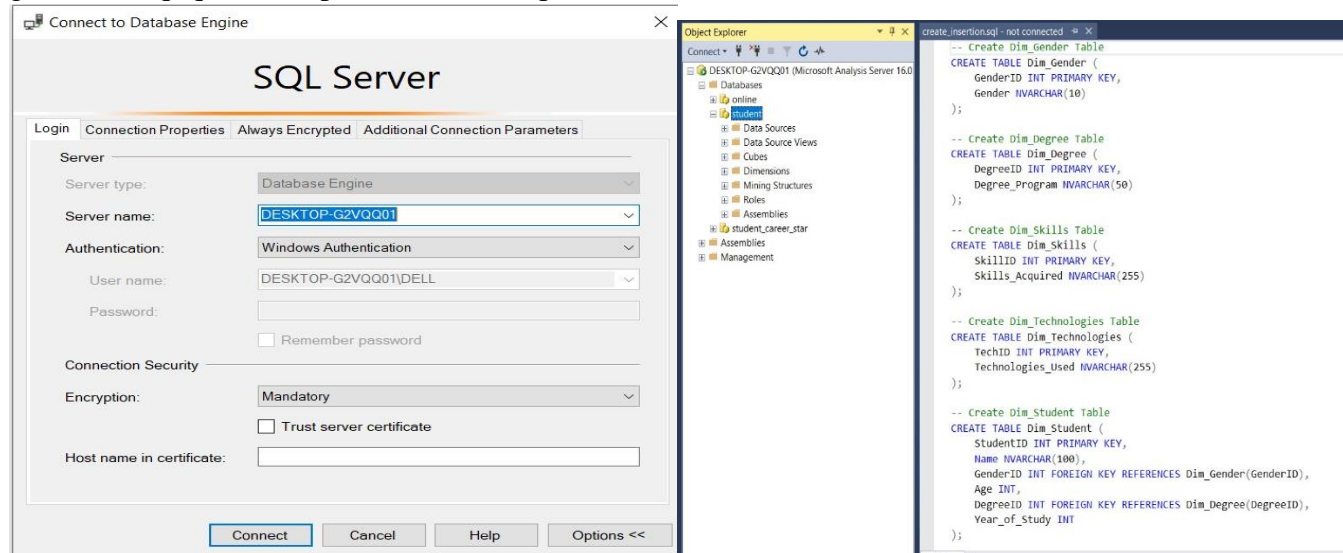
9. Fact_Employment

Attribute	Data Type	Key Type
EmploymentID	INT	Primary Key
StudentID	INT	Foreign Key → Dim_Student(StudentID)
CareerID	INT	Foreign Key → Dim_Career(CareerID)
InternshipID	INT	Foreign Key → Dim_Internship(InternshipID)
Employment_Status	NVARCHAR(50)	
Job_Search_Duration	NVARCHAR(20)	
Job_Search_Platforms_Used	NVARCHAR(255)	

10. Fact_Project_Details

Attribute	Data Type	Key Type
ProjectDetailID	INT	Primary Key
StudentID	INT	Foreign Key → Dim_Student(StudentID)
ProjectID	INT	Foreign Key → Dim_Project(ProjectID)
Project_Completion_Status	NVARCHAR(20)	

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

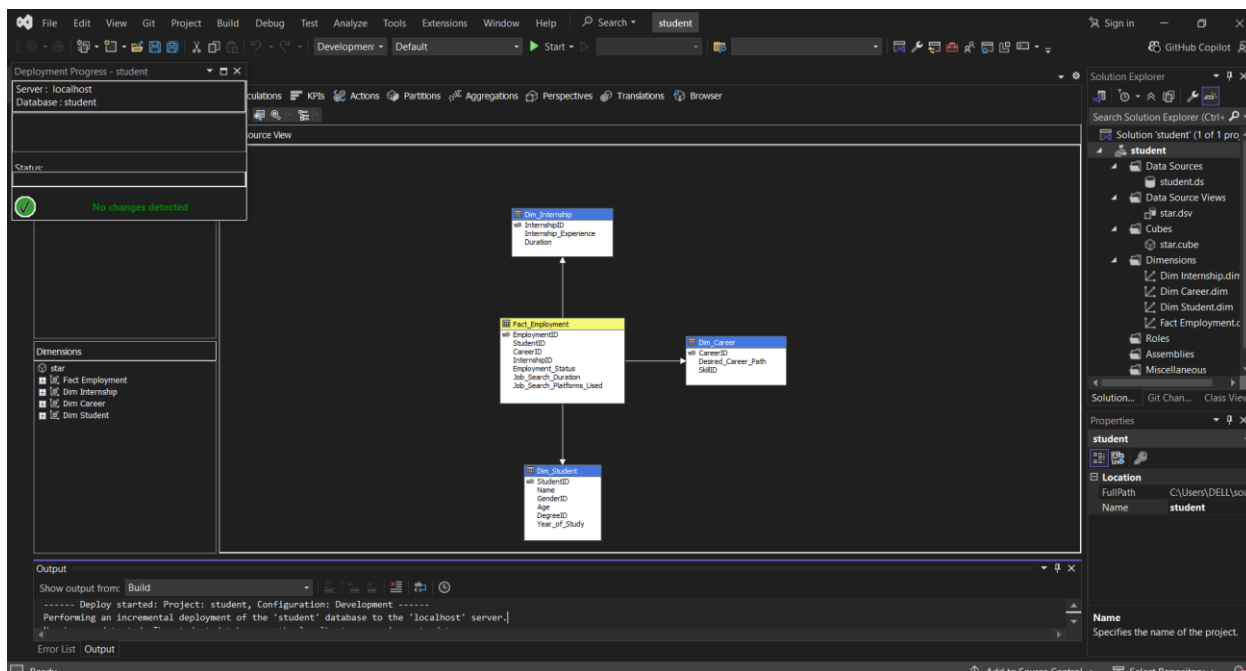
- Developed a multidimensional analysis project using Visual Studio.
- Configured Data Source & Data Source View, establishing connections to the database and defining table relationships.
- Designed database schema diagrams to visualize data structure.
- Validated table relationships to ensure data integrity.
- Created Cubes & Measures, defining fact tables, dimensions, and key performance measures for analysis.

4.1 STAR SCHEMA:

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

4.1.1 Design & Visualize the Schema

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



Schemas one after another

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

4.1.2 Deploy the Data Warehouse & Load Data

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

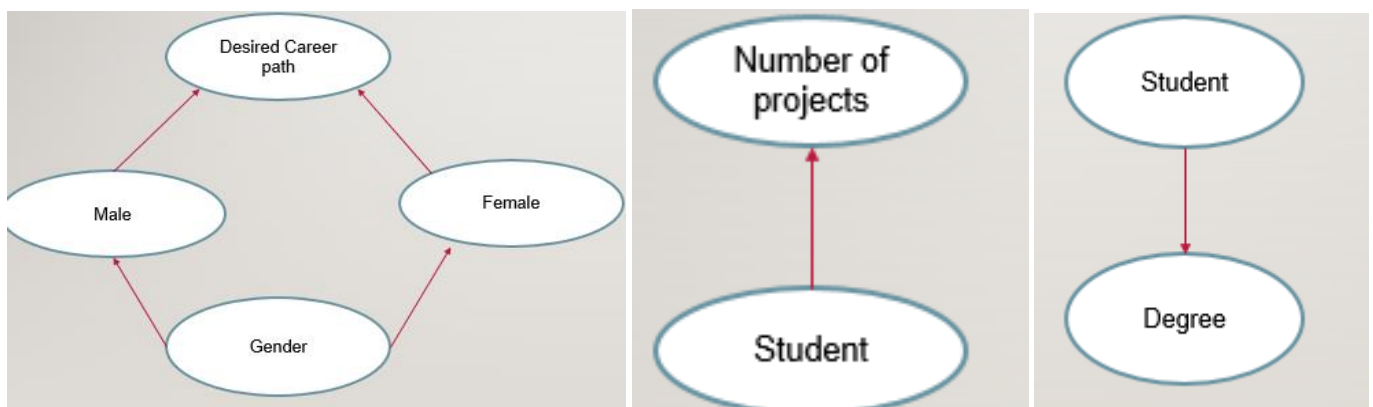
4.1.3 Create & Execute OLAP Queries

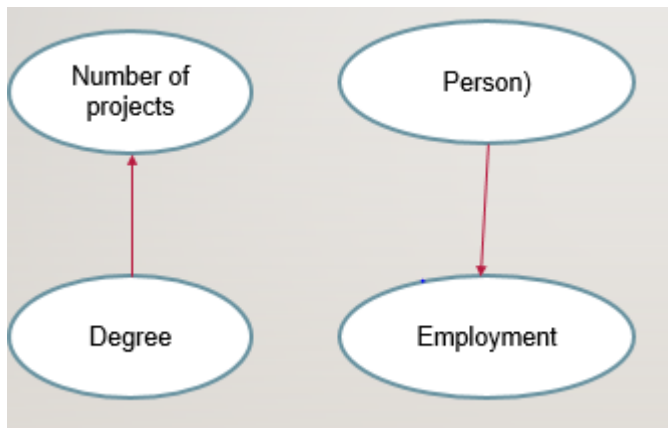
- Write OLAP queries to perform data analysis.
- Use ROLLUP, SLICE, DICE, DRILL-DOWN, and PIVOT operations for multi-dimensional analysis.

4.1.4 Perform OLAP Operations

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

Concept Hierarchies used :





MDX Queries OLAP operations in STAR SCHEMA:

A)What is the overall gender distribution of students related to career path?

Roll-Up (Aggregation of Genders):

SELECT

NON EMPTY [Dim Student].[Gender ID].Members ON COLUMNS,

NON EMPTY [Dim Career].[Desired Career Path].Members ON ROWS

FROM

[star]]

Output:

Messages		
	Male	Female
Software Development	15	12
Data Analysis	9	14
Cybersecurity	7	5
Business Management	10	11
UI/UX Design	4	6

B))list of all students under each Student ID and career path? ?

Drill-Down:

SELECT

[Dim Student].[Gender ID].[All] ON COLUMNS,

NON EMPTY

CROSSJOIN(

[Dim Student].[Student ID].Members,

[Dim Career].[Desired Career Path].Members

) ON ROWS

FROM [star] **OUTPUT:**

Messages Results		
	Student ID	(All)
1	1	Software Development
2	2	Data Analysis
3	3	Cybersecurity
4	4	Business Management
5	5	UI/UX Design
6	6	Software Development
7	7	Data Analysis

C) What is the internship experience based on different durations?- duration greater than 3 months.

Slice :

```
SELECT
NON EMPTY FILTER(
[Dim Internship].[Duration].MEMBERS,
[Dim Internship].[Duration].CurrentMember.Name > "2"
) ON COLUMNS,
NON EMPTY [Dim Internship].[Internship Experience].MEMBERS ON ROWS
FROM
[star]
```

OUTPUT:

	All	2 months	3 months
All	10	2	2
No	4	(null)	(null)
Yes	6	2	2

D) What are the desired career paths of students, filtered by gender - Gender ID is not equal to 2.Dice:

```
SELECT
NONEMPTY(
[Dim Career].[Desired Career Path].Members
) ON COLUMNS,
FILTER(
[Dim Student].[Gender ID].Members,
[Dim Student].[Gender ID].CurrentMember.Name <> "2"
AND [Dim Student].[Gender ID].CurrentMember.Name <> "Unknown"
AND [Dim Student].[Gender ID].CurrentMember.Name <> ""
) ON ROWS
FROM
[star]
```

OUTPUT:

	All	AI/ML	Cyber Security	Data Science	Software Development	Web Development
All	10	2	2	2	2	2
1	4	(null)	1	(null)	2	1

E) How do career paths vary by gender?

Pivot (Rearranging Dimensions):

```
SELECT
NONEMPTY(
[Dim Student].[Gender ID].Members
) ON COLUMNS,
NONEMPTY(
```

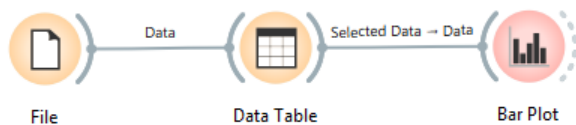
```
[Dim Career].[Career ID].Members
) ON ROWS
FROM
[star]
```

OUTPUT:

	All	1	2
All	10	4	6
1	2	1	1
2	2	2	(null)
3	2	(null)	2
4	2	1	1
5	2	(null)	2

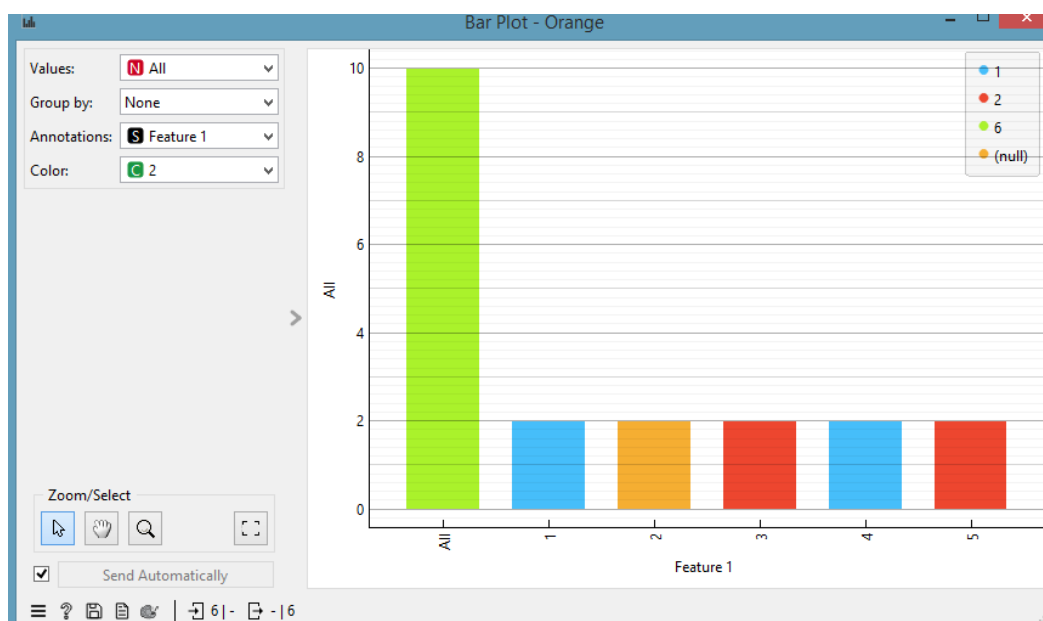
4.1.5 Visualize OLAP Results:

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



Bar Plot Configuration:

- **X-Axis:** Career ID (Different career categories)
- **Y-Axis:** Number of Students
- **Observation:**
 - The bar plot will display the count of students based on their **Career ID and Gender ID**.
 - **Each bar represents the total number of students** within a specific career category.
 - **Different gender groups (Gender ID: 1, 2) are shown as separate bars** for comparison.
 - The "All" category aggregates the total student count across all careers.

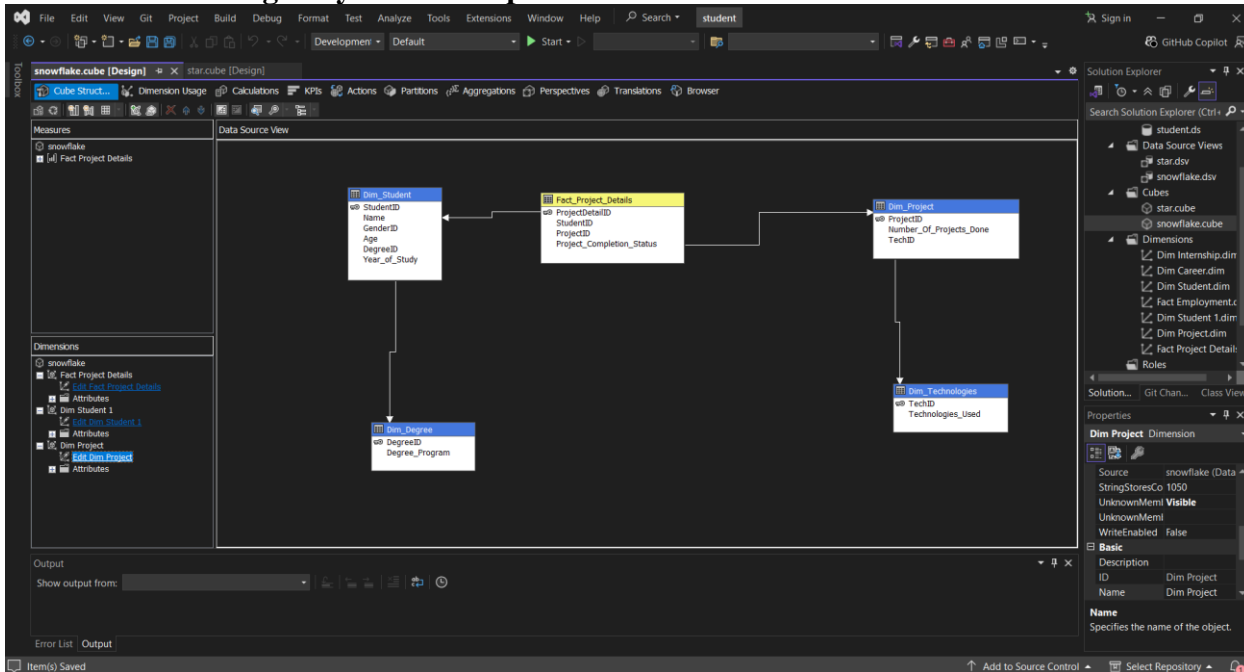


4.2 SNOWFLAKE SCHEMA:

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

4.2.1 Design & Visualize the Snowflake Schema

- Identify **Fact Tables** (e.g., Project_details)
- Identify **Dimension Tables** (e.g., student, project).
- Normalize dimension tables by breaking them into **sub-dimensions** (e.g., student → degree).
- 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.

MDX Queries OLAP operations in SNOWFLAKE SCHEMA:

(A) Total Project Count By each student?

