### Predicting AI Tool Usage Patterns Using Classification Models

### 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

## M. JNANA PRASANNA (22481A05E0)

## P. SANJANA (22481A05J1)

## M. SHRIVALLI (22481A05F8)

## M. RAJ VARDHAN (22481A05E7)

Under the Enviable and Esteemed Guidance of

# Dr. G.V.S.N.R.VPRASAD, M.Tech, M.S, Ph.D.

Professor of CSE & Director - PGCRD

## DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE

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

**SESHADRI RAO KNOWLEDGE VILLAGE**

**GUDLAVALLERU - 521356 ANDHRA PRADESH**

**2024-25**

## SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE

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

**SESHADRI RAO KNOWLEDGE VILLAGE, GUDLAVALLERU**

## DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

****

**CERTIFICATE**

This is to certify that the project report entitled **“Predicting AI Tool Usage Patterns Using Classification Models”** is a Bonafide ­record of work carried out by **M. JNANA PRASANNA (22481A05E0)**, **P. SANJANA (22481A05J1)**, **M. SHRIVALLI (22481A05F8)**, **M. RAJ VARDHAN** **(22481A05E7)**, under the guidance and supervision of **Dr. G.V.S.N.R.VPRASAD**, Professor of CSE & Director-PGCRD,M.S., M.Tech, Ph.D. 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. G.V.S.N.R.V Prasad) (Dr. M. BABU RAO)**

**External Examiner**

## ACKNOWLEDGEMENT

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

We would like to express our deep sense of gratitude and sincere thanks to

**Dr. G.V. S.N.V.R Prasad, Professor of Computer Science and Engineering & Director(PGCRD) M.S., M.Tech, Ph.D.** for his constant guidance, supervision, and motivation in completing the project work.

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

We would like to take this opportunity to thank our beloved principal, **Dr.B.Karuna Kumar, M.Tech, Ph.D.,** for providing great support for us in completing our project and giving us the opportunity to do the project.

Our Special thanks to the faculty of our department and the 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

**M. JNANA PRASANNA (22481A05E0)**

**P. SANJANA (22481A05J1)**

**M. SHRIVALLI (22481A05F8)**

**M. RAJ VARDHAN (22481A05E7)**

**INDEX:**

## TITLE

**LISTOFTABLES i**

[LISTOFFIGURES](#_bookmark0) ii

[ABSTRACT](#_bookmark1) iii

[CHAPTER1:INTRODUCTION](#_bookmark2) 01

* 1. [Introduction01- 06](#_TOC_250007)
  2. Problemdefinition 06

[CHAPTER2:PROPOSEDMETHOD 07](#_TOC_250006)

* 1. Methodology 07
     1. [BlockDiagram 08](#_TOC_250005)
     2. AlgorithmandExplanation 09
  2. [Data Preparation 10](#_TOC_250004)
     1. [DataSetDescription 10](#_TOC_250003)
     2. DataPre-processing 10

[CHAPTER3:RESULTS 12](#_TOC_250002)

* 1. ORANGEtooldescriptions 12
  2. Screenshots 12-23
  3. Interpretation 24

[CHAPTER4:CONCLUSIONANDFUTURESCOPE 25](#_TOC_250001)

[References 26](#_TOC_250000)

ListofProgramOutcomesandProgramSpecific Outcomes 27

**MappingofProgramOutcomeswithgraduatedPOSand PSOs**

## LISTOF TABLES

1. ClusteringModels
   1. K-Means
   2. Hierarchical
   3. DBScan
   4. MeanShift
   5. GaussianMixtureModels
   6. Agglomerative
   7. SpectralClustering
   8. AffinityPropagation
   9. Optics
   10. Birch 2.Classificationmethodology

3.BlockDiagram

## LISTOF FIGURES

* + 1. ClusteringDefinition
    2. K-Means
    3. DBSCAN
    4. HierarchicalClustering
    5. Correlation
    6. Outliers
    7. Clusteringmethodology
    8. :Scatterplot
    9. : Block Diagram
    10. DownloadandInstall Orange
    11. OpenanewFile
    12. LoadDataset
    13. DataInfoofdataset
    14. RankoftheDataset
    15. DataTablebeforefinding Outliers
    16. DataTableafter Finding Outliers

3.2.8.Applying Clustering Models

* + 1. VisualizationofScatterPlot
    2. Evaluationthrough Pythoncode
    3. EvaluationthroughSumofSquaredError
    4. Evaluationthrough SumofSquaredError
    5. Errorof Evaluations
    6. DataTablesafter Clustering
    7. CorrectlyClusteredwithK-Means
    8. Mis-ClassifiedwithDBSCAN,Hierarchical
    9. OverallWorkflow

## 

## ABSTRACT

The aim of this project is to predict AI tool usage patterns using classification models. By leveraging machine learning techniques, this study analyzes user behavior and adoption trends of AI tools across various domains. The objective is to identify key factors influencing AI tool usage and determine the most effective classification model for accurate predictions. This research helps organizations and developers understand user preferences and make data-driven decisions for AI tool development.

The dataset, containing AI tool usage survey responses, is preprocessed and analyzed using Orange, a data mining and machine learning tool. Various classification models, such as Decision Trees, Naïve Bayes, Random Forest, and Support Vector Machines, are applied to predict AI tool adoption. Feature selection techniques help in identifying the most significant attributes affecting AI tool usage.

Model performance is evaluated using accuracy, precision, recall, and F1-score metrics. A comparative analysis of different classification models is conducted to determine the most reliable approach for AI usage prediction. The study also explores the impact of demographic factors, industry preferences, and user experience on AI tool adoption trends.

The findings of this study provide valuable insights into AI tool usage patterns, helping businesses, researchers, and policymakers enhance AI adoption strategies. The results can be used to optimize AI tool development and marketing efforts, ensuring a better alignment with user needs. Future work can explore deep learning methods to improve prediction accuracy and extend the study to a larger dataset.

Additionally, this project highlights the advantages of using Orange as a visual machine learning platform, making predictive analysis more accessible to researchers and non-technical users. The tool’s interactive interface allows for easy implementation of classification models without requiring extensive coding expertise. By demonstrating Orange’s capabilities in predictive modeling, this study encourages wider adoption of data-driven decision-making in AI development.

**Keywords**:

* **AI Tool Usage**
* **Prediction**
* **Classification Models**
* **Machine Learning**
* **Orange Tool**

**PART-A**

### AI Tool Usage Patterns: A Demographic Analysis

### Using KDD Process

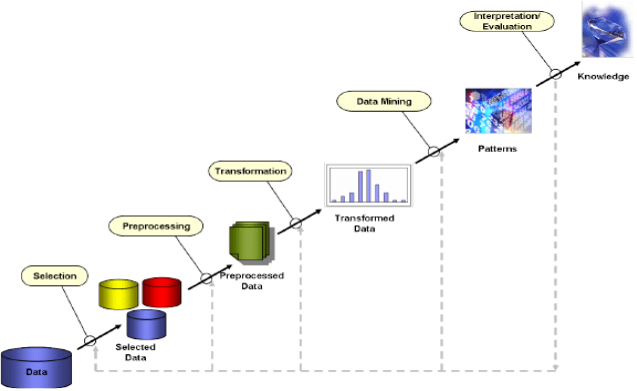
## CHAPTER 1: INTRODUCTION

## INTRODUCTION

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

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

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

****

* 1. **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.

### DATA WAREHOUSING

### In our AI tools usage analysis project, a data warehouse is used to store and analyze data collected from Survey. This centralized data repository allows for efficient trend analysis, user segmentation, and platform comparison.

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

**Fig 1.1 Data Warehousing Block Diagram**

**1.** **Data Source Layer (Extracting Data)**

* Data is collected from survey.
* Includes user demographics, Time spent, Frequency, Purpose, and device usage.

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

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

1. **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.

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

* Allows multi-dimensional analysis to identify **trends in user behavior**.
* Enables queries like:
* Which AI tool has the highest engagement?
* Do all Age groups use AI tools?
* What are the most common devices used ?

1. **Data Visualization & Reporting**

* Insights are presented using **dashboards, reports, and visual charts**.
* Helps AI tool users to optimize their services based on user trends.

### 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.

* **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 AI tools usage 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 which AI Tool platform a user is likely to prefer, based on attributes such as user demographics, Frequency, Time spent, Learning Priority.

#### 

**Fig 1.2 Data Mining Block Diagram**

**Fig Block Diagram of Data Mining**

### 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 AI tools usage dataset to grasp its structure and content.
* Identifying attributes such as user demographics, Frequency, Time spent, Learning Priority.

1. **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.

1. **Modelling**

* Applying various classification algorithms like Decision Trees, Random Forest, and Logistic Regression to predict a user's preferred platform.

1. **Evaluation**

* Assessing model performance using accuracy, precision, recall, and F1-score to ensure reliable predictions.

1. **Deployment**

* Integrating the best-performing model to provide insights into which platform users prefer based on their usage.

## SUPERVISED LEARNING

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

* **K-Nearest Neighbors (KNN)**
* **Decision Trees**
* **Random Forest**
* **Logistic Regression**

**Categories of Supervised Learning in This Project:**

1. **Classification:**

* The dataset contains categorical labels (e.g., Spotify, Apple Music, YouTube Music).
* Classification algorithms predict which platform a user prefers based on listening habits, demographics, and subscription type.

1. **Regression:**

* If we analyze listening duration as a continuous variable, regression models could predict how long a user is likely to listen on a platform.
* However, since our project focuses on platform prediction, classification is the primary approach.

| **Algorithm** | **Description** | **Type** |
| --- | --- | --- |
| Logistic Regression | Extension of linear regression that’s used for classification tasks. The output variable is 2 binary either yes or no | Classification rather 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 |

### 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.

### Association:

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

### How to Choose a Data Mining Algorithm?

Choosing the right data mining algorithm depends on:

* **If the data has labels:**

Use Supervised Learning (Classification/Regression)**.**

* **If the data has no labels:**

Use Unsupervised Learning (Clustering/Association)**.**

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



**Fig 1.3 Data Mining Basic Diagram**

* **CHALLENGES AND LIMITATIONS OF DATA MINING**

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

**Key Challenges:**

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

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

## APPLICATIONS OF DATA MINING

### 1. Customer Relationship Management (CRM)

Data mining helps businesses analyze customer demographics, purchase history, and behavioral trendsto 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.

**Technical Requirements**

**Dataset**:

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

**Software/Tools**:

* Orange Tool
* SQL Server Management Studio
* SSDT
* MDX

**Hardware**:

* Processor
* RAM
* Hard Disk

**Models**:

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

## PROBLEM STATEMENT

Predicting AI tool usage patterns by analyzing user demographics, behavior, and preferences using classification models. It focuses on identifying key factors that influence the frequency and likelihood of AI tool adoption.

### This project aims to analyze and predict user behavior related to AI tool usage using classification models. It seeks to identify key factors influencing AI adoption across different user demographics. The goal is to build a predictive model that categorizes users based on their likelihood of using AI tools.

### Objectives:

* Identify significant predictors of AI tool usage.
* Classify users into distinct usage categories.
* Provide actionable insights into the factors driving or hindering AI adoption.

By leveraging **classification models**, the system will enable **better audience segmentation**, leading to improved **user satisfaction and tool optimization**.

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

**2.1 METHODOLOGY:**

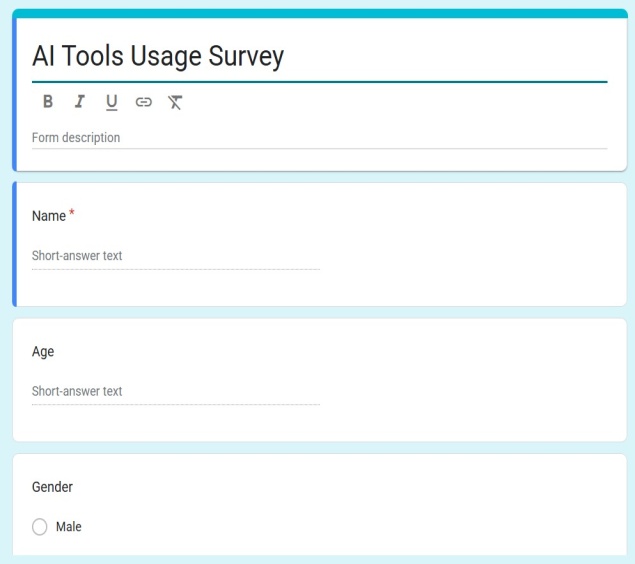
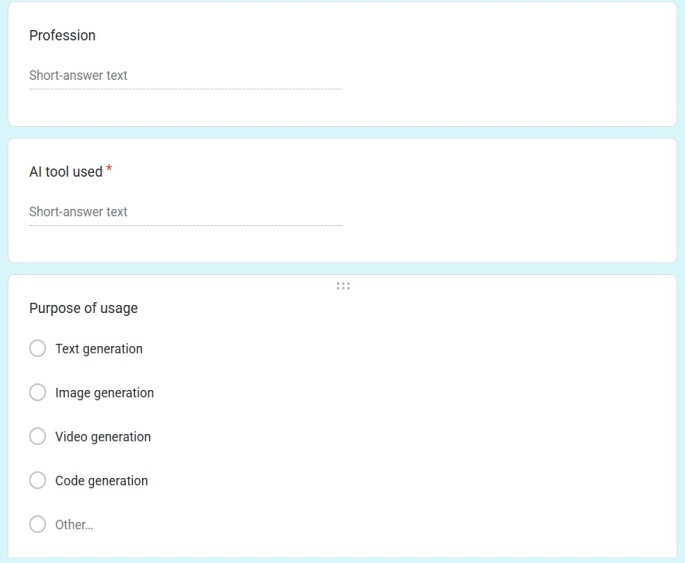
The KDD process is performed in step by step from collection of data set to the classification and developing the prediction model. There are some intermediatory steps in which we created all three schemas with the help of various tools like SSMS(SQL Server Management Services), Visual Studio and SSAS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

**Fig 2.1 Step by Step Process in KDD**

## STEP-1: COLLECTING & EXPLORING DATASET

### 1. Extracting the Form to Collect Information from Users

The dataset was created by gathering information on **AI Tool Usage** from various users. Data was collected through **Survey**.

**** ****

1. **(b)**

**Fig (a),(b) Images of Google Form**

### 2. Defining Survey or Data Collection Methods

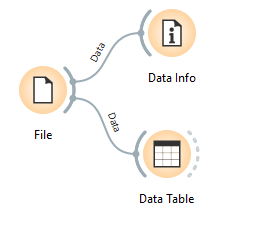
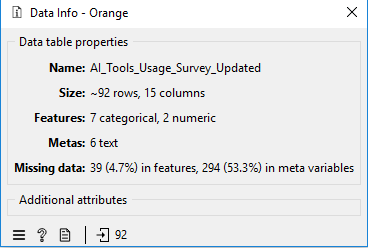
* **Online Surveys:** A structured questionnaire was distributed online, including **multiple-choice questions** to capture user preferences, platform choices, and listening habits.

