EXP NO: **DATA CLEANING: IMPLEMENT VARIOUS MISSING HANDLING MECHANISMS, IMPLEMENT VARIOUS NOISY HANDLING MECHANISMS**

DATE:

AIM:

BACKGROUND THEORY:

Data Cleaning:

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabelled.

Etl Process:

Data Cleaning is an important part of ETL processes as it ensures that only high-quality data is loaded into the Data Warehouse. This helps to improve the accuracy of security decisions.

Data Warehousing is a process of organizing and storing data in a centralized location for easy access and analysis. Data warehousing is used to store historical data from multiple sources in a single location. Data warehouses provide a single view of data that can be used for reporting and analysis. Data warehouses are often used in business intelligence applications.

PROCEDURE:

1. Load and Inspect Data:

o Use the “File” widget to load dataset.csv.

o Connect it to the “Data Table” widget to inspect the data.

2. Handle Missing Data:

o Connect the “File” widget to the “Impute” widget.

o In the “Impute” widget, choose “Mean/Median” for numeric features and “Most Frequent” for categorical features.

o Connect the “Impute” widget to a “Data Table” widget to inspect the imputed data.

3. Remove Columns with Excessive Missing Data:

o Connect the “Impute” widget to the “Select Columns” widget.

o In the “Select Columns” widget, manually remove columns with a high percentage of

missing values.

4. Handle Noisy Data:

o Connect the “Select Columns” widget to the “Smoothing” widget.

o In the “Smoothing” widget, apply “Binning” for continuous variables.

o Connect the “Smoothing” widget to a “Data Table” widget to inspect the smoothed data.

5. Detect and Remove Outliers:

o Connect the “Smoothing” widget to the “Outliers” widget.

o In the “Outliers” widget, choose the “Z-Score” method and set a threshold (e.g., 3).

o Connect the “Outliers” widget to a “Data Table” widget to inspect the data with outliers removed.

OUTPUT:

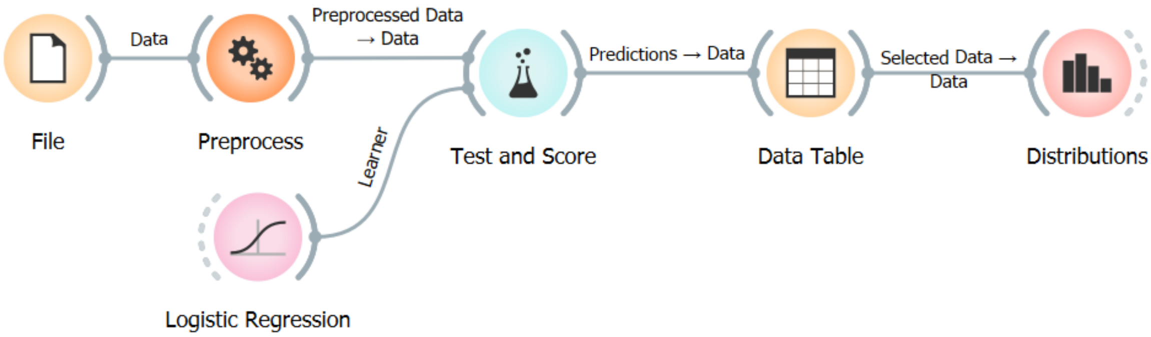
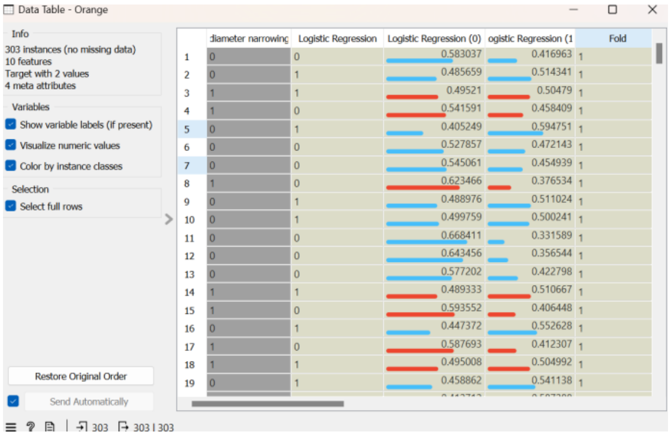


FIG 2.1: IMPLEMENTATION OF DATA PREPROCESSING AND LOGISTIC REGRESSION



Info
303 instances (no missing data)
10 features
Target with 2 values
4 meta attributes

