***Local vs. Online LLM Models: Evaluating User Engagement, Performance, and Usability***

#### 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

**Barama Haritha Sai (22481A0523)**

**Alaparthi Yaswanth Kiran (22481A0510)**

**Chandu Mounika (22481A0541)**

**Gosala Kartheek (22481A0565)**

Under the Enviable and Esteemed Guidance of

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

**Professor of CSE**

****

## DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

**SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE**

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

**SESHADRIRAO KNOWLEDGE VILLAGE**

**GUDLAVALLERU – 521356**

**ANDHRA PRADESH**

### 2024-25

**SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE**

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

**DEPARTMENT OF** **COMPUTER SCIENCE AND ENGINEERING**



**CERTIFICATE**

This is to certify that the project report entitled “**Local vs. Online LLM Models:Evaluating User Engagement, Performance, and Usability”** is a bonafide record of work carried out by **Barama Haritha Sai (22481A0523), Alaparthi Yaswanth Kiran(22481A0510), Chandhu.Mounika(22481A0541), Gosala.Kartheek (22481A0565),** under the guidance and supervision of **Dr.M.Babu Rao ,M.Tech,Ph.D, Professor,** Computer Science and Engineering, in the partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering of Jawaharlal Nehru Technological University Kakinada, Kakinada during the academic year 2024-25.

**Project Guide Head of the Department (Dr. M. BABU RAO) (Dr. M. BABU RAO)**

**External Examiner**

## ACKNOWLEDGEMENT

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

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

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

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

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

Team Members

**B.Haritha Sai (22481A0523)**

**A.Yaswanth Kiran(22481A0510)**

**Ch.Mounika (22481A0541)**

**G.Kartheek (22481A0565)**

|  |  |
| --- | --- |
| **INDEX** |  |
| **CONTENTS** | **PAGE NO** |
| **Abstract** | 1 |
| **PART A: KDD PROCESS** | 2-29 |
| **Chapter 1: Introduction** | 2-8 |
| 1.1 Introduction to KDD |  |
| 1.2 Data Warehousing |  |
| 1.3 Data Mining |  |
| **Chapter 2: Data Mining and Warehousing Process on Collected Dataset** | 9-26 |
| 2.1 Problem Statement |  |
| 2.2 Methodology |  |
| **Chapter 3: Experimental Analysis** | 26-29 |
| 3.1 Evaluation |  |
| 3.2 Conclusion |  |
| **PART B: DATA MINING IN DETAIL** | 30-42 |
| **Chapter 1: Introduction on data mining methodology** | 30-31 |
| 1.1 Problem Statement |  |
| 1.2 Identification of appropriate methodology |  |
| **Chapter 2: Analysis on Dataset** | 32-33 |
| **Chapter 3: Working on Dataset** | 34-39 |
| **Chapter 4: Experimental Analysis** | 40-42 |
| **PART C: FINAL ANALYSIS** | 43-44 |
| **Evaluation of Experimental Analysis** | 43 |
| **Conclusion** | 44 |
| **References** | 45 |
| **List of Program Outcomes and Program Specific Outcomes** | 46-47 |
| **Mapping of Program Outcomes with graduated POs and PSOs** | 48 |

## 

## ABSTRACT

With the rapid advancement of Artificial Intelligence, Large Language Models (LLMs) have gained significant traction among professionals and students alike. This study explores the usage patterns of LLMs, focusing on online and locally deployed models. The dataset comprises survey responses that provide insights into the preferred LLM tools, frequency of usage, and motivations for using AI-powered applications. By analysing these responses, we aim to understand how individuals interact with AI tools in their daily activities and professional environments.

The study highlights the diverse range of LLMs utilized, including ChatGPT, Gemini, LLaMA, DeepSeek, and Claude. Additionally, a segment of users prefers deploying LLMs locally, citing reasons such as privacy concerns and faster response times. The dataset also captures the variations in model sizes used locally, ranging from 7B to 65B+ parameters, reflecting different computational capabilities and application needs. Understanding these trends helps identify key factors influencing AI adoption and the specific preferences of users across various professions.

Furthermore, the findings reveal variations in prompt lengths, with respondents using short, medium, and long prompts based on their requirements. This indicates a growing awareness and adaptation of AI-powered writing assistants to enhance productivity. By analysing this dataset, organizations and AI developers can gain valuable insights into user preferences, leading to the development of more tailored AI models and interfaces that align with real-world use cases.

This research serves as a foundation for future studies on the effectiveness and challenges of AI-assisted workflows. As AI tools continue to evolve, tracking user behaviour and preferences will be crucial in optimizing LLM performance and improving accessibility across different user groups

## 

**PART A:** **Local vs. Online LLM Models: Evaluating User Engagement, Performance, and Usability Using KDD PROCESS**

## CHAPTER 1: INTRODUCTION

* 1. **INTRODUCTION TO KDD**

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

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

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

**A diagram of data processing

AI-generated content may be incorrect.**

Fig:1

**Data Selection**

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

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

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

**Data Cleaning**

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

* **Missing Values:**Gaps in data are filled with the mean or most probable value to maintain dataset completeness.
* **Noisy Data:**Noise is reduced using techniques like binning, regression, or clustering to smooth or group the data.
* **Removing Duplicates:**Duplicate records are removed to maintain consistency and avoid errors in analysis.

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

**Data Transformation and Reduction**

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

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

Data Reduction helps simplify the dataset while preserving key information.

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

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

**Data Mining**

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

Key characteristics of data mining patterns include:

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

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

**Evaluation and Interpretation of Results**

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

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

### 1.2 DATA WAREHOSUING

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

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

### A diagram of data storage AI-generated content may be incorrect.

### Fig:2

### Need for Data Warehousing

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

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

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

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

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

### Components of Data Warehouse

### The main components of a data warehouse include:

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

**1.3 DATA MINING**

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

One of the key breakthroughs in data mining is its ability to handle and analyze big data efficiently. With the increasing volume, velocity, and variety of data, traditional methods are often insufficient. Data mining techniques like clustering, classification, regression, and association rule learning are essential for extracting valuable insights from complex datasets quickly and accurately.

Data mining is closely related to machine learning and data analytics. While data mining focuses on discovering new patterns within large datasets, machine learning involves developing algorithms thatcan learn from and make predictions on data. These fields complement each other, enhancing data analysis and predictive modeling capabilities.

Diagram of a diagram of data

AI-generated content may be incorrect.

Fig:3 Data Mining Block Diagram

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

### Supervised Learning

Supervised learning is a type of machine learning where the model is trained on a labeled dataset, meaning each training example is paired with an output label. The model learns to map inputs to outputs, enabling it to predict labels for new, unseen data accurately. Common algorithms include K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees, and Logistic Regression.This approach is widely used for tasks like classification and regression.

There are two categories of Supervised Learning

* Classification
* Regression

#### Classification

Classification is a type of supervised machine learning technique used to categorize data into predefined labels or groups based on input features. It involves training a model on a labeled dataset, where each data point is associated with a specific category. The model learns patterns and relationships within the data and then applies this knowledge to classify new, unseen data. Classification can be binary (e.g., spam vs. not spam) or multiclass (e.g., classifying OTT users as Casual Viewers, Regular Viewers, or Binge Watchers).

#### Regression

Regression in supervised learning is a pivotal technique for predicting continuous output values based on input features. It encompasses a wide range of algorithms aimed at understanding and modeling the relationship between inputs and continuous outputs using labeled training data. Common regression algorithms, such as Linear Regression, Additionally, regression techniques play a crucial role in tasks such as forecasting, optimization, and trend analysis across diverse domains.

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Description** | **Type** |
| Random Forest | Extension of linear regression that’s used for classification tasks. The output variable is 2 binary either yes or no | Classification rather regression |
| SVM | It finds the optimal hyperplane that best separates data points of different classes in high-dimensional space. | 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 |

### 

### 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

**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.

**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 Data Mining Algorithm**

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

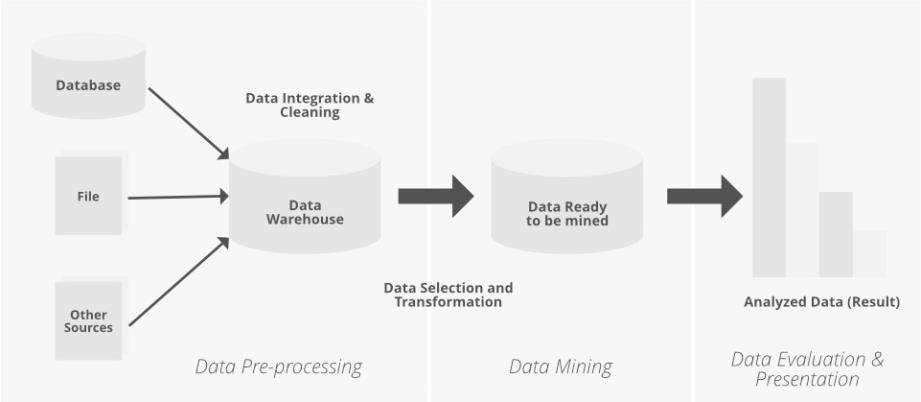


Fig:4 Data Mining Basic Diagram

#### Challenges and Limitations of Data Mining

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

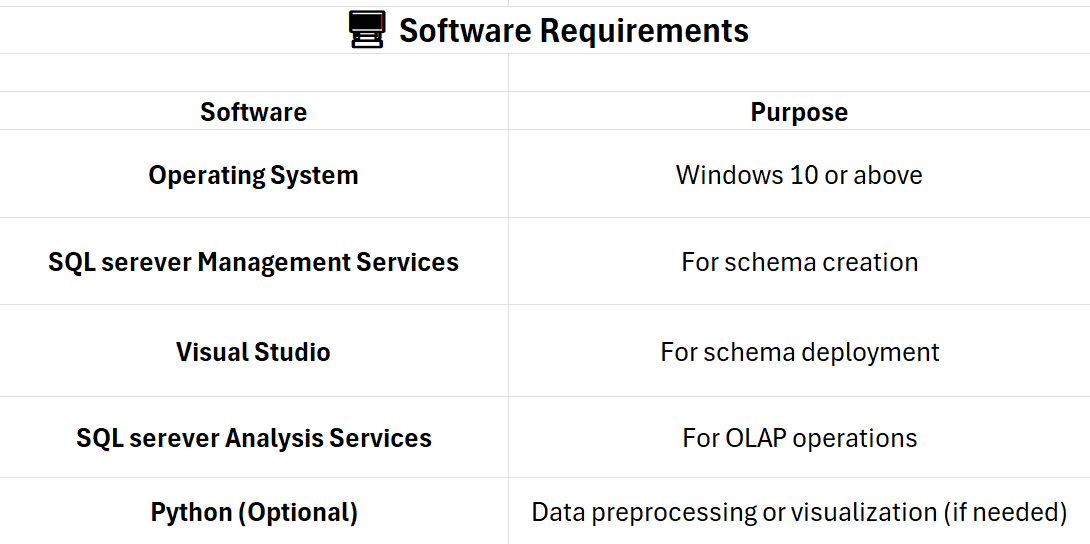
#### Applications of Data Mining

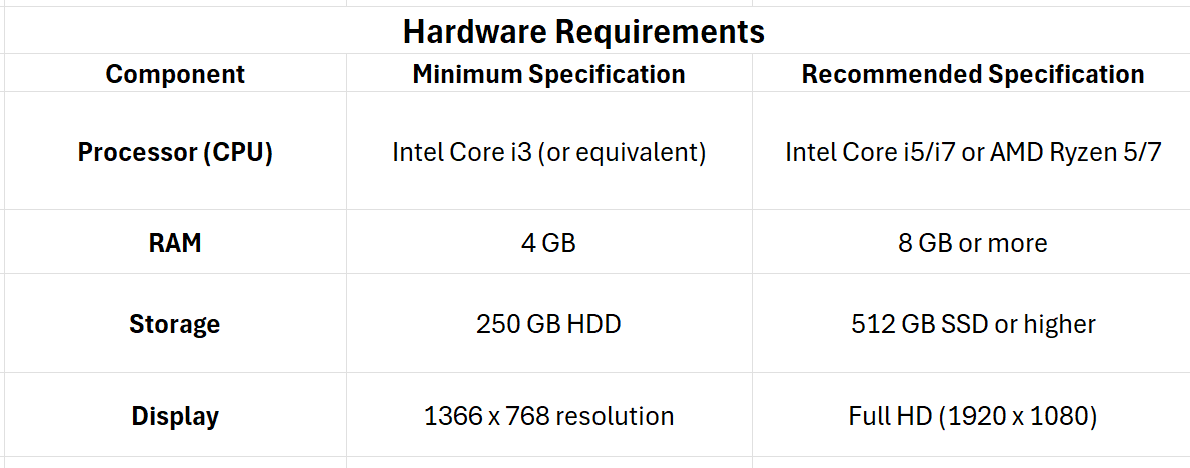
1. **Customer Relationship Management (CRM):** Data mining is a cornerstone in Customer Relationship Management (CRM), enabling businesses to delve deep into customer data for actionable insights. By scrutinizing diverse aspects such as demographics, purchase history, and behavioral patterns, companies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

### Data Mining vs Data Warehousing

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

**REQUIREMENTS FOR THIS PROJECT:**



****

****

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

**2.1 Problem Statement:**

The dataset focuses on understanding user engagement with Large Language Models (LLMs), both online and locally. It captures key aspects such as the preferred AI tools, frequency of use, and reasons for choosing locally hosted models over cloud-based alternatives. Factors like privacy concerns, response speed, and model efficiency play a crucial role in determining user preferences. Additionally, the dataset explores variations in model sizes and prompt lengths, shedding light on how different users interact with AI for various tasks. By analyzing these patterns, the study aims to provide insights into AI adoption trends across different professional backgrounds, helping developers and organizations optimize AI tools for better usability, accessibility, and performance.

**2.2 Methodology:**

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

📝 Step 1: Collect Dataset

│

▼

🛠 Step 2: Preprocess Data

│

▼

Step 3: Construct OLAP Schemas

│

▼

🔍 Step 4: Visualize Schemas

│

▼

Step 5: Deploy & Create OLAP Queries

│

▼

📊 Step 6: Perform OLAP Operations

│

▼

🎨 Step 7: Visualize OLAP Results

│

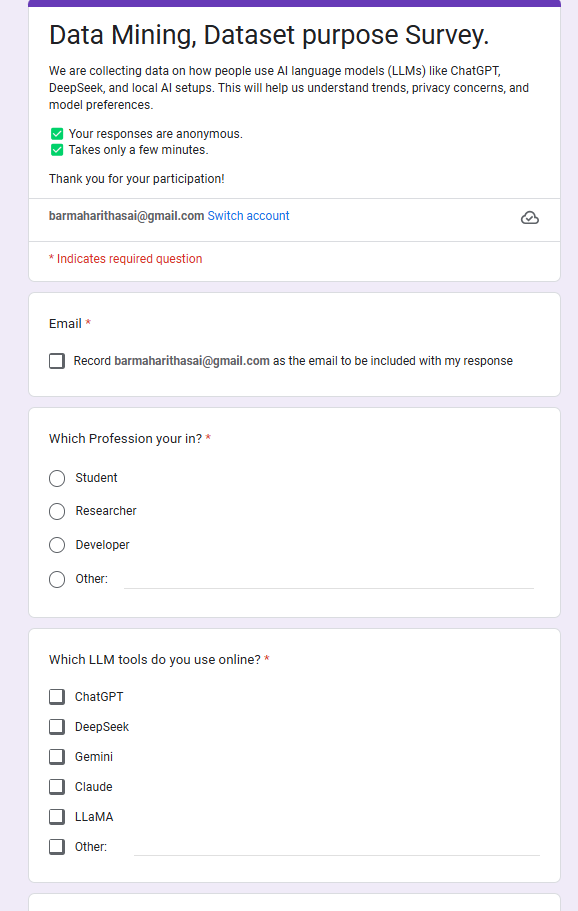
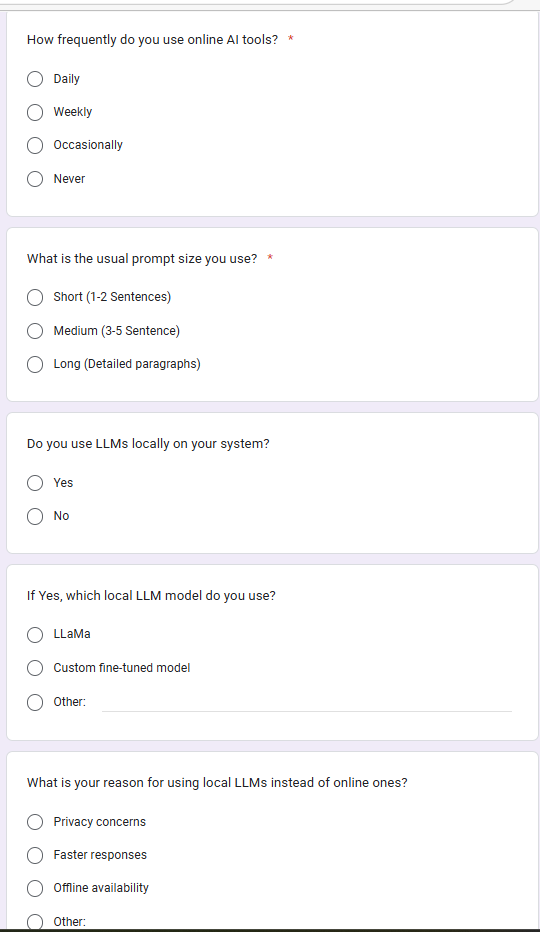
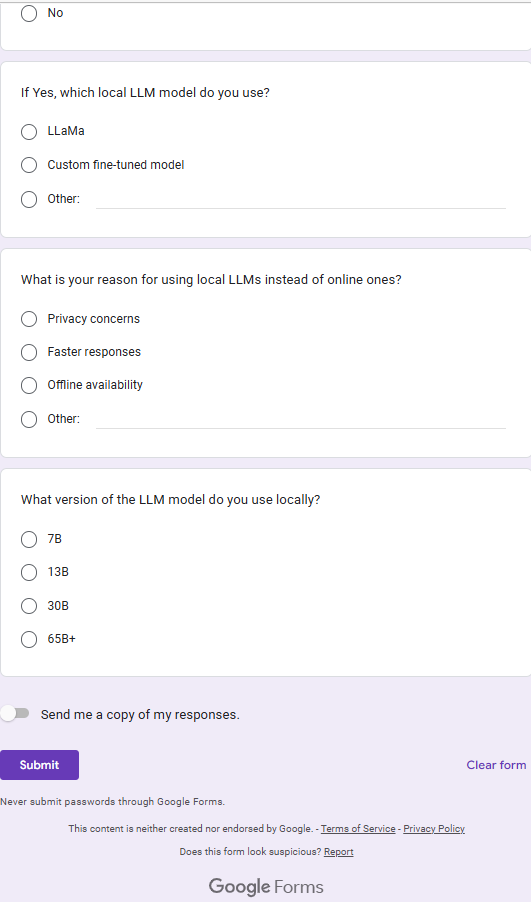
▼

Step 8: Perform Data Mining

## Step 1: Dataset Collection & Exploration

* Dataset is collected using google form to analyze whether LLMs are using online or offline.

