Analysis on Preferred Streaming Platforms Based on Viewer Demographics and Genre Preferences: A KDD Approach

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

P. Karuna (22481A05H7)

P. Varshitha(22481A05J4)

P. Tejaswi(22481A05I1)

M. Viswas Abhishikth (22481A05F1)

Under the Enviable and Esteemed Guidance of

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

Professor

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 **"ANALYSIS ON PREFERED STREAMING PLATFORMS BASED ON VIEWER DEMOGRAPHICS AND GENRE PREFEERENCES: A KDD APPROACH"** is a bonafide record of work carried out by **P. Karuna (22481A05H7), P. Varshitha (22481A05J4), P. Tejaswi(22481A05I1), M.Viswas Abhishikth(22481A05F1),** under the guidance and supervision of **Dr.G.V.S.N.R.V.Prasad,M.S.,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
(Dr.G.V.S.N.R.V.Prasad)

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

External Examiner

ACKNOWLEDGEMENT

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

We would like to express our deep sense of gratitude and sincere thanks to **Dr.G.V.S.N.R.V.Prasad,M.S.,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

P. Karuna (22481A05H7)

P. Varshitha(22481A05J4)

P. Tejaswi(22481A05I1)

M. Viswas Abhishikth (22481A05F1)

INDEX

ABSTRACT

The rapid growth of digital platforms and the increasing availability of user interaction data has made analytical processing and machine learning essential in deriving meaningful insights. This project presents a two-fold data analysis approach: performing OLAP (Online Analytical Processing) operations on a movie dataset and using the Orange tool for classifying the outcomes of horses based on medical data.

The first phase focuses on OLAP operations applied to a movie dataset modeled using star and snowflake schemas. The dimensions considered include User Age Group, Genre, Application Platform, and Duration, while measures such as Total Watch Time and View Count were analyzed. OLAP operations such as slice, drill-down, roll-up, and pivot allowed for dynamic and flexible exploration of user behavior patterns and movie preferences.

In the second phase, the Orange data mining tool was utilized to conduct a classification task on the Horse Colic dataset, a medical dataset containing information about the health condition and treatment of horses. The primary objective was to predict the outcome of the horse (e.g., lived, died, or was euthanized) based on input features like pulse, temperature, surgical lesion, and more. The Orange tool provided a user-friendly, drag-and-drop environment to preprocess the dataset, visualize feature relationships, and build classification models using algorithms such as Decision Tree, Naive Bayes, and k-Nearest Neighbors (k-NN),Gradient Boosting and Neural Network. The performance of these models was evaluated using confusion matrices, ROC curves, classification accuracy, and cross-validation scores.

The integration of OLAP operations for structured analytical querying and Orange for machine learning classification demonstrates a powerful synergy between traditional data warehousing techniques and modern AI tools. This dual approach not only facilitates in-depth data analysis but also supports predictive insights, making it valuable for real-world decision-making in domains ranging from digital media to veterinary health analytics.

PART A: Predicting the Most Preferred Movie Streaming App Based on User Demographics and Genre Preferences Using KDD PROCESS

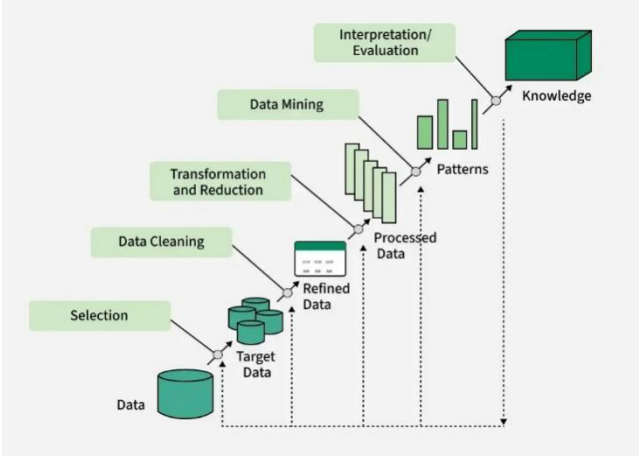
CHAPTER 1: INTRODUCTION

1.1 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



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

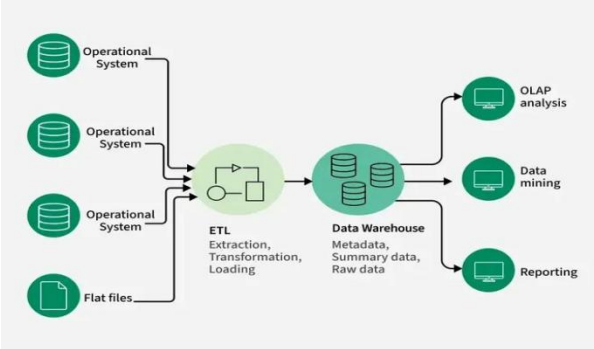
Evaluation in KDD involves assessing the patterns identified during data mining to determine their relevance and usefulness. It includes calculating the “interestingness score” for each pattern, which helps to identify valuable insights. Visualization and summarization techniques are then applied to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



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

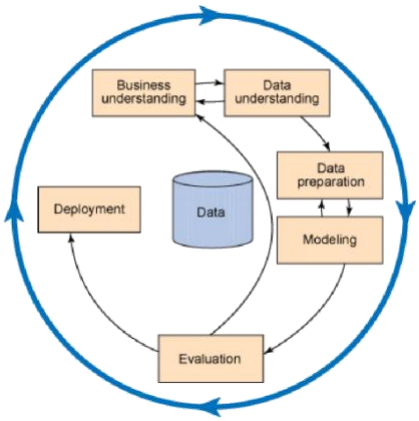
The main components of a data warehouse include:

- Data Sources: These are the various operational systems, databases, and external data feeds that provide raw data to be stored in the warehouse.
- ETL (Extract, Transform, Load) Process: The ETL process 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 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 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, 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.



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
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
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
Gradient Boosting	Gradient Boosting is an ensemble learning technique that builds models sequentially, where each new model corrects the errors of the previous ones. It combines weak learners, usually decision trees, to create a strong predictive model. The algorithm minimizes a loss function using gradient descent.	Regression and Classification
Neural Network	Neural networks are computational models inspired by the human brain, consisting of layers of interconnected nodes (neurons). They learn patterns from data by adjusting weights through training using algorithms like backpropagation. Neural networks are used for complex tasks like image recognition, natural language processing, and predictive modeling.	Regression and Classification
SVM	Support Vector Machine (SVM) is a supervised learning algorithm that finds the optimal hyperplane to separate data into classes with the maximum margin. It can be used for both classification and regression tasks using linear or non-linear kernels.	Regression and Classification
Random Forest	Random Forest is an ensemble learning method that builds multiple decision trees and combines their outputs for better accuracy and stability. It is used for both classification and regression by averaging or voting across trees.	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.

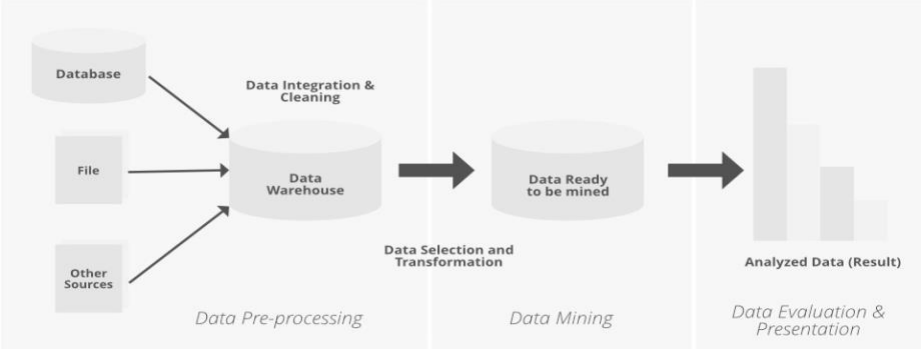
By applying clustering, we aim to identify distinct groups of individuals with similar visual characteristics, facilitating personalized recommendations for lens suitability. This unsupervised approach aids in data exploration and segmentation, providing insights into diverse needs and preferences among 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 item sets 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.



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 successful data 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 not only 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:

Software Requirements:

- Orange Data Mining Tool: Visual data analysis and modeling (version 3.x recommended)
- Microsoft SSMS: Managing and querying relational databases (SQL Server)
- Microsoft Visual Studio: Development and integration environment for coding
- Microsoft Excel / CSV Tool:For dataset formatting and conversion

Hardware Requirements:

- Processor: Intel i7 or higher (recommended: i5/i7 for faster performance)
- RAM: Minimum 4 GB (recommended: 8 GB or higher for handling larger datasets)
- Hard Disk: At least 1 GB of free space for Orange and dataset storage
- Display: Minimum 1024x768 resolution
- Operating System: Windows 10 or higher / macOS / Linux (64-bit)

CHAPTER 2:DATA MINING AND DATA WARE HOUSING

2.1 Problem Statement:

With the rise of multiple streaming platforms, understanding user preferences based on demographics and genre choices is essential. This project focuses on analyzing viewer data—such as age group, genre preference, and watch history—to identify trends and predict the preferred streaming platform (app) for each user. Using OLAP operations for analysis and machine learning for prediction, the goal is to help platforms better tailor content and improve user engagement.

Objectives:

- **Analyze Demographic-Based Viewing Patterns:**

Examine user watch history across different age groups to identify content consumption trends and genre preferences.

- **Evaluate Genre Popularity Across Platforms:**

Use OLAP operations to assess which genres perform best on specific streaming apps, segmented by user demographics.

- **Identify Platform Preference Trends:**

Determine the most preferred streaming platforms (apps) for different user groups based on their viewing behavior and content type.

- **Predict User Platform Preference Using Machine Learning:**

Develop and train a predictive model to forecast the preferred streaming app for individual users based on their demographic and viewing data.

- **Enhance Content Personalization Strategies:**

Provide actionable insights to streaming platforms to optimize content recommendations and improve user engagement and retention.

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.

The process begins by collecting the dataset from various sources, ensuring that all relevant information is gathered for further analysis.The data must be preprocessed, which typically involves cleaning, normalizing, or transforming the raw input to ensure consistency and accuracy. Once the data is prepared, schemas are constructed to define the structure and relationships within the database or data warehouse. These schemas are then visualized to give a clear overview of the data relationships and to confirm that the structure aligns with the project's goals. Following this, the deployment phase involves creating OLAP queries, which form the foundation for multi-dimensional data analysis. The next step is to perform OLAP operations, such as slicing, dicing, rolling up, or drilling down, to examine the data from various perspectives. Once the OLAP operations are completed, the resulting insights are visualized for better comprehension and decision-making. Finally, the process concludes with data mining tasks, where advanced techniques and algorithms are applied to uncover hidden patterns and trends for deeper, more predictive insights.

Step 1: Collect Dataset

↓

Step 2: Preprocess Data

↓

Step 3: Construct 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 is collected Using Google Forms

- Designed Google Form – with relevant questions (e.g., age, gender, viewing habits, OTT usage)

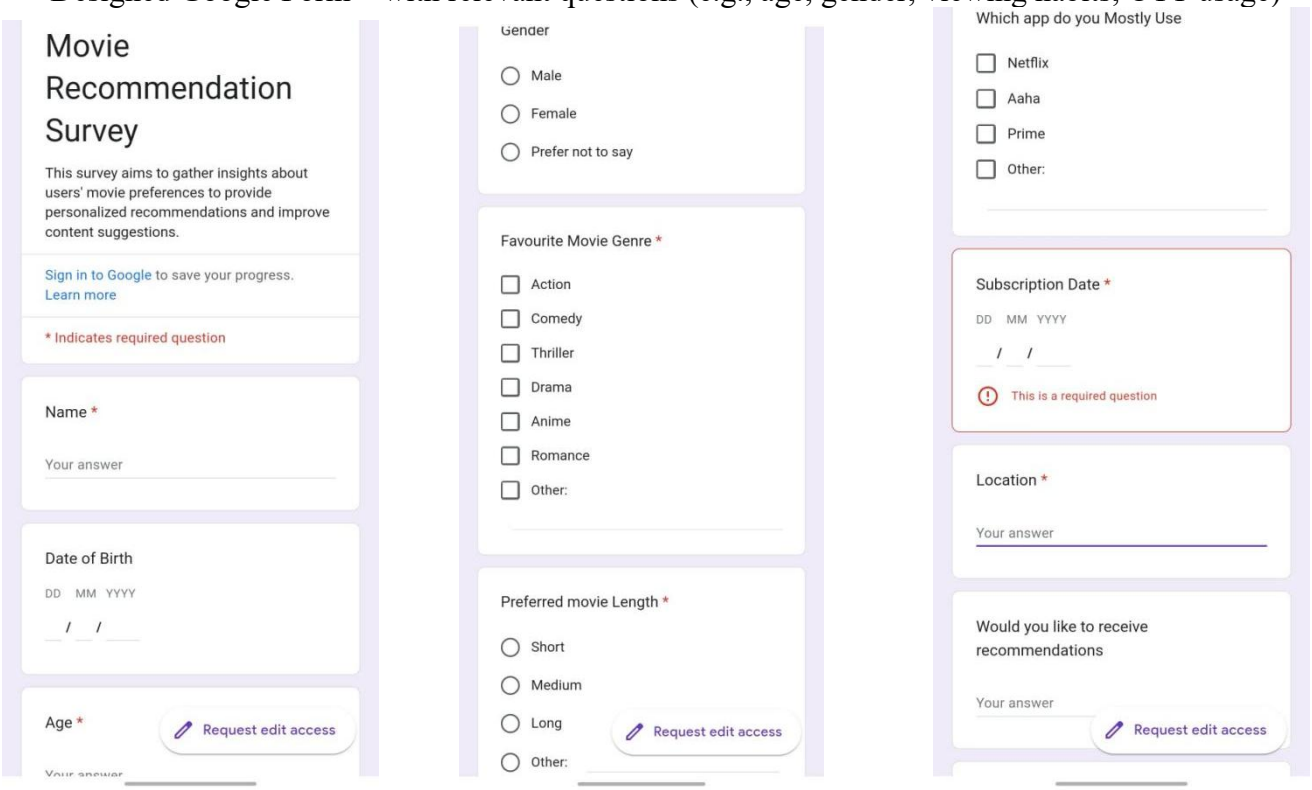
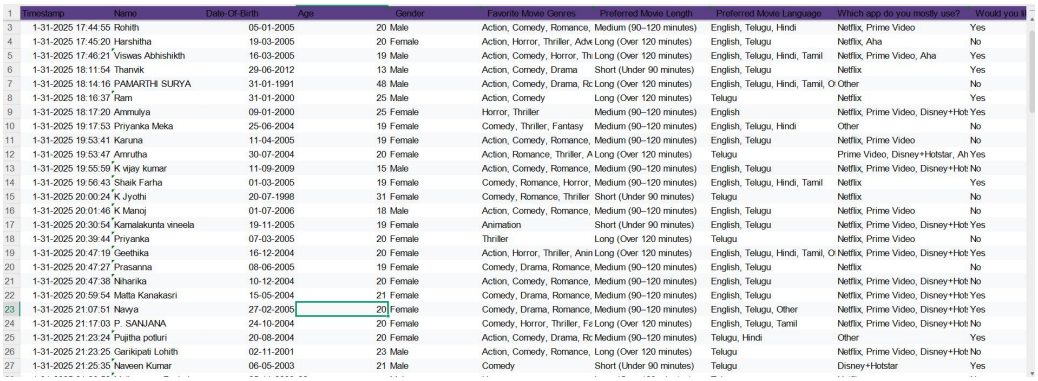


Fig:1.1 Google form for collection of dataset

- Shared it with participants through email, social media, or targeted groups.
  Link: https://docs.google.com/forms/d/1hslHVhpJcVMxhLWAhXOv5SKQzcantQzWERIpUEFX

- Collected Responses – Monitored responses and ensure enough data is gathered.

- Exported Data – Downloaded the responses as a CSV file for further processing.



