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?

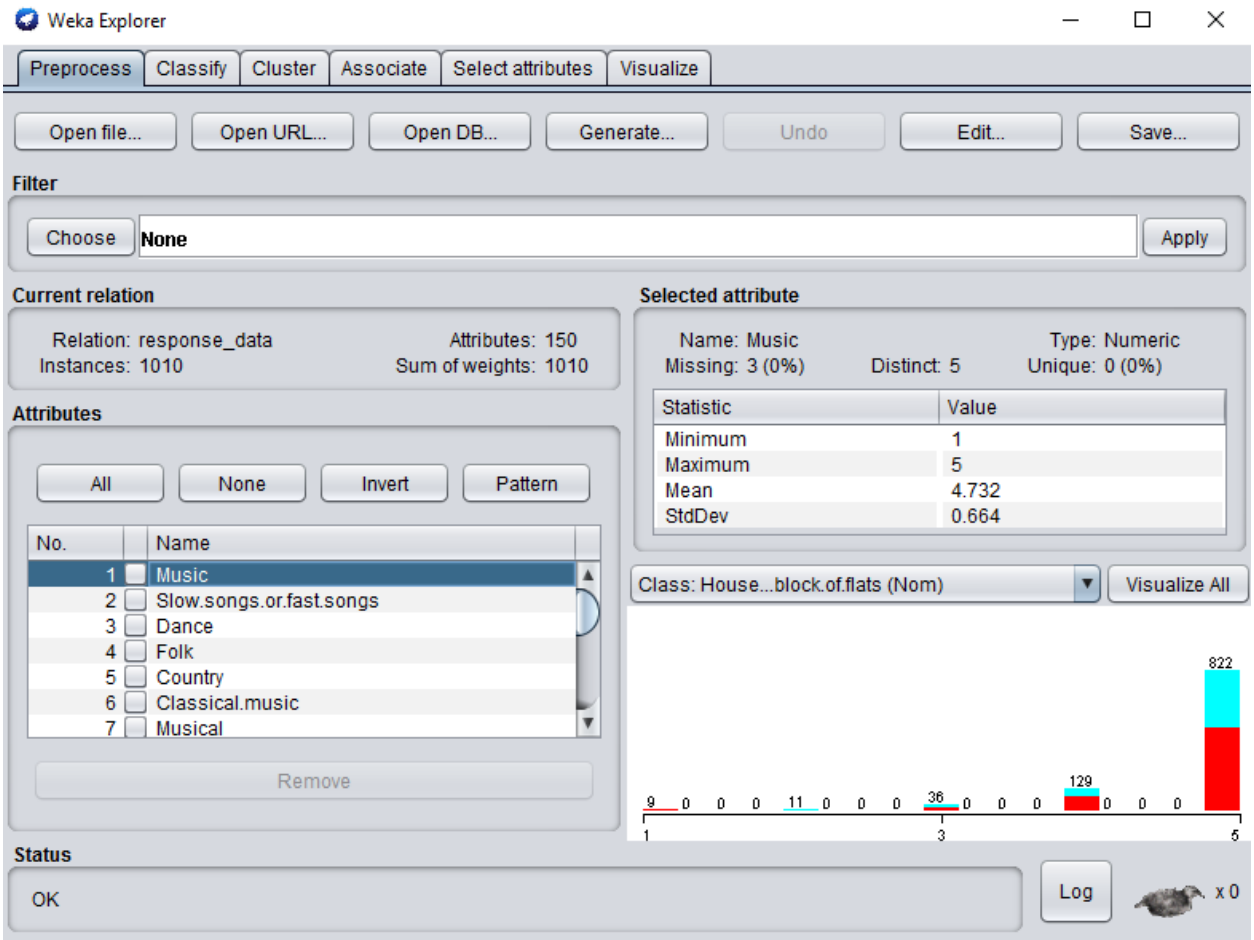
5. What can be difference in interests gender wise?

