St. Francis Institute of Technology, Mumbai-400 103

**Department Of Information Technology**

A.Y. 2024-2025

Class: TE-ITA/B, Semester: VI

Subject: **Business Intelligence Lab**

**Experiment – 3: Data Preprocessing in WEKA Tool**

1. **Aim:** Data Preprocessing in WEKA Tool.
2. **Objectives:** After study of this experiment, the students will be able to
   * Understand and know how data is preprocessed in Weka.
3. **Outcomes:**

After study of this experiment, the students will be able to

**CO2:** Organize and prepare the data needed for data mining using pre preprocessing

techniques.

**CO3:** Perform exploratory analysis of the data to be used for mining

1. **Prerequisite:** Introduction to steps in data preprocessing.
2. **Requirements:** Personal Computer, Windows XP operating system/Windows 7, Internet

Connection, Microsoft Word, WEKA tool.

1. **Theory:**
2. Introduction to Weka.
3. What is Data Preprocessing in data Mining?

Data preprocessing in data mining refers to the process of cleaning, transforming, and organizing raw data into a structured format suitable for analysis or modeling. It involves tasks like handling missing values, removing duplicates, normalizing data, encoding categorical variables, and scaling features. This step is crucial as it ensures the quality of the data and enhances the performance of machine learning models.

1. Why do you need Preprocessing?

Preprocessing is essential because raw data is often messy, inconsistent, or incomplete. Here's why it's needed:

**Improves Data Quality**: It helps in handling missing values, outliers, and noise, ensuring that the data is accurate and reliable.

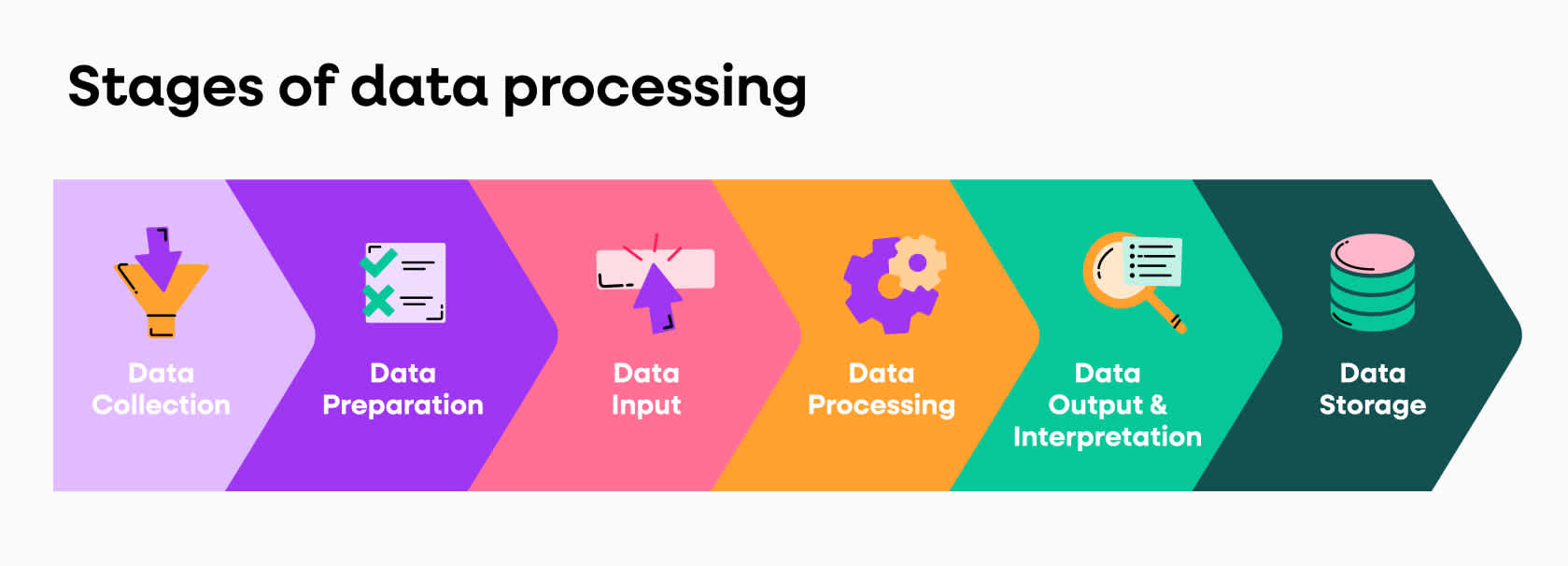
**Enhances Model Accuracy**: Cleaned and well-processed data leads to better model performance by reducing errors and improving predictions.

**Reduces Complexity**: By transforming data into a more usable format (like normalizing or encoding), it simplifies the modeling process.

**Ensures Consistency**: Preprocessing standardizes data from different sources, making it easier to analyze and compare.

**Faster Processing**: Well-prepared data can be processed faster, making it more efficient for analysis or machine learning algorithms.

4. Steps involved in Data Preprocessing.



Data Collection: Gathering raw data from various sources, such as databases, spreadsheets, APIs, etc.

Data Cleaning:

* Handling Missing Values: Fill, delete, or estimate missing data.
* Removing Duplicates: Identify and eliminate duplicate records.
* Handling Outliers: Detect and treat outliers that may skew results.

Data Transformation:

* Normalization/Standardization: Rescale data to a standard range (e.g., 0 to 1) or ensure features have a similar scale.
* Encoding Categorical Data: Convert categorical variables (e.g., gender, color) into numerical values (e.g., one-hot encoding).
* Feature Engineering: Create new features or modify existing ones to improve model performance.

Data Reduction:

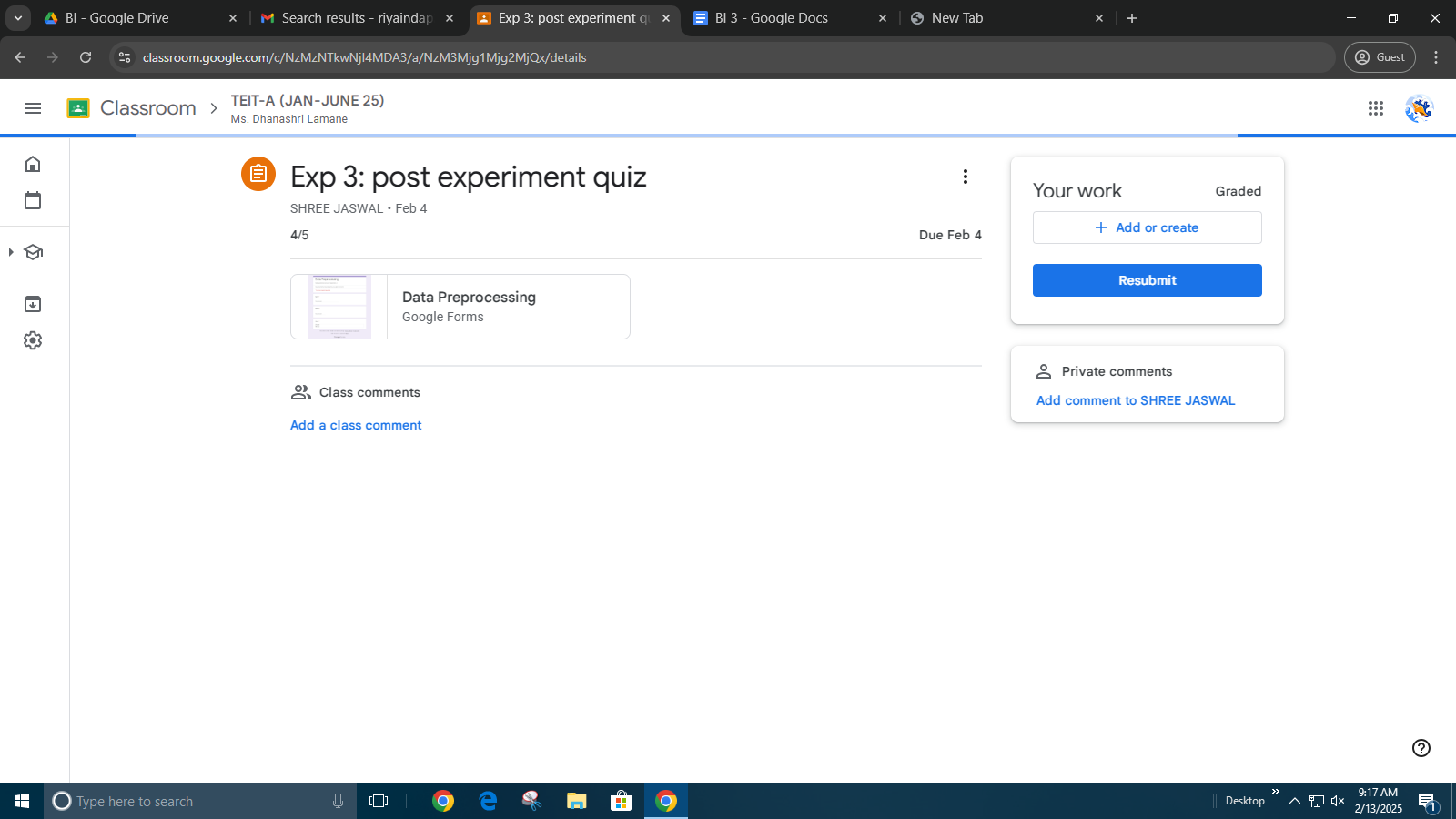
* Feature Selection: Remove irrelevant or redundant features to reduce dimensionality.
* Dimensionality Reduction: Use techniques like PCA (Principal Component Analysis) to reduce the number of features while preserving data variance.