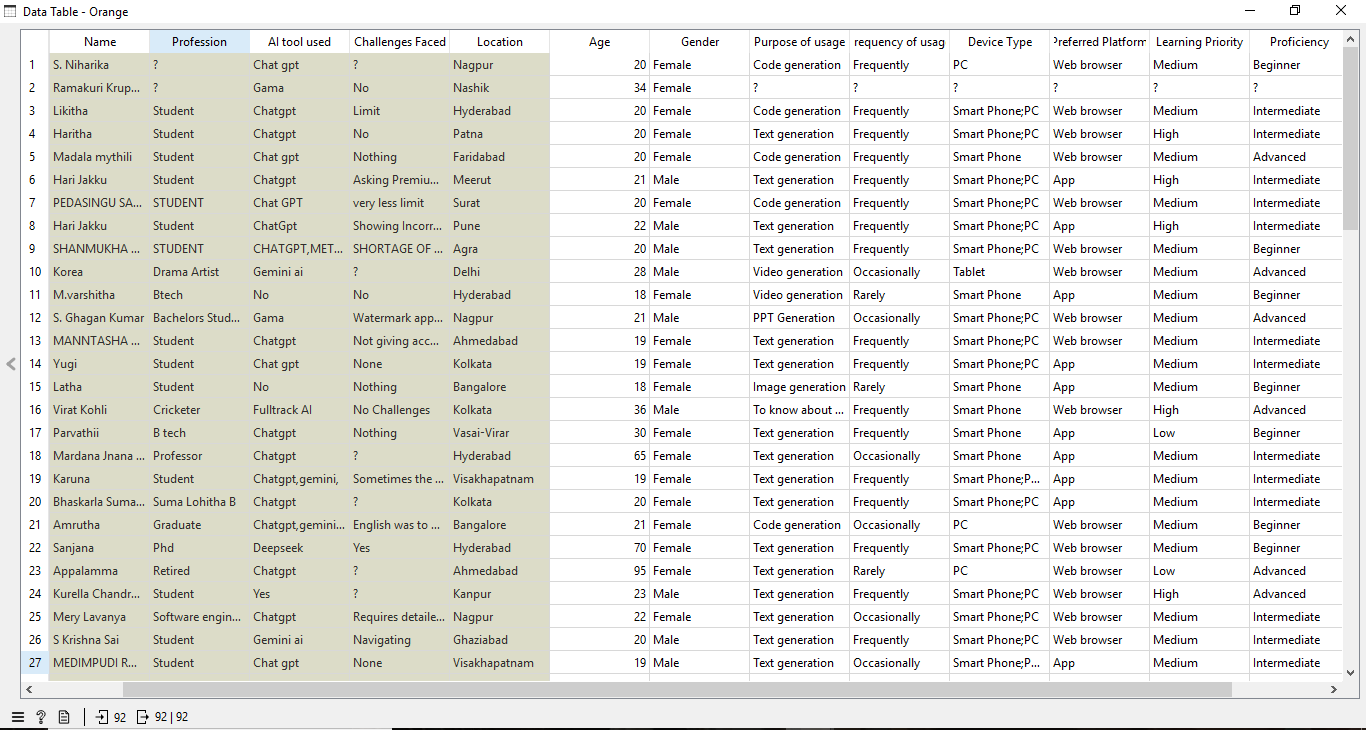
**Link:** [**https://docs.google.com/forms/AI-Tool usage Survey**](https://docs.google.com/forms/d/e/1FAIpQLSfFU5wlQijEpQZIdD0_0QB8tDETv7oyU247IDlJG9HyK4IPmA/viewform?usp=sharing)

**3. Choosing Attributes for Analysis**

The key attributes selected for analysis include:

* **AI tool used** – The specific AI tools that the respondent has used.
* **User Demographics**  – Includes **age, gender, and location** to analyze trends.
* **Purpose of usage**  – The primary reason or context in which the respondent uses AI tools.
* **Frequency of usage** – How often the respondent uses AI tools (e.g., daily, weekly).
* **Device Type**  – The type of device (e.g., laptop, smartphone) used to access AI tools.
* **Proficiency** – The respondent’s self-rated skill level in using AI tools.

** **

****

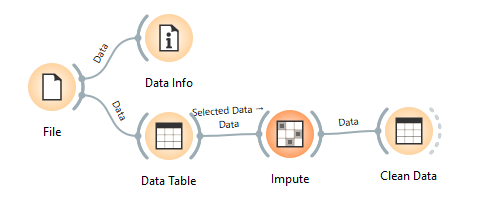
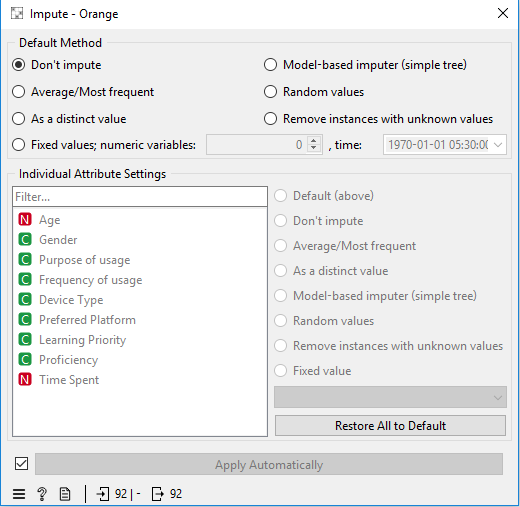
**Fig 2.1 Dataset Before Imputing and Preprocessing**

## Step-2: PREPROCESS THE DATA

## Preprocess the Dataset Using ORANGE TOOL

## 1. Handling Missing Values

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

**2. Data Cleaning & Transformation**

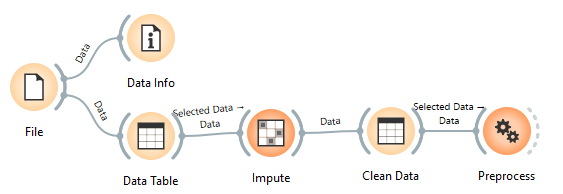
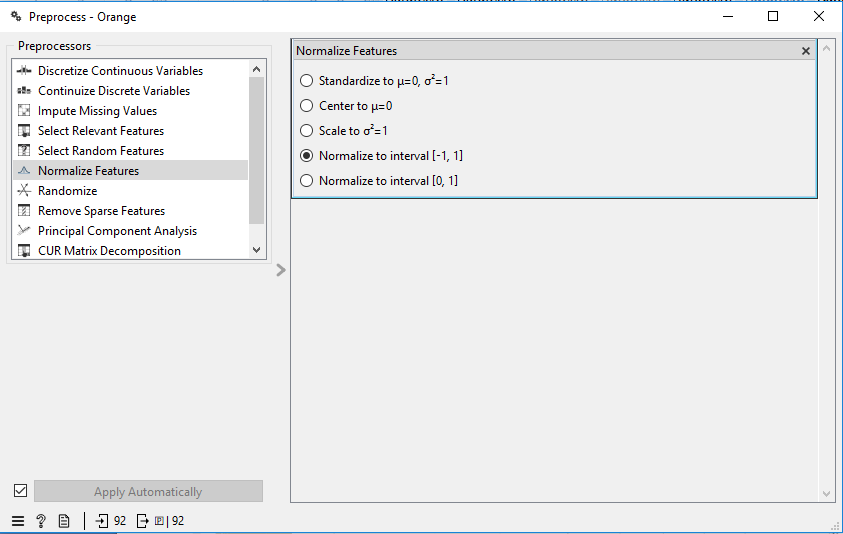
* Standardized text-based attributes.
* Converted categorical values into numerical formfor analysis.

**3. Removing Duplicates & Inconsistencies**

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

**4. Normalization & Encoding**

* **Categorical attributes** were encoded.
* **Normalization** was applied to standardize numerical values.

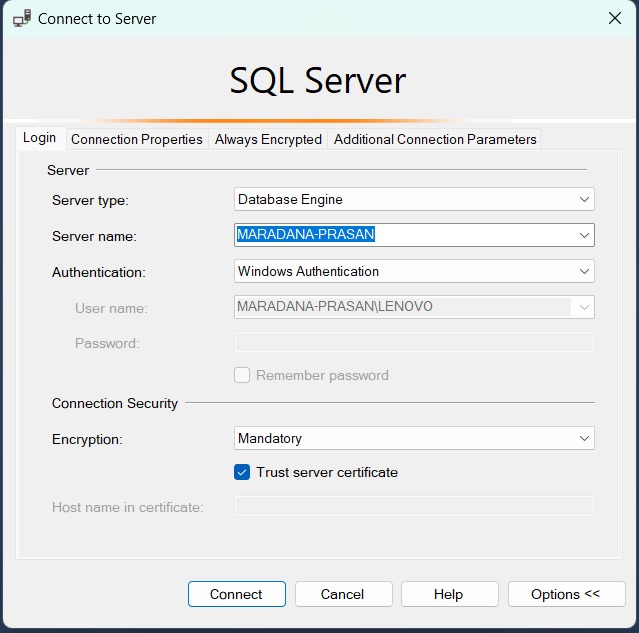
 



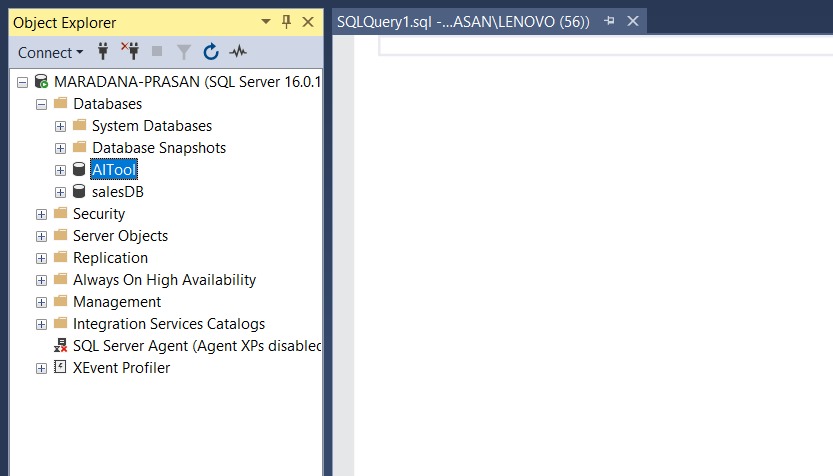
**Fig 2.2 Dataset After Imputing and Preprocessing**

**STEP-3: Creation of Database and Construction of OLAP Schemas**

* Generate SQL Queries for Schema Construction by Normalizing the collected Dataset (Using Database Engine).
* Designed SQL queries Star Schema, Snowflake Schema, and Fact Constellation Schema.
* SQL queries are separate to create each fact and dimension tables. Inserted data into tables using SQL in Database Engine and executed them in SSMS.

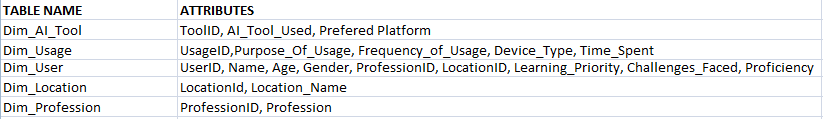


**Fig 2.3 Database Engine**

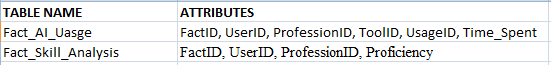
****

**Fig 2.4 Creation of DataBase**

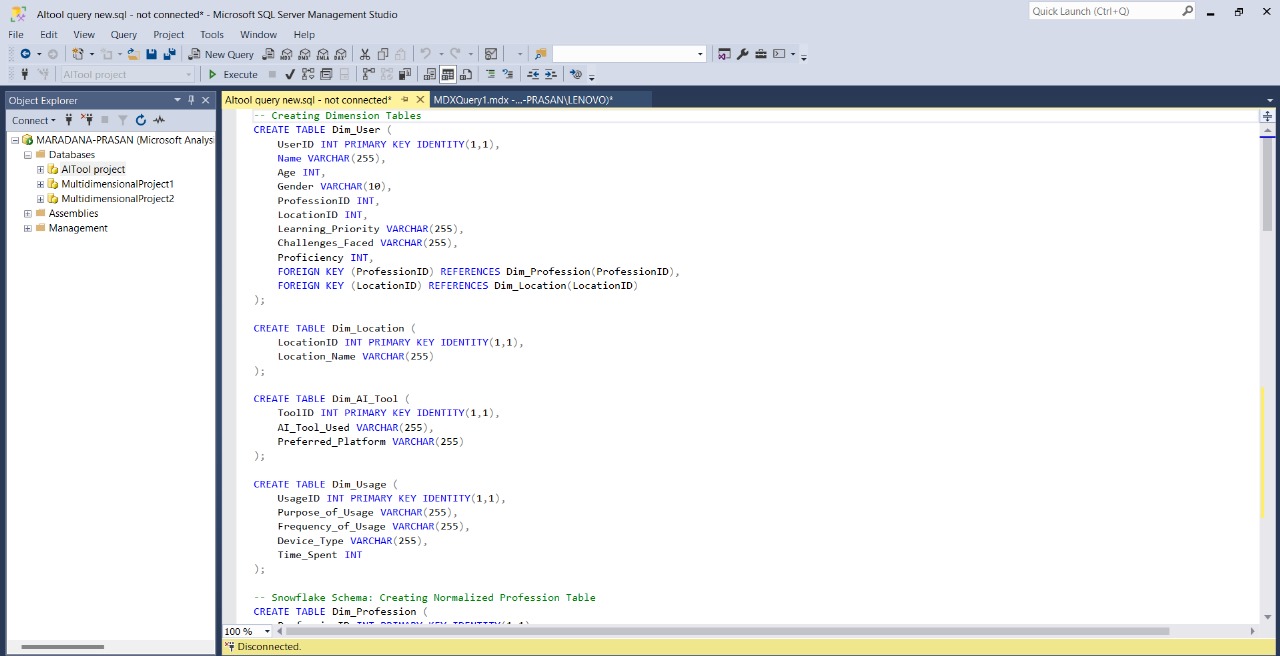
**3.1 Creation of Dimension Tables**

****

**3.2 Creation of Fact Tables**

****

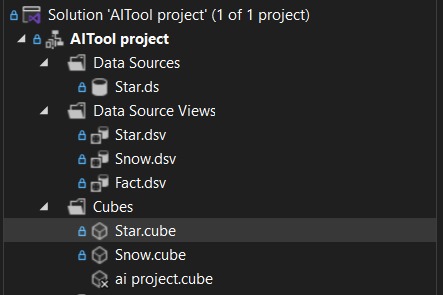
**3.3 Creating Tables & Inserting Values Using SQL Queries in AI Tool DataBase**



