

Title: Young Peoples Survey

Group Number: 34

First Name	Last Name	Online Students? (Y or N)	Shared with ITMD 527? (Y or N)
Prashant	Patil	Y	N
Shashank	Khede	N	N
Navtej Singh	Chawla	N	N

Table of Contents

1. Introduction and Motivations.....	2
2. Data Description.....	3
3. Research Problems and Solutions	3
4. KDD	4
4.1. Data Processing.....	4
4.2. Data Mining Tasks and Processes	19
5. Evaluations and Results	19
5.1. Evaluation Methods	19
5.2. Results and Findings.....	19
6. Conclusions and Future Work.....	19
6.1. Conclusions	19
6.2. Limitations.....	20
6.3. Potential Improvements or Future Work	20

1. Introduction and Motivations

People survey is an interesting collection of data where people were asked questions about their music preferences, movie preferences, their hobbies, phobias, views on life and personal traits, spending habits to analyze that is there any group of people with same interests or likings. People answer these questions on a range of 1 to 5 based on their likings. This survey was done electronically as well as on paper. This survey can be used in recommender systems where a person can be recommended movie or music based on his/her interests. The data file for this survey has around 1010 data with 150 attributes. This survey is anonymous and the person name is not taken so that we can answer multiple questions and take up this as a research project.

2. Data Description

I. We have taken our data set is collected from www.kaggle.com.

II. The dataset contains 1010 data where each row has around 150 columns of data.

There are multiple variables those are:

1. Music preferences
2. Movie preference
3. Hobbies & amp Interests
4. Phobias
5. Health habits
6. Personal traits, opinions and views on life
7. Spending habits and other demographics.

Each of these variables have multiple options like movie preference has further options which are horror, thriller, comedy, romantic, sci-fi, war, animated or action. Each row represents a person's response to question on the range of 1-5 where 1 is strongly is agree/dislike and 5 is strongly agree/like.

The dataset is quite large with the combination of numerical and categorical data along with missing data which makes this data set a challenging one to work on. This is basically a survey data conducted online and on paper both where data helps in analyzing that per a person's interest or preference what kind of music he/she likes or what phobia he/she has or do left-handed people have different kind of interests than that of right-handed people. This kind of datasets can be used to answer different questions on personal behaviors, phobias, interests and spending habits. The knowledge out of this data can be implemented in the recommender systems where people will be recommended, items based on their interests.