Data Splitting: Split the dataset into training and testing subsets to evaluate model performance later.

1. **Laboratory Exercise:** Implementation of Data Preprocessing in WEKA and take printout of related snapshots.
2. **Post-Experiments Exercise**

**A  Questions:**

                   In form of MCQ type test



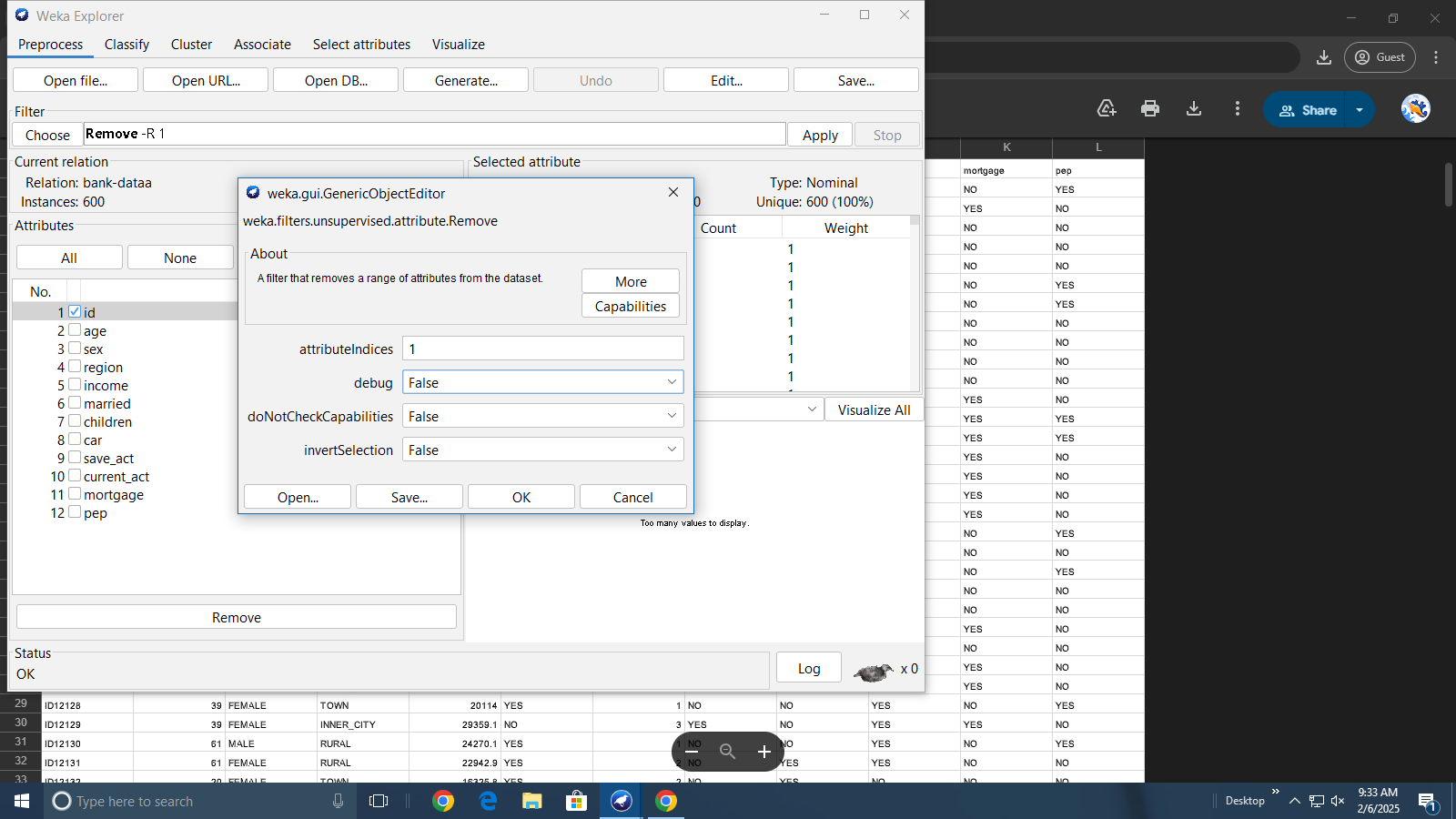
**B  Conclusion:**

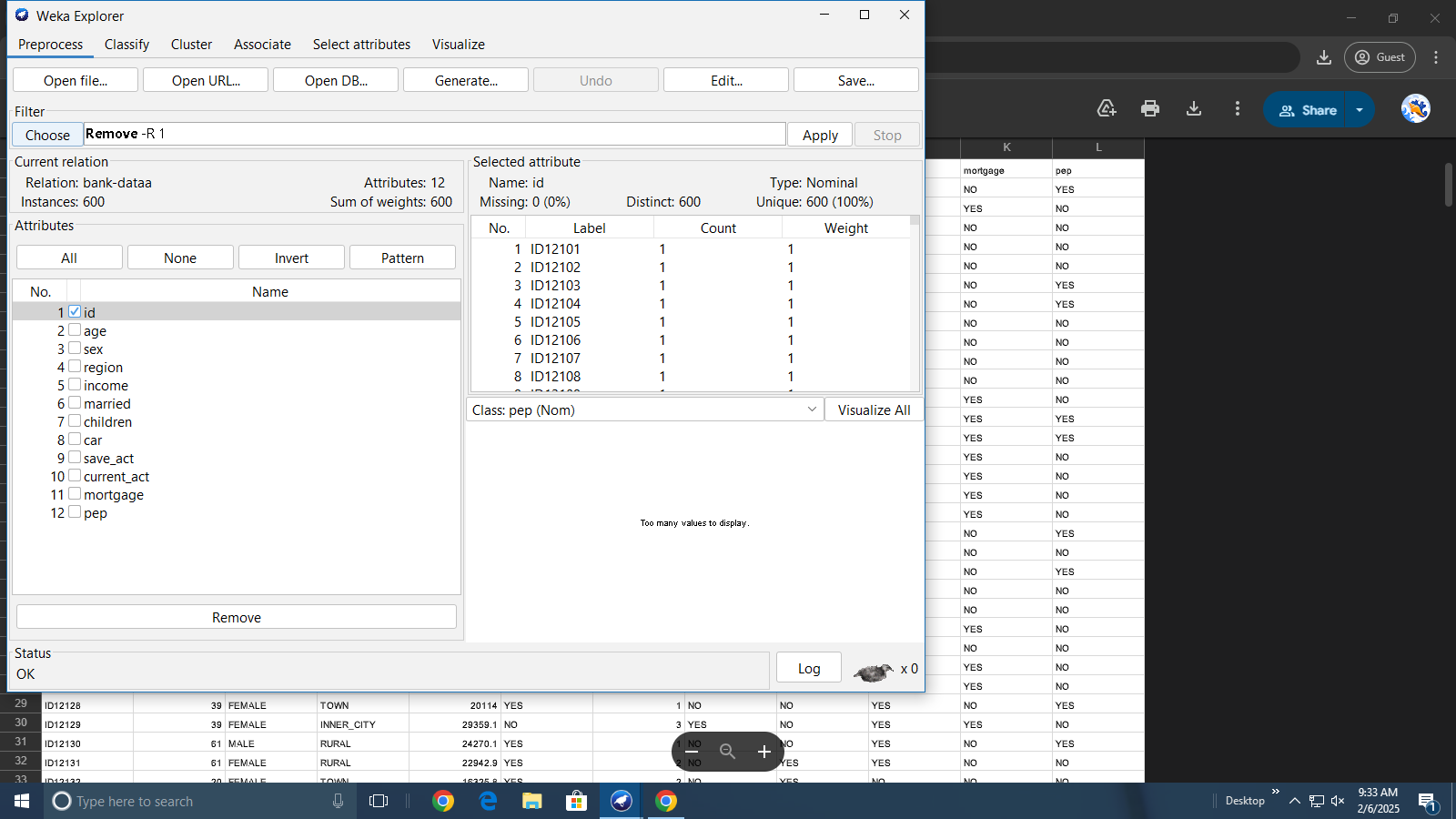
1. Summary of Experiment

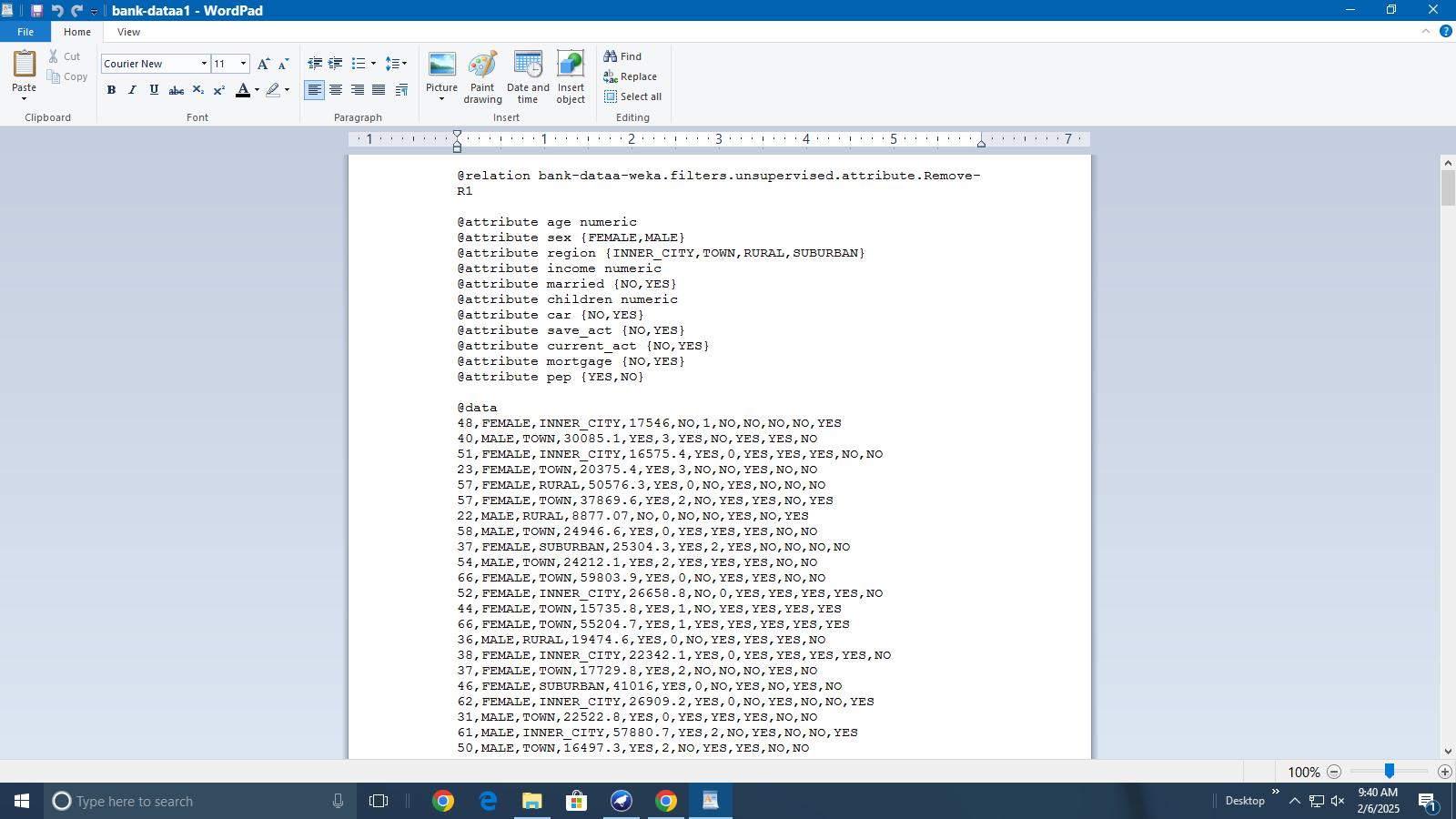