## Fig 2.5 SQL queries inserted for schema creation

**STEP-4: Visualize Schemas And Perform OLAP Operations using MDX Queries.**

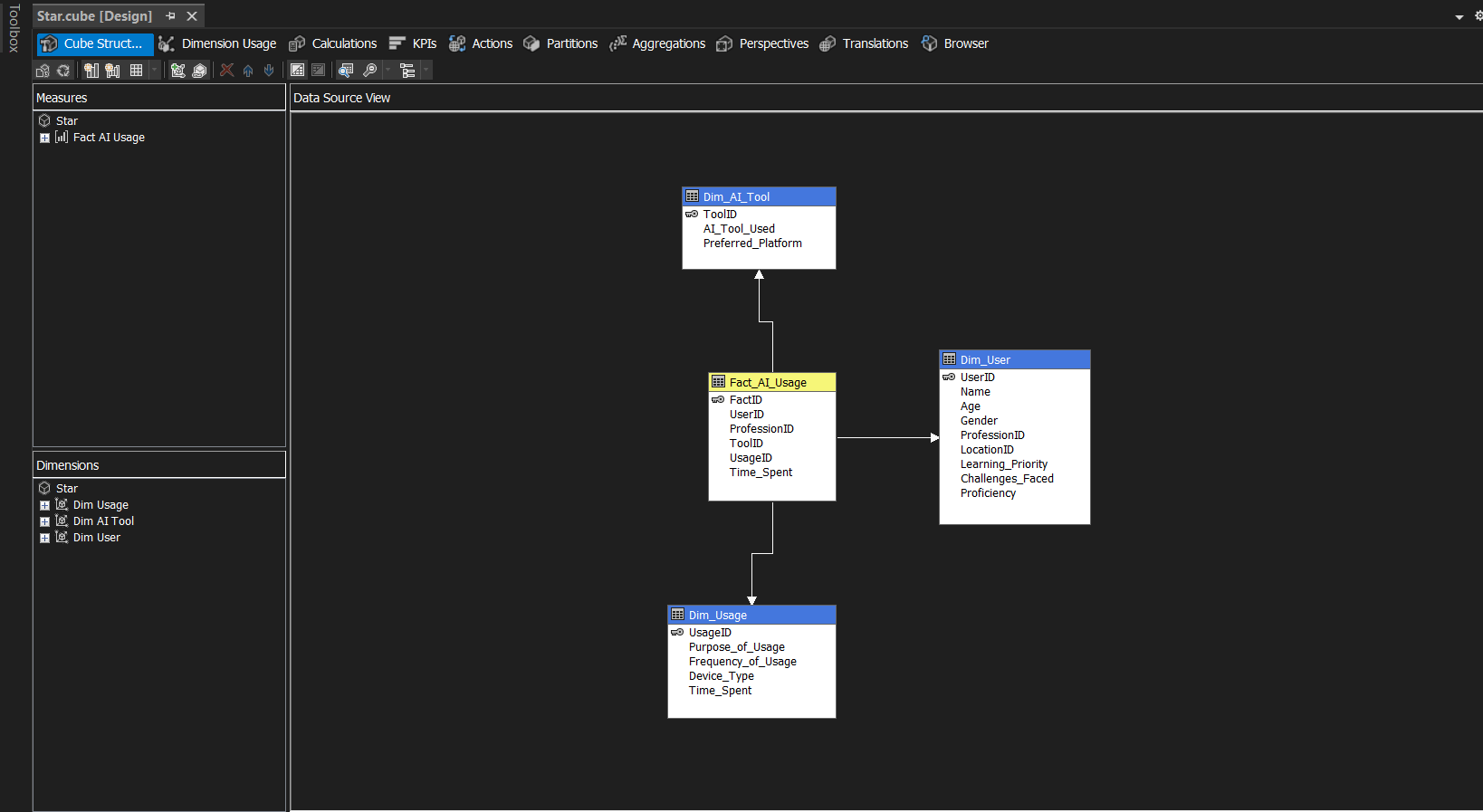
* Developed a multidimensional analysis project using Visual Studio.
* Configured Data Source & Data Source View, establishing connections to the database and defining table relationships.
* Designed database schema diagrams to visualize data structure.
* Validated table relationships to ensure data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

****

**1. STAR SCHEMA:**

Listener The **Star Schema** is a denormalized database schema used in OLAP, where a central **Fact Table** (containing measurable data Time Spent) is directly connected to multiple **Dimension Tables** (such as AI-tool, usage,) in a star-like structure.

* 1. **Design & Visualize the Schema**
* Create the **Star Schema** with Fact and Dimension tables.
* Define relationships between tables for efficient querying.



* 1. **Deploy the Data Warehouse & Load Data**
* Store structured data into the data warehouse.
* Ensure ETL (Extract, Transform, Load) processes are completed.
  1. **Create & Execute OLAP Queries**
* Write OLAP queries to perform data analysis.
* Use ROLLUP, CUBE, SLICE, DICE, DRILL-DOWN, and PIVOT operations for multi-dimensional analysis.
  1. **Perform OLAP Operations**
* Run the queries to process large datasets efficiently.
* Perform aggregations, filtering, and transformations on the stored data.

## MDX Queries OLAP operations in STAR SCHEMA:

1. Aggregate AI Usage purposes into 'Business AI' and 'Research AI' categories and compute the total Time Spent, average Time Spent, and total usage count for each.

**Concept Hierarchy Used:**

**Purpose of Usage**

**Business AI Research AI**

**AI Tools AI Tools**

**Roll-Up (Aggregation of Time):**

WITH

MEMBER [Dim Usage].[Purpose Of Usage].[Business AI] AS

SUM(

{

[Dim Usage].[Purpose Of Usage].[Automation],

[Dim Usage].[Purpose Of Usage].[Data Analysis],

[Dim Usage].[Purpose Of Usage].[AI in Finance]

}

)

MEMBER [Dim Usage].[Purpose Of Usage].[Research AI] AS

SUM(

{

[Dim Usage].[Purpose Of Usage].[Machine Learning],

[Dim Usage].[Purpose Of Usage].[Deep Learning],

[Dim Usage].[Purpose Of Usage].[AI in Healthcare]

}

)

MEMBER [Measures].[Avg Time Spent] AS

AVG(

{

[Dim Usage].[Purpose Of Usage].[Business AI],

[Dim Usage].[Purpose Of Usage].[Research AI]

},

[Measures].[Time Spent]

)

MEMBER [Measures].[Total Usage Count] AS

COUNT(

{

[Dim Usage].[Purpose Of Usage].[Business AI],

[Dim Usage].[Purpose Of Usage].[Research AI]

}

)

SELECT

{[Measures].[Time Spent], [Measures].[Avg Time Spent], [Measures].[Total Usage Count]} ON COLUMNS,

{

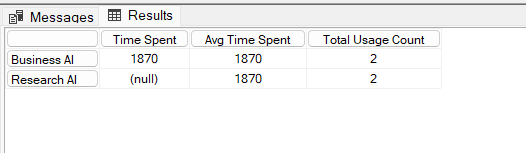
[Dim Usage].[Purpose Of Usage].[Business AI],

[Dim Usage].[Purpose Of Usage].[Research AI]

} ON ROWS

FROM [Star];

**Output:**

****

**Execution Time:** 1 ms

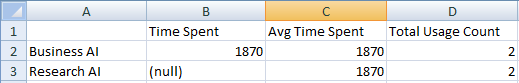
**Visualize OLAP Results**

Visualizing AI-Tool Preferences Using Orange Tool.

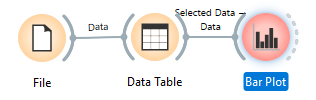
#### Prepare OLAP Output for Visualization

* Select key OLAP operation results related to AI-Tool preferences.
* Export the selected data as an **Excel sheet** for further visualization.
* Ensure the dataset includes relevant attributes such as **Time** **Spent**, **user preferences**, **Proficiency.**

**Roll-Up.csv:**

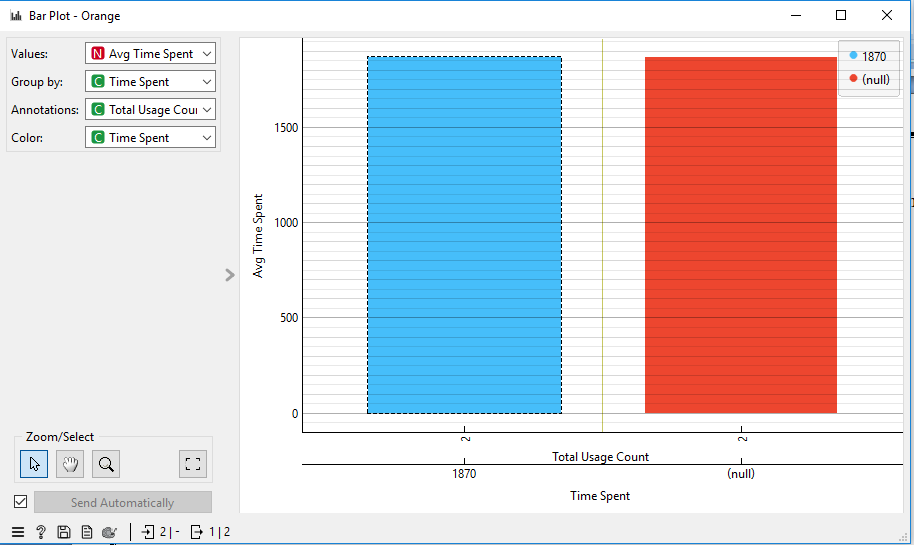


* To compare user engagement across different AI-Tools based on **Total usage Count** and **Avg.time spent**



#### Bar Chart Configuration:

* **X-Axis**: **Total usage count**
* **Y-Axis:** Average Time Spent
* **Grouped By:** Time Spent

****

**(B)**Time is spent using each AI tool, broken down in hours, minutes, and seconds.

**Concept Hierarchy Used:**

**Time in:**

**Hours**

**Minutes**

**Seconds**

**Drill-Down (Detailed Time Analysis):**

WITH

MEMBER [Measures].[Time in Hours] AS

([Measures].[Time Spent]) / 3600

MEMBER [Measures].[Time in Minutes] AS

([Measures].[Time Spent]) / 60

MEMBER [Measures].[Time in Seconds] AS

[Measures].[Time Spent]

SELECT

{

[Measures].[Time in Hours],

[Measures].[Time in Minutes],

[Measures].[Time in Seconds]

} ON COLUMNS,

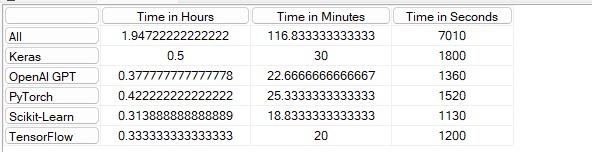
NONEMPTY(

[Dim AI Tool].[AI Tool Used].MEMBERS

) ON ROWS

FROM [Star];

**Output:**



**Execution Time:** 1 ms

1. Analyze the usage of AI tools by filtering only for users who use Laptops as their Device Type.

Display the total Time Spent, categorized by AI Tool Used, User Profession, and Gender."

**Slice (Filtering by Tool):**

WITH

MEMBER [Measures].[Selected AI Tool] AS

[Dim AI Tool].[AI Tool Used].CURRENTMEMBER.MEMBER\_CAPTION

MEMBER [Measures].[Selected Profession] AS

[Dim User].[Profession ID].CURRENTMEMBER.MEMBER\_CAPTION

MEMBER [Measures].[Selected Gender] AS

[Dim User].[Gender].CURRENTMEMBER.MEMBER\_CAPTION

SELECT

{

[Measures].[Time Spent],

[Measures].[Selected AI Tool],

[Measures].[Selected Profession],

[Measures].[Selected Gender]

}

ON COLUMNS,

NONEMPTY(

[Dim AI Tool].[AI Tool Used].[AI Tool Used].MEMBERS \*

[Dim User].[Profession ID].[Profession ID].MEMBERS \*

[Dim User].[Gender].[Gender].MEMBERS

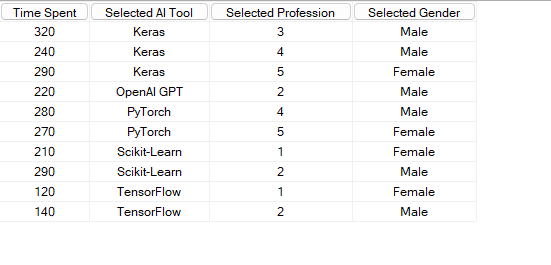
)

ON ROWS

FROM [Star]

WHERE ([Dim Usage].[Device Type].[Laptop]);

**Output:**



**Execution Time:** 1 ms

**(D)**Provide a breakdown of AI tool usage, displaying the total Time Spent along with AI Tool Used,

User Gender, and Device Type. Ensure that the columns are properly labeled for clarity.

**Pivot (Rearranging Dimensions):**

WITH

MEMBER [Measures].[Selected AI Tool] AS

[Dim AI Tool].[AI Tool Used].CURRENTMEMBER.MEMBER\_CAPTION

MEMBER [Measures].[Selected Gender] AS

[Dim User].[Gender].CURRENTMEMBER.MEMBER\_CAPTION

MEMBER [Measures].[Selected Device Type] AS

[Dim Usage].[Device Type].CURRENTMEMBER.MEMBER\_CAPTION

SELECT

{

[Measures].[Time Spent],

[Measures].[Selected AI Tool],

[Measures].[Selected Gender],

[Measures].[Selected Device Type]

} ON COLUMNS,

NONEMPTY(

[Dim AI Tool].[AI Tool Used].MEMBERS \*

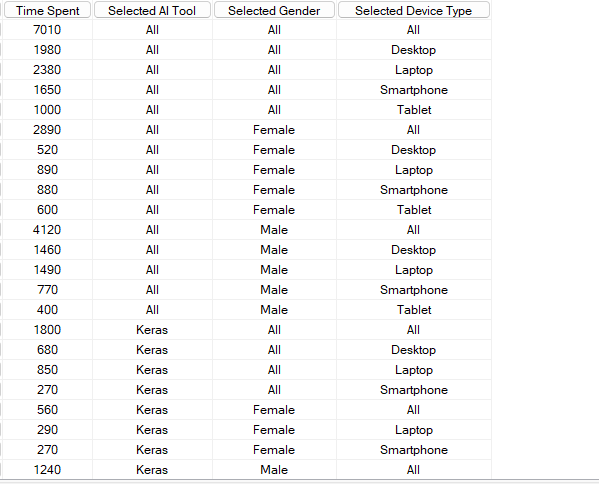
[Dim User].[Gender].MEMBERS \*

[Dim Usage].[Device Type].MEMBERS

) ON ROWS

FROM [Star];

**Output:**

****

**Execution Time:** 1 ms

**(E)**Analyze the AI tool usage by different users. Show the total Time Spent on AI tools, categorized

by AI Tool Used, User Gender, Profession, and Device Type (Laptop and Smartphone)."

