Lab Number: 3

Title

Prepare a numeric type primary dataset and perform discretization on the dataset

Objective

To prepare a purely numeric dataset and perform discretization on numeric attributes using Weka’s built-in unsupervised filters

IDE/Tools Used

Weka 3.8.6

Theory

Primary Dataset: The original, raw, operational/transactional data that comes directly from the source system (OLTP) is known as primary dataset. It contains all details in one big flat table with lots of redundancy, mixed data types, and no optimization for analysis.

Discretization: Discretization is the process of converting continuous (numeric) attributes into discrete (nominal/categorical) intervals or bins. It transforms values like 45, 28, 67 into categories like “(20–40]”, “(40–60]”, “(60–inf]”.

Discretization is important because:

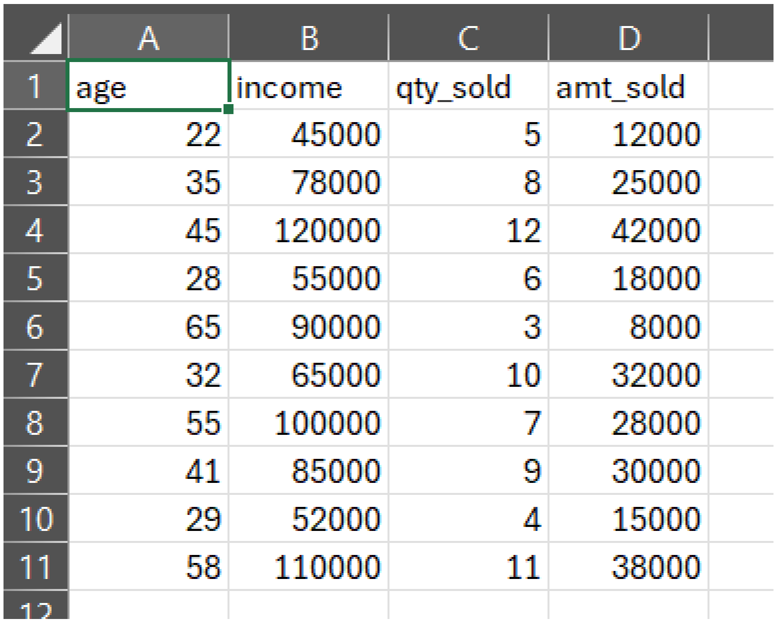
- Many algorithms in Weka work only or better with nominal attributes, like Naive Bayes (nominal version), Apriori, Decision Tables, etc
- Handles skewed distributions better than raw numeric values
- Improves interpretability of rules and decision trees
- Reduces sensitivity to outliers
- Required for association rule mining (Apriori, FPGrowth)

Implementation

The following steps were done inorder to perform discretization on dataset:

1. Generating the dataset

A purely numeric primary dataset was generated containing the sales data.

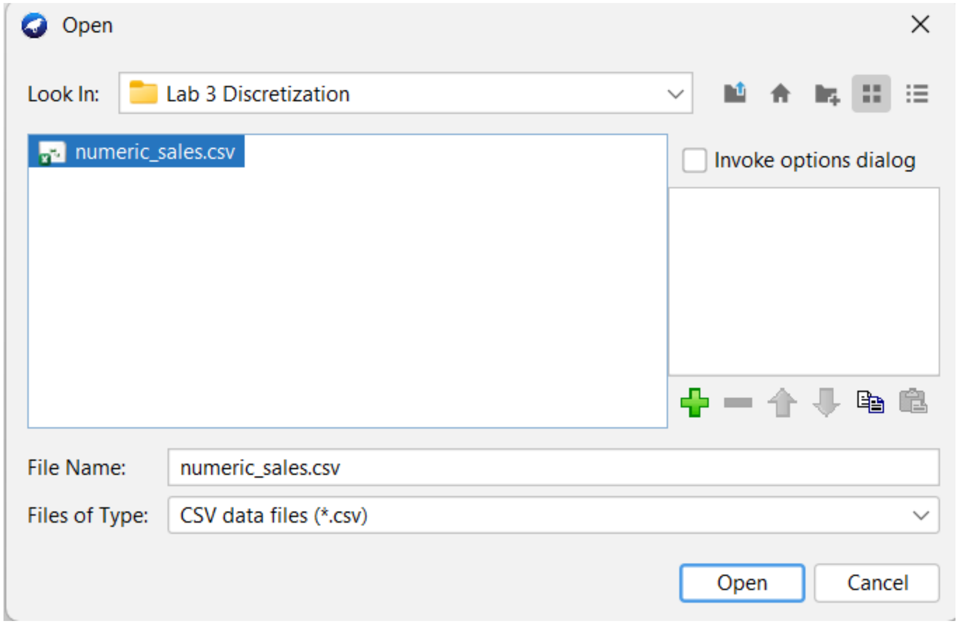


	A	B	C	D
1	age	income	qty_sold	amt_sold
2	22	45000	5	12000
3	35	78000	8	25000
4	45	120000	12	42000
5	28	55000	6	18000
6	65	90000	3	8000
7	32	65000	10	32000
8	55	100000	7	28000
9	41	85000	9	30000
10	29	52000	4	15000
11	58	110000	11	38000
12				

Figure 1: Visualization of the dataset

2. Conversion to .arff format

Using WEKA, and ARFF viewer, the dataset (in .csv format), was opened and saved as in .arff format after confirming all values are numeric.



Look In: Lab 3 Discretization

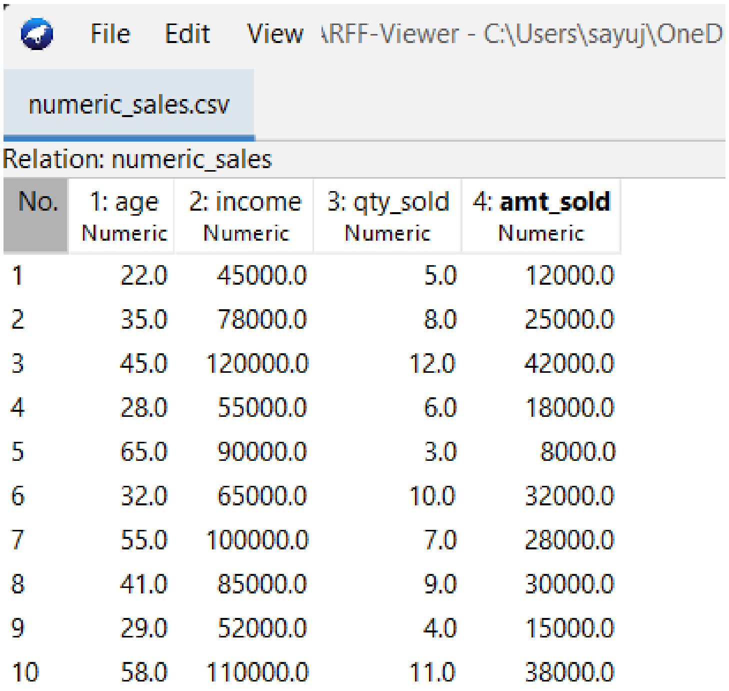
numeric_sales.csv

□ Invoke options dialog

File Name: numeric_sales.csv

Files of Type: CSV data files (*.csv)

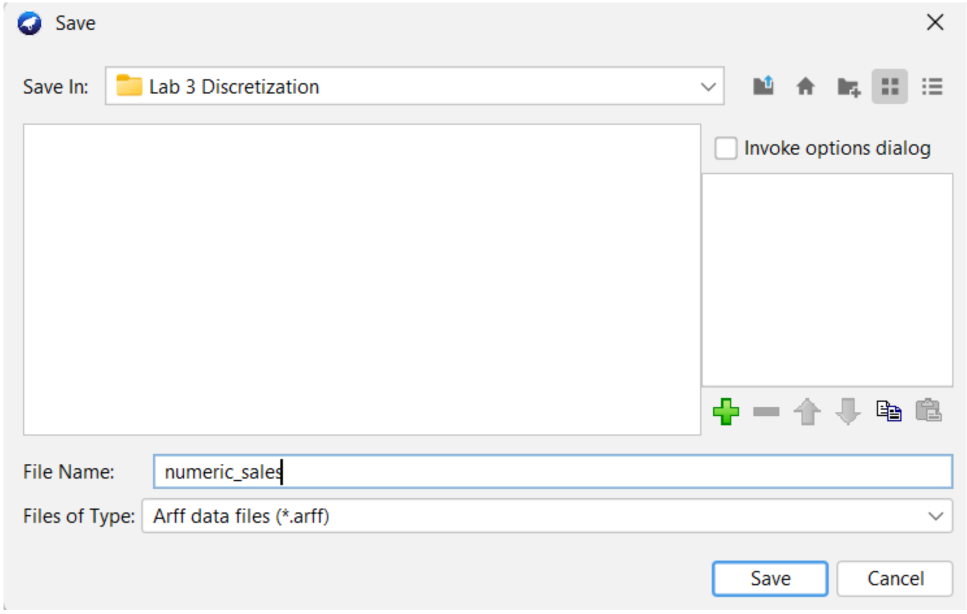
Figure 2: Opening the csv dataset in ARFF Viewer



Relation: numeric_sales

No.	1: age Numeric	2: income Numeric	3: qty_sold Numeric	4: **amt_sold** Numeric
1	22.0	45000.0	5.0	12000.0
2	35.0	78000.0	8.0	25000.0
3	45.0	120000.0	12.0	42000.0
4	28.0	55000.0	6.0	18000.0
5	65.0	90000.0	3.0	8000.0
6	32.0	65000.0	10.0	32000.0
7	55.0	100000.0	7.0	28000.0
8	41.0	85000.0	9.0	30000.0
9	29.0	52000.0	4.0	15000.0
10	58.0	110000.0	11.0	38000.0

Figure 3: Checking for data types



Save In: Lab 3 Discretization

□ Invoke options dialog

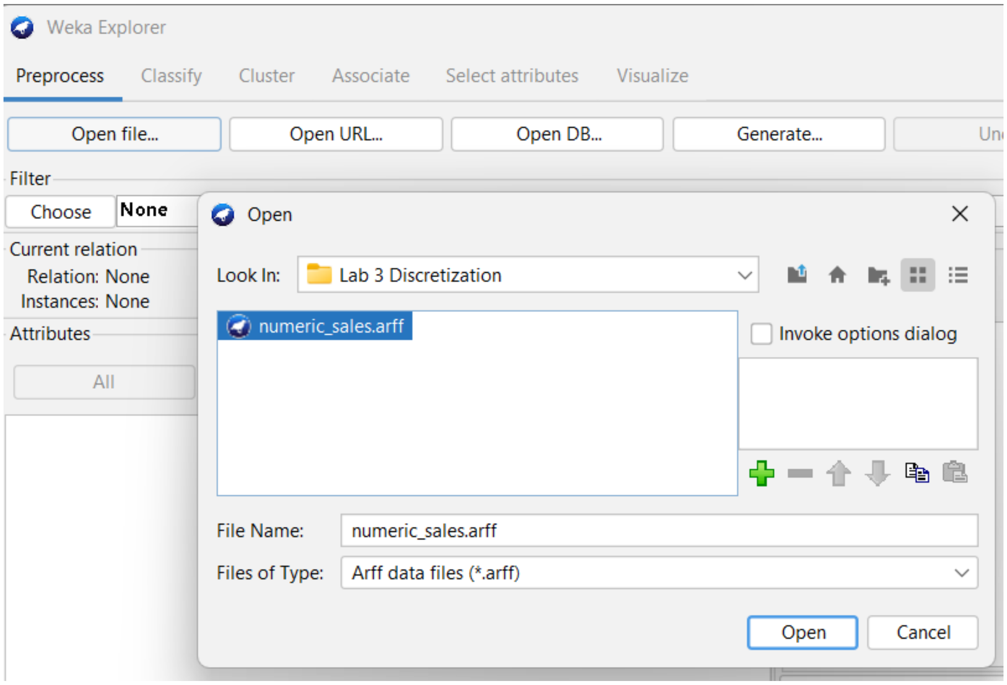
File Name: numeric_sales

Files of Type: Arff data files (*.arff)

Figure 4: Saving the dataset in .arff format

3. Open the dataset in Weka

Now open the recently created dataset in Weka explorer



Look In: Lab 3 Discretization

File Name: numeric_sales.arff

Files of Type: Arff data files (*.arff)

Figure 5: Open the dataset in Weka Explorer

4. Choose the "discretize" filter

Now open filter an choose unsupervised.attribute.discretize filter

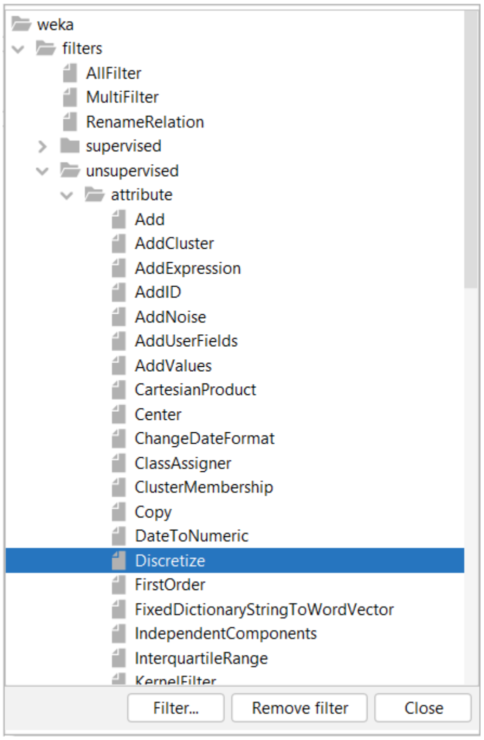
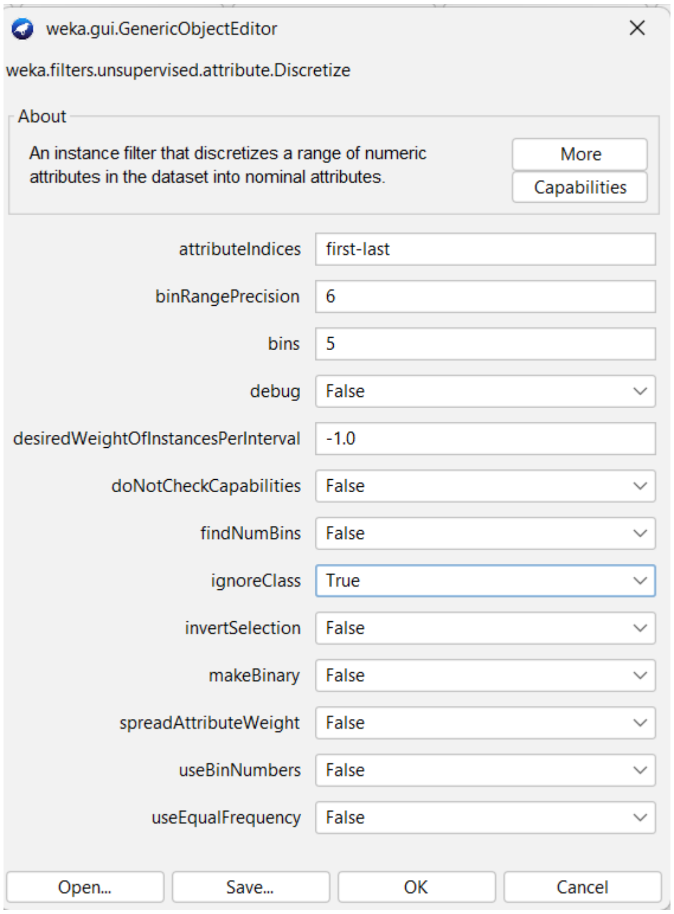


Figure 6: Discretize Filter

5. Choose the appropriate settings

Settings

- bins = 5
- useEqualFreq = False
- attributeIndices = first-last
- ignoreClass = True



weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.Discretize

About

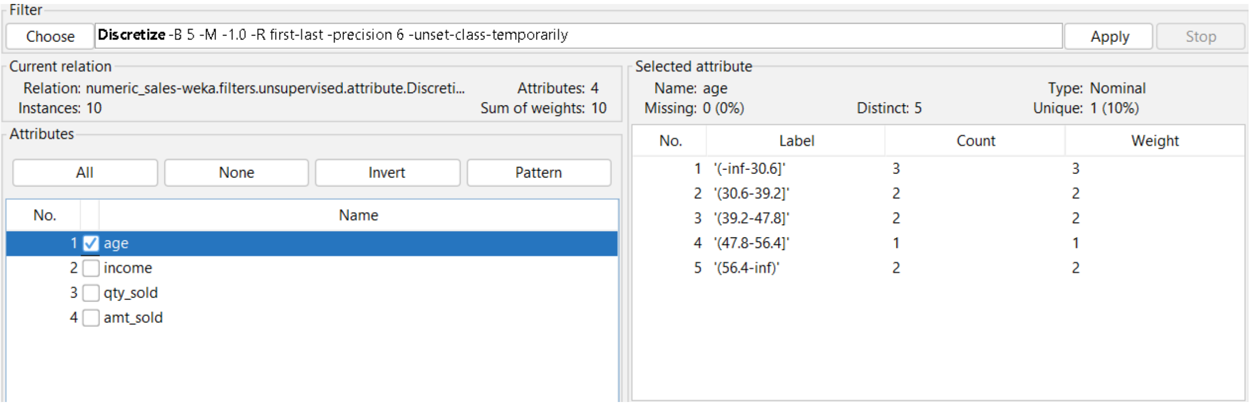
An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.

attributeIndices	first-last
binRangePrecision	6
bins	5
debug	False
desiredWeightOfInstancesPerInterval	-1.0
doNotCheckCapabilities	False
findNumBins	False
ignoreClass	True
invertSelection	False
makeBinary	False
spreadAttributeWeight	False
useBinNumbers	False
useEqualFrequency	False

Figure 7: Settings for discretize

6. Visualize the result

Each attribute is now divided into 5 equal-width intervals.



Discretize -B 5 -M -1.0 -R first-last -precision 6 -unset-class-temporarily

Current relation

Relation: numeric_sales-weka.filters.unsupervised.attribute.Discreti... Attributes: 4

Instances: 10 Sum of weights: 10

Attributes

No.	Name
1	☑ age
2	□ income
3	□ qty_sold
4	□ amt_sold

Selected attribute

Name: age Type: Nominal

Missing: 0 (0%) Distinct: 5 Unique: 1 (10%)

No.	Label	Count	Weight
1	'(-inf-30.6]'	3	3
2	'(30.6-39.2]'	2	2
3	'(39.2-47.8]'	2	2
4	'(47.8-56.4]'	1	1
5	'(56.4-inf)'	2	2

Figure 8: Final label of age class

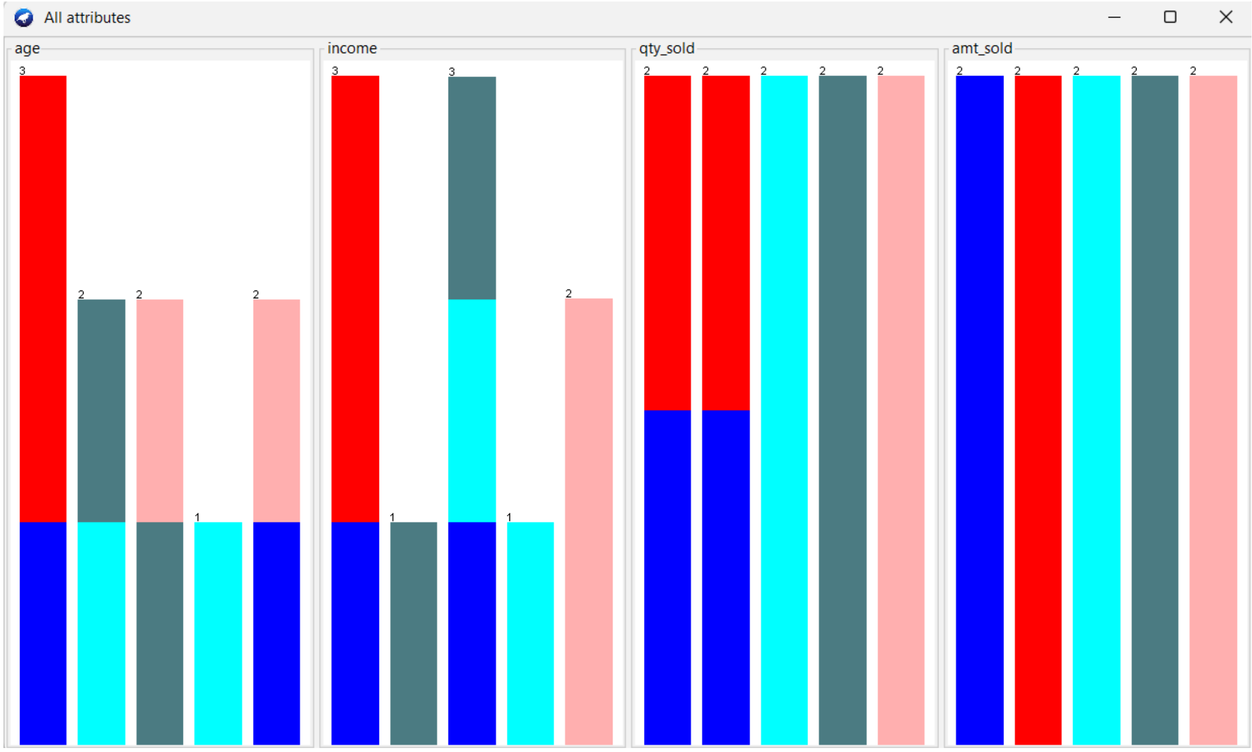
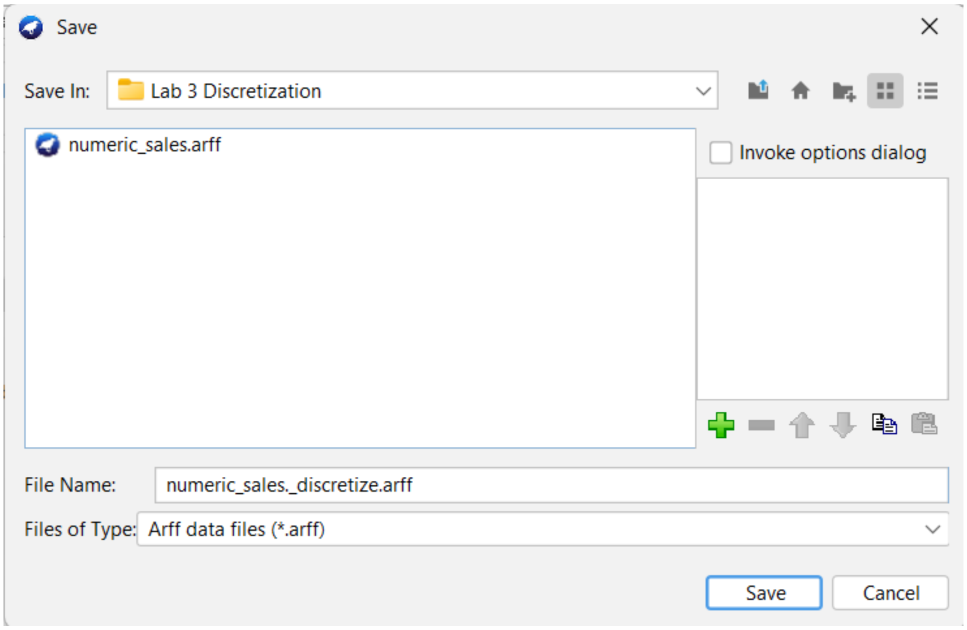


Figure 9: Visualization of dataset after discretization

7. Save the data in .arff format

Finally save the data in .arff format.



Save In: Lab 3 Discretization

numeric_sales.arff

□ Invoke options dialog

File Name: numeric_sales._discretize.arff

Files of Type: Arff data files (*.arff)

Figure 10: Saving data in .arff format

Discussion

This lab focused on converting continuous numeric attributes into discrete intervals using WEKA's discretization filter. The creation of a purely numeric dataset emphasized the need for consistent data types before discretization. Applying the unsupervised.attribute.Discretize filter with five equal-width bins demonstrated how numeric ranges can be grouped into meaningful intervals.

The transformation allowed attributes to shift from raw numeric values to categorical bins, which are often required for algorithms such as Apriori or Decision Trees. The visualization revealed how each attribute's distribution was segmented, improving interpretability and preparing the dataset for algorithms that do not accept continuous data. The exercise reinforced how discretization can reduce sensitivity to outliers and simplify downstream modeling tasks.

Conclusion

This lab successfully demonstrated the discretization process on a numeric dataset using WEKA. By generating a primary numeric dataset, converting it to ARFF format, and applying the Discretize filter, continuous values were transformed into five clear, consistent intervals. This process showed how discretization supports better interpretability and compatibility with algorithms requiring categorical attributes. The lab emphasized discretization as a crucial preprocessing technique in many data mining workflows.