DATA WAREHOUSING
&
DATA MINING
(01CE0723)

Lab Manual

A.Y. 2025-26

Name : Samarth Chavda
Er. No. : 92200103165
Semester: 7
Class : TC2
Batch : B

INDEX

Sr. No.	Experiments	Plan Date	Actual Date	Marks	Signature
1.	Explore data mining and data warehousing tools.				
2.	Explore Weka modules: Explorer, Experimenter, KnowledgeFlow, Workbench, Simple CLI. Exploring Explorer module with .csv and .arff files.				
3.	Prepare and analyse "student" dataset, also analyse "student", "weather.nominal" and "iris" dataset along with editing and visualization.				
4.	Apply Preprocessing techniques on dataset using filters: Remove, ReplaceMissingValues, ReplaceMissingWithUserConstant, ReplaceWithMissingValue, Descritize. Also do the result analysis before and after preprocessing.				
5.	Apply Preprocessing techniques on dataset using filters: NumericToNominal, StringToNominal, NominalToBinary, NumericToNominal. Also do the result analysis before and after preprocessing.				
6.	Demonstration on APRIORI algorithm along with frequent item sets, non-frequent item sets and stron & weak association rules.				
7.	Apply APRIORI algorithm on "weather.nominal" dataset and analyze the results.				
8.	Demonstration on "J48", "RandomForest" and "NaiveBayes" classification algorithms using test options.				
9.	Apply and analyze "J48", "RandomForest" and "NaiveBayes" classification algorithms on "weather.nominal" dataset and compare the results.				
10.	Demonstration on prediction algorithms "NaiveBayes" and "Logistic" by creating classification model and "Supplied Test Set" options.				

11.	Apply prediction "NaiveBayes" and "Logistic" by creating classification model and "Supplied Test Set" options on any suitable dataset and compare the results.				
12.	Demonstration on "SimpleKMeans" clustering algorithm using "EuclideanDistance".				
13.	Apply and analyze "SimpleKMeans" clustering algorithm on suitable dataset, with the observation of "maxIterations" and "numClusters" parameters along with visualization.				
14.	Case study on applications of Data Mining tools and techniques used for Business Intelligence.				

Experiment List

Sr. No.	Title	CO
1.	Explore data mining and data warehousing tools.	CO1, CO2
2.	Explore Weka modules: Explorer, Experimenter, KnowledgeFlow, Workbench, Simple CLI. Exploring Explorer module with .csv and .arff files.	CO2, CO3, CO4, CO5
3.	Prepare and analyse “student” dataset, also analyse “student”, “weather.nominal” and “iris” dataset along with editing and visualization.	CO2
4.	Apply Preprocessing techniques on dataset using filters: Remove, ReplaceMissingValues, ReplaceMissingWithUserConstant, ReplaceWithMissingValue, Descritize. Also do the result analysis before and after preprocessing.	CO3
5.	Apply Preprocessing techniques on dataset using filters: NumericToNominal, StringToNominal, NominalToBinary, NumericToNominal. Also do the result analysis before and after preprocessing.	CO3
6.	Demonstration on APRIORI algorithm along with frequent item sets, non-frequent item sets and stron & weak association rules.	CO4
7.	Apply APRIORI algorithm on “weather.nominal” dataset and analyze the results.	CO4
8.	Demonstration on “J48”, “RandomForest” and “NaiveBayes” classification algorithms using test options.	CO5
9.	Apply and analyze “J48”, “RandomForest” and “NaiveBayes” classification algorithms on “weather.nominal” dataset and compare the results.	CO5
10.	Demonstration on prediction algorithms “NaiveBayes” and “Logistic” by creating classification model and “Supplied Test Set” options.	CO5
11.	Apply prediction “NaiveBayes” and “Logistic” by creating classification model and “Supplied Test Set” options on any suitable dataset and compare the results.	CO5
12.	Demonstration on “SimpleKMeans” clustering algorithm using “EuclideanDistance”.	CO5
13.	Apply and analyze “SimpleKMeans” clustering algorithm on suitable dataset, with the observation of “maxIterations” and “numClusters” parameters along with visualization.	CO5
14.	Case study on applications of Data Mining tools and techniques used for Business Intelligence.	CO2, CO3, CO4, CO5

Experiment 1

Title: Explore data mining and data warehousing tools.

List of Tools Explored for Data Mining & Data Warehousing:
1. Microsoft SQL Server
2. SAP HANA
3. Oracle Autonomous Data Warehouse + Oracle Data Mining (ODM)

Tool 1: Microsoft SQL Server

- Introduction

  Microsoft SQL Server is a relational database management system (RDBMS) developed by Microsoft. It supports both data warehousing and data mining functionalities, especially when used with its integrated services like SQL server database engine, SSIS, SSAS and SSRS.
- Features

Description: - This image shows the Microsoft SQL Server Management Studio (SSMS) interface. The "Object Explorer" panel on the left is expanded and connected to a SQL Server instance named in the image. The explorer lists components such as Databases, Security, Server Objects, Replication, PolyBase, Management, and XEvent Profiler. The central workspace is currently empty, ready for query input or other operations.

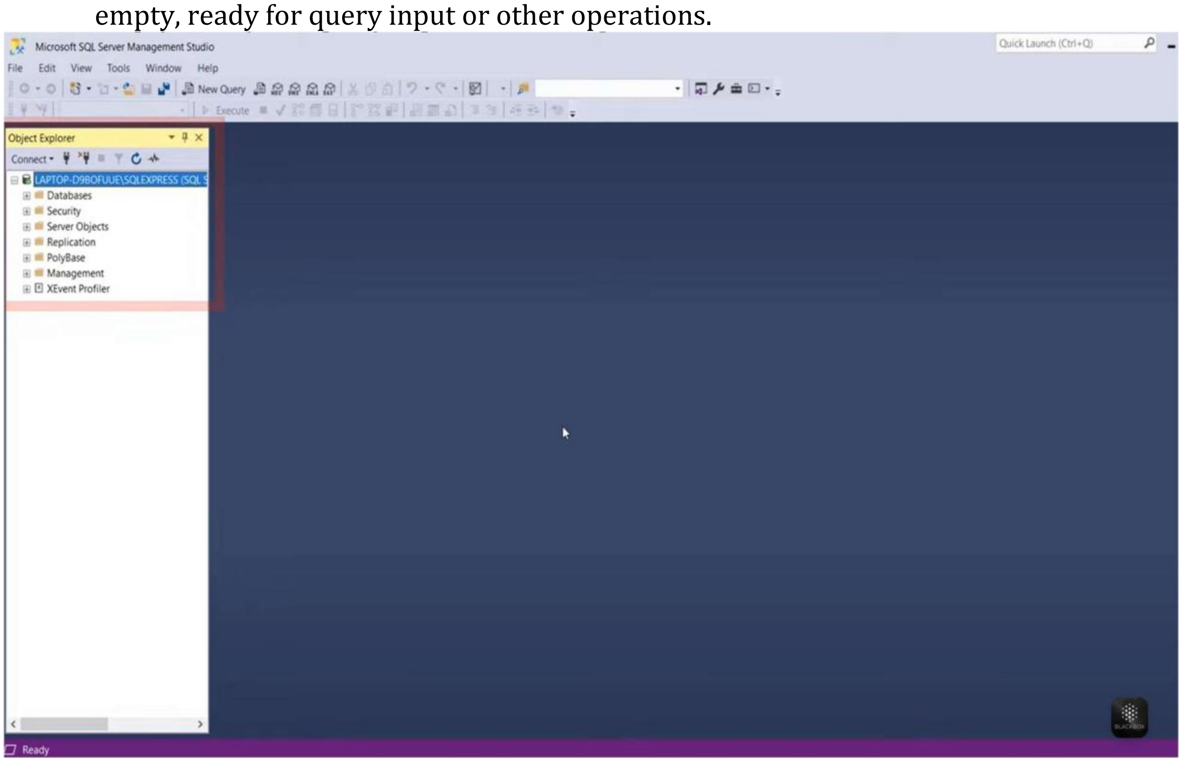

Figure 1.1

Description: - This image shows a simple SQL command has been run to create a table called demo_ssms with just one column named id, which stores integers. Also, the object explorer displays the connect server and its system databases.

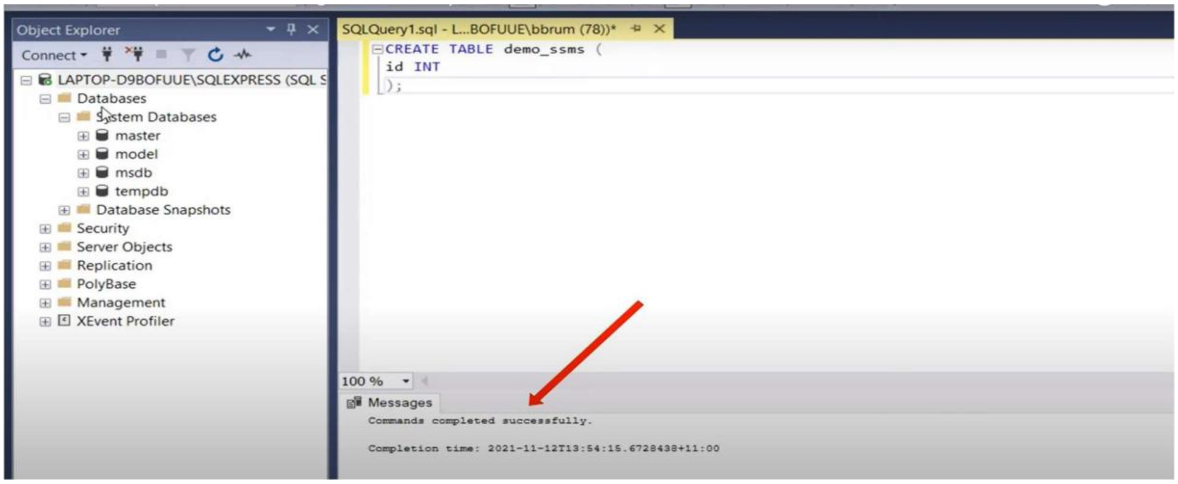

Figure 1.2

- Official website of tool

https://www.microsoft.com/en-in/sql-server

Tool 2: SAP HANA

- Introduction

SAP HANA (High-Performance Analytic Appliance) is an in-memory, column-oriented relational database developed by SAP. It is designed for real-time data processing and analytics, combining transactional and analytical workloads on a single platform.

- Features

Description: - This image shows the SAP HANA Studio interface. It is divided into three main sections: Systems view, Editor area, and Other views like Error Log and Properties . Each section provides different tools for managing, monitoring, and configuring SAP HANA systems.

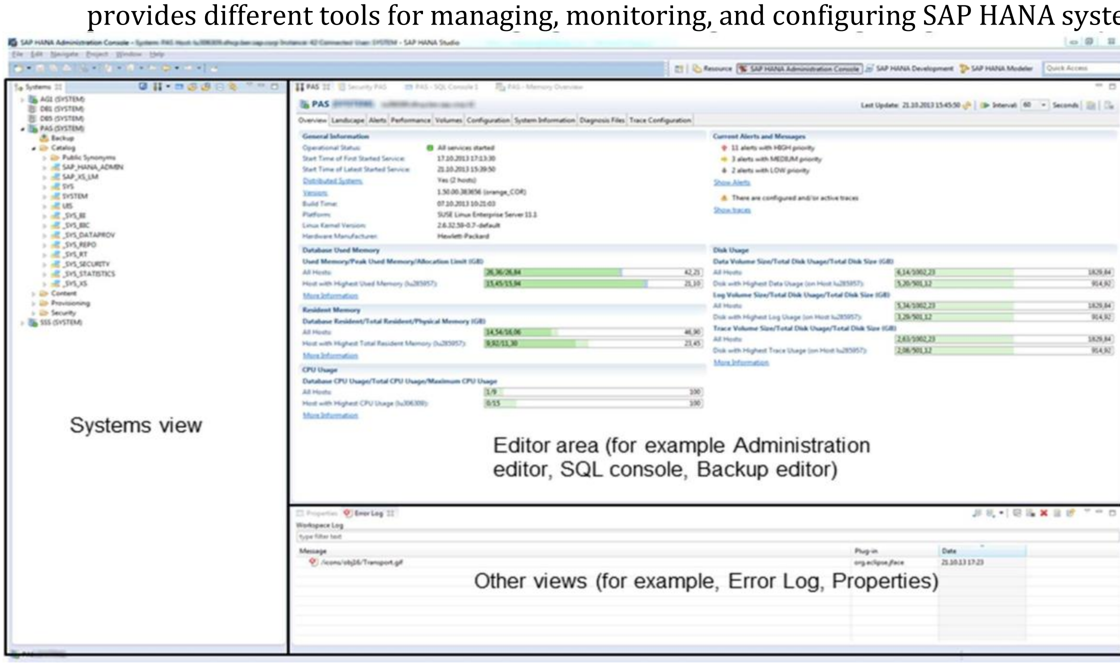

Figure 1.4

- Official website of tool

https://www.sap.com/products/data-cloud/hana.html

Tool 3: Oracle Autonomous Data Warehouse + Oracle Data Mining (ODM)

- Introduction

Oracle Autonomous Data Warehouse (ADW) is a fully managed, cloud-based data warehouse service optimized for analytic workloads, offering high performance, scalability, and automation. Oracle Data Mining (ODM), part of Oracle Advanced Analytics, enables in-database machine learning, allowing users to build predictive models directly within the Oracle Database. Together, they empower businesses to uncover insights and make data-driven decisions efficiently and securely.