The sample form is

* Shared it with participants through email, social media, or targeted groups.

Link to the Google form: https://forms.gle/wixCyu2w9EcGu1he6

* Collected Responses – Monitored responses and ensure enough data is gathered.
* Exported Data – Downloaded the responses as a CSV file for further processing.

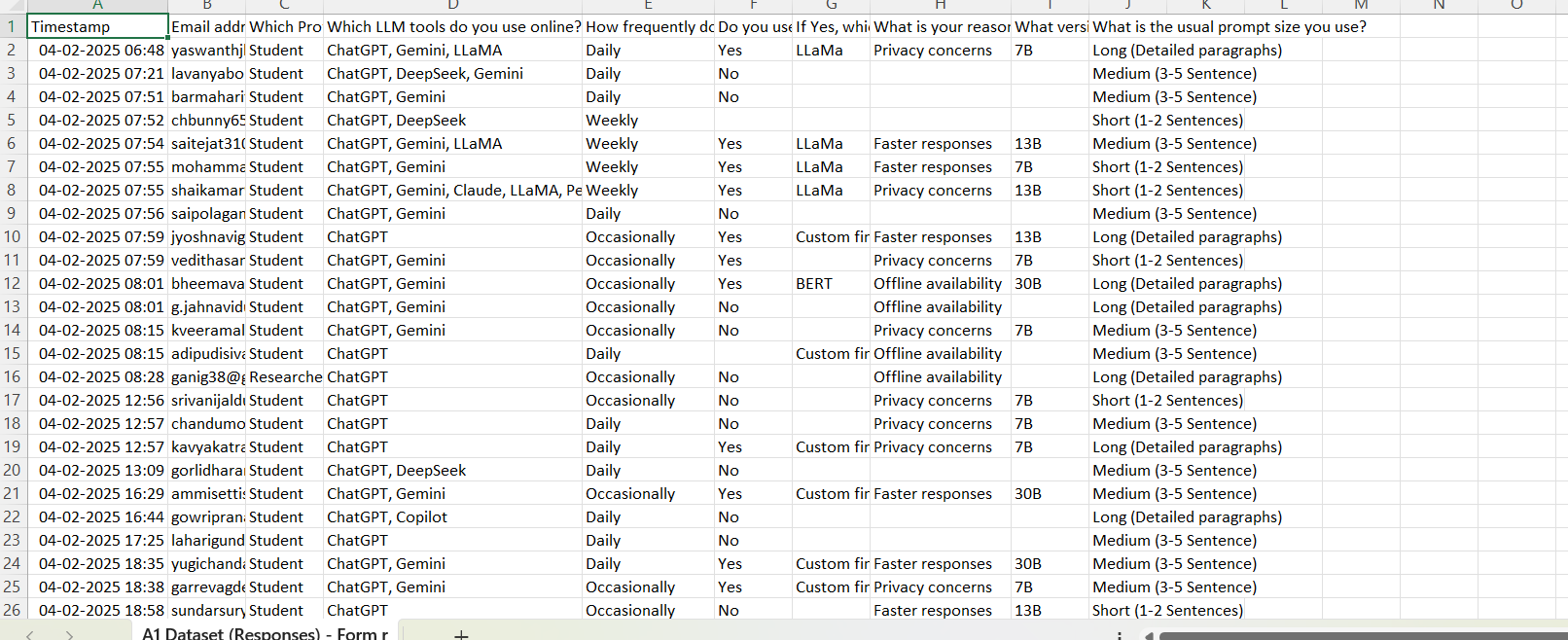
****

Fig:5

List of attributes in my google form:

* Which Profession you are in?
* Which LLM tools do you use online?
* How frequently do you use online AI tools?
* What is the usual prompt size you use?
* Do you use LLMs locally on your system?
* If Yes, which local LLM model do you use?
* What is your reason for using local LLMs instead of online ones?
* What version of the LLM model do you use locally?

**Step 2:** Preprocess the Dataset Using Orange Tool

(Select the best preprocessing technique using test and score)

* Load the Dataset – Import the collected CSV file into Orange.

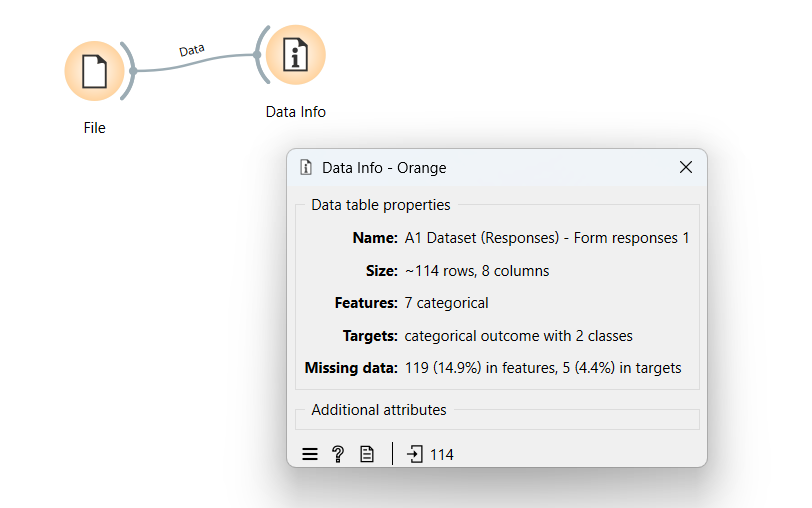


Fig:6

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

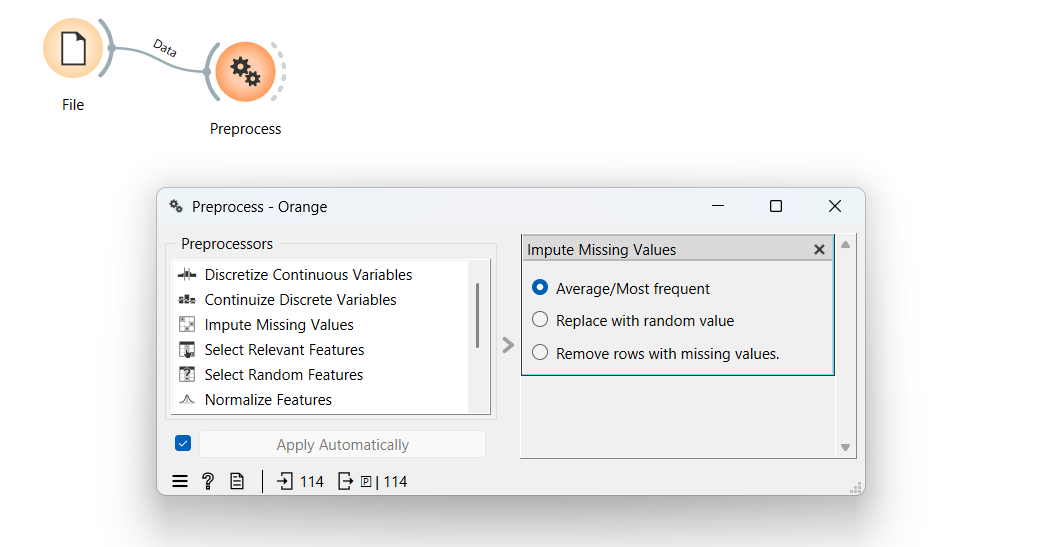


Fig:7

* Applying continuous discrete variables preprocessing technique where we used treat as ordinal attribute to change the categorical values into numerical values.

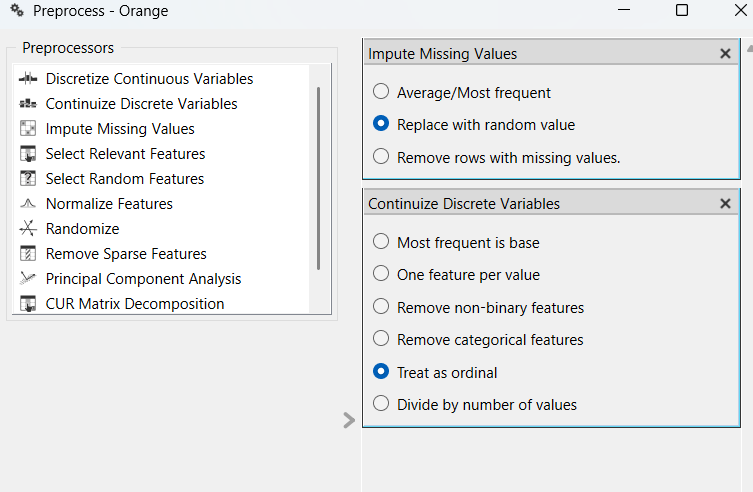


Fig:8

* Data table before preprocessing:

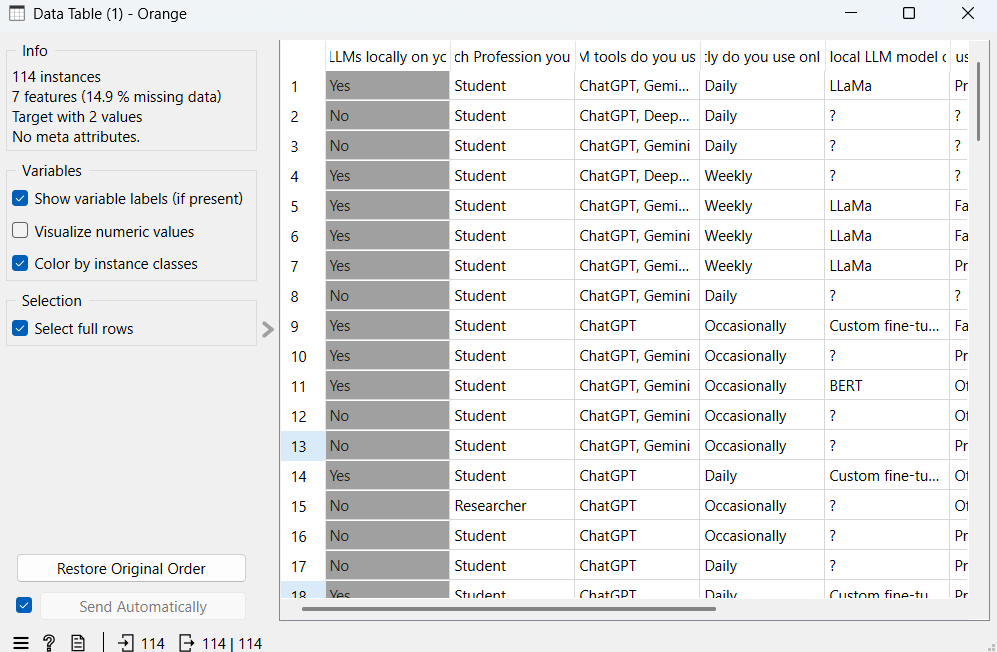


Fig:9

* Data table after preprocessing:

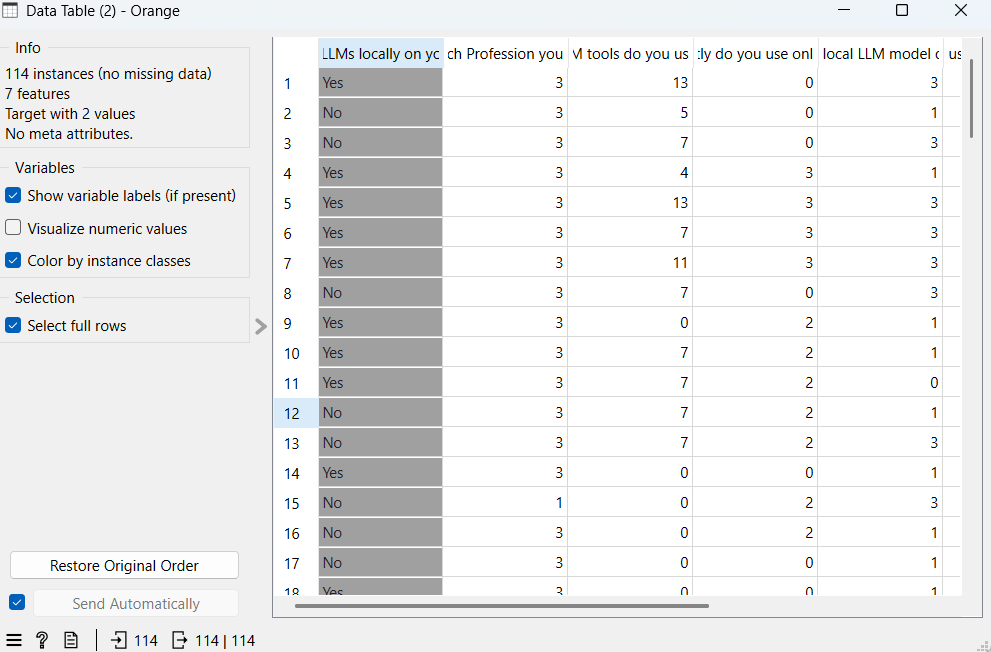
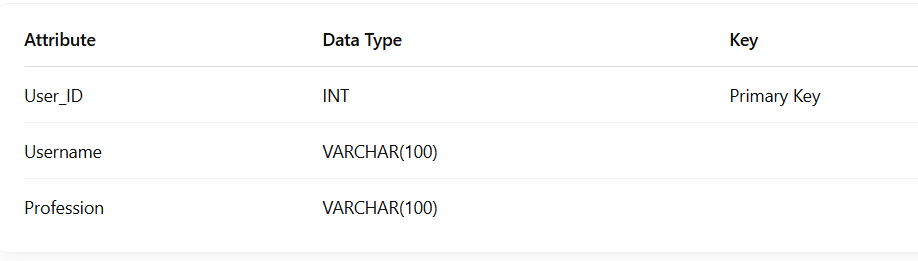


Fig:10

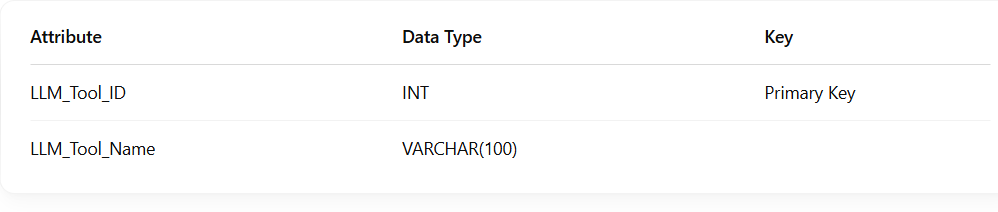
**Step 3**: Schema Constructionby Normalizing the Dataset (Using Database Engine)