Timestamp	Name	Date Of Birth	Age	Gender	Favorite Movie Genres	Preferred Movie Length	Preferred Movie Language	Which app do you mostly use?	Would you li
1-31-2025 17:44:55	Rohith	05-01-2005	20	Male	Action, Comedy, Romance,	Medium (90–120 minutes)	English, Telugu, Hindi	Netflix, Prime Video	Yes
1-31-2025 17:45:20	Harshitha	19-03-2005	20	Female	Action, Horror, Thriller, Adv	Long (Over 120 minutes)	English, Telugu	Netflix, Aha	No
1-31-2025 17:46:21	Viswas Abhishikth	16-03-2005	19	Male	Action, Comedy, Horror, Th	Long (Over 120 minutes)	English, Telugu, Hindi, Tamil	Netflix, Prime Video, Aha	Yes
1-31-2025 18:11:54	Thanvik	29-06-2012	13	Male	Action, Comedy, Drama	Short (Under 90 minutes)	English, Telugu	Netflix	Yes
1-31-2025 18:14:16	PAMARTHI SURYA	31-01-1991	48	Male	Action, Comedy, Drama, Ro	Long (Over 120 minutes)	English, Telugu, Hindi, Tamil, O	Other	No
1-31-2025 18:16:37	Ram	31-01-2000	25	Male	Action, Comedy	Long (Over 120 minutes)	Telugu	Netflix	Yes
1-31-2025 18:17:20	Ammulya	09-01-2000	25	Female	Horror, Thriller	Medium (90–120 minutes)	English	Netflix, Prime Video, Disney+Hot	Yes
1-31-2025 19:17:53	Priyanka Meka	25-06-2004	19	Female	Comedy, Thriller, Fantasy	Medium (90–120 minutes)	English, Telugu, Hindi	Other	No
1-31-2025 19:53:41	Karuna	11-04-2005	19	Female	Action, Comedy, Romance,	Medium (90–120 minutes)	English, Telugu	Netflix, Prime Video	No
1-31-2025 19:53:47	Amrutha	30-07-2004	20	Female	Action, Romance, Thriller, A	Long (Over 120 minutes)	Telugu	Prime Video, Disney+Hotstar, Ah	Yes
1-31-2025 19:55:59	K vijay kumar	11-09-2009	15	Male	Action, Comedy, Romance,	Medium (90–120 minutes)	English, Telugu	Netflix, Prime Video, Disney+Hot	No
1-31-2025 19:56:43	Shaik Farha	01-03-2005	19	Female	Comedy, Romance, Horror,	Medium (90–120 minutes)	English, Telugu, Hindi, Tamil	Netflix	Yes
1-31-2025 20:00:24	K Jyothi	20-07-1998	31	Female	Comedy, Romance, Thriller	Short (Under 90 minutes)	Telugu	Netflix	No
1-31-2025 20:01:46	K Manoj	01-07-2006	18	Male	Action, Comedy, Romance,	Medium (90–120 minutes)	English, Telugu	Netflix, Prime Video	No
1-31-2025 20:30:54	Kamalakunta vineela	19-11-2005	19	Female	Animation	Short (Under 90 minutes)	English, Telugu	Netflix, Prime Video, Disney+Hot	Yes
1-31-2025 20:39:44	Priyanka	07-03-2005	20	Female	Thriller	Long (Over 120 minutes)	Telugu	Netflix, Prime Video	No
1-31-2025 20:47:19	Geethika	16-12-2004	20	Female	Action, Horror, Thriller, Anin	Long (Over 120 minutes)	English, Telugu, Hindi, Tamil, O	Netflix, Prime Video, Disney+Hot	Yes
1-31-2025 20:47:27	Prasanna	08-06-2005	19	Female	Comedy, Drama, Romance,	Medium (90–120 minutes)	English, Telugu	Netflix	No
1-31-2025 20:47:38	Niharika	10-12-2004	20	Female	Action, Comedy, Romance,	Medium (90–120 minutes)	English, Telugu	Netflix, Prime Video, Disney+Hot	No
1-31-2025 20:59:54	Matta Kanakasri	15-05-2004	21	Female	Comedy, Drama, Romance,	Medium (90–120 minutes)	English, Telugu	Netflix, Prime Video, Disney+Hot	Yes
1-31-2025 21:07:51	Navya	27-02-2005	20	Female	Comedy, Drama, Romance,	Medium (90–120 minutes)	English, Telugu, Other	Netflix, Prime Video, Disney+Hot	Yes
1-31-2025 21:17:03	P. SANJANA	24-10-2004	20	Female	Comedy, Horror, Thriller, Fa	Long (Over 120 minutes)	English, Telugu, Tamil	Netflix, Prime Video, Disney+Hot	No
1-31-2025 21:23:24	Pujitha potluri	20-08-2004	20	Female	Action, Comedy, Drama, Ro	Medium (90–120 minutes)	Telugu, Hindi	Other	Yes
1-31-2025 21:23:25	Garikipati Lohith	02-11-2001	23	Male	Action, Comedy, Romance,	Long (Over 120 minutes)	Telugu	Netflix, Prime Video, Disney+Hot	No
1-31-2025 21:25:35	Naveen Kumar	06-05-2003	21	Male	Comedy	Short (Under 90 minutes)	Telugu	Disney+Hotstar	Yes

Fig:1.2 Collected dataset in csv form

Fields in Dataset:

- Name
- Age
- Date of Birth
- Gender
- Favorite movie genre
- Which app do you mostly use?
- Subscription date
- Location
- Would you like to receive recommendations?

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.

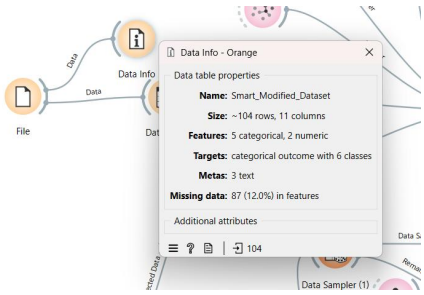


fig:2.1 Dataset imported to orange

- Handle Missing Values – Used imputation techniques (average/most frequent) to fill missing values.

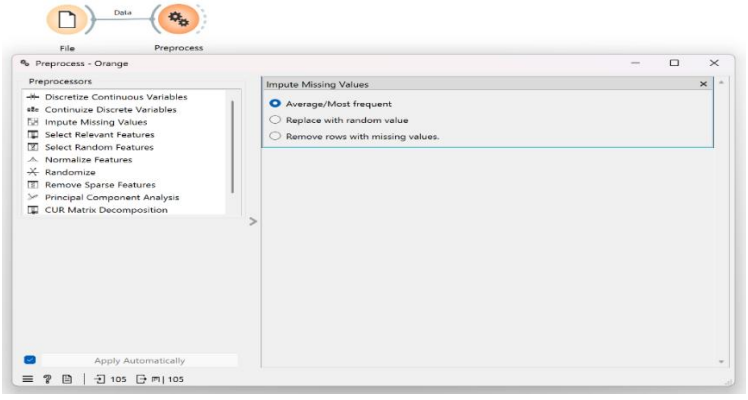


fig:2.2 Impute missing values

- Continuize Discrete Variables is used as another preprocess technique

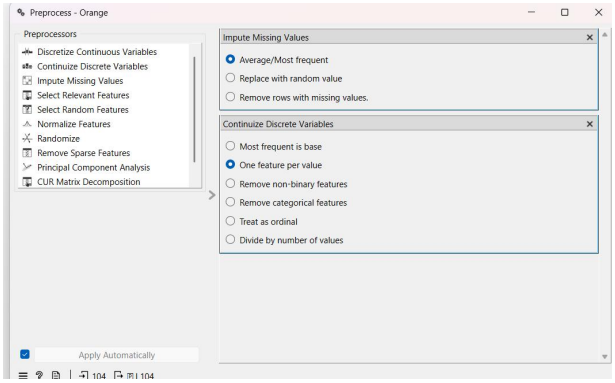


fig:2.3 Continuize Discrete Variables

Continuize Discrete Variables				
	Naïve Bayes	**KNN**	**Random forest**	**LOGISTIC REGRESSION**
Most Frequent in base	0.394	0.481	0.885	0.212
One feature per value	0.346	0.481	0.846	0.212
Remove non binary features	0.404	0.481	0.740	0.212
Treat as ordinal	0.423	0.481	0.875	0.212
Divide by number of values	0.423	0.481	0.856	0.212

- After analyzing the data from the above table we fix the continuize discrete variables to one feature per value
- After preprocessing the preprocessed dataset can be as shown in fig:2.4

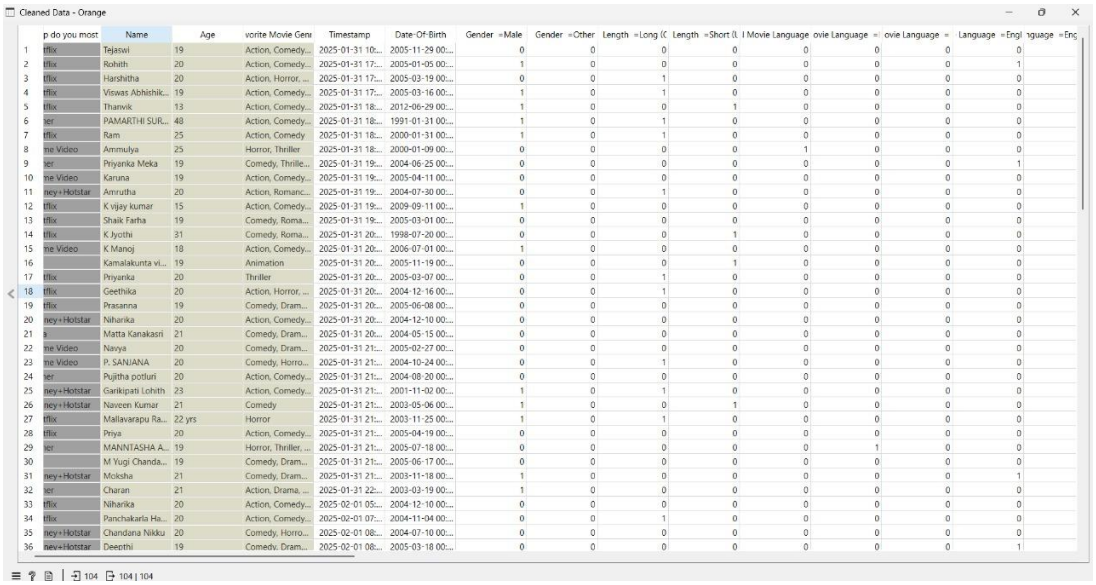


fig:2.4 Pre-processed dataset

- Save Preprocessed Data – Export the cleaned dataset for further use.

Step 3: 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.

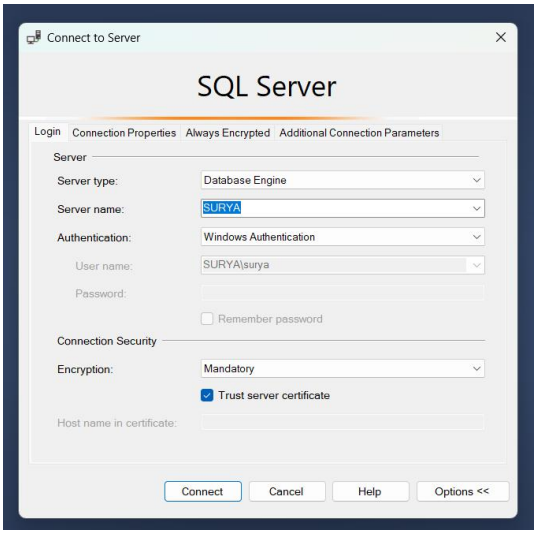


fig:3.1 Database Engine

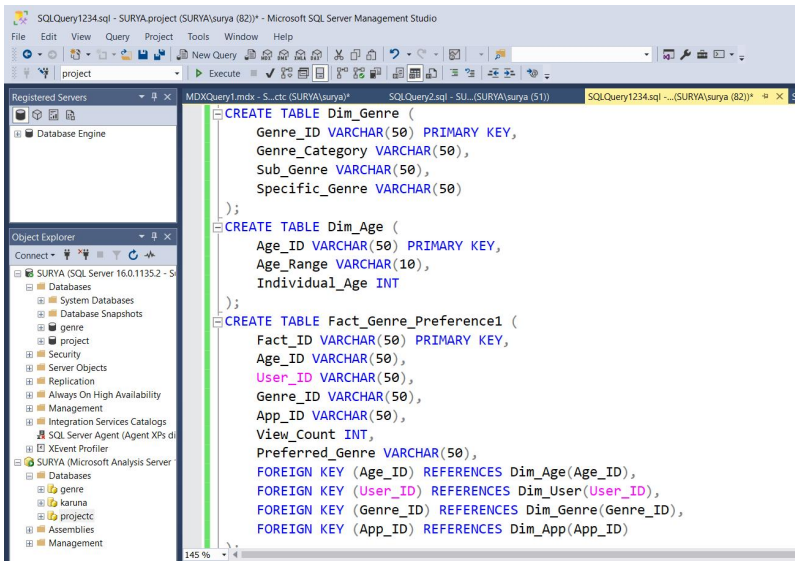


fig:3.2 SQL queries inserted for schema creation

Dimension Tables:

1.User_Dim:

Attribute Name	Type	Key
User_ID	VARCHAR(50)	Primary Key
Name	VARCHAR(100)	
Gender	VARCHAR(10)	
Language	VARCHAR(50)	
City	VARCHAR(50)	
State	VARCHAR(50)	
Country	VARCHAR(50)	

2.Genre_Dim:

Attribute Name	Type	Key
Genre_ID	VARCHAR(50)	Primary Key
Genre_Category	VARCHAR(50)	
Sub_Genre	VARCHAR(50)	
Specific_Genre	VARCHAR(50)	

3.App_Dim:

Attribute Name	Type	Key
App_ID	VARCHAR(50)	Primary Key
App_Name	VARCHAR(100)	
Subscription_Type	VARCHAR(50)	
App_Type	VARCHAR(50)	

4.Age_Dim:

Attribute Name	Type	Key
Age_ID	VARCHAR(50)	Primary Key
Age_Range	VARCHAR(10)	
Individual_Age	INT	

5.Duration_Dim:

Attribute Name	Type	Key
Duration_ID	VARCHAR(50)	Primary Key
Length_ID	VARCHAR(50)	Foreign Key (References Length_Dim)
Hours	INT	
Minutes	INT	
Seconds	INT	

6.Subscription_Dim:

Attribute Name	Type	Key
Subscription_ID	VARCHAR(50)	Primary Key
Subscription_Type	VARCHAR(50)	
Cost	INT	
Date_ID	INT	Foreign Key (References Date_Dim)

Sub-Dimension Tables:

1.Length_Dim:

Attribute Name	Type	Key
Length_ID	VARCHAR(50)	Primary Key
Total_Minutes	INT	

2.Date_Dim:

Attribute Name	Type	Key
Date_ID	INT	Primary Key
Start_Date	DATE	
Year	INT	
Quarter	VARCHAR(5)	
Month	VARCHAR(20)	
Day	INT	

Fact Tables:

1.Fact_Active_Subscribers:

Attribute Name	Type	Key
Fact_ID	VARCHAR(50)	Primary Key
Subscription_ID	VARCHAR(50)	Foreign Key (References Subscription_Dim)
App_ID	VARCHAR(50)	Foreign Key (References App_Dim)
User_ID	VARCHAR(50)	Foreign Key (References User_Dim)
Age_ID	VARCHAR(50)	Foreign Key (References Age_Dim)
Active_Subscriber_Count	INT	

2.Fact_User_App_Engagement:

Attribute Name	Type	Key
Fact_ID	VARCHAR(50)	Primary Key
Age_ID	VARCHAR(50)	Foreign Key (References Age_Dim)
User_ID	VARCHAR(50)	Foreign Key (References User_Dim)
Duration_ID	VARCHAR(50)	Foreign Key (References Duration_Dim)
App_ID	VARCHAR(50)	Foreign Key (References App_Dim)
View_Count	INT	
Preferred_Genre	VARCHAR(50)	
App_usage_by_age_group	VARCHAR(50)	

3.Fact_Genre_Preference:

Attribute Name	Type	Key
Fact_ID	VARCHAR(50)	Primary Key
Age_ID	VARCHAR(50)	Foreign Key (References Age_Dim)
User_ID	VARCHAR(50)	Foreign Key (References User_Dim)
Genre_ID	VARCHAR(50)	Foreign Key (References Genre_Dim)
App_ID	VARCHAR(50)	Foreign Key (References App_Dim)
View_Count	INT	
Preferred_Genre	VARCHAR(50)	

- Now Create a database “Project” to insert all these tables.
- Generate SQL queries to create and insert data into all the tables in SQL Server Database Engine

Step 4: Visualized all the Schemas in Visual Studio

- Created analysis service multidimensional project in Visual Studio.
- Defined Data Source & Data Source View – Connected the database and define table relationship
- Generated database diagrams for schemas.
- Verified relationships between tables.
- Created Cubes & Measures – Defined fact tables, measures, and dimensions.

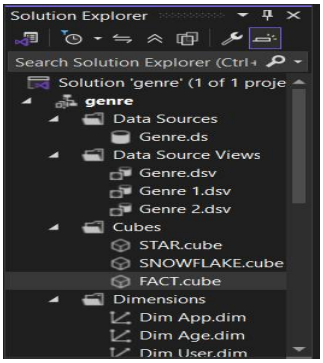
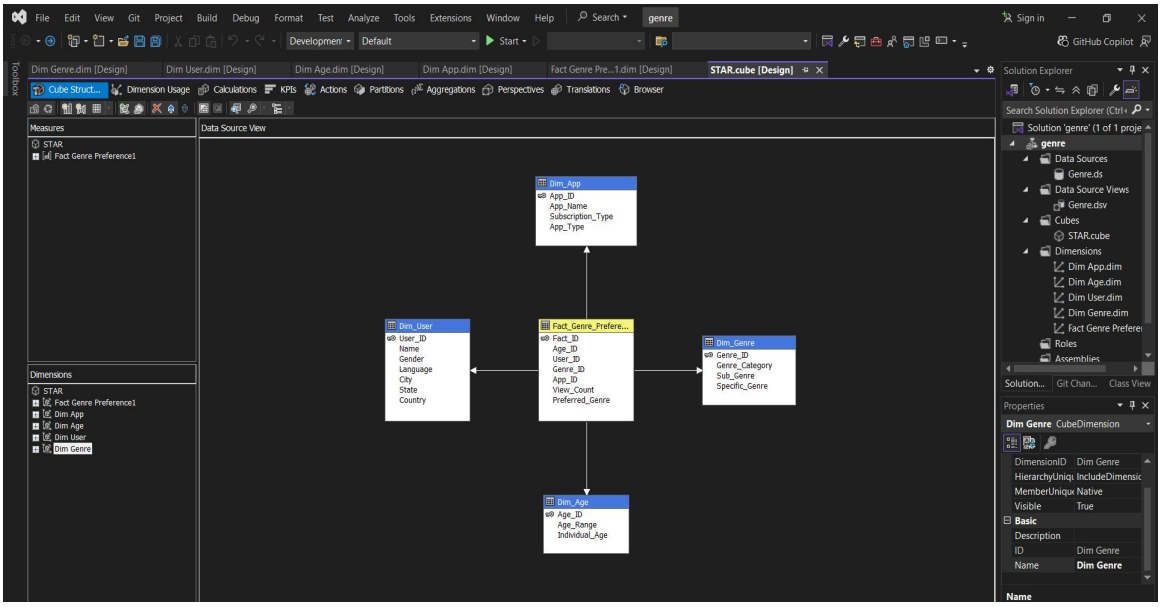


fig:4.1 Schema visualization in Visual Studio

These are the schemas that we have created

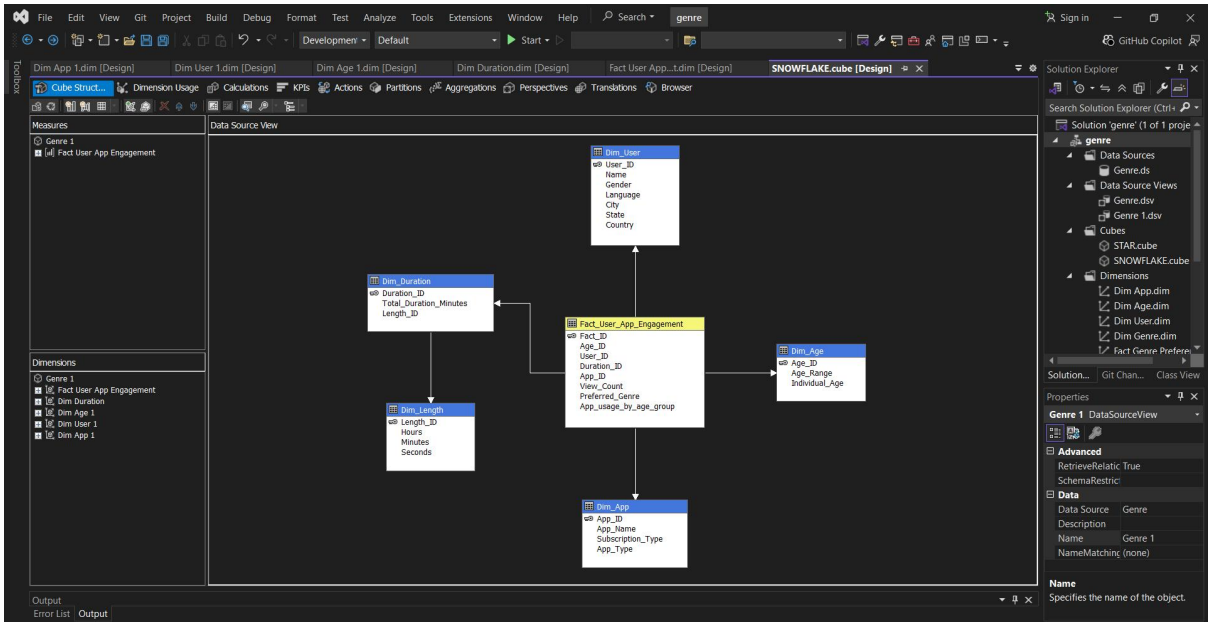
Star Schema:



Star schema has one Fact table and four Dimension tables.

