جامعة الأمير سلطان
PRINCE SULTAN UNIVERSITY

CCIS كلية علوم الحاسب والمعلومات
COLLEGE OF COMPUTER & INFORMATION SCIENCES

Assignment 2

Prepared by

Hussam Aldossary

Supervised by

Dr. Esam Mohammed Othman

Table of Contents

1. INTRODUCTION 3
   1.1. PURPOSE 3
   1.2. SCOPE 3

2. DATASET ANALYSIS AND PREPARATION 4
   2.1. HCV Data: Clustering Analysis 4
      2.1.1. Description 4
      2.1.2. features 4
      2.1.3. WEKA Analysis 5
      2.1.4. Results 13
   2.2. MathE Data: Clustering Analysis 15
      2.2.1. Description 15
      2.2.2. features 15
      2.2.3. WEKA Analysis 16
      2.2.4. Results 29
3. REFERENCES .....

1 INTRODUCTION

1.1 Purpose

The purpose of this assignment is to apply practical data mining techniques—specifically **Clustering**—using the WEKA (Waikato Environment for Knowledge Analysis) toolset. Unlike classification, which predicts predefined labels, clustering explores the data to find natural groupings and hidden structures. Through hands-on analysis of two distinct datasets (Health/Medicine and Education), this project aims to:

- Develop proficiency in unsupervised learning workflows, from dataset acquisition to cluster interpretation.
- Compare different clustering algorithms (e.g., SimpleKMeans, EM) to understand how they group data.
- Evaluate clustering performance using "Classes to Clusters" evaluation to determine how well the discovered clusters map to known categories.
- Demonstrate the utility of WEKA for exploratory data analysis and pattern recognition.

This report documents the methodology, analysis, and findings for each dataset, serving as a structured record of the analytical process and its outcomes.

1.2 SCOPE

The scope of this project is defined by the selection of two specific datasets and the use of the WEKA software for a Clustering task.

- **Datasets:**
    - **HCV Data:** A multivariate dataset concerning Hepatitis C diagnosis.
    - **MathE Data:** A dataset concerning student performance in higher education mathematics.
- **Tool:** All preprocessing and modeling tasks are performed within the WEKA Explorer environment.
- **Process:**
    - **Dataset Analysis:** Understanding the domain and attributes.
    - **Data Preparation:** Handling missing values and converting data to ARFF format.
    - **Model Development:** Applying clustering algorithms (SimpleKMeans, EM).
    - **Evaluation:** Using the "Classes to Clusters" mode to compare clusters against ground truth labels (Accuracy/Error rates).

Deliverable: The final output is this comprehensive report, which includes dataset descriptions, methodological steps, analysis of results supported by WEKA screenshots, and a comparative discussion of findings.

Out of Scope:

- Deployment of models in a production environment.
- Extensive feature engineering beyond basic WEKA filters.
- Comparative analysis with other data mining tools (e.g., R, Python's scikit-learn).
- In-depth theoretical derivation of the algorithms used.

2 DATASET ANALYSIS AND PREPARATION

This section provides a detailed overview of the two datasets selected from the UCI Machine Learning Repository for analysis. Each dataset represents a clustering machine learning task. For each dataset, a description of the problem domain, a summary of its features, and notes on data preparation for WEKA are provided.

2.1 Dataset 1: Binary Classification – Rice (Cammeo and Osmancik)

2.1.1 Description

The HCV (Hepatitis C Virus) dataset contains laboratory values from blood donors and Hepatitis C patients. The domain is **Health and Medicine**. The goal of this analysis is to use clustering to see if patients naturally group into their diagnostic categories based on their blood work data alone, without being explicitly told the diagnosis.

The dataset includes values for blood donors and patients in various stages of Hepatitis C (Hepatitis, Fibrosis, Cirrhosis). Successfully clustering these patients could indicate that these blood markers are strong indicators of disease progression.

2.1.2 Features

The dataset contains 615 instances and 12 features. Most features are numerical laboratory values.

Feature Name	Data Type	Description
Category (Target)	Nominal	Diagnosis (Blood Donor, Suspect, Hepatitis, Fibrosis, Cirrhosis).
Age	Integer	Age of the patient.
Sex	Binary	Gender (m/f).
ALB	Continuous	Albumin Blood Test.
ALP	Continuous	Alkaline Phosphatase.
ALT	Continuous	Alanine Transaminase.

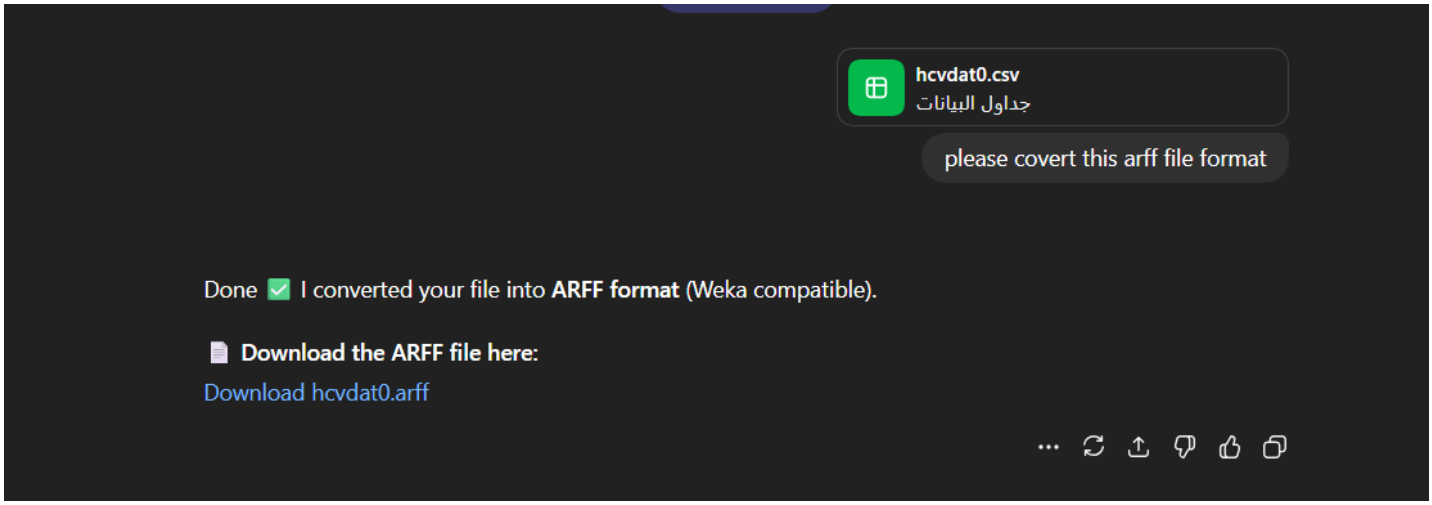
AST	Continuous	Aspartate Transaminase.
BIL	Continuous	Bilirubin.

Key Dataset Facts:

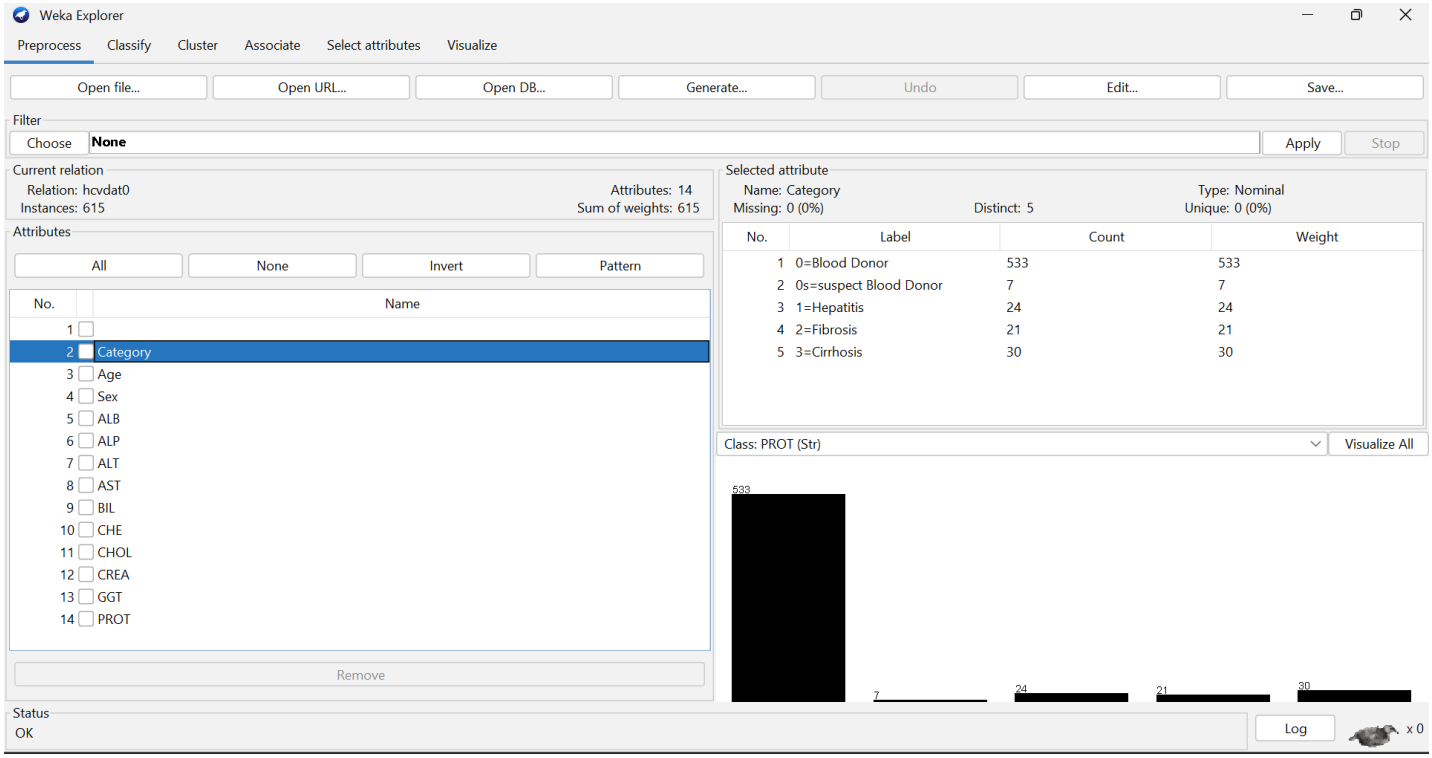
- **Number of Instances:** 615
- **Target:** Category (5 distinct values).
- **Missing Values:** Yes (in ALB, ALP, ALT, CHOL, PROT).
- **Source File:** hcvdat0.csv