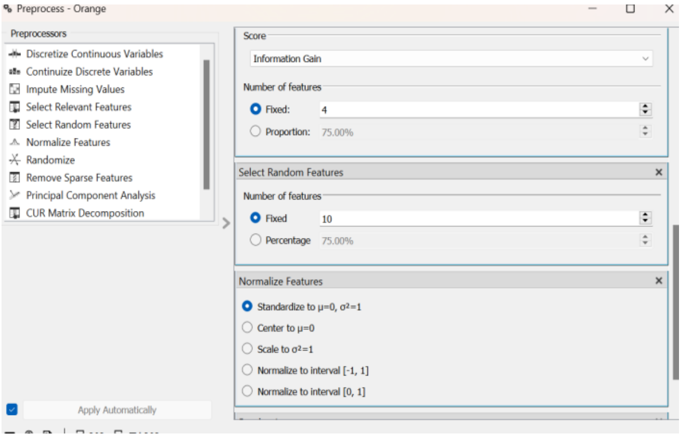
Variables
☑ Show variable labels (if present)
☑ Visualize numeric values
☑ Color by instance classes

Selection
☑ Select full rows

Restore Original Order
☑ Send Automatically

303 | 303 | 303

	diameter narrowing	Logistic Regression	Logistic Regression (0)	ogistic Regression (1	Fold
1	0	0	0.583037	0.416963	1
2	0	1	0.485659	0.514341	1
3	1	1	0.49521	0.50479	1
4	1	0	0.541591	0.458409	1
5	0	1	0.405249	0.594751	1
6	0	0	0.527857	0.472143	1
7	0	0	0.545061	0.454939	1
8	1	0	0.623466	0.376534	1
9	0	1	0.488976	0.511024	1
10	0	1	0.499759	0.500241	1
11	0	0	0.668411	0.331589	1
12	0	0	0.643456	0.356544	1
13	0	0	0.577202	0.422798	1
14	1	1	0.489333	0.510667	1
15	1	0	0.593552	0.406448	1
16	0	1	0.447372	0.552628	1
17	1	0	0.587693	0.412307	1
18	1	1	0.495008	0.504992	1
19	0	1	0.458862	0.541138	1



Preprocessors
Discretize Continuous Variables
Continuize Discrete Variables
Impute Missing Values
Select Relevant Features
Select Random Features
Normalize Features
Randomize
Remove Sparse Features
Principal Component Analysis
CUR Matrix Decomposition

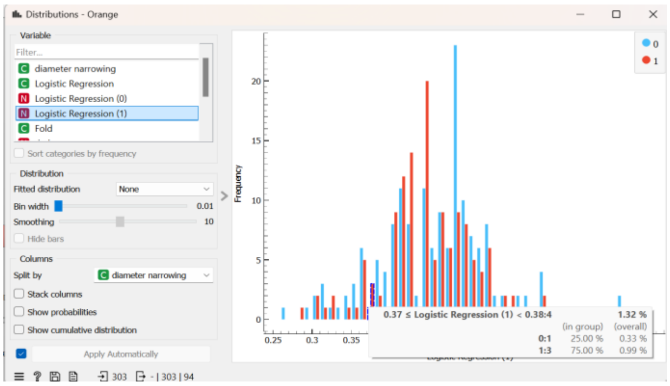
☑ Apply Automatically

Score
Information Gain
Number of features
● Fixed: 4
○ Proportion: 75.00%

Select Random Features
Number of features
● Fixed 10
○ Percentage 75.00%

Normalize Features
● Standardize to μ=0, σ²=1
○ Center to μ=0
○ Scale to σ²=1
○ Normalize to interval [-1, 1]
○ Normalize to interval [0, 1]

FIG 2.1.1: IMPLEMENTATION OF LOGISTIC REGRESSION FIG 2.1.2: IMPLEMENTATION OF DATA PREPROCESSING



Variable
Filter...
C diameter narrowing
C Logistic Regression
N Logistic Regression (0)
N Logistic Regression (1)
C Fold
□ Sort categories by frequency

Distribution
Fitted distribution: None
Bin width 0.01
Smoothing 10
□ Hide bars

Columns
Split by: C diameter narrowing
□ Stack columns
□ Show probabilities
□ Show cumulative distribution
☑ Apply Automatically

303 | - | 303 | 94

0.37 ≤ Logistic Regression (1) < 0.38:4 1.32 %

	(in group)	(overall)
0:1	25.00 %	0.33 %
1:3	75.00 %	0.99 %

FIG 2.2: DISTRIBUTION OF LOGISTIC REGRESSION

RESULT: