Parshvanath Charitable Trust's
A. P. SHAH INSTITUTE OF TECHNOLOGY, THANE
(All Programs Accredited by NBA)
Department of Information Technology

Business Intelligence Mini Project
Cervical Cancer Behaviour Risk

ITL602 BI LAB Semester VI

AY : 2023-2024

Data Set Name-UCI

Submitted By

1. **Harmi Mathukiya 21104044**
2. **Avantika More 21104033**
3. **Atharva Mohape 21104121**

1. Problem Definition

- To classify a cervical cancer behavior risk project involves identifying and understanding the factors contributing to cervical cancer risk behaviors and designing interventions to mitigate these risks effectively.
- The project must comprehensively analyze the behaviors and practices that contribute to cervical cancer risk, including but not limited to lack of screening, delay in seeking medical care, high-risk sexual behaviors, smoking, poor diet, and lack of awareness about the disease and preventive measures.

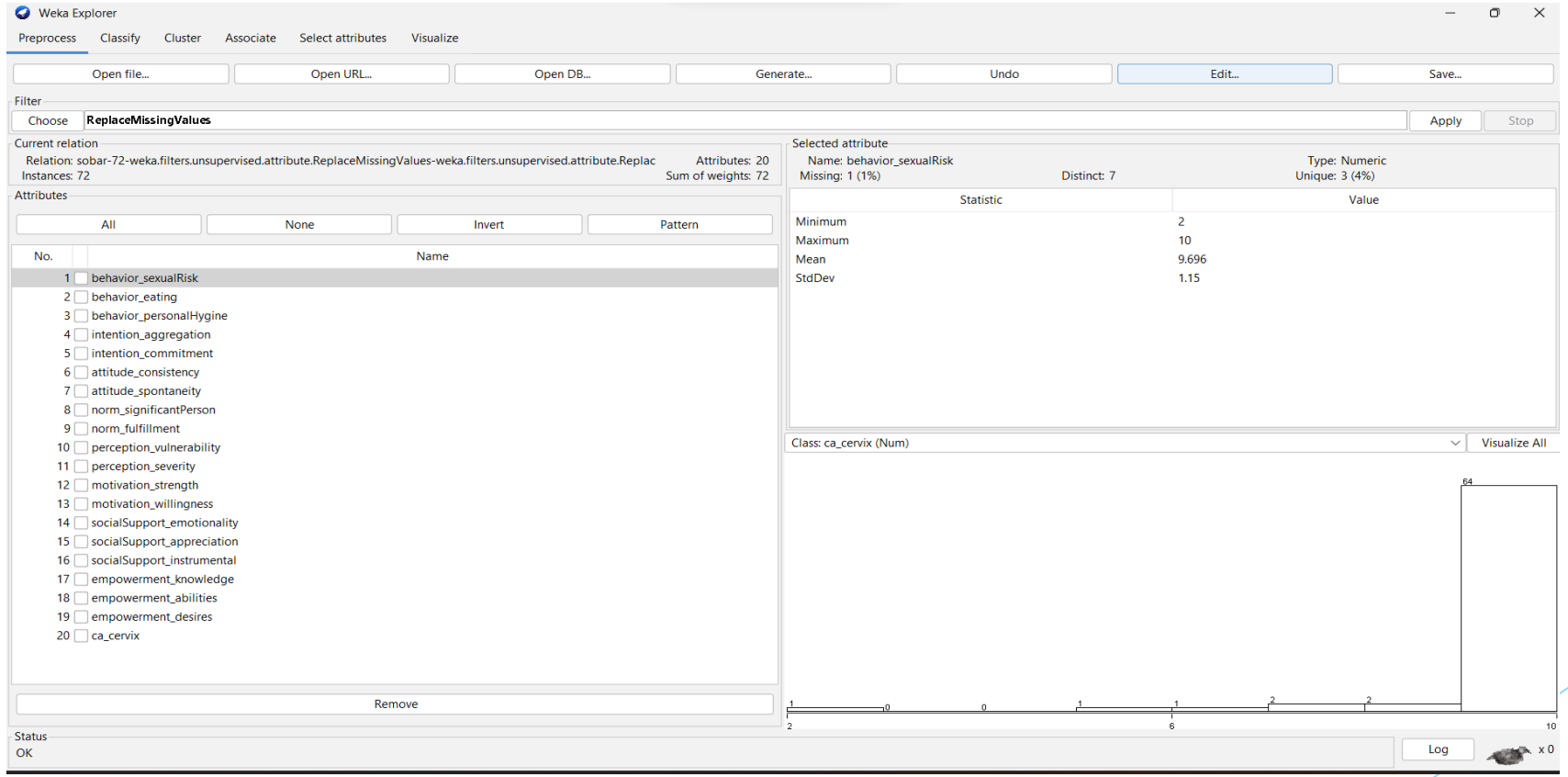
2. Dataset identified

- Name of dataset-UCI
- Dataset source-https://archive.ics.uci.edu/static/public/537/cervical+cancer+behavior+risk.zip
- Brief description of data-The dataset from the UCI Machine Learning Repository comprises anonymized patient information related to cervical cancer. It includes a range of attributes such as demographic details, habits, medical history, and clinical information. These features cover factors like age, number of sexual partners, smoking habits, hormonal contraceptive use, and more. The dataset aims to provide insights into the behavior risk associated with cervical cancer, offering valuable information for predictive modeling and early intervention strategies to mitigate the impact of this prevalent health concern among women.

