



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-

The screenshot shows the Weka Explorer interface with the 'Preprocess' tab selected. The 'ReplaceMissingValues' filter is applied to the 'behavior_sexualRisk' attribute. The 'Current relation' is 'sobar-72-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Replac', with 20 attributes and 72 instances. The 'Selected attribute' is 'behavior_sexualRisk', which is numeric with 7 distinct values and 1 missing value (1%).

Attributes:

No.	Name
1	behavior_sexualRisk
2	behavior_eating
3	behavior_personalHygiene
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

Selected attribute statistics:

Statistic	Value
Minimum	2
Maximum	10
Mean	9.696
StdDev	1.15

Class: ca_cervix (Num) **Visualize All**

Status: OK

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_personalHygiene 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
24	10.0	15.0	12.0	10.0	15.0	7.0	10.0

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: sobar-72-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Replac
Instances: 72 | Attributes: 20 | Sum of weights: 72

Attributes: All | None | Invert | Pattern

No.	Name
1	<input type="checkbox"/> behavior_sexualRisk
2	<input checked="" type="checkbox"/> behavior_eating
3	<input checked="" type="checkbox"/> behavior_personalHygiene
4	<input type="checkbox"/> intention_aggregation
5	<input type="checkbox"/> intention_commitment
6	<input type="checkbox"/> attitude_consistency
7	<input type="checkbox"/> attitude_spontaneity
8	<input type="checkbox"/> norm_significantPerson
9	<input type="checkbox"/> norm_fulfillment
10	<input type="checkbox"/> perception_vulnerability
11	<input type="checkbox"/> perception_severity
12	<input type="checkbox"/> motivation_strength
13	<input type="checkbox"/> motivation_willingness
14	<input type="checkbox"/> socialSupport_emotionality
15	<input type="checkbox"/> socialSupport_appreciation
16	<input type="checkbox"/> socialSupport_instrumental
17	<input type="checkbox"/> empowerment_knowledge
18	<input type="checkbox"/> empowerment_abilities
19	<input type="checkbox"/> empowerment_desires
20	<input type="checkbox"/> ca_cervix

Remove

Status: OK

Log x 0

Selected attribute: Name: behavior_personalHygiene | 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

Bin	Count
1	2
2	2
3	0
4	6
5	11
6	6
7	13
8	7
9	5
10	20

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 Sum of weights: 72 Instances: 72

Attributes: All None Invert Pattern

No.	Name
1	<input type="checkbox"/> behavior_sexualRisk
2	<input type="checkbox"/> behavior_eating
3	<input type="checkbox"/> behavior_personalHygiene
4	<input type="checkbox"/> intention_aggregation
5	<input type="checkbox"/> intention_commitment
6	<input checked="" type="checkbox"/> attitude_consistency
7	<input type="checkbox"/> attitude_spontaneity
8	<input type="checkbox"/> norm_significantPerson
9	<input type="checkbox"/> norm_fulfillment
10	<input type="checkbox"/> perception_vulnerability
11	<input type="checkbox"/> perception_severity
12	<input type="checkbox"/> motivation_strength
13	<input type="checkbox"/> motivation_willingness
14	<input type="checkbox"/> socialSupport_emotionality
15	<input type="checkbox"/> socialSupport_appreciation
16	<input type="checkbox"/> socialSupport_instrumental
17	<input type="checkbox"/> empowerment_knowledge
18	<input type="checkbox"/> empowerment_abilities
19	<input type="checkbox"/> empowerment_desires
20	<input type="checkbox"/> ca_cervix

Remove

Status: OK

Selected attribute: Name: attitude_consistency Missing: 0 (0%) Distinct: 8 Type: Nominal 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

Bin	Count
1	1
2	0
3	1
4	5
5	18
6	0
7	15
8	21
9	5
10	6

Log x 0

► Algorithms Used:

Linear Regression:

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier

Choose **NaiveBayes**

Test options

☐ Use training set

☐ Supplied test set

☒ Cross-validation Folds

☐ Percentage split %

(Num) ca_cervix

Result list (right-click for options)