2.1.3 WEKA Analysis

Using chatGPT to convert csv to arff

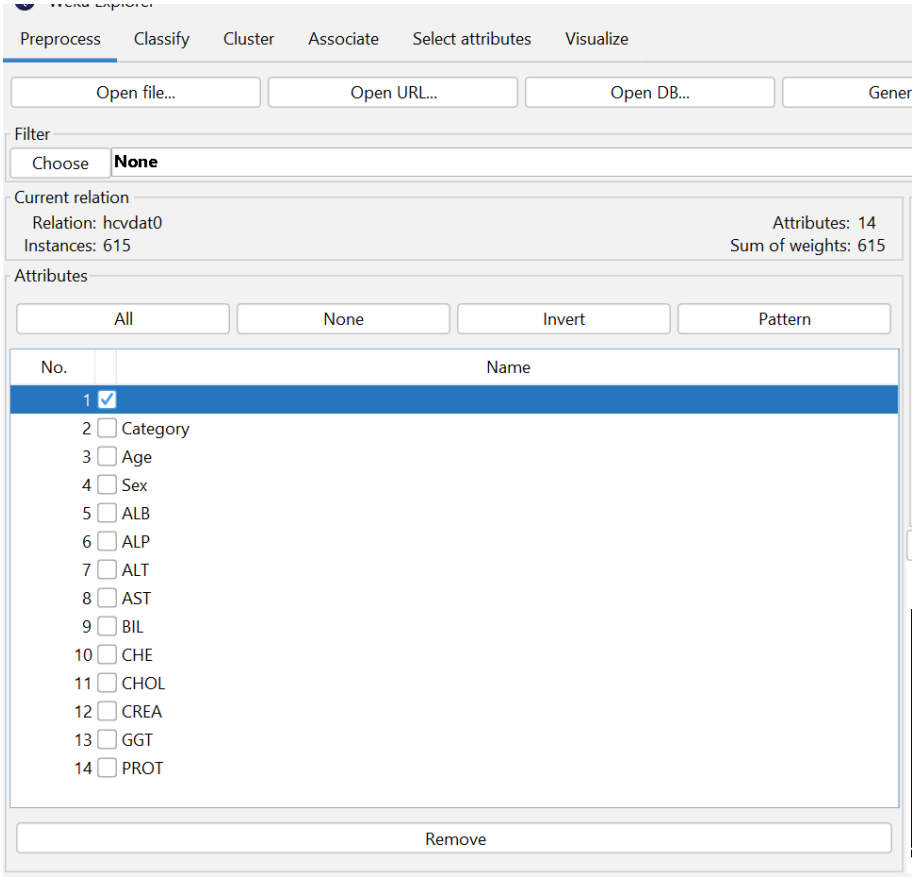


Loading data



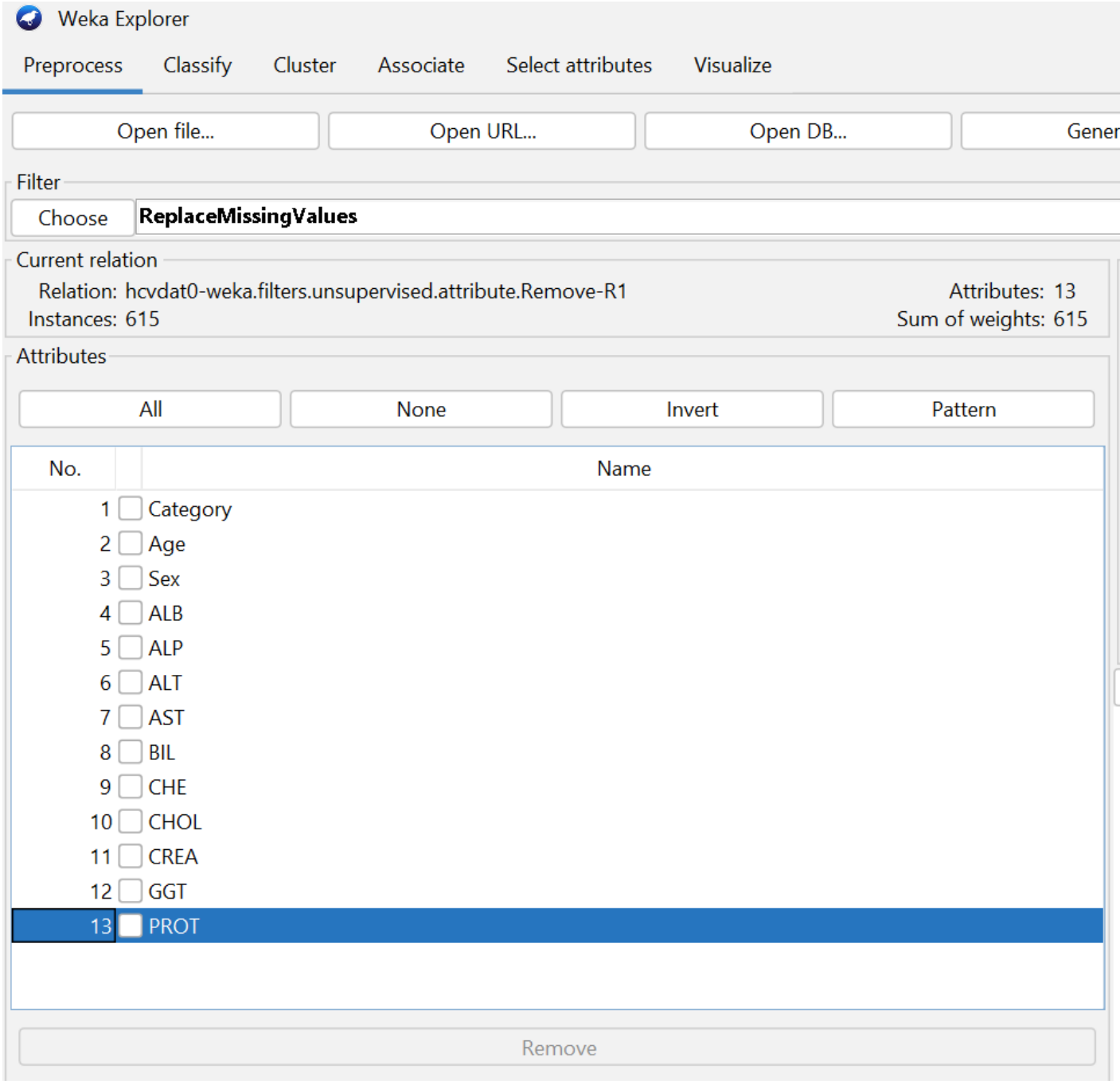
Preprocessing

- **Remove ID**

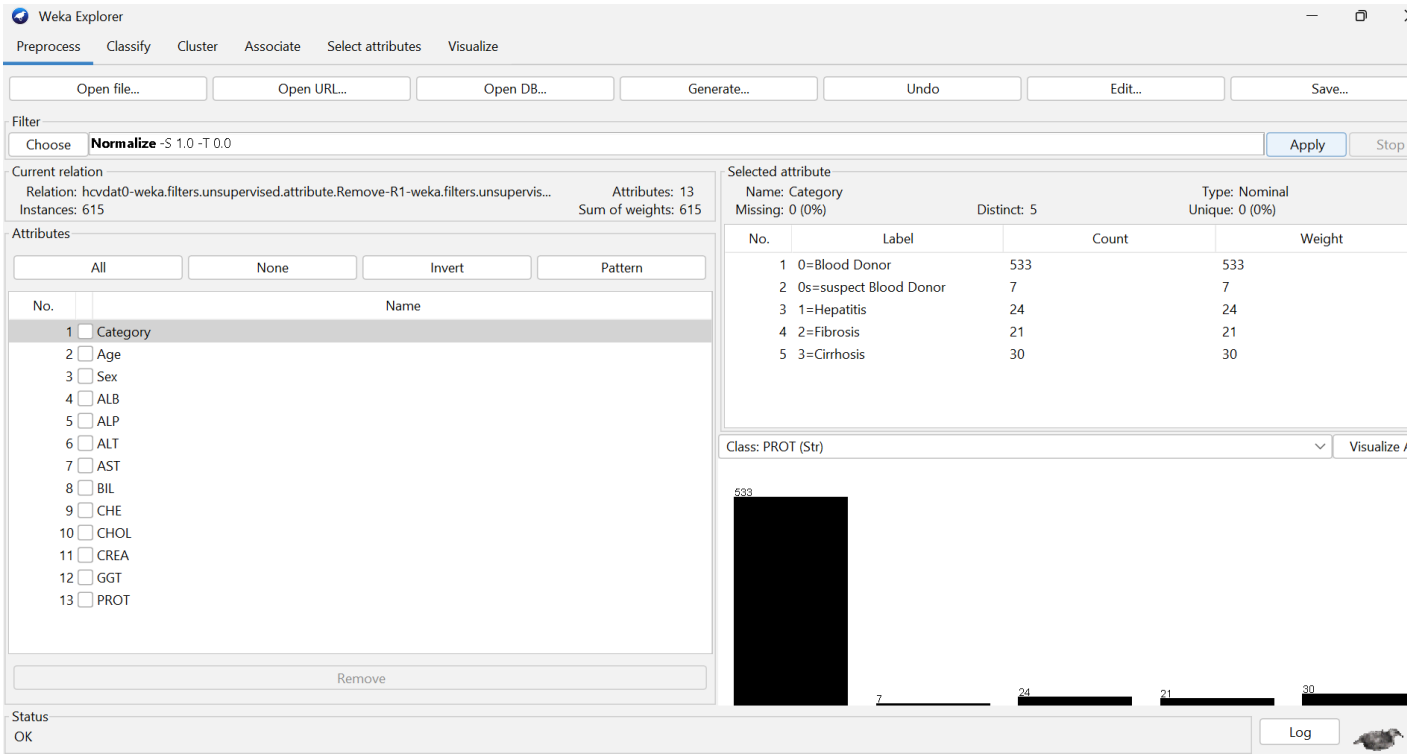


- **Handle Missing Values by ReplaceMissingValues filter**

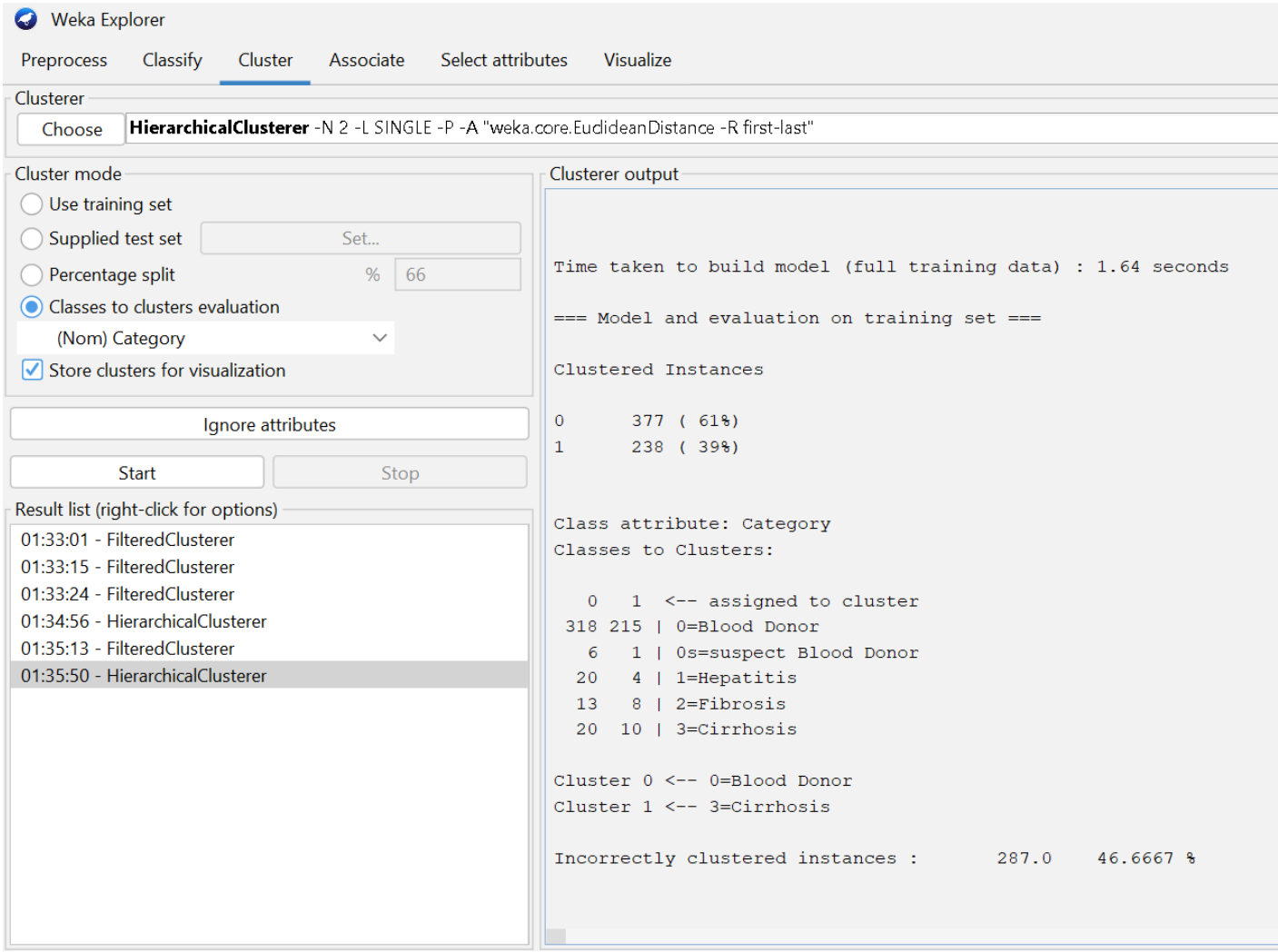
  Navigate to: filters > unsupervised > attribute > ReplaceMissingValues.



- **Normalize (Scale Numbers) by Normalize filter**
  Navigate to: filters > unsupervised > attribute > Normalize.



FilteredClusterer (Classes to clusters evaluation based on category)



2.1.4 results

Table 1: Clustering Performance Summary

Metric	Value
Number of Clusters	2
Time to Build Model	1.64 seconds
Incorrectly Clustered Instances	287 (46.67%)
Clustering Accuracy	53.33%
Cluster 0 Assignment	0=Blood Donor
Cluster 1 Assignment	3=Cirrhosis

Table 2: Classes to Clusters Confusion Matrix

Class	Assigned to Cluster 0	Assigned to Cluster 1	Total Instances
0=Blood Donor	318	215	533
0s=suspect Blood Donor	6	1	7
1=Hepatitis	20	4	24
2=Fibrosis	13	8	21
3=Cirrhosis	20	10	30

Observations:

- **The model achieved an accuracy of roughly 53%, which indicates that separating patients into just two groups (Clusters) is not sufficient to distinguish between the 5 different disease stages.**
- **Cluster 0 captured the majority of Blood Donors (318 instances), but a significant number of donors (215) were also placed in Cluster 1, showing that healthy patients have varied blood profiles.**
- **Cluster 1 was assigned to Cirrhosis, but it actually contains more Blood Donors (215) than Cirrhosis patients (10). This suggests the algorithm struggled to find a distinct pattern for Cirrhosis when limited to only 2 clusters.**

2.2 Dataset 2: Multi-class Classification – Car Evaluation

2.2.1 Description

The MathE dataset comes from a mathematical platform for higher education. The domain is **Education**. It contains 9,546 answers to questions across various topics (Linear Algebra, Calculus, etc.).
The goal of this analysis is to cluster student performance records to identify patterns in how students answer questions. For example, do students cluster based on their Country, or does the Difficulty Level of the question drive the grouping?

2.2.2 Features

The dataset contains mixed types (categorical and numerical) with 9,546 instances..

Feature Name	Data Type	Description	Feature Name
Student Country	Nominal	The country of origin of the student.	**Student Country**
Question ID	Integer	Unique ID for the question.	**Question ID**
Type of Answer	Binary	**Target:** 1 = Correct, 0 = Incorrect.	**Type of Answer**
Question Level	Nominal	Difficulty (Basic, Advanced).	**Question Level**
Topic	Nominal	Main math topic (e.g., Differentiation, Linear Algebra).	**Topic**
Subtopic	Nominal	Specific sub-area.	**Subtopic**

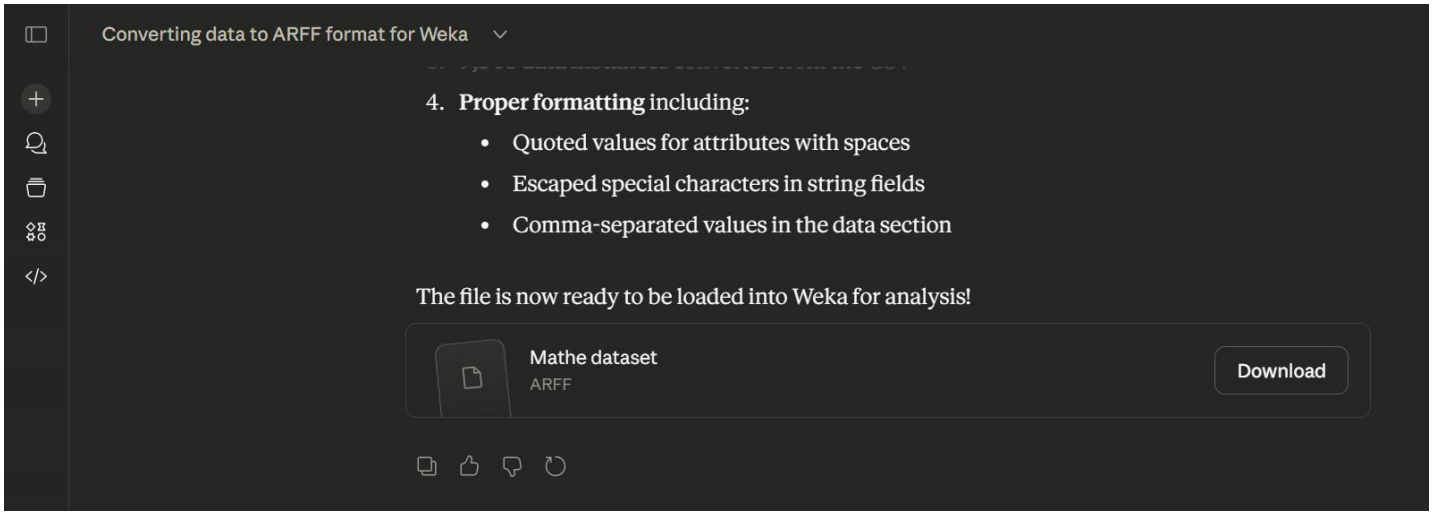
Keywords	Nominal	Tags associated with the question.	**Keywords**

Key Dataset Facts:

- **Instances:** 9,546
- **Missing Values:** None.
- **Features:** Heavily categorical/nominal.
- **Source File:** MathE dataset (4).csv

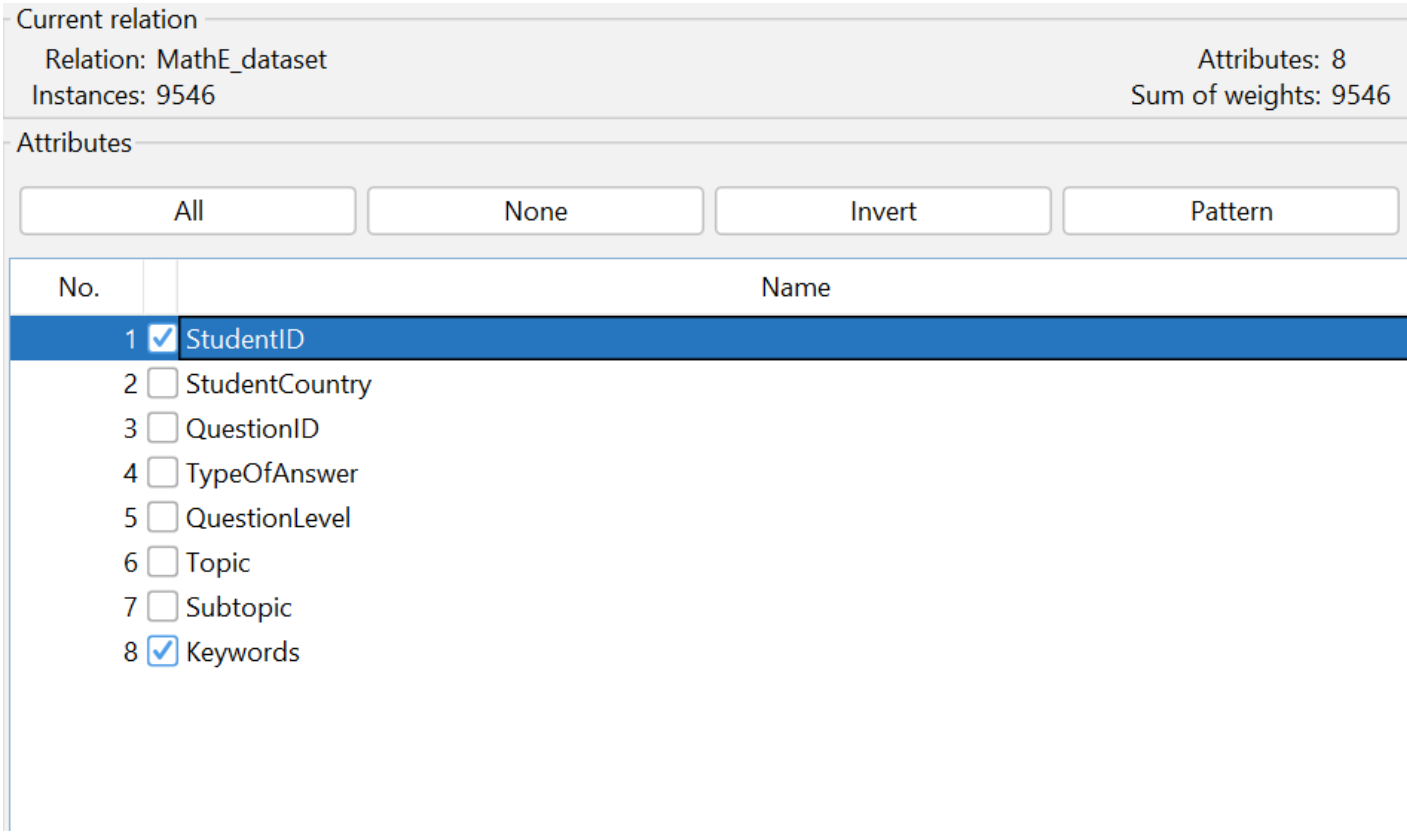
2.2.3 WEKA Analysis

Converting to arff using claude

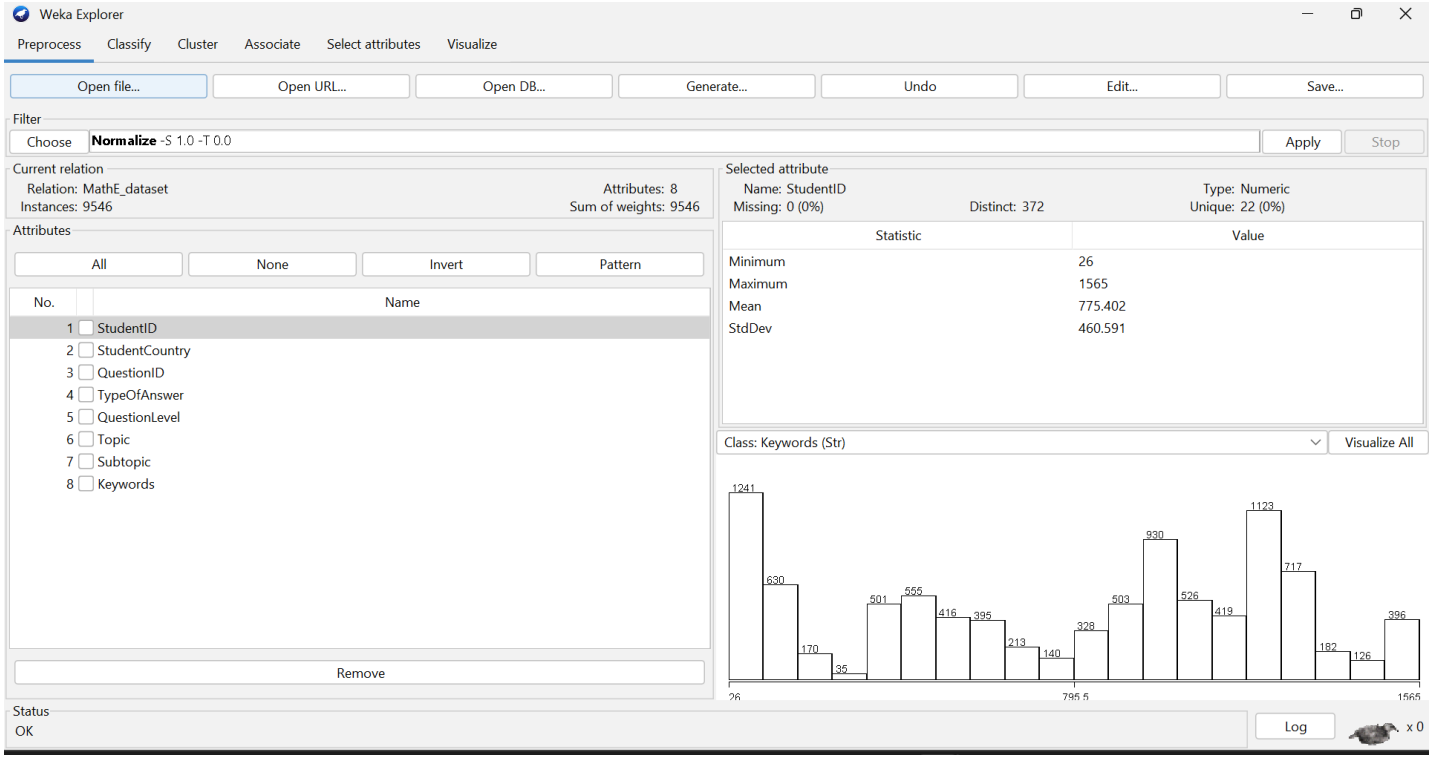


Preprocessing:

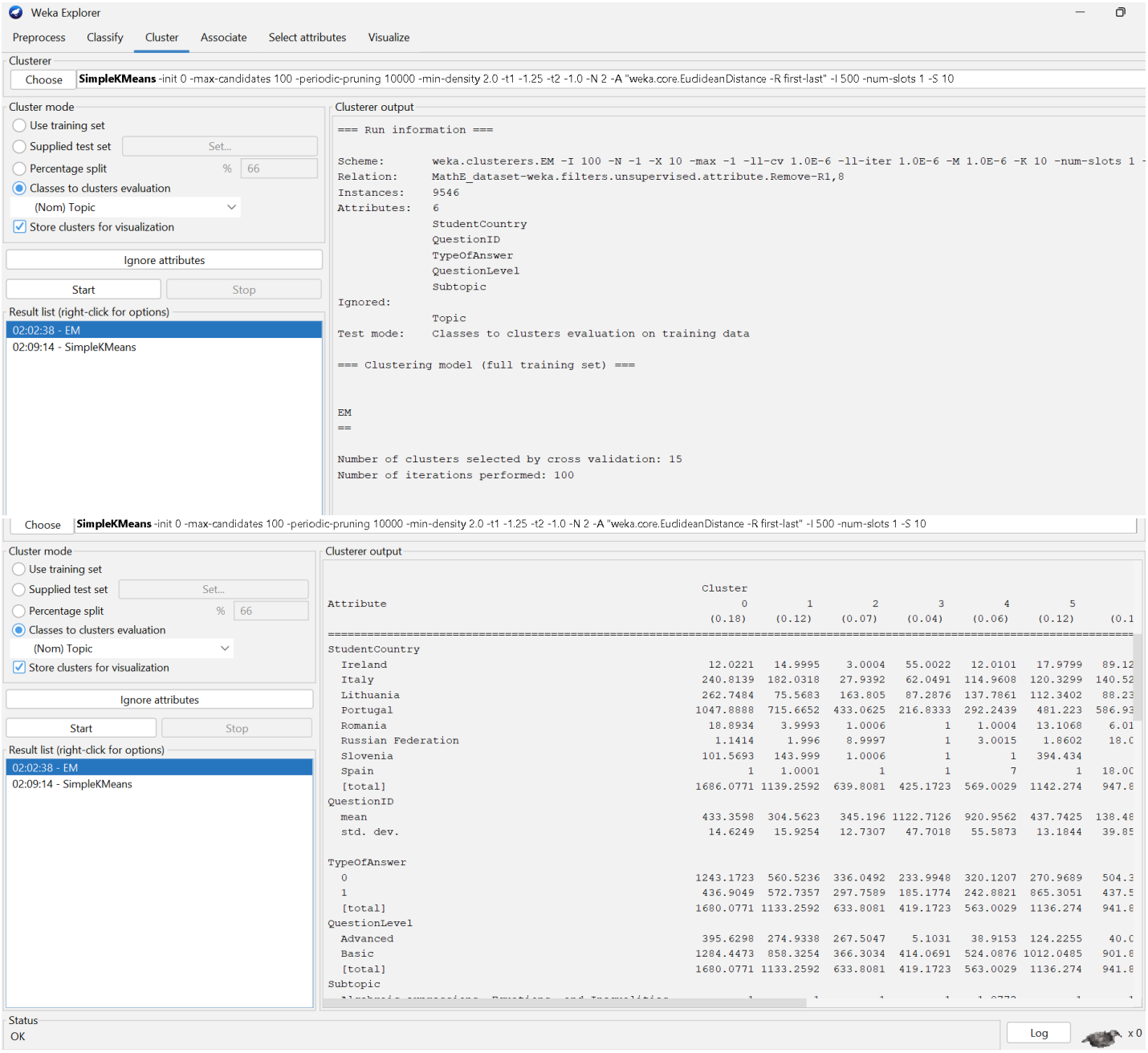
- **Removing studentID +keywords (because textbased)**



Loading data



em(Classes to clusters evaluation based on Topic)



Attribute	Cluster 0 (0.18)	1 (0.12)	2 (0.07)	3 (0.04)	4 (0.06)	5 (0.12)	(0.1
StudentCountry							
Ireland	12.0221	14.9995	3.0004	55.0022	12.0101	17.9799	89.12
Italy	240.8139	182.0318	27.9392	62.0491	114.9608	120.3299	140.52
Lithuania	262.7484	75.5683	163.805	87.2876	137.7861	112.3402	88.23
Portugal	1047.8888	715.6652	433.0625	216.8333	292.2439	481.223	586.93
Romania	18.8934	3.9993	1.0006	1	1.0004	13.1068	6.01
Russian Federation	1.1414	1.996	8.9997	1	3.0015	1.8602	18.0
Slovenia	101.5693	143.999	1.0006	1	1	394.434	
Spain	1	1.0001	1	1	7	1	18.00
[total]	1686.0771	1139.2592	639.8081	425.1723	569.0029	1142.274	947.8
QuestionID							
mean	433.3598	304.5623	345.196	1122.7126	920.9562	437.7425	138.48
std. dev.	14.6249	15.9254	12.7307	47.7018	55.5873	13.1844	39.85
TypeOfAnswer							
0	1243.1723	560.5236	336.0492	233.9948	320.1207	270.9689	504.3
1	436.9049	572.7357	297.7589	185.1774	242.8821	865.3051	437.5
[total]	1680.0771	1133.2592	633.8081	419.1723	563.0029	1136.274	941.8
QuestionLevel							
Advanced	395.6298	274.9338	267.5047	5.1031	38.9153	124.2255	40.0
Basic	1284.4473	858.3254	366.3034	414.0691	524.0876	1012.0485	901.8
[total]	1680.0771	1133.2592	633.8081	419.1723	563.0029	1136.274	941.8
Subtopic							

Clusterer

Choose **SimpleKMeans** -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Cluster mode

- Use training set
- Supplied test set — Set...
- Percentage split — % 66
- Classes to clusters evaluation
- (Nom) Topic
- Store clusters for visualization

Ignore attributes

Start | Stop

Result list (right-click for options)

02:02:38 - EM
02:09:14 - SimpleKMeans

Clusterer output

Subtopic							
Algebraic expressions, Equations, and Inequalities	1	1	1	1	1.0772	1	1
Analytic Geometry	1	1	1	1	1.0002	1	
Complex Numbers	1	1.2902	555.7098	14.6992	1.3008	1	
Definite Integrals	1	1	1	16	1	1	
Derivatives	1.008	1.0997	1	6.9996	1.0024	1.0007	109.78
Differential Equations	1	1	1	1	61.9999	1	
Domain, Image and Graphics	1	1	1	1.0001	87.0055	1	
Double Integration	1	1	1	2	1.0104	1	
Eigenvalues and Eigenvectors	1	1	1	1	25.9982	1	
Elementary Geometry	1.2634	1.0018	69.8965	1	1	1.2483	
Graph Theory	1.1242	1.0004	6.7587	1.0018	15.9982	1.0445	
Integration Techniques	66.3891	12.9535	1.0824	4.9991	1.0017	1.3025	7.99
Limits and Continuity	1	1	1	1.0153	11.9866	1	
Linear Optimization	1	4.5725	1	1.1148	22.9267	1	1.00
Linear Systems	1	1.0005	1	1.0022	92.9978	1	66.29
Linear Transformations	71.8904	857.5275	1.1687	1	1.0048	111.1542	1.00
Matrices and Determinants	1	1	1	1	1	1	1
Nonlinear Optimization	1	1	1	13.7755	107.2244	1	
Numerical Methods	1	1	1	310.9951	1.0049	1	
Partial Differentiation	1	1.0061	1	10.5699	143.4301	1	16.01
Probability	1.0022	75.91	3.0891	1	1.0324	1.001	1.00
Set Theory	1	1	1	42.9996	1.0004	1	
Statistics	1	1	1	4	1	1	337.76
Vector Spaces	1542.3999	184.897	1.103	1	1	1024.5227	
[total]	1702.0771	1155.2592	655.8081	441.1723	585.0029	1158.274	963.8

Clusterer

Choose **SimpleKMeans** -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Cluster mode

- Use training set
- Supplied test set — Set...
- Percentage split — % 66
- Classes to clusters evaluation
- (Nom) Topic
- Store clusters for visualization

Ignore attributes

Start | Stop

Result list (right-click for options)

02:02:38 - EM
02:09:14 - SimpleKMeans

Clusterer output

```
Time taken to build model (full training data) : 365.62 seconds

=== Model and evaluation on training set ===

Clustered Instances

 0      1649 ( 17%)
 1      1131 ( 12%)
 2       633 (  7%)
 3       414 (  4%)
 4       564 (  6%)
 5      1163 ( 12%)
 6       937 ( 10%)
 7       633 (  7%)
 8       145 (  2%)
 9       202 (  2%)
10       265 (  3%)
11       623 (  7%)
12       844 (  9%)
13       317 (  3%)
14        26 (  0%)


Log likelihood: -9.61335
```

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Clusterer

Choose **SimpleKMeans** -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Cluster mode

- Use training set
- Supplied test set — Set...
- Percentage split — % 66
- Classes to clusters evaluation
- (Nom) Topic
- Store clusters for visualization

Ignore attributes

Start | Stop

Result list (right-click for options)

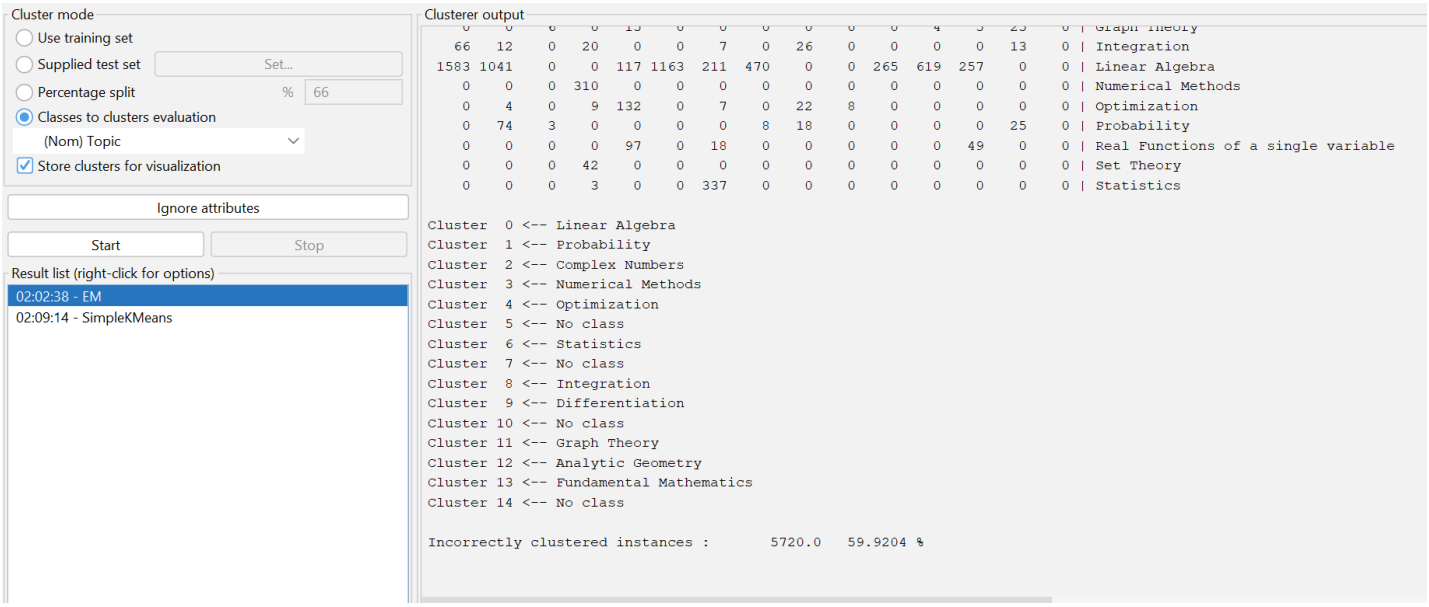
02:02:38 - EM
02:09:14 - SimpleKMeans

Clusterer output

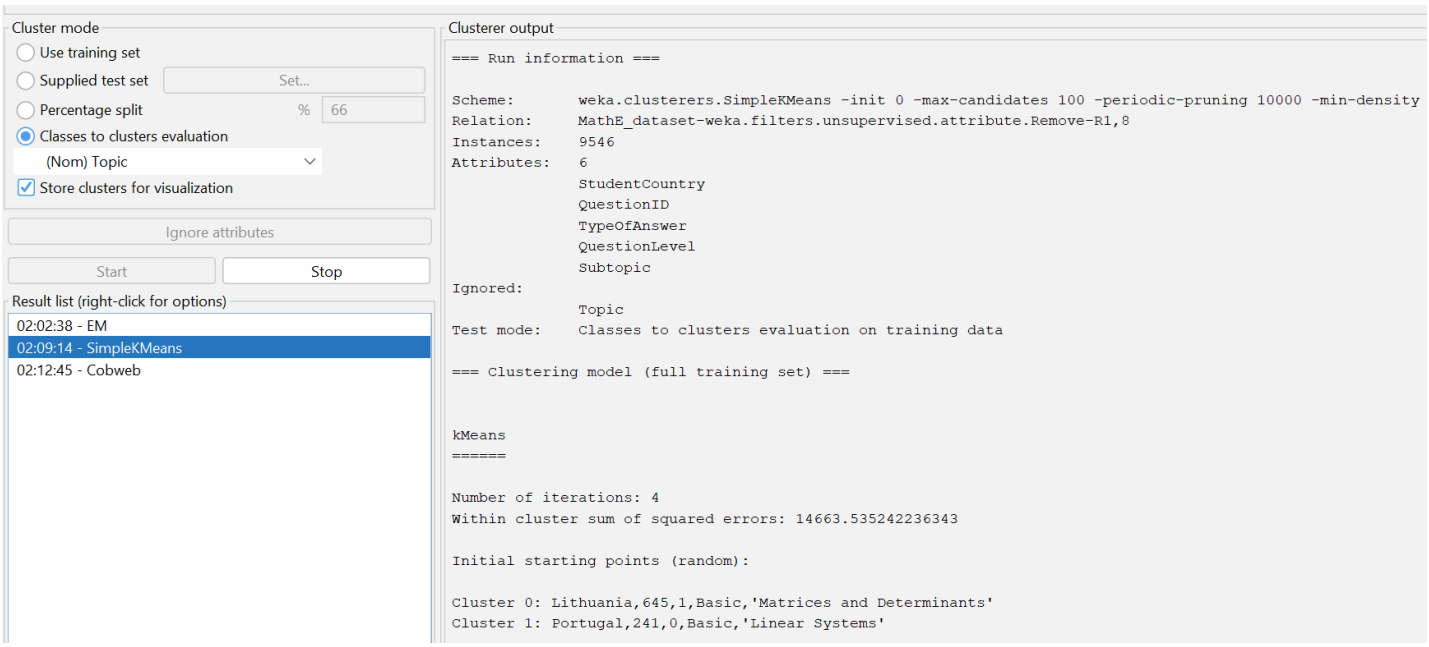
```
Class attribute: Topic
Classes to Clusters:

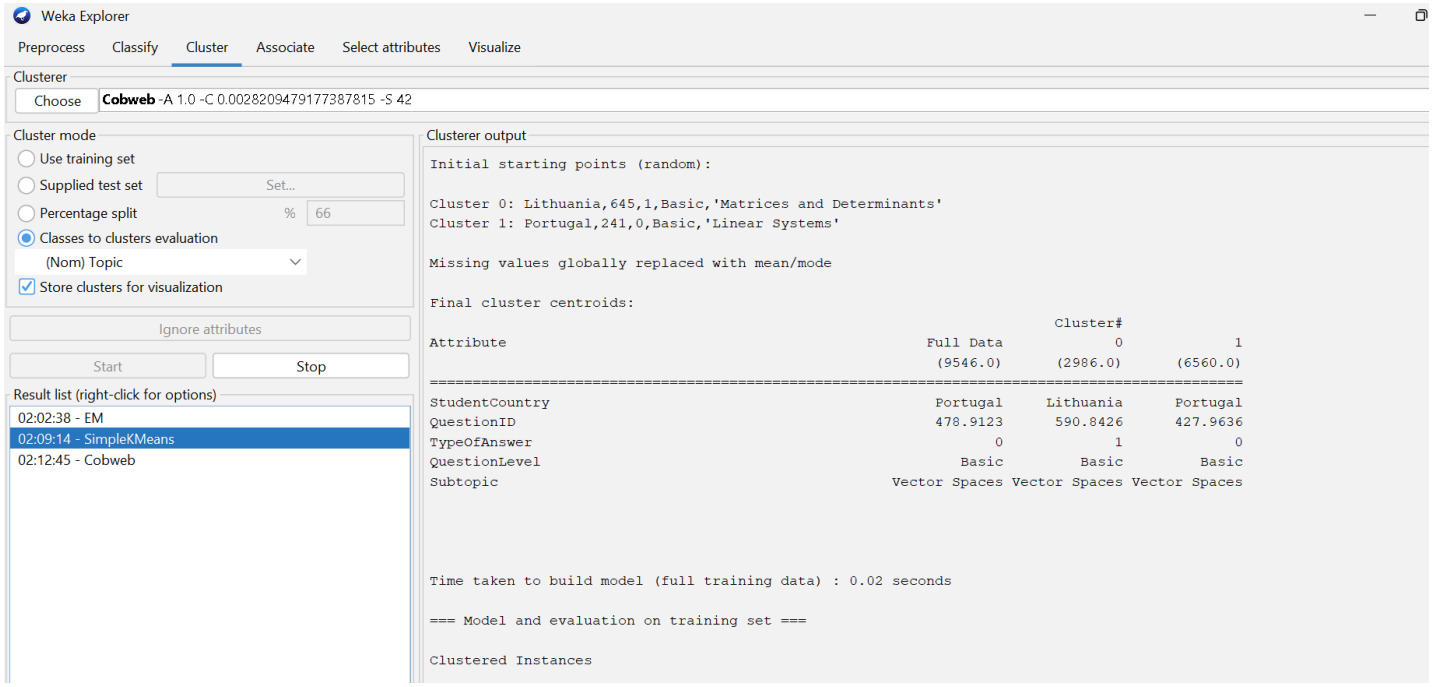
    0    1    2    3    4    5    6    7    8    9   10   11   12   13   14  <-- assigned to cluster
    0    0    0    0    0    0    0    0    0    0    0    0  358    0    0 | Analytic Geometry
    0    0  555   14    0    0    0    0    0    0    0    0    0    0   23 | Complex Numbers
    0    0    0    0   61    0   47    0    0    0    0    0    0    0    0 | Differential Equations
    0    0    0   16  142    0  123    0   79  194    0    0   24    1    0 | Differentiation
    0    0   69    0    0    0  187  155    0    0    0    0  151  253    3 | Fundamental Mathematics
    0    0    6    0   15    0    0    0    0    0    0    4    5   25    0 | Graph Theory
   66   12    0   20    0    0    7    0   26    0    0    0    0   13    0 | Integration
 1583 1041    0    0  117 1163  211  470    0    0  265  619  257    0    0 | Linear Algebra
    0    0    0  310    0    0    0    0    0    0    0    0    0    0    0 | Numerical Methods
    0    4    0    9  132    0    7    0   22    8    0    0    0    0    0 | Optimization
    0   74    3    0    0    0    0    8   18    0    0    0    0   25    0 | Probability
    0    0    0    0   97    0   18    0    0    0    0    0   49    0    0 | Real Functions of a single variable
    0    0    0   42    0    0    0    0    0    0    0    0    0    0    0 | Set Theory
    0    0    0    3    0    0  337    0    0    0    0    0    0    0    0 | Statistics

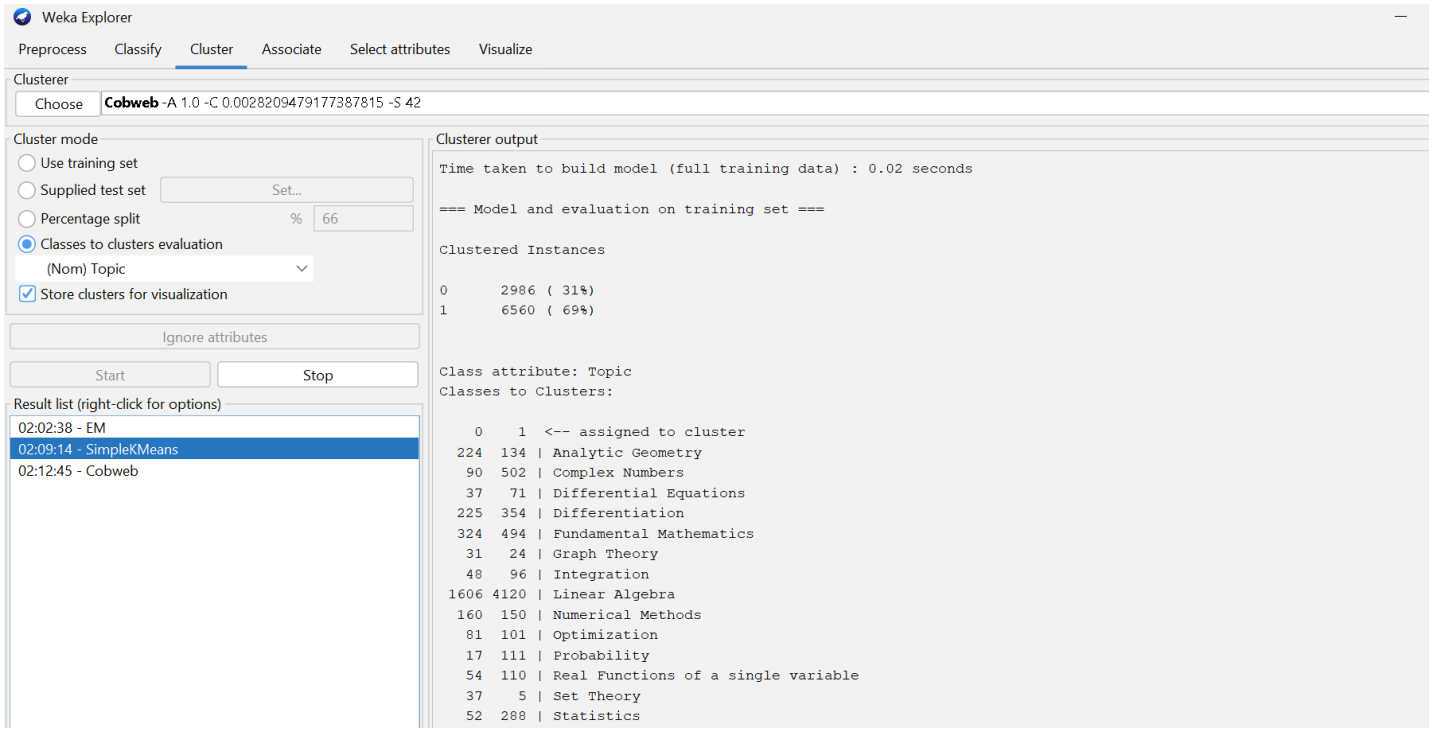
Cluster  0 <-- Linear Algebra
Cluster  1 <-- Probability
Cluster  2 <-- Complex Numbers
Cluster  3 <-- Numerical Methods
Cluster  4 <-- Optimization
Cluster  5 <-- No class
Cluster  6 <-- Statistics
```

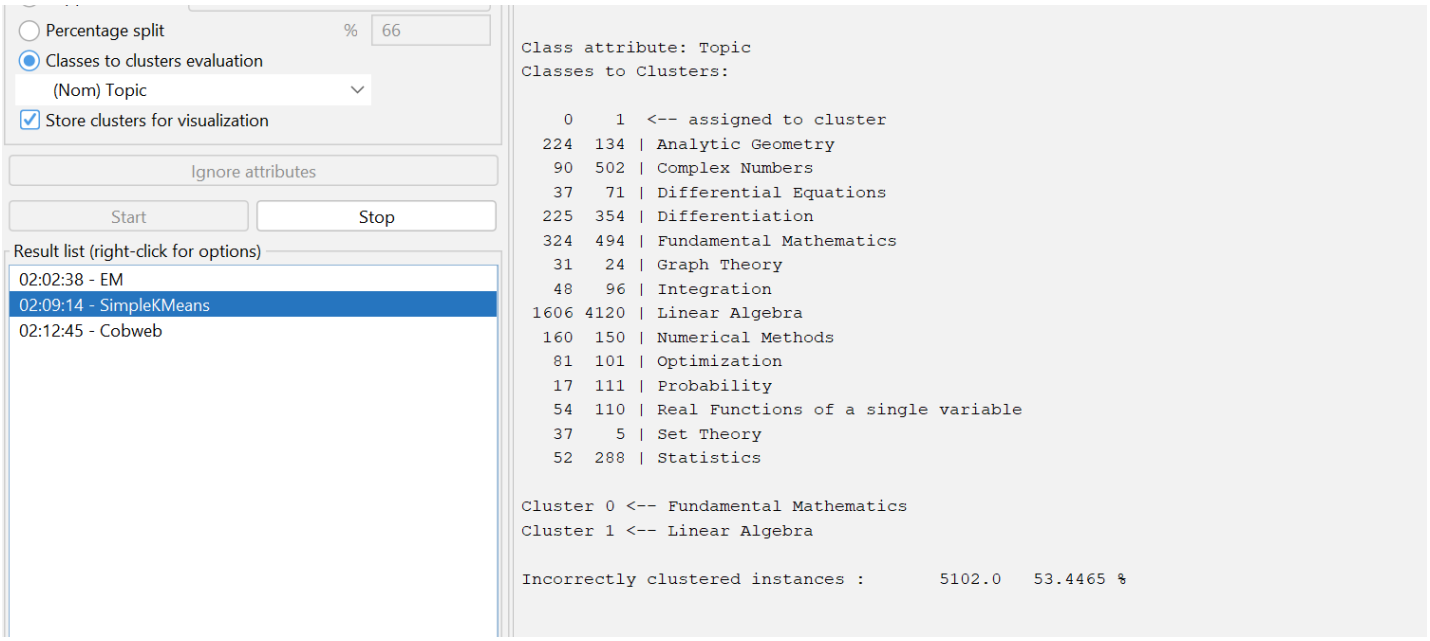


simpleKmeans(Classes to clusters evaluation based on Topic)