**Dice (Filtering Across Multiple Dimensions):**

WITH

MEMBER [Measures].[Selected AI Tool] AS

[Dim AI Tool].[AI Tool Used].CURRENTMEMBER.MEMBER\_CAPTION

MEMBER [Measures].[Selected Gender] AS

[Dim User].[Gender].CURRENTMEMBER.MEMBER\_CAPTION

MEMBER [Measures].[Selected Profession] AS

[Dim User].[Profession ID].CURRENTMEMBER.MEMBER\_CAPTION

SELECT

{

[Measures].[Time Spent],

[Measures].[Selected AI Tool],

[Measures].[Selected Gender],

[Measures].[Selected Profession]

} ON COLUMNS,

NONEMPTY(

CROSSJOIN(

{[Dim AI Tool].[AI Tool Used].MEMBERS},

{[Dim User].[Gender].MEMBERS},

{[Dim User].[Profession ID].MEMBERS}

)

) ON ROWS

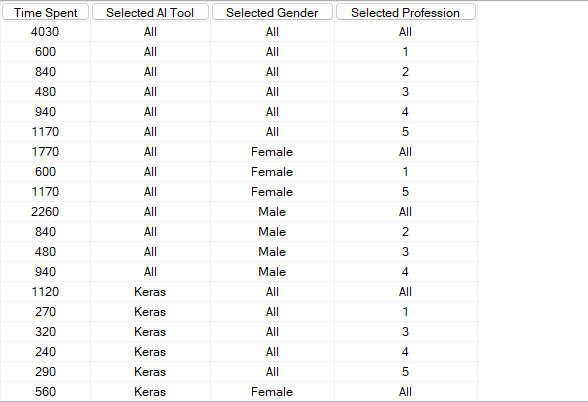
FROM [Star]

WHERE (

{[Dim Usage].[Device Type].[Laptop], [Dim Usage].[Device Type].[Smartphone]}

);

**Output:**



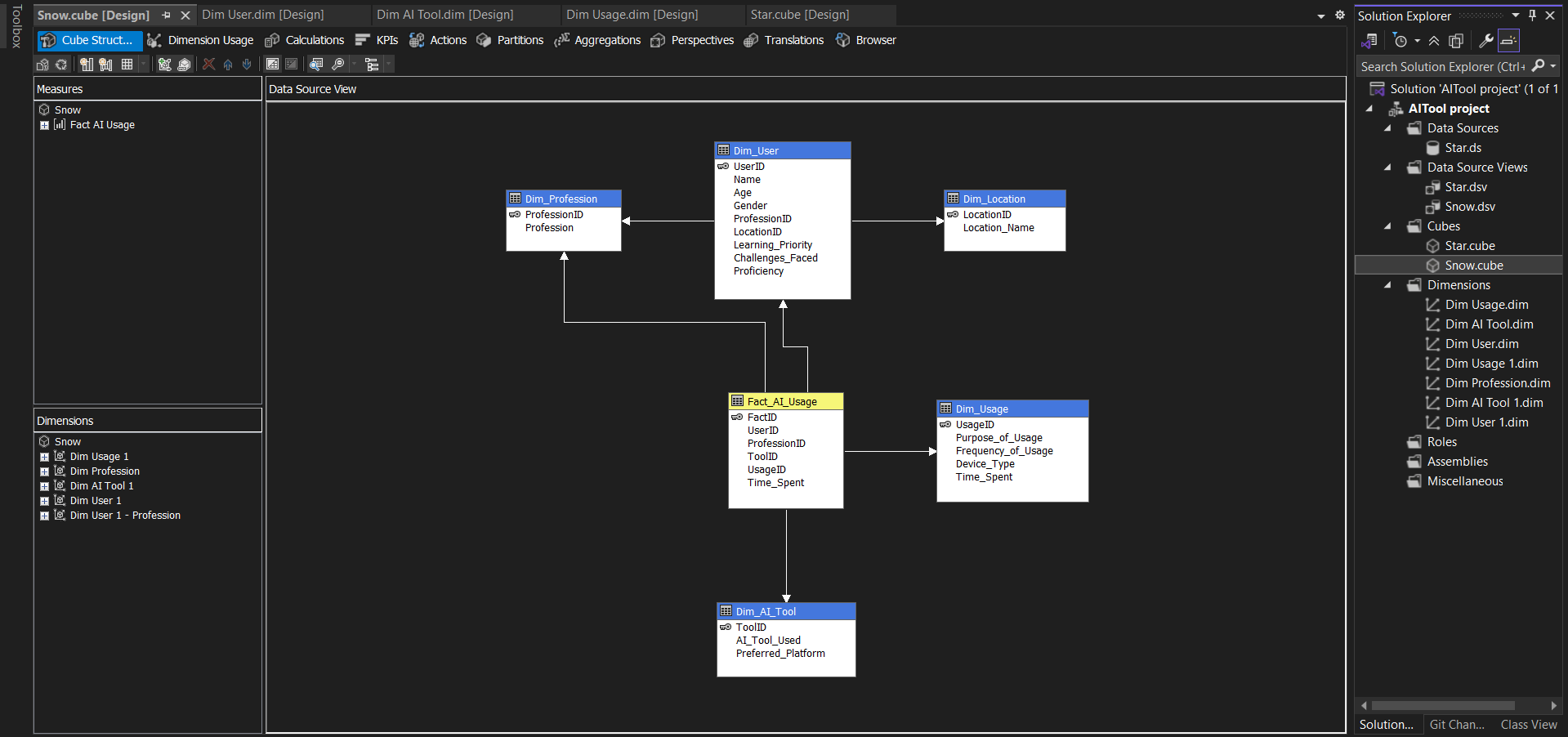
**Execution Time:** 1 ms

**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:

**2.1 Design & Visualize the Snowflake Schema**

* Identify **Fact Tables** (e.g., AI-usage)
* Identify **Dimension Tables** (e.g., User, Tool).
* Normalize dimension tables by breaking them into **sub-dimensions** (e.g., User → Location ).
* Ensure **foreign key relationships** between tables.



**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.

**2.3 Create & Execute OLAP Queries**

Write **SQL queries** for analytical processing:

* **ROLLUP** – Aggregate data across different levels.
* **CUBE** – Compute multi-dimensional aggregates.
* **DRILL-DOWN** – View data at finer granularity.
* **SLICE & DICE** – Filter and analyze subsets of data.

**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)**Summarize AI tool usage by categorizing Purpose of Usage into Business AI and Research AI. Show

the total and average Time Spent for each category."

**Concept Hierarchy Used:**

**Purpose of Usage**

**Business AI Research AI**

**AI Tools AI Tools**

### Roll-Up (Aggregation of Tools):

WITH

MEMBER [Dim Usage 1].[Purpose Of Usage].[Business AI] AS

SUM(

{

[Dim Usage 1].[Purpose Of Usage].[Automation],

[Dim Usage 1].[Purpose Of Usage].[Data Analysis],

[Dim Usage 1].[Purpose Of Usage].[AI in Finance]

}

)

MEMBER [Dim Usage 1].[Purpose Of Usage].[Research AI] AS

SUM(

{

[Dim Usage 1].[Purpose Of Usage].[Machine Learning],

[Dim Usage 1].[Purpose Of Usage].[Deep Learning],

[Dim Usage 1].[Purpose Of Usage].[AI in Healthcare]

}

)

MEMBER [Measures].[Avg Time Spent] AS

AVG(

{

[Dim Usage 1].[Purpose Of Usage].[Business AI],

[Dim Usage 1].[Purpose Of Usage].[Research AI]

},

[Measures].[Time Spent]

)

MEMBER [Measures].[Total Usage Count] AS

COUNT(

{

[Dim Usage 1].[Purpose Of Usage].[Business AI],

[Dim Usage 1].[Purpose Of Usage].[Research AI]

}

)

SELECT

{ [Measures].[Time Spent], [Measures].[Avg Time Spent], [Measures].[Total Usage Count] } ON COLUMNS,

{

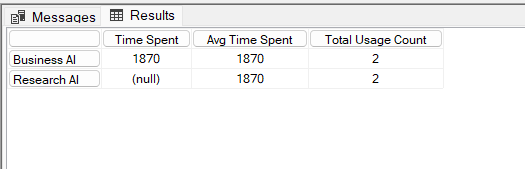
[Dim Usage 1].[Purpose Of Usage].[Business AI],

[Dim Usage 1].[Purpose Of Usage].[Research AI]

} ON ROWS

FROM [Snow];

**Output:**

****

**Execution Time:** 1 ms

1. Analysis of Time Spent by City and Street Location

**Concept Hierarchy Used:**

**Location**

**City**

**Street**

**Drill-Down (Time Analysis):**

WITH

MEMBER [Measures].[City] AS

[Dim User 1].[Location Name].CURRENTMEMBER.MEMBER\_CAPTION

MEMBER [Measures].[Street] AS

[Dim User 1].[Location ID].CURRENTMEMBER.MEMBER\_CAPTION

SELECT

{

[Measures].[Time Spent],

[Measures].[City],

[Measures].[Street]

} ON COLUMNS,

NONEMPTY(

CROSSJOIN(

[Dim User 1].[Location Name].Children,

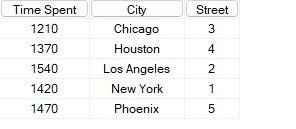
[Dim User 1].[Location ID].Children

)

) ON ROWS

FROM [Snow];

**Output:**

****

**Execution Time:** 1 ms

**(C)**"Retrieve Time Spent on AI tools for users using Laptops, categorized by AI Tool Used,

Profession, and Gender."

**Slice (Filtering by Profession):**

WITH

MEMBER [Measures].[Selected AI Tool] AS

[Dim AI Tool 1].[AI Tool Used].CURRENTMEMBER.MEMBER\_CAPTION

MEMBER [Measures].[Selected Profession] AS

[Dim Profession].[Profession ID].CURRENTMEMBER.MEMBER\_CAPTION

MEMBER [Measures].[Selected Gender] AS

[Dim User 1].[Gender].CURRENTMEMBER.MEMBER\_CAPTION

SELECT

{

[Measures].[Time Spent],

[Measures].[Selected AI Tool],

[Measures].[Selected Profession],

[Measures].[Selected Gender]

} ON COLUMNS,

NONEMPTY(

CROSSJOIN(

[Dim AI Tool 1].[AI Tool Used].MEMBERS,

[Dim Profession].[Profession ID].MEMBERS,

[Dim User 1].[Gender].MEMBERS

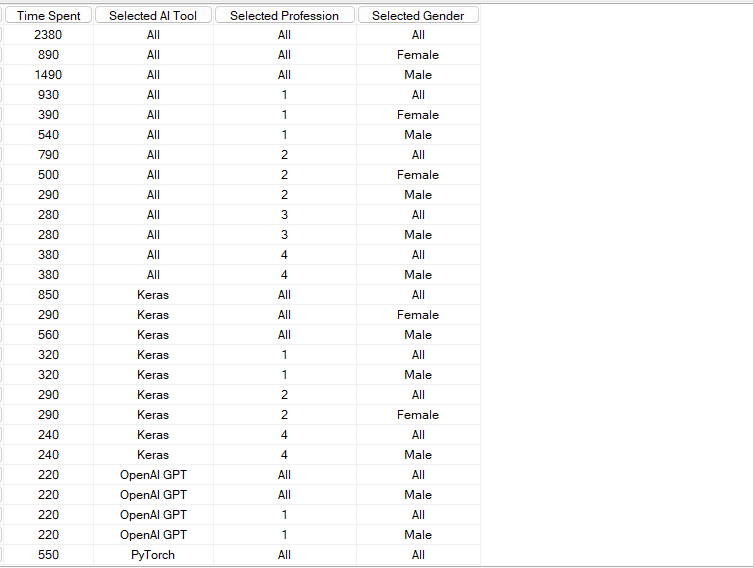
)

) ON ROWS

FROM [Snow]

WHERE ([Dim Usage 1].[Device Type].[Laptop]);

**Output:**

****

**Execution Time:** 1 ms

**(D)**"Analyze Time Spent on AI tools for users by AI Tool Used, Gender, Profession, and selected Device

Types (Laptop and Smartphone)."

**Dice (Filtering by Multiple Dimensions):**

WITH

MEMBER [Measures].[Selected AI Tool] AS

[Dim AI Tool 1].[AI Tool Used].CURRENTMEMBER.MEMBER\_CAPTION

MEMBER [Measures].[Selected Gender] AS

[Dim User 1].[Gender].CURRENTMEMBER.MEMBER\_CAPTION

MEMBER [Measures].[Selected Profession] AS

[Dim Profession].[Profession ID].CURRENTMEMBER.MEMBER\_CAPTION

MEMBER [Measures].[Selected Device Type] AS

[Dim Usage 1].[Device Type].CURRENTMEMBER.MEMBER\_CAPTION

SELECT

{

[Measures].[Time Spent],

[Measures].[Selected AI Tool],

[Measures].[Selected Gender],

[Measures].[Selected Profession],

[Measures].[Selected Device Type]

} ON COLUMNS,

NONEMPTY(

CROSSJOIN(

[Dim AI Tool 1].[AI Tool Used].MEMBERS,

[Dim User 1].[Gender].MEMBERS,

[Dim Profession].[Profession ID].MEMBERS,

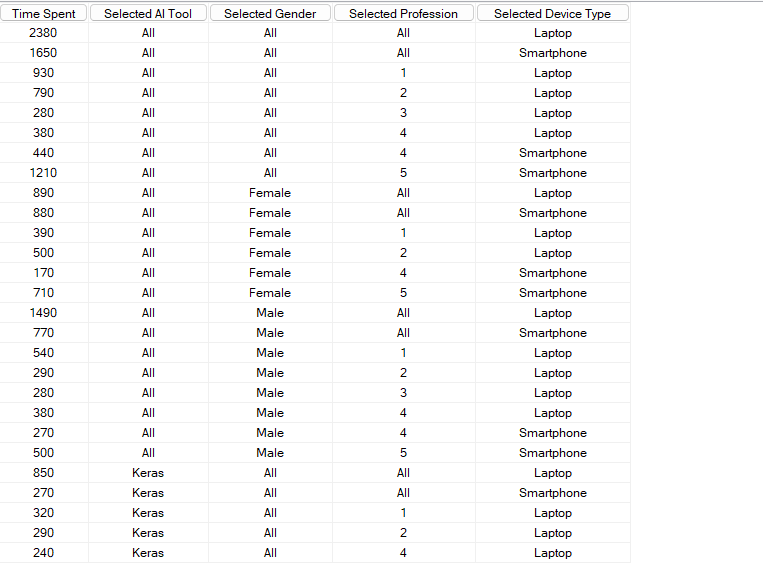
{[Dim Usage 1].[Device Type].[Laptop], [Dim Usage 1].[Device Type].[Smartphone]}

)

) ON ROWS

FROM [Snow];