- Fact table is Fact_Genre_Preference.
- Dimension tables are Dim_User, Dim_Genre, Dime_Age, Dim_app.

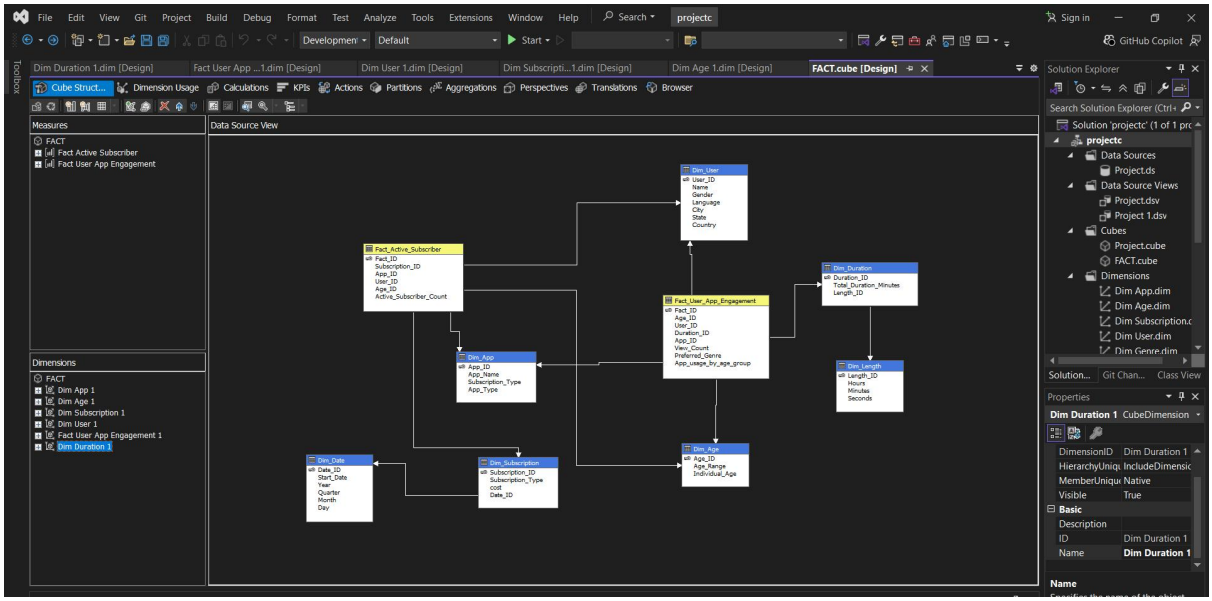
Snowflake Schema:



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

- Fact table is Fact_User_App_Preference.
- Dimension tables are Dim_User, Dim_Age, Dim_App, Dim_Duration, Dim_length.

Fact Constellation:



Fact Constellation has two Fact tables, seven Dimension tables and two Sub-Dimension tables.

- Fact tables are Fact_User_App_Preference, Fact_Active_Subscriber.
- Dimension tables are Dim_User, Dim_Age, Dim_App, Dim_Duration, Dim_length, Dim_Date,Dim_Sbscription.

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

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

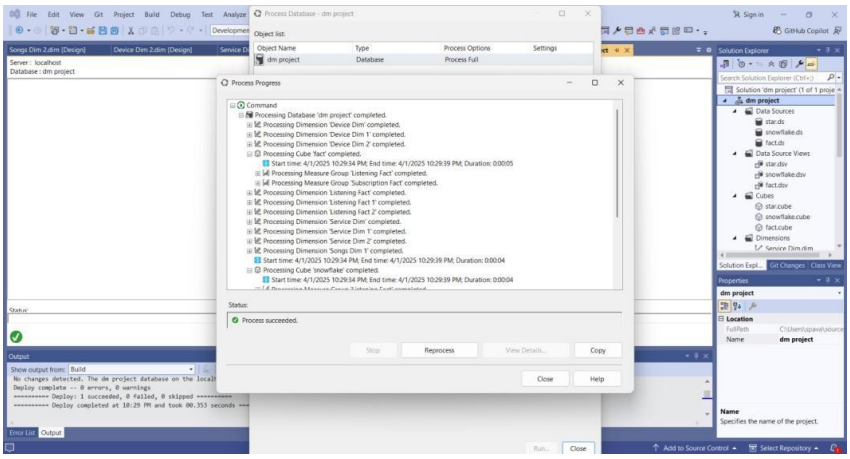


fig:5.1 Build & Deploying the project

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

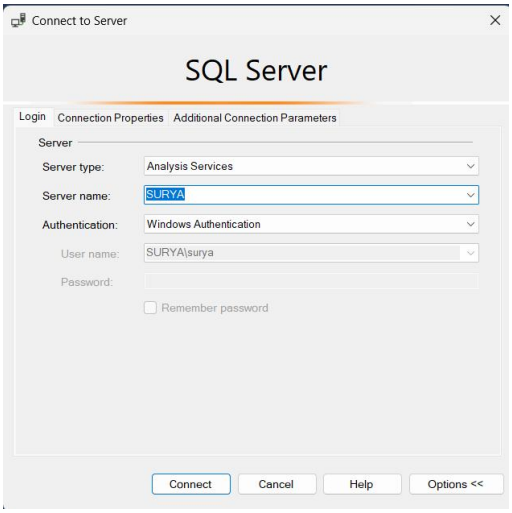


fig:5.2 Connection to Analysis Services

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

- 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.

Concept hierarchies:

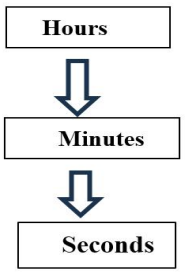


fig: Duration Hierarchy

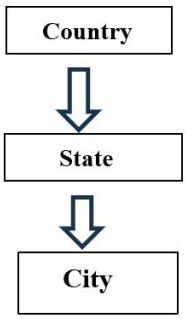


fig: Geographical Hierarchy

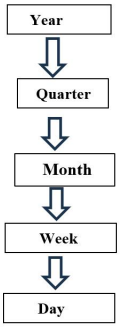
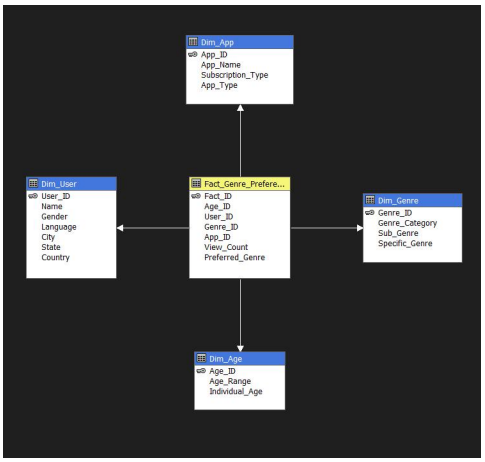


fig: Time Hierarchy

MDX Queries OLAP operations in star Schema:



Roll-up: Roll up on Cities to States

A(i): Which combination of streaming platform, age group, and genre generates the highest views and preferences in each state?"

Query:

```
SELECT
   {[Measures].[Fact Genre Preference1 Count]} ON COLUMNS,
   NON EMPTY
   (
      [Dim App].[App Name].[App Name] *
      [Dim Age].[Age Range].[Age Range] *
      [Dim Genre].[Genre Category].[Genre Category] *
      [Dim User].[State].[State]
   ) ON ROWS
FROM [STAR]
```

Output:

				Fact Genre Preference1 Count
Amazon Prime Video	13-19	Comedy	England	1
Amazon Prime Video	20-30	Action	Andhra Pradesh	1
Amazon Prime Video	20-30	Action	Texas	1
Amazon Prime Video	20-30	Animation	Maharashtra	1
Amazon Prime Video	20-30	Comedy	Maharashtra	1
Amazon Prime Video	20-30	Fantasy	Texas	1
Disney+	13-19	Comedy	Andhra Pradesh	1
Disney+	13-19	Comedy	New York	1
Disney+	13-19	Horror	Tamil Nadu	1
Disney+	20-30	Action	Andhra Pradesh	2
Disney+	20-30	Action	Illinois	1
Disney+	20-30	Comedy	Michigan	1
Disney+	20-30	Drama	Pennsylvania	1
HBO Max	13-19	Comedy	Andhra Pradesh	1
HBO Max	13-19	Horror	Madhya Pradesh	1
HBO Max	20-30	Action	Scotland	1
HBO Max	20-30	Horror	Andhra Pradesh	1
HBO Max	20-30	Romance	Delhi	1
HBO Max	40+	Action	Karnataka	1
Hulu	20-30	Comedy	Tamil Nadu	1
Netflix	13-19	Action	California	2
Netflix	13-19	Action	Telangana	1
Netflix	13-19	Action	Texas	1
Netflix	13-19	Animation	England	1
Netflix	13-19	Comedy	Andhra Pradesh	1
Netflix	13-19	Comedy	California	1

Execution Time: 1ms

Roll-up: Roll up on States to Countries

A(ii): How do genre and view counts vary across age groups and countries, and which platforms dominate in specific regions?

Query:

```
SELECT
   {[Measures].[Fact Genre Preference1 Count]} ON COLUMNS,
   NON EMPTY
   (
      [Dim App].[App Name].[App Name]*
      [Dim Age].[Age Range].[Age Range] *
      [Dim Genre].[Genre Category].[Genre Category] *
      [Dim User].[Country].[Country]
   ) ON ROWS
FROM [STAR]
SELECT
   {[Measures].[View Count], [Measures].[Fact Genre Preference Count]} ON COLUMNS,
   NON EMPTY
   [Dim User].[Continent].[Asia]ON ROWS
FROM [star]
```

Output:

				Fact Genre Preference1 Count
Amazon Prime Video	13-19	Comedy	UK	1
Amazon Prime Video	20-30	Action	India	1
Amazon Prime Video	20-30	Action	USA	1
Amazon Prime Video	20-30	Animation	India	1
Amazon Prime Video	20-30	Comedy	India	1
Amazon Prime Video	20-30	Fantasy	USA	1
Disney+	13-19	Comedy	India	1
Disney+	13-19	Comedy	USA	1
Disney+	13-19	Horror	India	1
Disney+	20-30	Action	India	2
Disney+	20-30	Action	USA	1
Disney+	20-30	Comedy	USA	1
Disney+	20-30	Drama	USA	1
HBO Max	13-19	Comedy	India	1
HBO Max	13-19	Horror	India	1
HBO Max	20-30	Action	UK	1
HBO Max	20-30	Horror	India	1
HBO Max	20-30	Romance	India	1
HBO Max	40+	Action	India	1
Hulu	20-30	Comedy	India	1
Netflix	13-19	Action	India	1
Netflix	13-19	Action	USA	3
Netflix	13-19	Animation	UK	1
Netflix	13-19	Comedy	India	2
Netflix	13-19	Comedy	UK	1

Execution Time: 1ms

Drill down: Drill down on Age group to individual age

B: How does genre preference shift between age ranges and individual ages and which streaming platforms best cater to these granular preferences?

Query:

```
SELECT
  {
    [Measures].[Fact Genre Preference1 Count]
  } ON COLUMNS,
  NON EMPTY
  (
    [Dim App].[App Name].[App Name] *
    [Dim Genre].[Genre Category].[Genre Category] *
    [Dim Age].[Age Range].[Age Range] *
    [Dim Age].[Individual Age].[Individual Age]
  ) ON ROWS
FROM [STAR]
```

Output:

				Fact Genre Preference1 Count
Amazon Prime Video	Action	20-30	20	2
Amazon Prime Video	Animation	20-30	21	1
Amazon Prime Video	Comedy	13-19	19	1
Amazon Prime Video	Comedy	20-30	21	1
Amazon Prime Video	Fantasy	20-30	20	1
Disney+	Action	20-30	20	2
Disney+	Action	20-30	23	1
Disney+	Comedy	13-19	19	2
Disney+	Comedy	20-30	20	1
Disney+	Drama	20-30	20	1
Disney+	Horror	13-19	19	1
HBO Max	Action	20-30	21	1
HBO Max	Action	40+	48	1
HBO Max	Comedy	13-19	19	1
HBO Max	Horror	13-19	19	1

Execution Time: 1ms

Slice: Slice for Genre Category

C:What are the view counts and genre preference counts for each major genre category (e.g., Action, Comedy, Drama)?

Query:

```
SELECT
  {[Measures].[ Fact Genre Preference1]} ON COLUMNS,
  NON EMPTY
  CROSSJOIN(
    {[Dim Age].[Individual Age].[Individual Age].MEMBERS},
    CROSSJOIN(
      {[Dim App].[App Name].[App Name].MEMBERS},
      CROSSJOIN(
        {[Dim User].[Gender].[Gender].MEMBERS}
      )
    )
  ) ON ROWS
FROM [STAR]
WHERE
  ([Dim Genre].[Genre Category].[Genre Category].&[Action])
```

Output:

				Fact Genre Preference1 Count
15	Netflix	Male	U12	1
18	Netflix	Male	U15	1
19	Netflix	Female	U01	1
19	Netflix	Female	U38	1
20	Amazon Prime Video	Male	U02	1
20	Amazon Prime Video	Male	U39	1
20	Disney+	Female	U11	1
20	Disney+	Female	U18	1
20	Netflix	Female	U20	1
20	Netflix	Female	U33	1
21	HBO Max	Male	U32	1
23	Disney+	Male	U25	1
25	Netflix	Male	U07	1
48	HBO Max	Male	U06	1

Execution Time:1ms

Dice: Dice for (App Name=”Netflix, Amazon Prime” , Age =”13,18,25” , Genre Category=”Comedy, Action”, Gender=”Male, Female”) and (Measures= “view Count” , ” Fact Genre Preference1 Count”)

D:"What is the view count of Action and Comedy genres for users aged 18, 25, and 13, using Netflix and Amazon Prime apps, segmented by gender and user ID?"

Query:

```
SELECT
  {[Measures].[View Count]} ON COLUMNS,
  NON EMPTY
  CROSSJOIN(
    {
     [Dim Age].[Individual Age].[Individual Age].&[18],
     [Dim Age].[Individual Age].[Individual Age].&[25],
     [Dim Age].[Individual Age].[Individual Age].&[30]
    },
    CROSSJOIN(
      {
       [Dim App].[App Name].[App Name].&[Netflix],
       [Dim App].[App Name].[App Name].&[Amazon Prime]
      },
      CROSSJOIN(
        {
         [Dim Genre].[Genre Category].[Genre Category].&[Action],
         [Dim Genre].[Genre Category].[Genre Category].&[Comedy]
        },
        CROSSJOIN(
          {
           [Dim User].[Gender].[Gender].&[Male],
           [Dim User].[Gender].[Gender].&[Female]
          },
          {[Fact Genre Preference1].[User ID].[User ID].MEMBERS}
        )
      )
    )
  ) ON ROWS
FROM [STAR]
```

Output:

					Fact Genre Preference1 Count
18	Netflix	Action	Male	U15	2
25	Netflix	Action	Male	U07	1
13	Netflix	Comedy	Male	U05	1

Pivot:

E:When pivoting the data, how do the view counts and genre preference counts distribute across different Apps when displayed on columns instead of rows?