2. Importance of Experiment

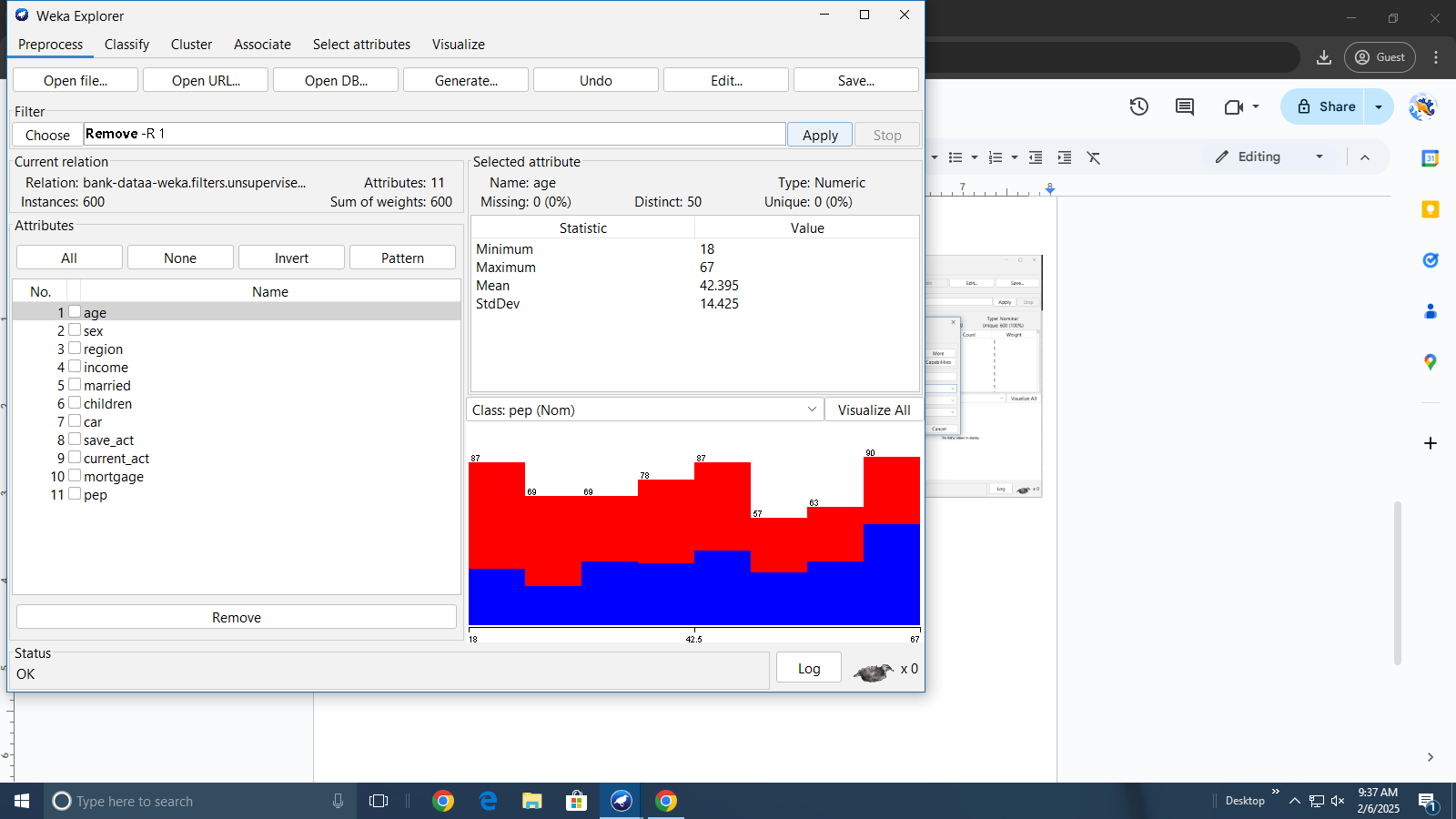
3. Application of Experiment

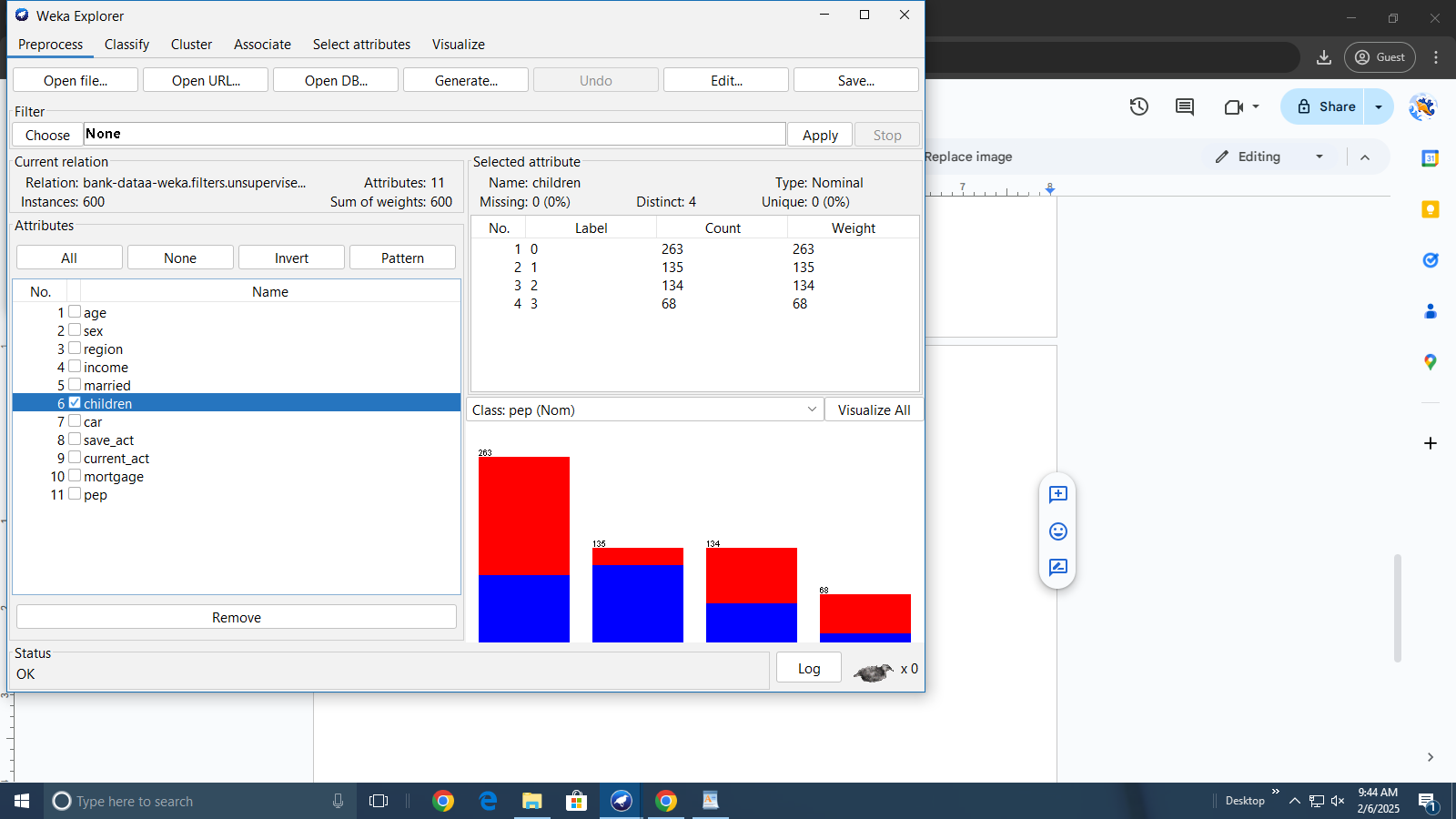
1. **Reference:** Data Mining: Concept & Techniques, 3rd Edition, Jiawei Han, Micheline Kamber, Jian Pei, Elsevier.