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

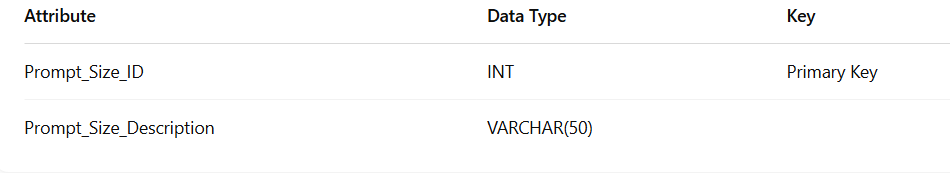
1. **Dim\_User**

****

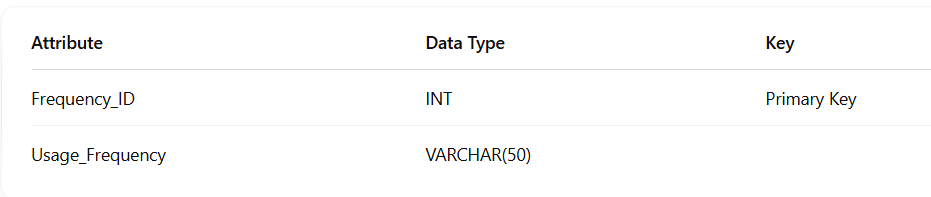
1. **Dim\_LLM\_Tools**

****

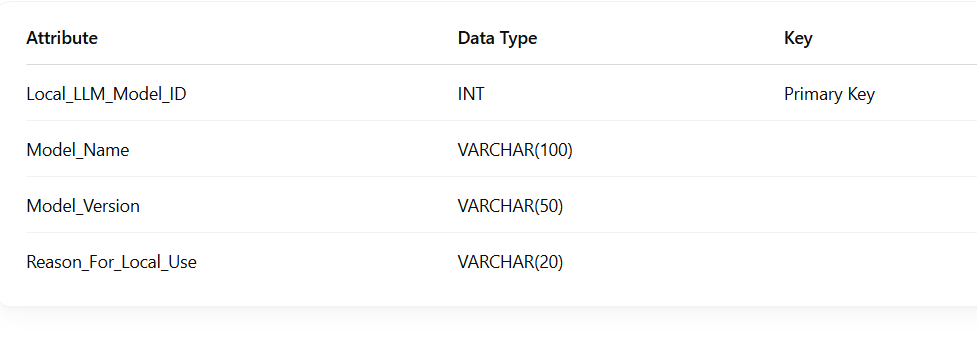
1. **Dim\_Prompt\_Size**

****

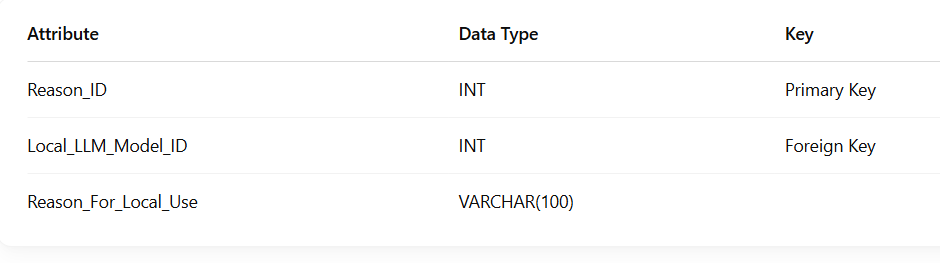
1. **Dim\_Frequency**



1. **Dim\_Local\_LLM\_Models**

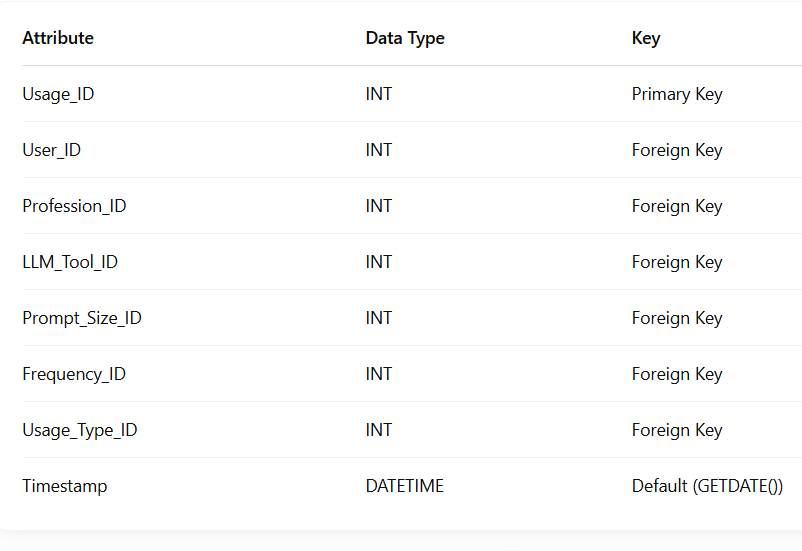


**6. Dim\_Model\_Reason**



* **Fact Tables:**

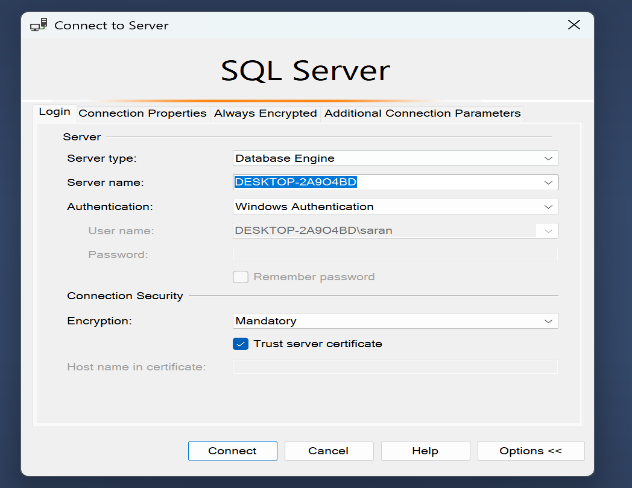
1. **Fact\_Usage**

****

1. **Fact\_Model\_Performance**

****

* Now Create a database “LLMDB” to insert all these tables.
* Generate SQL queries to create and insert data into all the tables in SQL Server Database Engine (SSMS).



## Fig:11 Database Engine

## 

## Fig:12 SQL queries inserted for schema creation

**Step 4:** Schema Visualization in Visual Studio

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

These are the schemas

**Star Schema:**

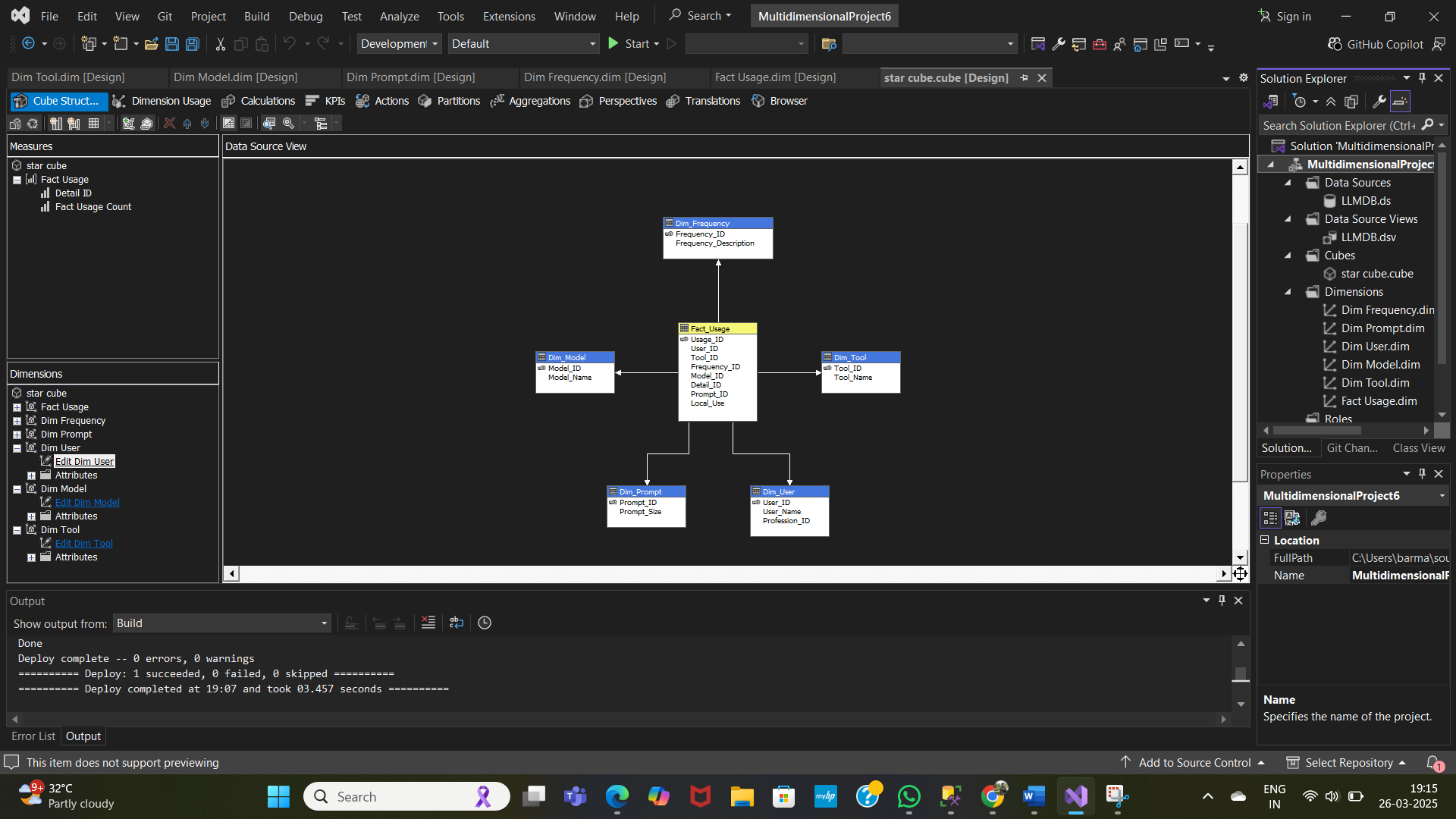


Fig:13

Star schema has one Fact table and five Dimension tables.

* Fact table is Fact\_Usage
* Dimension tables are Dim\_Model, Dim\_Frequency, Dim\_Tool, Dim\_User, Dim\_Prompt

**Snowflake Schema:**

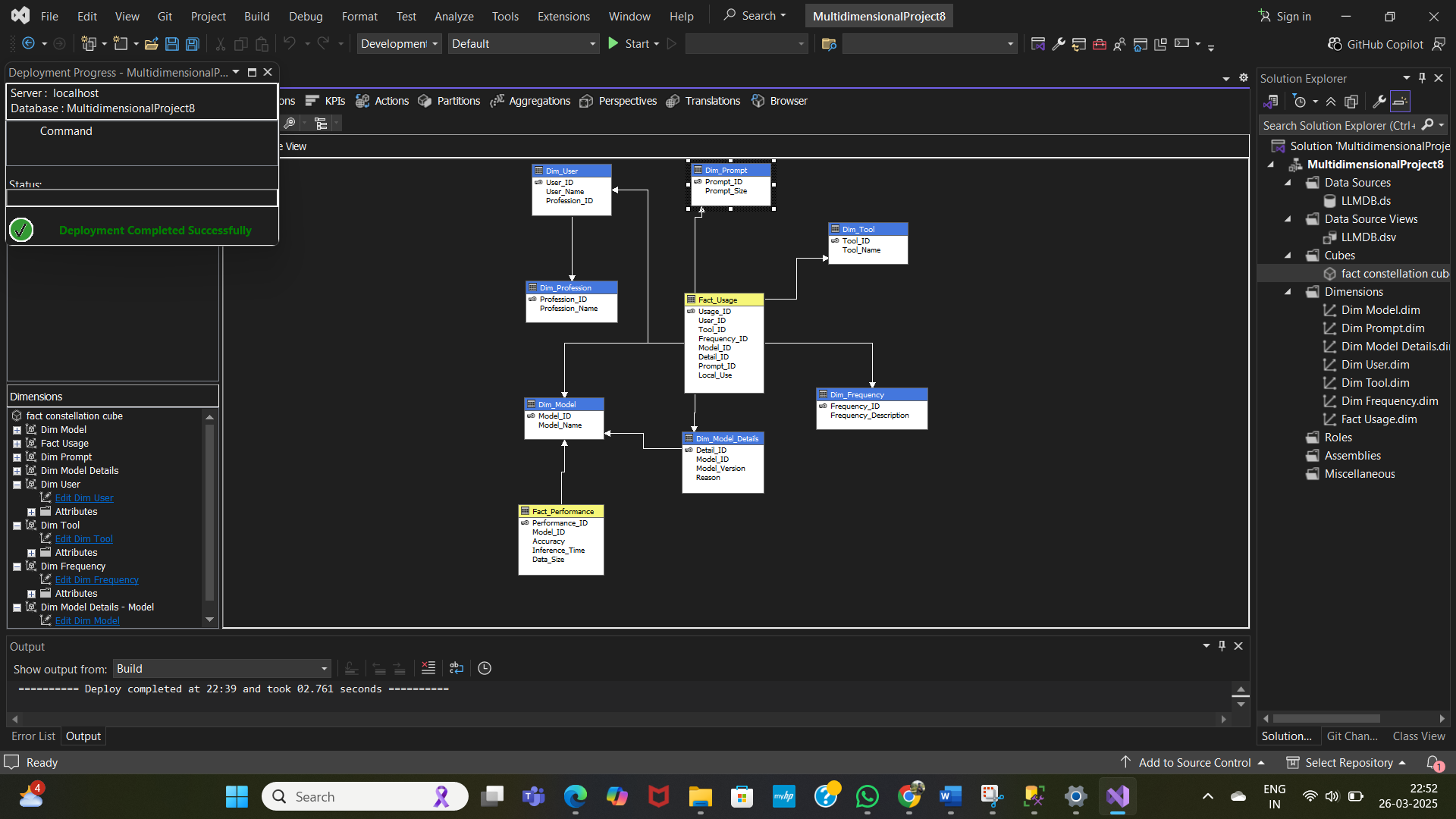
****

Fig:14

Snowflake schema has one Fact table, six Dimension tables and one Sub-Dimension tables.

* Fact table is Fact\_Usage
* Dimension tables are Dim\_Model, Dim\_Frequency, Dim\_Tool, Dim\_User, Dim\_Prompt,Dim\_Model,Dim\_Profession

**Fact Constellation:**

****

**Fig:15**

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

* Fact table is Fact\_Usage,Fact\_Performance
* Dimension tables are Dim\_Model, Dim\_Frequency, Dim\_Tool, Dim\_User, Dim\_Prompt,Dim\_Model,Dim\_Profession

**Step 5:** Build ,Deploy and process all the multi dimensional cubes in VS code and visualize them in SSAS server. Perform various olap operations.

* Write MDX Queries for Drill-Down, Roll-Up, Slice, Dice operations in SSAS.

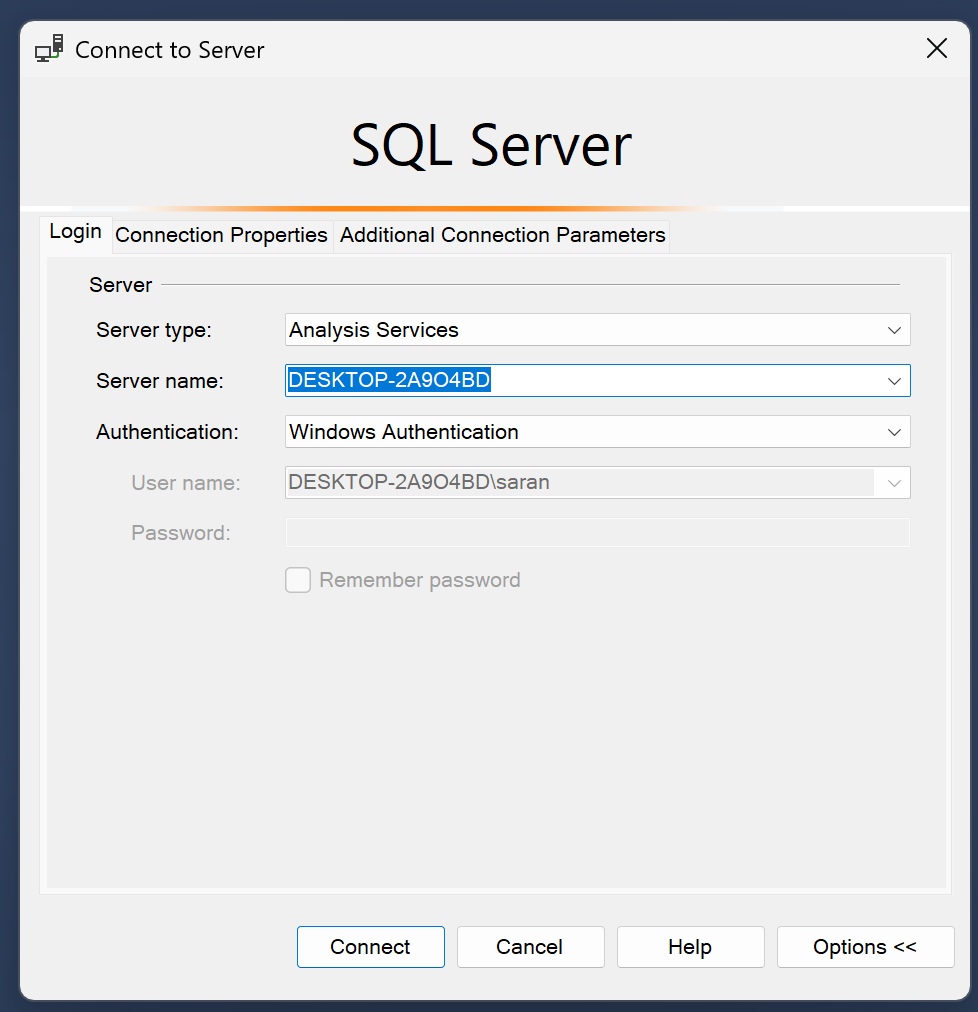


Fig:16 Analysis Services

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

## Generated image Generated imageGenerated image Generated image

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

## MDX Queries OLAP operations in star Schema:

## 

## Fig:17

**Roll-up**:

**(A)** How can we analyze the usage frequency of different AI models by retrieving their individual usage counts ?

**Query:**

## SELECT

## {[Measures].[Fact Usage Count]} ON COLUMNS,

## NON EMPTY

## CROSSJOIN(

## [Dim Model].[Model Name].Members,

## CROSSJOIN(

## [Dim Prompt].[Prompt Size].Members,

## CROSSJOIN(

## [Dim Tool].[Tool Name].Members,

## [Dim Frequency].[Frequency Description].Members

## )

## )

## ) ON ROWS

## FROM [star cube]

## Output:

## 

## Execution Time: 0.01 s

**Sample Visualization:**

* Since this OLAP output has most of the Null values we filled the missing values manually for better visualization. Visualization is done using the orange tool.
* First the OLAP results are used to create a excel sheet and then this sheet is loaded in the orange tool to perform visualization.

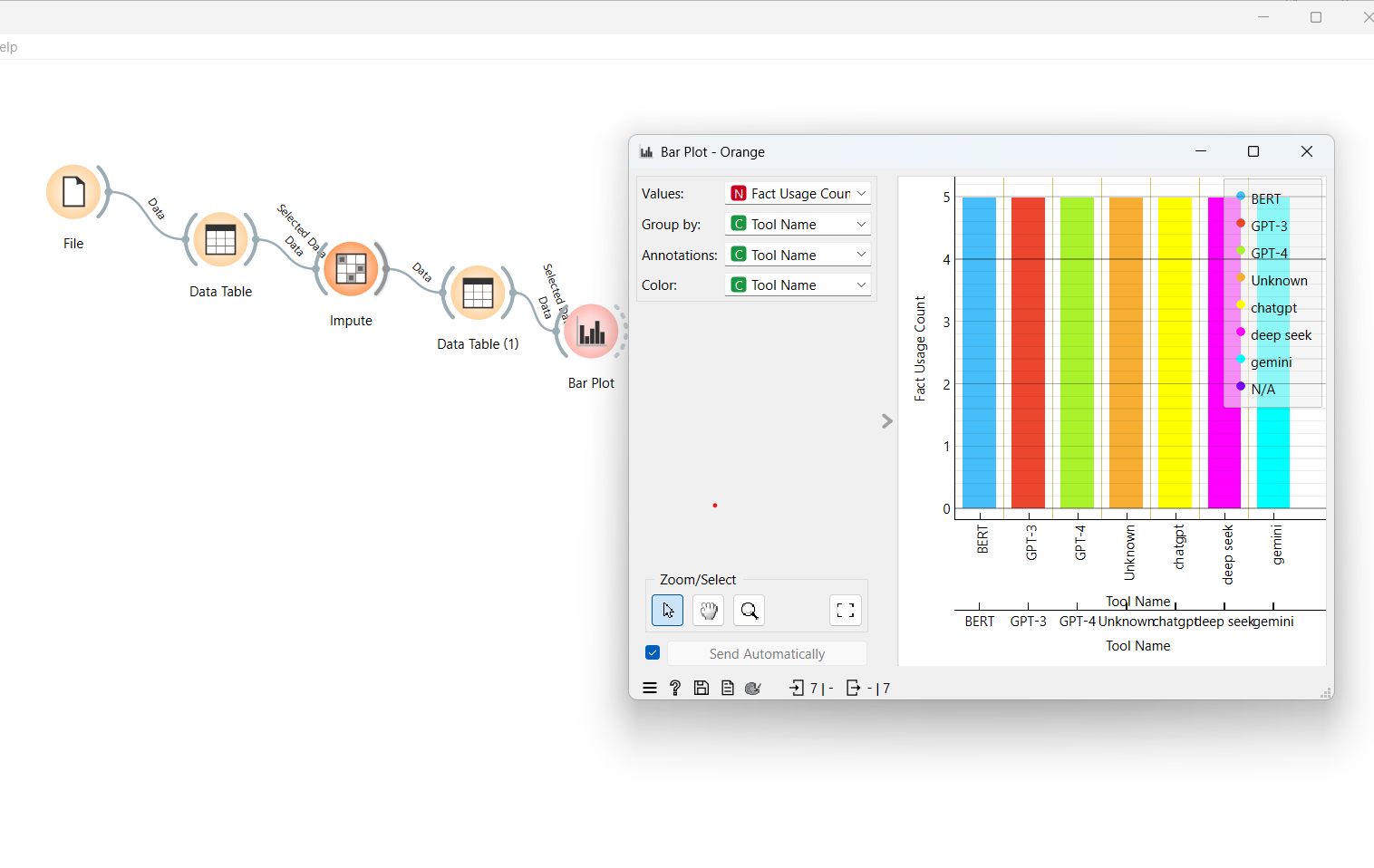


Fig:18

**Drill-Down:**

**(B)** How can we analyze the total Fact usage count by **User\_Name** and drill down further to see which **tools** each user is using and their **frequency of usage**?

## Query:

SELECT

{[Measures].[Fact Usage Count]} ON COLUMNS,

NON EMPTY (

[Dim User].[User Name].[User Name].MEMBERS \*

[Dim Tool].[Tool Name].[Tool Name].MEMBERS \*

[Dim Frequency].[Frequency Description].[Frequency Description].MEMBERS

) ON ROWS

FROM [star cube]

## Output:

## 

## Execution Time: 0.01 s

## Slice: Slice for Tool Name = “Keras”

**(C)** It filters the OLAP cube by selecting only the Keras tool, providing a focused view.

## Query:

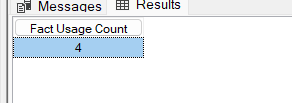
SELECT

{[Measures].[Fact Usage Count]} ON COLUMNS

FROM [star cube]

WHERE ([Dim Tool].[Tool Name].[Keras])

**Output:**



**Execution Time:** 0.01 ms

**(D)** How can we filter the cube to create a **sub-cube** that includes only users and tools where the **Frequency\_Description** is set to **Weekly**?

**Dice:**

**Query:**

SELECT

{[Measures].[Fact Usage Count]} ON COLUMNS,

NON EMPTY (

[Dim User].[User Name].[User Name].MEMBERS \*

[Dim Tool].[Tool Name].[Tool Name].MEMBERS

) ON ROWS

FROM [star cube]

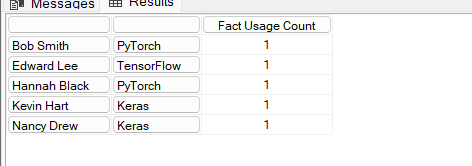
WHERE

(

[Dim Frequency].[Frequency Description].&[Weekly]

)

**Output:**



**Execution Time:** 0.01s

## MDX Queries for OLAP operations in Snow Flake Schema:

## 

## Fig:19

## Roll-up:

**(A)** Summarizing the Usage Count by frequency.

## Query:

SELECT

{[Measures].[Fact Usage Count]} ON COLUMNS,

NON EMPTY

CROSSJOIN(

[Dim Frequency].[Frequency Description].Members,

CROSSJOIN(

[Dim Tool].[Tool Name].Members,

CROSSJOIN(

[Dim Prompt].[Prompt Size].Members,

CROSSJOIN(

[Dim User].[User Name].Members,

[Dim User].[Profession Name].Members

)

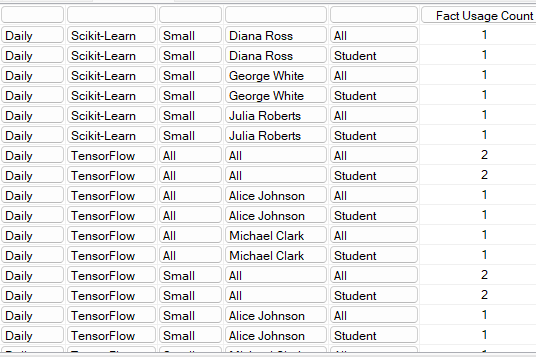
)

)

) ON ROWS

FROM [snowflake cube]

**Output:**



**Execution Time:** 3 ms

**Drill Down:**

**(B)** Viewing local usage for each user within their respective professions.

**Query:**

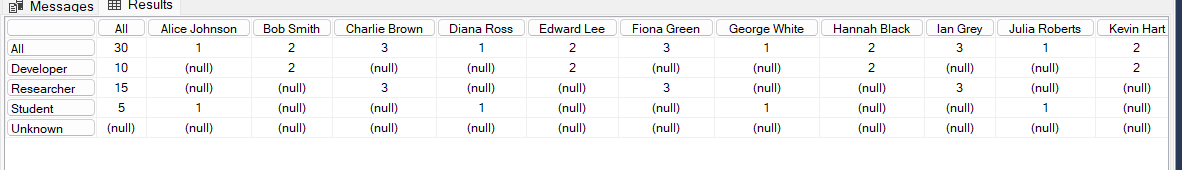
SELECT

{[Dim User].[Profession Name].Members} ON ROWS,

{[Dim User].[User Name].Members} ON COLUMNS

FROM [snowflake cube]

**Output:**



**Execution Time:** 5 ms

**(C)** Viewing the local usage for a specific tool (e.g., PowerBI).

**Slice:** Slice for tool name=”Keras”

**Query:**

SELECT

{[Measures].[Fact Usage Count]} ON COLUMNS,

{[Dim User].[Profession Name].Members} ON ROWS

FROM [snowflake cube]

WHERE ([Dim Tool].[Tool Name].[Keras])

**Output:**

## 

**Execution Time:** 0.01 ms

**(D)** A table showing the Fact Usage Count (some form of event usage) for each Profession Name, but only for monthly frequency**.**

**Dice:**

## Query:

SELECT

{[Dim User].[Profession Name].Members} ON ROWS,

{[Measures].[Fact Usage Count]} ON COLUMNS

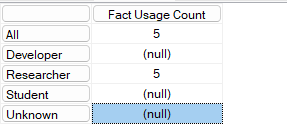
FROM [snowflake cube]

WHERE (

[Dim Frequency].[Frequency Description].[Monthly]

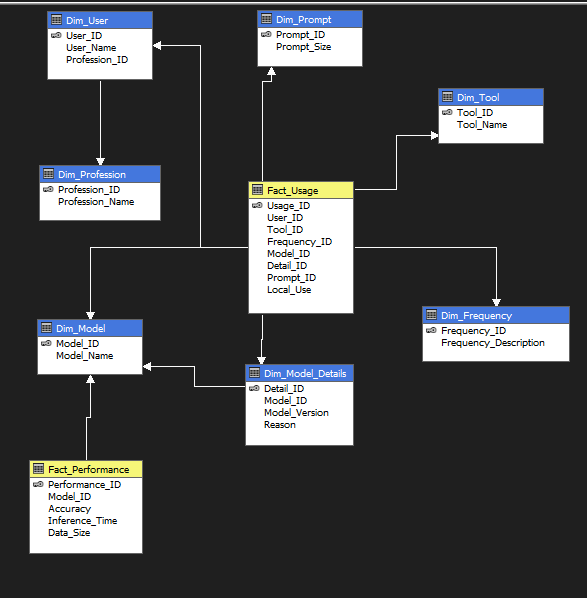
)

**Output:**



**Execution Time:** 0.01 ms

**MDX Queries for OLAP operations in Fact Constellation:**

****

**Fig:20**

**(A**) Summarize the Accuracy by Tool name and then roll up to the tool level.

## Roll-up:

## Query:

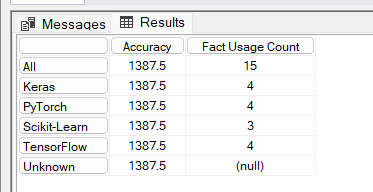
SELECT

{[Dim Tool].[Tool Name].Members} ON ROWS,

{[Measures].[Accuracy]} ON COLUMNS

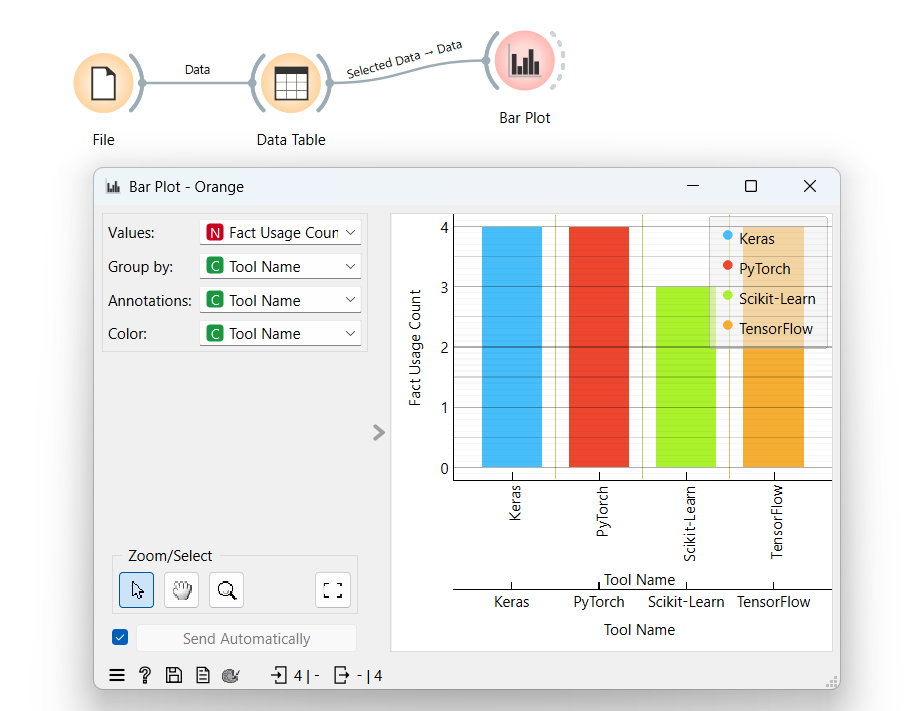
FROM [fact constellation cube]

**Output:**



**Execution Time:** 0.01 ms

**Sample Visualization:**

****

**Fig:21**

**Drill Down:**

**(B)** A table displaying models and their child entities (versions, sub-models, etc.) along with Fact Usage Count, Accuracy, and Data Size, allowing for hierarchical drill-down exploration

**Query:**

SELECT

{[Dim Model].[Model Name].children} ON ROWS,

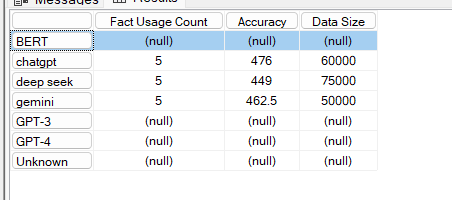
{[Measures].[Fact Usage Count],

[Measures].[Accuracy],

[Measures].[data size]} ON COLUMNS

FROM [fact constellation cube]

**Output:**



**Execution Time:** 0.02 ms

**(C)** View the fact usage count of a specific model (chatgpt) from the Fact\_Performance table.

**Slice:** Slice for model=”chatgpt”

**Query:**

SELECT

{[Measures].[Fact Usage Count],[Measures].[Accuracy],

[Measures].[data size]} ON COLUMNS

FROM [fact constellation cube]

WHERE ([Dim Model].[Model Name].[chatgpt])

**Output:**

## 

**Execution Time:** 0.03 ms

**(D)** A table showing professions along with their weekly Fact Usage Count and Fact Performance Count.

**Dice:**

**Query:**

SELECT

{[Dim User].[Profession Name].Members} ON ROWS,

{[Measures].[Fact Usage Count],[Measures].[fact performance count]

} ON COLUMNS

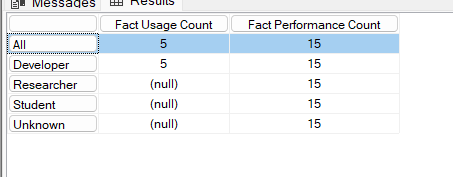
FROM [fact constellation cube]

WHERE (

[Dim Frequency].[Frequency Description].[Weekly]

)

**Output:**



**Execution Time:** 4 ms

## Step 7: Perform Data Mining on the Dataset

One can perform any data mining technique like Classification, Regression, Clustering and Association rule Mining. According to my data set we choose to perform Classification on data set we want to Classify whether the person is using LLM models locally or online. We mentioned this also in our problem Statement.

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

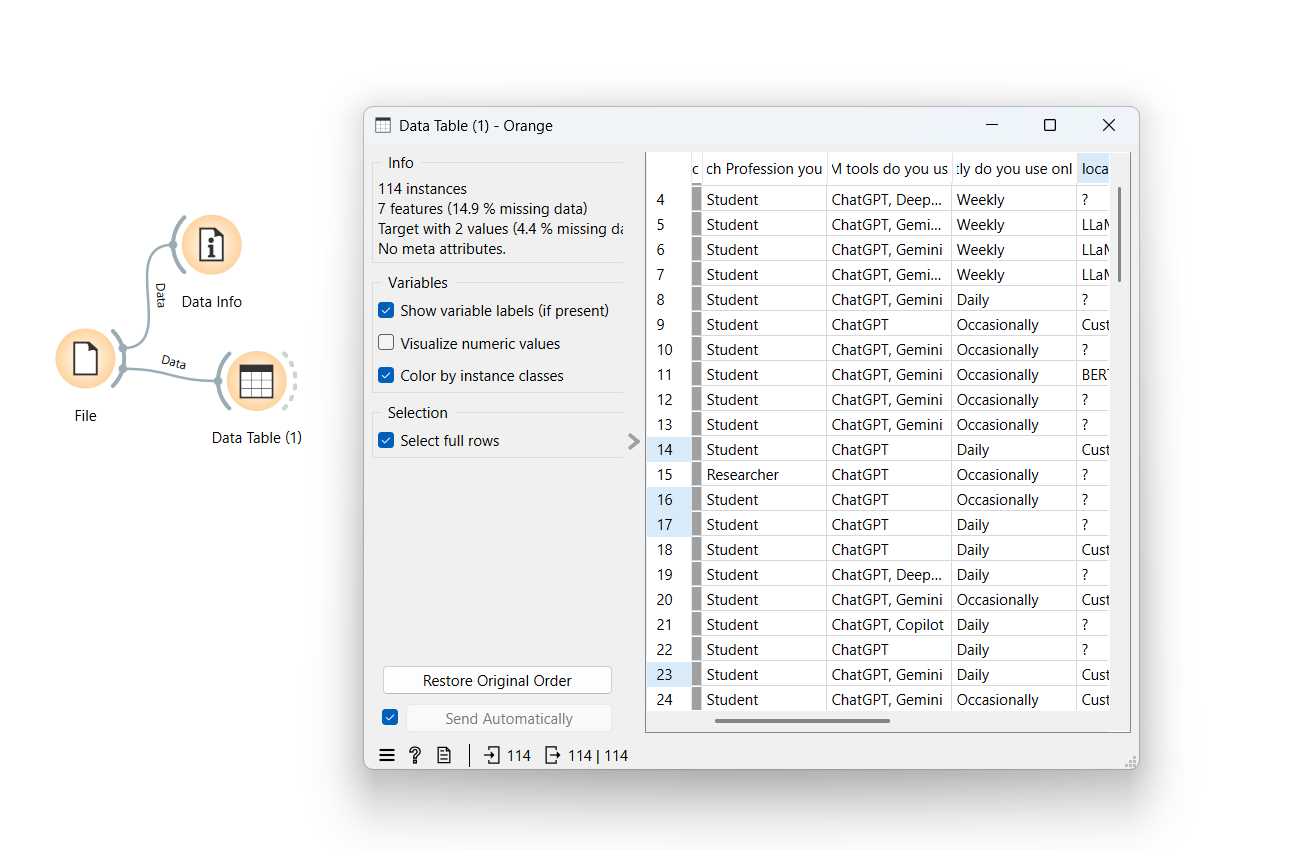


Fig:22

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

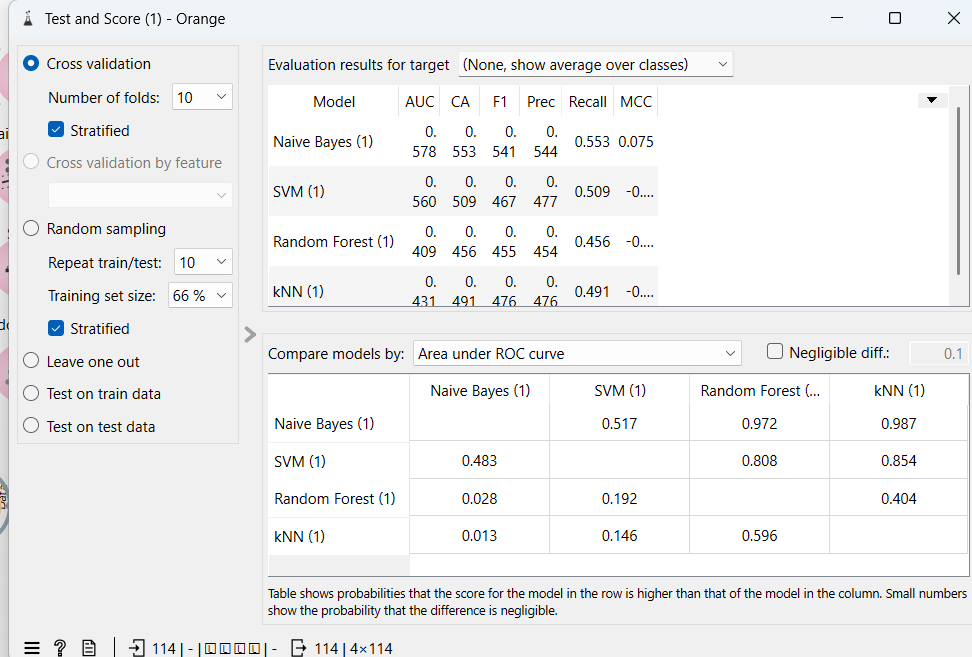
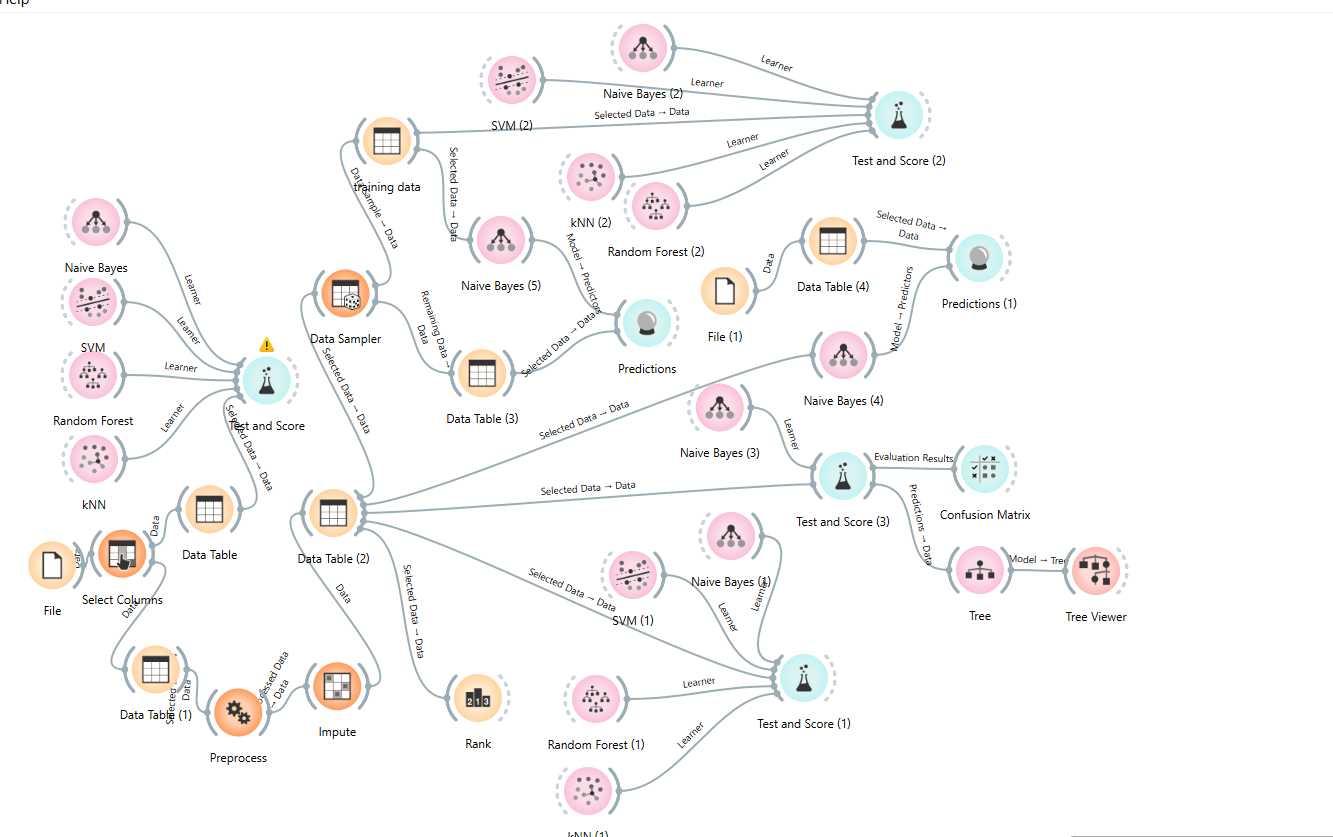


Fig:23

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



**Fig:24**

**CHAPTER 3: EXPERIMENTAL ANALYSIS**

* Study Classifier Accuracy

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

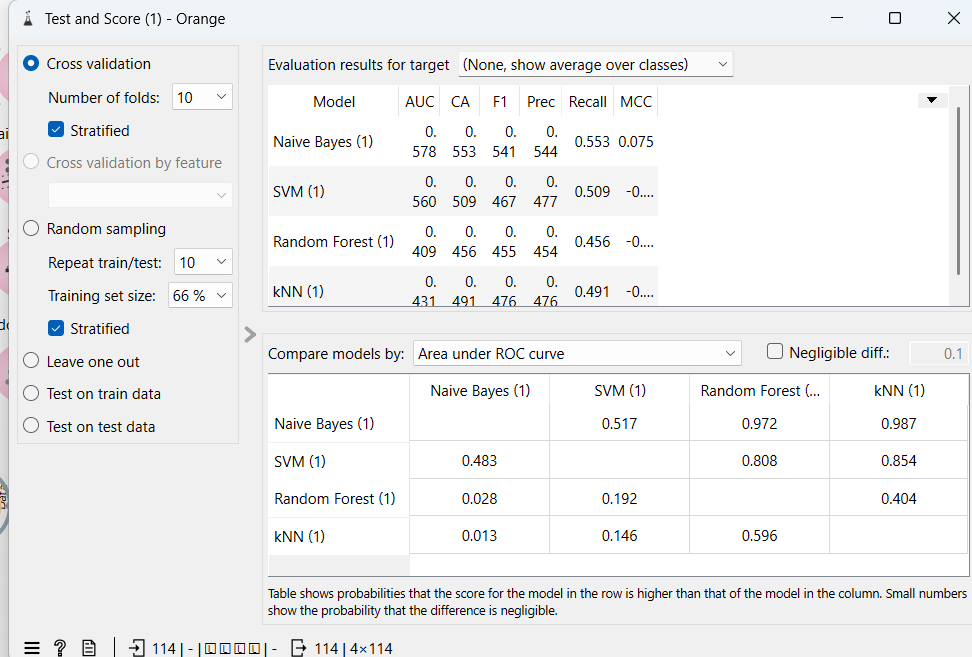


Figure 25: Test & Score measurements

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

* Evaluate Model Performance

Observe the confusion matrix and derive metrics such as Accuracy, F- measure, True Positive Rate (TPR), False Positive Rate (FPR), Precision, and Recall.

Apply cross-validation strategy with various fold levels in the Test & Score widget to compare accuracy results.

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

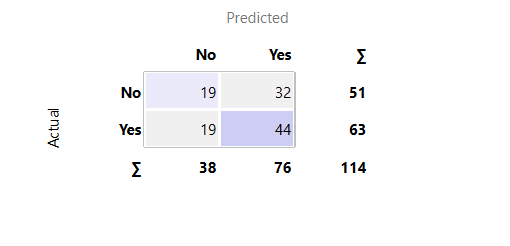


Figure 26: Confusion matrix of tree

The confusion matrix evaluates a model predicting whether a user is utilizing a **Local LLM** or an **Online LLM**. It correctly classifies **19 Local LLM** and **44 Online LLM** users but misclassifies **32 Local LLM** as Online and **19 Online LLM** as Local. With an accuracy of **55.26%**, the model struggles with false positives, often misidentifying Local LLM users.

**Prediction model:**

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

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

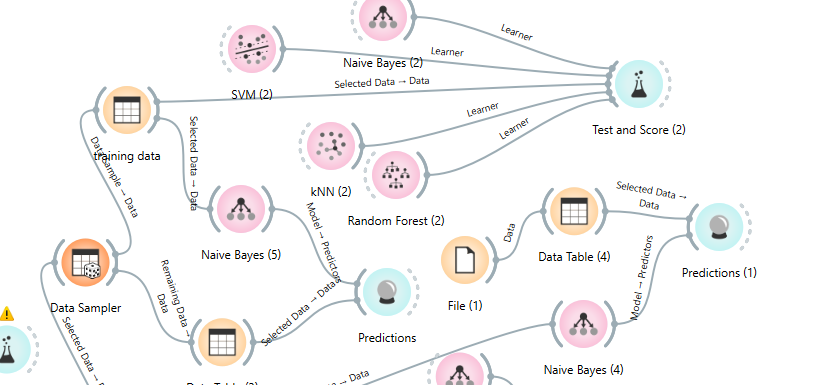


Fig:27

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

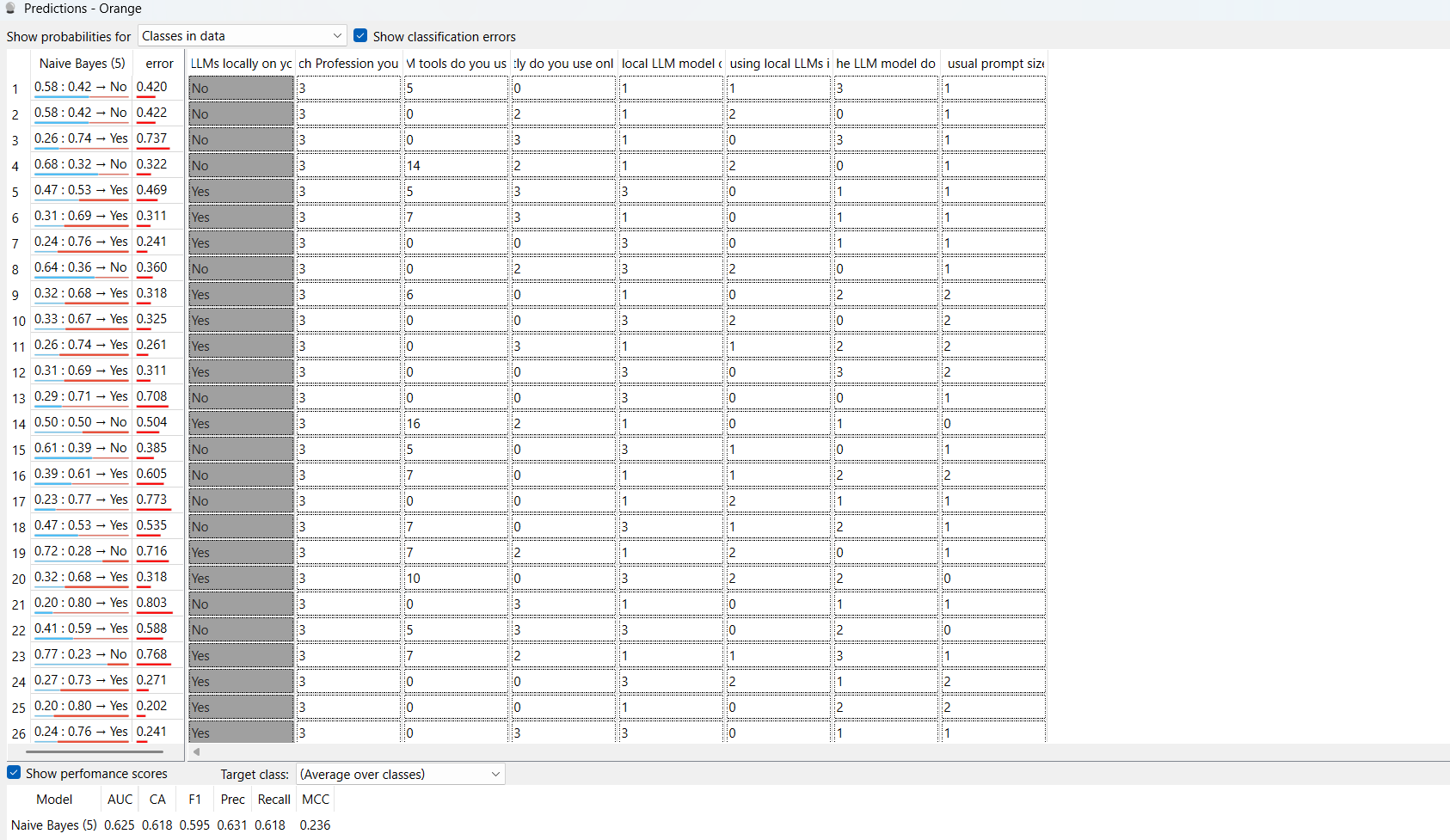


Fig:28

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

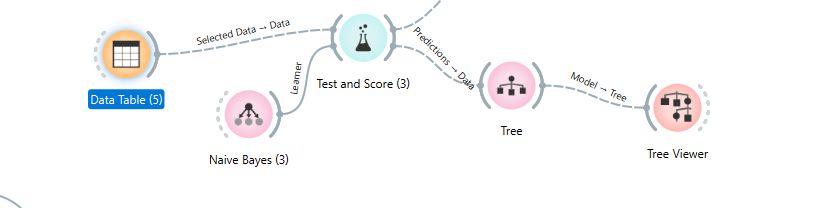


Fig:29

* The classification tree for the Classification model is as follows

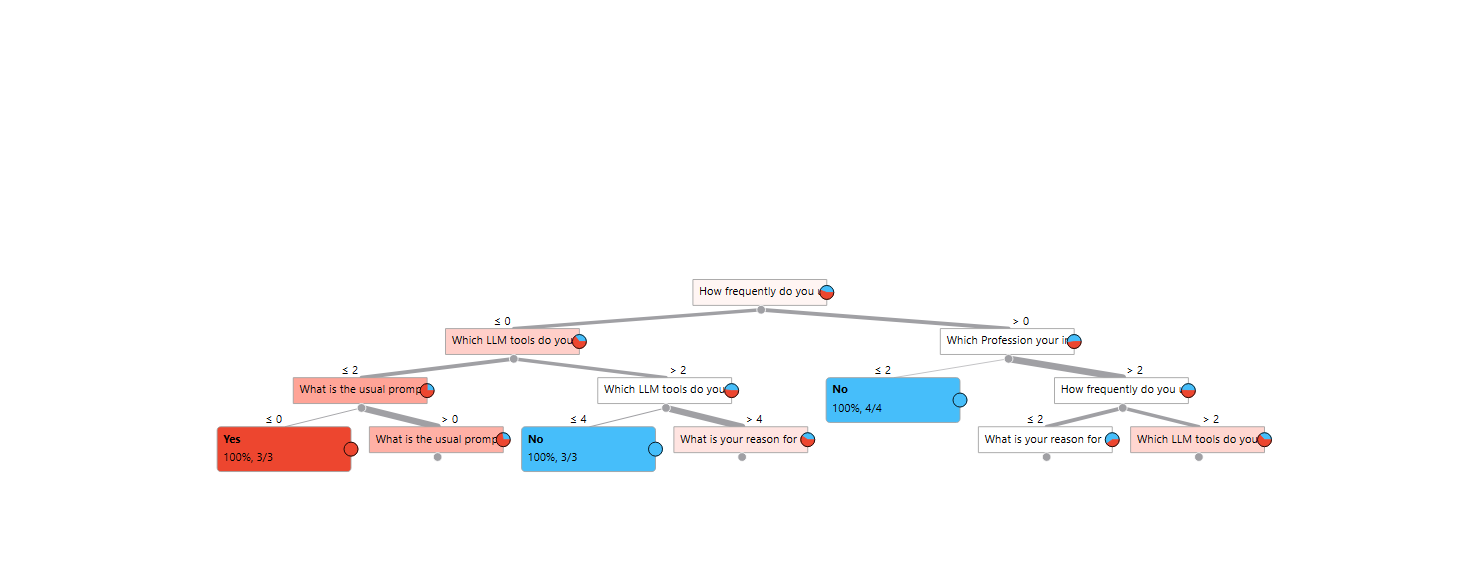


Figure 30: Classification tree

Figure 10 shows that the classification tree has four levels and the leaf nodes contains the values in the target class.