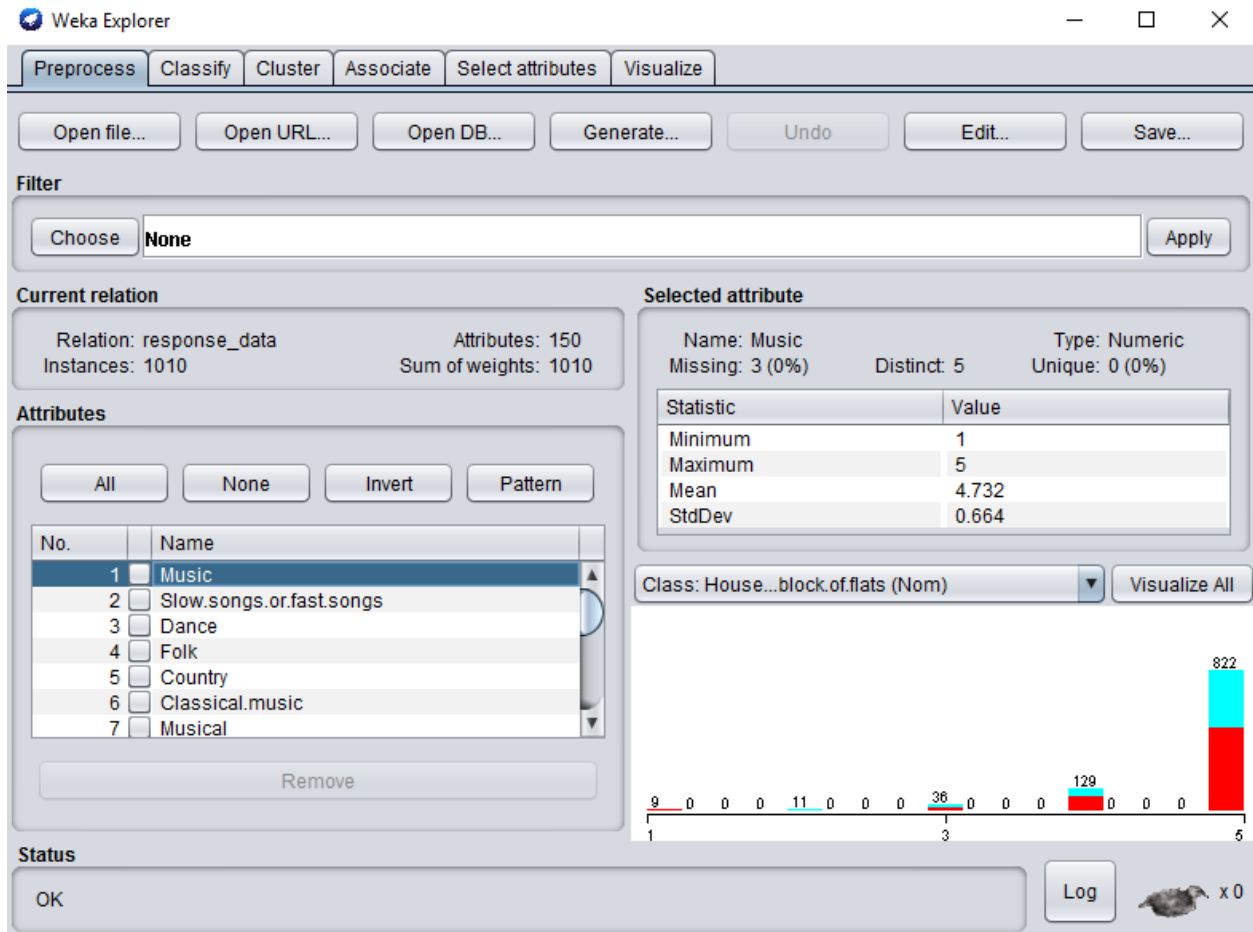
3. Research Problems and Solutions

1. Is there any gender wise difference in music/movie preferences?
2. Can we find patterns in different phobias?
3. How can we handle missing values to do our analysis more effectively?
4. What are the spending habits of people according to gender and their location?
5. What can be difference in interests gender wise?
6. Is there a possibility that we can find those people from the survey who cheated and answered questions randomly?

4. KDD

4.1. Data Processing

We have in all 150 columns containing 139 numerical data and 11 categorical data. Now we don't need these columns so we remove the unnecessary columns and use only those which are required for us research/mining task. We choose 84 columns which we require and these columns are related to music, movie, phobia, money spent data. Some more columns selected are age, gender, left-right handed, village or city Original Response Data file with 150 attributes:



We remove unused attributes:

Weka Explorer

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

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter

Choose Remove -R 74-133 Apply

Current relation

Relation: response_data-weka.filt... Attributes: 90
Instances: 1010 Sum of weights: 1010

Selected attribute

Name: Music
Missing: 3 (0%) Distinct: 5 Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	5
Mean	4.732
StdDev	0.664

Class: House...block.of.flats (Nom) Visualize All

Attributes

All None Invert Pattern

No.	Name
84	<input type="checkbox"/> Number.of.siblings
85	<input type="checkbox"/> Gender
86	<input type="checkbox"/> Left...right.handed
87	<input type="checkbox"/> Education
88	<input type="checkbox"/> Only.child
89	<input type="checkbox"/> Village...town
90	<input type="checkbox"/> House...block.of.flats

Remove

Status

OK Log x 0

Value	Count
1	9
2	0
3	11
4	36
5	129

We do feature selection using Information Gain before removing the unused attributes

Weka Explorer

Preprocess Classify Cluster Associate **Select attributes** Visualize

Attribute Evaluator

Choose InfoGainAttributeEval

Search Method

Choose Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode

☒ Use full training set
☐ Cross-validation Folds 10
Seed 1

(Nom) Gender

Start Stop

Result list (right-click for options)

21:57:23 - Ranker + InfoGainAttributeEval
21:58:47 - GreedyStepwise + Wrapper
22:09:11 - Ranker + InfoGainAttributeEval
22:09:23 - Ranker + InfoGainAttributeEval
22:09:31 - Ranker + InfoGainAttributeEval
22:09:37 - Ranker + InfoGainAttributeEval
22:09:44 - Ranker + InfoGainAttributeEval
22:09:51 - GreedyStepwise + Wrapper
22:10:05 - Ranker + InfoGainAttributeEval
23:25:39 - Ranker + InfoGainAttributeEval

Attribute selection output

```

=== Run information ===

Evaluator:   weka.attributeSelection.InfoGainAttributeEval
Search:     weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:   mydata-weka.filters.unsupervised.instance.RemovePercentage-P80.0-V
Instances:  808
Attributes: 84
Music
Slow.songs.or.fast.songs
Dance
Folk
Country
Classical.music
Musical
Pop
Rock
Metal.or.Hardrock
Punk
Hiphop..Rap
Reggae..Ska
Swing..Jazz
Rock.n.roll
Alternative