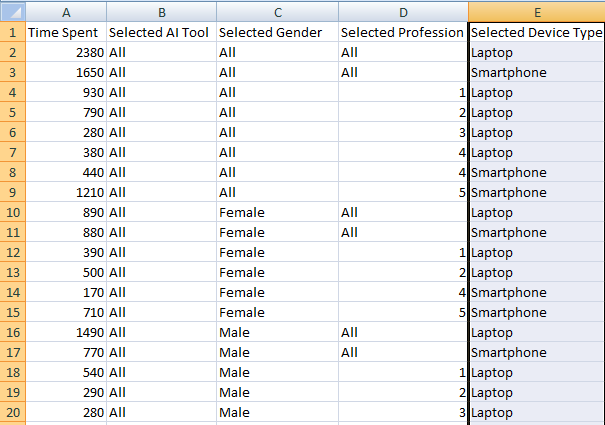
**Output:**

****

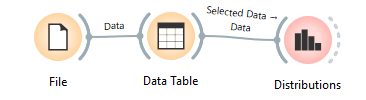
**Execution Time:** 1 ms

**Visualize OLAP Results**

**Dice(snow).csv:**

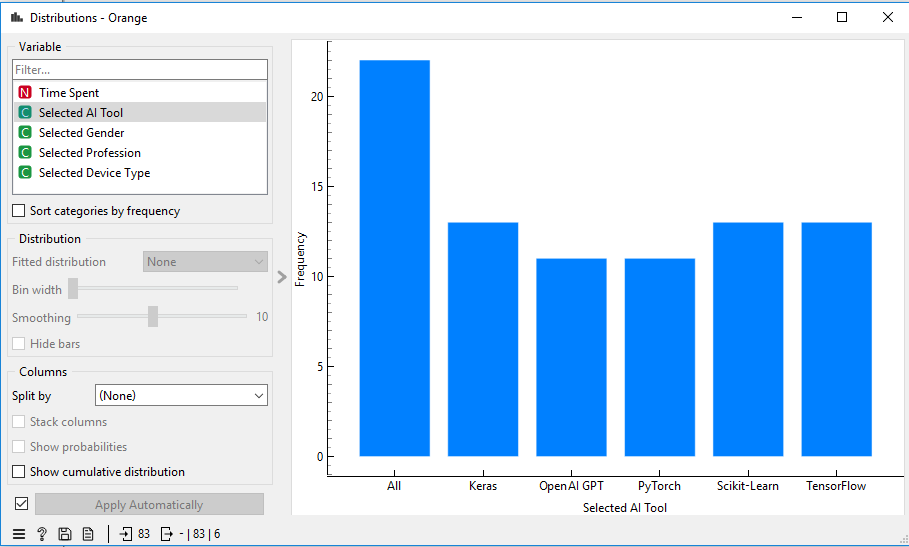
****

* To compare user engagement across different AI-Tools based on **Gender** and **Time spent.**

****

#### Distributions Chart Configuration:

* **X-Axis**:AI Tool
* **Y-Axis:** Frequency

****

**(E)**"Pivot AI Tool Used against Device Type and User Profession to analyze Time Spent."

**Pivot (Rearranging Dimensions):**

WITH

MEMBER [Measures].[AI Tool Used] AS

[Dim AI Tool 1].[AI Tool Used].CURRENTMEMBER.MEMBER\_CAPTION

MEMBER [Measures].[Device Type] AS

[Dim Usage 1].[Device Type].CURRENTMEMBER.MEMBER\_CAPTION

MEMBER [Measures].[User Profession] AS

[Dim Profession].[Profession ID].CURRENTMEMBER.MEMBER\_CAPTION

SELECT

{

[Measures].[Time Spent],

[Measures].[AI Tool Used],

[Measures].[Device Type],

[Measures].[User Profession]

} ON COLUMNS,

NONEMPTY(

CROSSJOIN(

[Dim AI Tool 1].[AI Tool Used].MEMBERS,

[Dim Usage 1].[Device Type].MEMBERS,

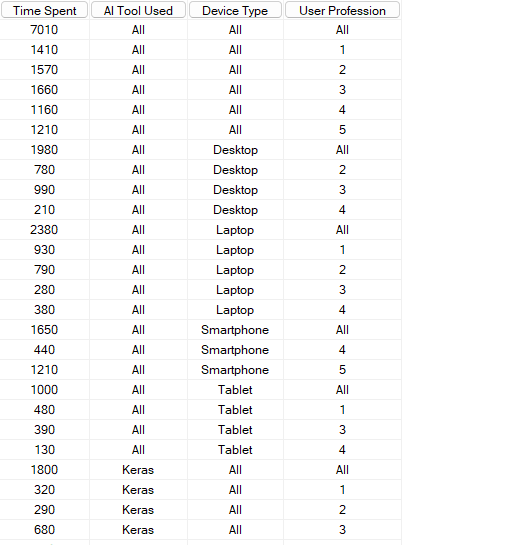
[Dim Profession].[Profession ID].MEMBERS

)

) ON ROWS

FROM [Snow];

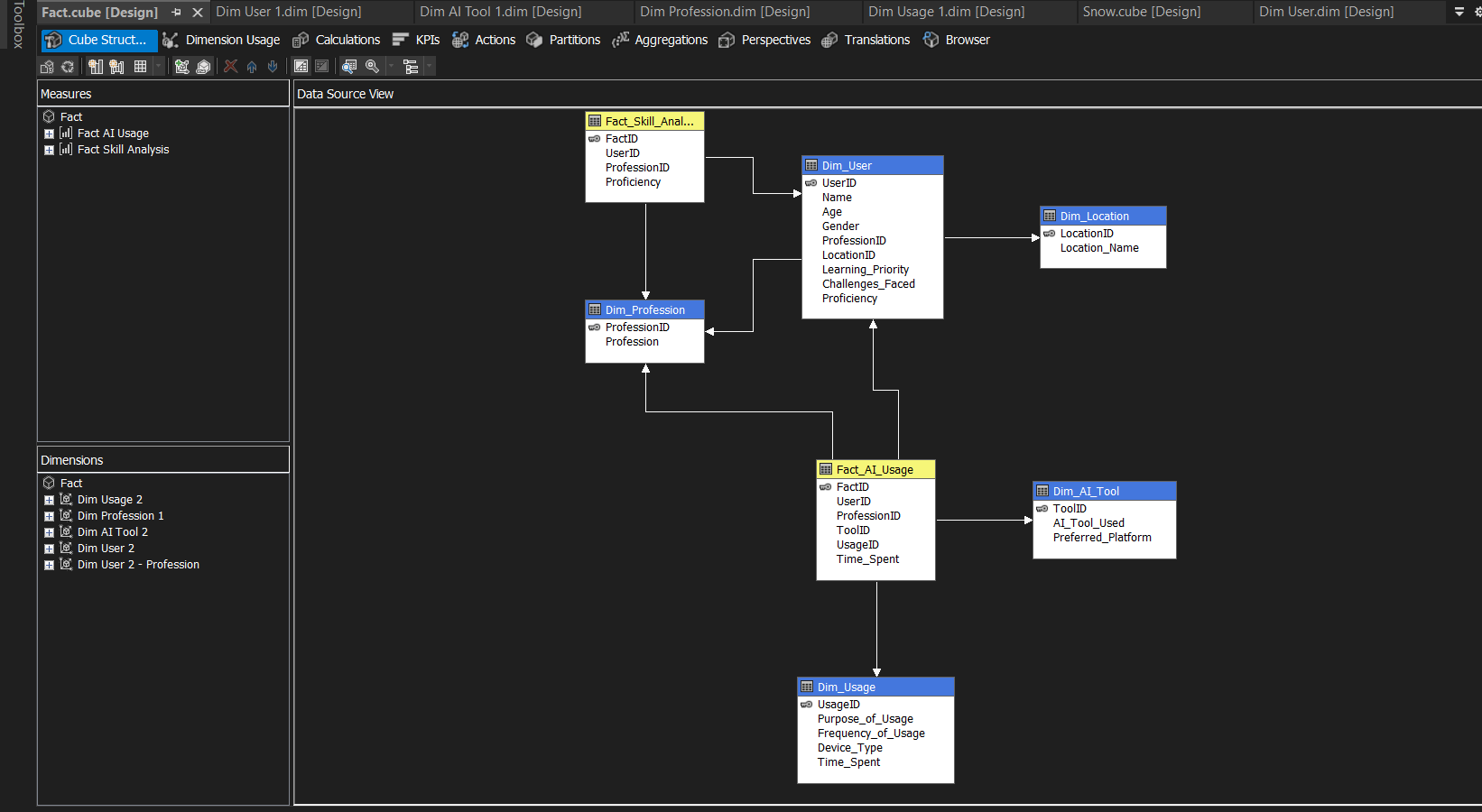
**Output:**

****

**Execution Time:** 1 ms

**3. FACT CONSTELLETION SCHEMA:**

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.



**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.

**3.2 Create & Execute OLAP Queries**

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

**3.3 Perform OLAP Operations**

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

**MDX Queries OLAP operations in FACTCONSTELLATION SCHEMA:**

1. Aggregate AI tool usage into ‘Web-Based AI’ and ‘Desktop AI’ categories based on Preferred

Platform. Compute the total Time Spent, average Proficiency, and total tool usage count for each category.

**Concept Hierarchy Used:**

**Purpose of Usage**

**Business AI Research AI**

**AI Tools AI Tools**

**Roll-Up (Aggregation by Usage Count):**

WITH

MEMBER [Dim AI Tool 2].[Preferred Platform].[Web-Based AI] AS

SUM(

{

[Dim AI Tool 2].[Preferred Platform].[Cloud],

[Dim AI Tool 2].[Preferred Platform].[Online]

}

)

MEMBER [Dim AI Tool 2].[Preferred Platform].[Desktop AI] AS

SUM(

{

[Dim AI Tool 2].[Preferred Platform].[Windows],

[Dim AI Tool 2].[Preferred Platform].[Mac],

[Dim AI Tool 2].[Preferred Platform].[Linux]

}

)

MEMBER [Measures].[Avg Proficiency] AS

AVG(

{

[Dim AI Tool 2].[Preferred Platform].[Web-Based AI],

[Dim AI Tool 2].[Preferred Platform].[Desktop AI]

},

[Measures].[Proficiency]

)

MEMBER [Measures].[Total Usage Count] AS

COUNT(

{

[Dim AI Tool 2].[Preferred Platform].[Web-Based AI],

[Dim AI Tool 2].[Preferred Platform].[Desktop AI]

}

)

SELECT

{[Measures].[Time Spent], [Measures].[Avg Proficiency], [Measures].[Total Usage Count]} ON COLUMNS,

{

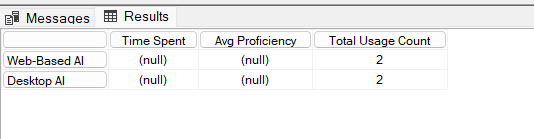
[Dim AI Tool 2].[Preferred Platform].[Web-Based AI],

[Dim AI Tool 2].[Preferred Platform].[Desktop AI]

} ON ROWS

FROM [Fact];

**Output:**

****

**Execution Time:** 1 ms

1. Total time spent (in hours, minutes, and seconds) on different AI tools distributed across gender and

Profession

**Concept Hierarchy Used:**

**Time Spent in:**

**Hours**

**Minutes**

**Seconds**

**Drill-Down (Time Analysis):**

WITH

MEMBER [Measures].[Time in Hours] AS

[Measures].[Time Spent] / 3600

MEMBER [Measures].[Time in Minutes] AS

([Measures].[Time Spent] - (INT([Measures].[Time Spent] / 3600) \* 3600)) / 60

MEMBER [Measures].[Time in Seconds] AS

[Measures].[Time Spent] -

(INT([Measures].[Time Spent] / 60) \* 60)

MEMBER [Measures].[Selected AI Tool] AS

[Dim AI Tool 2].[AI Tool Used].CURRENTMEMBER.MEMBER\_CAPTION

MEMBER [Measures].[Selected Gender] AS

[Dim User 2].[Gender].CURRENTMEMBER.MEMBER\_CAPTION

MEMBER [Measures].[Selected Profession] AS

[Dim User 2 - Profession].[Profession].CURRENTMEMBER.MEMBER\_CAPTION

SELECT

{

[Measures].[Time in Hours],

[Measures].[Time in Minutes],

[Measures].[Time in Seconds],

[Measures].[Selected AI Tool],

[Measures].[Selected Gender],

[Measures].[Selected Profession]

}

ON COLUMNS,

NONEMPTY(

[Dim AI Tool 2].[AI Tool Used].[AI Tool Used].MEMBERS \*

[Dim User 2].[Gender].[Gender].MEMBERS \*

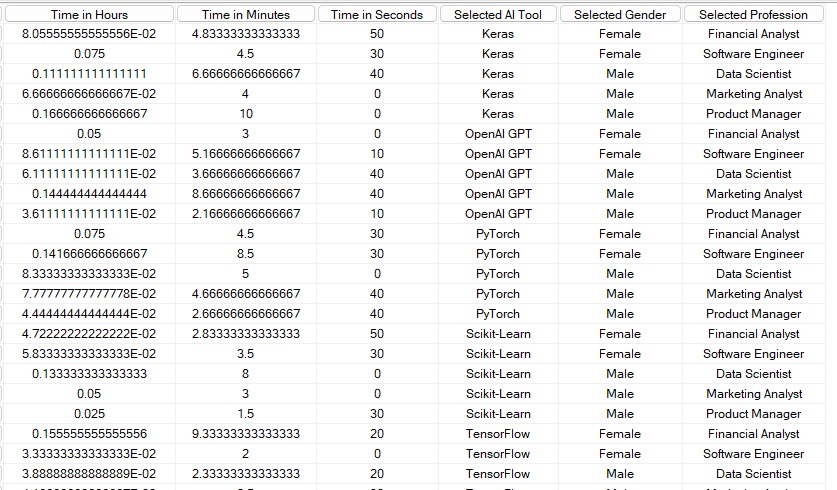
[Dim User 2 - Profession].[Profession].[Profession].MEMBERS

)

ON ROWS

FROM [Fact];

**Output:**

****

**Execution Time:** 1 ms

**(C)**Slice the data to display Time Spent, along with selected AI Tool Used, Profession, and Gender, only

for users who accessed AI tools using a **Smartphone**.

**Slice (Filtering):**

WITH

MEMBER [Measures].[Selected AI Tool] AS

[Dim AI Tool 2].[AI Tool Used].CURRENTMEMBER.MEMBER\_CAPTION

MEMBER [Measures].[Selected Profession] AS

[Dim User 2 - Profession].[Profession].CURRENTMEMBER.MEMBER\_CAPTION

MEMBER [Measures].[Selected Gender] AS

[Dim User 2].[Gender].CURRENTMEMBER.MEMBER\_CAPTION

SELECT

{

[Measures].[Time Spent],

[Measures].[Selected AI Tool],

[Measures].[Selected Profession],

[Measures].[Selected Gender]

}

ON COLUMNS,

NONEMPTY(

[Dim AI Tool 2].[AI Tool Used].[AI Tool Used].MEMBERS \*

[Dim User 2 - Profession].[Profession].[Profession].MEMBERS \*

[Dim User 2].[Gender].[Gender].MEMBERS

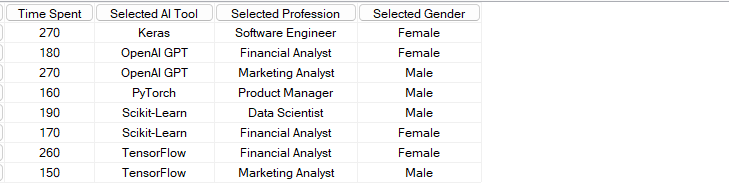
)

ON ROWS

FROM [Fact]

WHERE ([Dim Usage 2].[Device Type].[Smartphone]);

**Output:**

****

**Execution Time:** 1 ms

**(D)**Filter the dataset to display Time Spent, AI Tool Used, Gender, and Profession for users who

accessed AI tools using either **Desktop** or **Smartphone**, and whose **Proficiency Level is Advanced**.