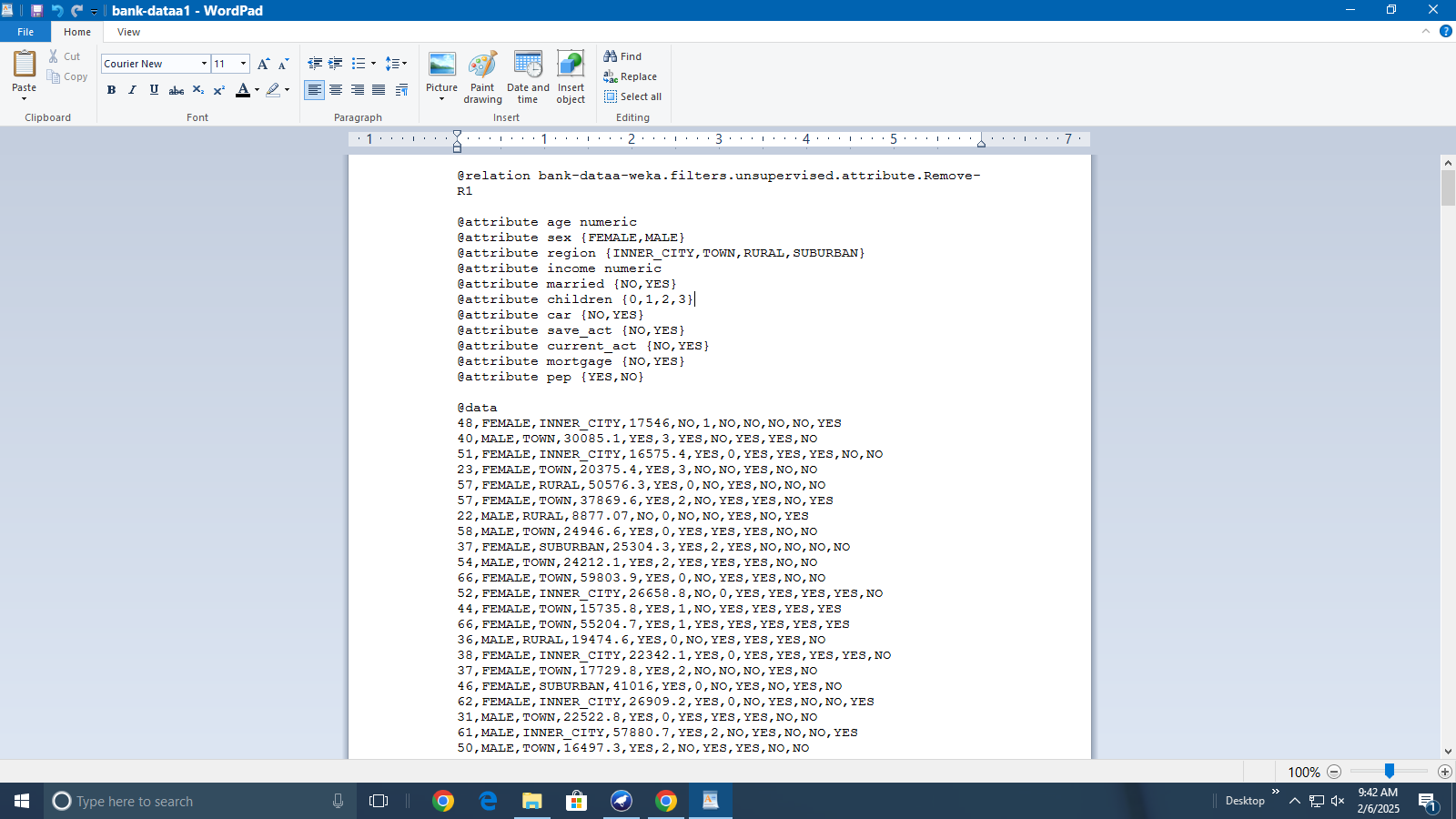
Dataset Loading Removing an attribute

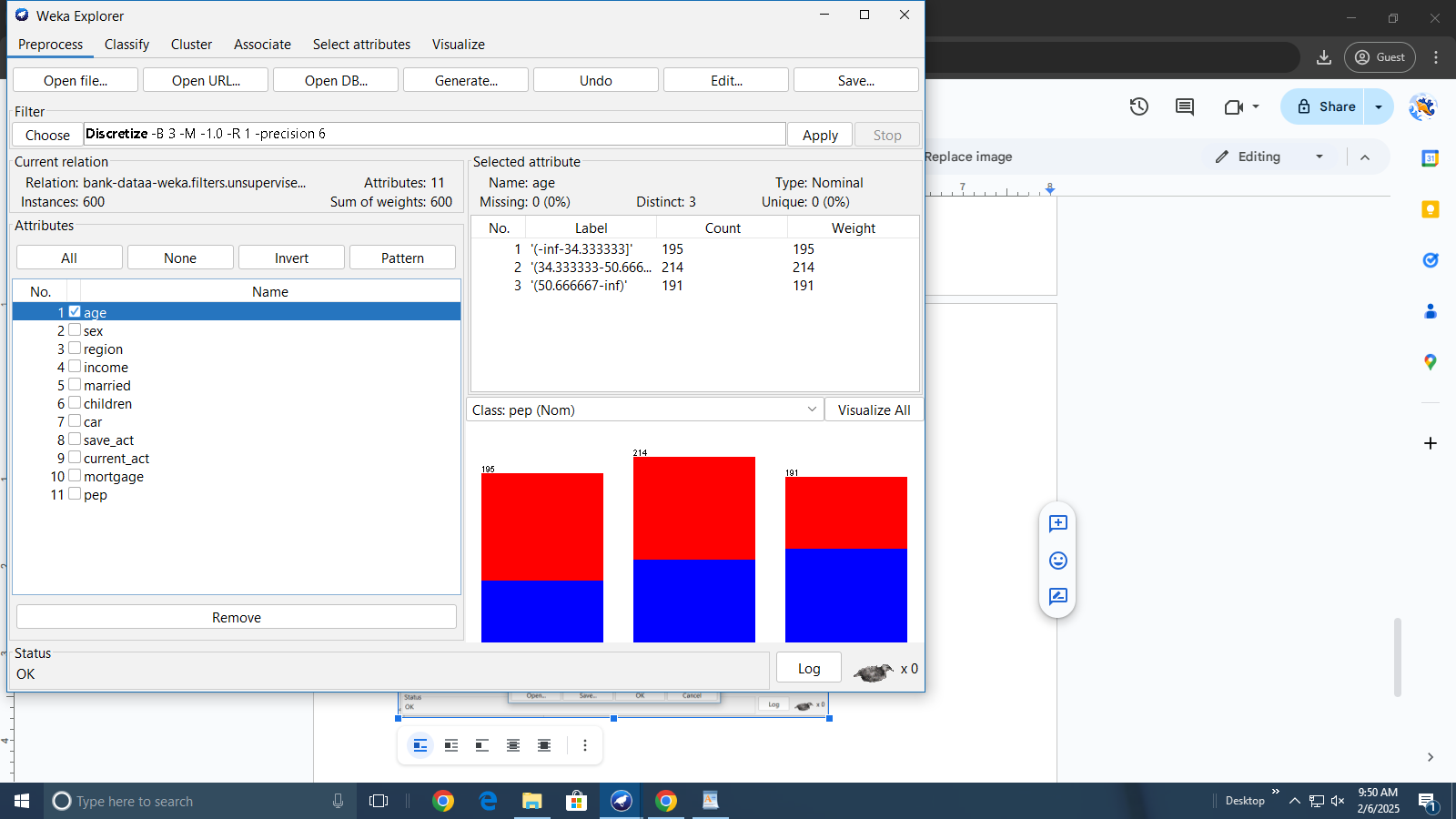


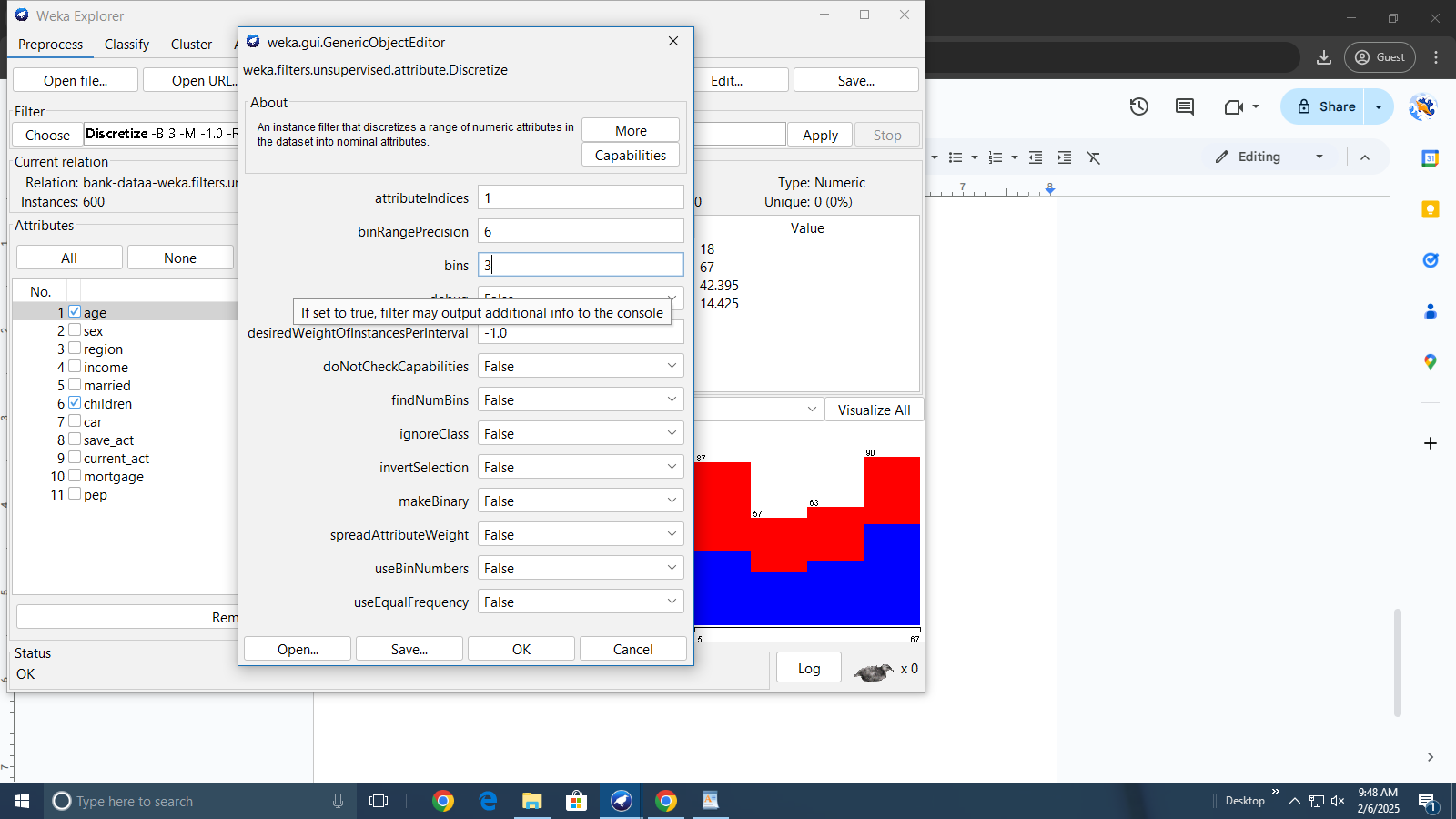
Removed attribute Dataset with children attribute in numeric format