Roll-Up (Aggregation):

SELECT


```

{[Measures].[Fact Project Details Count]} ON COLUMNS,
NON EMPTY
CROSSJOIN(
  [Dim_Student].[Name].MEMBERS,
  CROSSJOIN(
    [Dim_Degree].[Degree Program].MEMBERS,
    CROSSJOIN(
      [Dim_Project].[Number Of Projects Done].MEMBERS,
      [Dim_Technologies].[Technologies Used].MEMBERS
    )
  )
) ON ROWS
FROM [snowflake]

```

OUTPUT:

All	Student ID	Project ID	Degree	Technology	Fact Project
All	1	2	BTech	Technology	8 d00
All	4	2	MTech	Technology	10
All	5	3	MTech	Technology	7
All	7	3	MTech	Technology	12
All	10	3	PhD	Technology	6

(B) View All Degree Program

Drill-Down

```

SELECT
EXCEPT(
  [Dim Student 1].[Degree Program].Members,
  {[Dim Student 1].[Degree Program].[Unknown]}
) ON COLUMNS
FROM [snowflake]

```

OUTPUT:

Messages		Results	
All	BTech	MTech	
10	9	1	

(C) Display Degree Programs and Project Completion Status

Slice

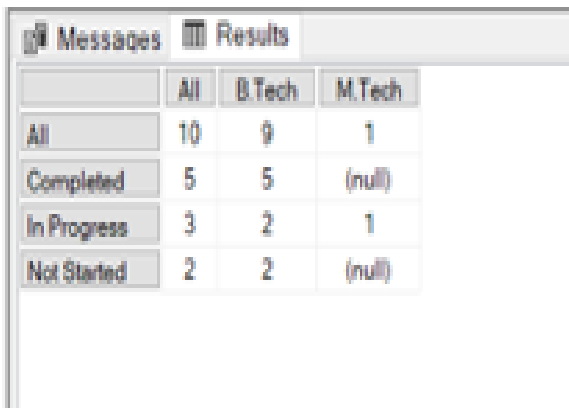
```

SELECT
NON EMPTY [Dim Student 1].[Degree Program].Members ON COLUMNS,

```

NON EMPTY [Fact Project Details].[Project Completion Status].Members ON ROWS
FROM [snowflake]

OUTPUT:



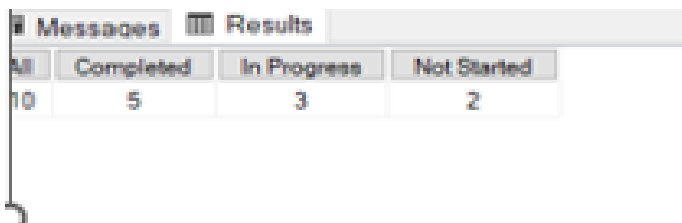
	All	B.Tech	M.Tech
All	10	9	1
Completed	5	5	(null)
In Progress	3	2	1
Not Started	2	2	(null)

(D) View Project Completion Status

Dice

SELECT
[Fact Project Details].[Project Completion Status].Members ON COLUMNS
FROM [snowflake]

OUTPUT:



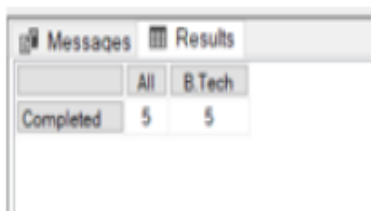
	Completed	In Progress	Not Started
All	5	3	2

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

Pivot (Rearranging Dimensions):

SELECT
FILTER(
[Dim Student 1].[Degree Program].Members,
[Dim Student 1].[Degree Program].CurrentMember.Name < "M"
AND [Dim Student 1].[Degree Program].CurrentMember.Name <> "Unknown"
) ON COLUMNS,
FILTER(
[Fact Project Details].[Project Completion Status].Members,
[Fact Project Details].[Project Completion Status].CurrentMember.Name = "Completed"
) ON ROWS
FROM [snowflake]

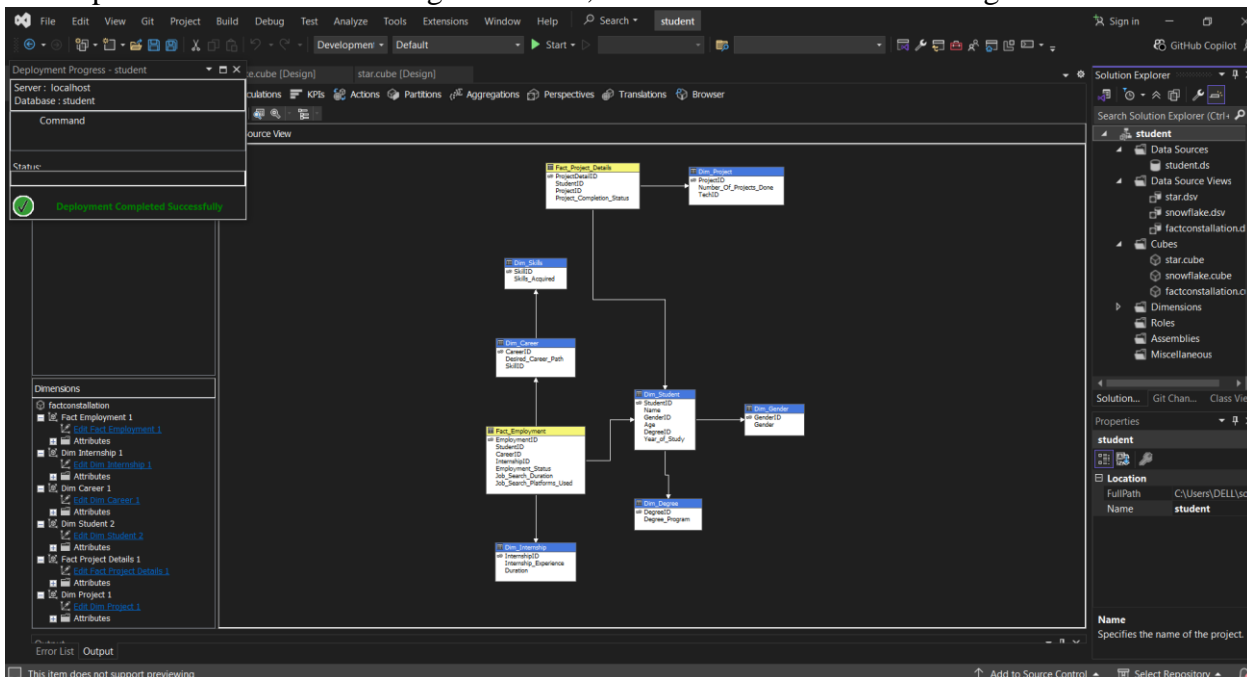
OUTPUT:



	All	B.Tech
Completed	5	5

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) Display the summarized project completion by degree program

Roll-Up

SELECT

NON EMPTY

CROSSJOIN(

[Dim Student 2].[Degree Program].Members,

[Dim Student].[Year of Study].Members

) ON COLUMNS,

NON EMPTY

CROSSJOIN([Fact Project Details 1].[Project Completion Status].Members,

[Dim Project 1].[Number of Projects Done].Members,

[Dim Internship 1].[Duration].Members,

[Dim Career 1].[Desired Career Path].Members,

[Dim Skills].[Skills Acquired].Members,

[Fact Employment].[Employment Status].Members

) ON ROWS

FROM [factconstallation]

WHERE ([Measures].[Fact Project Details 1 Count])

OUTPUT:

	Arts		Year 2		Year 1		Engineering	
	Year of Study		Year of 1	Year 2	Year of Study	Year 1	Year 2	
Project Completed	1 Prodiect	1		4		12	12	17
Duration of Desired Career Path	1	2	Technology	4	Healthcare	4	3	6
	6-months	6	Technology	6	Education	4	6-month	Technical
	Employed	5	Employed	0		4	4	3
Immpetely Completed	1 Progiect	3	Healthcare	4	Healthcare	4	4	3
	3	3	Pducation	4	Technical	4	4-month	Anlytical
	2-months	2	Problem-Solt	0	Problem-Solv		3	0
Not Completed	1 Progiect	4	Business	3		3	3	3
	1	4	Leadership	4	Teamwork	4	3	Teamwork
	Employed	6	Unemployed	0	Unemployed		5-month	5
Not Completed	1 Progiect	1	Business	3	Technology	6	3	6
	3	3	Technology	4	6-months	6	6-month	Technical
	Unemploye	0	Employed	0		0	3	3
	1 Progiect	3	Business	3	Technology	6	2	3
	3	3	Technology	4	6-months	6	6-month	Technical
	Employed	7	Technical Skiv	0	Technical		3	0
	Unemploye	6	Unemployed	7		3	3	5

(B) student names and their employment status :

Drill-Down

SELECT

FILTER([Fact Employment 1].[Employment Status].Members, [Fact Employment 1].[Employment Status].CurrentMember.Name <> "Unknown") ON COLUMNS,

FILTER([Dim Student 2].[Name].Members, [Dim Student 2].[Name].CurrentMember.Name <> "Unknown") ON ROWS