**Dice (Filtering by Country and Language):**

WITH

MEMBER [Measures].[Selected AI Tool] AS

[Dim AI Tool 2].[AI Tool Used].CURRENTMEMBER.MEMBER\_CAPTION

MEMBER [Measures].[Selected Gender] AS

[Dim User 2].[Gender].CURRENTMEMBER.MEMBER\_CAPTION

MEMBER [Measures].[Selected Profession] AS

[Dim User 2 - Profession].[Profession].CURRENTMEMBER.MEMBER\_CAPTION

SELECT

{

[Measures].[Time Spent],

[Measures].[Selected AI Tool],

[Measures].[Selected Gender],

[Measures].[Selected Profession]

} ON COLUMNS,

NONEMPTY(

CROSSJOIN(

{[Dim AI Tool 2].[AI Tool Used].MEMBERS},

{[Dim User 2].[Gender].MEMBERS},

{[Dim User 2 - Profession].[Profession].MEMBERS}

),

[Measures].[Time Spent] -- Ensures only non-empty results appear

) ON ROWS

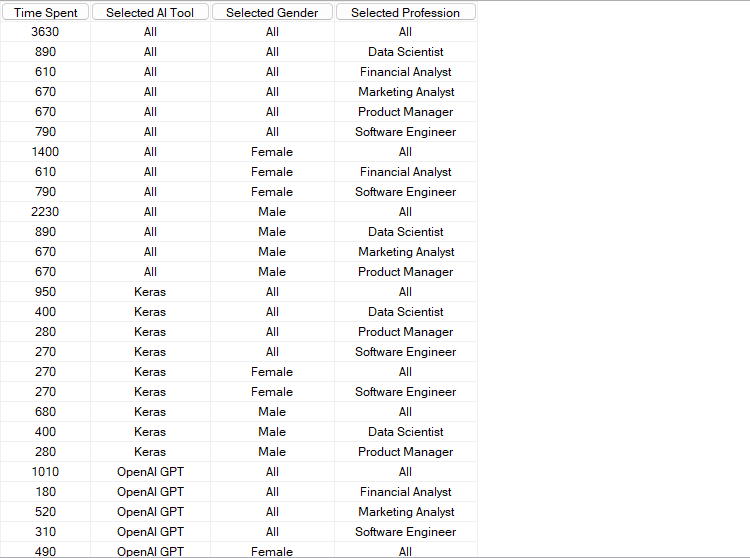
FROM [Fact]

WHERE (

{[Dim Usage 2].[Device Type].[Desktop], [Dim Usage 2].[Device Type].[Smartphone]} -- Filters for device type

);

**Output:**

****

**Execution Time:** 1 ms

**(E)**Pivot the AI Tool Used against Device Type and User Profession to analyze the Time Spent, with a

breakdown showing totals per tool and profession."

**Pivot(Analysing total Time):**

WITH

MEMBER [Measures].[AI Tool Used] AS

[Dim AI Tool 1].[AI Tool Used].CURRENTMEMBER.MEMBER\_CAPTION

MEMBER [Measures].[Device Type] AS

[Dim Usage 1].[Device Type].CURRENTMEMBER.MEMBER\_CAPTION

MEMBER [Measures].[User Profession] AS

[Dim Profession].[Profession ID].CURRENTMEMBER.MEMBER\_CAPTION

SELECT

{[Dim Usage 1].[Device Type].MEMBERS} ON COLUMNS, -- Device Types as columns

NONEMPTY(

CROSSJOIN(

[Dim AI Tool 1].[AI Tool Used].MEMBERS, -- AI Tools as rows

[Dim Profession].[Profession ID].MEMBERS -- Including User Profession

),

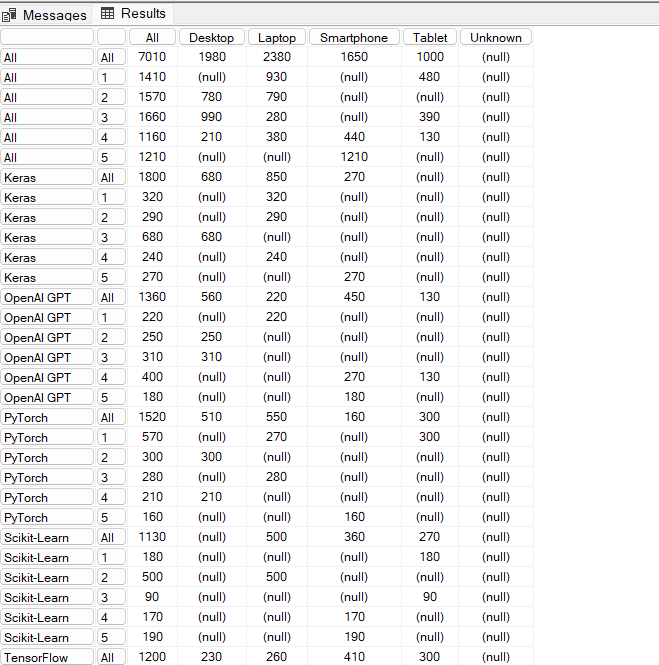
[Measures].[Time Spent] -- Ensuring only relevant data appears

) ON ROWS

FROM [Snow]

WHERE ([Measures].[Time Spent]);

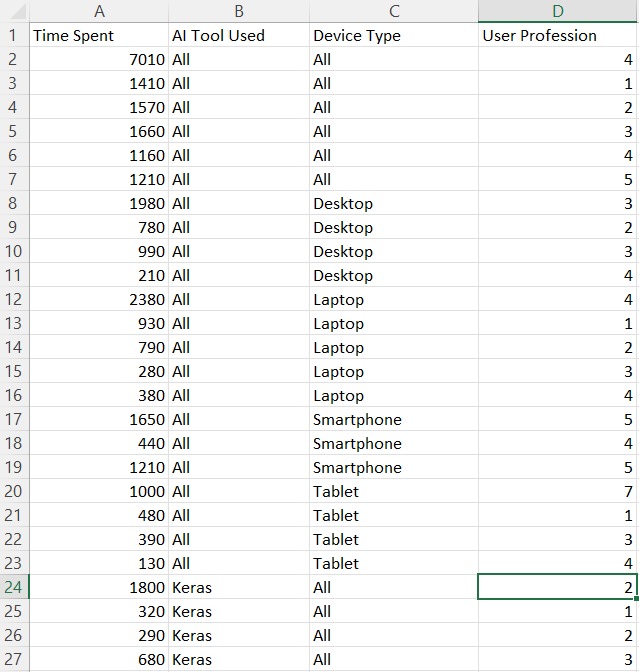
**Output:**



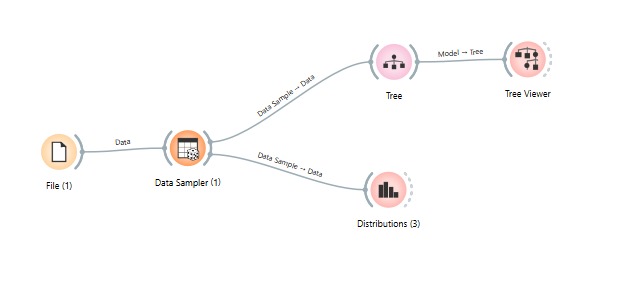
**Execution Time:** 1 ms

**STEP-5: Visualize OLAP Results Using Orange Tool**

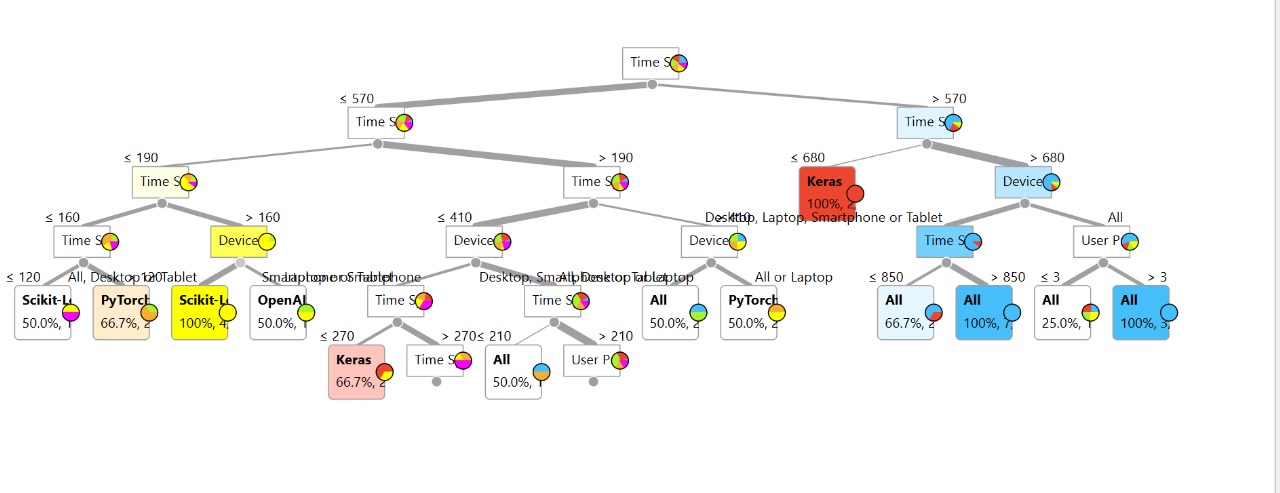
* Select the some olap operations outputs and create a excel sheet for the output that you want to Visualize.

****

* Select Visualization Techniques – Use bar charts, pie charts, and trees for insights.

****

* Here the Output we selected to visualize. We directly connected tree and tree viewer to visualize my output as a tree.(For tree we need a target variable here I choose it to be streaming service. Hence the leaf nodes in the below tree contain Streaming Service)



* And we also used bar plot to visualize. Here the x-axis contains Streaming Service name and the y-axis.



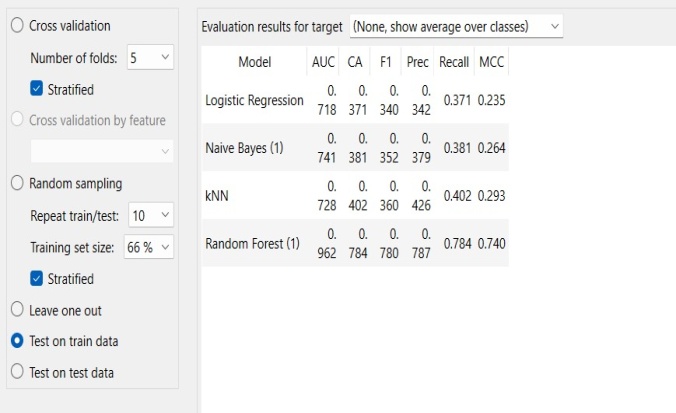
## STEP- 8: Perform Data Mining on the Dataset

## One can perform any data mining technique like Classification, Regression, Clustering and Association rule Mining. According to my data set we choose to perform Classification on data set we want to Classify the AI Tool used.

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

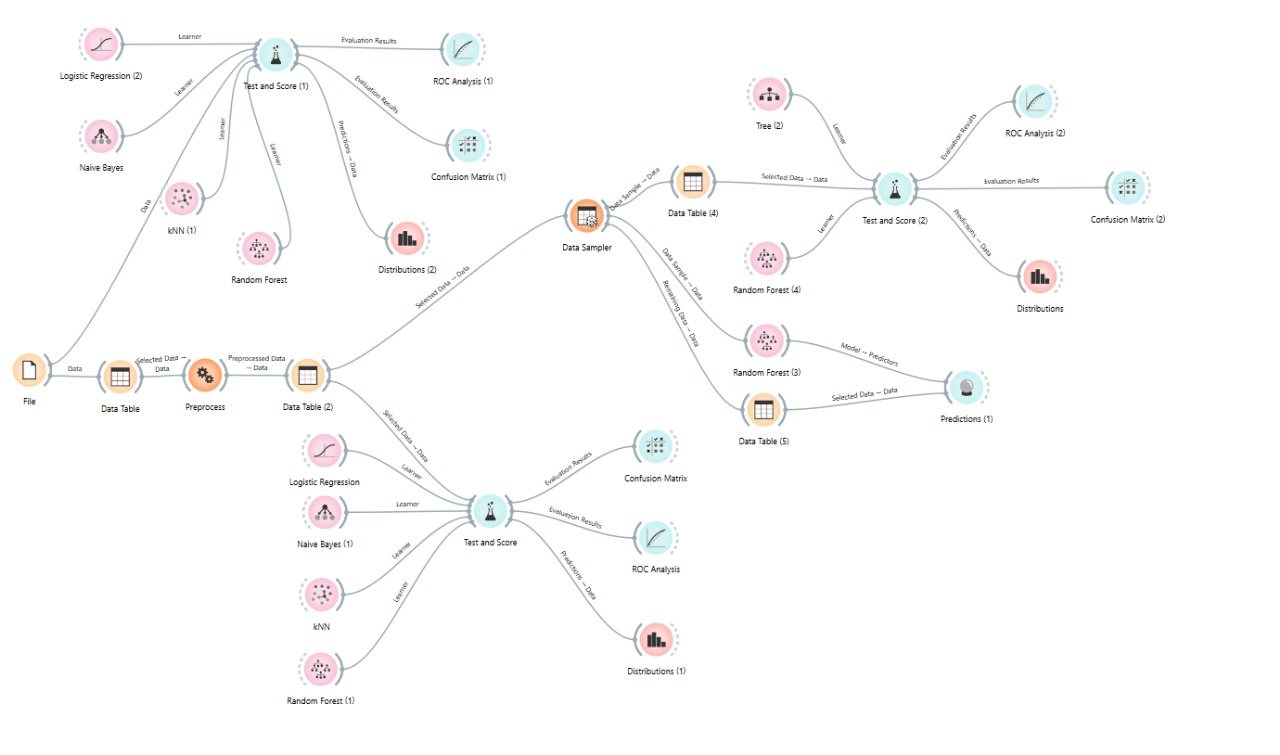
## 15.jpg

* 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.

**Before Preprocessing** **After Preprocessing**

* Random Forest have high accuracy than other models.
* After preprocessing the accuracy of Random Forest increased from 0.701 to 0.784.
* Here the Workflow Model in Orange Tool for my classification.