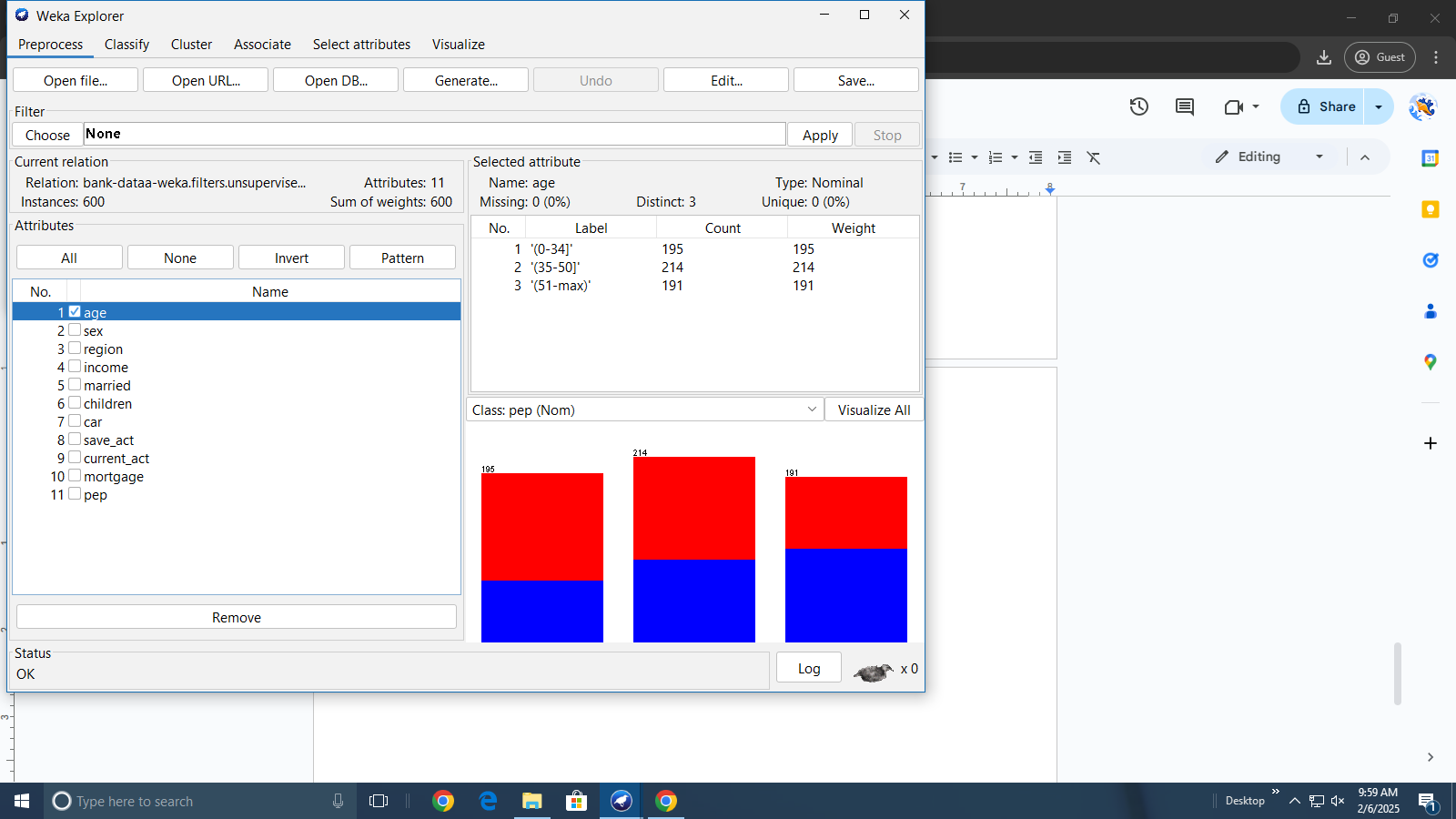


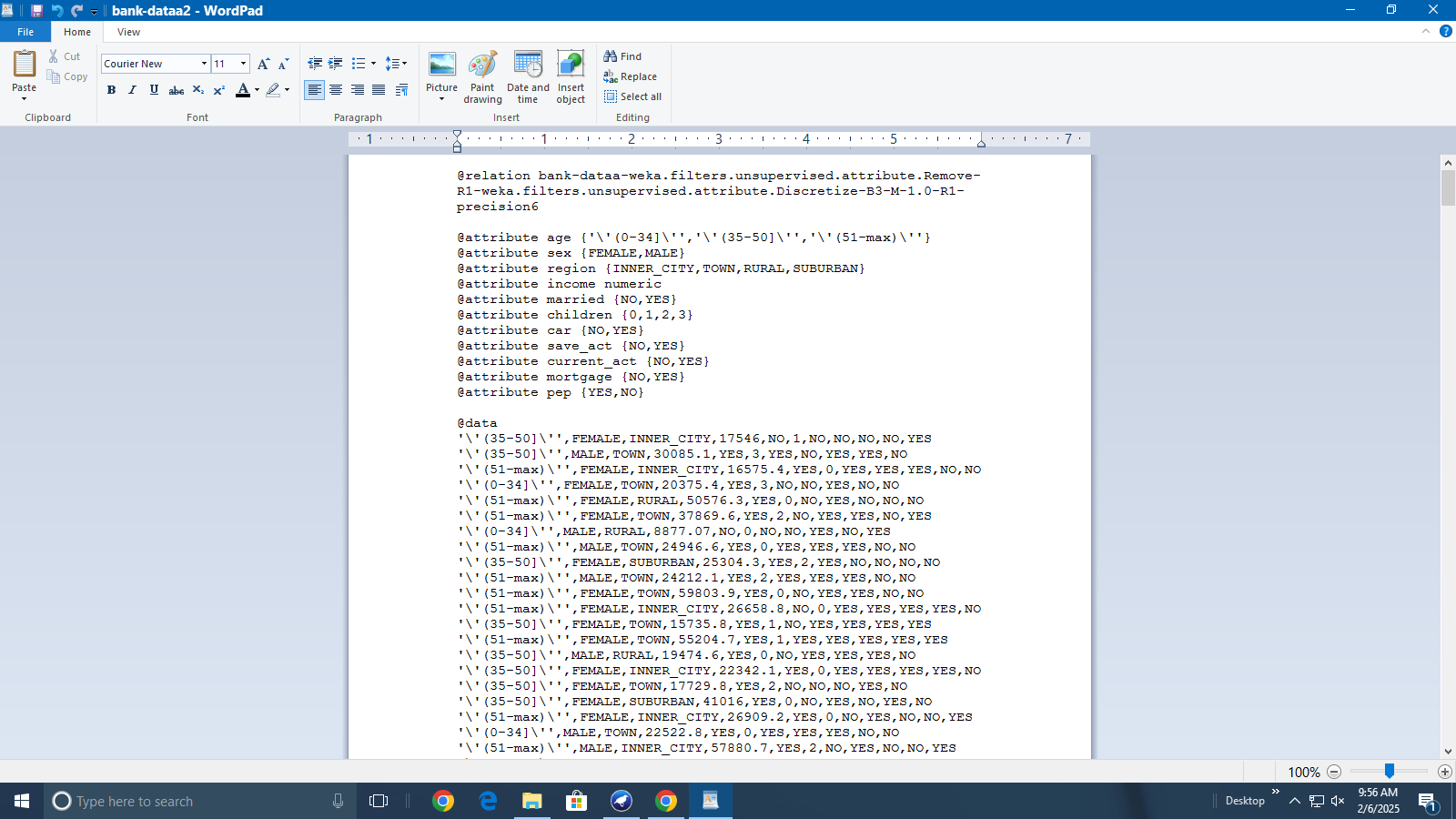
Discretization of children attribute without WEKA Discretized children attribute visible in WEKA