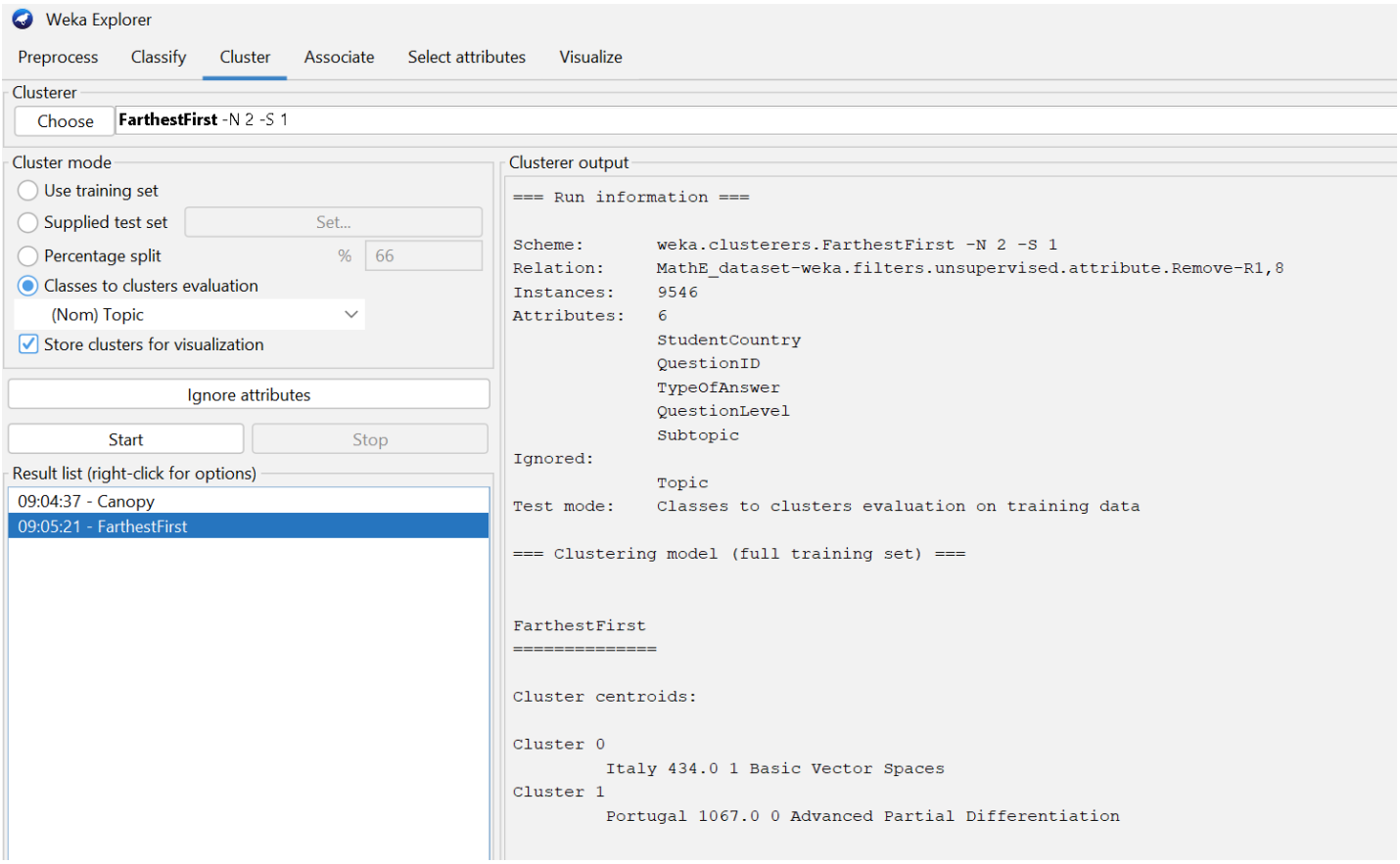


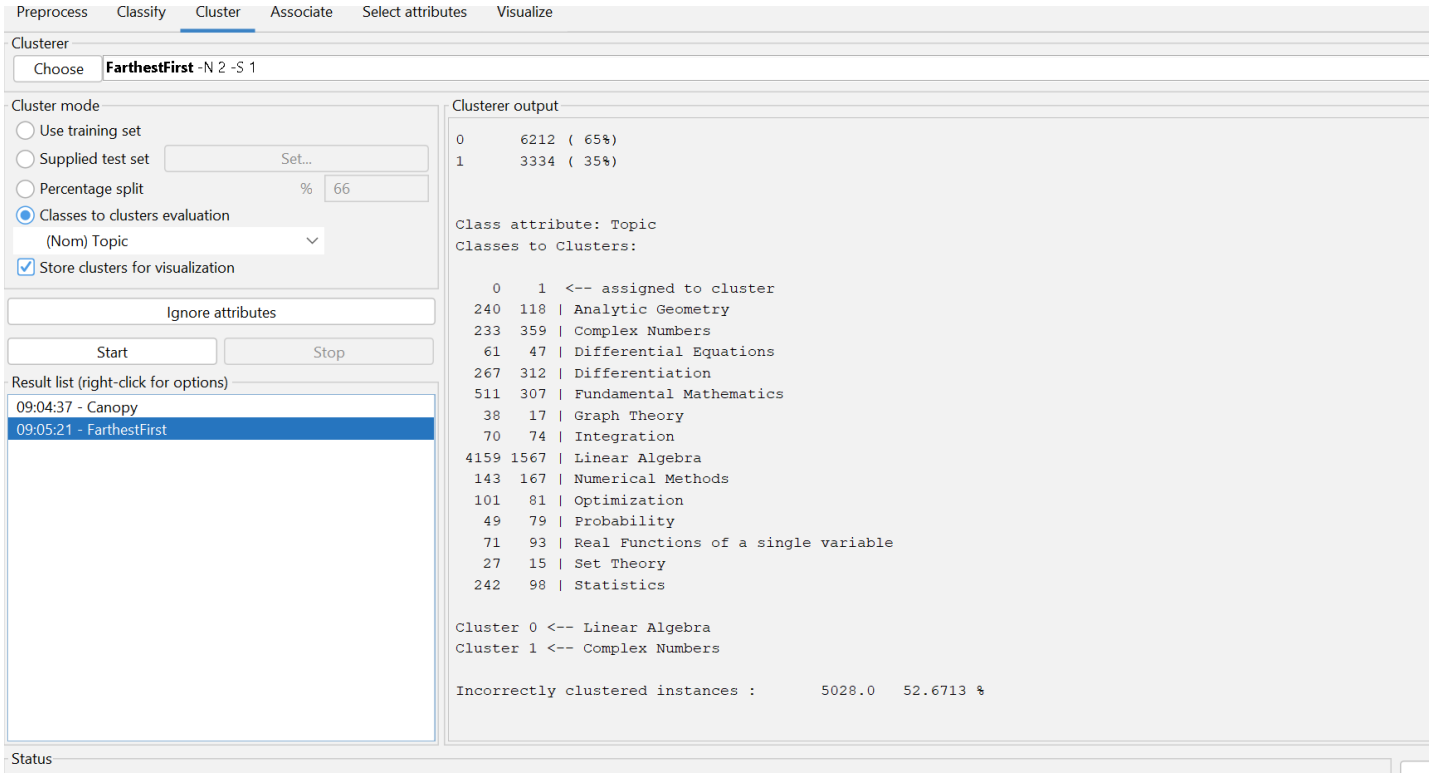

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
Cobweb -A 1.0 -C 0.0028209479177387815 -S 42
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Nom) Topic
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
02:02:38 - EM
02:09:14 - SimpleKMeans
02:12:45 - Cobweb
Clusterer output
Initial starting points (random):
Cluster 0: Lithuania,645,1,Basic,'Matrices and Determinants'
Cluster 1: Portugal,241,0,Basic,'Linear Systems'
Missing values globally replaced with mean/mode
Final cluster centroids:
Cluster#
Attribute Full Data 0 1
(9546.0) (2986.0) (6560.0)
StudentCountry Portugal Lithuania Portugal
QuestionID 478.9123 590.8426 427.9636
TypeOfAnswer 0 1 0
QuestionLevel Basic Basic Basic
Subtopic Vector Spaces Vector Spaces Vector Spaces
Time taken to build model (full training data) : 0.02 seconds
=== Model and evaluation on training set ===
Clustered Instances


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
Cobweb -A 1.0 -C 0.0028209479177387815 -S 42
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Nom) Topic
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
02:02:38 - EM
02:09:14 - SimpleKMeans
02:12:45 - Cobweb
Clusterer output
Time taken to build model (full training data) : 0.02 seconds
=== Model and evaluation on training set ===
Clustered Instances
0 2986 (31%)
1 6560 (69%)
Class attribute: Topic
Classes to Clusters:
0 1 <-- assigned to cluster
224 134 | Analytic Geometry
90 502 | Complex Numbers
37 71 | Differential Equations
225 354 | Differentiation
324 494 | Fundamental Mathematics
31 24 | Graph Theory
48 96 | Integration
1606 4120 | Linear Algebra
160 150 | Numerical Methods
81 101 | Optimization
17 111 | Probability
54 110 | Real Functions of a single variable
37 5 | Set Theory
52 288 | Statistics



fathestFirst(Classes to clusters evaluation based on Topic)




Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
FarthestFirst -N 2 -S 1
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Nom) Topic
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
09:04:37 - Canopy
09:05:21 - FarthestFirst
Clusterer output
0 6212 (65%)
1 3334 (35%)
Class attribute: Topic
Classes to Clusters:
0 1 <-- assigned to cluster
240 118 | Analytic Geometry
233 359 | Complex Numbers
61 47 | Differential Equations
267 312 | Differentiation
511 307 | Fundamental Mathematics
38 17 | Graph Theory
70 74 | Integration
4159 1567 | Linear Algebra
143 167 | Numerical Methods
101 81 | Optimization
49 79 | Probability
71 93 | Real Functions of a single variable
27 15 | Set Theory
242 98 | Statistics
Cluster 0 <-- Linear Algebra
Cluster 1 <-- Complex Numbers
Incorrectly clustered instances : 5028.0 52.6713 %
Status

2.2.4 Result

Table 1: Performance Comparison of Clustering Algorithms

Metric	FarthestFirst	SimpleKMeans	EM (Expectation Maximization)
Number of Clusters	2 (Fixed)	2 (Fixed)	15 (Auto-detected)
Time to Build Model	0.06 seconds (in the second attempt)	0.09 seconds (in the second attempt)	273.79 seconds (in the second attempt)
Incorrectly Clustered	5,028 (52.67%)	5,102 (53.45%)	5,720 (59.92%)
Clustering Accuracy	**47.33%**	**46.55%**	**40.08%**
Primary Insight	Separated *Linear Algebra* from *Complex Numbers*	Separated *Fundamental Math* from *Linear Algebra*	Detailed breakdown of specific topics (Statistics, Probability, etc.)

Table 2: Cluster Interpretations (k=2 Models)

Algorithm	Cluster 0 Label	Cluster 1 Label	Observation
FarthestFirst	Linear Algebra	Complex Numbers	It picked two very distinct topics as "anchors" for the clusters.
SimpleKMeans	Fundamental Mathematics	Linear Algebra	It grouped the data based on the two largest topic categories.

Table 3: EM Algorithm Detailed Topic Mapping

Unlike the other models, EM automatically detected 15 clusters. This table shows which clusters successfully identified a specific Math Topic.

Cluster ID	Identified Topic (Majority Class)
Cluster 0	**Linear Algebra**

Cluster 1	**Probability**
Cluster 2	**Complex Numbers**
Cluster 3	**Numerical Methods**
Cluster 4	**Optimization**
Cluster 6	**Statistics**
Cluster 8	**Integration**
Cluster 9	**Differentiation**
Cluster 11	**Graph Theory**
Cluster 12	**Analytic Geometry**
Cluster 13	**Fundamental Mathematics**
Unassigned	**Clusters 5, 7, 10, 14 (mostly contained duplicates of Linear Algebra)**

References

University of California, Irvine. (n.d.). UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

Lichtinghagen, R., Klawonn, F., & Hoffmann, G. (2020). *HCV data* [Data set]. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/571/hcv+data

Azevedo, B., Pacheco, M., Fernandes, F., & Pereira, A. (2024). *Assessing Mathematics Learning in Higher Education (MathE)* [Data set]. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/1031/dataset+for+assessing+mathematics+learning+in+higher+education