- Features

Description: - This screenshot shows the Oracle Data Miner GUI within Oracle SQL Developer. It features multiple data visualizations such as scatter plots, histograms, and box plots used for exploratory data analysis. The interface supports interactive graphing to help users identify trends and patterns before building predictive models.

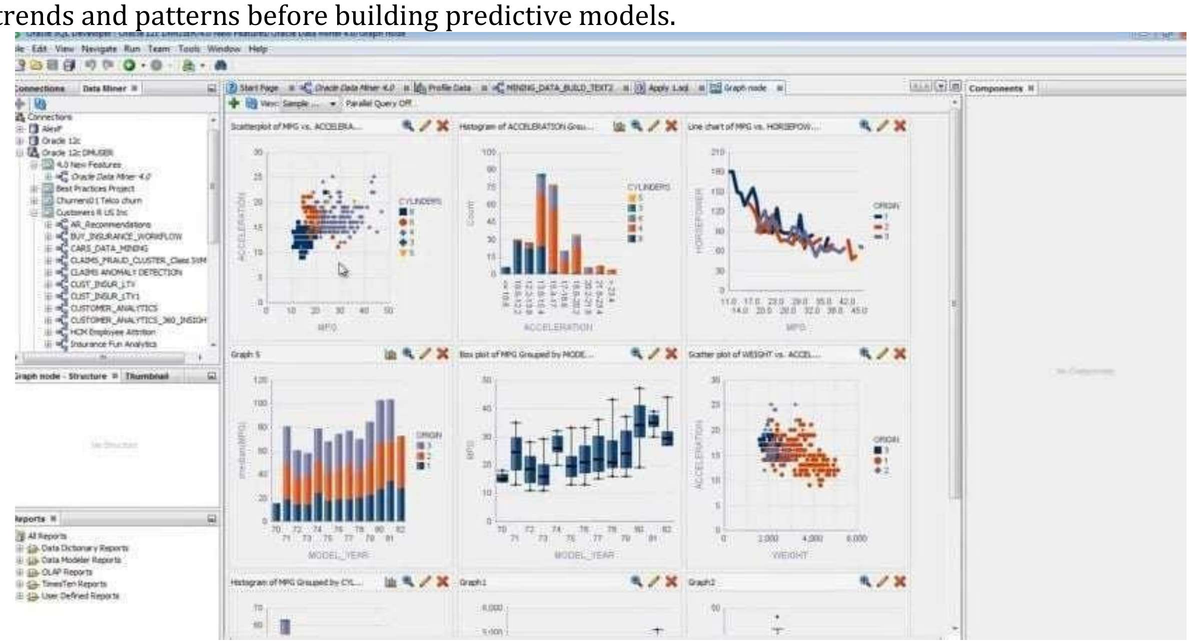

Figure 1.5

- Official website of tool

https://www.oracle.com/in/autonomous-database/

Comparison of all the tools:

Feature	Oracle Autonomous Data Warehouse + ODM	SAP HANA	Microsoft SQL Server
Vendor	Oracle Corporation	SAP SE	Microsoft
Type	Cloud-based Autonomous Data Warehouse + Advanced Analytics	In-memory database & platform	Relational DBMS
Deployment	Cloud (Autonomous)	On-premise, Cloud, Hybrid	On-premise, Cloud (Azure), Hybrid
Primary Focus	Self-managing data warehouse + integrated data mining	Real-time analytics, in-memory computing	Traditional OLTP/OLAP with BI tools

Experiment Outcome:

The experiment demonstrated that data mining and warehousing tools like Oracle Data Miner (ODM), SAP HANA, and SQL Server offer robust capabilities for data analysis and storage. ODM excels in predictive modeling, SAP HANA provides real-time analytics with in-memory processing, and SQL Server offers integrated services for efficient data warehousing and mining.

Experiment 2

Title: Explore Weka modules: Explorer, Experimenter, KnowledgeFlow, Workbench, Simple CLI.

WEKA History & Introduction:

WEKA (Waikato Environment for Knowledge Analysis) is a popular open-source machine learning software written in Java, developed at the University of Waikato in New Zealand. It provides a comprehensive suite of tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is widely used in research, education, and industrial applications for data mining and machine learning tasks.

History: -

• In 1993: Development of WEKA began at the University of Waikato, New Zealand, as part of a government-funded research project.
• 1997: The original version of WEKA was a Tcl/Tk-based prototype.
• 1999: WEKA was rewritten entirely in Java, making it platform-independent and more user-friendly.
• 2001: The software became popular after the publication of the book "Data Mining: Practical Machine Learning Tools and Techniques" by Ian H. Witten and Eibe Frank, which used WEKA extensively.
• 2005–present: WEKA has continuously evolved, with contributions from the open-source community and updates from the University of Waikato team

WEKA Applications:

Applications of WEKA are: -

i. Educational and Research Purposes
ii. Healthcare and Medical Diagnosis
iii. Business Intelligence and Marketing
iv. Text Mining and Natural Language Processing (NLP)
v. Fraud Detection and Cybersecurity

Modules in WEKA:

1. Explorer
2. Experimenter
3. KnowledgeFlow
4. Workbench
5. Simple CLI

Module 1: Explorer

- Purpose of the Module

The Explorer is one of the main user interfaces in WEKA and serves as a central platform for conducting interactive data analysis and machine learning experiments. It allows users to easily explore, preprocess, model, and evaluate datasets without needing to write any code.

- Screenshots with description

Description: - This image shows the preprocess tab which is used for loading and preparing data before applying ML algorithm. It’s designed to guide you step-by-step through data preprocessing, classification, clustering, and evaluation, making it perfect for both beginners and researchers exploring data.

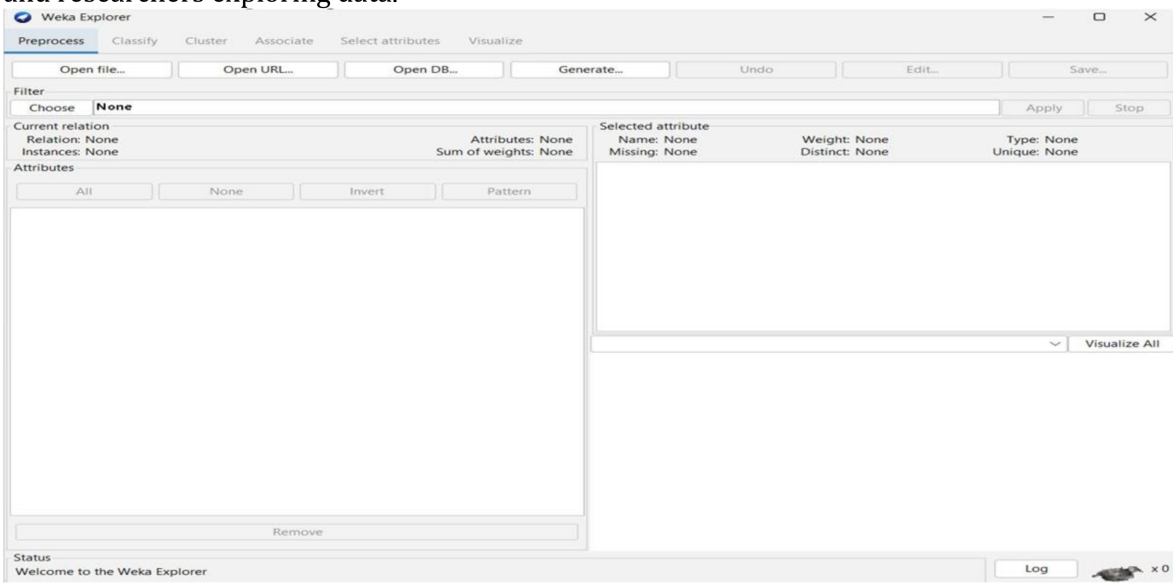

Figure 2.1

- Applications of the module

The WEKA Explorer module is excellent for rapidly evaluating machine learning concepts on actual data. It can be used to clean up your dataset, test various algorithms, such as clustering or decision trees, and quickly assess how well they work.

Module 2: Experimenter

- Purpose of the Module

The WEKA Experimenter makes it easy to test and compare different machine learning models without doing everything by hand. It uses consistent testing methods like cross-validation and even shows you which models perform better with clear, statistical results.

- Screenshots with description

Description: - This image shows the WEKA experimenter environment specially setup tab. It’s where you configure your machine learning experiments by choosing datasets, algorithms, and how many times to repeat the tests. The interface lets you easily set up cross-validation, compare models, and organize everything before running the experiments.

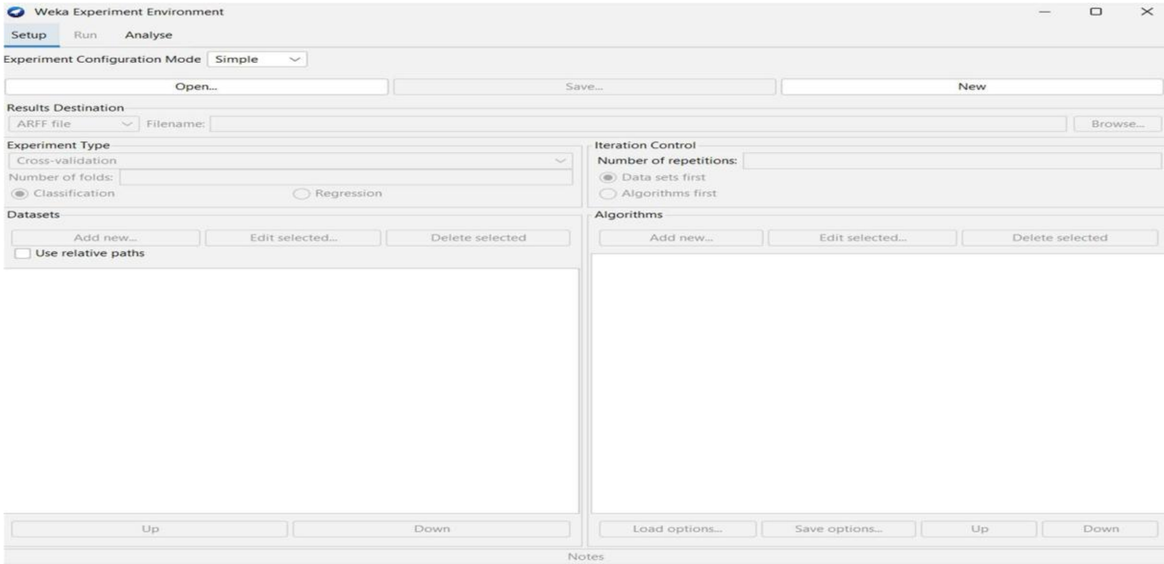

Figure 2.2

- Applications of the module

Its main application is to compare the performance of different algorithms (or their settings) across datasets using consistent evaluation methods like cross-validation. It's especially useful for research, academic projects, or anyone needing to test and validate models in a systematic and repeatable way.

Module 3: KnowledgeFlow

- Purpose of the Module

The KnowledgeFlow module in Weka provides a visual programming environment for designing and executing machine learning workflows. It allows users to connect components like data sources, filters, classifiers, and evaluators in a flowchart-style layout.

- Screenshots with description

Description : - This image shows the WEKA KnowledgeFlow Environment, a visual workspace where you can build machine learning workflows by dragging and connecting components. Instead of writing code, you can design the entire data mining process from loading data and applying filters to training models and visualizing results using a simple flowchart-style interface.

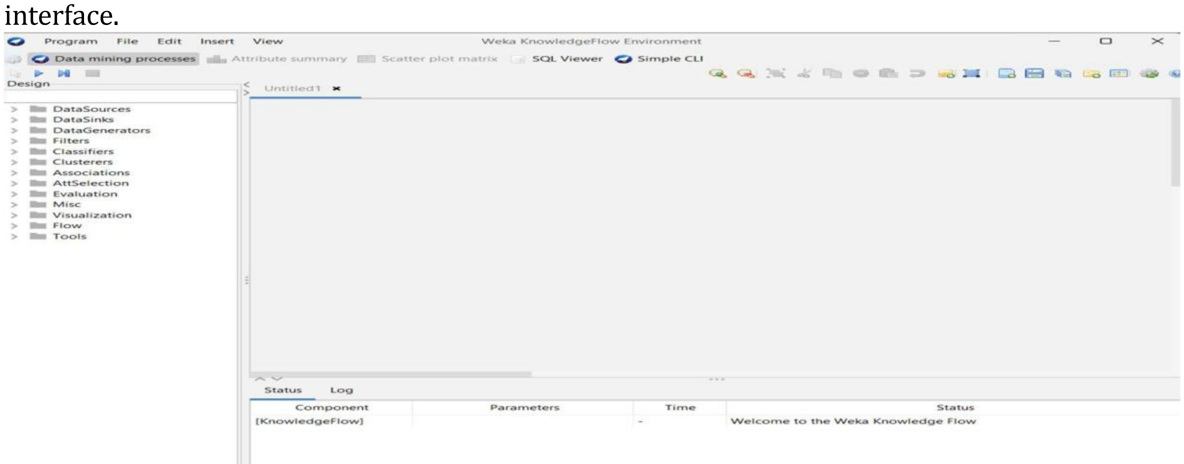

Figure 2.3

- Applications of the module

The WEKA KnowledgeFlow module is mainly used to visually design and automate machine learning workflows. It allows users to connect components like data sources, filters, classifiers, and evaluation tools in a flowchart format.

Module 4: Workbench

- Purpose of the Module

WEKA workbench is the tool used for data mining and ML. It helps to preprocess data, apply various algorithms for classification, regression, clustering and evaluate model performance.

- Screenshots with description

Description: - This image displays the Weka Workbench interface in its initial "Preprocess" state. It’s ready for users to load a dataset using options like “Open file...” or “Open URL...”. The interface is clean and structured, offering tools for filtering data, viewing attributes, and preparing for further analysis like classification or clustering. No data is loaded yet, as indicated by the empty fields.

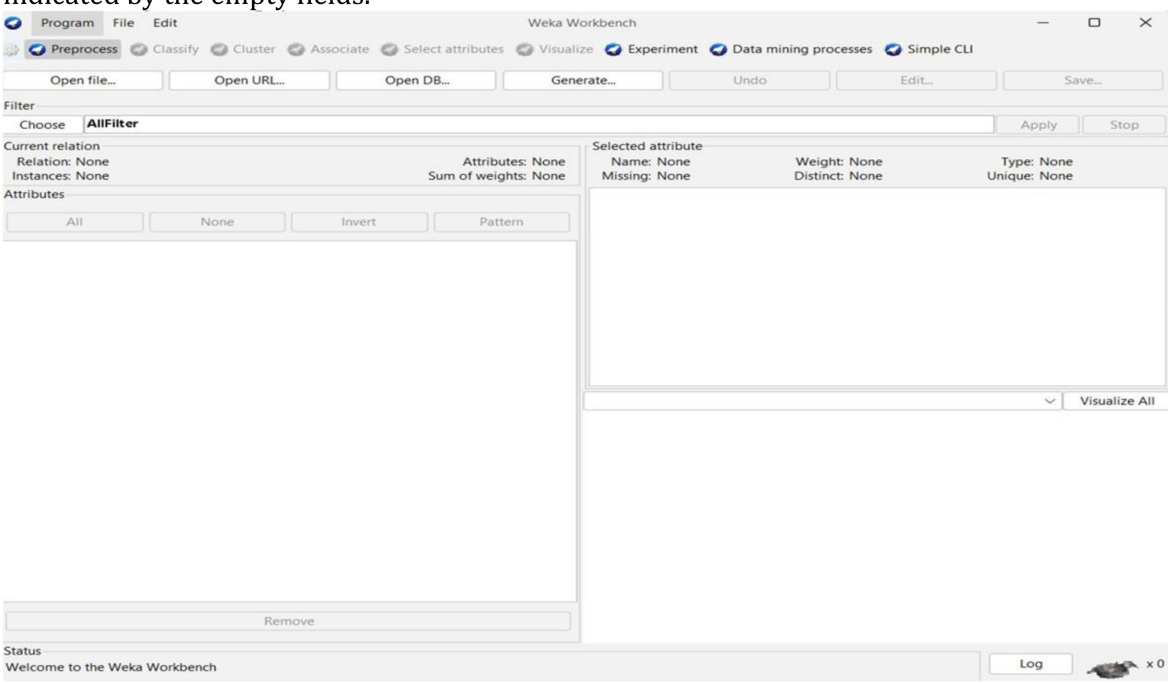

Figure 2.4

- Applications of the module

The Weka Workbench is widely used for data mining, machine learning, and predictive analytics. It helps users preprocess data, build models (e.g., classification, clustering), and evaluate their performance. Common applications include academic research, business intelligence, medical diagnosis, and fraud detection. Its intuitive GUI makes complex analysis accessible without programming.

Experiment 3

Title: Prepare and analyse "student" dataset, also analyse "student", "weather.nominal" and "iris" dataset along with editing and visualization.

- ## File formats and data types supported by WEKA

  - File formats supported by WEKA
    - arff
    - arff.gz
    - bsi
    - csv
    - dat
    - data
    - json
    - json.gz
    - libsvm
    - m
    - names
    - xrff
    - xrff.gz

  - Data types supported by WEKA
- Numeric (Integer and Real), String, Date, and Relational

- ## Preparation and analysis of "student.arff" dataset

  - Dataset Code

```
@relation StudentData

@attribute NAME string
@attribute CLASS {TC1,TC2,TC3,TC4,TC5,TC6}
@attribute CITY {JAMNAGAR,RAJKOT,AHEMDABAD,SURAT,VADODARA}
@attribute DOB string
@attribute ER_NO numeric
@attribute GENDER {MALE,FEMALE}
@attribute EMAILID string
@attribute SPI numeric
@attribute BACKLOG {YES,NO}
@attribute PLACEMENT_STATUS {YES,NO}

@data
'Ravi Patel',TC1,RAJKOT,12-06-2001,101,MALE,ravi.patel@example.com,8.2,NO,YES
'Neha Sharma',TC2,JAMNAGAR,05-11-
2000,102,FEMALE,neha.sharma@example.com,7.4,NO,NO
```

'Amit Joshi',TC3,AHEMDABAD,22-03-2001,103,MALE,amit.joshi@example.com,6.9,YES,NO
'Priya Mehta',TC4,VADODARA,17-08-2000,104,FEMALE,priya.mehta@example.com,9.1,NO,YES
'Karan Desai',TC5,SURAT,01-01-2002,105,MALE,karan.desai@example.com,8.7,NO,YES
'Anjali Thakkar',TC6,JAMNAGAR,09-05-2001,106,FEMALE,anjali.thakkar@example.com,7.8,NO,YES
'Manav Shah',TC1,RAJKOT,14-02-2000,107,MALE,manav.shah@example.com,6.5,YES,NO
'Divya Patel',TC2,AHEMDABAD,30-06-2001,108,FEMALE,divya.patel@example.com,7,YES,NO
'Jay Soni',TC3,VADODARA,23-09-2000,109,MALE,jay.soni@example.com,8.4,NO,YES
'Pooja Gohil',TC4,SURAT,07-12-2000,110,FEMALE,pooja.gohil@example.com,7.6,NO,YES
'Hitesh Bhatt',TC5,JAMNAGAR,18-04-2001,111,MALE,hitesh.bhatt@example.com,6.3,YES,NO
'Rina Shah',TC6,AHEMDABAD,11-10-2000,112,FEMALE,rina.shah@example.com,8.1,NO,YES
'Vishal Patel',TC1,SURAT,02-07-2001,113,MALE,vishal.patel@example.com,8,NO,YES
'Meena Doshi',TC2,VADODARA,15-05-2000,114,FEMALE,meena.doshi@example.com,7.2,NO,NO
'Rahul Trivedi',TC3,RAJKOT,19-11-2001,115,MALE,rahul.trivedi@example.com,6.7,YES,NO
'Kajal Chauhan',TC4,JAMNAGAR,25-01-2000,116,FEMALE,kajal.chauhan@example.com,8.3,NO,YES
'Nikhil Shah',TC5,AHEMDABAD,03-08-2001,117,MALE,nikhil.shah@example.com,7.9,NO,YES
'Seema Patel',TC6,SURAT,29-03-2000,118,FEMALE,seema.patel@example.com,7.5,NO,YES
'Bhavin Mehta',TC1,VADODARA,06-09-2001,119,MALE,bhavin.mehta@example.com,6.4,YES,NO
'Rupal Desai',TC2,JAMNAGAR,21-12-2000,120,FEMALE,rupal.desai@example.com,8.6,NO,YES
'Harsh Panchal',TC3,RAJKOT,10-10-2001,121,MALE,harsh.panchal@example.com,7.1,NO,NO
'Jinal Patel',TC4,AHEMDABAD,27-02-2001,122,FEMALE,jinal.patel@example.com,8.9,NO,YES
'Dhruv Shah',TC5,VADODARA,13-07-2000,123,MALE,dhruv.shah@example.com,7.7,NO,YES
'Komal Thakkar',TC6,SURAT,08-11-2001,124,FEMALE,komal.thakkar@example.com,6.6,YES,NO
'Rakesh Chauhan',TC1,JAMNAGAR,04-04-2001,125,MALE,rakesh.chauhan@example.com,8.8,NO,YES
'Vaishali Joshi',TC2,AHEMDABAD,16-06-2000,126,FEMALE,vaishali.joshi@example.com,7.3,NO,NO
'Yash Patel',TC3,RAJKOT,20-08-2001,127,MALE,yash.patel@example.com,6.8,YES,NO
'Nidhi Shah',TC4,VADODARA,12-09-2000,128,FEMALE,nidhi.shah@example.com,8.5,NO,YES
'Sahil Trivedi',TC5,SURAT,05-02-2001,129,MALE,sahil.trivedi@example.com,7,NO,NO
'Hetal Mehta',TC6,JAMNAGAR,31-03-2000,130,FEMALE,hetal.mehta@example.com,8.2,NO,YES
Harsh,TC4,AHEMDABAD,19-07-1988,110,MALE,harsh@gmail.com,7.8,NO,YES

- Analysis of "student.arff" with weka

- Gender with different class

Description: - The gender distribution is fairly balanced, and the visualization aids in understanding how this attribute relates to another, likely "CLASS".

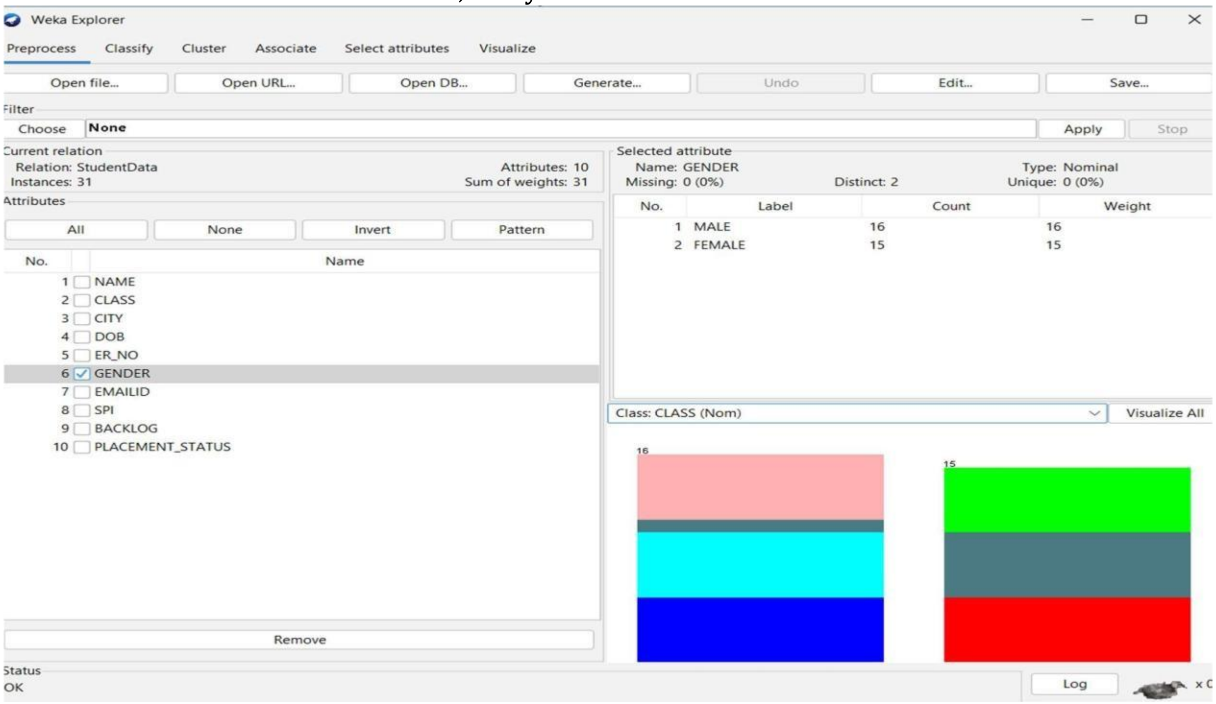

Fig 3.1

- Gender with Placement status

Description: - The bar chart indicates a visual comparison of placement outcomes across genders, useful for analyzing gender-based placement trends.