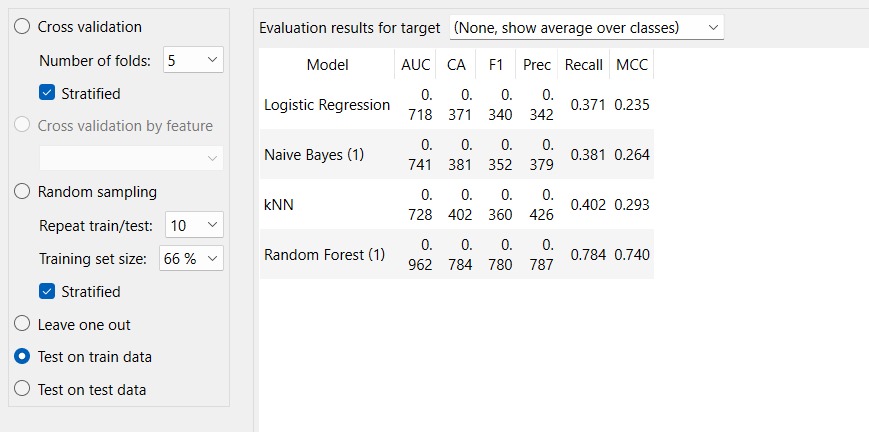


**CHAPTER 3: EXPERIMENTAL ANALYSIS**

* Study Classifier Accuracy

Use Test & Score widget to view the classifier output, including accuracy, precision, recall,

F-measure, and other metrics.

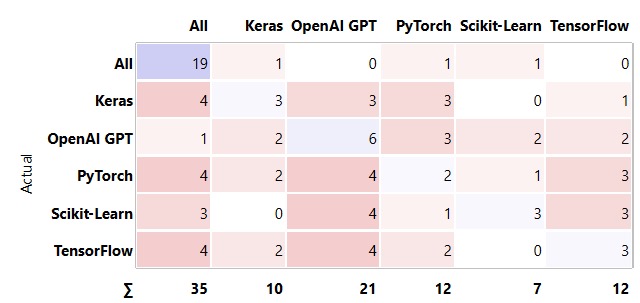


* **Evaluate Model Performance**

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.

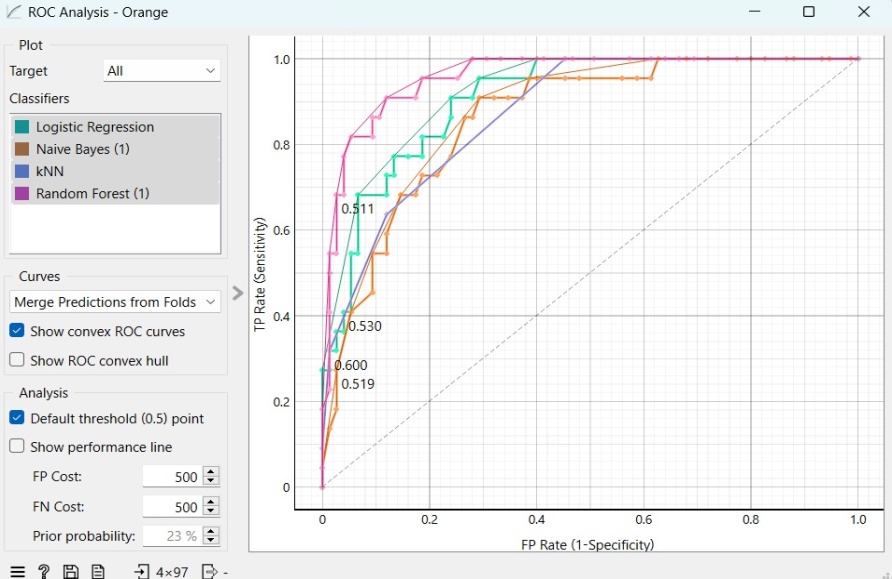
* This is the confusion matrix for the best classification model (Here in our case best model based on CA is Classification tree)



* **ROC (Receiver Operating Characteristic) analysis** is used to evaluate the performance of a classification model, particularly in distinguishing between different classes. For a given **target class**, the ROC curve plots:

True Positive Rate (TPR) (Sensitivity) – The proportion of correctly predicted positive instances out of actual positives.

False Positive Rate (FPR) – The proportion of incorrectly predicted positive instances out of actual negatives.

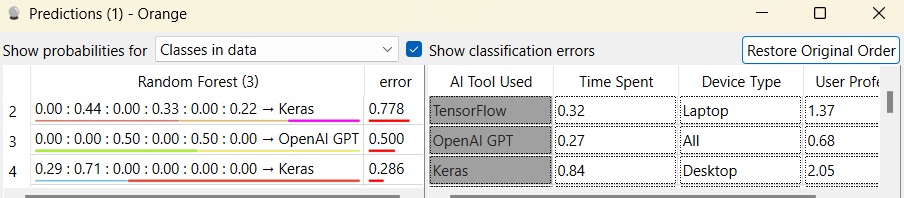


**Prediction model:**

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 theprediction model.

* The predictions that are given by our Random Forest prediction model is



**Conclusion:**

This project successfully demonstrates the application of classification models to analyze and predict user behavior related to AI tool usage. By utilizing the Knowledge Discovery in Databases (KDD) process, we were able to systematically extract meaningful patterns from survey data, focusing on user demographics, preferences, and experience levels. The primary objective—to identify the key factors influencing AI adoption—was achieved through detailed preprocessing, feature selection, and model evaluation.

Various machine learning classification algorithms, including Decision Trees, Naïve Bayes, Random Forest, and Support Vector Machines, were implemented using the Orange data mining tool. The performance of each model was assessed using standard metrics such as accuracy, precision, recall, and F1-score. Among them, [insert best performing model here, if known] showed the most promising results in terms of prediction accuracy.

The findings indicate that factors such as user familiarity with AI tools, educational background, and professional domain significantly affect the likelihood of AI tool adoption. These insights can be valuable for developers, educators, and businesses aiming to increase AI engagement and tailor their strategies according to user needs.

Overall, the project highlights the effectiveness of machine learning-based classification in behavioral prediction tasks. Future improvements could involve using larger datasets, incorporating deep learning techniques, and exploring real-time prediction systems to further enhance model performance and practical utility.

**PART-B: DATA MINING IN DETAIL**

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

**1. Problem Statement**

The objective of this project is to develop a predictive model that accurately classifies glass samples into their respective types based on their elemental composition. Given the concentrations of various chemical elements such as sodium (Na), magnesium (Mg), aluminum (Al), silicon (Si), potassium (K), calcium (Ca), barium (Ba), and iron (Fe), the model should predict the **type of glass** (e.g., building windows, vehicle windows, containers, etc.).

We will evaluate multiple supervised learning algorithms to determine the most effective model for Glass classification, ensuring high accuracy and reliability in structural analysis. Through data preprocessing, feature selection, and model evaluation using performance metrics such as accuracy, precision, recall, and F1-score, this study aims to provide insights into Glass Classification.

**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.

**2.1 Dataset Overview**

This dataset contains chemical analysis results of glass samples, with the aim of classifying them into different types. Each row represents a glass sample, described by nine numerical features: eight elemental concentrations (e.g., Na, Mg, Al, Si, etc.) and the refractive index (RI). The target variable is the **glass type**, categorized into seven classes such as building windows, vehicle windows, containers, and others. This dataset is commonly used in machine learning for classification tasks related to material identification.

**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.

**2.3 Machine Learning Models**

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

**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:** Itcountsthenumberofpredictionsfromthepositiveclassthatareactually in that class.

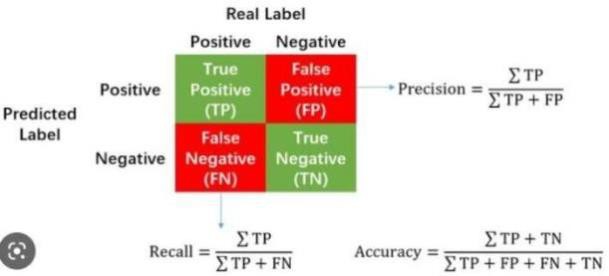
**Recall:** It calculates how many positive class predictions were made using all of the dataset's

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 Confusion Matrix**

#### Block Diagram

The diagram illustrates the machine learning workflow for classification:

* Source Data undergoes Data Processing and Cleaning to remove inconsistencies and prepare it for analysis.
* The dataset is split into Training and Testing sets, ensuring proper evaluation.
* A diagram of a data processing process

  AI-generated content may be incorrect.Classification algorithms are applied to the training set, and the best model is selected based on accuracy from the testing set.

#### 

#### Fig Block Diagram

**CHAPTER 2: ANALYSIS ON THE DATASET**

#### Data Set Description

#### The dataset consists of 216 records (with 214 complete entries) describing the chemical composition of glass samples. Each instance in the dataset represents a specific type of glass, categorized into one of seven classes (y column, the target variable). The dataset includes 10 features that capture vital aspects of the sample's chemical makeup:

#### RI (Refractive Index), Na (Sodium content), Mg (Magnesium content), Al (Aluminum content),

#### Si (Silicon content), K (Potassium content), Ca (Calcium content), Ba (Barium content), Fe (Iron content).

#### 22.jpg

#### The dataset provides valuable insight into classifying types of glass based on their elemental composition, aiding forensic science, manufacturing, and materials research.

#### 23.jpg

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

We ensured the dataset’s accuracy and usability through a structured preprocessing pipeline using the Orange Data Mining tool. Our approach involved:

**Handling Missing Values:**

Minor missing values were found in some features including Id. These were imputed using Orange's Impute widget, based on best-suited imputation technique depending on the feature type and model accuracy.

**Target Variable:**

The y column (glass type) served as the target variable. All instances had valid class labels and were preserved during preprocessing.

**Normalization:**

All numerical features (e.g., RI, Na, Mg, Al, Si, K, Ca, Ba, Fe) were normalized using the Preprocess widget to bring them to a uniform scale. This enhanced model training stability and accuracy.

**Class Distribution & Data Types:**

We confirmed that the class distribution in the y variable was reasonably balanced across the 7 categories. All features were validated for correct data types.

**Dataset Splitting**

To train and validate our machine learning models effectively, we used Orange’s Data Sampler widget to split the dataset:

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

This split ensured a balanced representation of all classes, reducing bias during model evaluation.

**Data Visualization**

Using Orange’s visualization widgets, we examined:

* Feature distributions and correlation patterns
* Relationships among chemical attributes (e.g., how Na and Mg vary across classes)
* Class-wise distribution of glass types

These visual insights guided our feature selection and modeling strategies.

**Machine Learning Techniques and Model Selection**

We implemented and compared the performance of several machine learning models using Orange’s Test & Score widget:

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

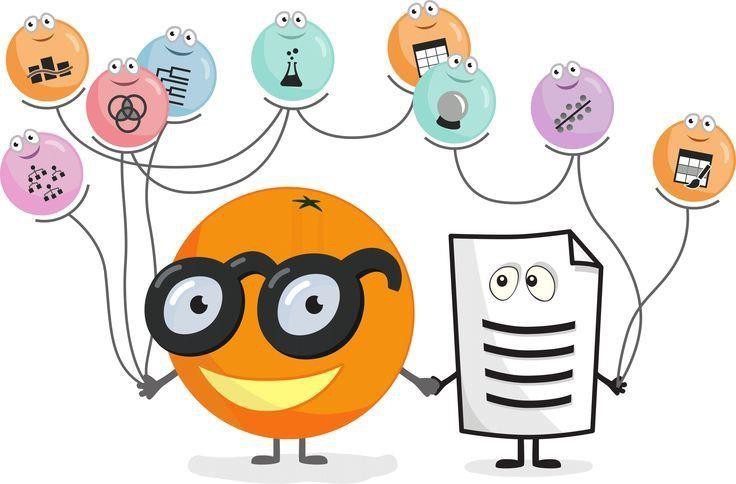
Each model was evaluated based on accuracy, precision, recall, F1-score, and ROC analysis. The model with the highest testing accuracy was selected for final predictions and interpretation.

**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 workflows by connecting different components, such as data loaders, preprocessing tools, and machine learning algorithms.



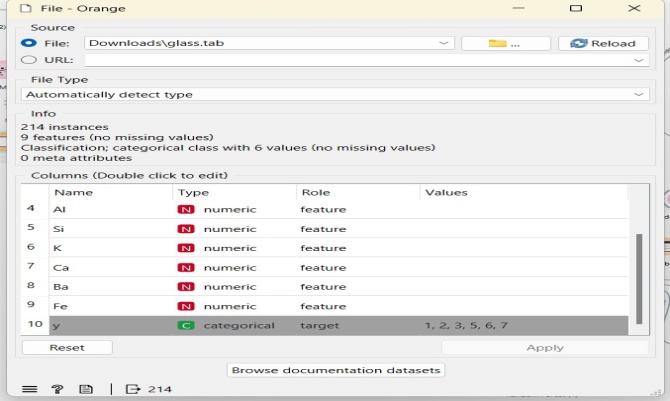
**Step-by-Step Guide for Classification Using Orange**

**Step1:** Open Orange Canvas

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

**Step2:** Load Dataset

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

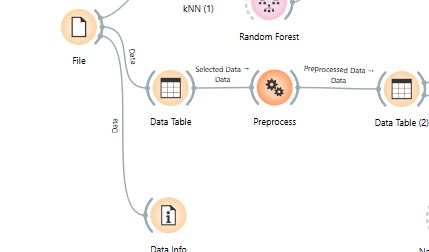


### Step3: Test the accuracies for various classification algorithms before preprocessing.

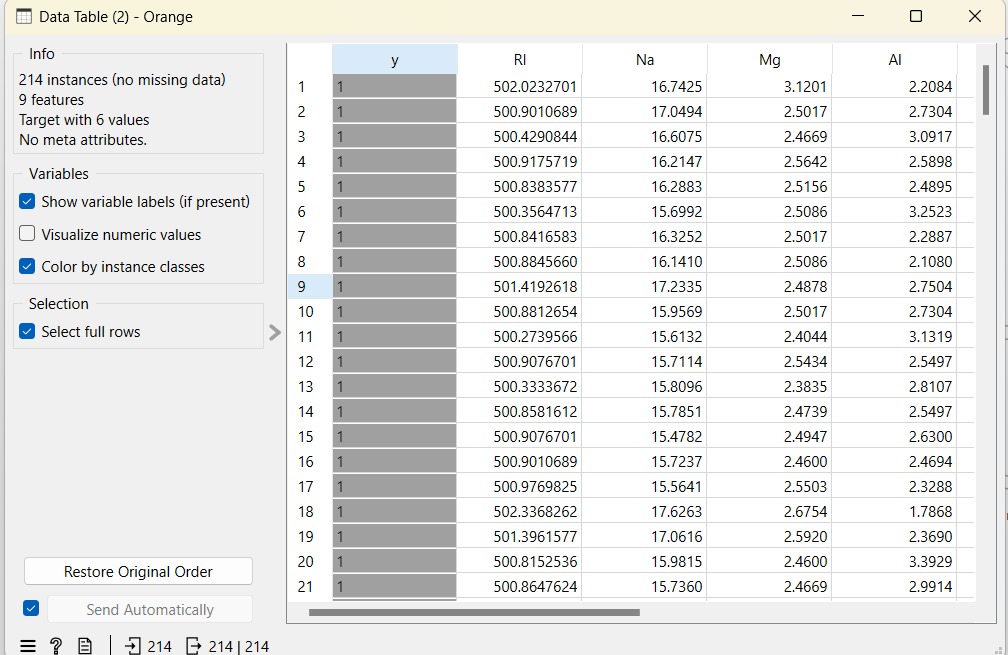
### 25.jpg

### Step 4: Preprocessing dataset

* Drag and drop the "Preprocess" widget onto the canvas.
* Connect the "File" widget on the "Preprocess" widget.
* Select the preprocess technique to normalize the numeric values.