**CONCLUSION:**

In this phase of our project, we developed a data-driven approach to classify users based on whether they utilize a Local LLM or an Online LLM, using their interaction patterns and system attributes. The process began with data collection and preprocessing, ensuring quality through cleaning, normalization, and feature engineering. We then designed and implemented OLAP schemas (Star, Snowflake, and Fact Constellation), structured fact and dimension tables using Oracle SQL in SSMS, and visualized these schemas in Visual Studio.

Once deployed on the SSAS server, we performed OLAP operations such as Drill-Down, Roll-Up, Slice, and Dice to analyze user trends in LLM usage. The insights derived were visualized using Orange Tool, enabling a deeper understanding of how users interact with Local and Online LLM models.

Further, we implemented classification analysis to categorize users based on their LLM preference, considering factors such as response time, processing speed, network dependency, and interaction frequency. Various machine learning models were evaluated, and the best-performing model was selected using accuracy, precision, recall, F1-score, and ROC analysis.For my project Naïve Bayes model gave highest accuracy.

This project demonstrates how OLAP and machine learning can effectively distinguish between Local and Online LLM users. The insights gained can be utilized for resource allocation, optimizing LLM performance, and enhancing user experience, ensuring efficient AI model deployment strategies.

**KEY FINDINGS:**

* The analysis reveals that a significant majority of users prefer using **online LLMs** over offline ones, highlighting the convenience and accessibility of cloud-based models.
* **Students and developers** form the largest group of online LLM users, utilizing tools like ChatGPT and Gemini for learning, coding, and productivity enhancement.
* **Researchers** are more inclined towards using **locally deployed LLMs**, driven by concerns related to **data privacy**, **customization**, and **faster processing** without internet dependency.
* The choice between online and offline LLMs is influenced by factors such as **profession, frequency of use, model version, and prompt size**, helping tailor LLM tools to user-specific needs.

**PART-B: POST-OPERATIVE PATIENT DATA ANALYSIS**

**CHAPTER 1: INTRODUCTION ON DATA MINING METHODOLOGY**

**1.1 Problem Statement :**

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

**1.2 Identification of appropriate Methodology:**

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

**1.2.1 Dataset Overview**

The dataset contains **post-operative patient data**, including **vital signs (temperature, oxygen saturation, blood pressure), stability indicators, and comfort levels** to assess patient conditions. Based on these attributes, the **discharge decision (ADM-DECS)** classifies patients into three categories: **Intensive Care (I), Home Discharge (S), or General Hospital Admission (A)**. This dataset can be analyzed to identify patterns in patient stability and optimize post-operative care decisions.

**1.2.2 Methodology**

A diagram of a model

AI-generated content may be incorrect.We need the processes the dataset and make sure there are no redundancies test various classification algorithms and then develop the prediction model by training with training dataset and testing it with the testing dataset.

**Fig:1**

**1.2.3 Machine Learning Models**

We intend to use Supervised Machine Learning models to classify Patients by selecting target variable as I,S,A. They are Naïve Bayes, Random Forest, Decision Tree, KNN, SVM.

**1.2.4 Evaluation Metrics**

Since this is a classification problem, we can use:

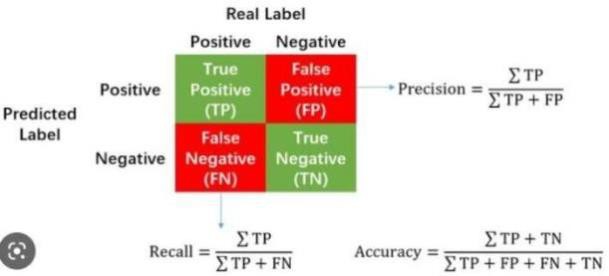
**Accuracy:** Accuracy is Calculated and Compared and best one should be noticed.

**Precision:** It counts the number of predictions from the positive class that are actually in that class.

**Recall:** It calculates how many positive class predictions were made using all of the dataset's positive examples.

**F-Measure:** It offers a single score that evenly weighs issues of precision and recall.

**Confusion Matrix:** It is used to determine the classification models performance for a set of test data.



**Fig:2** Confusion matrix

#### Block Diagram

The diagram illustrates the machine learning workflow for classification:

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

A diagram of a data processing process

AI-generated content may be incorrect.

Fig:3Block Diagram

**CHAPTER 2: ANALYSIS ON THE DATASET**

**TITLE:** “**Post-Operative Patient Data Analysis”**

**Data Set Description:**

The dataset contains information about post-operative patients, focusing on their vital signs, stability indicators, and comfort levels. It includes 90 records with features like temperature, blood pressure, and oxygen levels, along with an admission decision outcome for each patient.

These features include:

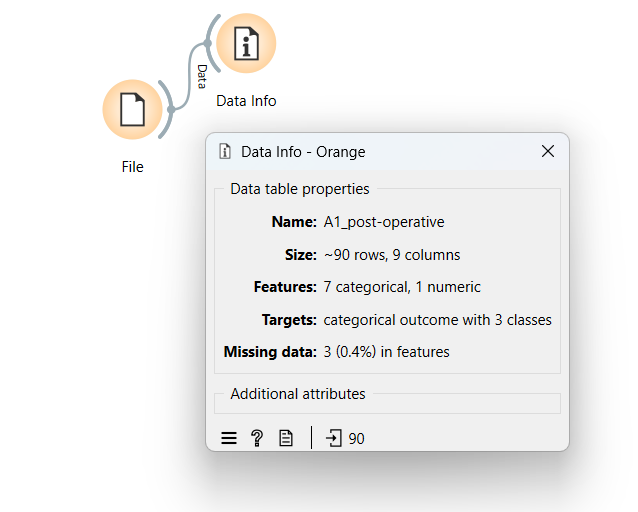


Fig:4

Each entry in the data provides valuable insights into post-operative patient conditions and medical decisions, aiding in the classification of admission outcomes based on key health indicators. This data is instrumental in analyzing patient stability, supporting clinical decision-making, and enhancing post-operative care strategies.

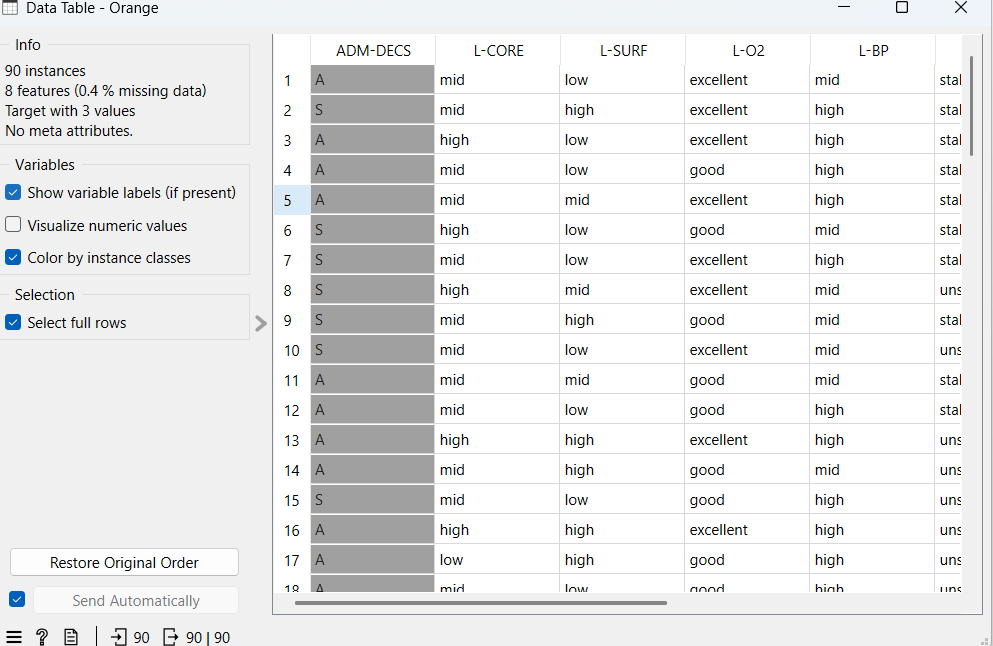


Fig:5

**Dataset Splitting:**

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

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

This split was designed to **maintain a balanced representation of class label**  in both sets, allowing us to effectively train and test machine learning models despite the limited number of records.

**Data Visualization:**

For data visualization, we utilized the **Orange Data Mining** tool to analyze relationships and correlations among key features such as temperature, blood pressure, and oxygen levels. This helped us detect dependencies between features and identify influential attributes for classification.

Before proceeding with modeling, we performed **data cleaning** to ensure no **duplicate** or **missing values** were present. Next, we analyzed the **distribution of admission decision outcome**. This **simplified the model**, improving both interpretability and performance.

**Data Validation, Cleaning, and Preparation Process**

We meticulously assessed the data to ensure its accuracy and readiness for analysis. We began by identifying relevant variables such as admission decision (target variable), core and surface temperatures, oxygen levels, blood pressure, and patient comfort. Through careful examination, we addressed missing values and ensured the consistency of each feature, maintaining the data’s integrity for effective classification.

To enhance the reliability and performance of our classification model, we applied several **data preprocessing techniques** to the A1 Post-operativedataset using the **Orange Data Mining tool**:

* **Handling Missing Values:s**  
  The dataset included missing values in features like L-CORE, L-SURF, L-O2, L-BP, and others. We used the **Impute widget** to handle these: i) **Average Imputation** was applied to numeric-like fields (e.g., COMFORT which contained values like ‘10’, ‘15’) to maintain the statistical consistency of the data. Ii) **Frequent Value Imputation** was used for categorical attributes such as L-CORE, L-SURF, and L-O2, replacing missing entries with the most common category in each column.
* **Discretization of Continuous Variables:**  
  The COMFORT feature, which included continuous values (like 10, 15), was transformed into categorical intervals using the **Entropy-MDL (Minimum Description Length)** discretization method. This allowed the model to effectively learn from grouped levels of patient comfort rather than individual numeric values.
* **Randomization of Features:**  
  To reduce data bias and test model robustness, the **Randomize widget** was used on select feature columns. This helped assess whether the model was overly reliant on any specific feature ordering and supported more generalized performance evaluation.

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

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

**Machine Learning Techniques and Model Selection**

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

1. **K-Nearest Neighbors (KNN)**
2. **Random Forest**
3. **Naïve Bayes**
4. **Support Vector Machine (SVM)**

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

**CHAPTER 3: WORKING ON THE DATASET**

**(DEVELOPING PREDICTION MODEL)**

# Orange Data Mining tool description:

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

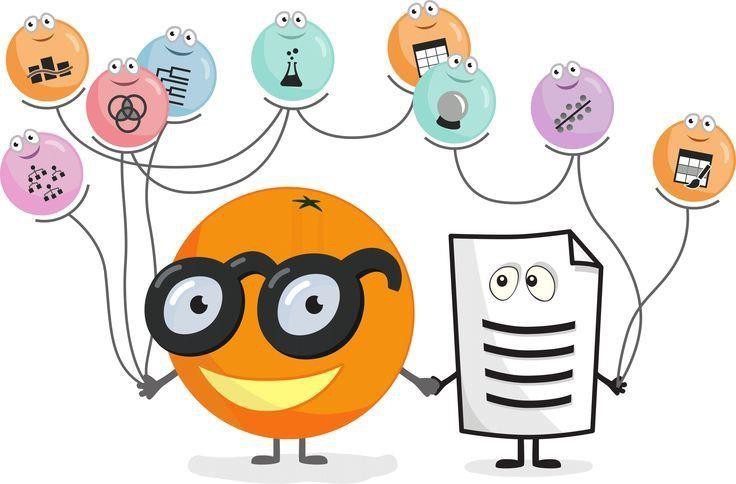


Fig:6

Step-by-Step Guide for Classification Using Orange

**Step 1:** Open Orange Canvas

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

**Step 2:** Load Dataset

* Drag and drop the "File" widget onto the canvas.
* Click on the "File" widget and then click on the "Browse" button.
* Choose your dataset(e.g.”A1\_post\_operative.tab”) and open it.