Query:

```
SELECT
NON EMPTY
   [Dim App].[App Name].Children ON COLUMNS,
```

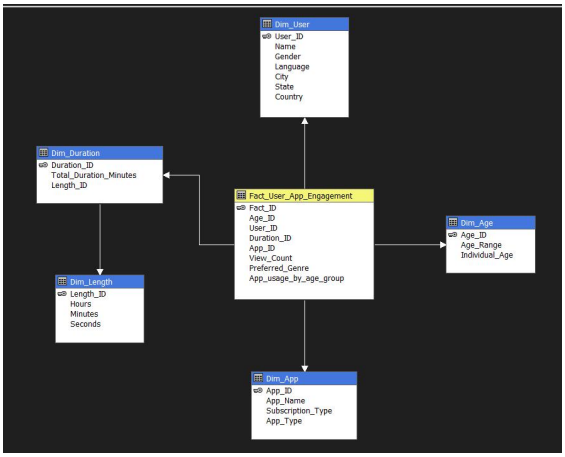
```
NON EMPTY
   {
      [Measures].[Fact Genre Preference1 Count]
   }  ON ROWS
FROM [STAR]
```

Output:

	Amazon Prime Video	Disney+	HBO Max	Hulu	Netflix
Fact Genre Preference1 Count	6	8	6	1	19

Execution Time: 1ms

MDX Queries for OLAP operations in Snow Flake Schema:



Roll-up: Roll up on Seconds to Minutes

A(i): How does user engagement time(minutes) vary across different states, age groups, and apps based on the duration of usage?

Query:

```
SELECT
  {
    [Measures].[Fact User App Engagement Count]
  } ON COLUMNS,
  NON EMPTY
  (
    [Dim Duration].[Minutes].[Minutes] *
    [Dim Age 1].[Age Range].[Age Range] *
    [Dim User 1].[State].[State] *
    [Dim App 1].[App Name].[App Name]
  ) ON ROWS
FROM [SNOWFLAKE]
```

Output:

				Fact User App Engagement Count
0	20-30	Michigan	Amazon Prime Video	1
0	20-30	Telangana	JioCinema	1
0	20-30	Texas	Amazon Prime Video	1
10	13-19	California	MX Player	1
10	20-30	Andhra Pradesh	SonyLIV	1
10	20-30	Maharashtra	SonyLIV	1
10	20-30	Scotland	MX Player	1
15	13-19	New York	Apple TV+	1
15	40+	Karnataka	Apple TV+	1
20	20-30	Andhra Pradesh	Peacock	1
20	20-30	England	Peacock	1
20	20-30	Tamil Nadu	ManoramaMAX	1
20	20-30	Telangana	ManoramaMAX	1
25	20-30	Scotland	Paramount+	1
25	20-30	Tamil Nadu	Paramount+	1
25	20-30	Texas	Voot	1
25	30-40	Andhra Pradesh	Voot	1
30	13-19	Telangana	Netflix	1
30	20-30	Andhra Pradesh	HBO Max	1
30	20-30	Maharashtra	Netflix	1
30	Under 13	California	HBO Max	1
35	13-19	California	Discovery+	1
35	20-30	Illinois	Discovery+	1
40	13-19	Andhra Pradesh	ALTBalaji	1
40	13-19	Texas	ALTBalaji	1

Execution Time: 1ms

Roll-up: Roll up on Seconds to Hours

A(ii): How does user engagement time(hours) vary across different states, age groups, and apps based on the duration of usage?

Concept Hierarchy: Duration(Seconds->Minutes->Hours)

Query:

```
SELECT
  {
    [Measures].[Fact User App Engagement Count]
  } ON COLUMNS,
  NON EMPTY
  (
    [Dim Duration].[Hours].[Hours] *
    [Dim Age 1].[Age Range].[Age Range] *
    [Dim User 1].[State].[State] *
    [Dim App 1].[App Name].[App Name]
  ) ON ROWS
FROM [SNOWFLAKE]
```

Output:

				Fact User App Engagement Count
0	Under 13	California	HBO Max	1
1	13-19	Andhra Pradesh	Hotstar	1
1	13-19	California	MX Player	1
1	13-19	Madhya Pradesh	Hotstar	1
1	13-19	New York	Apple TV+	1
1	13-19	Telangana	Netflix	1
1	20-30	Andhra Pradesh	Eros Now	1
1	20-30	England	Disney+	1
1	20-30	Maharashtra	Netflix	1
1	20-30	Scotland	MX Player	1
1	20-30	Tamil Nadu	Disney+	1
1	20-30	Telangana	Eros Now	1
1	20-30	Texas	Voot	1
1	30-40	Andhra Pradesh	Voot	1
1	40+	Karnataka	Apple TV+	1
2	13-19	England	Aha	1
2	13-19	England	Sun NXT	1
2	20-30	Andhra Pradesh	Aha	1
2	20-30	Andhra Pradesh	Peacock	1
2	20-30	Andhra Pradesh	SonyLIV	1
2	20-30	England	Peacock	1
2	20-30	Maharashtra	SonyLIV	1
2	20-30	Michigan	Amazon Prime Video	1
2	20-30	Pennsylvania	Sun NXT	1

Execution Time: 1ms

Drill down: Drill down on Country to City

B: "How does user engagement with different apps vary across countries, states, cities, and age groups?"

Query:

```
SELECT
  {
    [Measures].[Fact User App Engagement Count]
  } ON COLUMNS,
  NON EMPTY
  (
    [Dim User 1].[Country].[Country] *
    [Dim User 1].[State].[State] *
    [Dim User 1].[City].[City] *
    [Dim App 1].[App Name].[App Name] *
    [Dim Age 1].[Age Range].[Age Range]
  ) ON ROWS
FROM [SNOWFLAKE]
```

Output:

					Fact User App Engagemen
India	Andhra Pradesh	Guntur	HBO Max	20-30	1
India	Andhra Pradesh	Kadapa	Eros Now	20-30	1
India	Andhra Pradesh	Kakinada	Voot	30-40	1
India	Andhra Pradesh	Nellore	ALTBalaji	13-19	1
India	Andhra Pradesh	Rajahmundry	SonyLIV	20-30	1
India	Andhra Pradesh	Tirupati	JioCinema	13-19	1
India	Andhra Pradesh	Vijayawada	Aha	20-30	1
India	Andhra Pradesh	Vijayawada	Hotstar	13-19	1
India	Andhra Pradesh	Visakhapatnam	Peacock	20-30	1
India	Delhi	Delhi	Hulu	20-30	1
India	Karnataka	Bengaluru	Apple TV+	40+	1
India	Madhya Pradesh	Bhopal	Hotstar	13-19	1
India	Maharashtra	Nagpur	SonyLIV	20-30	1
India	Maharashtra	Pune	Netflix	20-30	1
India	Tamil Nadu	Chennai	Disney+	20-30	1
India	Tamil Nadu	Chennai	Hulu	13-19	1
India	Tamil Nadu	Coimbatore	Paramount+	20-30	1
India	Tamil Nadu	Trichy	ManoramaMAX	20-30	1
India	Telangana	Hyderabad	JioCinema	20-30	1
India	Telangana	Hyderabad	ManoramaMAX	20-30	1
India	Telangana	Hyderabad	Netflix	13-19	1
India	Telangana	Warangal	Eros Now	20-30	1
UK	England	Birmingham	ZEE5	13-19	1
UK	England	Leeds	Peacock	20-30	1

Execution Time: 1ms

Slice: Slice for App Name

C:For users who watched content on Netflix, how does user engagement (view count and app engagement count) vary across different age groups, states, and durations?

Query:

```
SELECT
  {[Measures].[Fact User App Engagement1 Count]} ON COLUMNS,
  NON EMPTY
  CROSSJOIN(
    {[Dim User 1].[User ID].[User ID].MEMBERS},
    CROSSJOIN(
      {[Dim Age 1].[Age Range].[Age Range].MEMBERS},
      CROSSJOIN(
        {[Dim User 1].[State].[State].MEMBERS},
                    {[Dim Duration3].[Duration ID].[Duration ID].MEMBERS}

      )
    )
  ) ON ROWS
FROM [SNOWFLAKE]
WHERE
  ([Dim App 1].[App Name].[App Name].&[Netflix])
```

Output:

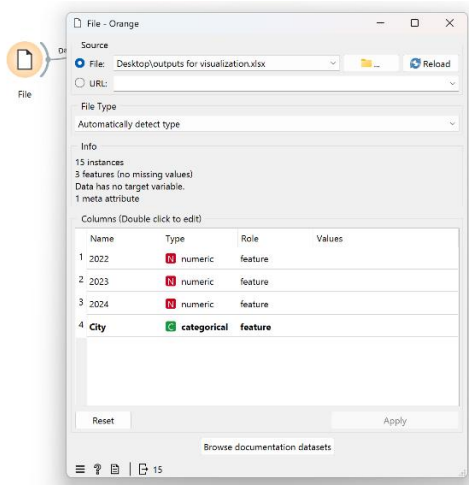
Messages | Results

				Fact User App Engagement1 Count
U01	13-19	Telangana	D01	1
U21	20-30	Maharashtra	D01	1

Execution Time: 1ms

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.



- Then using the bar plot we visualized them.

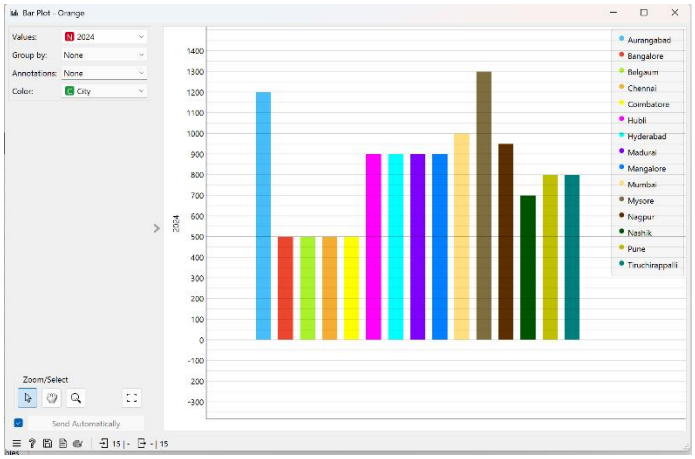


Figure: 1: Visualization Bar plot

The above Figure 1 shows the bar plot of view count for the year 2024 across various cities in the states of Karnataka, Maharashtra, etc...

Dice: Dice for (App Name=”Netflix” or “Aha”, using user id, duration id and age id)

D: MDX DICE query to retrieve a sub cube that displays the View Count and total number of records (Fact User App Engagement1 Count) for all combinations of User ID, Age ID, and Duration ID, but only for the apps Netflix and Aha.

Query:

```
SELECT
   {[Measures].[Fact User App Engagement1 Count]} ON COLUMNS,
   NON EMPTY
```

```
    CrossJoin(
      [Dim User 1].[User ID].[User ID].Members,
      CrossJoin(
        {[Dim App 1].[App Name].[Netflix],[Dim App 1].[App Name].[Aha]},
        CrossJoin(
          [Dim Age 1].[Age ID].[Age ID].Members,
          [Dim Duration3].[Duration ID].[Duration ID].Members
        )
      )
    ) ON ROWS
FROM
    [SNOWFLAKE]
```

Output:

				Fact User App Engagement1 Count
U01	Netflix	AG01	D01	1
U19	Aha	AG19	D19	1
U21	Netflix	AG21	D01	1
U39	Aha	AG39	D19	1

Execution Time:1 ms

Pivot:

E: When pivoting the data, how do the view counts and engage counts distribute across different Apps and age ranges when displayed on columns instead of rows?

Query:

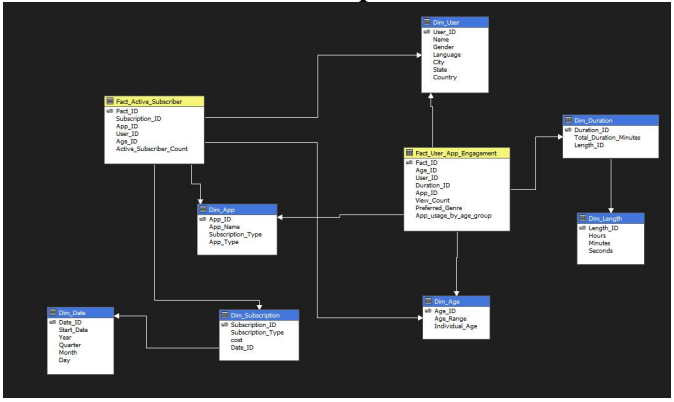
```
SELECT
  {
    [Measures].[Fact User App Engagement Count]
    } ON ROWS,
  NON EMPTY
  (
    [Dim App 1].[App Name].[App Name] *
    [Dim Age 1].[Age Range].[Age Range]
  ) ON COLUMNS
FROM [SNOWFLAKE]
```

Output:

	Aha	Aha	ALTBalaji	Amazon Prime Video	Apple TV+	Apple TV+	Discovery+	Discovery+	Disney+	Eros Now	HBO Max	HBO Max	Hotstar	Hulu	Hulu	JioCinema
	13-19	20-30	13-19	20-30	13-19	40+	13-19	20-30	20-30	20-30	20-30	Under 13	13-19	13-19	20-30	13-19
Fact User App Engagement Count	1	1	2	2	1	1	1	1	2	2	1	1	2	1	1	1

Execution Time: 1ms

MDX Queries for OLAP operations in Fact Constellation:



Roll-up: Roll up on Day to Quarters

A(i): What is the total app engagement, view count, and active subscriber count grouped by user, age range, app name, and subscription quarter?

Query:

```
SELECT
   NON EMPTY {
      [Measures].[Fact User App Engagement Count],
      [Measures].[Fact Active Subscriber Count]
   } ON COLUMNS,
   NON EMPTY (
      [Dim User 1].[Name].[Name] *
      [Dim App 1].[App Name].[App Name] *
      [Dim Subscription 1].[Subscription Type].[Subscription Type] *
      [Dim Subscription 1].[Quarter].[Quarter].Members
   ) ON ROWS
FROM [FACT]
```

Output:

				Fact User App Engagement Count	Fact Active Subscriber Count
Ammulya	Paramount+	Weekly	Q4	1	(null)
Ammulya	Paramount+	Yearly	Q1	1	(null)
Ammulya	Paramount+	Yearly	Q2	1	(null)
Ammulya	Paramount+	Yearly	Q3	1	(null)
Ammulya	Paramount+	Yearly	Q4	1	(null)
Ammulya	Paramount+	Unknown	Unknown	1	(null)
Amrutha	SonyLIV	Monthly	Q1	1	1
Amrutha	SonyLIV	Monthly	Q2	1	(null)
Amrutha	SonyLIV	Monthly	Q3	1	(null)
Amrutha	SonyLIV	Monthly	Q4	1	(null)
Amrutha	SonyLIV	Quarterly	Q1	1	(null)
Amrutha	SonyLIV	Quarterly	Q3	1	(null)
Amrutha	SonyLIV	Quarterly	Q4	1	(null)
Amrutha	SonyLIV	Weekly	Q1	1	(null)
Amrutha	SonyLIV	Weekly	Q2	1	(null)
Amrutha	SonyLIV	Weekly	Q4	1	(null)
Amrutha	SonyLIV	Yearly	Q1	1	(null)
Amrutha	SonyLIV	Yearly	Q2	1	(null)
Amrutha	SonyLIV	Yearly	Q3	1	(null)
Amrutha	SonyLIV	Yearly	Q4	1	(null)
Amrutha	SonyLIV	Unknown	Unknown	1	(null)
Chandana Nikku	Acorn TV	Weekly	Q1	(null)	1
Chandana Nikku	ALTBalaji	Monthly	Q1	1	(null)
Chandana Nikku	ALTBalaji	Monthly	Q2	1	(null)
Chandana Nikku	ALTBalaji	Monthly	Q3	1	(null)
Chandana Nikku	ALTBalaji	Monthly	Q4	1	(null)
Chandana Nikku	ALTBalaji	Quarterly	Q1	1	(null)
Chandana Nikku	ALTBalaji	Quarterly	Q3	1	(null)

Execution Time: 1ms

Roll-up: Roll up on Day to Years

A(ii): What is the total app engagement, view count, and active subscriber count grouped by user, age range, app name, and subscription year?

Query:

```
SELECT
   NON EMPTY {[Measures].[Fact User App Engagement Count],[Measures].[Fact Active Subscriber Count]} ON COLUMNS,
   NON EMPTY
         (

                  [Dim User 1].[Name].[Name]*
                  [Dim Age 1].[Age Range].[Age Range]*
                  [Dim App 1].[App Name].[App Name]*
                  [Dim Subscription 1].[Year].[Year].Members
         )        ON ROWS
FROM [FACT]
```

Output:

				Fact User App Engagement Count	Fact Active Subscriber Count
Ammulya	20-30	Paramount+	2022	1	(null)
Ammulya	20-30	Paramount+	2023	1	(null)
Ammulya	20-30	Paramount+	2024	1	1
Ammulya	20-30	Paramount+	Unknown	1	(null)
Amrutha	20-30	SonyLIV	2022	1	(null)
Amrutha	20-30	SonyLIV	2023	1	(null)
Amrutha	20-30	SonyLIV	2024	1	1
Amrutha	20-30	SonyLIV	Unknown	1	(null)
Chandana Nikku	13-19	Acorn TV	2024	(null)	1
Chandana Nikku	13-19	ALTBalaji	2022	1	(null)
Chandana Nikku	13-19	ALTBalaji	2023	1	(null)
Chandana Nikku	13-19	ALTBalaji	2024	1	(null)
Chandana Nikku	13-19	ALTBalaji	Unknown	1	(null)
Charan	13-19	Popcornflix	2024	(null)	1
Charan	20-30	MX Player	2022	1	(null)
Charan	20-30	MX Player	2023	1	(null)
Charan	20-30	MX Player	2024	1	(null)
Charan	20-30	MX Player	Unknown	1	(null)
Deepthi	13-19	BritBox	2023	(null)	1
Deepthi	20-30	Sun NXT	2022	1	(null)
Deepthi	20-30	Sun NXT	2023	1	(null)
Deepthi	20-30	Sun NXT	2024	1	(null)
Deepthi	20-30	Sun NXT	Unknown	1	(null)
Dhru	13-19	Xumo	2024	(null)	1
Dhru	20-30	Aha	2022	1	(null)
Dhru	20-30	Aha	2023	1	(null)
Dhru	20-30	Aha	2024	1	(null)
Dhru	20-30	Aha	Unknown	1	(null)
Garikipati Lohith	20-30	HBO Max	2022	1	(null)
Garikipati Lohith	20-30	HBO Max	2023	1	(null)

Execution Time: 1ms

Drill down: Drill down on Year to Day

B: "How does user engagement with different user and app id's vary across Year, Quarter, Month, Day?"