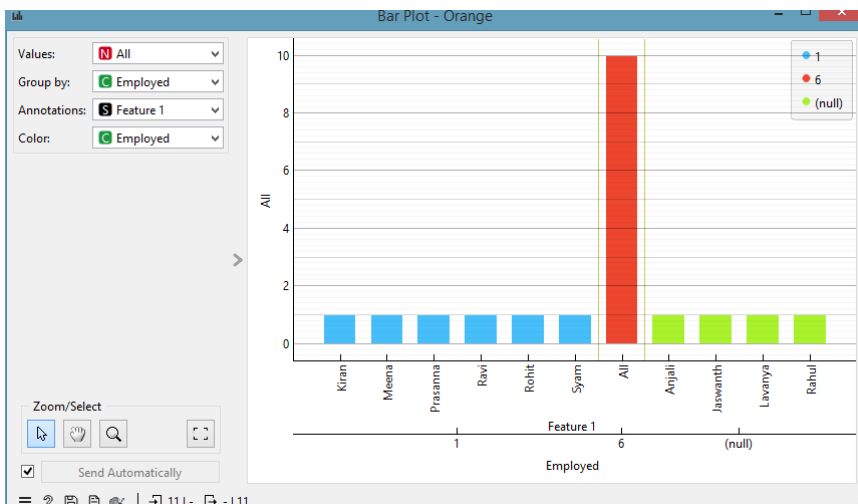
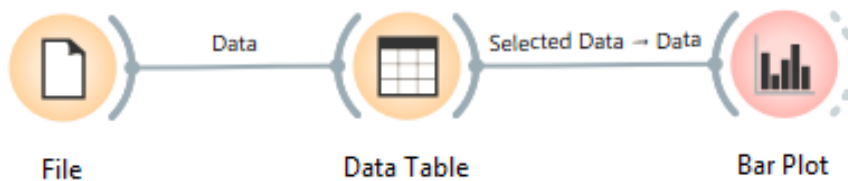
FROM [factconstallation]

OUTPUT:

Results			
	All	Employed	Unemployed
All	10	6	4
Anjali	1	(null)	1
Jaswanth	1	(null)	1
Kiran	1	1	(null)
Lavanya	1	(null)	1
Meena	1	1	(null)
Prasanna	1	1	(null)
Rahul	1	(null)	1
Ravi	1	1	(null)
Rohit	1	1	(null)
Syam	1	1	(null)

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) Display all male students whose name is not 'Unknown' and age is greater than 18.

Slice

SELECT

FILTER(

[Dim Student 2].[Name].Members,

```
[Dim Student 2].[Name].CurrentMember.Name <> "Unknown"
AND [Dim Student 2].[Age].CurrentMember.Name > 18
) ON COLUMNS
FROM [factconstallation]
WHERE [Dim Student 2].[Gender].[Male]
OUTPUT:
```

OUTPUT:

Messages Results						
All	Jaswanth	Kiran	Rahul	Ravi	Rohit	Syam
6	1	1	1	1	1	1

(D) View Project Completion Status

Dice

```
SELECT
[Fact Project Details 1].[Project Completion Status].Members ON COLUMNS
FROM [factconstallation]
```

OUTPUT:

Messages Results			
All	Completed	In Progress	Not Started
10	10	10	10

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

Pivot

```
SELECT
FILTER([Dim Student 2].[Gender].Members, [Dim Student 2].[Gender].CurrentMember.Name <>
"Unknown") ON COLUMNS,
FILTER([Dim Student 2].[Age].Members, [Dim Student 2].[Age].CurrentMember.Name <> "Unknown")
ON ROWS
FROM [factconstallation]
OUTPUT:
```

Messages Results			
	All	Female	Male
All	10	4	6
18	1	1	(null)
19	3	2	1
20	3	(null)	3
21	2	(null)	2
22	1	1	(null)

STEP 5: PERFORM DATA MINING

Classification of Students Based on Employment Status

Objective:

The goal is to classify students based on their **employment status** (Employed, Unemployed) using **Supervised Machine Learning techniques**.

5.1 DATA PREPARATION FOR CLASSIFICATION

- **Dataset Features:**

1. **Student Attributes:**

- Name, Gender, Age, Department

2. **Academic Performance:**

- CGPA, Attendance Percentage, Number of Backlogs

3. **Skill Set & Certifications:**

- Technical Skills (e.g., Python, Java, SQL)
- Certifications (e.g., AWS, Data Science)

4. **Placement Preparation:**

- Internship Experience
- Number of Mock Interviews Attended
- Number of Companies Applied

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**

- Normalizing numerical values like **CGPA, Attendance Percentage, and Internships** using **Min-Max Scaling**.

3. **Balancing the Dataset**

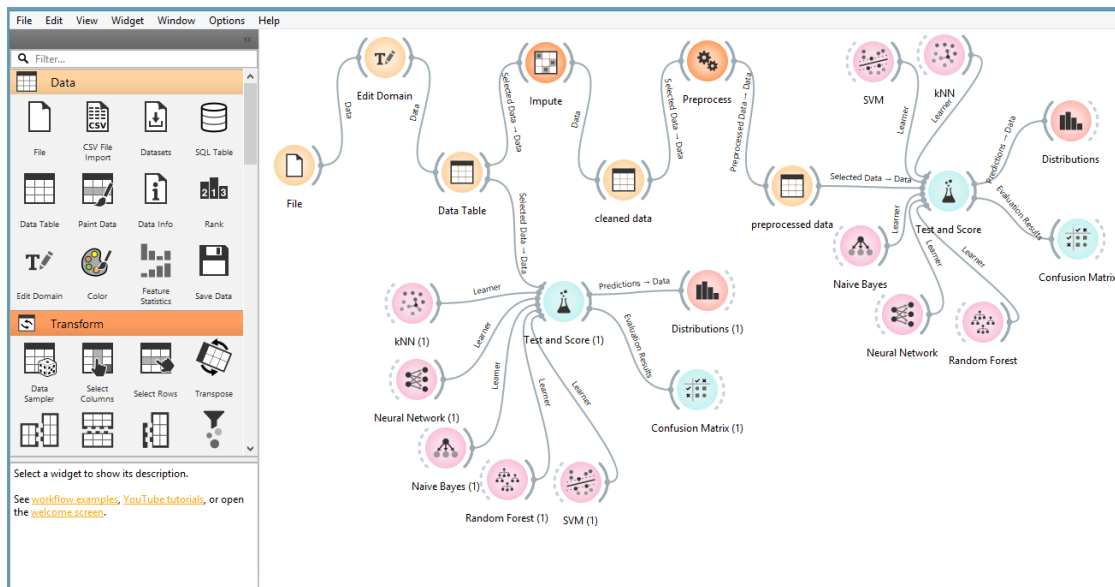
- If the dataset is **imbalanced** (i.e., more employed students than unemployed), apply **SMOTE (Synthetic Minority Over-sampling Technique)** to balance the classes.

5.2 SELECTING CLASSIFICATION ALGORITHMS

We use **Supervised ML models** to predict the user's preferred music streaming platform, including:



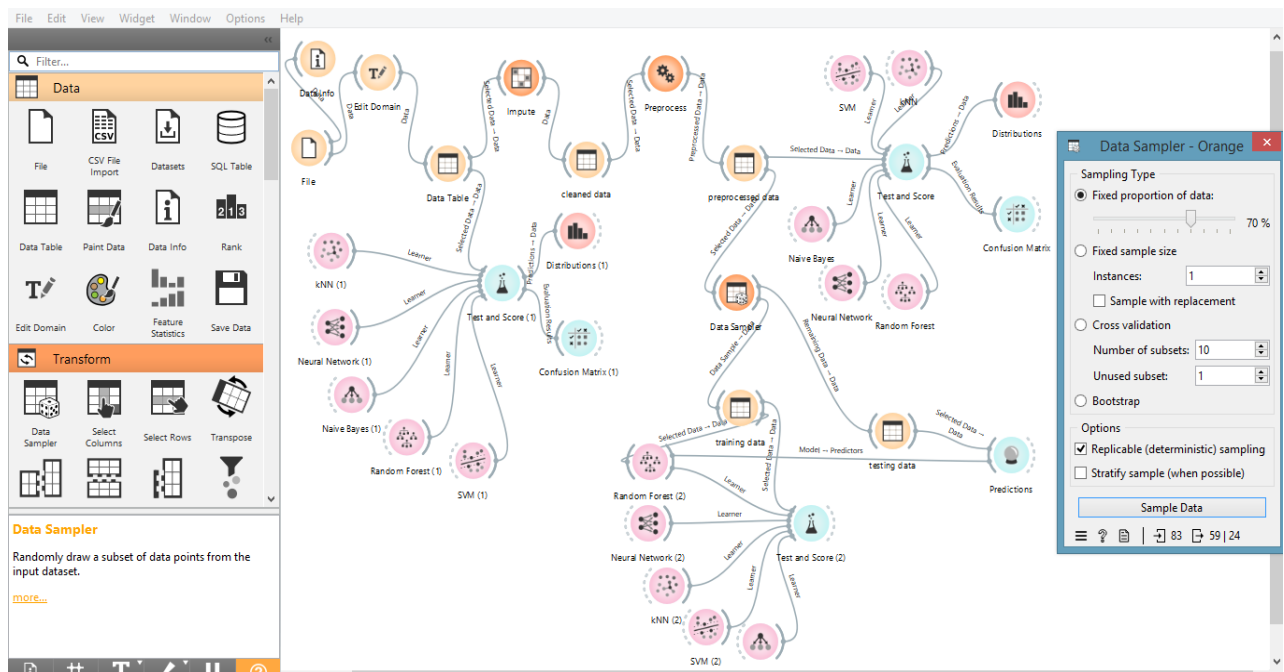
- **Random Forest** 🌲
- **Decision Tree** 🌳
- **Neural Networks** •
- **Navie Bayes** 📊
- **KNN (K-Nearest Neighbors)** 🏠
- **Support Vector Machine (SVM)** 📊



- 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 data including academic performance, technical skills, certifications, and placement preparation efforts.
- Label data based on the preferred streaming platform.
- Label each record with the **employment status** (Employed / Unemployed).
- Handle missing values and normalize numeric values

Step 2: Model Training & Classification

- Train classification models to **predict employment status** of students.