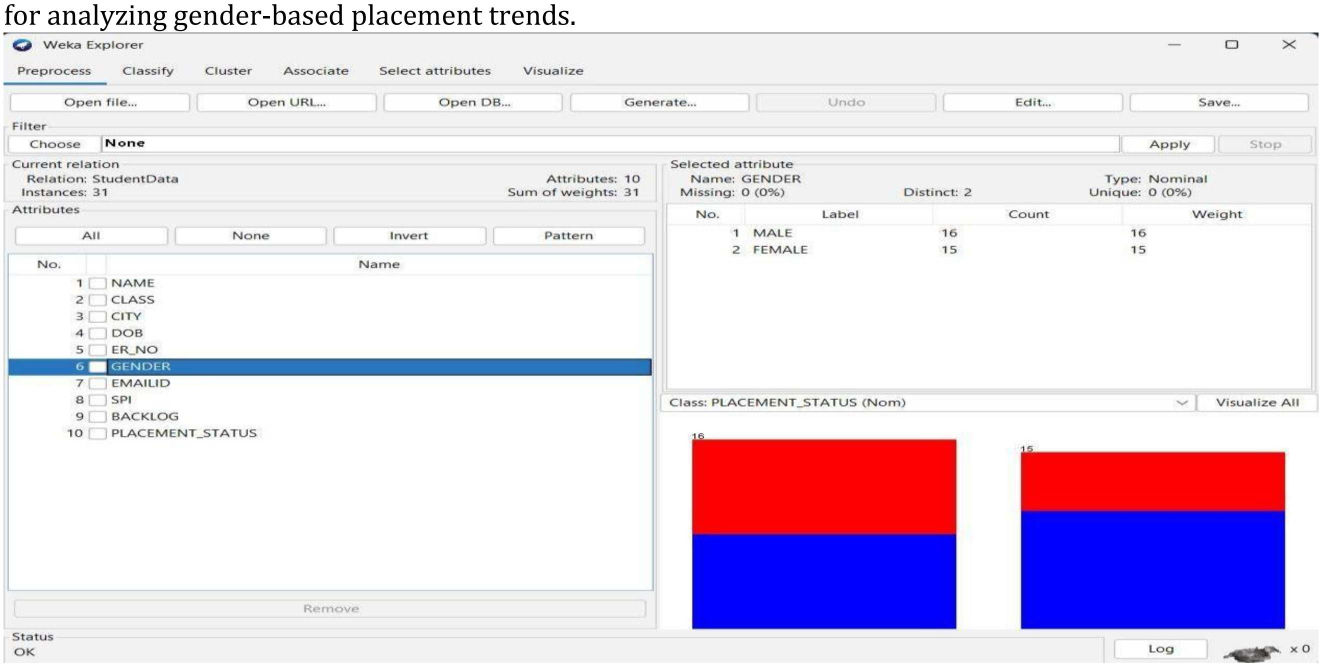

Fig 3.2

- Gender with different city

Description: - The visualization shows a diverse mix of cities for both male and female students. This can help in identifying regional patterns or preferences among students.

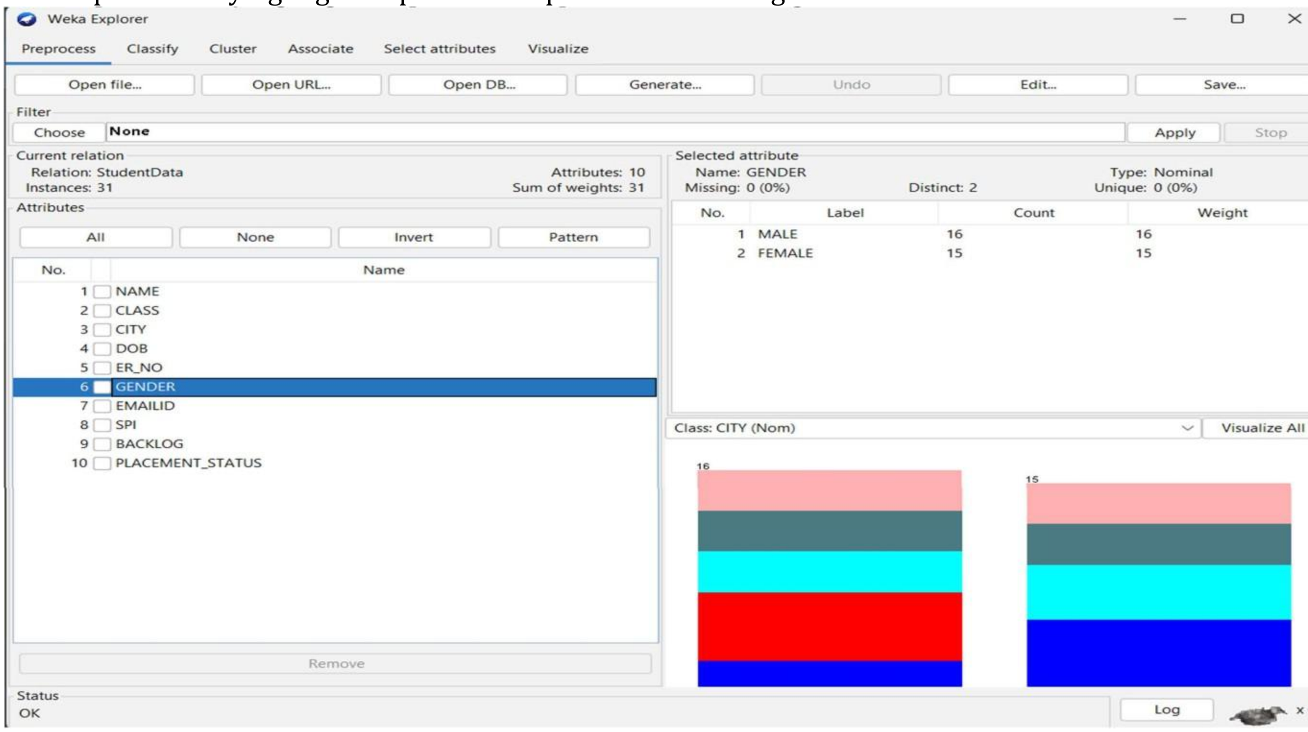

Fig 3.3

- Gender with student's backlog count

Description: - The chart suggests female students might have more backlogs than male students in this group.

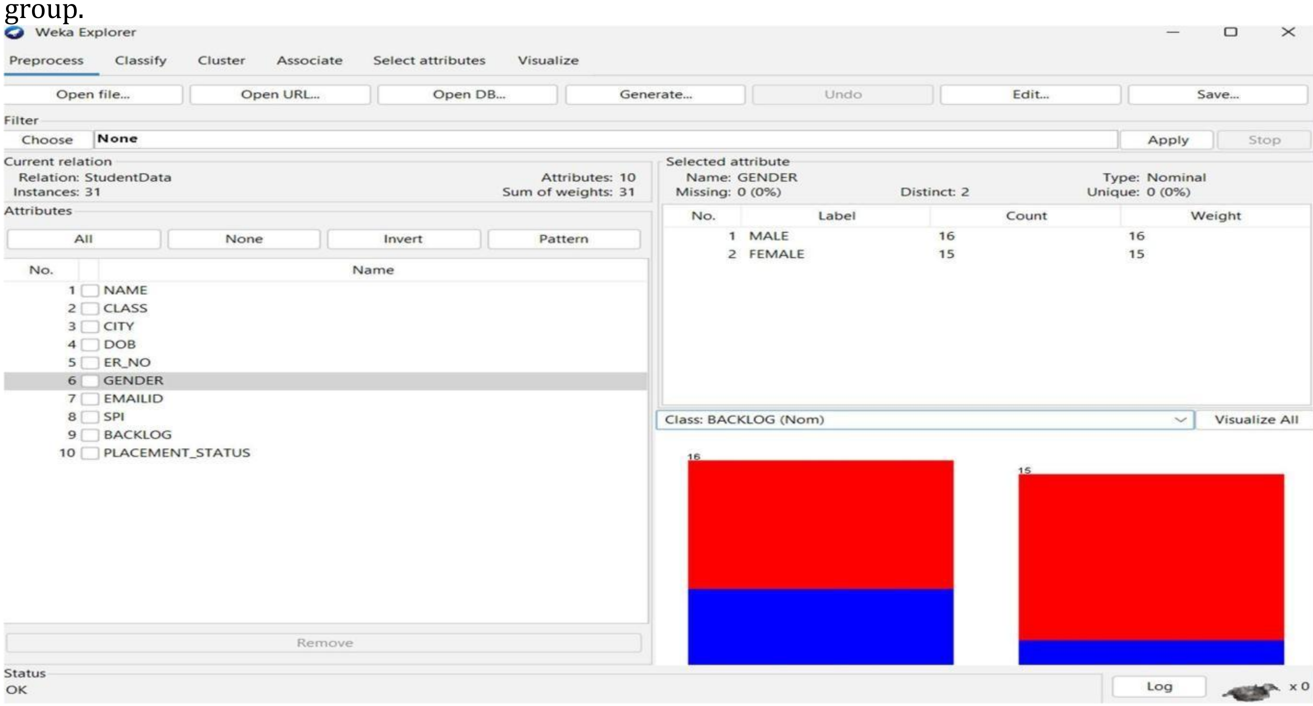

Fig 3.2

- Gender distribution
  Description: - The gender data is complete and almost evenly split.

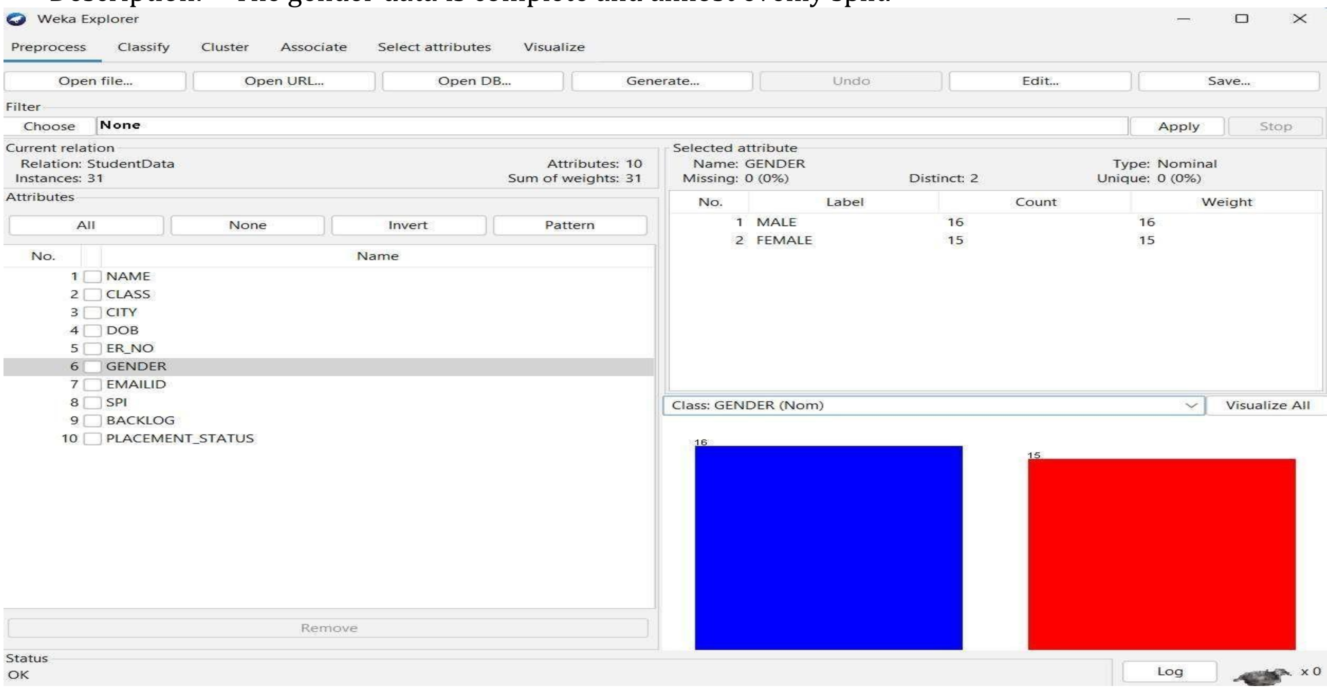

Fig 3.5

- Class Visualization of "class" attribute for all attribute

Description: - Each tile in the grid represents a single attribute from the dataset. Bar charts and histograms are used to visualize how the data is distributed.

Fig 3.6

- Class Visualization of "city" attribute for all attribute

Description: - This screenshot presents multiple bar charts and histograms for each attribute in the dataset. These charts help us understand how data values are distributed and how they relate to a selected class.

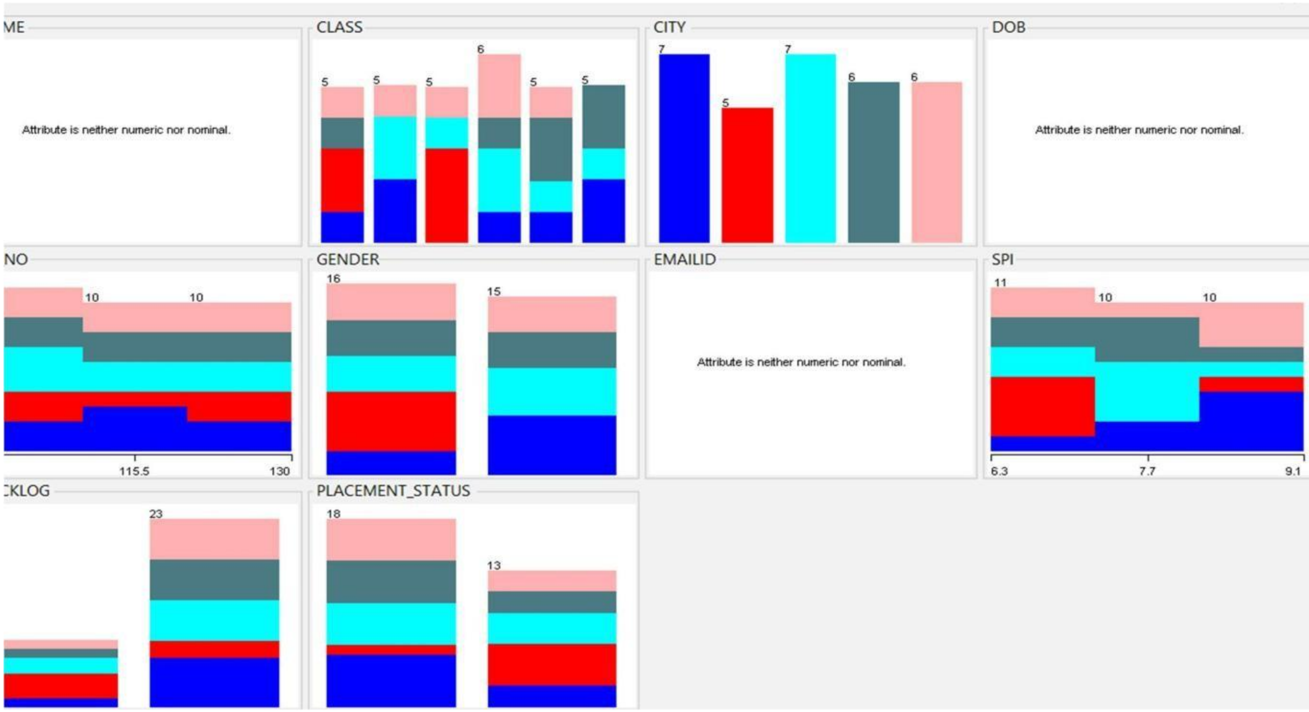

Fig 3.7

- **Analysis of "weather.nominal.arff" dataset**

  - Dataset Code

```
relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes
rainy,cool,normal,FALSE,yes
rainy,cool,normal,TRUE,no
overcast,cool,normal,TRUE,yes
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
rainy,mild,normal,FALSE,yes
sunny,mild,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,mild,high,TRUE,no
```

- Analysis of “weather.nominal.arff” with weka

- Analysis of “play” attribute with outlook

Description: - The bar chart at the bottom displays the distribution of the play attribute. The dropdown "Class: outlook (Nom)" indicates that this visualization is color-coded based on the outlook attribute.

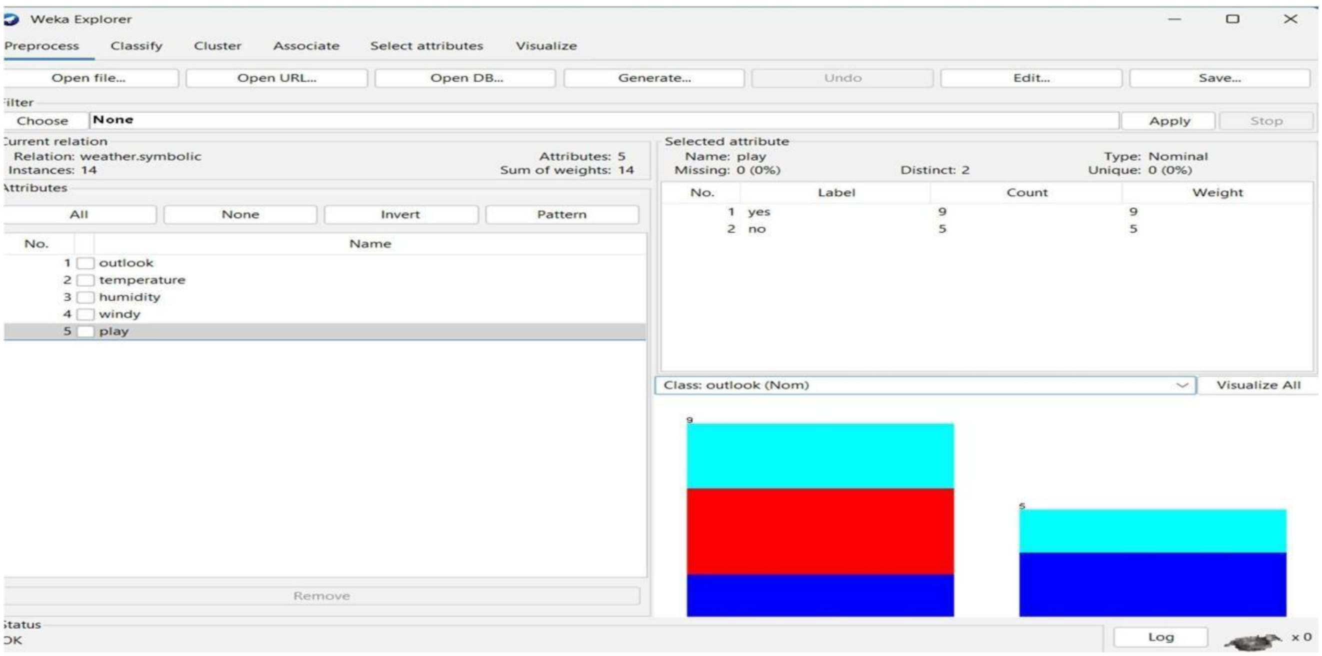

Fig 3.8

- Analysis of “play” attribute with temperature

Description: - Each bar is color-coded, likely showing how different values of temperature attribute are distributed within each "play" decision.

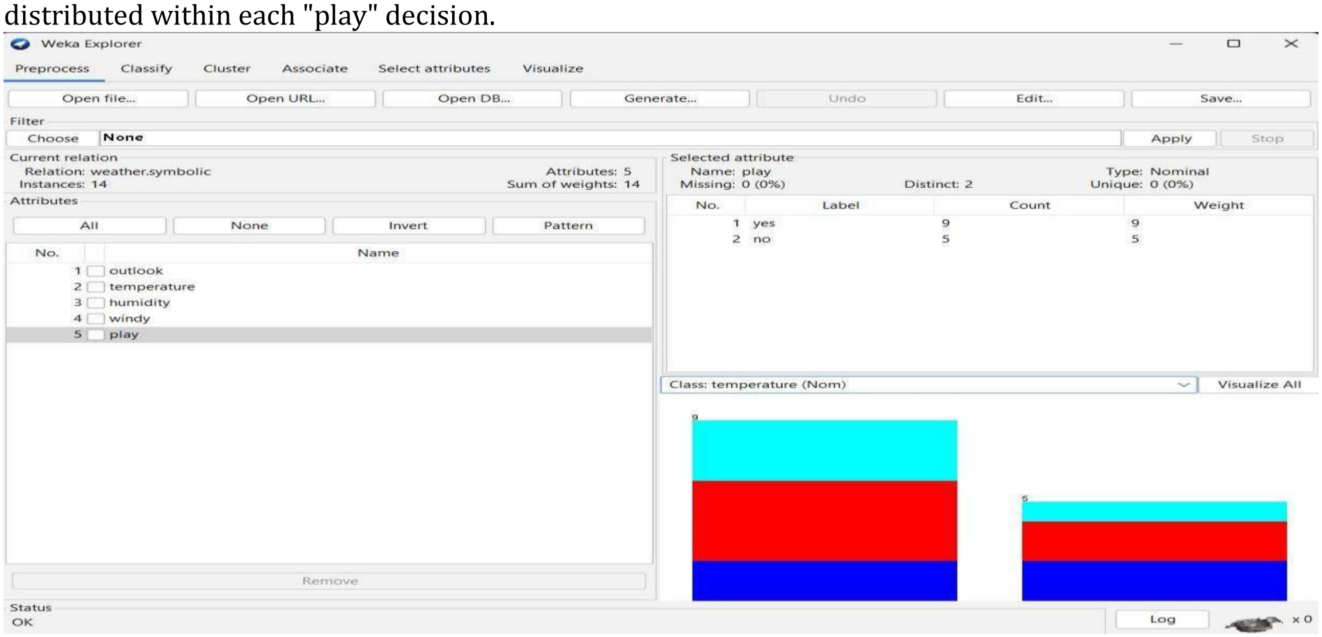

Fig 3.9

- Analysis of "play" attribute with humidity

Description: - This snapshot provides a helpful visual breakdown of how the humidity feature could be influencing decisions in the dataset.

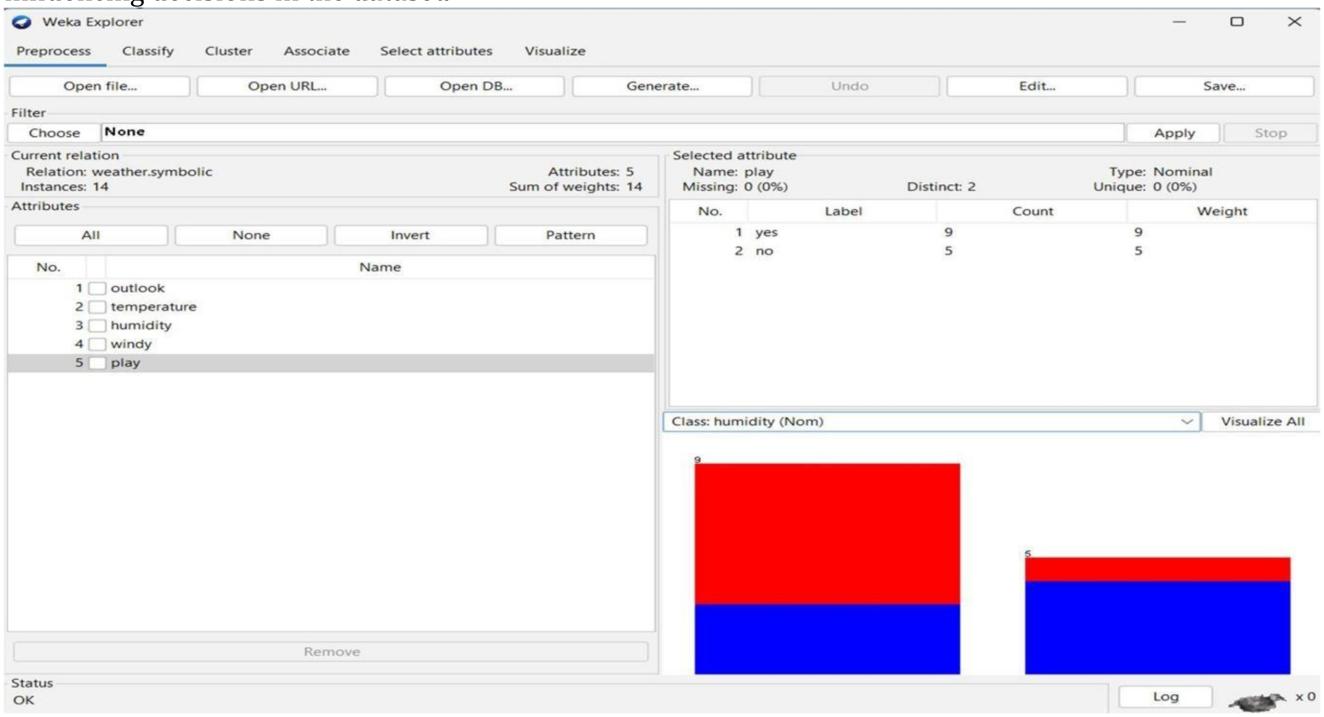

Fig 3.10

- Analysis of "play" attribute with windy

Description: - A bar graph visualizing the distribution of the play attribute with respect to the selected class attribute windy. Bars are color-coded, likely with red and blue representing different values

Fig 3.9

- Analysis of "play" attribute with play class

Description: - The play attribute is being analyzed, likely as the target class for a future classification task. The visualization shows that more instances are labeled as yes (play) than no (don't play).

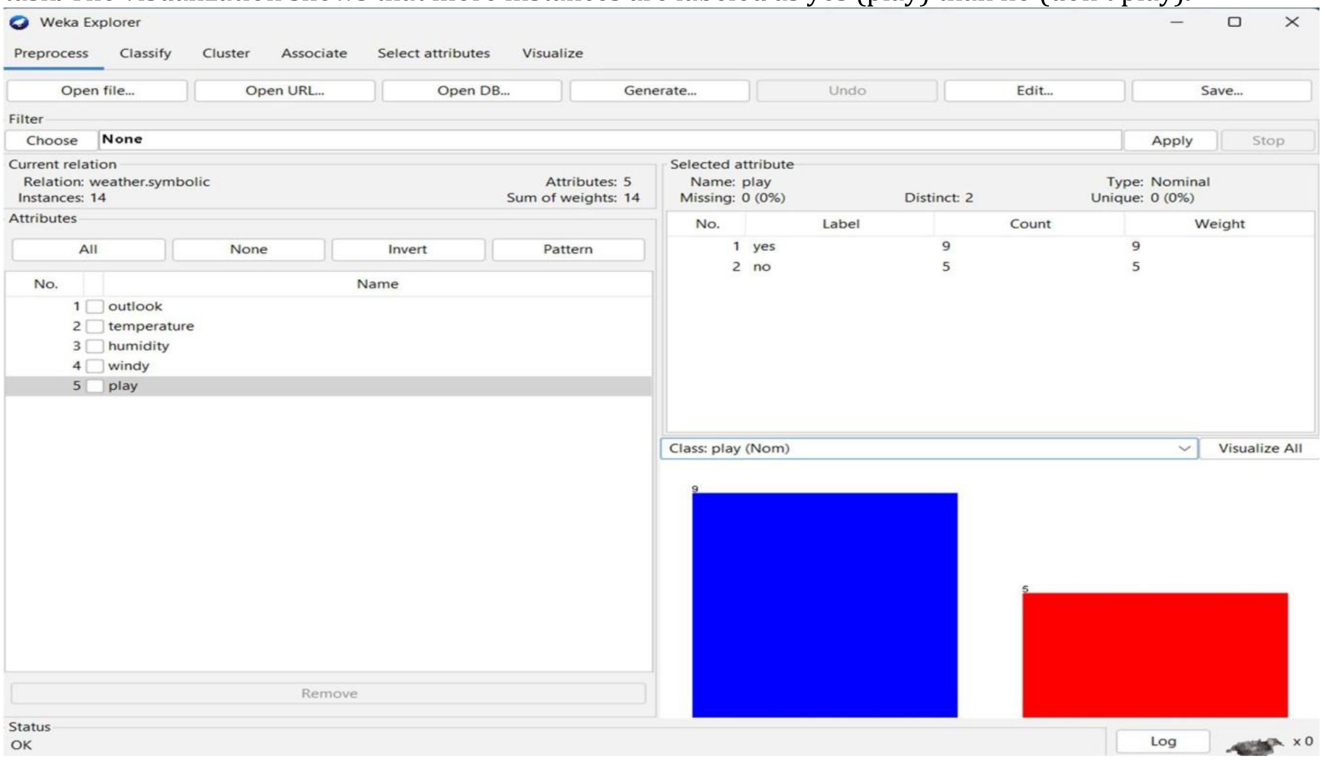

Fig 3.12

- Class visualization of "windy" attribute with temperature class withal other attribute

Description: - Each plot displays the distribution of values for an attribute, with **color-coded bars** representing different classes of temperature.

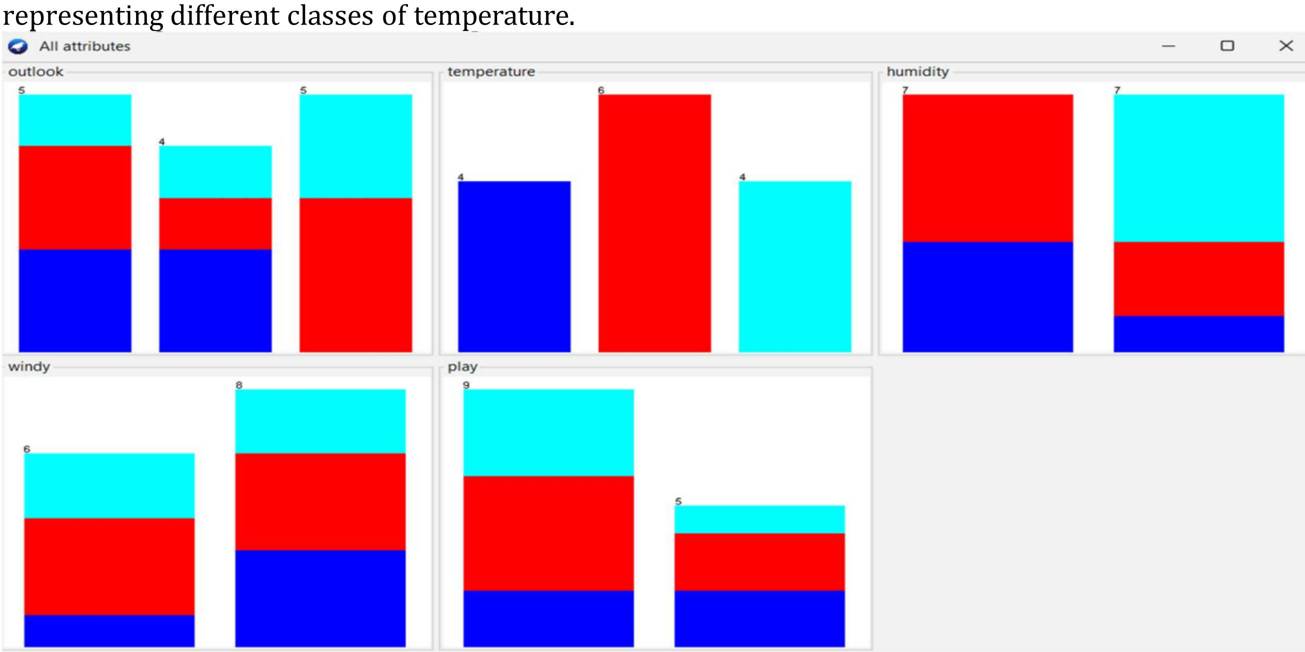

Fig 3.13

- Class visualization of "humidity" over "outlook" class with all attribute

Description: - Each bar in the charts corresponds to a value of the attribute shown on the x-axis. The bars are stacked and color-coded to represent different classes of humidity (most likely high and normal, coded in red, blue, and cyan, though exact mapping may vary depending on your data).

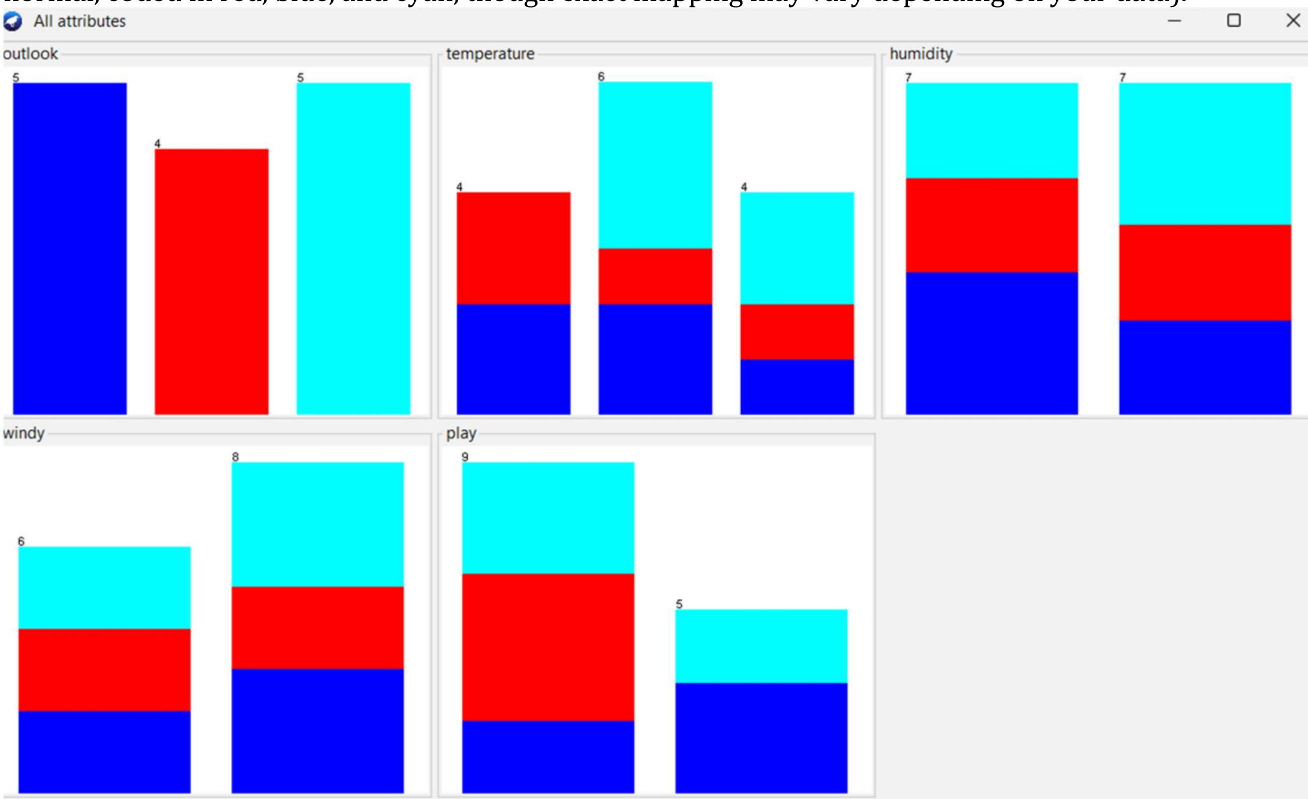

Fig 3.14

- **Analysis of "iris.arff" dataset**

  - Dataset Code

```
@RELATION iris

@ATTRIBUTE  sepallengthREAL
@ATTRIBUTE sepalwidth  REAL
@ATTRIBUTE petallength          REAL
@ATTRIBUTE petalwidth  REAL
@ATTRIBUTE class          {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
```

5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
4.8,3.0,1.4,0.1,Iris-setosa
4.3,3.0,1.1,0.1,Iris-setosa
5.8,4.0,1.2,0.2,Iris-setosa
5.7,4.4,1.5,0.4,Iris-setosa
5.4,3.9,1.3,0.4,Iris-setosa
5.1,3.5,1.4,0.3,Iris-setosa
5.7,3.8,1.7,0.3,Iris-setosa
5.1,3.8,1.5,0.3,Iris-setosa
5.4,3.4,1.7,0.2,Iris-setosa
5.1,3.7,1.5,0.4,Iris-setosa
4.6,3.6,1.0,0.2,Iris-setosa
5.1,3.3,1.7,0.5,Iris-setosa
4.8,3.4,1.9,0.2,Iris-setosa
5.0,3.0,1.6,0.2,Iris-setosa
5.0,3.4,1.6,0.4,Iris-setosa
5.2,3.5,1.5,0.2,Iris-setosa
5.2,3.4,1.4,0.2,Iris-setosa
4.7,3.2,1.6,0.2,Iris-setosa
4.8,3.1,1.6,0.2,Iris-setosa
5.4,3.4,1.5,0.4,Iris-setosa
5.2,4.1,1.5,0.1,Iris-setosa
5.5,4.2,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.0,3.2,1.2,0.2,Iris-setosa
5.5,3.5,1.3,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
4.4,3.0,1.3,0.2,Iris-setosa
5.1,3.4,1.5,0.2,Iris-setosa
5.0,3.5,1.3,0.3,Iris-setosa
4.5,2.3,1.3,0.3,Iris-setosa
4.4,3.2,1.3,0.2,Iris-setosa
5.0,3.5,1.6,0.6,Iris-setosa
5.1,3.8,1.9,0.4,Iris-setosa
4.8,3.0,1.4,0.3,Iris-setosa
5.1,3.8,1.6,0.2,Iris-setosa
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
s6.9,3.1,4.9,1.5,Iris-
versicolor 5.5,2.3,4.0,1.3,Iris-
versicolor

6.5,2.8,4.6,1.5,Iris-versicolor
5.7,2.8,4.5,1.3,Iris-versicolor
6.3,3.3,4.7,1.6,Iris-versicolor
4.9,2.4,3.3,1.0,Iris-versicolor
6.6,2.9,4.6,1.3,Iris-versicolor
5.2,2.7,3.9,1.4,Iris-versicolor
5.0,2.0,3.5,1.0,Iris-versicolor
5.9,3.0,4.2,1.5,Iris-versicolor
6.0,2.2,4.0,1.0,Iris-versicolor
6.1,2.9,4.7,1.4,Iris-versicolor
5.6,2.9,3.6,1.3,Iris-versicolor
6.7,3.1,4.4,1.4,Iris-versicolor
5.6,3.0,4.5,1.5,Iris-versicolor
5.8,2.7,4.1,1.0,Iris-versicolor
6.2,2.2,4.5,1.5,Iris-versicolor
5.6,2.5,3.9,1.1,Iris-versicolor
5.9,3.2,4.8,1.8,Iris-versicolor
6.1,2.8,4.0,1.3,Iris-versicolor
6.3,2.5,4.9,1.5,Iris-versicolor
6.1,2.8,4.7,1.2,Iris-versicolor
6.4,2.9,4.3,1.3,Iris-versicolor
6.6,3.0,4.4,1.4,Iris-versicolor
6.8,2.8,4.8,1.4,Iris-versicolor
6.7,3.0,5.0,1.7,Iris-versicolor
6.0,2.9,4.5,1.5,Iris-versicolor
5.7,2.6,3.5,1.0,Iris-versicolor
5.5,2.4,3.8,1.1,Iris-versicolor
5.5,2.4,3.7,1.0,Iris-versicolor
5.8,2.7,3.9,1.2,Iris-versicolor
6.0,2.7,5.1,1.6,Iris-versicolor
5.4,3.0,4.5,1.5,Iris-versicolor
6.0,3.4,4.5,1.6,Iris-versicolor
6.7,3.1,4.7,1.5,Iris-versicolor
6.3,2.3,4.4,1.3,Iris-versicolor
5.6,3.0,4.1,1.3,Iris-versicolor
5.5,2.5,4.0,1.3,Iris-versicolor
5.5,2.6,4.4,1.2,Iris-versicolor
6.1,3.0,4.6,1.4,Iris-versicolor
5.8,2.6,4.0,1.2,Iris-versicolor
5.0,2.3,3.3,1.0,Iris-versicolor
5.6,2.7,4.2,1.3,Iris-versicolor
5.7,3.0,4.2,1.2,Iris-versicolor
5.7,2.9,4.2,1.3,Iris-versicolor
6.2,2.9,4.3,1.3,Iris-versicolor
5.1,2.5,3.0,1.1,Iris-versicolor
5.7,2.8,4.1,1.3,Iris-versicolor
6.3,3.3,6.0,2.5,Iris-virginica

5.8,2.7,5.1,1.9,Iris-virginica
7.1,3.0,5.9,2.1,Iris-virginica
6.3,2.9,5.6,1.8,Iris-virginica
6.5,3.0,5.8,2.2,Iris-virginica
7.6,3.0,6.6,2.1,Iris-virginica
4.9,2.5,4.5,1.7,Iris-virginica
7.3,2.9,6.3,1.8,Iris-virginica
6.7,2.5,5.8,1.8,Iris-virginica
7.2,3.6,6.1,2.5,Iris-virginica
6.5,3.2,5.1,2.0,Iris-virginica
6.4,2.7,5.3,1.9,Iris-virginica
6.8,3.0,5.5,2.1,Iris-virginica
5.7,2.5,5.0,2.0,Iris-virginica
5.8,2.8,5.1,2.4,Iris-virginica
6.4,3.2,5.3,2.3,Iris-virginica
6.5,3.0,5.5,1.8,Iris-virginica
7.7,3.8,6.7,2.2,Iris-virginica
7.7,2.6,6.9,2.3,Iris-virginica
6.0,2.2,5.0,1.5,Iris-virginica
6.9,3.2,5.7,2.3,Iris-virginica
5.6,2.8,4.9,2.0,Iris-virginica
7.7,2.8,6.7,2.0,Iris-virginica
6.3,2.7,4.9,1.8,Iris-virginica
6.7,3.3,5.7,2.1,Iris-virginica
7.2,3.2,6.0,1.8,Iris-virginica
6.2,2.8,4.8,1.8,Iris-virginica
6.1,3.0,4.9,1.8,Iris-virginica
6.4,2.8,5.6,2.1,Iris-virginica
7.2,3.0,5.8,1.6,Iris-virginica
7.4,2.8,6.1,1.9,Iris-virginica
7.9,3.8,6.4,2.0,Iris-virginica
6.4,2.8,5.6,2.2,Iris-virginica
6.3,2.8,5.1,1.5,Iris-virginica
6.1,2.6,5.6,1.4,Iris-virginica
7.7,3.0,6.1,2.3,Iris-virginica
6.3,3.4,5.6,2.4,Iris-virginica
6.4,3.1,5.5,1.8,Iris-virginica
6.0,3.0,4.8,1.8,Iris-virginica
6.9,3.1,5.4,2.1,Iris-virginica
6.7,3.1,5.6,2.4,Iris-virginica
6.9,3.1,5.1,2.3,Iris-virginica
5.8,2.7,5.1,1.9,Iris-virginica
6.8,3.2,5.9,2.3,Iris-virginica
6.7,3.3,5.7,2.5,Iris-virginica
6.7,3.0,5.2,2.3,Iris-virginica
6.3,2.5,5.0,1.9,Iris-virginica
6.5,3.0,5.2,2.0,Iris-virginica

6.2,3.4,5.4,2.3,Iris-virginica
5.9,3.0,5.1,1.8,Iris-virginica

- Analysis of "student.arff" with weka

- Analysis of "sepallength" attribute

Description: - The histogram reveals a clear class separation: sepallength is a strong predictor, especially for distinguishing Setosa (blue) from the other species. Versicolor and Virginica overlap more, but still show distinguishable distribution tendencies.

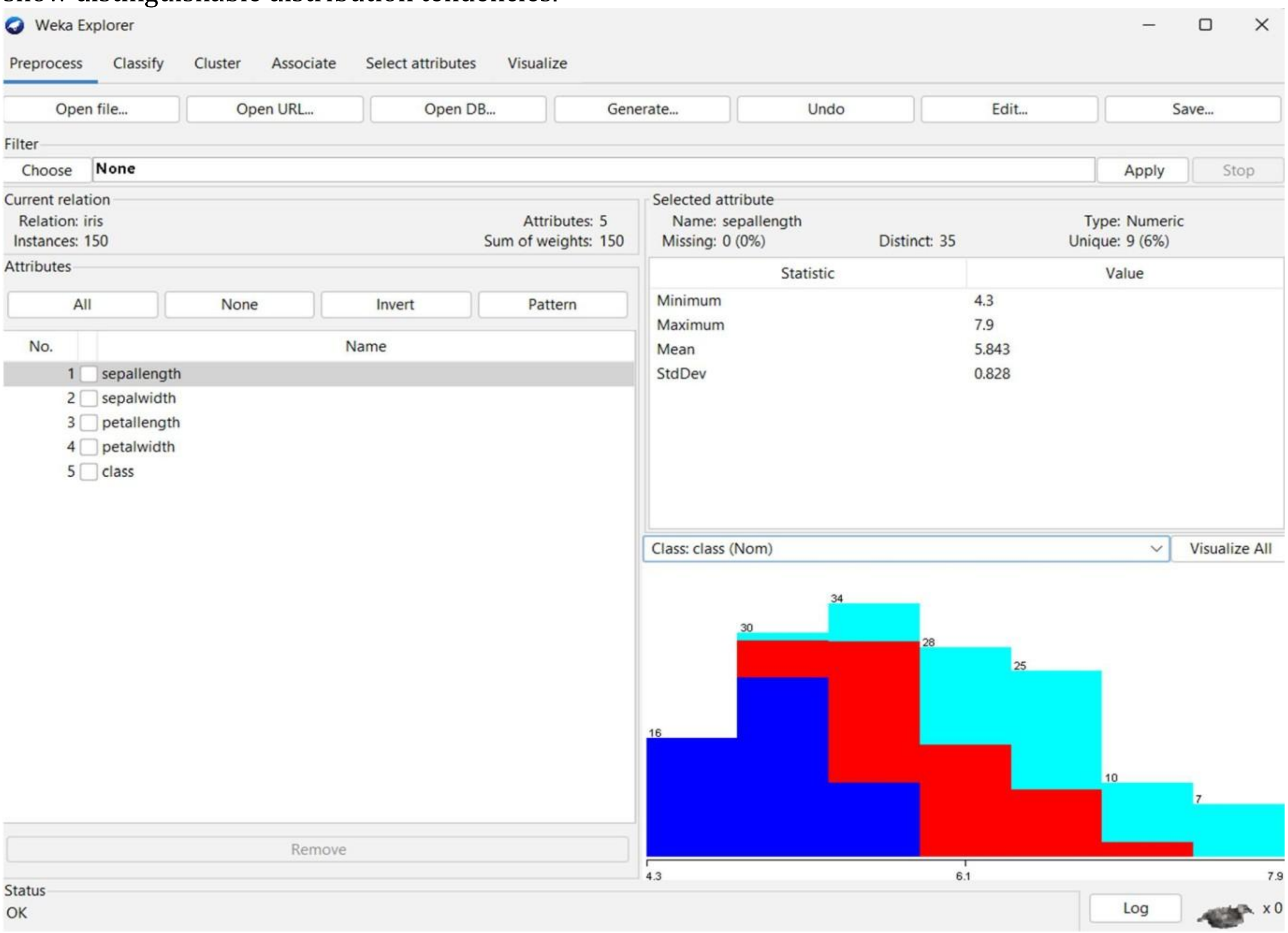

Fig 3.15

- Analysis of "sepalwidth" attribute

Description: - The screenshot provides statistical and visual insight into how the sepalwidth attribute is distributed across different iris species. While it shows some trends (e.g., Setosa tends to have higher sepal width), the overlapping ranges suggest it should be combined with other attributes for accurate classification.

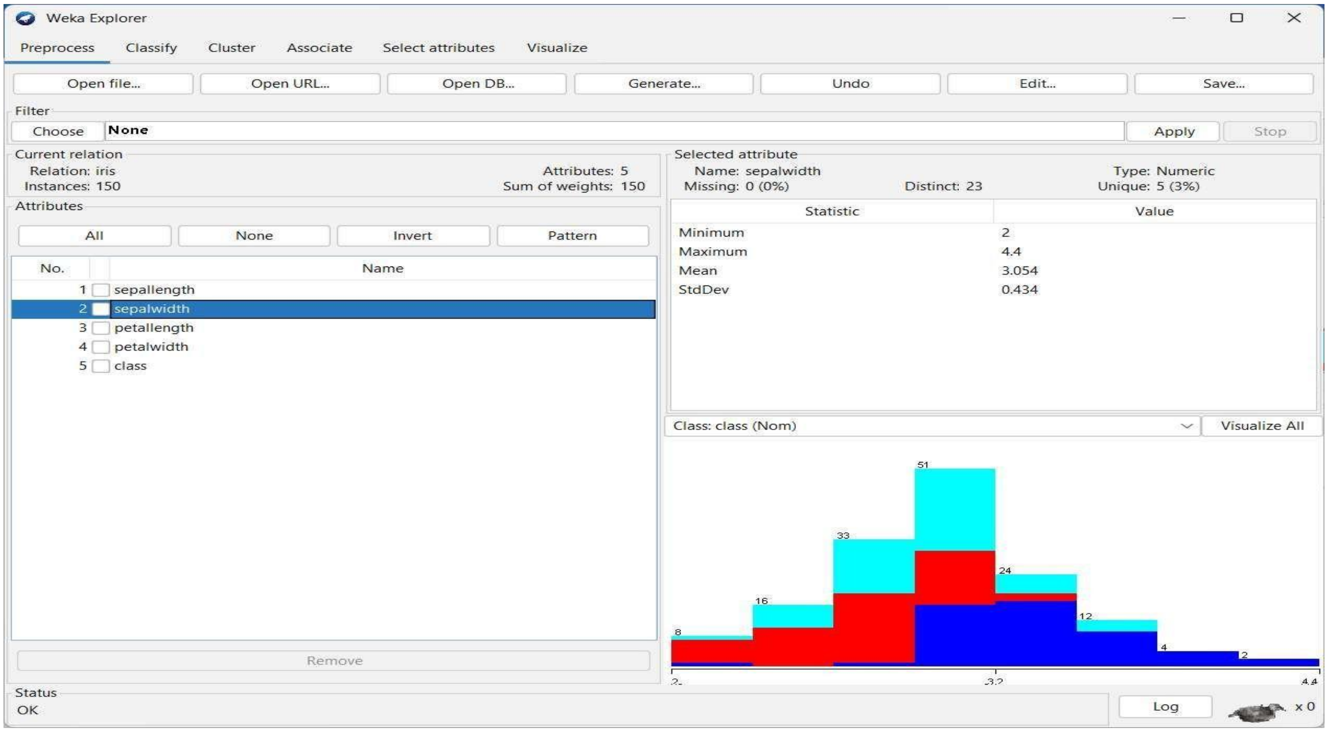

Fig 3.16

- Analysis of “petallength” attribute

Description: - The attribute petallength is a strong discriminating feature in the Iris dataset: Some overlap exists between *versicolor* and *virginica*, but it is still reasonably distinct. High standard deviation indicates wide variability, making it informative for classification.

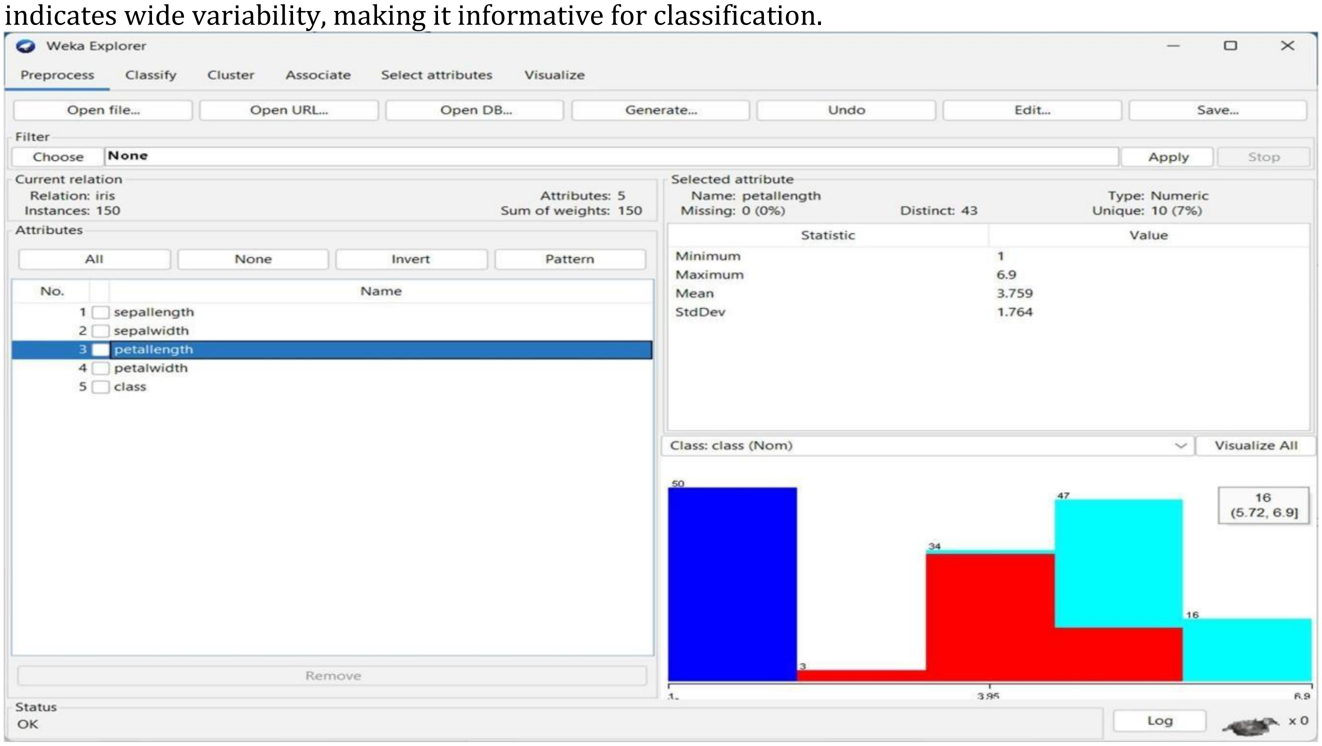

Fig 3.17

- Analysis of “petalwidth” attribute

Description: - The attribute petalwidth is selected, displaying its statistical summary and a histogram showing its distribution across different iris species.

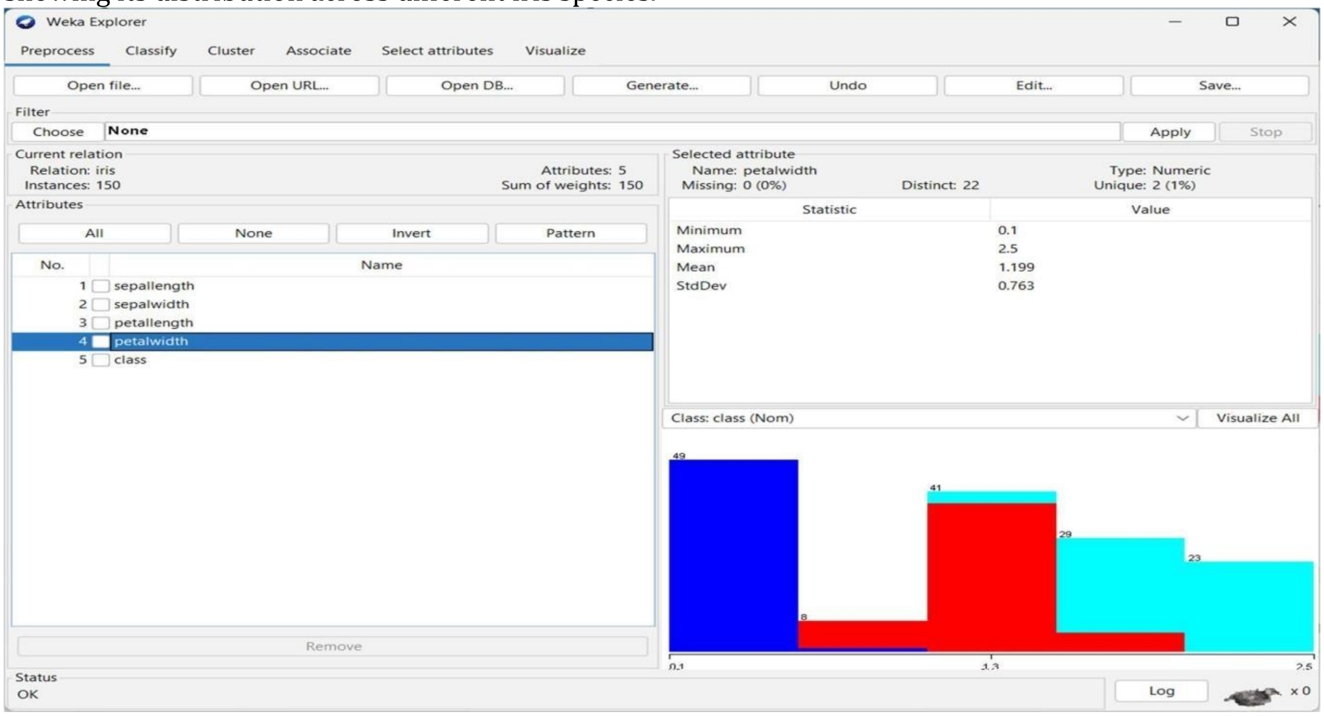

Fig 3.18

- Analysis of “class” attribute

Description: - It analysis the three different species with respect to class. The three species are iris-sentosa, iris-versicolor, iris-virginica.

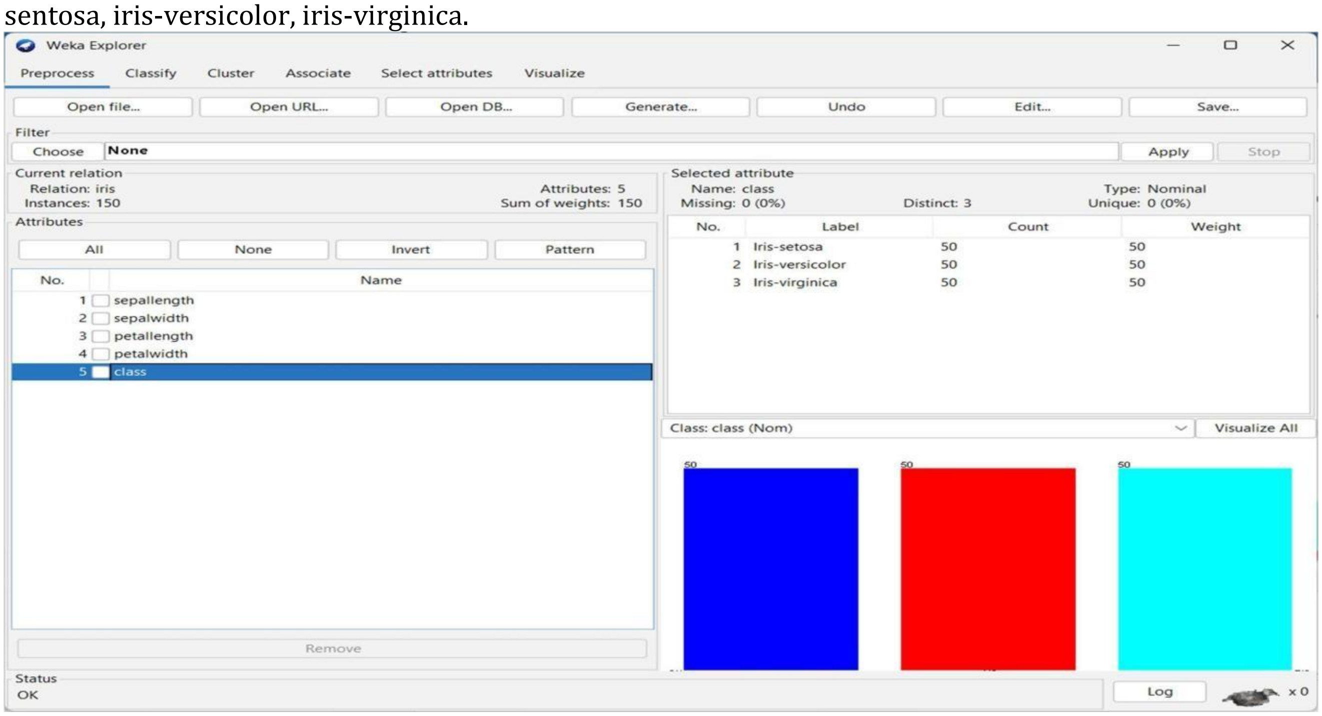

Fig 3.19

- Class visualization of “class” attribute over “class” class with all other attribute

Description: - The petal length and width are the most discriminative features for distinguishing between Iris species, while sepal features show moderate overlap. Equal class distribution ensures balanced learning in classification tasks.

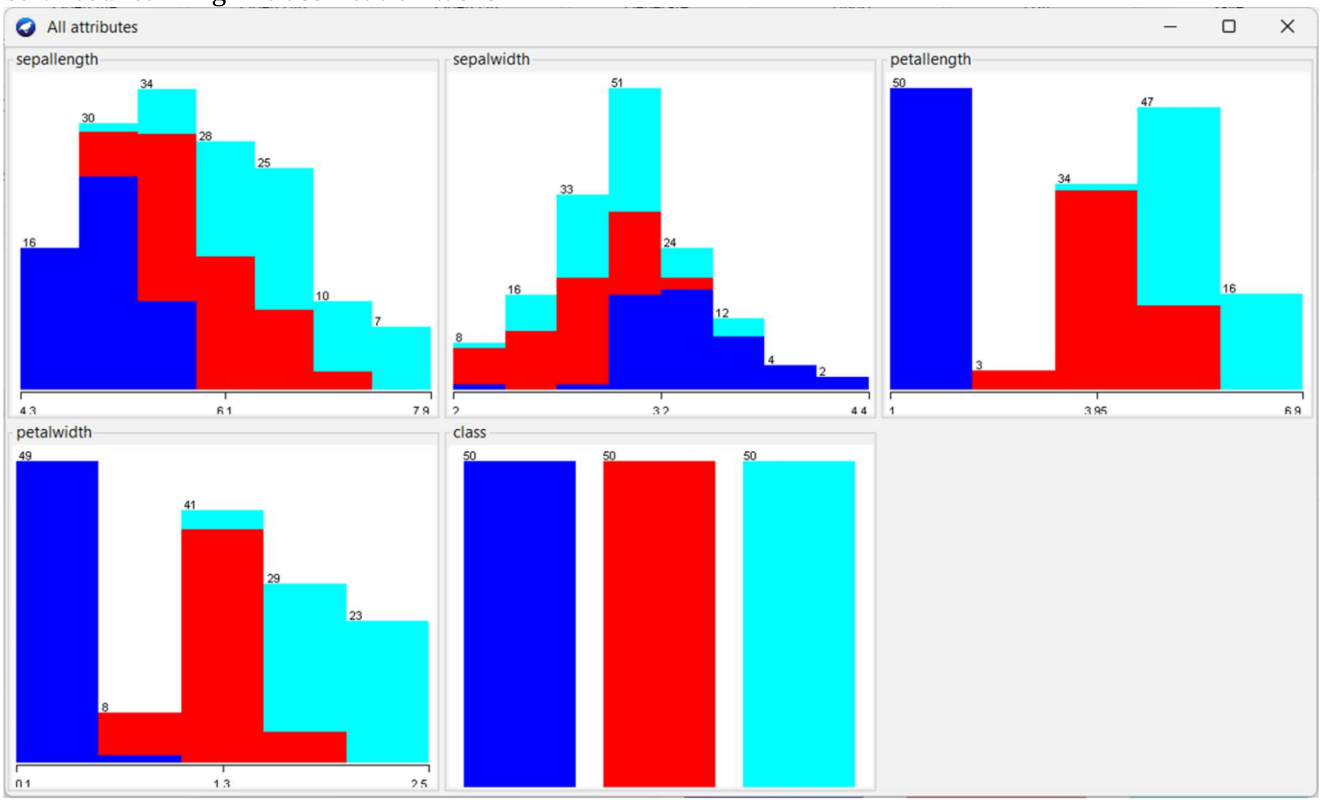

Fig 3.20

Experiment Outcome:

The experiment involved preparing and analyzing the "student", "weather.nominal", and "iris" datasets using the WEKA tool. Various data attributes were visualized to understand patterns such as gender-based placement, class distribution, and regional trends. In the weather dataset, the "play" decision was analyzed against attributes like temperature and humidity to infer decision influences. For the iris dataset, petal and sepal measurements were found to be effective in classifying species. Overall, the exercise enhanced skills in data preprocessing, visualization, and attribute-based analysis for classification purposes.