Query:

```
SELECT
   NON EMPTY {[Measures].[Fact User App Engagement Count], [Measures].[Fact Active Subscriber Count]} ON COLUMNS,
   NON EMPTY
   (
            [Dim User 1].[User ID].[User ID] *
               [Dim App 1].[App ID].[App ID] *
             [Dim Subscription 1].[Year].[Year]*
            [Dim Subscription 1].[Quarter].[Quarter] *
            [Dim Subscription 1].[Month].[Month]*
            [Dim Subscription 1].[Day].[Day]
       ) ON ROWS
FROM [FACT];
```

Output:

						Fact User App Engagement Count	Fact Active Subscriber Count
U01	A01	2022	Q2	June	6	1	(null)
U01	A01	2022	Q4	November	1	1	(null)
U01	A01	2023	Q2	June	30	1	(null)
U01	A01	2023	Q2	May	10	1	(null)
U01	A01	2023	Q3	August	10	1	(null)
U01	A01	2023	Q3	August	8	1	(null)
U01	A01	2023	Q3	July	1	1	(null)
U01	A01	2023	Q3	September	14	1	(null)
U01	A01	2023	Q3	September	9	1	(null)
U01	A01	2023	Q4	December	15	1	(null)
U01	A01	2023	Q4	December	20	1	(null)
U01	A01	2023	Q4	December	30	1	(null)
U01	A01	2023	Q4	November	11	1	(null)
U01	A01	2023	Q4	November	21	1	(null)
U01	A01	2023	Q4	October	1	1	(null)
U01	A01	2023	Q4	October	18	1	(null)
U01	A01	2023	Q4	October	25	1	(null)
U01	A01	2024	Q1	February	10	1	(null)

Execution Time: 1ms

Slice: Slice for Subscription Type

C: What are the view counts, Active Subscriber Counts for Subscription Type (Yearly)?

Query:

```
SELECT
   {[Measures].[Fact Active Subscribers Count]} ON COLUMNS,
   NON EMPTY
   CrossJoin(
      [Dim User].[User ID].[User ID].Members,
      CrossJoin(
```

```
        [Dim Age].[Age ID].[Age ID].Members,
        CrossJoin(
          [Dim App].[App Name].[App Name].Members,
          [Dim Duration].[Duration ID].[Duration ID].Members
        )
      )
  ) ON ROWS
FROM
  [Fct]
WHERE
  ([Dim Subscription].[Subscription Type].[Yearly])
```

Output:

				Fact Active Subscribers Count
U01	AG01	Netflix	D01	(null)
U02	AG02	Amazon Prime Video	D01	1
U02	AG02	Amazon Prime Video	D02	1
U02	AG02	Amazon Prime Video	D03	1
U02	AG02	Amazon Prime Video	D04	1
U02	AG02	Amazon Prime Video	D05	1
U02	AG02	Amazon Prime Video	D06	1
U02	AG02	Amazon Prime Video	D07	1
U02	AG02	Amazon Prime Video	D08	1
U02	AG02	Amazon Prime Video	D09	1
U02	AG02	Amazon Prime Video	D10	1

Execution Time:1ms

Dice: Dice On (App Name = Netflix, Hotstar), (Subscription Type = Monthly), (Age ID = AG01, AG02), (User ID=U01,U02), (Duraton ID=D01,D02), and Measures = View Count, Fact User App Engagement Count, Fact Active Subscriber Count).

D: How do engagement metrics compare for users U01 and U02, aged between AG01 and AG02, using Netflix and Hotstar on monthly subscriptions, during short (D01) and extended (D05) viewing sessions?

Query:

```
SELECT
  NON EMPTY {
    [Measures].[Fact User App Engagement Count],
    [Measures].[Fact Active Subscribers Count]
  } ON COLUMNS,

  NON EMPTY
  (
    {
      [Dim Age].[Age ID].[AG01],
      [Dim Age].[Age ID].[AG02]
    } *
    {
      [Dim App].[App Name].[Netflix],
      [Dim App].[App Name].[Hotstar]
    } *
    {
      [Dim Duration].[Duration ID].[D01],
      [Dim Duration].[Duration ID].[D05]
    } *
    {
      [Dim Subscription].[Subscription Type].[Monthly]
    } *
    {
      [Dim User].[User ID].[U01],
      [Dim User].[User ID].[U02]
    }
  ) ON ROWS

FROM [Fact]
```

Output:

Messages | Results

					Fact User App Engagement Count	Fact Active Subscribers Count
AG01	Netflix	D01	Monthly	U01	1	1
AG01	Netflix	D05	Monthly	U01	(null)	1

Execution Time:1 ms

Pivot:

E: When pivoting the data, how do the view counts, engage counts, subscriber counts distribute across different Subscription costs when displayed on columns instead of rows?

Query:

```
SELECT
  NON EMPTY
  {
    [Measures].[Fact User App Engagement Count],
    [Measures].[Fact Active Subscriber Count]
  } ON ROWS,
  NON EMPTY
  [Dim Subscription 1].[Cost]. Members ON COLUMNS
FROM [FACT]
```

Output:

	All	1099	399	599	Unknown
Fact User App Engagement Count	40	40	40	40	40
Fact Active Subscriber Count	40	7	18	15	(null)

Execution Time: 1ms

Step 7: 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.

	User ID	Age ID	App	Length	View Coun	Fact Active	Subscribers Count
2	U01	AG01	Netflix	100	120	(null)	
3	U02	AG02	Amazon Pr	100	(null)	1	
4	U02	AG02	Amazon Pr	102	80	1	
5	U02	AG02	Amazon Pr	105	(null)	1	
6	U02	AG02	Amazon Pr	108	(null)	1	
7	U02	AG02	Amazon Pr	112	(null)	1	
8	U02	AG02	Amazon Pr	115	(null)	1	
9	U02	AG02	Amazon Pr	116	(null)	1	
10	U02	AG02	Amazon Pr	117	(null)	1	
11	U02	AG02	Amazon Pr	119	(null)	1	
12	U02	AG02	Amazon Pr	120	(null)	1	
13	U02	AG02	Amazon Pr	121	(null)	1	
14	U02	AG02	Amazon Pr	122	(null)	1	
15	U02	AG02	Amazon Pr	124	(null)	1	
16	U02	AG02	Amazon Pr	125	(null)	1	
17	U02	AG02	Amazon Pr	126	(null)	1	
18	U02	AG02	Amazon Pr	127	(null)	1	

fig:7.1 Dataset in excel sheet

Fields available in the above data set:

- User id
- Age id
- App name

- Length(total mins)
- Fact active subscribers count

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

fig:7.2 Workflow for Distributions

- Here for 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)

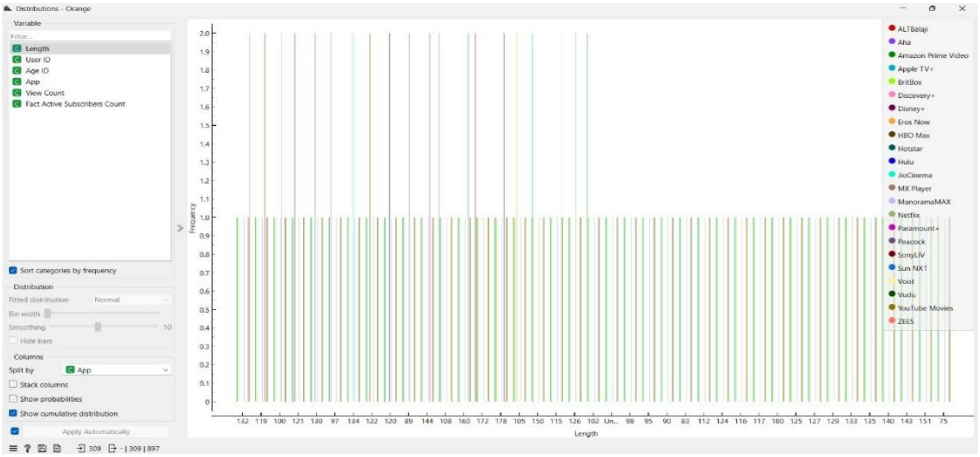


fig:7.3 Bar Plot Visualization

- This bar plot presents the frequency of movie durations (lengths) viewed on various streaming platforms, with each bar color-coded according to a specific app.

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 type of viewers based on the movie length (Eg: long, short, medium etc..)
- The Most suitable model for my preprocessed dataset is Classification. The target class has 104 values in my classification model.

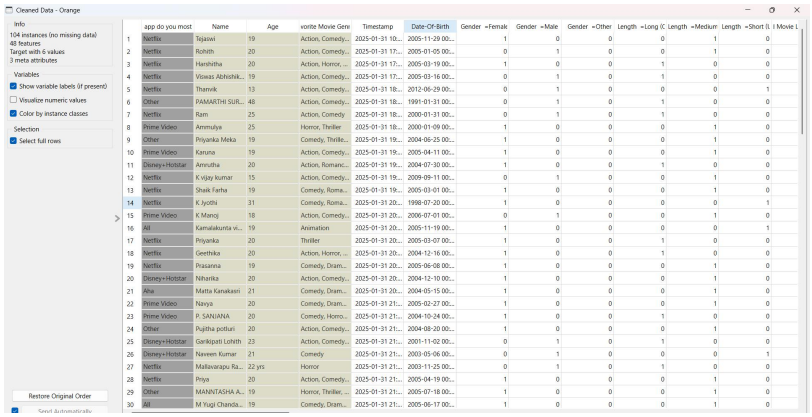
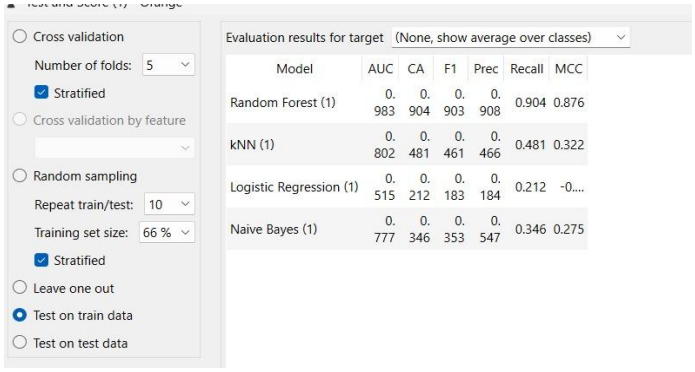


fig:8.3 Preprocessed Dataset

Fig 8.3 shows the data after applying preprocessing techniques and continuize discrete variables(One feature per value)

- 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.



Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest (1)	0.983	0.904	0.903	0.908	0.904	0.876
kNN (1)	0.802	0.481	0.461	0.466	0.481	0.322
Logistic Regression (1)	0.515	0.212	0.183	0.184	0.212	-0....
Naive Bayes (1)	0.777	0.346	0.353	0.547	0.346	0.275

fig:8.4 Test & Score of Preprocessed Dataset

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

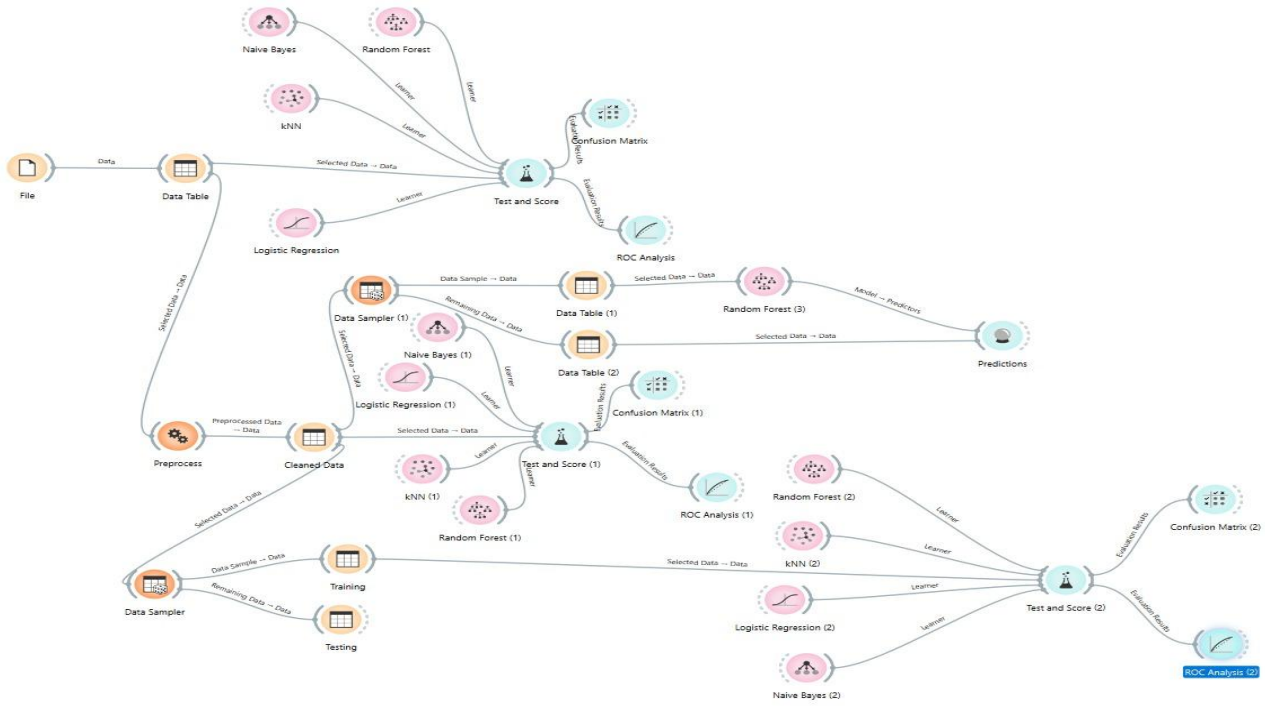


Fig:8.5 Entire Work Flow

Fig 8.5 demonstrates a multi-model approach for classification, where data is sampled, fed into various algorithms (e.g., Logistic Regression, Random Forest, Naive Bayes), and then evaluated using metrics such as Test and Score, Confusion Matrices, and ROC Analysis. It allows quick comparison of model performance on the same dataset, enabling a data scientist to select the most effective classifier.

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.

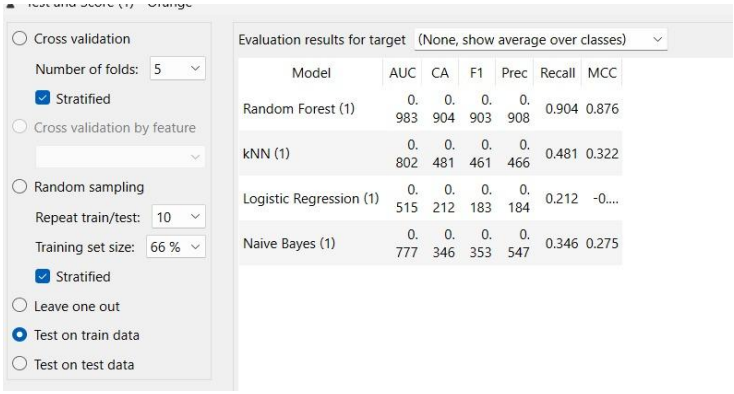


Figure 3 Test & Score measurements

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

	MODELS	CA	F1	PREC	RECALL
WITHOUT PREPROCESSING	KNN	0.481	0.461	0.466	0.481
	NAÏVE BAYES	0.327	0.365	0.508	0.327
	RANDOM FOREST	0.788	0.785	0.808	0.788
	LOGISTIC REGRESSION	0.212	0.183	0.184	0.212
WITH PREPROCESSING (WITHSAMPLER)	KNN	0.481	0.461	0.466	0.481
	NAÏVE BAYES	0.346	0.353	0.547	0.346
	LOGISTIC REGRESSION	0.212	0.183	0.184	0.212
	RANDOM FOREST	0.904	0.903	0.908	0.904

Table 1: Model Performance Comparison with and without Preprocessing & Sampling

3.1 PREDICTION ANALYSIS :

Prediction model:

- Now based on the classifier accuracy we developed a prediction model by splitting 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.

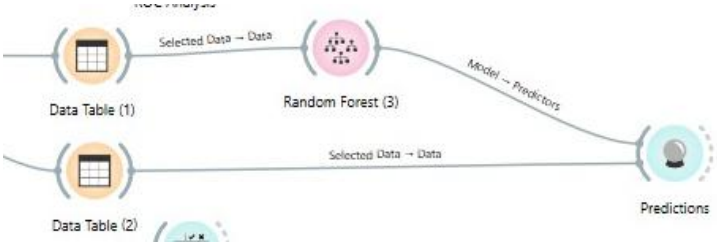


fig:3.1.1 Prediction Workflow

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

Predictions - Orange

Show probabilities for: Classes in data ☑ Show classification errors

	Random Forest (3)	error	app do you most	Name	Age	vorite Movie Genr
1	0.00 : 0.02 : 0.02 : 0.41 : 0.41 : 0.15 → Netflix	0.588	Netflix	Harshitha	20	Action, Horror, ...
2	0.00 : 0.03 : 0.22 : 0.39 : 0.27 : 0.10 → Netflix	0.733	Other	Pujitha potluri	20	Action, Comedy...
3	0.15 : 0.00 : 0.30 : 0.28 : 0.22 : 0.05 → Disney+Hotstar	0.695	Disney+Hotstar	srinivas	21	Drama, Horror, ...
4	0.17 : 0.00 : 0.30 : 0.32 : 0.05 : 0.15 → Netflix	0.681	Netflix	tarun	50	Action, Comedy...
5	0.08 : 0.12 : 0.15 : 0.22 : 0.37 : 0.05 → Other	0.628	Other	Chinna	17	Action, Drama, ...
6	0.00 : 0.06 : 0.22 : 0.10 : 0.58 : 0.03 → Other	0.780	Disney+Hotstar	saranya	14	Comedy, Anima...
7	0.05 : 0.07 : 0.07 : 0.07 : 0.57 : 0.16 → Other	0.929	Disney+Hotstar	phani	5	Animation, Doc...
8	0.00 : 0.03 : 0.11 : 0.64 : 0.20 : 0.02 → Netflix	1.000	Aha	Matta Kanakasri	21	Comedy, Dram...
9	0.05 : 0.03 : 0.30 : 0.17 : 0.33 : 0.12 → Other	0.704	Disney+Hotstar	T Varun	18	Action, Horror, ...
10	0.11 : 0.05 : 0.09 : 0.19 : 0.53 : 0.03 → Other	0.813	Netflix	Priyanka	20	Comedy, Dram...
11	0.00 : 0.08 : 0.11 : 0.58 : 0.06 : 0.17 → Netflix	0.827	Prime Video	K Manoj	18	Action, Comedy...
12	0.05 : 0.03 : 0.06 : 0.32 : 0.55 : 0.00 → Other	0.455	Other	charulata	24	Horror, Thriller, ...
13	0.12 : 0.00 : 0.30 : 0.11 : 0.33 : 0.15 → Other	0.883	Aha	VEMURI BHARA...	20	Action, Comedy...
14	0.09 : 0.00 : 0.32 : 0.10 : 0.32 : 0.17 → Disney+Hotstar	0.684	Other	T Lakshmi sri	20	Romance, Horror

Fig:3.1.2 Prediction Classifications

- 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 Random Forest)

Predicted

Actual	Aha	All	Disney+Hotstar	Netflix	Other	Prime Video	Σ
Aha	5	0	0	2	0	0	7
All	0	3	0	0	0	0	3
Disney+Hotstar	0	0	27	1	1	0	29
Netflix	0	0	2	26	1	0	29
Other	0	0	0	1	22	0	23
Prime Video	0	0	2	0	0	11	13
Σ	5	3	31	30	24	11	104

Figure 3.1.3: Confusion matrix of Random Forest

Figure 3.1.3 shows that the classification performance of the model is generally good at predicting users who prefer Disney+Hotstar, Netflix, and Prime Video, but it struggles more with less popular categories like "Aha" and "Other".

3.2 VISUALISATION METRICS FOR CLASSIFICATION MODEL

- Now coming to the Visualization of the model we selected a barplot for our classification model.

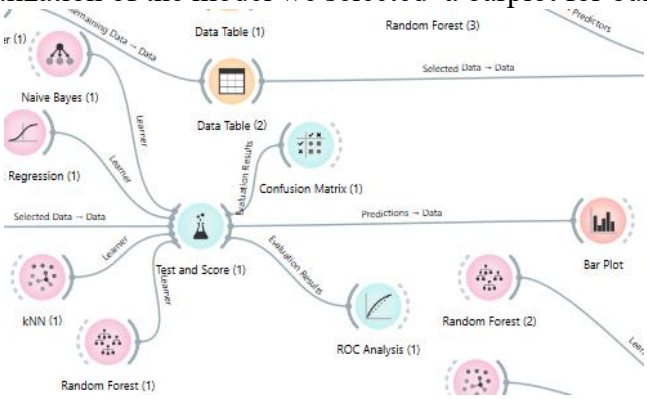


fig 3.2.1:Workflow for Bar Plot

- The bar plot for the Classification model is as follows

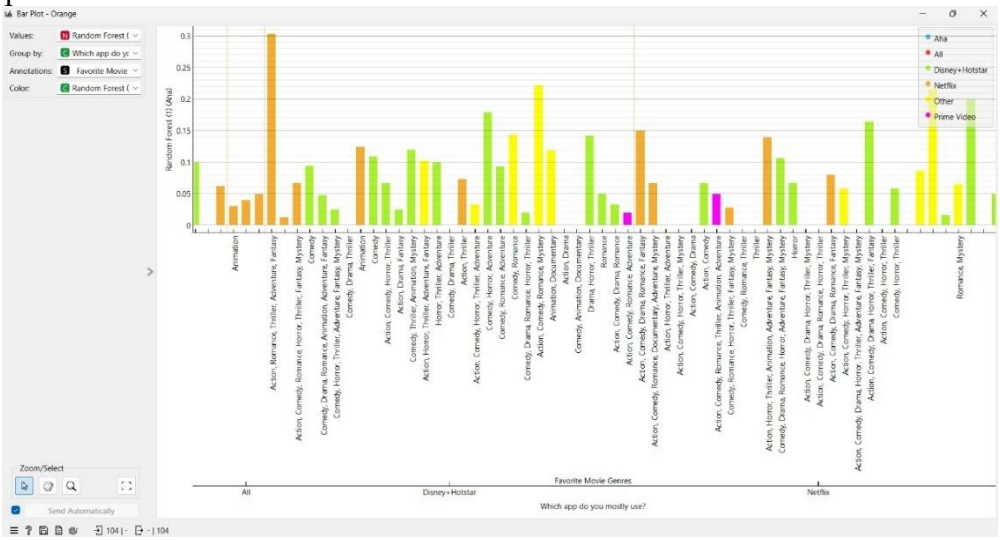


Figure 3.2.2:Visualisation of Bar plot for classification model

Figure 3.2.2 shows that the Random Forest is able to **identify key patterns** for certain classes more clearly (e.g., Netflix).

CONCLUSION:

In this project , we successfully implemented an end-to-end data-driven approach to classify or analyze users' preferred streaming platforms based on demographic data. A total of 104 records were analyzed, representing users' actual preferences for popular OTT platforms like **Aha**, **Disney+Hotstar**, **Netflix**, **Prime Video**, and others. I began by collecting and preprocessing the dataset, ensuring data quality through cleaning, normalization, and feature engineering. Following this, I designed and implemented OLAP schemas (Star, Snowflake, and Fact Constellation), inserted data into fact and dimension tables using Oracle SQL in SSMS, and visualized these schemas in Visual Studio.

After deploying the OLAP schemas on the SSAS server, we performed OLAP operations such as Drill-Down, Roll-Up, Slice, and Dice to extract meaningful insights from user data. The OLAP results were then visualized using Orange Tool, providing a clear representation of user behavior trends.
Finally, we conducted a classification analysis to categorize OTT users as Genre Viewers, Subscribers preferred the app. Multiple machine learning models were evaluated, and the best-performing model was identified using accuracy, precision, recall, F1-score.

This part of the project demonstrates the effectiveness of OLAP and machine learning in understanding user behavior on Movie viewers. The insights gained can be leveraged for personalized recommendations, targeted marketing, and content optimization, ultimately enhancing the user experience.

KEY FINDINGS:

- **Netflix's Broad Appeal:** Dominates across multiple genres—especially action, thriller, and drama—thanks to a diverse and extensive content library that attracts a wide range of viewers.

- **Amazon Prime's Versatility:** Shows strong performance in comedy and drama with exclusive releases, appealing to a mixed demographic through its varied offerings.

- **Disney+ Hotstar's Family Focus:** Leads in family-friendly and animated genres, leveraging its strong brand identity (Disney, Marvel, etc.) to capture households with younger viewers.

- **Genre-Specific Strengths:** Each platform exhibits niche strengths—Netflix for action and thrillers, Amazon Prime for comedy and drama, and Disney+ Hotstar for kid-friendly content—with some smaller platforms like Voot capturing regional or specialized interests.

- **Viewer Demographic Trends:** Teen and young adult audiences prefer trending series on Netflix, while families are drawn to the safe, brand-backed content on Disney+ Hotstar, and Amazon Prime maintains a balanced appeal across multiple age groups.

Even though the chart does not explicitly break down data by age or other demographic factors, some inferences can be made:

- **Teen and Young Adult Audiences**: These viewers often gravitate toward **Netflix** for trending series, teen dramas, and YA-friendly content.
- **Families and Younger Children**: The data spikes for **Disney+ Hotstar** imply a strong preference in households with children (due to Pixar, Marvel, Disney classics, etc.).
- **Broad or Mixed Demographics**: **Amazon Prime's** consistent presence across multiple genres suggests that it appeals to a wide range of viewer ages and tastes, potentially due to its bundled ecosystem (e.g., Prime shipping + video).

PART-B: Predicting Horse Health Outcomes Using Classification Models
CHAPTER 1: INTRODUCTION ON DATA MINING METHODOLOGY

1.1 Problem Statement:

This project aims to develop a classification-based predictive model that analyzes various physiological and clinical features of horses to determine their health status. By identifying key indicators associated with health deterioration or recovery, the model seeks to support veterinarians and caretakers in making informed, data-driven decisions for improved equine healthcare management.

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

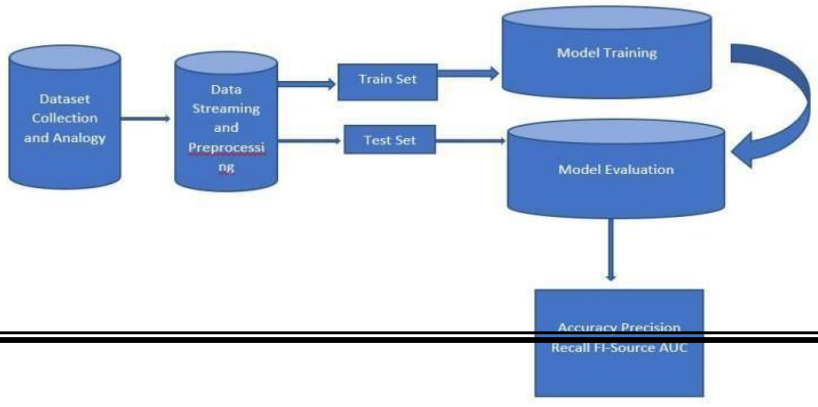
The Horse-Colic dataset contains attributes about individuals. These include:

- Surgery
- Age
- Pulse, RR, RectalTemp
- Outcome
- SurgLes, LesTypeSite1-3

The objective is to use these features to classify the **final health result** of the horse.

1.2.2 Methodology

We need the processes the dataset and make sure there are no redundancies test various classification algorithms and then develop the prediction model by training with training dataset



and testing it with the testing dataset.

1.2.3 Machine Learning Models

We intend to use Supervised Machine Learning models:

- Neural Network
- Gradient Boosting
- Random Forest
- SVM.

1.2.4 Evaluation Metrics

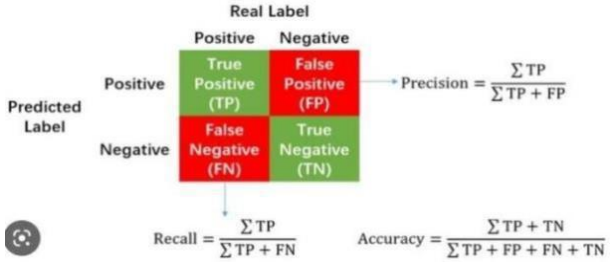
Since this is a classification problem, we can use:

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.

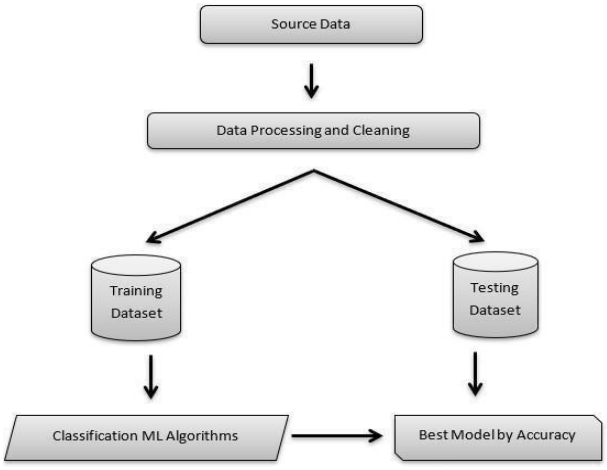


Confusion matrix

Block Diagram

The diagram illustrates the machine learning workflow for classification:

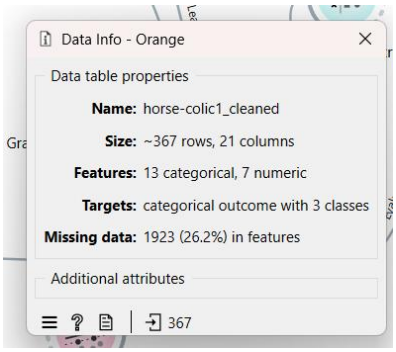
1. Source Data undergoes Data Processing and Cleaning to remove inconsistencies and prepare it for analysis.
2. The dataset is split into Training and Testing sets, ensuring proper evaluation.
3. Classification algorithms are applied to the training set, and the best model is selected based on accuracy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



CHAPTER 2: ANALYSIS ON THE DATASET

Data Set Description:

This dataset contains **370 records** of horses with colic symptoms, featuring **28 attributes** including clinical, physiological, and pathological features. It includes variables like temperature, pulse, mucous membrane color, and abdominal conditions, along with the **final health outcome** (e.g., lived, died).
These features include:



Each entry in the **horse_colic** dataset provides valuable insights into equine health conditions and veterinary decision-making, aiding in the classification of colic severity based on key medical attributes. This dataset is instrumental in analyzing health dependencies, optimizing diagnosis processes, and improving future treatment methodologies for horses suffering from colic.

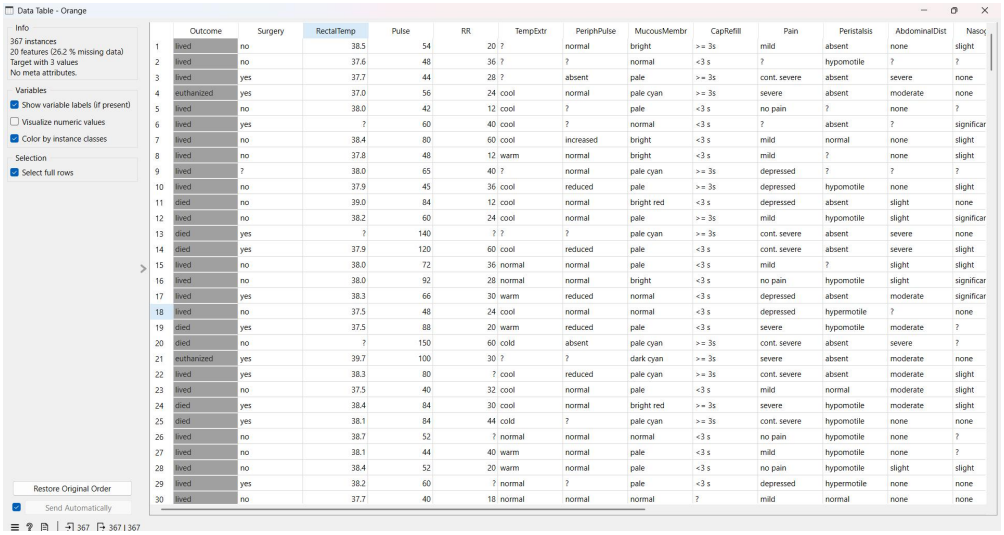
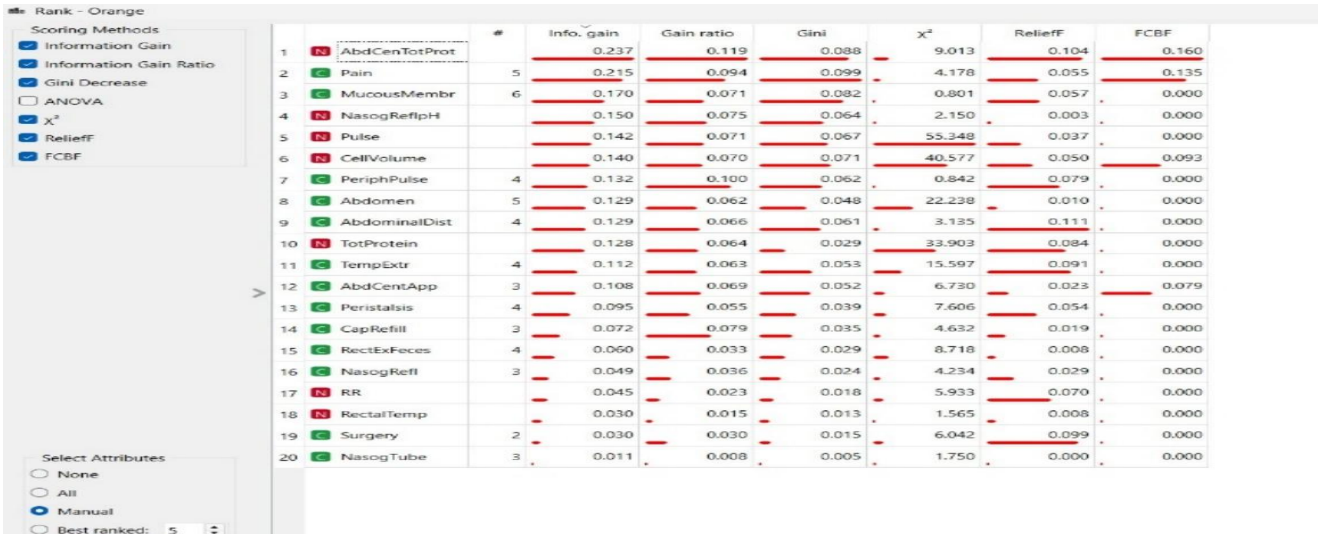


fig: Dataset before preprocessing

Feature Engineering:

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



		#	Info. gain	Gain ratio	Gini	χ^2	ReliefF	FCBF
1	AbdCenTotProt		0.237	0.119	0.088	9.013	0.104	0.160
2	Pain	5	0.215	0.094	0.099	4.178	0.055	0.135
3	MucousMembr	6	0.170	0.071	0.082	0.801	0.057	0.000
4	NasogReflpH		0.150	0.075	0.064	2.150	0.003	0.000
5	Pulse		0.142	0.071	0.067	55.348	0.037	0.000
6	CellVolume		0.140	0.070	0.071	40.577	0.050	0.093
7	PeriphPulse	4	0.132	0.100	0.062	0.842	0.079	0.000
8	Abdomen	5	0.129	0.062	0.048	22.238	0.010	0.000
9	AbdominalDist	4	0.129	0.066	0.061	3.135	0.111	0.000
10	TotProtein		0.128	0.064	0.029	33.903	0.084	0.000
11	TempExtr	4	0.112	0.063	0.053	15.597	0.091	0.000
12	AbdCentApp	3	0.108	0.069	0.052	6.730	0.023	0.079
13	Peristalsis	4	0.095	0.055	0.039	7.606	0.054	0.000
14	CapRefill	3	0.072	0.079	0.035	4.632	0.019	0.000
15	RectExFeces	4	0.060	0.033	0.029	8.718	0.008	0.000
16	NasogRefl	3	0.049	0.036	0.024	4.234	0.029	0.000
17	RR		0.045	0.023	0.018	5.933	0.070	0.000
18	RectalTemp		0.030	0.015	0.013	1.565	0.008	0.000
19	Surgery	2	0.030	0.030	0.015	6.042	0.099	0.000
20	NasogTube	3	0.011	0.008	0.005	1.750	0.000	0.000

fig : Feature Ranking in Orange

- The image shows the "Rank" widget in Orange, which ranks features based on different scoring methods.
- Various metrics like Information Gain, Gain Ratio, Gini, ANOVA, Chi-Square (χ^2), and ReliefF are used for ranking.
- The table displays features with their respective scores, visually represented by red bars.