3. Data mining task performed

- **Filters used:**

ReplaceMissingValue-



- **Remove:**

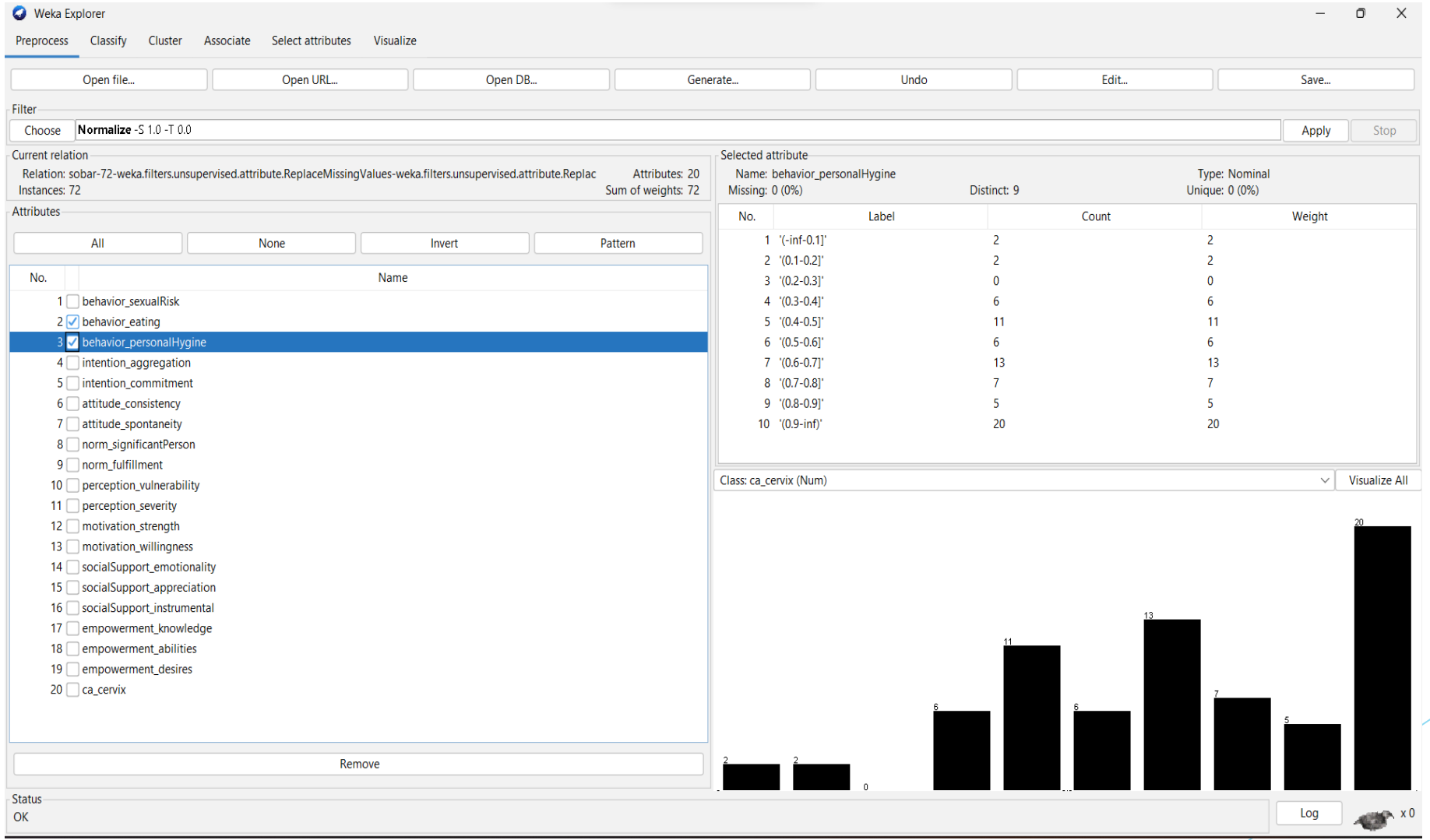
Viewer

Relation: sobar-72-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.ReplaceMissingVal

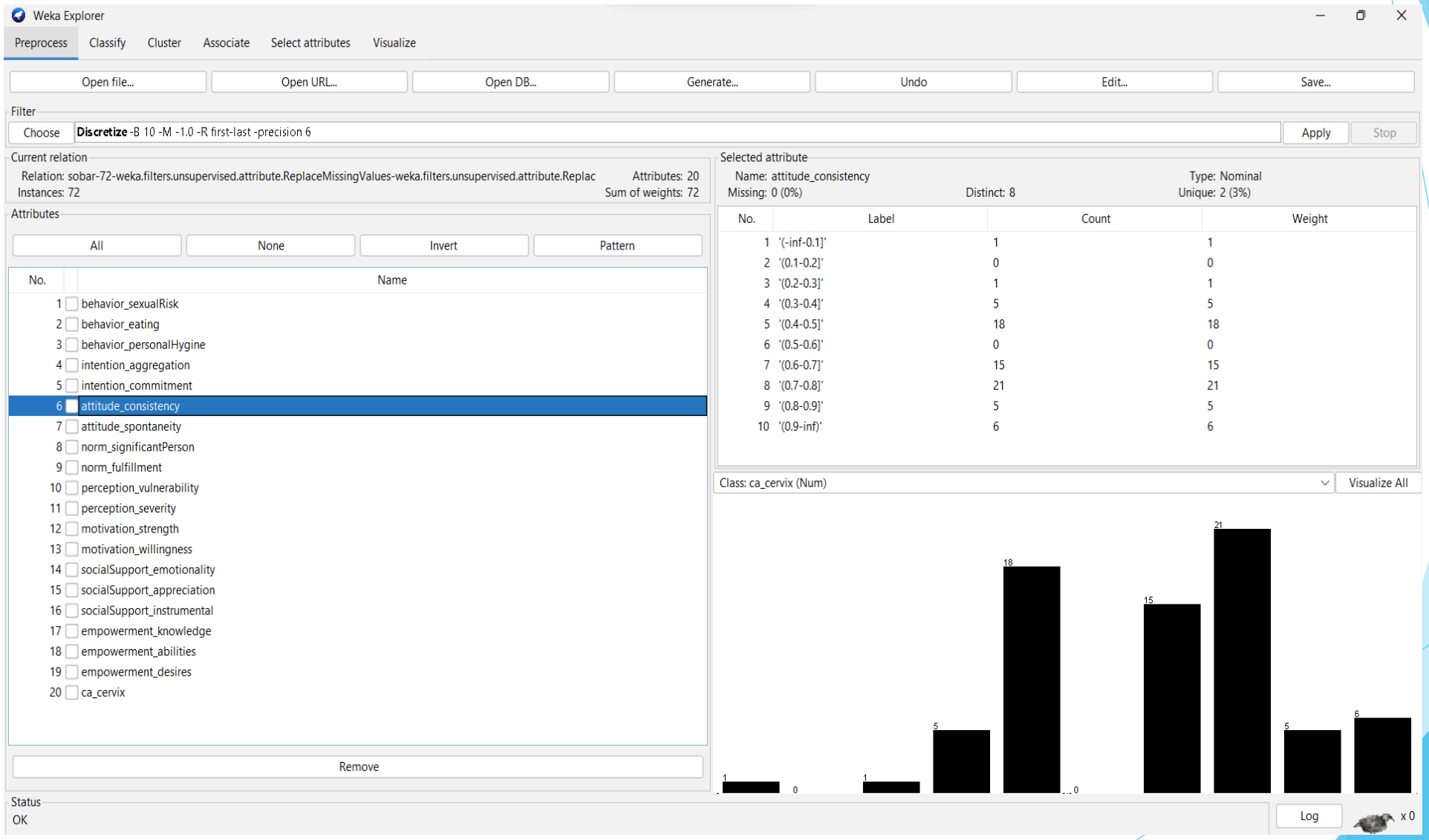
No.	1: behavior_sexualRisk Numeric	2: behavior_eating Numeric	3: behavior_personalHygine Numeric	4: intention_aggregation Numeric	5: intention_commitment Numeric	6: attitude_consistency Numeric	7: attitude_spontaneity Numeric
1	10.0	13.0	12.0	4.0	7.0	9.0	10.0
2	10.0	11.0	11.0	10.0	14.0	7.0	7.0
3		15.0	3.0	2.0	14.0	8.0	10.0
4	10.0	11.0	10.0	10.0	15.0	7.0	7.0
5	8.0	11.0	7.0	8.0	10.0	7.0	8.0
6	10.0	14.0	8.0	6.0	15.0	8.0	10.0
7	10.0	15.0	4.0	6.0	14.0	6.0	10.0
8	8.0	12.0	9.0	10.0	10.0	5.0	10.0
9	10.0	15.0	7.0	2.0	15.0	6.0	10.0
10		15.0	7.0	6.0	11.0	8.0	8.0
11	7.0	15.0	7.0	10.0	14.0	7.0	9.0
12	10.0		8.0	9.0	15.0	7.0	10.0
13	10.0	15.0	12.0	10.0	15.0	6.0	10.0
14	9.0	12.0	14.0	9.0	15.0	10.0	9.0
15	2.0	15.0	15.0	6.0	13.0	8.0	9.0
16	10.0	15.0	7.0	6.0	14.0	8.0	8.0
17	10.0	15.0	9.0	7.0	6.0	8.0	8.0
18	10.0	12.0	7.0	5.0	10.0	8.0	8.0
19	10.0	11.0	12.0	2.0	10.0	8.0	8.0
20	10.0	12.0	12.0	8.0	10.0	8.0	6.0
21	10.0	15.0	15.0	4.0	15.0	8.0	10.0
22	10.0	12.0	11.0	10.0	15.0	7.0	8.0
23	10.0	13.0	14.0	10.0	15.0	6.0	8.0

Add instance | Undo | OK | Cancel

Normalize:

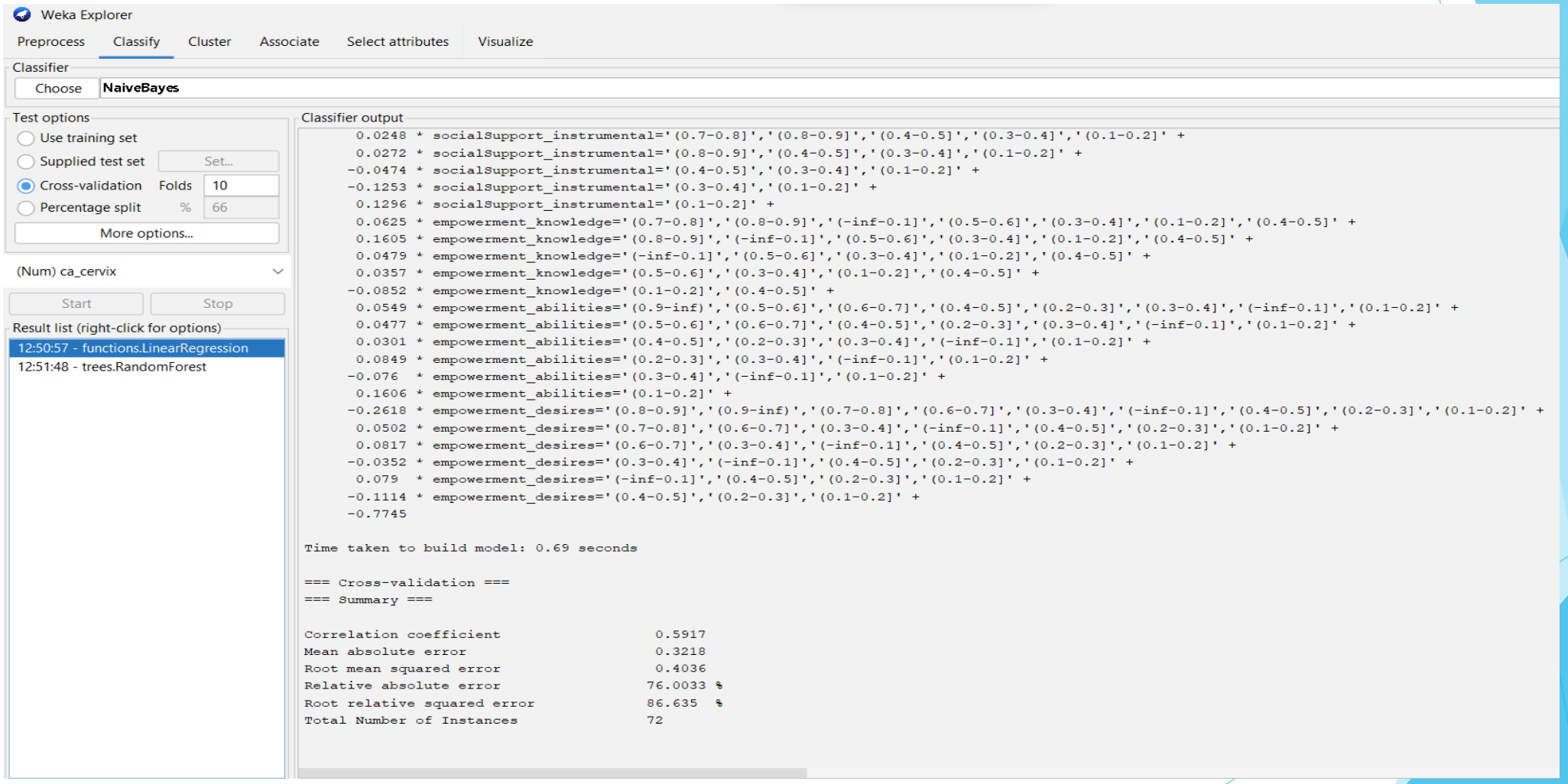

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
Normalize -S 1.0 -T 0.0
Apply
Stop
Current relation
Relation: sobar-72-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Replac
Attributes: 20
Instances: 72
Sum of weights: 72
Attributes
All
None
Invert
Pattern
No.
Name
1 behavior_sexualRisk
2 behavior_eating
3 behavior_personalHygine
4 intention_aggregation
5 intention_commitment
6 attitude_consistency
7 attitude_spontaneity
8 norm_significantPerson
9 norm_fulfillment
10 perception_vulnerability
11 perception_severity
12 motivation_strength
13 motivation_willingness
14 socialSupport_emotionality
15 socialSupport_appreciation
16 socialSupport_instrumental
17 empowerment_knowledge
18 empowerment_abilities
19 empowerment_desires
20 ca_cervix
Remove
Selected attribute
Name: behavior_personalHygine
Type: Nominal
Missing: 0 (0%)
Distinct: 9
Unique: 0 (0%)
No. Label Count Weight
1 '(-inf-0.1]' 2 2
2 '(0.1-0.2]' 2 2
3 '(0.2-0.3]' 0 0
4 '(0.3-0.4]' 6 6
5 '(0.4-0.5]' 11 11
6 '(0.5-0.6]' 6 6
7 '(0.6-0.7]' 13 13
8 '(0.7-0.8]' 7 7
9 '(0.8-0.9]' 5 5
10 '(0.9-inf)' 20 20
Class: ca_cervix (Num)
Visualize All
2
2
0
6
11
6
13
7
5
20
Status
OK
Log
x 0

Discretize:

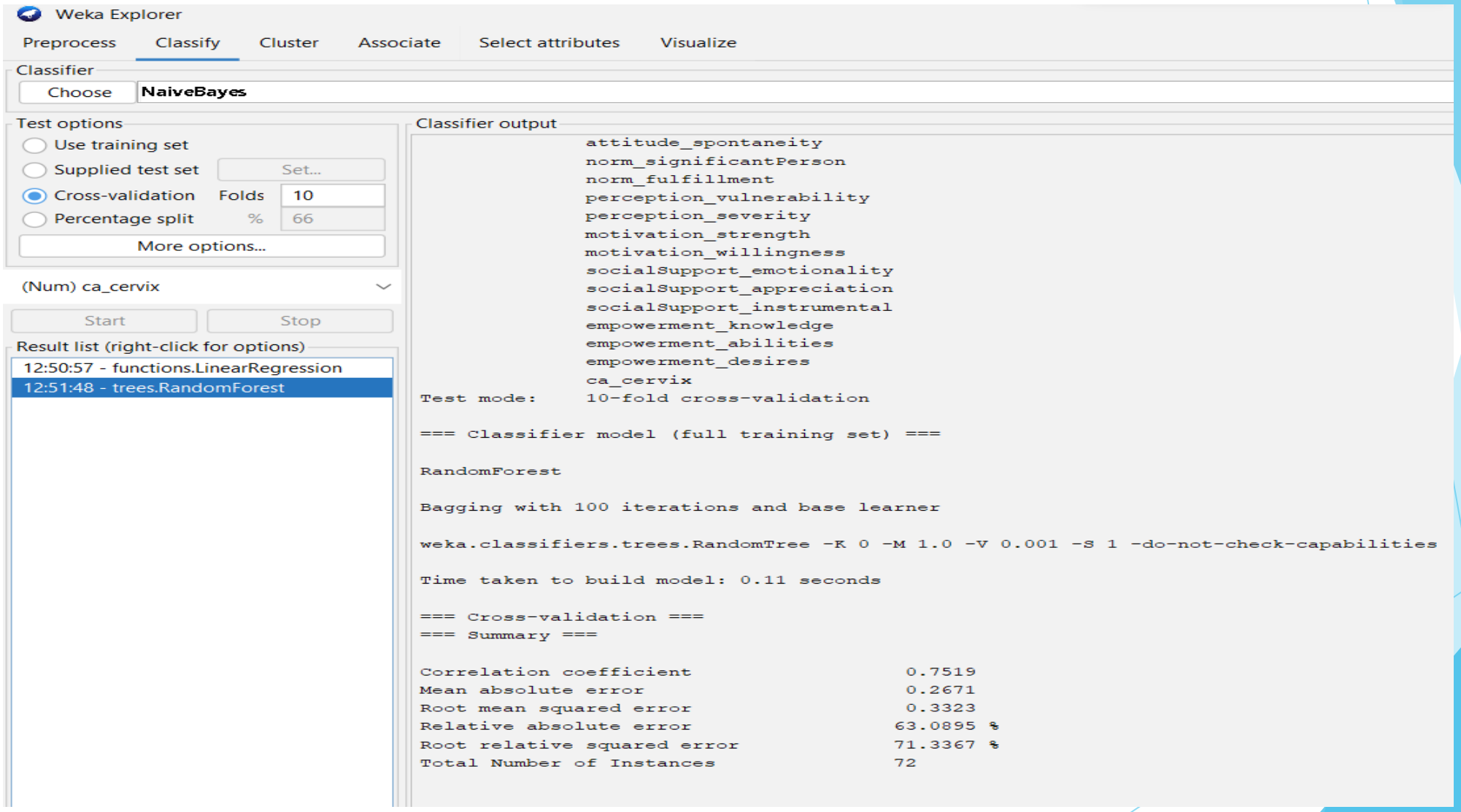

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
Discretize -B 10 -M -1.0 -R first-last -precision 6
Apply
Stop
Current relation
Relation: sobar-72-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Replac
Attributes: 20
Instances: 72
Sum of weights: 72
Attributes
All
None
Invert
Pattern
No.
Name
1 behavior_sexualRisk
2 behavior_eating
3 behavior_personalHygine
4 intention_aggregation
5 intention_commitment
6 attitude_consistency
7 attitude_spontaneity
8 norm_significantPerson
9 norm_fulfillment
10 perception_vulnerability
11 perception_severity
12 motivation_strength
13 motivation_willingness
14 socialSupport_emotionality
15 socialSupport_appreciation
16 socialSupport_instrumental
17 empowerment_knowledge
18 empowerment_abilities
19 empowerment_desires
20 ca_cervix
Remove
Selected attribute
Name: attitude_consistency
Type: Nominal
Missing: 0 (0%)
Distinct: 8
Unique: 2 (3%)
No. Label Count Weight
1 '(-inf-0.1]' 1 1
2 '(0.1-0.2]' 0 0
3 '(0.2-0.3]' 1 1
4 '(0.3-0.4]' 5 5
5 '(0.4-0.5]' 18 18
6 '(0.5-0.6]' 0 0
7 '(0.6-0.7]' 15 15
8 '(0.7-0.8]' 21 21
9 '(0.8-0.9]' 5 5
10 '(0.9-inf)' 6 6
Class: ca_cervix (Num)
Visualize All
1
0
1
5
18
0
15
21
5
6
Status
OK
Log
x 0

- **Algorithms Used:**

Linear Regression:



Random Forest:



4. Conclusion

The Random Forest algorithm demonstrated several advantages in this context:

- Accuracy: Random Forest tends to offer high accuracy in classification tasks, which is crucial for effectively identifying behavior risk levels associated with cervical cancer.
- Robustness: Random Forest is less prone to overfitting compared to individual decision trees, thanks to its ensemble nature. It can handle noise and complex relationships in the data, making it well-suited for real-world datasets like the one used in this project.
- Feature Importance: Random Forest provides a measure of feature importance, allowing us to identify the most influential attributes contributing to behavior risk prediction. This can offer valuable insights for healthcare practitioners in understanding the factors associated with cervical cancer risk.
- Interpretability: While Random Forest is not as interpretable as a single decision tree, it still provides a degree of interpretability that allows us to understand the decision-making process behind behavior risk classification.