6. Is there a possibility that we can find those people from the survey who cheated and answered questions randomly?

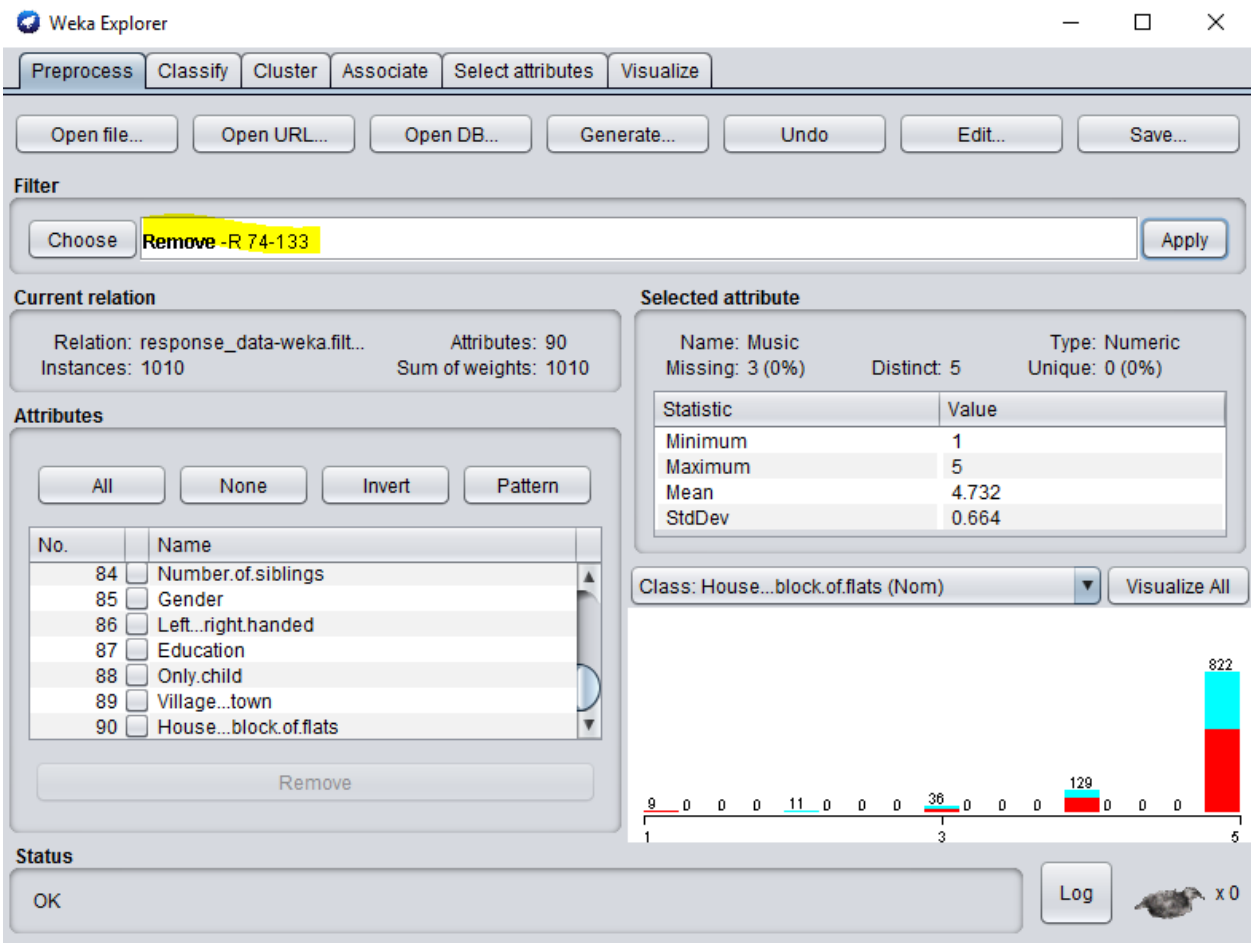
4. KDD

4.1. Data Processing

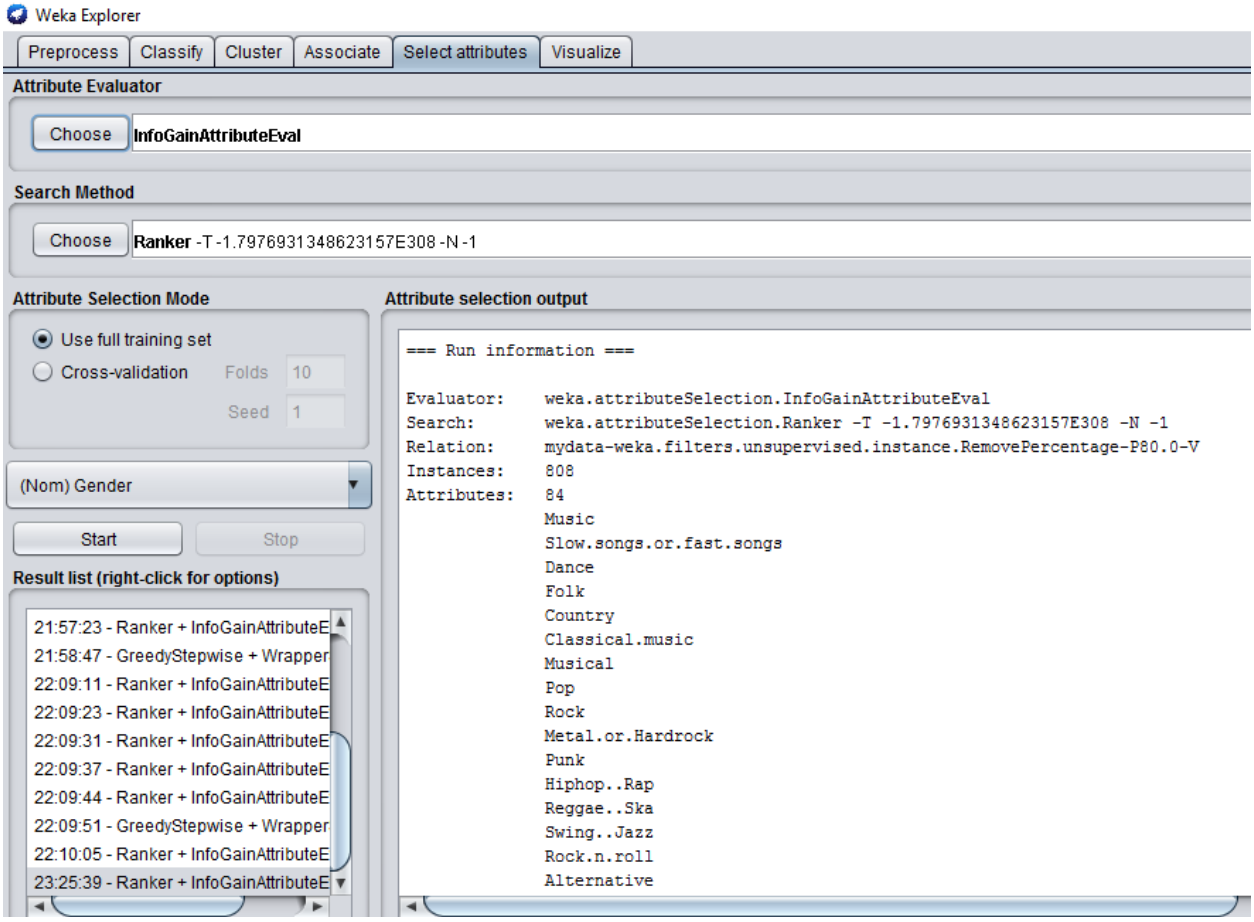
We have in all 150 columns containing 139 numerical data and 11 categorical data. Now we don't need these columns so we remove the unnecessary columns and use only those which are required for us research/mining task. We choose 84 columns which we require and these columns are related to music, movie,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