Data Validation, Cleaning, and Preparation Process

We meticulously assessed the horse colic dataset to ensure its accuracy and readiness for medical analysis. The primary objective was to classify the severity or presence of colic in horses based on key clinical and physiological indicators. We began by identifying relevant variables such as pain levels, mucous membrane color, pulse, abdominal conditions, and others that are significant for diagnosis.

To enhance the dataset's reliability and improve model performance, the following data preprocessing techniques were applied using the Orange Data Mining tool:

- Handling Missing Values: Missing values were prevalent in several attributes such as rectal temperature, pulse, and abdominal distension. These were imputed using the most appropriate method (mean, median, or mode) based on model accuracy and variable type.
- Target Variable Imputation: Missing values in the target variable (colic outcome) were handled using the Impute widget to maintain consistency across samples.
- Normalization of Numerical Features: Attributes such as pulse rate, temperature, packed cell volume, and total protein were normalized to ensure a uniform scale, thus enhancing model convergence and comparability between features.

In real-world medical datasets, inaccuracies and inconsistencies are common. Hence, data validation was conducted by:

- Verifying data types (categorical vs. numerical)
- Ensuring a balanced distribution across target classes
- Checking for and removing duplicates or anomalies

Dataset Splitting

Given the relatively small sample size of the dataset, careful consideration was given to how the data was split for training and testing:

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

This approach allowed for effective training while preserving a sufficient portion for unbiased evaluation. Stratified sampling was applied to ensure balanced representation of colic severity classes across both sets.

Data Visualization

Using Orange's powerful visualization widgets, we analyzed the interdependencies and distributions of critical features:

- Scatter plots, box plots, and heatmaps were employed to observe the relationships between key features such as pain, pulse, mucous membranes, and abdominal conditions.
- Visual exploration helped identify outliers, clusters, and patterns linked to colic outcomes, guiding both feature selection and model design.

Machine Learning Techniques and Model Selection

We implemented and evaluated multiple machine learning algorithms to classify colic outcomes accurately. The following models were tested using the Test & Score widget in Orange:

1. K-Nearest Neighbor (KNN)
2. Random Forest
3. Logistic Regression
4. Naive Bayes
5. Stochastic Gradient Descent (SGD)

Model performance was compared based on:

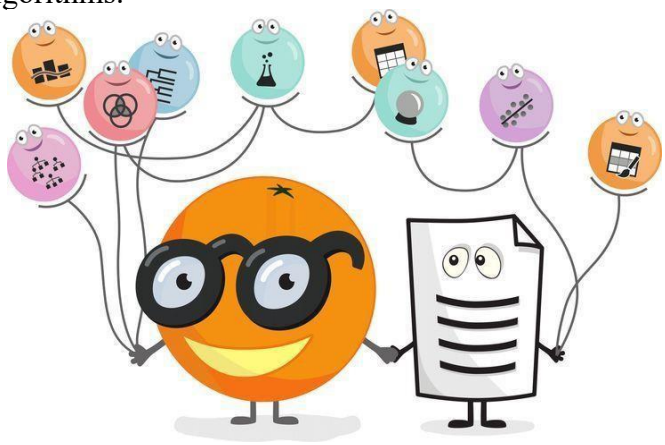
- Accuracy
- Precision/Recall
- F1 Score
- Area Under ROC Curve (AUC)

The Random Forest model emerged as the top performer with the highest classification accuracy, effectively capturing nonlinear relationships and feature interactions.

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.



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.”horse_colic_cleaned.csv”) and open it.

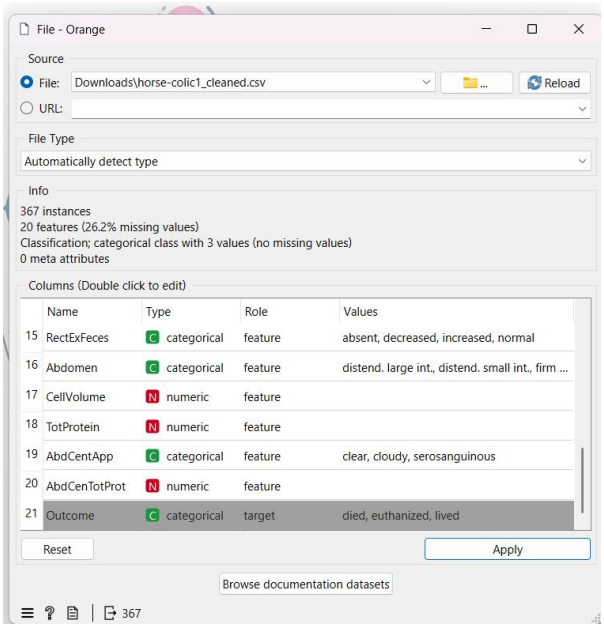


fig:2.1 Loading the dataset

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

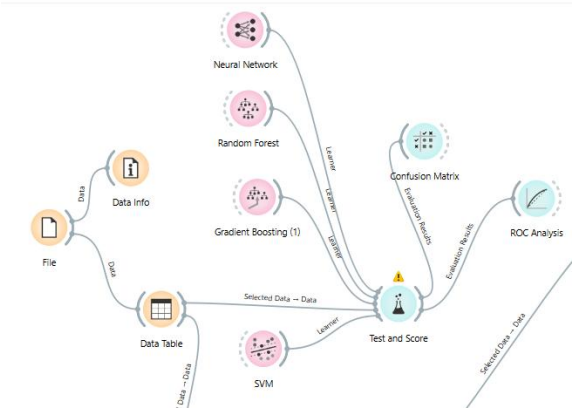
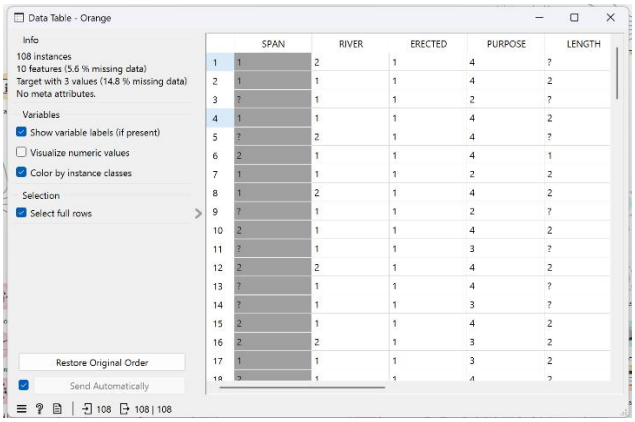


Fig:3.1 Before preprocessing

Data set before preprocessing:



- Accuracy,Precision,Recall,F1 are known for the dataset before preprocessing by using Test and Score widget

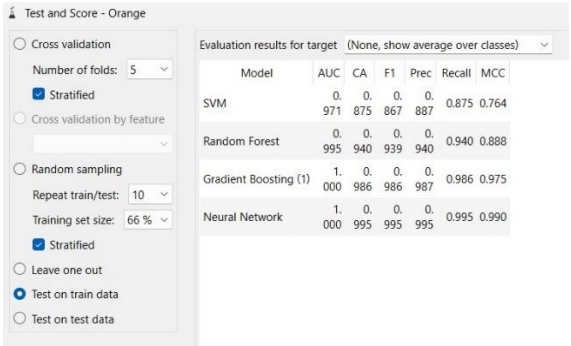


Fig:3.3 Test and Score before preprocessing

Step 4: Preprocessing dataset

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

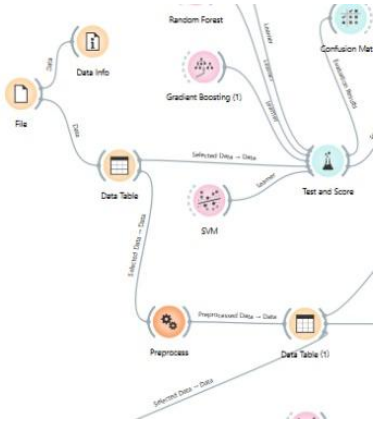


fig:4.1 Preprocessing

- Principal Component Analysis is performed in the pre-processing with various components as

Principal Component Analysis				
COMPONENTS	**SVM**	**GRADIENT BOOSTING**	**RANDOM FOREST**	**NEURAL NETWORKS**
10	0.801	0.992	0.946	0.853
20	0.856	0.995	0.967	0.973
30	0.869	0.995	0.962	0.992
50	0.8886	0.995	0.975	0.992

shown in the table

- After performing all the preprocessing techniques on the given dataset the data can be modified as shown in the fig:4.2

Step 5: Testing accuracy of various classification algorithms

- Drag and drop the "Test & Score" widget
- Connect the "KNN", ”Random Forest”, “Stochastic Gradient Descendent”, “Gradient Boosting”, ”Logistic Regression” 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.

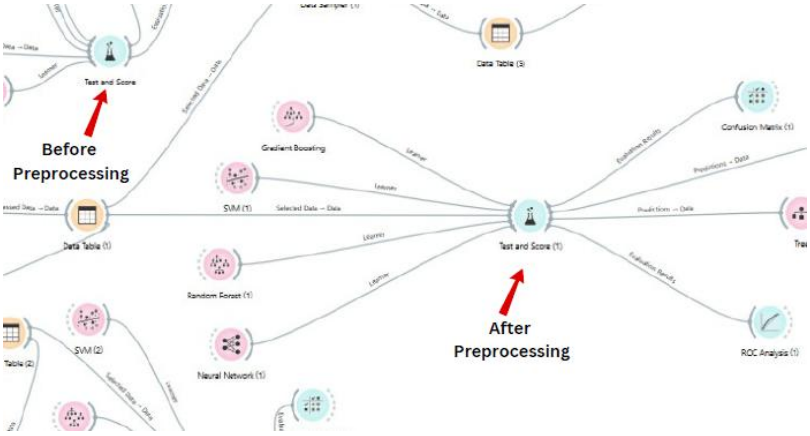


fig:5.1 Workflow for test & Score after preprocessing

- The Test and Score Widget is displayed as shown in the fig: 5.2

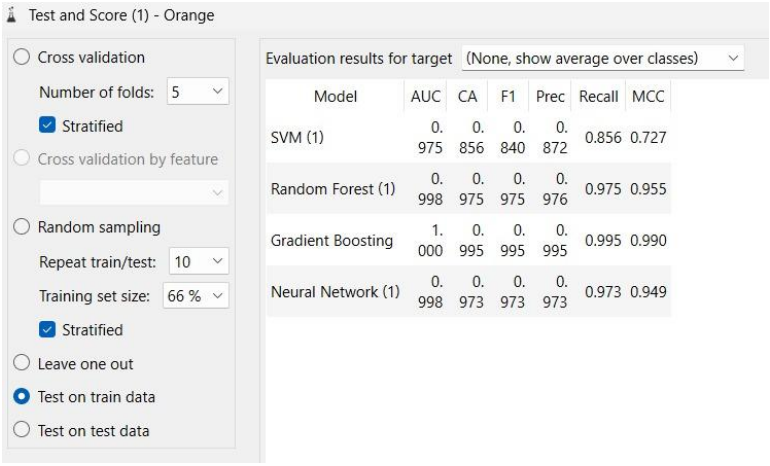


fig:5.2 Test & Score Measurements

- 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.

The below table shows readings of various evaluation metrics like AUC, CA, F1, Frequency, etc…

	MODELS	CA	F1	PREC	RECALL
WITHOUT PREPROCESSING	SVM	0.875	0.867	0.887	0.875
	NEURAL NETWORK	0.995	0.995	0.995	0.995
	RANDOM FOREST	0.940	0.939	0.940	0.940
	GRADIENT BOOSTING	0.986	0.986	0.987	0.986
WITH PREPROCESSING (WITHSAMPLER)	SVM	0.856	0.840	0.872	0.856
	NEURAL NETWORK	0.973	0.973	0.973	0.973
	GRADIENT BOOSTING	0.995	0.995	0.995	0.995
	RANDOM FOREST	0.975	0.975	0.976	0.975

Table : Model Performance Comparison with and without Preprocessing & Sampling

- **Neural Network** showed best CA before preprocessing and **Gradient Boosting** showed best CA

after applying preprocessing techniques.

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

- The prediction model needs both training and test data. Based on the training and test data the prediction model can be developed in two ways-
  First way is by splitting the dataset into training and test datasets using the data sampler
  This is clearly explained the figure below:

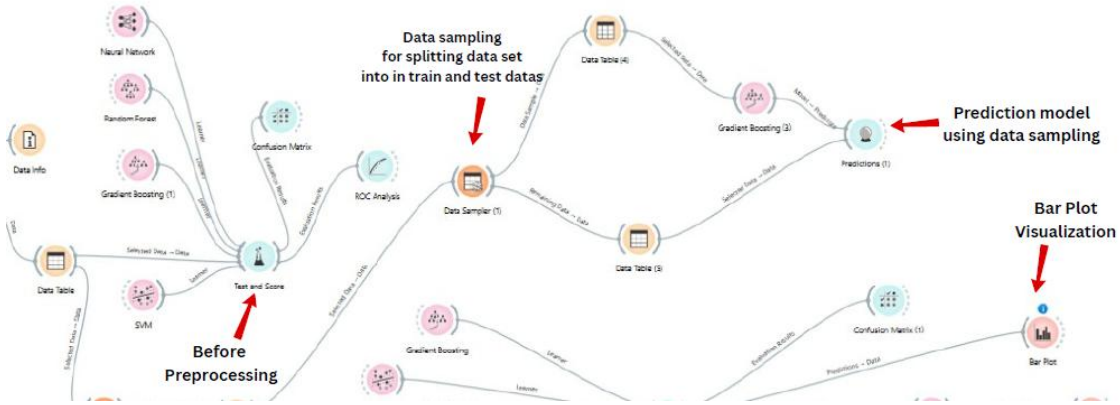


fig: 6.1Prediction model by data sampler

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:

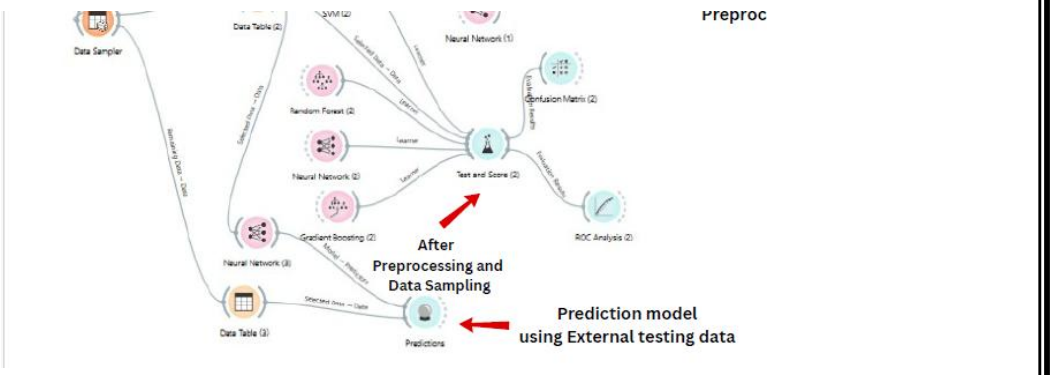
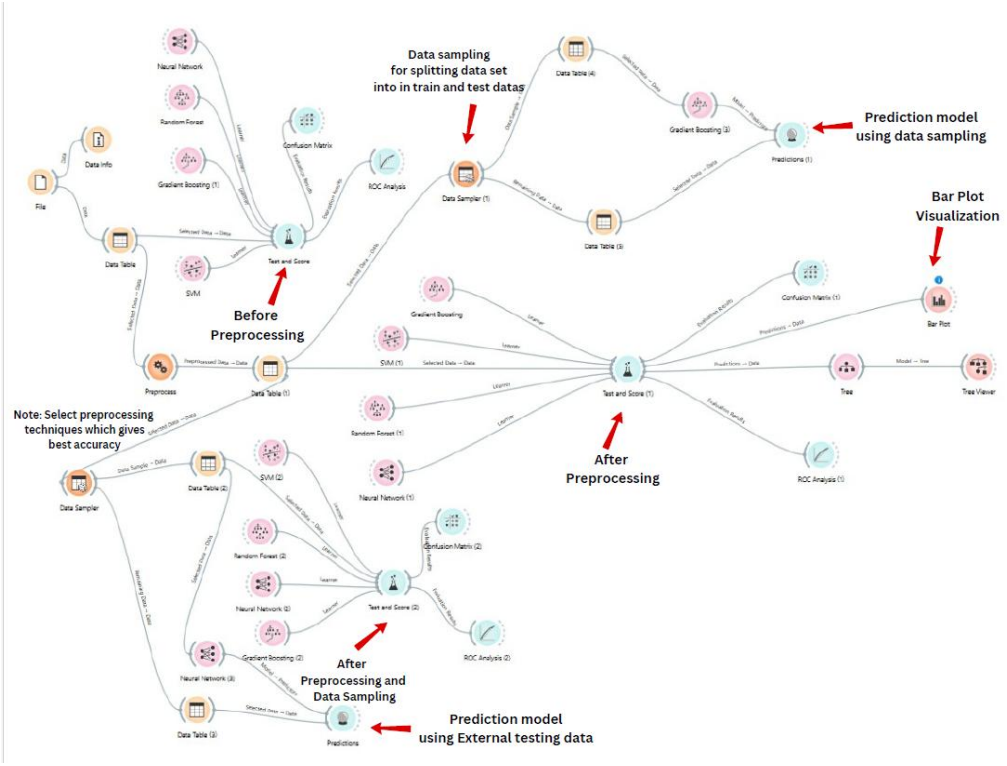


fig:6.2 Prediction model by separate train and test data

Entire Workflow:


Data sampling
for splitting data set
into in train and test datas
Prediction model
using data sampling
Bar Plot
Visualization
Before
Preprocessing
After
Preprocessing
Note: Select preprocessing
techniques which gives
best accuracy
After
Preprocessing and
Data Sampling
Prediction model
using External testing data

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

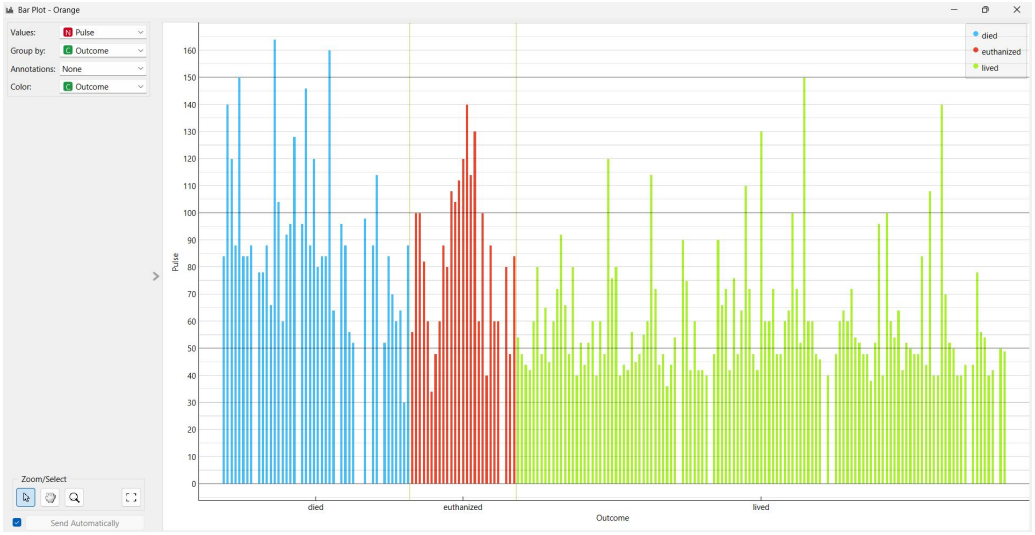


Figure 11: Bar plot for outcome (lived, Euthanized, died)

Figure 11 describes the x-axis of the bar plot is divided into three distinct outcome categories died, euthanized and lived each representing a different end result for the subjects in the dataset. The y-axis quantifies the frequency or number of cases corresponding to each outcome, allowing for a visual comparison of how commonly each result occurs. The height of each bar indicates the count of occurrences within that outcome category, where taller bars imply higher frequencies. This clear delineation along the two axes enables easy interpretation of the distribution and relative prominence of each outcome observed in the data.

CHAPTER 4: EXPERIMENTAL ANALYSIS

- Based on the Classifier accuracy that is shown in the Test & Score widget we choose to evaluate KNN and Logistic Regression algorithms using various metrics like confusion matrix

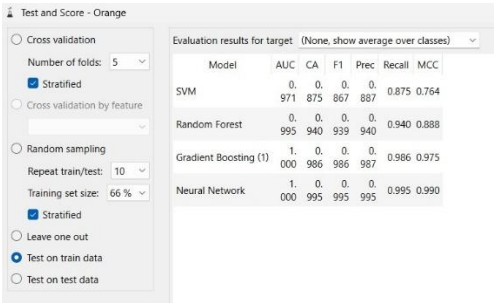


fig 4.1: CA before preprocessing

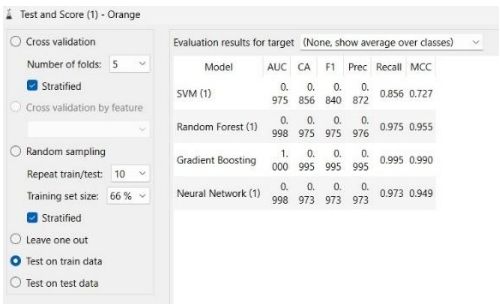


fig 4.2: CA after preprocessing

Figure 4.1 shows Test & Score before preprocessing (5-fold Cross Validation, 66% Training Size)

- SVM: 0.875
- Neural Network: 0.995
- Gradient Boosting: 0.986
- Random Forest: 0.940

Figure 4.2 shows Test & Score after preprocessing (5-fold Cross Validation, 66% Training Size)

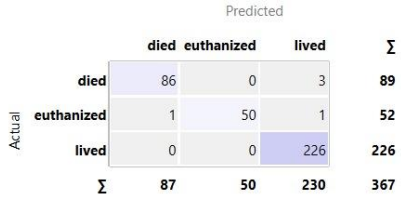
- SVM: 0.856 (Decreased)
- Neural Network: 0.973 (Decreased)
- Gradient Boosting: 0.995 (Increased)
- Random Forest: 0.975 (Increased)

Metric	Without Preprocessing (CA)	With Preprocessing (CA)	Improvement
SVM	0.875	0.856	1.9%
Random Forest	0.940	0.975	↑ 3.5%
Neural Network	0.995	0.973	↑ 2.2%
Gradient boosting	0.986	0.995	↑ 0.9%

Table: Comparison of Model Performance: Without Preprocessing vs With Preprocessing

Analysis on Confusion matrices:

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



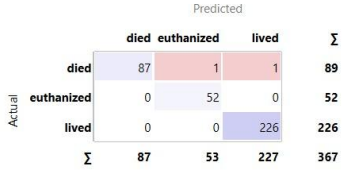


Fig 4.3: Confusion matrix before preprocessing fig 4.4: Confusion matrix after preprocessing

The observations that can be made by Fig 4.3 & 4.4 are as follows

Overall Accuracy:

- **Correct classifications** = (87 + 52 + 226) = **365** out of **367**
- **Accuracy** = 365367×100≈\frac{365}{367} \times 100 \approx367365 ×100≈ **99.45%**

Class-wise Performance:

• **Class:** Died

- **Correctly classified**: 87
- **Misclassified**:
  - 1 as Euthanized
  - 1 as Lived

• **Class:** Euthanized

- **Correctly classified**: 52
- **Misclassified**: None

• **Class:** Lived

- **Correctly classified**: 226
- **Misclassified**: None

Misclassification Trends:

- Gradient Boosting has **misclassifications** in minimal only in died class (1 instance)

- In conclusion, Gradient Boosting demonstrates exceptional performance with a remarkably high accuracy of 99.45% (365 out of 367 correct predictions). It performs perfectly for Class 2 (Euthanized) and Class 3 (Lived), with no misclassifications in either class. Only two instances from Class 1 (Died) were misclassified — one as Euthanized and one as Lived.
- If the primary objective is overall classification accuracy and reliability across all classes, Gradient Boosting is the most effective model. Its ability to correctly classify nearly all instances makes it highly suitable for critical applications where misclassification costs are significant.

Prediction model:

- Now based on the classifier accuracy we developed a prediction model by splitting 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.

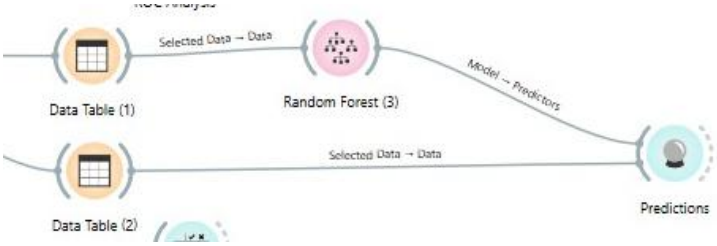


fig:3.1.1 Prediction Workflow

- The predictions that are given by random forest prediction model is
  The prediction model also gives error rate and accuracy.

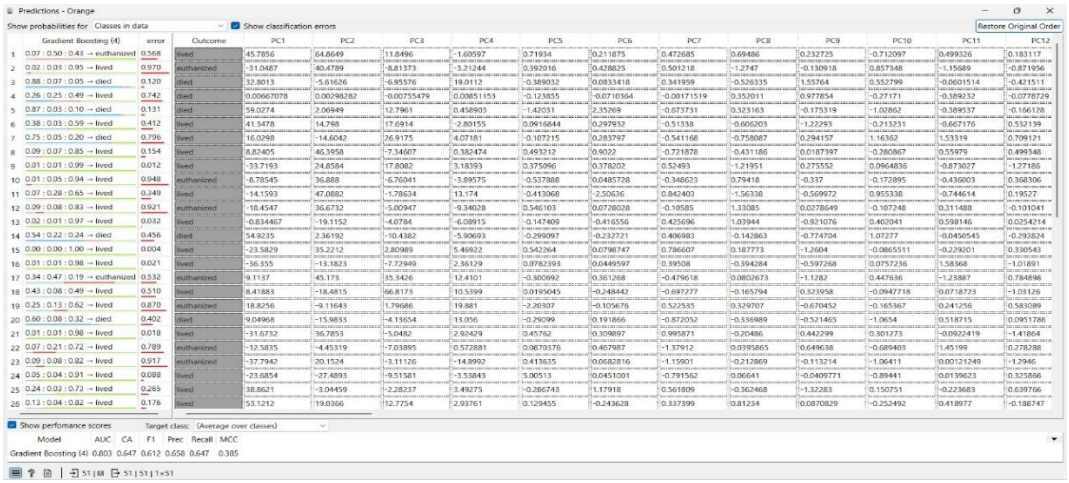


Fig: Prediction Classifications

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 improved model performance by handling missing values and normalizing numeric features.
   - Gradient Boosting showed the best classification accuracy (CA) after preprocessing, while Neural Network performed best before preprocessing.
2. **Overall Classification Accuracy**
   - Gradient Boosting achieved the best accuracy (82/108) compared to other techniques
3. **Confusion Matrix Insights**

   - Gradient Boosting achieved nearly perfect classification, with only two misclassifications out of 367 instances.

   - Class 2 (Euthanized) and Class 3 (Lived) were classified with 100% accuracy, indicating the model's strong ability to distinguish between non-fatal and moderate outcomes.

   - Class 1 (Died) showed minor misclassifications, with 1 instance predicted as Euthanized and 1 as Lived.

   - This suggests that Gradient Boosting slightly struggles with edge cases in Class 1, but maintains outstanding overall precision across all classes.

Final Decision

- If overall classification accuracy is the top priority, Gradient Boosting is the best choice, achieving 99.45% accuracy with only 2 misclassifications out of 367 instances.
- If the goal is to minimize misclassification across all classes, especially for critical outcomes like Class 2 (Euthanized) and Class 3 (Lived), Gradient Boosting still stands out, with zero errors in those classes.
- While minor confusion was observed in Class 1 (Died), the impact is minimal and far lower than other models like KNN or Logistic Regression.
- Therefore, the final model selection strongly favours Gradient Boosting, especially in scenarios where both high accuracy and balanced class performance are essential.

Future Scope for the Horse Colic Dataset with Multi-Target Classification

The Horse Colic dataset, containing clinical and pathological information on horses with colic, provides numerous opportunities for advanced analysis and real-world veterinary applications. Given its relevance to life-critical decision-making, the dataset can be leveraged in the following ways:

1. Improving Predictive Accuracy with Advanced ML Models

- Implement ensemble models like Random Forest, XGBoost, and Gradient Boosting to increase accuracy and handle feature complexity.
- Explore multi-target learning architectures (e.g., Multi-Task Neural Networks) to predict related outcomes such as survival, surgical intervention, and recovery time concurrently.

2. Feature Engineering & Dimensionality Reduction

- Apply Principal Component Analysis (PCA) or Autoencoders to reduce noise and retain only the most informative features.
- Use feature selection techniques to identify clinical signs (e.g., pulse rate, abdominal pain, temperature) that most influence survival outcomes.

3. Temporal and Pattern Analysis

- Incorporate time-series modelling to evaluate how patient vitals evolve over hours or days, aiding in dynamic prognosis.
- Identify early warning patterns from initial symptom records that may indicate severe or fatal cases.

4. Veterinary Decision Support Systems

- Build a clinical decision support tool that assists veterinarians in real-time diagnosis and treatment planning.
- Use predictive insights to recommend immediate actions (e.g., surgery, medication) based on patient condition.

5. Multi-Objective Optimization for Treatment Planning

- Optimize treatment strategies using evolutionary algorithms that balance survival rate, cost, and treatment duration.
- Apply Bayesian optimization for fine-tuning predictive models and identifying the best intervention paths.

6. Integration with Veterinary IoT and Real-Time Monitoring

- Connect with IoT-based monitoring systems (e.g., wearable devices tracking heart rate or temperature) for continuous updates.
- Develop an AI model that adapts based on live data, enabling early detection of deterioration.

7. Deployment as an Interactive Clinical Tool

- Design a dashboard (e.g., with Power BI or Streamlit) for veterinarians to visualize patient risk scores, vital signs, and model predictions.
- Create a machine learning API that integrates with veterinary hospital management systems to provide instant survival assessments.

By expanding into these future directions, the Horse Colic dataset can become instrumental in building smart, data-driven veterinary tools. These advancements could revolutionize how colic cases are diagnosed and treated, leading to better survival rates, reduced intervention costs, and more informed decision-making in animal healthcare.

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, Logistic Regression, Gradient Boosting, Neural Network, and Random Forest, etc…) were tested, and their performances were evaluated using classification accuracy (CA), and confusion matrices.

2.3. Key Findings from Model Comparisons

- KNN consistently performed the best across both datasets, benefiting from its ability to classify instances effectively when properly tuned.
- Logistic Regression showed improved performance after preprocessing, suggesting that real-world data benefits significantly from preprocessing techniques.
- Random Forest and Gradient Boosting had a lower initial accuracy but improved post-preprocessing, highlighting their dependence on high-quality input data.

3. Preprocessing Differences

Part A: Minimal Preprocessing

- Data set generated through google form is used to classify or analyze the streaming app platforms based on user demographics.
- Here we used classification models as:
  1.Random Forest
  2.Naive Bayes
  3.KNN
  4.Logistic Regression
- We used Average /frequent values to remove missing values which helps in increasing classification Accuracy.
- Finally Random Forest has an highest accuracy of 90.4%.

Part B: Extensive Preprocessing Required

- A online data set is provided which is used to classify or analyze the horses with colic which include details like age, temperature, pulse, pain level, and surgery status lived, died, or euthanized.
- Here we used classification models as:
  1.Neural Network
  2.Gradient Boosting

3.Random Forest
4.SVM

- We used Average /frequent values to remove missing values which helps in increasing classification Accuracy.
- Finally Gradient Boosting has an highest accuracy of 99.5%

4. Conclusion

- In Part A, we classified user preferences for streaming apps based on demographics. Among the models used, Random Forest gave the best result with an accuracy of 90.4%.
- In Part B, we analyzed medical data of horses with colic to predict their outcomes.
- Here, Gradient Boosting outperformed all other models with an impressive accuracy of 99.5%.
- Data preprocessing (handling missing values with average/frequent values) played a key role in improving model performance in both cases.
- Overall, the comparison highlights how different machine learning models perform across diverse datasets, and shows that model selection and data quality significantly impact classification accuracy.

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. **Equine Health Analysis & Diagnosis**
   - Cohen, N. D. (2005). *Equine internal medicine approaches to colic*. Saunders Elsevier.
   - Proudman, C. J. (1999). "A two-year, prospective survey of equine colic in general practice." *Equine Veterinary Journal*, 31(6), 456-460.
   - Archer, D. C., Pinchbeck, G. L., French, N. P., Proudman, C. J. (2008). "Long-term survival and health status after surgery for large colon volvulus in horses." *Equine Veterinary Journal*, 40(5), 396-401.

3. **Geospatial & Health Monitoring for Equines (EHM)**
   - Tinker, M. K., White, N. A., Lessard, P., Thatcher, C. D., Pelzer, K. D., Davis, B., & Carmel, D. K. (1997). "Prospective study of equine colic incidence and risk factors." *Journal of the American Veterinary Medical Association*, 210(10), 1425-1431.
   - Reeves, M. J., & Salman, M. D. (1995). "Multivariable analysis of risk factors for acute colic in horses." *Preventive Veterinary Medicine*, 24(3), 285-302.
   - Traub-Dargatz, J. L., Kopral, C. A., & Seitzinger, A. H. (2001). "Estimate of the national incidence of and operation for colic in U.S. horses." *Journal of the American Veterinary Medical Association*, 219(1), 67-71.

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 the information to provide valid conclusions to meet the desired needs.
5. **Modern tool usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
6. **The engineer and society**: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
7. **Environment and sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
8. **Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
9. **Individual and team work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

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

Program Specific Outcomes (PSOs)

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

PROJECT PROFORMA

Classification of Project	Application	Product	Research	Review
	√			

Note: Tick Appropriate category

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

Mapping Table

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

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

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