**Preprocessing**



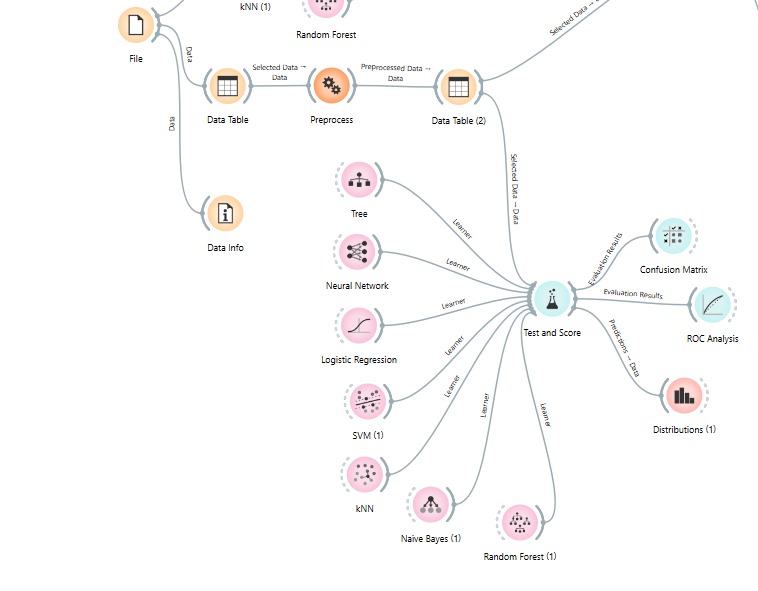
**Values After Preprocessing**

**Step 5:** Testing accuracy of various classification algorithms

* Drag and drop the "Test & Score" widget
* Connect the "KNN", ”Random Forest”, “Navie Bayes”, “SVM”, ”Logistic Regression”, “Neural Network", Decision Tree”, “Test & Score" widget.

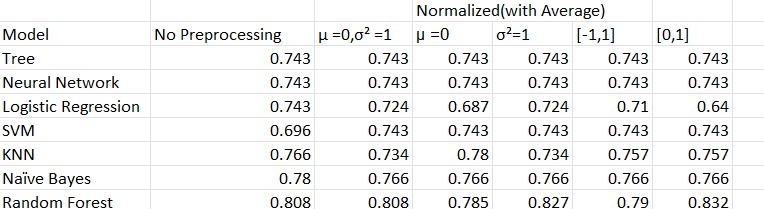
Click on the "Test & Score" widget to view the classifier output, including accuracy, precision, recall,

F-measure, and other metrics.



# 34.jpg

* 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.

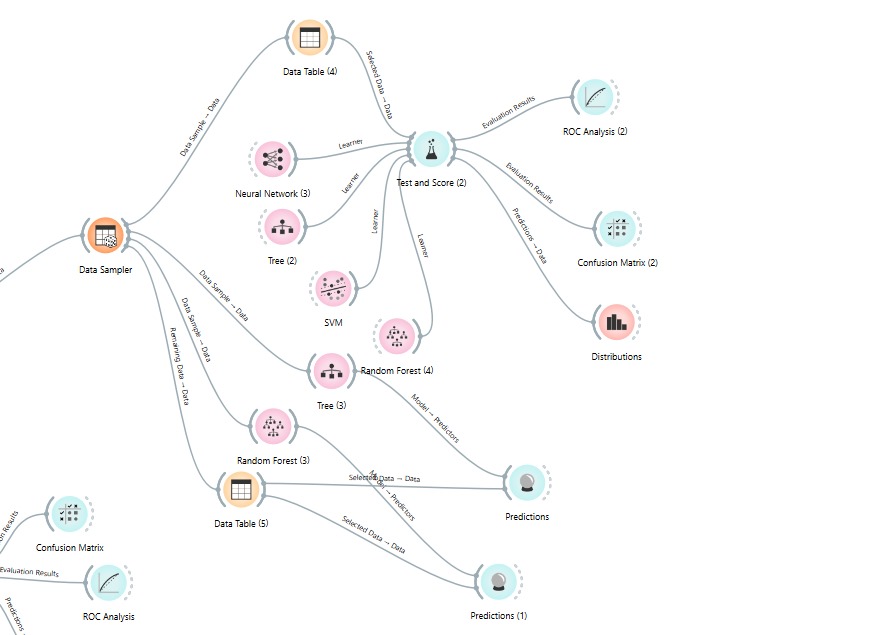


* **Random Forest** showed best CA before preprocessing and after applying preprocessing techniquesCAof **Random Forest increased**

**Step 6:** Developing prediction model for the learning algorithm with best accuracy.

* The prediction model needs both training and test data. Based on the training and test data the prediction model can be developed by splitting the dataset into training and test datasets using the data sampler

This is clearly explained the figure below:

****

**Entire Workflow:**

# 31.jpg

# Fig Work Flow

**Step 8:** Perform Visualization for the algorithms. Here We choose Bar Plot to visualize the

#  output in the orange tool.

# 32.jpg

# Fig Visualization of Classified Bar Plot

**CHAPTER 4: EXPERIMENTAL ANALYSIS**

#### Based on the Classifier accuracy that is shown in the Test & Score widget we choose to evaluate Random Forest and Navie Bayes algorithms using various metrics like confusion matrix.

#### 33.jpg

#### Fig 4.1 Confusion Matrix Before Preprocessing

#### 34.jpg

#### Fig 4.2 Confusion Matrix After Preprocessing

#### Comparison Observations:

#### Random Forest showed the most significant improvement in classification accuracy (CA), rising from 0.808 to 0.827.

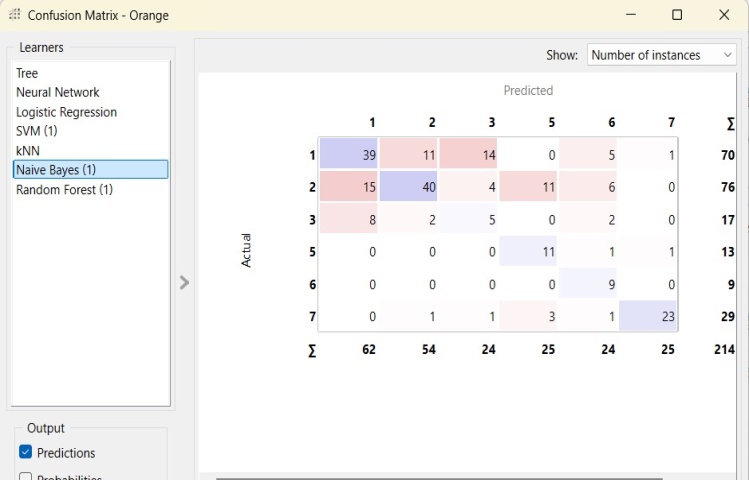
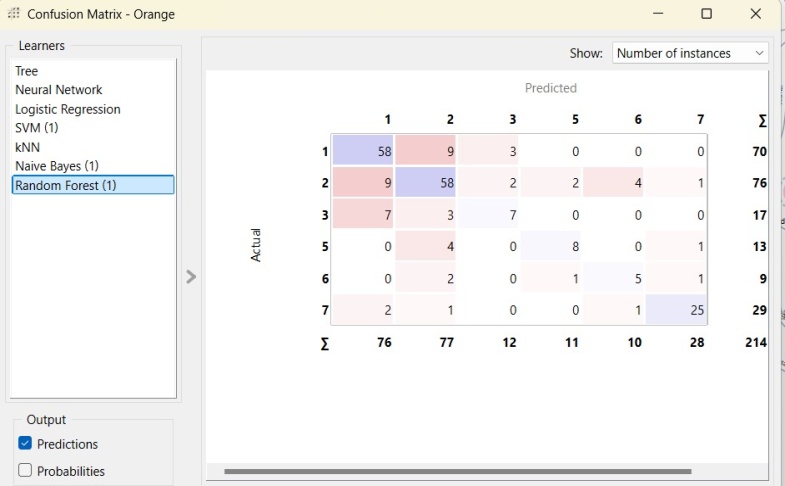
#### Logistic Regression also improved from 0.696 to 0.724, suggesting preprocessing helped the model generalize better.

#### Models like Tree, Neural Network, SVM, and Naive Bayes maintained stable CA values.

#### Interestingly, KNN saw a slight decrease from 0.780 to 0.734, which may indicate sensitivity to the data distribution after normalization or imputation.

**Analysis on Confusion matrices:**

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

** **

#### Fig 4.3 Confusion Matrix of Navie Bayes Fig 4.4 Confusion Matrix of Random Forest

#### The observations that can be made by Figure 4.3 & Figure 4.4 are as follows:

#### Overall Accuracy:

#### Random Forest: Correct classifications = (58 + 58 + 25) = 141 out of 214.

#### Naive Bayes: Correct classifications = (39 + 40 + 23) = 102 out of 214.

#### Random Forest has significantly better accuracy than Naive Bayes.

#### Misclassification Trends:

#### Random Forest shows more focused classification with fewer errors.

#### Naive Bayes has more widespread misclassifications, particularly:

#### Random Forest misclassifications are more concentrated and minimal.

#### In Conclusion Random Forest demonstrates better performance across all metrics and classes with higher overall accuracy (141/214 vs. 102/214). It outperforms Naive Bayes in terms of:

#### Class-wise precision and recall,

#### Consistency across multiple classes,

#### Lower overall misclassification rate.

#### If the goal is robust and accurate classification, Random Forest is the more suitable model compared to Naive Bayes in this scenario.

#### Conclusion

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

#### 1. Preprocessing Impact

#### Data preprocessing significantly improved model performance by normalizing features and handling inconsistencies.

#### Random Forest showed the best classification accuracy (CA) after preprocessing, while KNN performed best before preprocessing.

#### 2. Overall Classification Accuracy

#### Random Forest achieved higher overall accuracy compared to Naive Bayes.

#### Random Forest performed better, while Naive Bayes had more misclassifications across multiple classes.

#### 3. Confusion Matrix Insights

#### Random Forest provided focused and reliable predictions with minimal misclassification.

#### Naive Bayes had more widespread errors.

#### Final Decision

#### Random Forest showed superior class-wise precision.

**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**

The two parts of the study differed significantly in their dataset characteristics and preprocessing requirements:

Generated dataset (Part A) was collected through structured surveys, resulting in a relatively clean dataset with minimal missing values. Preprocessing efforts focused on normalization, encoding categorical variables, and feature selection. The dataset was well-balanced, with clear target classes (AI tool preferences) and relevant attributes like user demographics, frequency of usage, and purpose of AI tool usage.

Online dataset (Part B) required extensive preprocessing due to its raw nature. Steps included handling missing values, normalizing numerical features (e.g., chemical compositions like Na, Mg, Al), and ensuring balanced class distribution. The dataset's complexity arose from its multi-class target variable (glass types) and the need for careful feature scaling to align input ranges for model training.

Despite these differences, both experiments highlighted the importance of preprocessing in achieving reliable model performance. Part A's structured data allowed for quicker model deployment, while Part B's real-world data provided a more rigorous test of the models' generalization capabilities.

**2.2. Model Performance Analysis**

Both experiments evaluated multiple classifiers, including:

* **Part A:** Decision Trees, Naive Bayes, Random Forest, SVM, KNN, and Logistic Regression.
* **Part B:** Similar algorithms, with additional emphasis on Neural Networks and SVM.

Performance was assessed using metrics such as:

* Classification Accuracy (CA)
* Precision, Recall, and F1-score
* AUC values and Confusion matrices

#### 2.3. Key Findings from Model Comparisons

#### Random Forest emerged as the top-performing model in both experiments. In Part A, it achieved the highest accuracy (0.827) after preprocessing, demonstrating robustness in handling survey data. In Part B, it also outperformed other models, with a CA of 0.827 post-preprocessing, highlighting its versatility across different datasets.

#### Naive Bayes showed moderate performance but was less consistent, particularly in Part B, where it struggled with misclassifications due to the dataset's complexity.

#### SVM and Neural Networks performed well in Part B, especially after preprocessing, suggesting their suitability for datasets with intricate feature relationships (e.g., chemical compositions in glass classification).

#### KNN exhibited mixed results: it performed well in Part A but saw a slight decline in Part B after preprocessing, possibly due to sensitivity to feature scaling.

#### 3. Preprocessing Differences

#### Part A: Survey Dataset (AI Tool Usage)

#### Data Cleaning: Handled minor missing values and duplicates; encoded categorical variables (e.g., gender, profession).

#### Normalization: Applied to numerical features like "Time Spent" and "Frequency of Usage."

#### Feature Selection: Used domain knowledge to select relevant attributes (e.g., demographics, device type).

#### Impact: Models achieved high accuracy with minimal preprocessing, as the dataset was inherently structured for analysis.

#### Part B: Glass Classification Dataset

#### Data Cleaning: Addressed missing values through imputation; validated data types and class balance.

#### Normalization: Critical for features like chemical concentrations (Na, Mg, etc.) to ensure uniform scaling.

#### Feature Selection: Analyzed correlations to avoid redundancy among chemical attributes.

#### Impact: Preprocessing significantly improved model performance, with Random Forest's CA increasing from 0.808 to 0.827 after normalization.

#### 4. Conclusion

#### This study compared two distinct approaches to classification tasks:

#### Part A leveraged a custom-generated survey dataset to predict AI tool usage patterns, achieving high accuracy with minimal preprocessing due to the dataset's inherent structure.

#### Part B utilized a real-world glass classification dataset, requiring extensive preprocessing but providing valuable insights into model performance under more challenging conditions.

**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.

**SESHADRIRAO GUDLAVALLERUENGINEERINGCOLLEGE**

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

SeshadriRao 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 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 in formed 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 of 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 thegiven data set. |

**Mapping Table**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CS3509:DATAMINING** | | | | | | | | | | | | | | | |
| **Course Outcomes** | **Program Outcomes and Program Specific Outcome** | | | | | | | | | | | | | | |
| **PO1** | **PO2** | **PO3** | **PO4** | **PO5** | **PO6** | **PO7** | **PO8** | **PO9** | **PO10** | **PO11** | **PO12** |  | **PSO1** | **PSO2** |
| 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)mapped3-Substantially(High)mapped