```

We get in total 84 attributes

Weka Explorer

Preprocess Classify Cluster Associate **Select attributes** Visualize

Attribute Evaluator

Choose InfoGainAttributeEval

Search Method

Choose Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode

☒ Use full training set
☐ Cross-validation Folds 10
Seed 1

(Nom) Gender

Start Stop

Result list (right-click for options)

21:57:23 - Ranker + InfoGainAttributeEval
21:58:47 - GreedyStepwise + Wrapper
22:09:11 - Ranker + InfoGainAttributeEval
22:09:23 - Ranker + InfoGainAttributeEval
22:09:31 - Ranker + InfoGainAttributeEval
22:09:37 - Ranker + InfoGainAttributeEval
22:09:44 - Ranker + InfoGainAttributeEval
22:09:51 - GreedyStepwise + Wrapper
22:10:05 - Ranker + InfoGainAttributeEval
23:25:39 - Ranker + InfoGainAttributeEval

Attribute selection output

```

0      13 Reggae..Ska
0      5 Country
0      2 Slow.songs.or.fast.songs
0      3 Dance
0      4 Folk
0      61 Fun.with.friends
0      39 Economy.Management
0      41 Chemistry
0      15 Rock.n.roll
0      46 Law
0      23 Comedy
0      43 Geography
0      49 Religion
0      20 Movies
0      19 Opera
0      16 Alternative
0      53 Writing
0      52 Musical.instruments
0      14 Swing..Jazz

Selected attributes: 24,38,47,26,31,58,59,42,30,60,65,68,66,51,25,7,27,36,17,70,57,79,44,72,69,28,63,76,56,55,22,77,18,48,71,75,81,21,34,32,33,54,8,

```

Further removing unused values, we get 84 attributes and we save this file as arff.

Weka Explorer

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

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter

Choose Apply

Current relation

Relation: response_data-weka.filt... Attributes: 84
Instances: 1010 Sum of weights: 1010

Selected attribute

Name: Music Type: Numeric
Missing: 3 (0%) Distinct: 5 Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	5
Mean	4.732
StdDev	0.664

Class: Village...town (Nom) Visualize All

Attributes

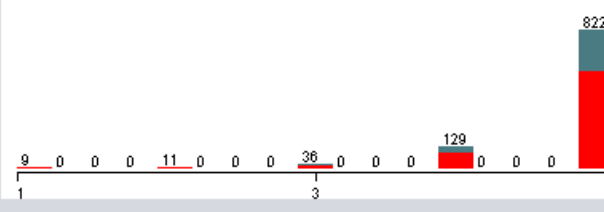
All None Invert Pattern

No.	Name
78	<input type="checkbox"/> Spending.on.looks
79	<input type="checkbox"/> Spending.on.gadgets
80	<input type="checkbox"/> Spending.on.healthy.eating
81	<input type="checkbox"/> Age
82	<input type="checkbox"/> Gender
83	<input type="checkbox"/> Left...right.handed
84	<input type="checkbox"/> Village...town

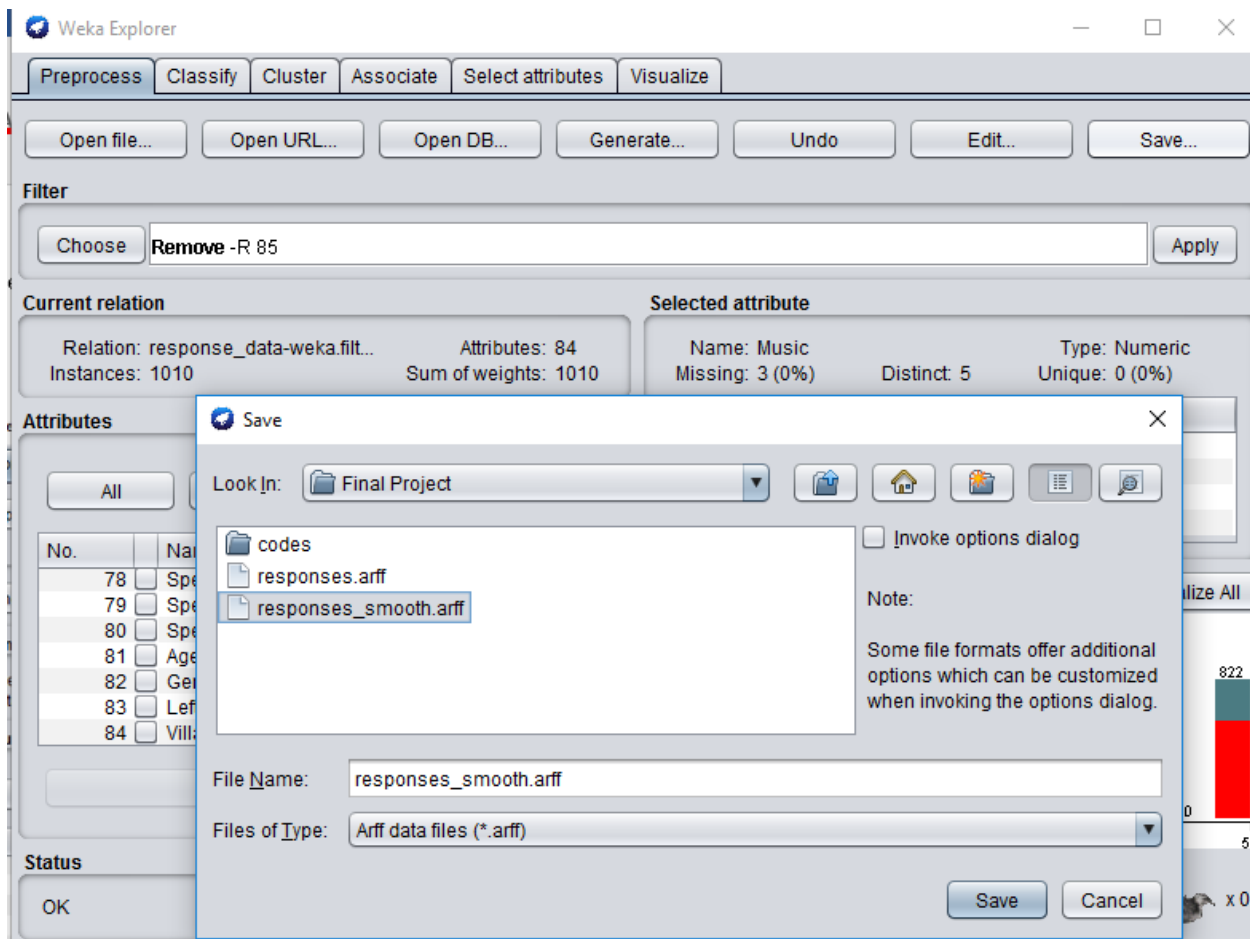
Remove

Status

OK Log x 0



We save the smooth file as arff.



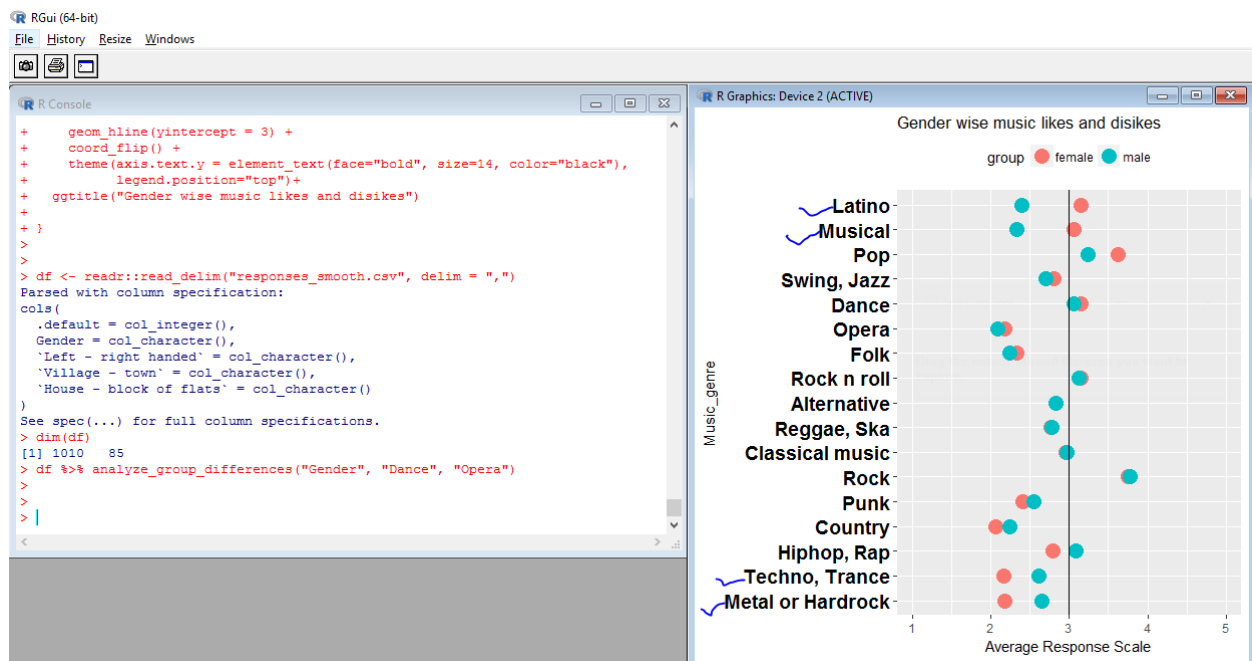
Next, we replace the missing values in the data. We replaced the numeric data by mean value because we have numeric data from range 1-5 so we replaced the missing values using value 3. Below are the steps:

[illegible][illegible][illegible]

responses_smooth - Excel													
File Home Insert Page Layout Formulas Data Review View Tell me what you want to do													
Clipboard		Font			Alignment			Number			Styles		
CF998													
	BV	BW	BX	BY	BZ	CA	CB	CC	CD	CE	CF	CG	CH
1	Finance	Shoppi	Brande	Enterta	Spendi	Spendi	Spendi	Age	Gender	Left - right handed	Village	House	lock of flats
3	3	4	1	4	2	5	2	19	female	right handed	village	block of flats	
47	4	5	4	4	5	4	5	27	male	right handed	city	block of flats	
170	4	3	4	3	3	3	4	18	female	right handed	village	house/bungalow	
198	3	1	4	2	3	3	3	20	male	right handed	city	block of flats	
199	3	1	1	5	2	2	2	19	female	right handed	village	block of flats	
012													
013													

Plotting a ggplot for male and female data to verify what kind of music they like.

- Females like “Latino” and “Musical” kind of music more than that of mens.
- Mens like to hear Metal or Hardrock and Techno, Trance music more often than that of women.



We keep only attributes related to music and keep gender data.

Removing remaining attributes using filter “Remove”

Weka Explorer

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

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter

Choose Remove -R 18-79 Apply

Current relation

Relation: mydata-weka.filters.unsupervised.attribute.Remove-R1-2-weka.filters.unsupe... Attributes: 20
Instances: 1010 Sum of weights: 1010

Attributes

All None Invert Pattern

No.	Name
5	<input type="checkbox"/> Musical
6	<input type="checkbox"/> Pop
7	<input type="checkbox"/> Rock
8	<input type="checkbox"/> Metal or Hardrock
9	<input type="checkbox"/> Punk
10	<input type="checkbox"/> Hiphop_Rap
11	<input type="checkbox"/> Reggae_Ska
12	<input type="checkbox"/> Swing_Jazz
13	<input type="checkbox"/> Rock.n.roll
14	<input type="checkbox"/> Alternative
15	<input type="checkbox"/> Latino
16	<input type="checkbox"/> Techno_Trance
17	<input type="checkbox"/> Opera
18	<input type="checkbox"/> Gender
19	<input type="checkbox"/> Left_right.handed
20	<input type="checkbox"/> Village..town

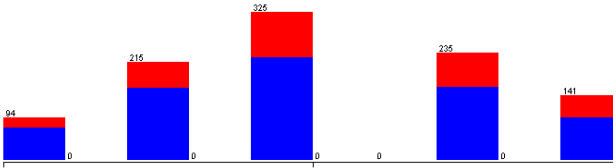
Remove

Selected attribute

Name: Dance
Missing: 0 (0%) Distinct: 5 Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	5
Mean	3.113
StdDev	1.168

Class: Village..town (Nom) Visualize All



“Gender” is our “Nominal” data which we will use for classification purpose.

Weka Explorer

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

Open file... Open URL... Open DB... Generate... Undo Edit...

Filter

Choose Remove -R 19-20

Current relation

Relation: mydata-weka.filters.unsupervised.attribute.Remove-R1-2-weka.filters.unsupe... Attributes: 18
Instances: 1010 Sum of weights: 1010

Attributes

All None Invert Pattern


No.	Name
3	<input type="checkbox"/> Country
4	<input type="checkbox"/> Classical.music
5	<input type="checkbox"/> Musical
6	<input type="checkbox"/> Pop
7	<input type="checkbox"/> Rock
8	<input type="checkbox"/> Metal or Hardrock
9	<input type="checkbox"/> Punk
10	<input type="checkbox"/> Hiphop_Rap
11	<input type="checkbox"/> Reggae_Ska
12	<input type="checkbox"/> Swing_Jazz
13	<input type="checkbox"/> Rock.n.roll
14	<input type="checkbox"/> Alternative
15	<input type="checkbox"/> Latino
16	<input type="checkbox"/> Techno_Trance
17	<input type="checkbox"/> Opera
18	<input checked="" type="checkbox"/> Gender

Selected attribute

Name: Gender
Missing: 0 (0%) Distinct: 2 Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	female	596	596.0
2	male	414	414.0

Class: Gender (Nom)



We divide our data into training and testing sets and perform classification.

Weka Explorer

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

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose RemovePercentage - P 80.0 - V Apply

Current relation: Relation: mydata-weka.filters.unsupervised.attribute.Remove-R1-2-weka.filters.unsuper... Attributes: 18 Instances: 808 Sum of weights: 808

Attributes: All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> Dance
2	<input type="checkbox"/> Folk
3	<input type="checkbox"/> Country
4	<input type="checkbox"/> Classical.music
5	<input type="checkbox"/> Musical
6	<input type="checkbox"/> Pop
7	<input type="checkbox"/> Rock
8	<input type="checkbox"/> Metal.or.Hardrock
9	<input type="checkbox"/> Punk
10	<input type="checkbox"/> Hiphop.Rap
11	<input type="checkbox"/> Reggae.Ska
12	<input type="checkbox"/> Swing.Jazz
13	<input type="checkbox"/> Rock.n.roll
14	<input type="checkbox"/> Alternative
15	<input type="checkbox"/> Latino
16	<input type="checkbox"/> Techno.Trance

Remove

Selected attribute

Name: Dance
Missing: 0 (0%)
Distinct: 5
Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	5
Mean	3.13
StdDev	1.179

Class: Gender (Nom) Visualize All

Dance Value	Class 1 (Blue)	Class 2 (Red)
1	77	0
2	0	184
3	0	264
4	0	193
5	0	120

We perform accuracy test on training data set. First, we supply the training data set and check the different algorithm accuracy.

Using NaiveBayes classification we get 69.67% accuracy

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose J48 -C 0.5 -M 2

Test options

☒ Use training set
☐ Supplied test set Set...
☐ Cross-validation Folds 20
☐ Percentage split % 66
More options...

(Nom) Gender

Start Stop

Result list (right-click for options)

- 16:11:41 - bayes.NaiveBayes
- 16:20:29 - bayes.NaiveBayes
- 16:21:27 - trees.J48
- 16:22:45 - trees.J48
- 16:23:31 - trees.RandomForest
- 20:25:38 - trees.J48
- 20:25:46 - trees.J48
- 20:25:58 - trees.J48
- 20:26:12 - trees.J48

Classifier output

Time taken to test model on training data: 0.03 seconds

=== Summary ===

Correctly Classified Instances	694	68.7129 %
Incorrectly Classified Instances	316	31.2871 %
Kappa statistic	0.3489	
Mean absolute error	0.3779	
Root mean squared error	0.4488	
Relative absolute error	78.1043 %	
Root relative squared error	91.2572 %	
Total Number of Instances	1010	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Weighted Avg.	0.750	0.403	0.728	0.750	0.739	0.349	0.748	0.805	female
	0.597	0.250	0.624	0.597	0.610	0.349	0.748	0.649	male

=== Confusion Matrix ===

a	b	-- classified as	
447	149	a =	female
167	247	b =	male

Now we use classification algorithm J48 decision tree to check accuracy which comes 89.97% with confidence factor=0.25.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose J48 -C 0.25 -M 2

Test options

☒ Use training set
☐ Supplied test set Set...
☐ Cross-validation Folds 20
☐ Percentage split % 66
More options...

(Nom) Gender

Start Stop

Result list (right-click for options)

- 22:10:26 - lazy.IBk
- 22:28:34 - lazy.IBk
- 22:29:07 - lazy.IBk
- 22:29:33 - trees.J48

Classifier output

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances	727	89.9752 %
Incorrectly Classified Instances	81	10.0248 %
Kappa statistic	0.7873	
Mean absolute error	0.1665	
Root mean squared error	0.2885	
Relative absolute error	34.6949 %	
Root relative squared error	58.9054 %	
Total Number of Instances	808	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Weighted Avg.	0.953	0.180	0.888	0.953	0.919	0.791	0.934	0.939	female
	0.820	0.047	0.920	0.820	0.867	0.791	0.934	0.921	male

=== Confusion Matrix ===

a	b	-- classified as	
462	23	a =	female
58	265	b =	male

For J48 decision tree classification with confidence factor =0.5 we get accuracy 93.19%.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose J48 -C 0.5 -M 2

Test options

☒ Use training set
☐ Supplied test set Set...
☐ Cross-validation Folds 20
☐ Percentage split % 66
More options...

(Nom) Gender

Start Stop

Result list (right-click for options)

- 22:10:26 - lazy.IBk
- 22:28:34 - lazy.IBk
- 22:29:07 - lazy.IBk
- 22:29:33 - trees.J48
- 22:30:17 - trees.J48

Classifier output

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances	753	93.1931 %
Incorrectly Classified Instances	55	6.8069 %
Kappa statistic	0.8573	
Mean absolute error	0.1131	
Root mean squared error	0.2378	
Relative absolute error	23.568 %	
Root relative squared error	48.5494 %	
Total Number of Instances	808	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.955	0.102	0.933	0.955	0.944	0.858	0.972	0.978	female
	0.898	0.045	0.929	0.898	0.913	0.858	0.972	0.962	male
Weighted Avg.	0.932	0.079	0.932	0.932	0.932	0.858	0.972	0.972	

=== Confusion Matrix ===

a	b	-- classified as
463	22	a = female
33	290	b = male

Using Random Forest algorithm, we get accuracy 99.75%

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Test options

☒ Use training set
☐ Supplied test set Set...
☐ Cross-validation Folds 20
☐ Percentage split % 66
More options...

(Nom) Gender

Start Stop

Result list (right-click for options)

- 22:10:26 - lazy.IBk
- 22:28:34 - lazy.IBk
- 22:29:07 - lazy.IBk
- 22:29:33 - trees.J48
- 22:30:17 - trees.J48
- 22:30:48 - trees.RandomForest

Classifier output

Time taken to test model on training data: 0.14 seconds

=== Summary ===

Correctly Classified Instances	806	99.7525 %
Incorrectly Classified Instances	2	0.2475 %
Kappa statistic	0.9948	
Mean absolute error	0.1488	
Root mean squared error	0.1702	
Relative absolute error	31.0038 %	
Root relative squared error	34.7464 %	
Total Number of Instances	808	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	FRC Area	Class
	0.998	0.003	0.998	0.998	0.998	0.995	1.000	1.000	female
	0.997	0.002	0.997	0.997	0.997	0.995	1.000	1.000	male
Weighted Avg.	0.998	0.003	0.998	0.998	0.998	0.995	1.000	1.000	

=== Confusion Matrix ===

```

a  b  <-- classified as
484  1 | a = female
  1 322 | b = male

```

For KNN Algorithm using Key=1 we get accuracy 99.75%

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A "weka.core.EuclideanDistance -R first-last"

Test options

☒ Use training set

☐ Supplied test set Set...

☐ Cross-validation Folds 20

☐ Percentage split % 66

More options...

(Nom) Gender

Start Stop

Result list (right-click for options)

- 16:23:31 - trees.RandomForest
- 20:25:38 - trees.J48
- 20:25:46 - trees.J48
- 20:25:58 - trees.J48
- 20:26:12 - trees.J48
- 21:05:54 - lazy.IBk
- 21:08:24 - lazy.IBk
- 21:09:23 - lazy.IBk
- 21:10:45 - trees.J48
- 21:10:58 - trees.J48
- 22:10:26 - lazy.IBk

Classifier output

Time taken to test model on training data: 0.39 seconds

=== Summary ===

Correctly Classified Instances	806	99.7525 %
Incorrectly Classified Instances	2	0.2475 %
Kappa statistic	0.9948	
Mean absolute error	0.0041	
Root mean squared error	0.038	
Relative absolute error	0.8574 %	
Root relative squared error	7.7613 %	
Total Number of Instances	808	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.998	0.003	0.998	0.998	0.998	0.995	1.000	1.000	female
	0.997	0.002	0.997	0.997	0.997	0.995	1.000	1.000	male
Weighted Avg.	0.998	0.003	0.998	0.998	0.998	0.995	1.000	1.000	

=== Confusion Matrix ===

a	b	<-- classified as	
484	1	a = female	
1	322	b = male	

We change and put Key=3 our accuracy comes to 80%.

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier

Choose **IBk -K 3 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A "weka.core.EuclideanDistance -R first-last"**

Test options

☒ Use training set
☐ Supplied test set Set...
☐ Cross-validation Folds 20
☐ Percentage split % 66
More options...

(Nom) Gender

Start Stop

Result list (right-click for options)

22:10:26 - lazy.IBk
22:28:34 - lazy.IBk

Classifier output

Time taken to test model on training data: 0.27 seconds

=== Summary ===

Correctly Classified Instances	651	80.5693 %
Incorrectly Classified Instances	157	19.4307 %
Kappa statistic	0.5859	
Mean absolute error	0.2751	
Root mean squared error	0.3647	
Relative absolute error	57.3172 %	
Root relative squared error	74.4494 %	
Total Number of Instances	808	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.882	0.310	0.811	0.882	0.845	0.590	0.886	0.909	female
	0.690	0.118	0.796	0.690	0.740	0.590	0.886	0.815	male
Weighted Avg.	0.806	0.233	0.805	0.806	0.803	0.590	0.886	0.871	

=== Confusion Matrix ===

a	b	<-- classified as
428	57	a = female
100	223	b = male

Changing key=5 we get accuracy 76.13%.

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier

Choose **IBk -K 5 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A "weka.core.EuclideanDistance -R first-last"**

Test options

☒ Use training set
☐ Supplied test set Set...
☐ Cross-validation Folds 20
☐ Percentage split % 66
More options...

(Nom) Gender

Start Stop

Result list (right-click for options)

22:10:26 - lazy.IBk
22:28:34 - lazy.IBk
22:29:07 - lazy.IBk

Classifier output

Time taken to test model on training data: 0.39 seconds

=== Summary ===

Correctly Classified Instances	609	75.3713 %
Incorrectly Classified Instances	199	24.6287 %
Kappa statistic	0.4723	
Mean absolute error	0.3353	
Root mean squared error	0.4028	
Relative absolute error	69.8679 %	
Root relative squared error	82.2303 %	
Total Number of Instances	808	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.849	0.390	0.766	0.849	0.805	0.477	0.834	0.875	female
	0.610	0.151	0.730	0.610	0.664	0.477	0.834	0.758	male
Weighted Avg.	0.754	0.294	0.751	0.754	0.749	0.477	0.834	0.828	

=== Confusion Matrix ===

a	b	<-- classified as
412	73	a = female
126	197	b = male

We create a graph for these different Algorithms accuracy on training data set.

4.2. Data Mining Tasks and Processes

5. Evaluations and Results

5.1. Evaluation Methods

We split our data into training and testing where we check the accuracy at 80% training and 20% testing dataset.

5.2. Results and Findings

From all our problem statements, we can see that our conclusions are as bellows:

Preferences	Male	Female
Music/Movie	Rock/Comedy	Rock/Comedy
Phobia	Snake	Dangerous Dogs
Spending Habits	Healthy Eatings	Healthy Eatings
Interests	Fun with friends	Fun with friends

6. Conclusions and Future Work

6.1. From all the problem statements we addressed, we have following conclusions:

- Females like “Latino” and “Musical” kind of music more than that of men.
- Men like to hear “Metal or Hard rock” and “Techno, Trance” music more often than that of women.
- Females watch “Romantic” and “Fantasy/Fairy tales” kind of movie more than that of men.
- Men like to watch “Action” and “War” genre movie more often than that of women.
- Females fear more of “Spiders” than that of male.
- Interestingly, male don’t fear of anything more than female but they fear “Dangerous Dogs” most.

- According to the plot, women spend more on “Shopping Centers”.
- Men spend more on “Gadgets”.
- People living in cities spend more on “Looks”.
- People living in villages spend more on “Finances”.
- Women are interested more in “Reading” and “Shopping” than that of men.
- Men are interested more in PC and Cars more than that of women.

6.2. Limitations

- Our dataset is limited to 1010 rows and we wish to have more data for further analysis.
- We wish to perform more analysis on more classification algorithms for good accuracy on training and testing datasets.
- Limited analysis on algorithms for accurate results.

6.3. Potential Improvements or Future Work

- We can implement other algorithms where we can get accurate results on classification results.
- We can use bigger dataset for deeper analysis on the dataset.

Important Notes:

1. Each team only submits a single copy to the blackboard system by a same team member. If more than 1 team members made the submissions, deduct 5 points
2. Two submissions: Report.pdf and Codes.zip. If your submissions are not in the correct format, deduct 2 points
3. You must produce your reports based on this template. If not, deduct 2 points
4. In the codes.zip, you need to include your codes if you use some programming language, such as python or R, to complete the project. If you used Weka, you should simply copy the outputs in Weka into a document, and submit them as Codes.zip
5. Your project will be rated by 4 sections: report, codes, presentation and value. The deducted points mentioned above, will be applied to the final total grade of your project. Not to the “report” section.

6. If you use the same project for ITMD 525 and 527, you need to submit a copy to each class.
7. For more details, you should refer to W15_PPT_02.pdf