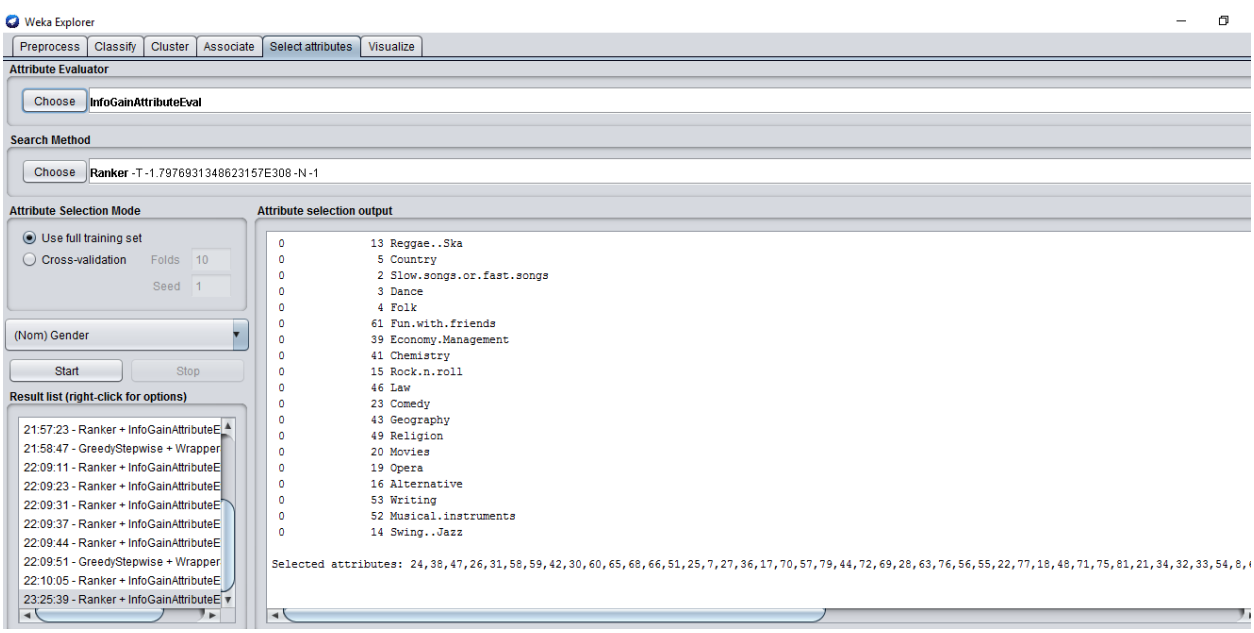
We remove unused attributes:



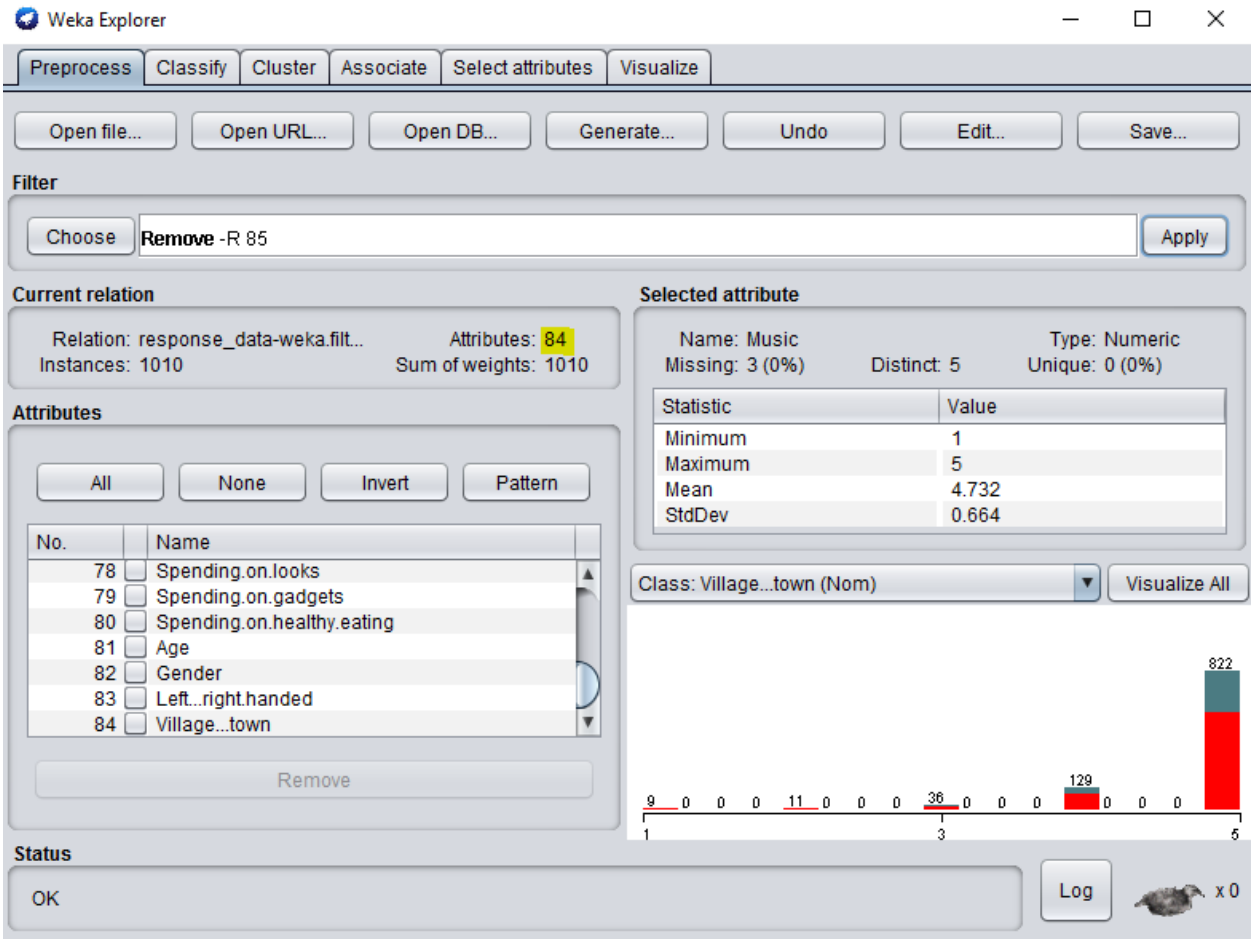
We do feature selection using Information Gain before removing the unused attributes



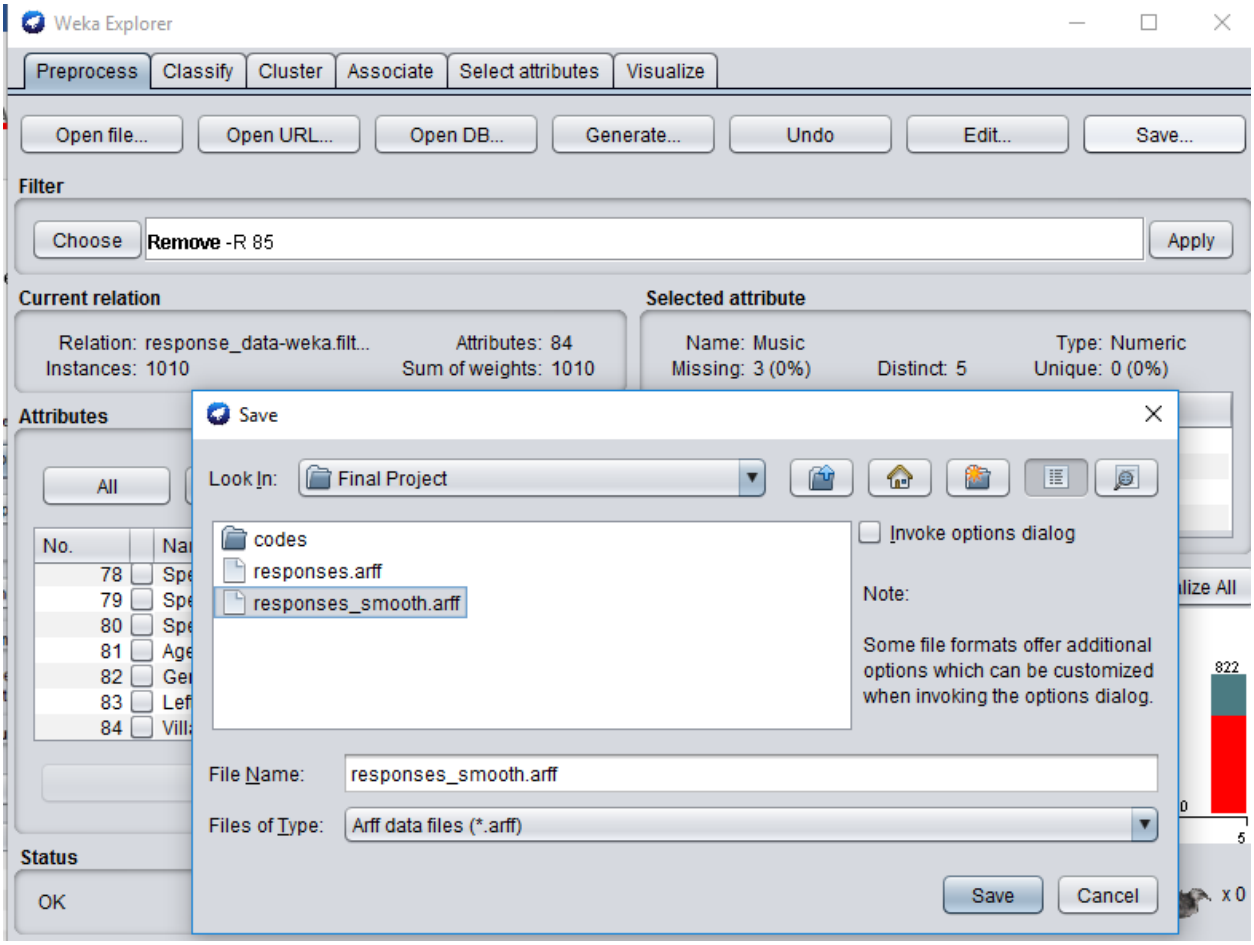
We get in total 84 attributes



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

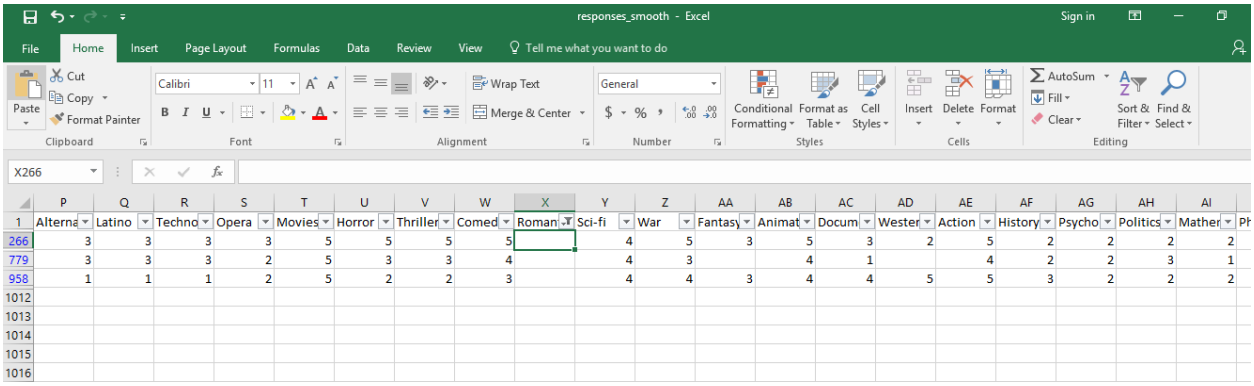


We save the smooth file as arff.



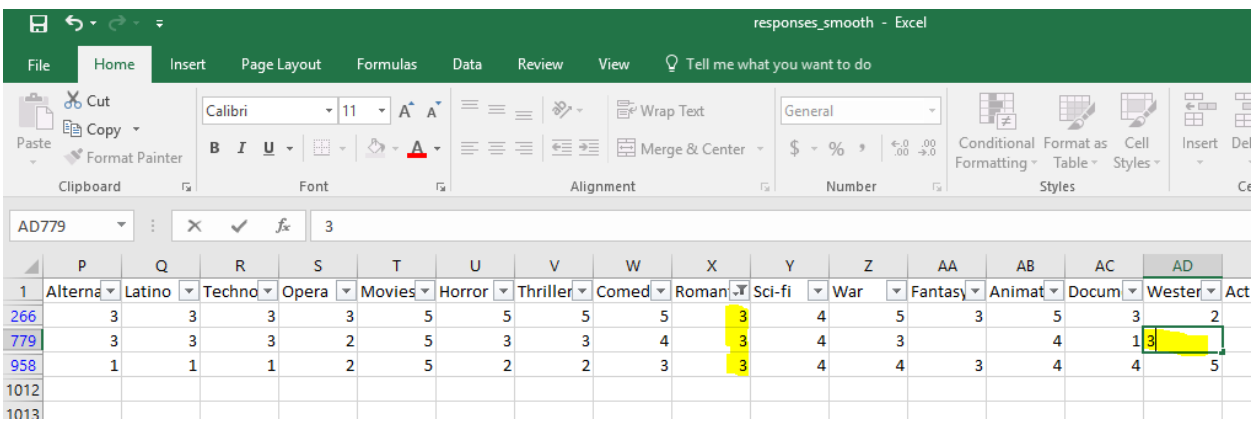
Next, we replace the missing values in the data. We replaced the numeric data by mean value because we have numeric data from range 1-5 so we replaced the missing values using value 3. Below are the steps:

Missing values are replaced by mean value:



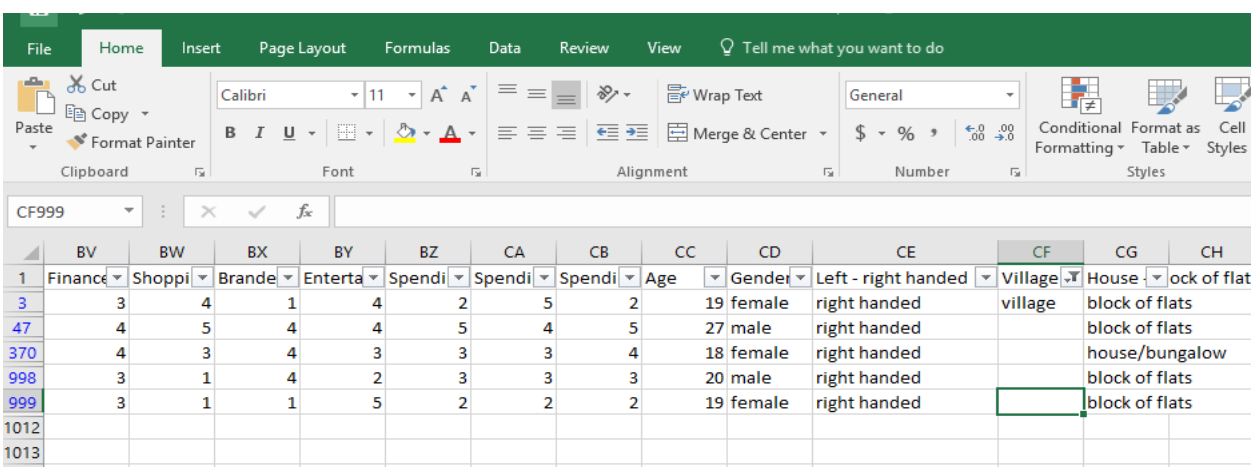
	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI
1	Alterna	Latino	Techno	Opera	Movies	Horror	Thriller	Comed	Roman	Sci-fi	War	Fantasy	Animat	Docum	Wester	Action	History	Psycho	Politics	Mather
266	3	3	3	3	5	5	5	5		4	5	3	5	3	2	5	2	2	2	2
779	3	3	3	2	5	3	3	4		4	3		4	1		4	2	2	3	1
958	1	1	1	2	5	2	2	3		4	4	3	4	4	5	5	3	2	2	2

Replacing by mean value:

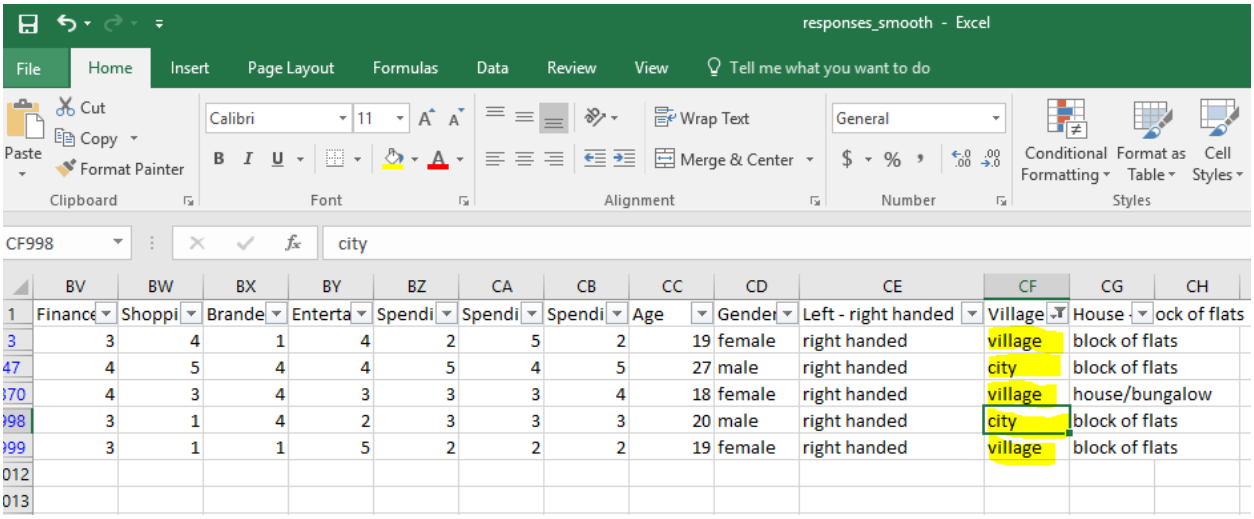


	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
1	Alterna	Latino	Techno	Opera	Movies	Horror	Thriller	Comed	Roman	Sci-fi	War	Fantasy	Animat	Docum	Wester
266	3	3	3	3	5	5	5	5	3	4	5	3	5	3	2
779	3	3	3	2	5	3	3	4	3	4	3		4	1	3
958	1	1	1	2	5	2	2	3	3	4	4	3	4	4	5