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	0.795	0.867	0.806	0.753	0.867	0.000
	Neural Network	0.841	0.692	0.692	0.697	0.692	0.584
	RANDOM FOREST	0.905	0.880	0.833	0.894	0.880	0.283
	KNN	0.842	0.855	0.849	0.844	0.855	0.321
	Naive Bayes	0.955	0.723	0.767	0.910	0.723	0.469
WITH PREPROCESSING	SVM	0.821	0.867	0.806	0.753	0.867	0.000
	Neural Network	0.847	0.867	0.840	0.835	0.867	0.244
	RANDOM FOREST	0.849	0.892	0.857	0.904	0.892	0.402
	KNN	0.770	0.867	0.840	0.835	0.867	0.244
	Naive Bayes	0.947	0.711	0.757	0.909	0.711	0.458

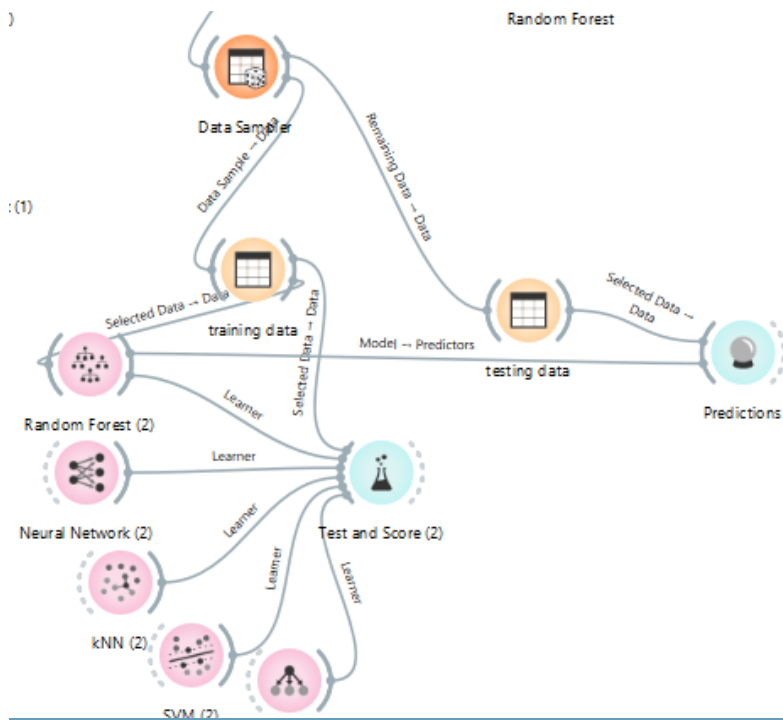
“ RANDOM FOREST Model Achieved the Highest Accuracy “

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

VISUALIZATION & PREDICTION ANALYSIS:

- **Data Preprocessing & Sampling:**

The 30% testing data is used for Prediction



Predictions - Orange

Show probabilities for: Classes in data ☒ Show classification errors Restore Original Order

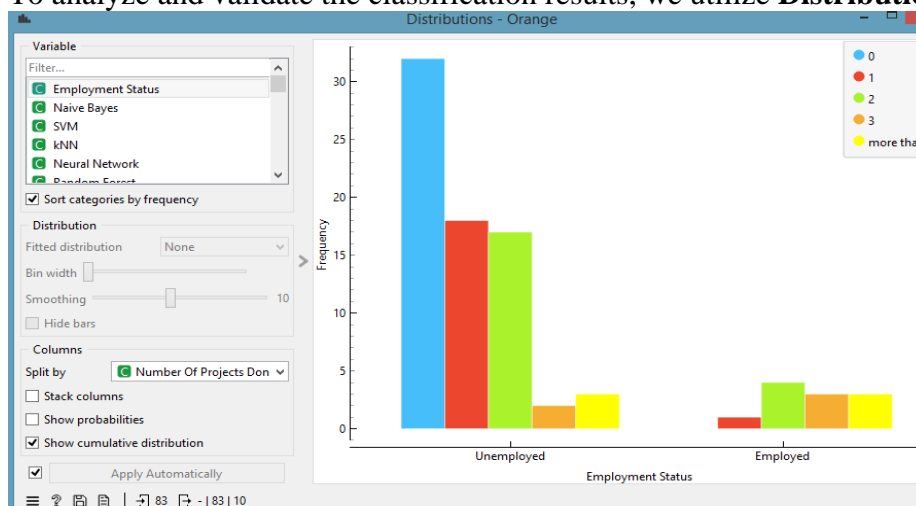
	Random Forest (2)	error	employment Status	Student Name	Gender	Degree Program
1	0.00 : 1.00 → Unemployed	0.000	Unemployed	Nayak	Male	B.Tech
2	0.32 : 0.68 → Unemployed	0.680	Employed	Alekhyia	Female	B.Tech
3	0.22 : 0.78 → Unemployed	0.220	Unemployed	Aakash	Male	B.Tech
4	0.00 : 1.00 → Unemployed	0.000	Unemployed	C anvesh	Male	B.Tech
5	0.47 : 0.53 → Unemployed	0.475	Unemployed	Mounika	Female	Other
6	0.00 : 1.00 → Unemployed	0.000	Unemployed	T. Vasanth	Male	B.Tech
7	0.00 : 1.00 → Unemployed	0.000	Unemployed	K.Dilip Kumar	Male	B.Tech

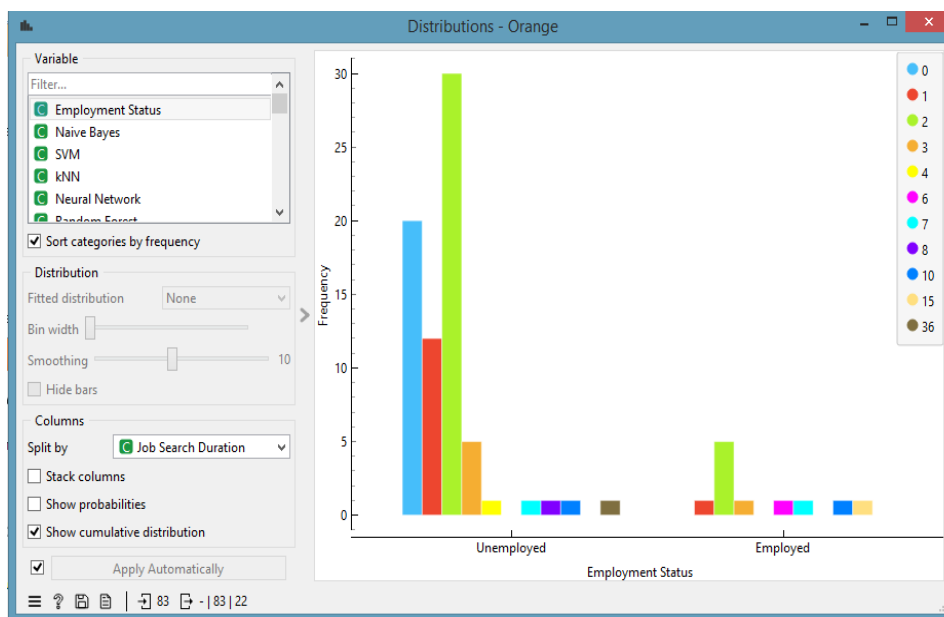
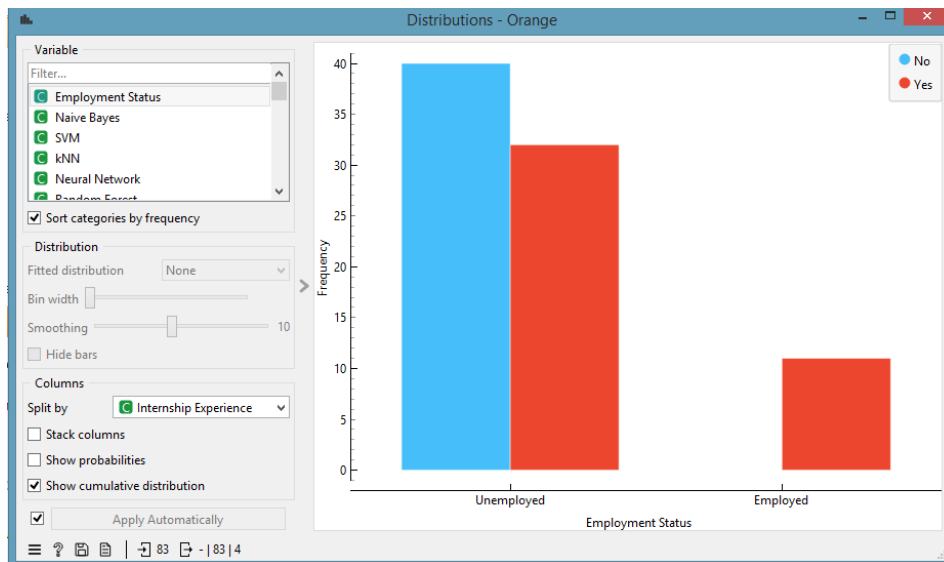
☒ Show performance scores Target class: (Average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest (2)	0.889	0.792	0.700	0.627	0.792	0.000

5.6 VISUALIZATION METRICS FOR CLASSIFICATION:

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

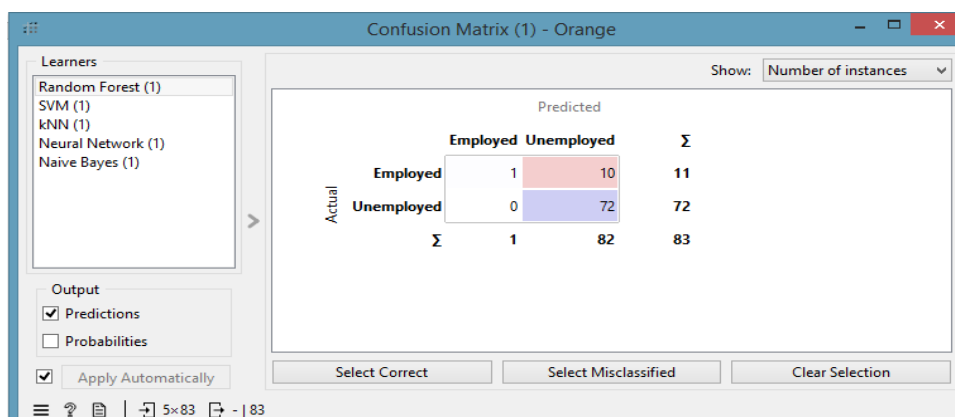




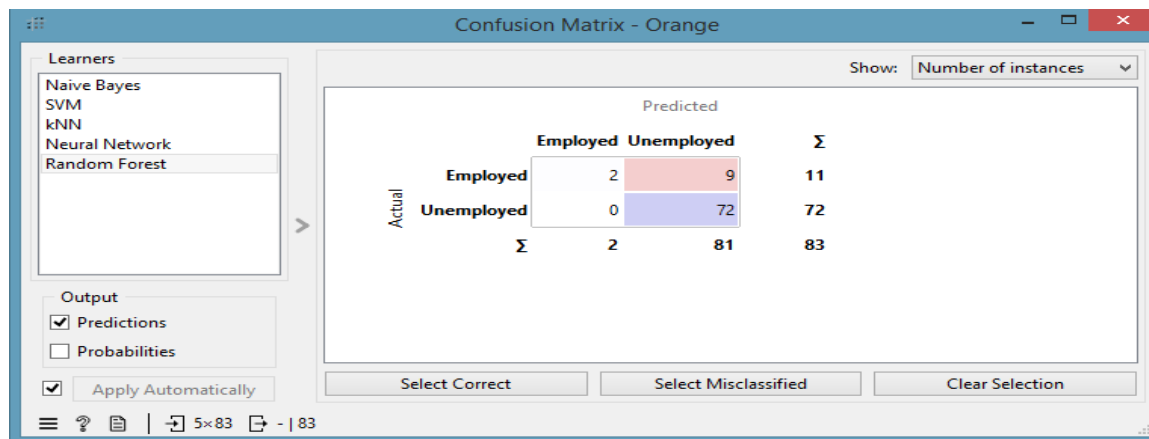
5.7 EVALUATION METRICS:

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

WITHOUT PREPROCESSING



WITH PREPROCESSING



Through these evaluations, we successfully classified users into their Employed or Unemployed

5.8 EXPERIMENT ANALYSIS :

The experiment aimed to classify students based on their **employment status** (Employed/Unemployed) using **supervised machine learning models**. Various models, including **Support Vector Machine (SVM)**, **Random Forest**, **KNN**, **Navie Bayes** and **Neural Networks**, were trained using student-related attributes such as **Degree, internships, technical skills, projects, and career preferences**.

After preprocessing the dataset—which included **handling missing values, normalization of numeric attributes**—the models were evaluated using standard classification metrics like **accuracy, precision, recall, and F1-score**. Among all the models tested, **Random Forest achieved the highest accuracy** and was selected as the best-performing model for classifying students based on their employability.

visualization techniques such as Distributions and confusion matrices were used to validate and interpret the classification results effectively.

Test and score image

CONCLUSION:

In this project, we successfully classified students based on their **employment status** using **Supervised Machine Learning techniques**. Among the various models tested, the **Random Forest** model demonstrated the highest accuracy and outperformed other classifiers. This classification model can help identify employability trends and assist institutions in tailoring training programs based on student profiles and career preferences.

KEY FINDINGS

- The Students Who Are Not Having atleast One Internship Are Almost Unemployed.
- The Students Who Did Less Number Of Projects Are Almost Unemployed.
- The Student Who Got Job Are Searched For The Job On Average 5 – 7 Months.

Our Recommendation For The Students Is : To Get The Job After Graduation They Must Aquire Skills Based On Their Desired Career Path And Do Some Projects Related To Their Skills And Complete Atleast One Internship On Their Career Domain.

PART-B

TITLE: Classifying DNA Sequence Using Promoters Data

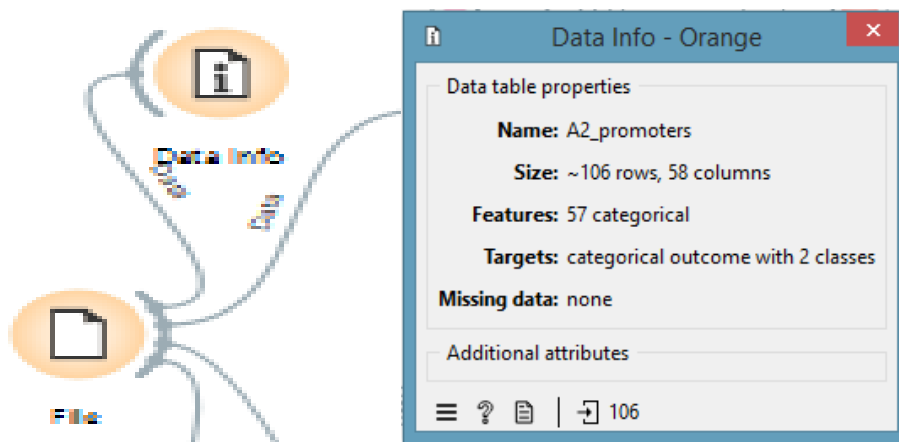
Abstract:

This project leverages machine learning techniques in **Orange Data Mining** to classify DNA sequences into two categories: **Promoters** and **Non-Promoters**. The dataset, A2_promoters.tab, contains labeled DNA sequences along with biological features that serve as predictors for classification.

Methodology:

1. DATA IMPORT AND CLEANING

- **File Widget:** Loads the dataset containing DNA sequence labeled as pp(promoters) or mm(not promoters)
- **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.



Feature Engineering

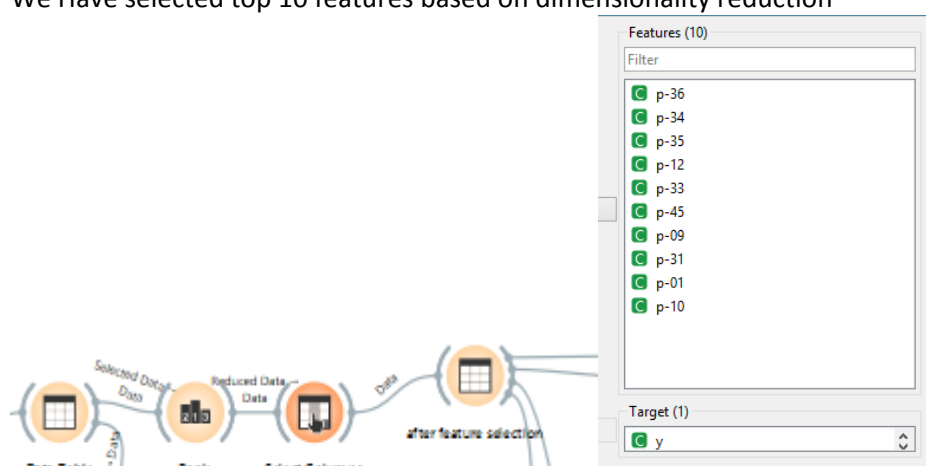
- **Rank Widget:** Selects the most relevant features using statistical ranking techniques, ensuring that only important predictors are used.
- **Feature Table:** Displays the top-ranked features based on their importance.

		#	Info. gain	Gain ratio	Gini
1	p-36	4	0.347	0.198	0.210
2	p-34	4	0.320	0.181	0.188
3	p-35	4	0.283	0.162	0.183
4	p-12	4	0.235	0.122	0.146
5	p-33	4	0.179	0.093	0.108
6	p-45	4	0.148	0.077	0.098
7	p-09	4	0.119	0.062	0.080
8	p-31	4	0.114	0.057	0.076
9	p-01	4	0.109	0.055	0.072
10	p-10	4	0.100	0.054	0.067
11	p-41	4	0.084	0.043	0.056
12	p-13	4	0.079	0.040	0.053
13	p-43	4	0.077	0.039	0.052
14	p-32	4	0.077	0.041	0.052
15	p-20	4	0.070	0.036	0.047

2.DATA PREPROCESSING:

Handling High-Dimensional Data

- The dataset contains **57 features**, making it computationally expensive to run the **Test and Score** widget efficiently.
- To address this, I used the **Preprocess Widget** and applied **Select Relevant Features** to reduce dimensionality.
- Dimensionality reduction methods were compared, selecting the **top 10 most relevant features** based on ranking techniques such as:
 - **Information Gain** (entropy-based dimensionality reduction)
 - **Gini Index** (decision tree-based ranking)
 - **Information Gain Ratio**
- We Have selected top 10 features based on dimensionality reduction



- 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. **Decision Tree**
5. **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.
Decision Tree	Splits data based on feature values for classification	Easy to interpret but prone to overfitting

Each model learns patterns in the training data to distinguish between **pp** and **mm** in DNA sequence

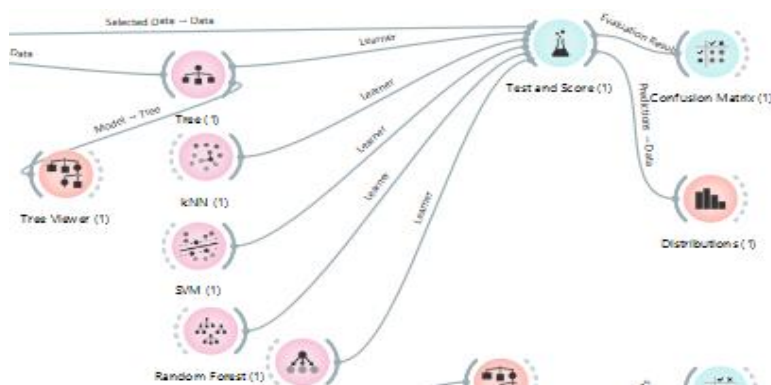
4. MODEL EVALUATION

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 **pp** and **mm** DNA sequence, 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** for DNA classification.



WITHOUT DIMENSIONALITY REDUCTION

Model	AUC	CA	F1	Prec	Recall	MCC
kNN	0.938	0.783	0.774	0.836	0.783	0.616
SVM	0.971	0.896	0.896	0.903	0.896	0.799
Random Forest	0.934	0.840	0.838	0.850	0.840	0.689
Naive Bayes	0.972	0.925	0.925	0.925	0.925	0.850
Tree	0.811	0.802	0.802	0.802	0.802	0.604

WITH DIMENSIONALITY REDUCTION

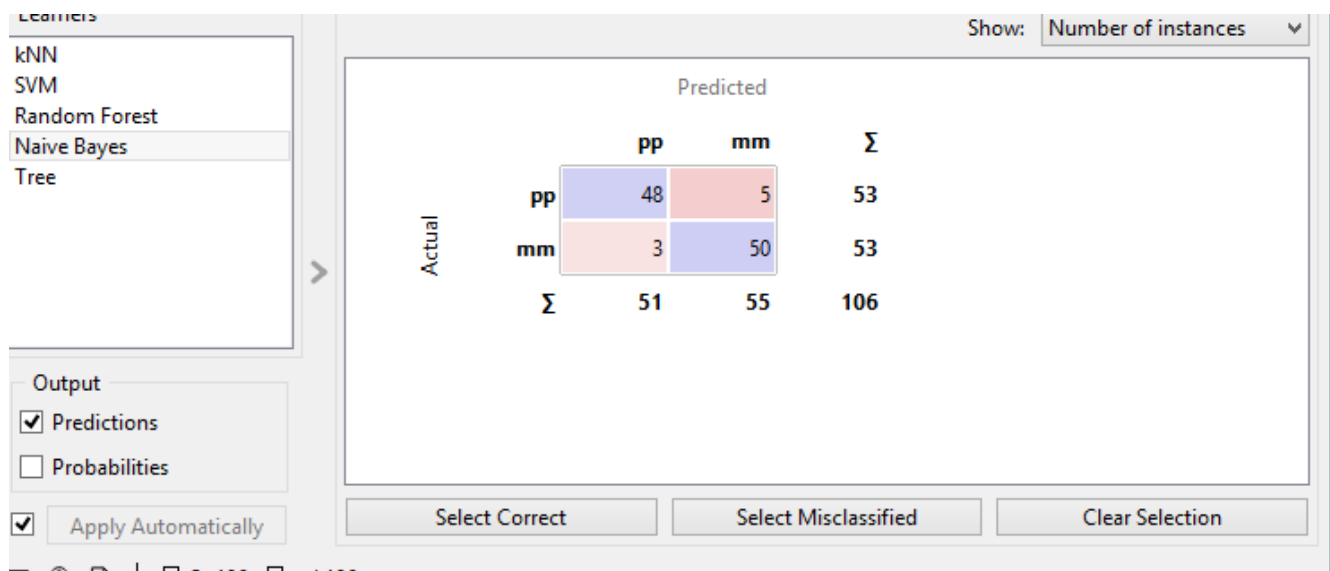
Evaluation results for target (None, show average over classes) ▼						
Model	AUC	CA	F1	Prec	Recall	MCC
kNN (1)	0.956	0.821	0.815	0.868	0.821	0.687
SVM (1)	0.997	0.962	0.962	0.963	0.962	0.925
Tree (1)	0.826	0.821	0.821	0.821	0.821	0.642
Random Forest (1)	0.969	0.887	0.887	0.887	0.887	0.774
Naive Bayes (1)	0.995	0.972	0.972	0.972	0.972	0.944

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

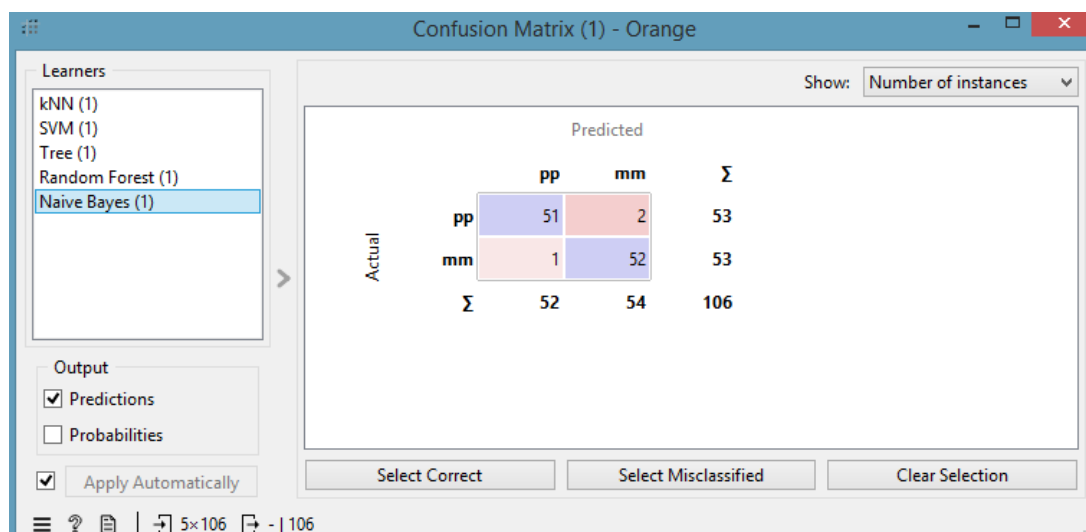
- Navie Bayes demonstrated high accuracy.

❑ **Confusion Matrix Widget:**

WITHOUT_DIMENSIONALITY_REDUCTION



WITH_DIMENSIONALITY_REDUCTION



5. PREDICTIONS & RESULTS

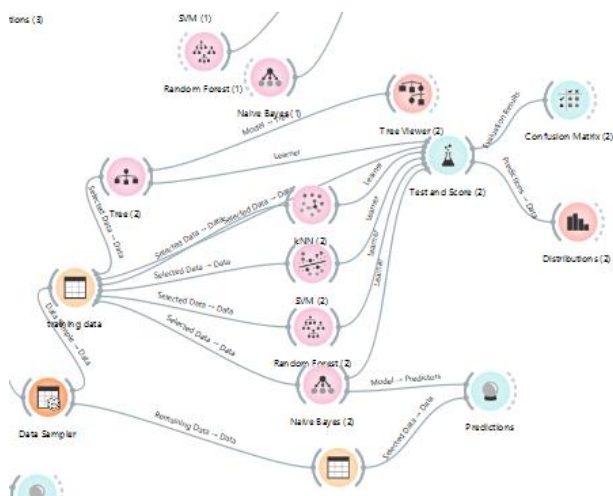
Predictions and Final Model Deployment

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

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 Navie Bayes model along with the remaining data from the **Data Sampler Widget**.
- The Navie Bayes model then classified these new samples as **pp** or **mm** based on learned patterns.

Predictions - Orange									
Show probabilities for		Classes in data		<input checked="" type="checkbox"/> Show classification errors		Restore Original Order			
	Naive Bayes (2)	error	y	p-36	p-34	p-35			
1	0.96 : 0.04 → pp	0.036	pp	t	t	t	t		
2	0.00 : 1.00 → mm	0.002	mm	a	c	a	c		
3	1.00 : 0.00 → pp	0.001	pp	t	g	t	t		
4	0.01 : 0.99 → mm	0.007	mm	a	t	c	g		
5	0.03 : 0.97 → mm	0.032	mm	c	t	c	g		
6	0.00 : 1.00 → mm	0.000	mm	t	c	a	c		
7	0.00 : 0.00 → mm	0.018							
<input checked="" type="checkbox"/> Show performance scores		Target class: (Average over classes)							
Model	AUC	CA	F1	Prec	Recall	MCC			
Naive Bayes (2)	1.000	0.935	0.934	0.942	0.935	0.873			

6. VISUALIZATION

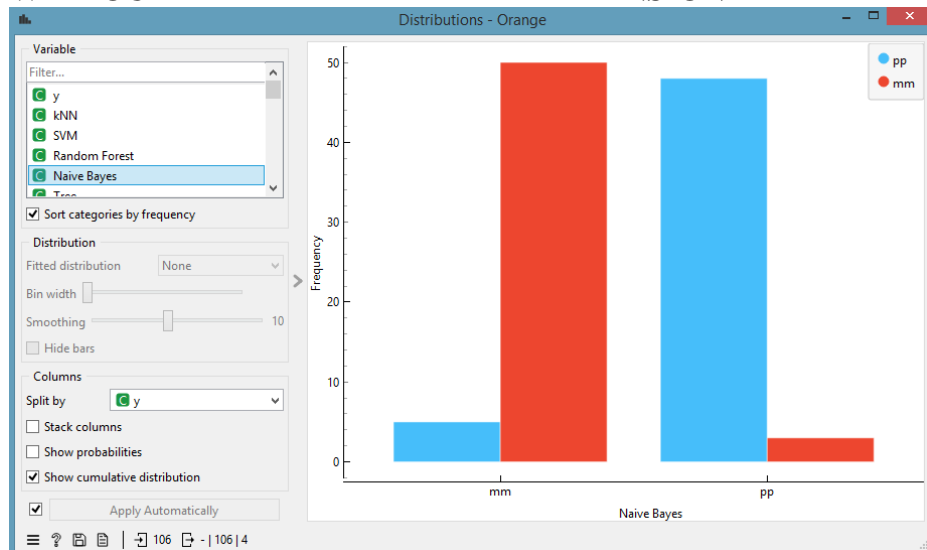
1. Distributions

Visualizes how different DNA sequence samples were classified based on the most relevant features.

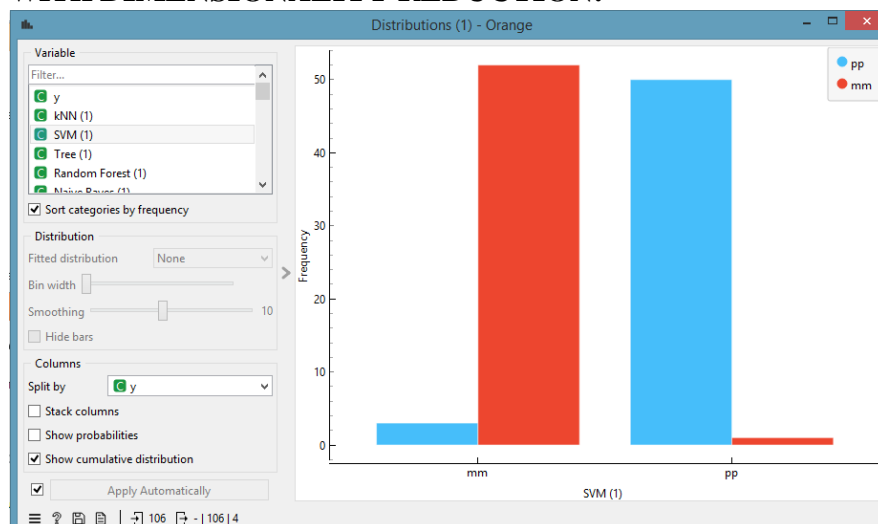
WITHOUT

DIMENSIONALITY

REDUCTION:



WITH DIMENSIONALITY REDUCTION:



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 DIMENSIONALITY REDUCTION

		Predicted		
		pp	mm	Σ
Actual	pp	48	5	53
	mm	3	50	53
Σ		51	55	106

WITH DIMENSIONALITY REDUCTION

		Predicted		
		pp	mm	Σ
Actual	pp	51	2	53
	mm	1	52	53
Σ		52	54	106

FINAL FLOW DIAGRAM USING ORANGE TOOL

Experimental analysis

PART-C EXPERIMENT ANALYSIS

1. Introduction

In data mining and machine learning, dataset selection plays a crucial role in determining the effectiveness of the models applied. This study analyzed two different experimental setups:

- **Part A** used a **real-world dataset** collected from institutional sources and surveys.
- **Part B** used a **generated DNA dataset**.

This section integrates insights from both experiments to provide final conclusions regarding their performance, applicability, and limitations.

2. Key Observations from Experimental Analysis

2.1. Data Characteristics and Preprocessing

- The **generated dataset (Part B)** was well-structured with no missing values. Preprocessing /Dimensionality reduction was straightforward, and the data was ready for model training with little effort.
- The **real-world dataset (Part A)** required extensive preprocessing, such as handling missing values, normalization, and dimensionality reduction due to inconsistencies.
- Although more challenging, Part B reflects real-world scenarios and helps in building more robust and generalized models.

2.2. Model Performance Analysis

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

2.3. Key Findings from Model Comparisons

- **KNN** consistently performed the best across both datasets, especially when hyperparameters were tuned effectively.
- **Navie Bayes** showed more accuracy after Dimensionality reduction (dimensionality reduction).
- **Support Vector Machine** showed improved accuracy after preprocessing, especially in Part B, highlighting the importance of data quality.
- **Random Forest** showed noticeable improvements after addressing missing data and normalization.

3. Preprocessing Differences

Part A: Minimal Preprocessing

- Data was balanced and pre-structured.
- No missing values.
- Feature engineering was already aligned with model requirements.

Part B: Extensive Preprocessing Required

- Missing values were handled using imputation.
- Normalization and feature scaling were applied.
- Redundant attributes were removed.
- Class imbalance issues were addressed using resampling techniques.

Impact:

- Initial performance on Part B was lower due to raw data quality.
- After preprocessing, Dimensionality reduction significant improvements were observed in model accuracy and reliability.

Part A (Student Dataset - Synthetic):

- The **confusion matrix** showed **high true positive and true negative rates** for both *Employed* and *Unemployed* students.
- **UnEmployed** were classified with near-perfect accuracy due to the structured and balanced nature of the synthetic dataset.
- Very few **true positives** or **true negatives** were observed, indicating excellent classification performance.
- The clarity of the synthetic features led to minimal confusion between classes.

Part B (DNA Sequence Dataset - Real-World Data):

- The **initial confusion matrix** showed **imbalanced performance** across classes due to the high dimensionality and variability of raw DNA sequence data.
- **Misclassification** occurred frequently between sequences with similar motifs or patterns, especially in closely related species or gene groups.
- The model initially exhibited a **high number of false negatives**, meaning it failed to detect some true class labels

After Dimensionality Reduction:

- Dimensionality Reduction helped reduce noise and improve model interpretability.
- The **updated confusion matrix** displayed:
 - A **notable increase in true positives** and **true negatives** across categories.
 - A **drop in false positives**, improving model specificity.
 - Some **false negatives** persisted, suggesting the presence of sequence variations or mutations not captured by basic encoding methods.

Impact:

- Confusion matrix results indicated that **Dimensionality Reduction played a vital role** in improving classification performance.
- Despite improved accuracy, the **complex nature of biological sequences** means further enhancement may be achieved through:
 - Advanced models (e.g., BiLSTM, Transformers),
 - Inclusion of biological context (e.g., GC content, codon usage),

- and **data augmentation** techniques for underrepresented classes.

5. Conclusion

This project demonstrated the comparison between a **generated dataset (Part A)** and a **real-world dataset (Part B)** in classifying student career outcomes and DNA sequence.

Key Takeaways:

- Part B achieved high accuracy quickly due to clean and pre-processed data.
- Part A required extensive cleaning and transformation but provided insights closer to real-world applications.
- **Rondom Forest** was the top-performing classifier overall for **student dataset**.

This study successfully applied **machine learning techniques** to student data in order to classify students based on their **employment status** (Employed or Unemployed). Among the various models tested,

The results highlight the potential of **AI and data mining approaches in educational analytics**, enabling institutions to better understand student outcomes, predict employment trends, and make data-driven decisions for academic and career support

- **Navie Bayes** was the top-performing classifier overall for **DNA Sequence dataset**.

This study successfully applied machine learning techniques to DNA sequence data to classify sequences as Promoters or Non-Promoters. Among the tested models, **Naïve Bayes** achieved the highest accuracy, confirming its effectiveness for sequence-based classification tasks. The results support the use of AI tools in **bioinformatics** for enhancing genetic research and understanding regulatory DNA elements.

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)

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 the information 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 for sustainable development.
8. **Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of 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 on level of mapping as follows:

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