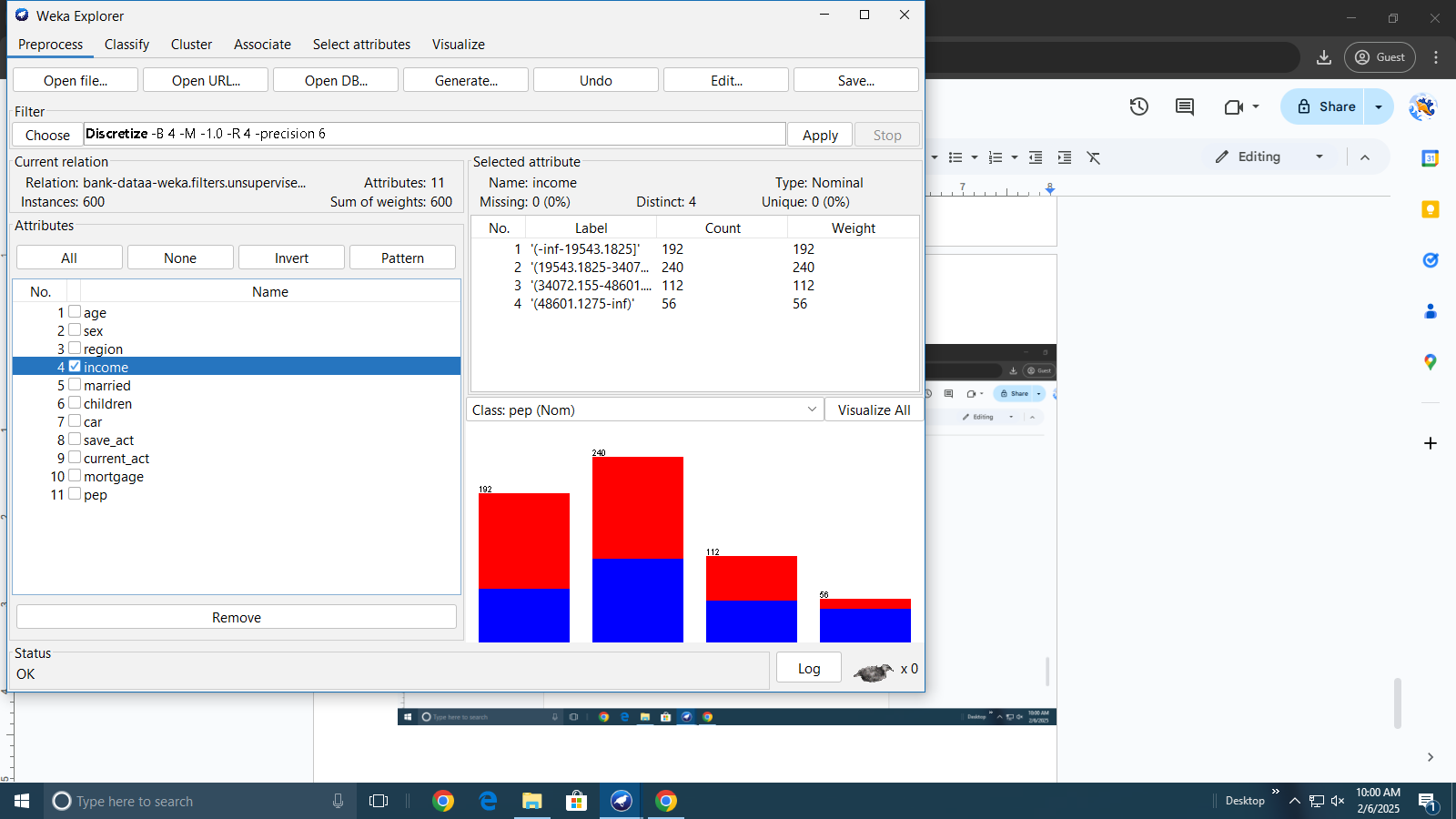


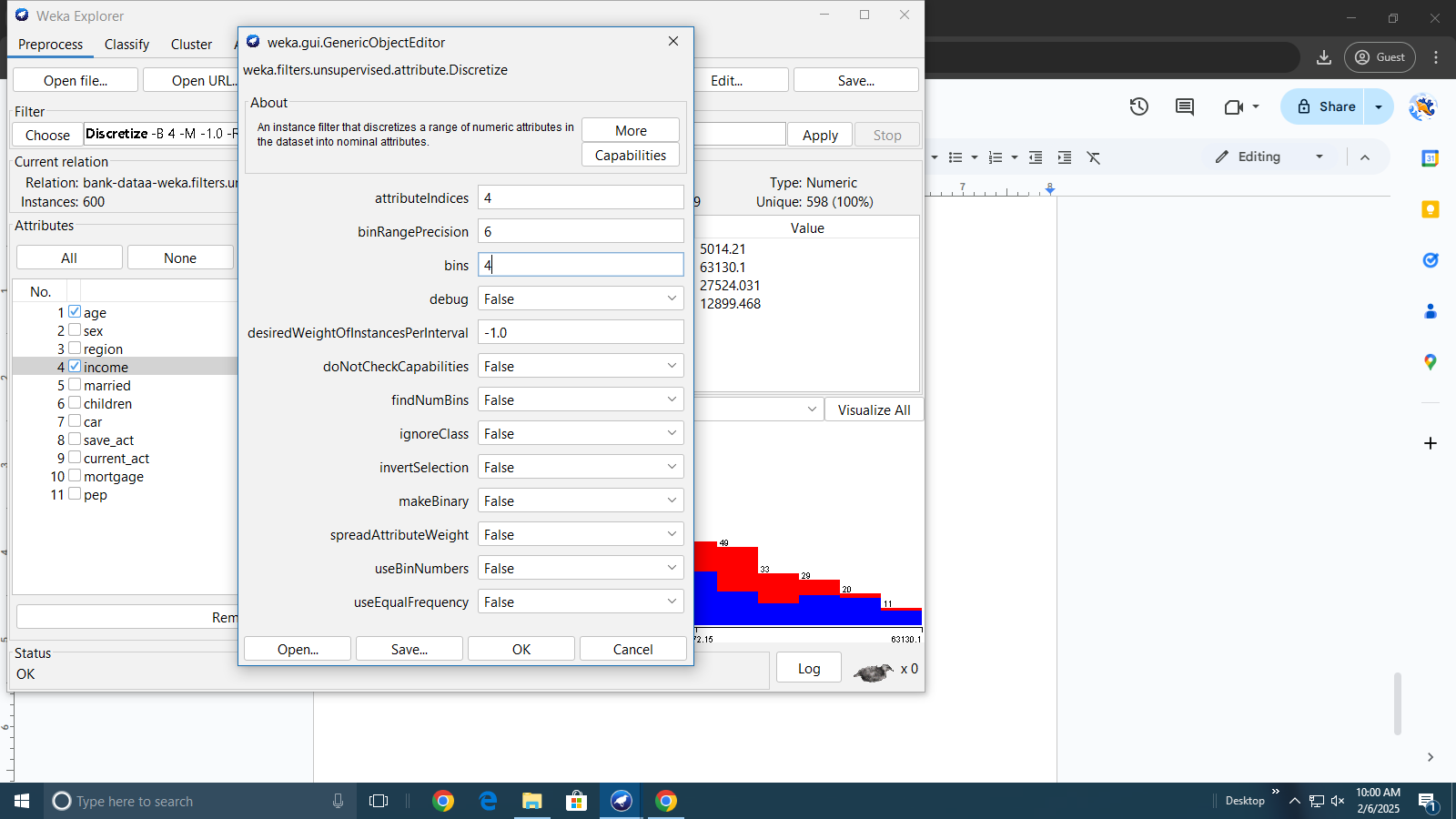
Discretizing age attribute in 3 bins after discretization