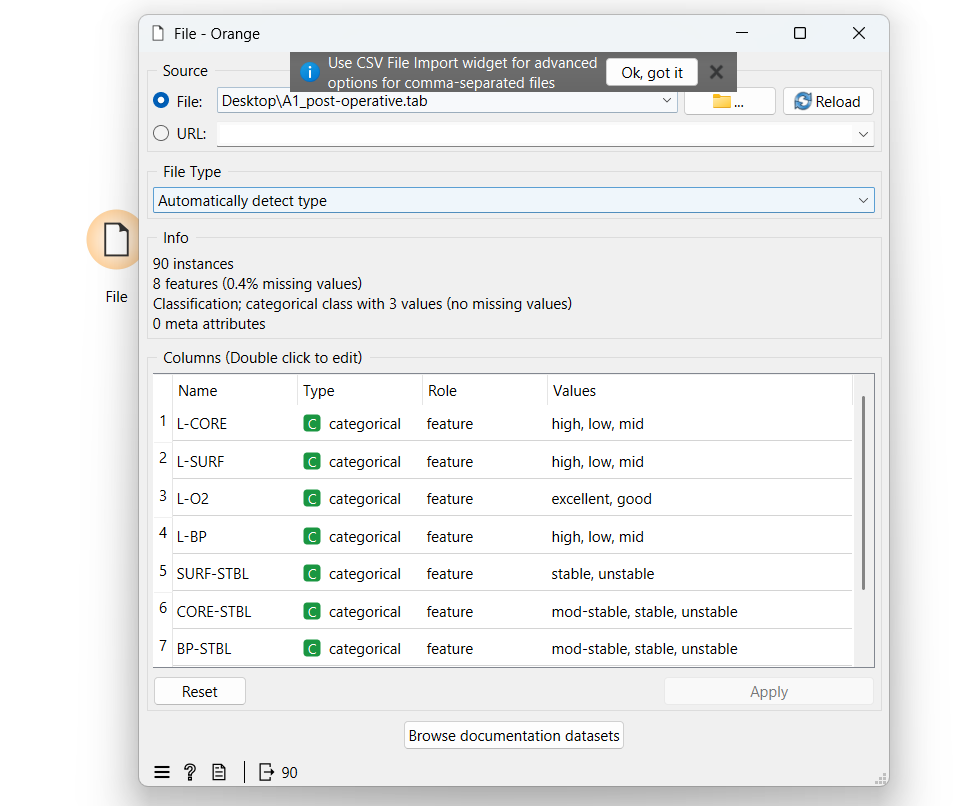


Fig:7

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

### 

### Fig:8

### Step 4: Preprocessing dataset

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

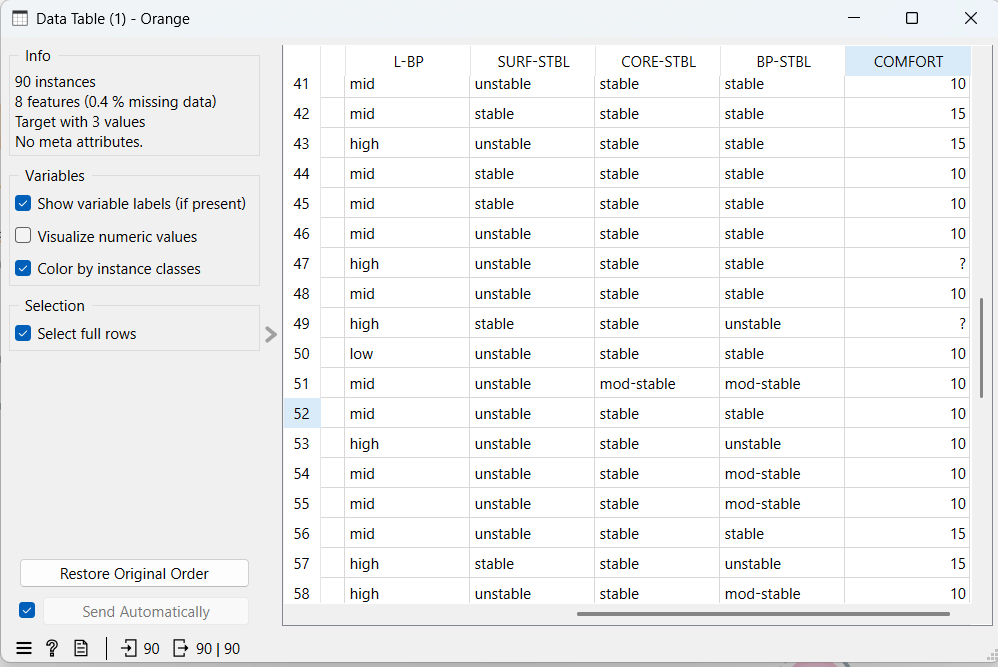


Fig:9Missing values before preprocessing

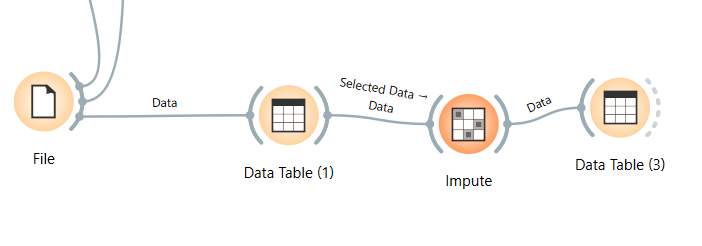


Fig:10 Preprocessing and imputing

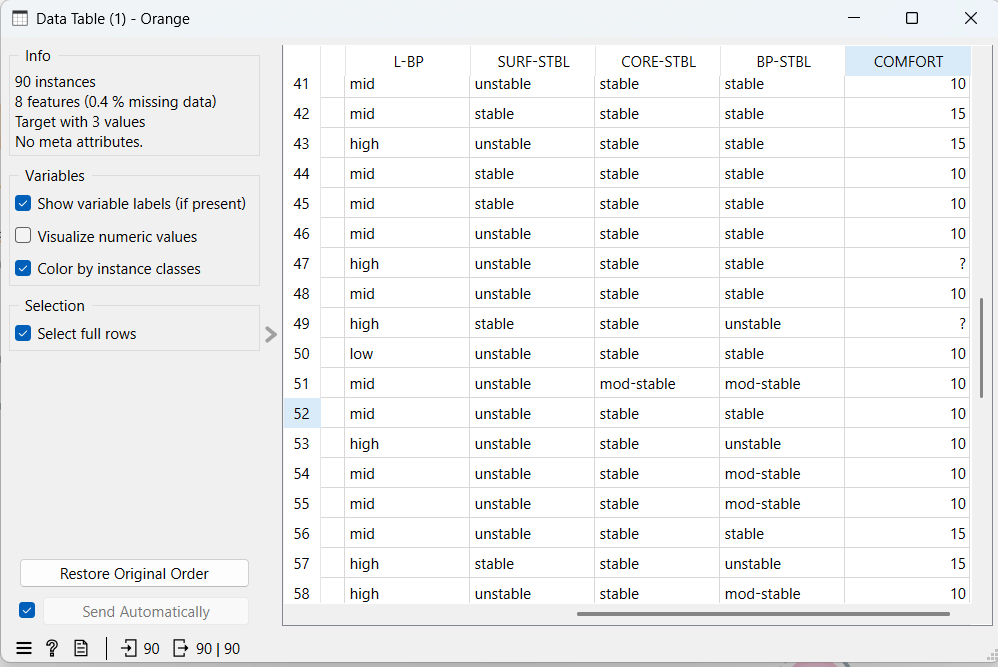


Fig:11 Missing values after preprocessing

**Step 5:** Imputing target variables

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

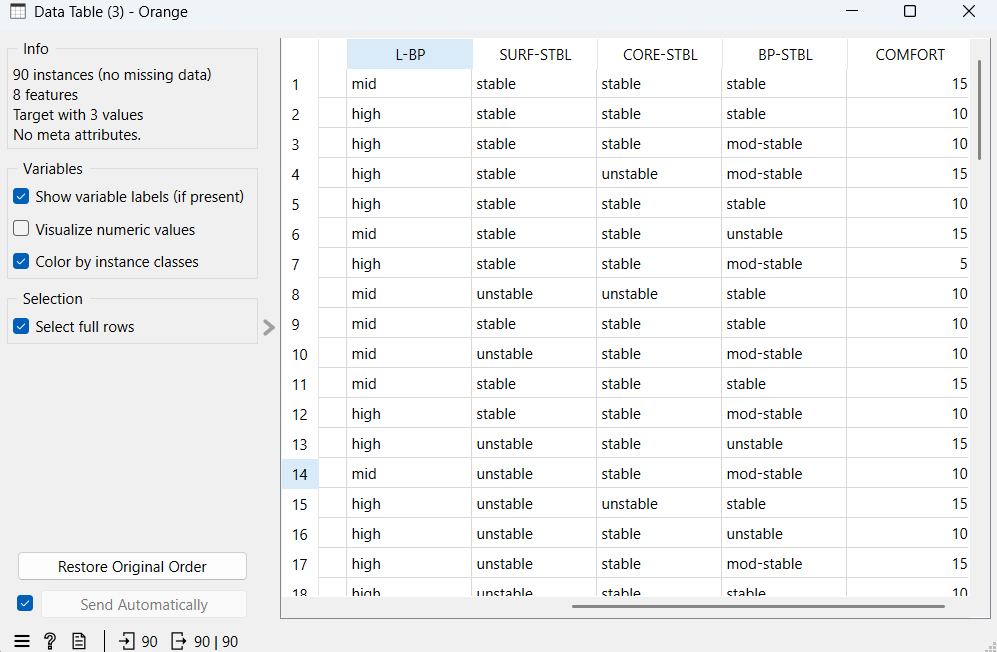


Fig:12 Dataset now has no missing values after imputing the target variable

**Step 6:** Testing accuracy of various classification algorithms

* Drag and drop the "Test & Score" widget
* Connect the "KNN", ”Random Forest”, “Naïve Bayes”, “Support Vector Machine” widgets to the "Test & Score" widget.
* Click on the "Test & Score" widget to view the classifier output, including accuracy, precision, recall, F-measure, and other metrics.

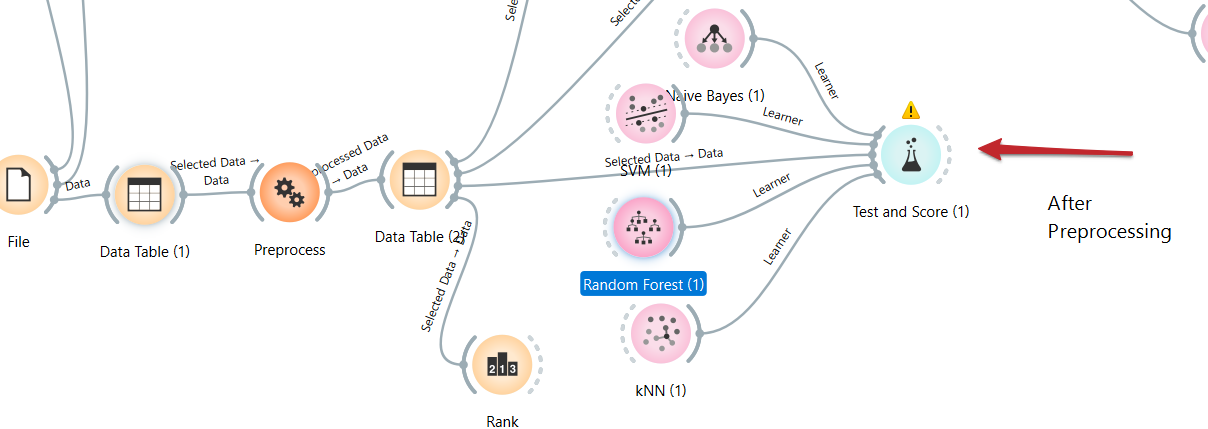


Fig:13

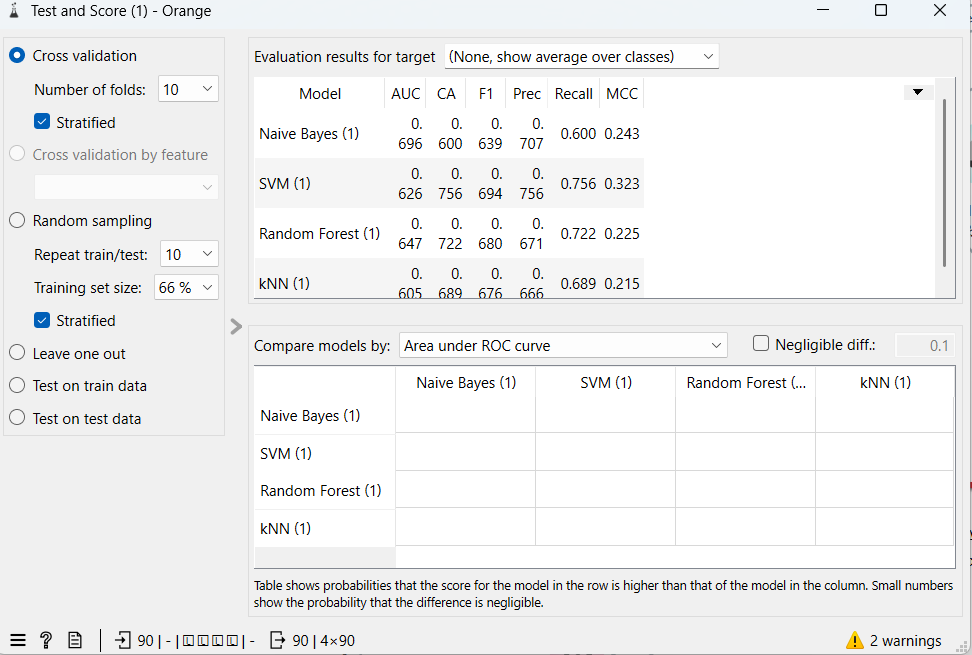


Fig:14

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

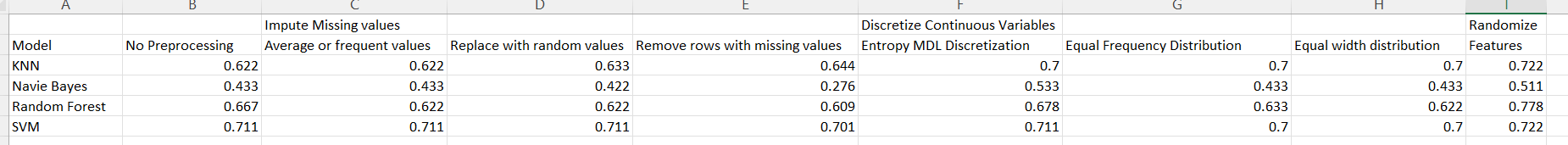


Fig:15

* **SVM** showed best CA before preprocessing and **Random Forest** showed best CA after applying preprocessing techniques.

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

* The prediction model needs both training and test data. Based on the training and test data the prediction model can be developed in two ways-

**First way** is by splitting the dataset into training and test datasets using the data sampler

This is clearly explained the figure below:

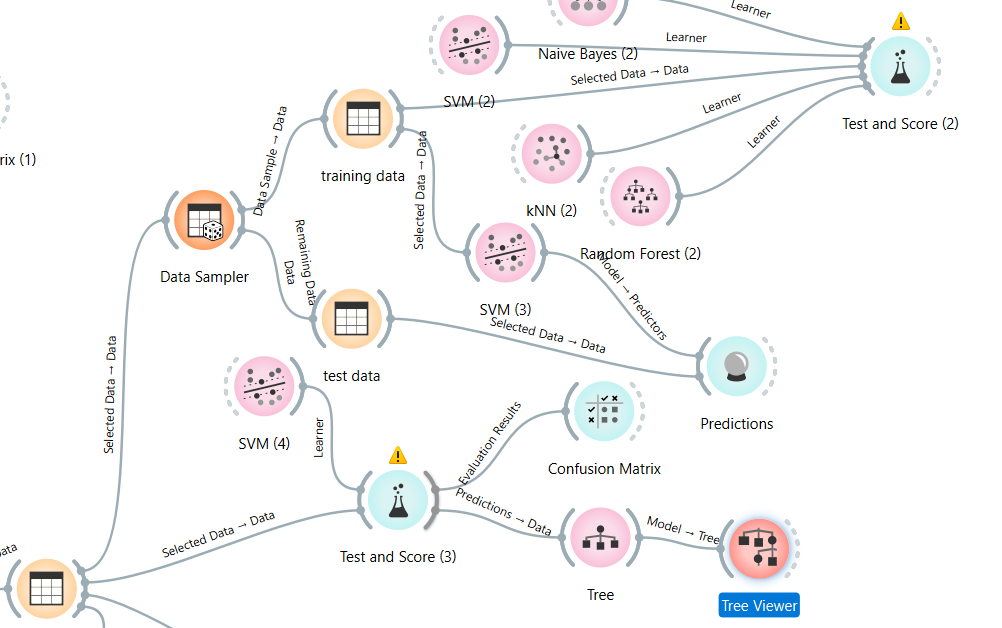


Fig:16 Prediction model by data sampler

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

This is explained the figure below:

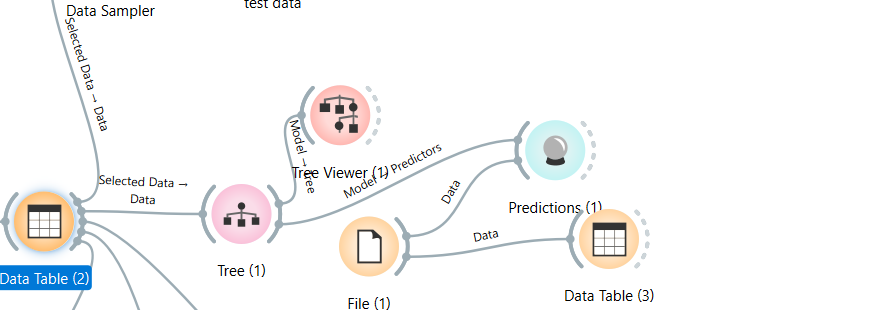
777

Fig:17 Prediction model by separate train and test data

**Entire Workflow:**

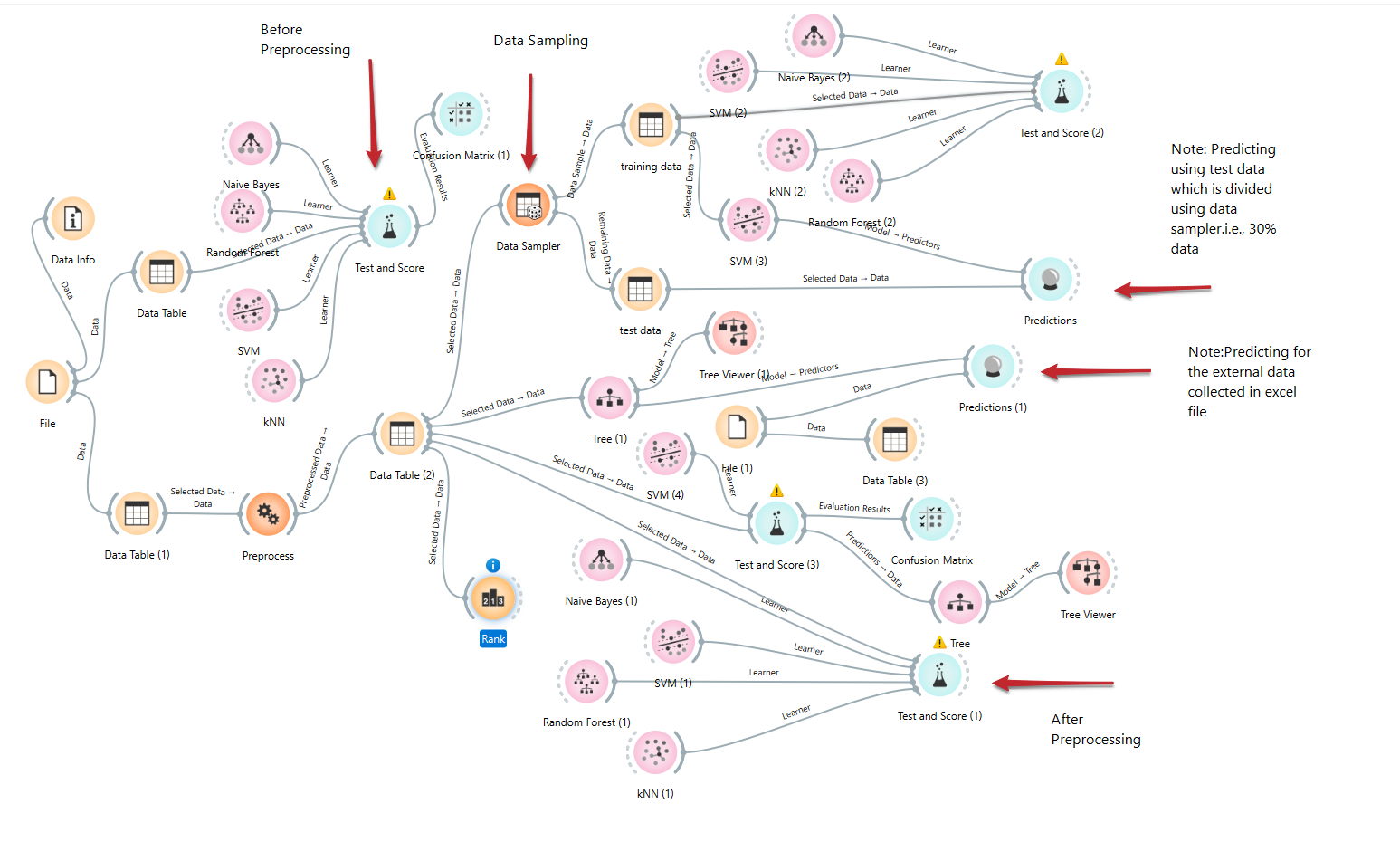
****

Fig:18

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

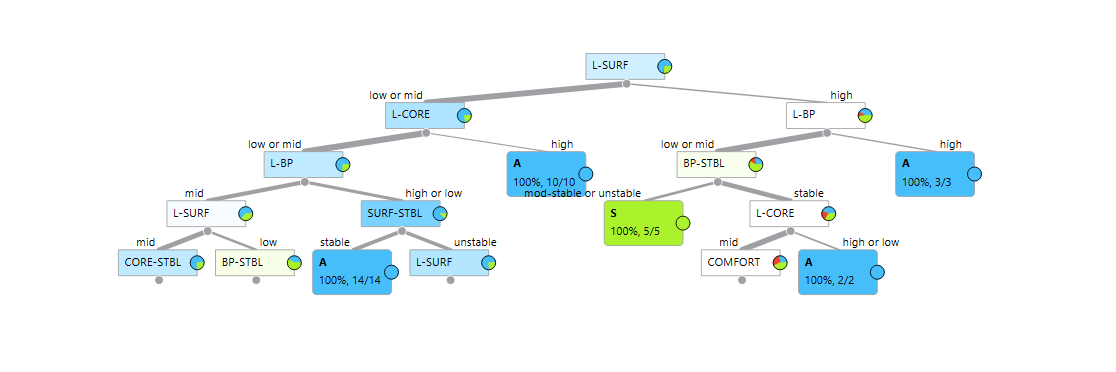


Figure 19: Classification tree for admission decision attribute

Figure 11 depicts the decision tree visually represents classification based on multiple attributes like **L-SURF,L-CORE,L-BP,SURF-STBL..**. Nodes split based on attribute values, with leaf nodes showing classification percentages. Green nodes indicate higher confidence in classification, while blue nodes represent lower confidence, suggesting possible misclassifications or mixed results in those branches.

**CHAPTER 4: EXPERIMENTAL ANALYSIS**

* Based on the Classifier accuracy that is shown in the Test & Score widget we choose to evaluate

SVM and Random Forest algorithms using metric like confusion matrix .

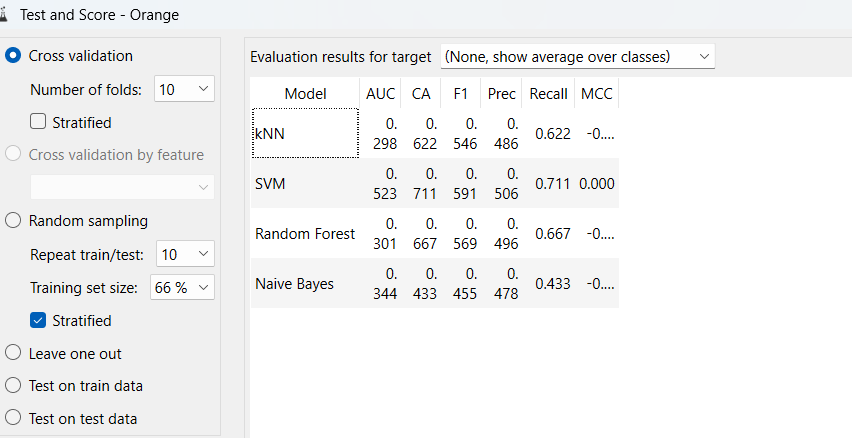


Figure 20: CA before preprocessing

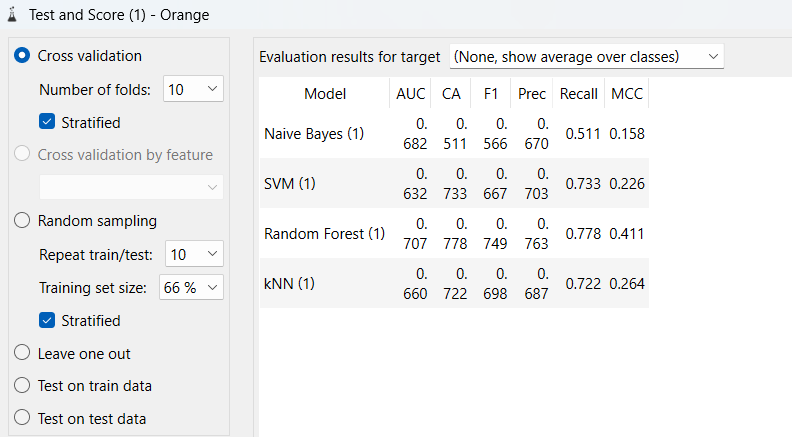


Figure 21: CA after preprocessing

Figure 12 shows Test & Score before preprocessing (10-fold Cross Validation, 66% Training Size)

* SVM: 0.711
* Random Forest: 0.667
* KNN:0.622
* Naïve Bayes:0.433

Figure 13 shows Test & Score after preprocessing (5-fold Cross Validation, 80% Training Size)

* Random Forest: 0.778
* SVM: 0.733
* KNN:0.722
* Naïve Bayes:0.511

**Comparison Observations**:

Random Forest saw the highest improvement, from 0.667 to 0.778.

SVM,KNN and Naïve Bayes also showed significant improvements (0.711 → 0.733 and 0.622 → 0.722,0.433→0.511 respectively).

Random Forest had a notable jump from 0.667 to 0.778.

**Analysis on Confusion matrices:**

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

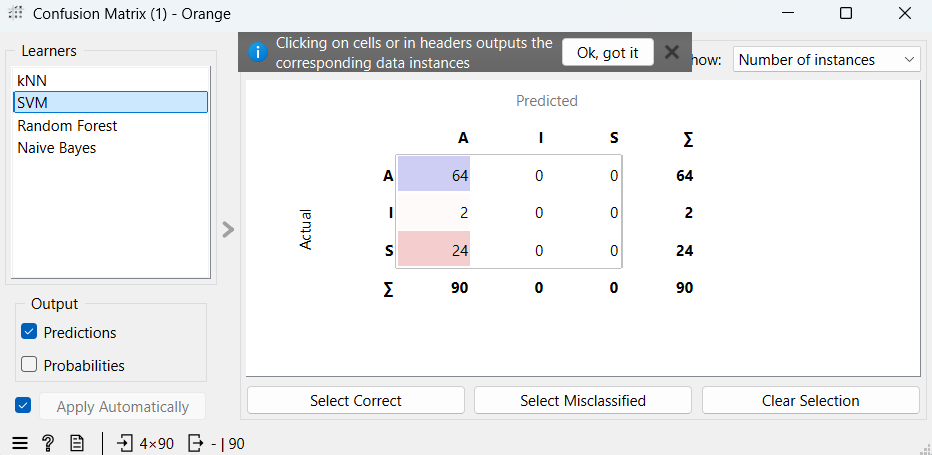
****

Figure 22: Confusion matrix of SVM

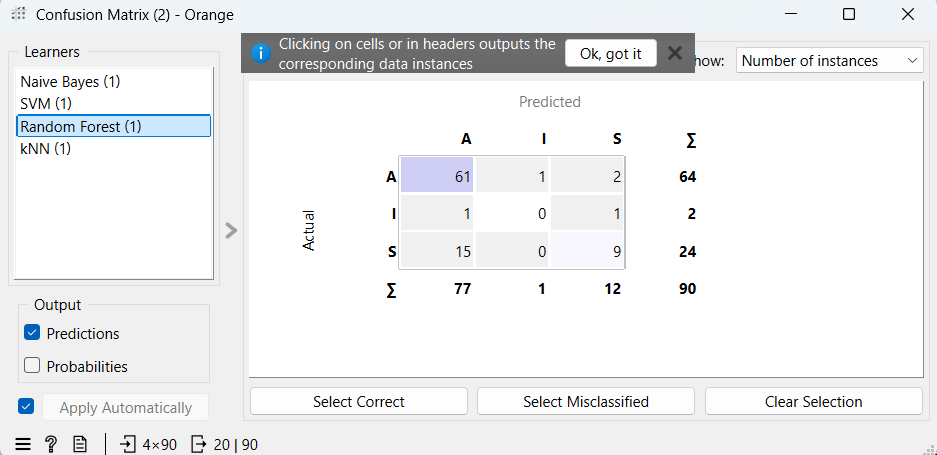
****

Figure 23: Confusion matrix of Random Forest

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

**Overall Accuracy:**

* SVM: Correct classifications = **(64+0+0) = 64** out of **90**.
* Random Forest: Correct classifications = **(61+0+9) = 70** out of **90**.
* **Random Forest has slightly better accuracy than SVM.**

**Class-wise Performance:**

* **Class A:**

 **SVM:** 64 correctly classified, 0 misclassified.

 **Random Forest:** 61 correctly classified, 3 misclassified (1 as Class I, 2 as Class S).

 **Observation:** SVM slightly better for Class A.

* **Class I:**

 **SVM:** 0 correctly classified, 2 misclassified as Class A.

 **Random Forest:** 0 correctly classified, 1 misclassified as Class A.

 **Observation:** Random Forest slightly better for Class I.

* **Class S:**
  + **SVM:** 0 correctly classified, 24 misclassified as Class A.
  + **Random Forest:** 9 correctly classified, 15 misclassified as Class A.
  + **Observation:** Random Forest significantly better for Class S.

**Misclassification Trends:**

* SVM has a higher misclassification of Class S as Class A (24 cases).
* Random Forest reduces the misclassification of Class S into Class A (15 cases), but introduces a few misclassifications in Class A (3 cases total).
* Random Forest improves the classification of Class S, while SVM performs better for Class A with perfect accuracy.

Random Forest has a slightly better accuracy (70/90 vs. 64/90).Random Forest performs better for Class S and slightly for Class I, while SVM performs slightly better for Class A with perfect classification.If the main priority is Class A accuracy, SVM might be preferable.If the goal is overall classification improvement and better handling of minority classes like Class S, Random Forest is the better choice.

**CONCLUSION:**

In this classification analysis using Orange Data Mining, we explored various machine learning models on the Post-Operative Patient Dataset, both before and after applying data preprocessing techniques. Below are the key findings and future research directions:

**1. Preprocessing Impact**

* Preprocessing steps such as imputing missing values (with average/frequent values), normalization of numerical features, discretization using Entropy-MDL, and randomization of features significantly enhanced the performance of several models.
* These steps helped improve model consistency, reduce bias from incomplete data, and achieve better generalization.

**2. Overall Classification Accuracy**

* Random Forest outperformed other models with an accuracy of 70 out of 90, while SVM followed closely with 64 out of 90.
* Random Forest performed well across all classes, especially in handling misclassification of Class S, which was entirely misclassified by SVM.

**3. Confusion Matrix Insights**

* SVM:
  + Perfectly classified Class A (64/64) but failed to classify Class I and S correctly.
* Random Forest:
  + Correctly classified Class A (61/64) and Class S (9/24), showing more balanced performance across classes.
* Misclassification trends revealed that SVM heavily misclassified minority classes into Class A, whereas Random Forest distributed predictions more appropriately.

**4. Model-wise Class Performance**

* Class A: SVM had slightly better accuracy (64 correct vs. 61 in Random Forest).
* Class I: Both models struggled, but Random Forest had fewer misclassifications.
* Class S: Random Forest significantly outperformed SVM (9 correct vs. 0).

**Final Decision**

* If overall classification accuracy and balanced performance across all classes is the goal, Random Forest is the preferred model.
* If maximum precision for Class A is critical, SVM may be considered due to its perfect classification in that category.
* Model selection should align with the specific healthcare objective, such as minimizing misdiagnosis for certain postoperative conditions.

**PART C: FINAL ANALYSIS**

**1. Introduction**

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

Part A, which used a generated dataset, and

Part B, which used an online dataset collected from external sources.

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

**2. Key Observations from Experimental Analysis**

**2.1. Data Characteristics and Preprocessing**

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

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

**2.2. Model Performance Analysis**

* Across both experiments, various classifiers (KNN, SVM, Naïve Bayes, and Random Forest, etc…) were tested, and their performances were evaluated using classification accuracy (CA)`, and confusion matrices.

**2.3Key Findings from Model Comparisons**

* Naïve Bayes performed the best for Part-A (LLM Preference Classification), showing strong accuracy in predicting user preference between Local and Online LLM models based on interaction and system-level features.
* Random Forest outperformed other models in Part-B (Post-Operative Dataset) with 70/90 correct classifications, demonstrating its robustness in handling imbalanced medical data and multiple classes.
* SVM performed exceptionally for Class A in the Post-Operative dataset but failed to identify minority classes (Class I and S), highlighting its limitation in unbalanced data contexts.

**2.4 Preprocessing Differences**

Part A: Extensive Preprocessing Required *(LLM Model Classification)*

* Raw interaction data required missing value imputation, normalization, and feature engineering.
* Preprocessing played a crucial role in improving the performance of classification models like Naïve Bayes and Logistic Regression.
* The dataset contained categorical and numerical attributes, needing careful scaling and transformation.

Part B: Moderate Preprocessing *(Post-Operative Dataset)*

* Data was already relatively clean but required categorical encoding, class balancing, and scaling.
* Handling imbalanced class distribution (especially for Class I and S) was essential for fair model evaluation.

Impact:

* In Part A, preprocessing was critical for transforming abstract system interactions into meaningful model features.
* In Part B, it enhanced the models' ability to learn from underrepresented classes, improving accuracy.

**2.5 Conclusion**

This study explored classification performance across two distinct datasets and application areas:

* Part A (LLM User Preference Classification) focused on differentiating user interaction behavior with Local vs. Online LLMs. Naïve Bayes was most effective, benefiting from simple yet structured probabilistic modelling.
* Part B (Post-Operative Dataset) dealt with real-world healthcare data. Random Forest showed balanced performance across all classes and was particularly better at identifying minority class (Class S), which SVM completely misclassified.

**2.6 Final Takeaways:**

* Part A required a data-driven, feature-engineered approach and highlighted the importance of preprocessing in behavioural classification tasks.
* Part B demanded robust handling of class imbalance and showcased Random Forest’s strength in multi-class medical diagnosis.
* Overall, Naïve Bayes (Part A) and Random Forest (Part B) emerged as the best-suited models, each fitting their domain’s unique requirements.

**REFERENCES**

**1.LLM Models: Online vs Local Usage Analysis (using Data Mining & Orange Tool)**

* Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Foundational reference for understanding deep learning models, including transformers used in LLMs.

* Rothman, D. (2021). *Transformers for Natural Language Processing*. Packt Publishing.

Explains LLM architecture, training, and deployment strategies including local vs cloud-based usage.

* Huyen, C. (2022). *Designing Machine Learning Systems*. O'Reilly Media.

Discusses practical deployment of ML systems including online and offline model performance.

* Zupan, J., & Zupan, B. (n.d.). *Visual Data Mining with Orange: Data Mining Fruitful and Fun*.

https://orange.biolab.si  
Official guide for using Orange tool in practical ML applications like LLM behavior analysis.

* Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

Covers classification, clustering, and model evaluation methods applicable to LLM usage analysis.

**2.Post-Operative Patient Analysis (using Data Mining & Data Warehousing)**

* Larose, D. T., & Larose, C. D. (2015). *Data Mining and Predictive Analytics* (2nd ed.). Wiley.

Includes healthcare-related predictive analysis with real datasets; aligns well with patient health data mining.

* Zupan, J., & Zupan, B. (n.d.). *Visual Data Mining with Orange*.

https://orange.biolab.si  
Visual interface and workflows in Orange for analyzing medical datasets.

* Inmon, W. H. (2005). *Building the Data Warehouse* (4th ed.). Wiley.

A definitive guide to structuring and using data warehouses for complex analysis tasks.

* Anahory, S., & Murray, D. (1997). *Data Warehousing in the Real World*. Pearson Education.

Focuses on integrating, cleaning, and preparing data for large-scale analysis in domains like healthcare.

* Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). *From Data Mining to Knowledge Discovery in Databases*. AI Magazine, 17(3), 37–54.

A classic paper that introduces the KDD process — essential when working with patient datasets.

**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