12:50:57 - functions.LinearRegression

12:51:48 - trees.RandomForest

Classifier output

```
0.0248 * socialSupport_instrumental='(0.7-0.8]', '(0.8-0.9]', '(0.4-0.5]', '(0.3-0.4]', '(0.1-0.2]' +
0.0272 * socialSupport_instrumental='(0.8-0.9]', '(0.4-0.5]', '(0.3-0.4]', '(0.1-0.2]' +
-0.0474 * socialSupport_instrumental='(0.4-0.5]', '(0.3-0.4]', '(0.1-0.2]' +
-0.1253 * socialSupport_instrumental='(0.3-0.4]', '(0.1-0.2]' +
0.1296 * socialSupport_instrumental='(0.1-0.2]' +
0.0625 * empowerment_knowledge='(0.7-0.8]', '(0.8-0.9]', '(-inf-0.1]', '(0.5-0.6]', '(0.3-0.4]', '(0.1-0.2]', '(0.4-0.5]' +
0.1605 * empowerment_knowledge='(0.8-0.9]', '(-inf-0.1]', '(0.5-0.6]', '(0.3-0.4]', '(0.1-0.2]', '(0.4-0.5]' +
0.0479 * empowerment_knowledge='(-inf-0.1]', '(0.5-0.6]', '(0.3-0.4]', '(0.1-0.2]', '(0.4-0.5]' +
0.0357 * empowerment_knowledge='(0.5-0.6]', '(0.3-0.4]', '(0.1-0.2]', '(0.4-0.5]' +
-0.0852 * empowerment_knowledge='(0.1-0.2]', '(0.4-0.5]' +
0.0549 * empowerment_abilities='(0.9-inf)', '(0.5-0.6]', '(0.6-0.7]', '(0.4-0.5]', '(0.2-0.3]', '(0.3-0.4]', '(-inf-0.1]', '(0.1-0.2]' +
0.0477 * empowerment_abilities='(0.5-0.6]', '(0.6-0.7]', '(0.4-0.5]', '(0.2-0.3]', '(0.3-0.4]', '(-inf-0.1]', '(0.1-0.2]' +
0.0301 * empowerment_abilities='(0.4-0.5]', '(0.2-0.3]', '(0.3-0.4]', '(-inf-0.1]', '(0.1-0.2]' +
0.0849 * empowerment_abilities='(0.2-0.3]', '(0.3-0.4]', '(-inf-0.1]', '(0.1-0.2]' +
-0.076 * empowerment_abilities='(0.3-0.4]', '(-inf-0.1]', '(0.1-0.2]' +
0.1606 * empowerment_abilities='(0.1-0.2]' +
-0.2618 * empowerment_desires='(0.8-0.9]', '(0.9-inf)', '(0.7-0.8]', '(0.6-0.7]', '(0.3-0.4]', '(-inf-0.1]', '(0.4-0.5]', '(0.2-0.3]', '(0.1-0.2]' +
0.0502 * empowerment_desires='(0.7-0.8]', '(0.6-0.7]', '(0.3-0.4]', '(-inf-0.1]', '(0.4-0.5]', '(0.2-0.3]', '(0.1-0.2]' +
0.0817 * empowerment_desires='(0.6-0.7]', '(0.3-0.4]', '(-inf-0.1]', '(0.4-0.5]', '(0.2-0.3]', '(0.1-0.2]' +
-0.0352 * empowerment_desires='(0.3-0.4]', '(-inf-0.1]', '(0.4-0.5]', '(0.2-0.3]', '(0.1-0.2]' +
0.079 * empowerment_desires='(-inf-0.1]', '(0.4-0.5]', '(0.2-0.3]', '(0.1-0.2]' +
-0.1114 * empowerment_desires='(0.4-0.5]', '(0.2-0.3]', '(0.1-0.2]' +
-0.7745
```

Time taken to build model: 0.69 seconds

=== Cross-validation ===

=== Summary ===

Correlation coefficient	0.5917
Mean absolute error	0.3218
Root mean squared error	0.4036
Relative absolute error	76.0033 %
Root relative squared error	86.635 %
Total Number of Instances	72

Random Forest:

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier
Choose **NaiveBayes**

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds 10

☐ Percentage split % 66

More options...

(Num) ca_cervix ▾

Start Stop

Result list (right-click for options)

- 12:50:57 - functions.LinearRegression
- 12:51:48 - trees.RandomForest**

Classifier output

```
attitude_spontaneity
norm_significantPerson
norm_fulfillment
perception_vulnerability
perception_severity
motivation_strength
motivation_willingness
socialSupport_emotionality
socialSupport_appreciation
socialSupport_instrumental
empowerment_knowledge
empowerment_abilities
empowerment_desires
ca_cervix

Test mode:   10-fold cross-validation

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.11 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                      0.7519
Mean absolute error                         0.2671
Root mean squared error                     0.3323
Relative absolute error                    63.0895 %
Root relative squared error                71.3367 %
Total Number of Instances                 72
```

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.