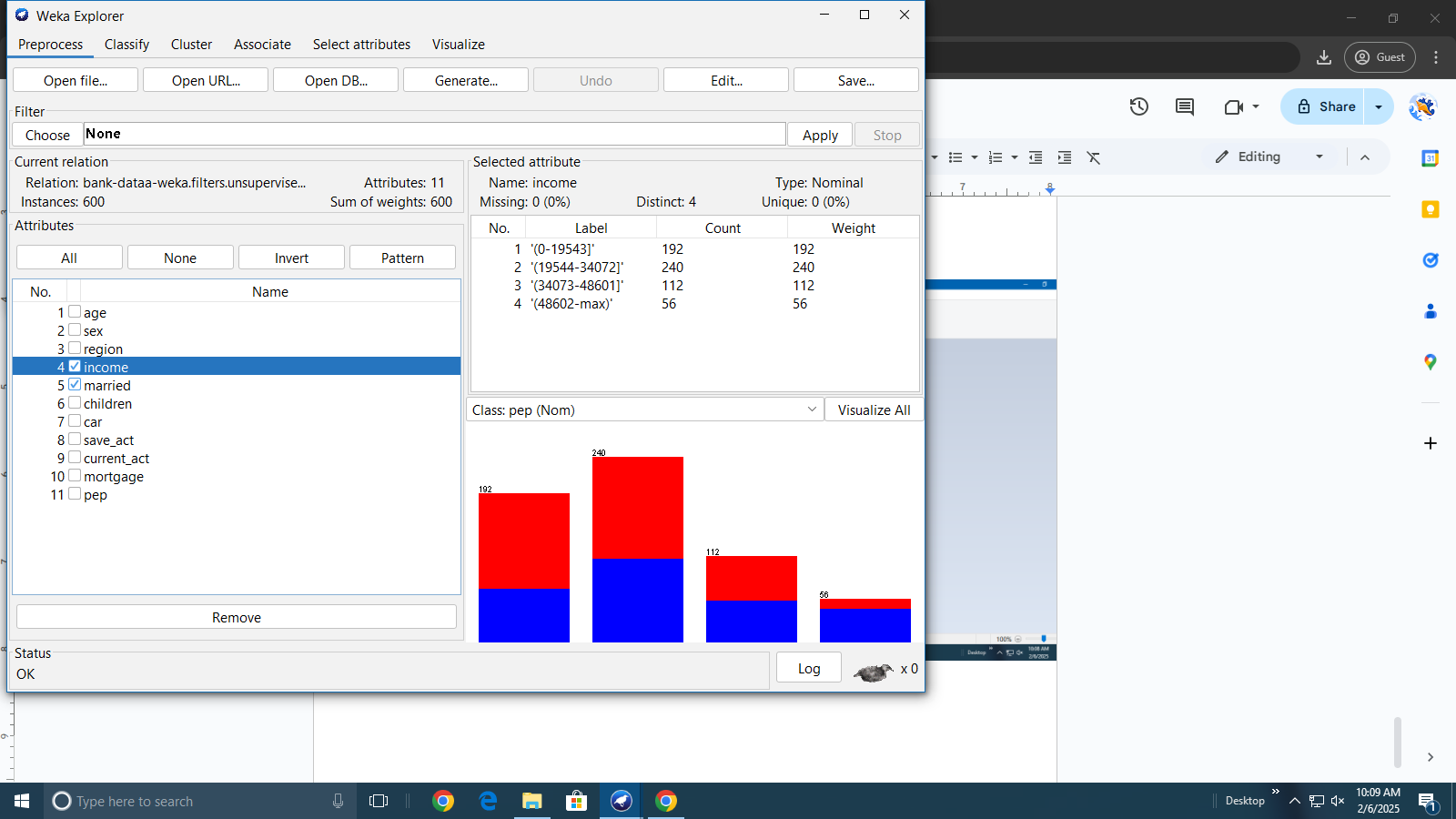


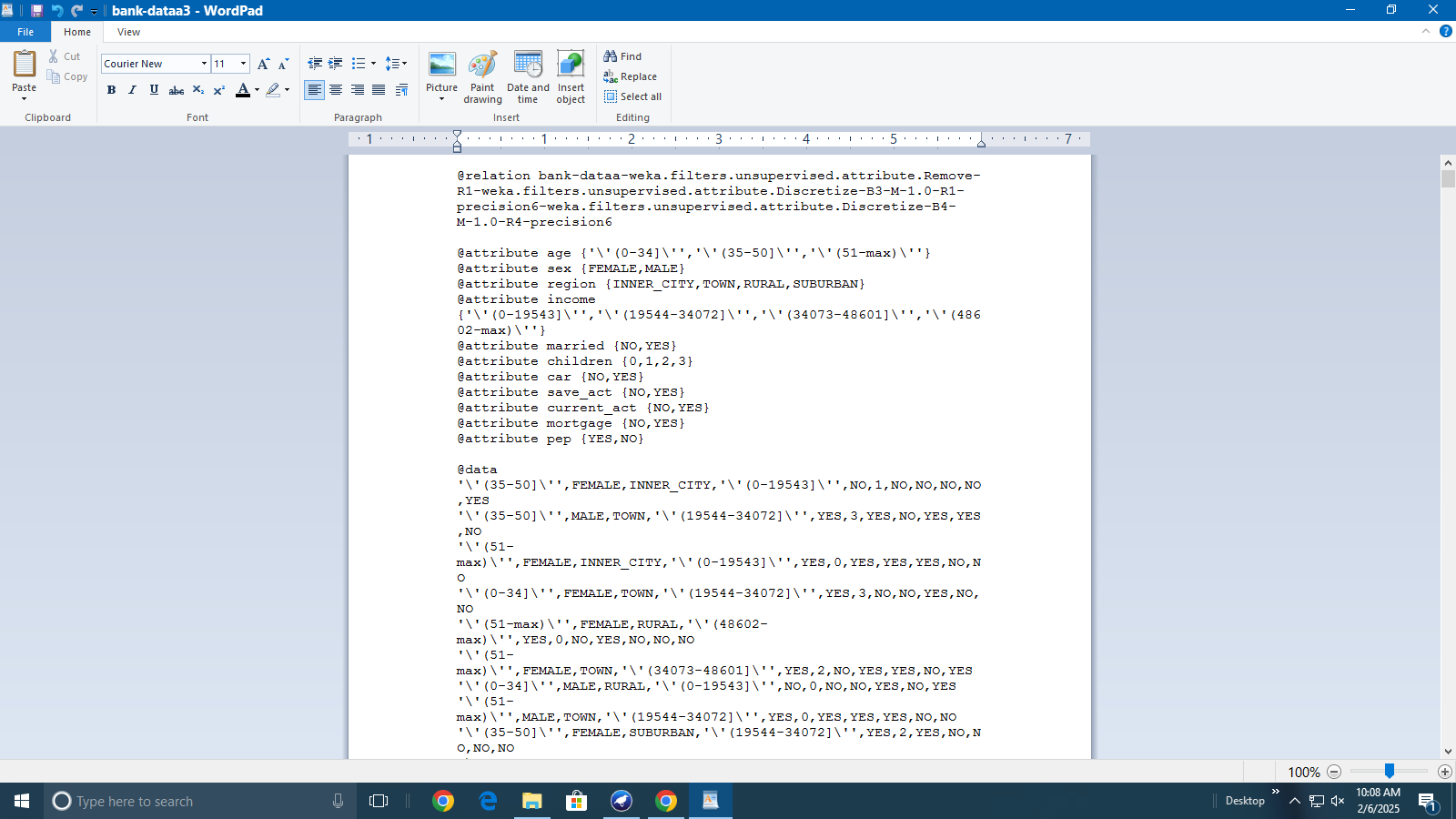
Assigning new labels to the bins New labels assigned to ‘age’ attribute visible in WEKA



Discretizing income attribute in 4 bins after discretization



Assigning new labels to the bins New labels assigned to ‘income’ attribute visible in WEKA



dataset with missing values Replaced missing values