Replacing missing nominal values with both responses (i.e. village & city)



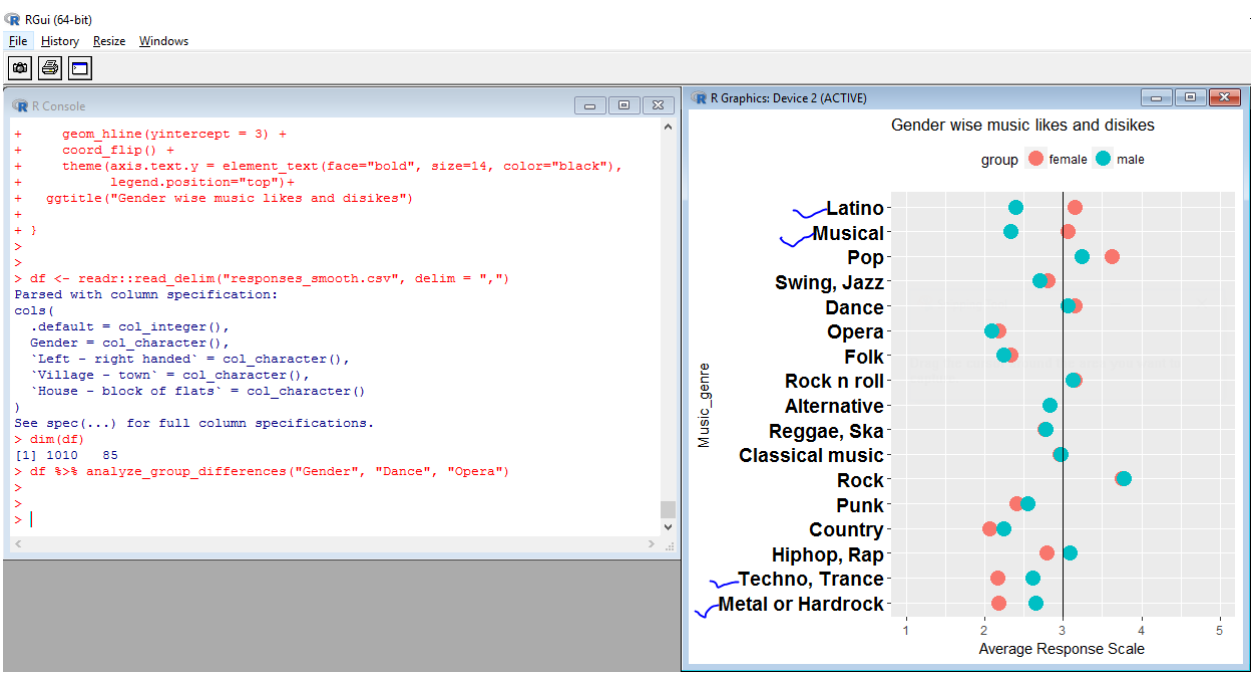
	BV	BW	BX	BY	BZ	CA	CB	CC	CD	CE	CF	CG	CH
1	Finance	Shoppi	Brande	Enterta	Spendi	Spendi	Spendi	Age	Gender	Left - right handed	Village	House -	ock of flat
3	3	4	1	4	2	5	2	19	female	right handed	village	block of flats	
47	4	5	4	4	5	4	5	27	male	right handed		block of flats	
370	4	3	4	3	3	3	4	18	female	right handed		house/bungalow	
998	3	1	4	2	3	3	3	20	male	right handed		block of flats	
999	3	1	1	5	2	2	2	19	female	right handed		block of flats	



	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 -	ock 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	
370	4	3	4	3	3	3	4	18	female	right handed	village	house/bungalow	
998	3	1	4	2	3	3	3	20	male	right handed	city	block of flats	
999	3	1	1	5	2	2	2	19	female	right handed	village	block of flats	
012													
013													

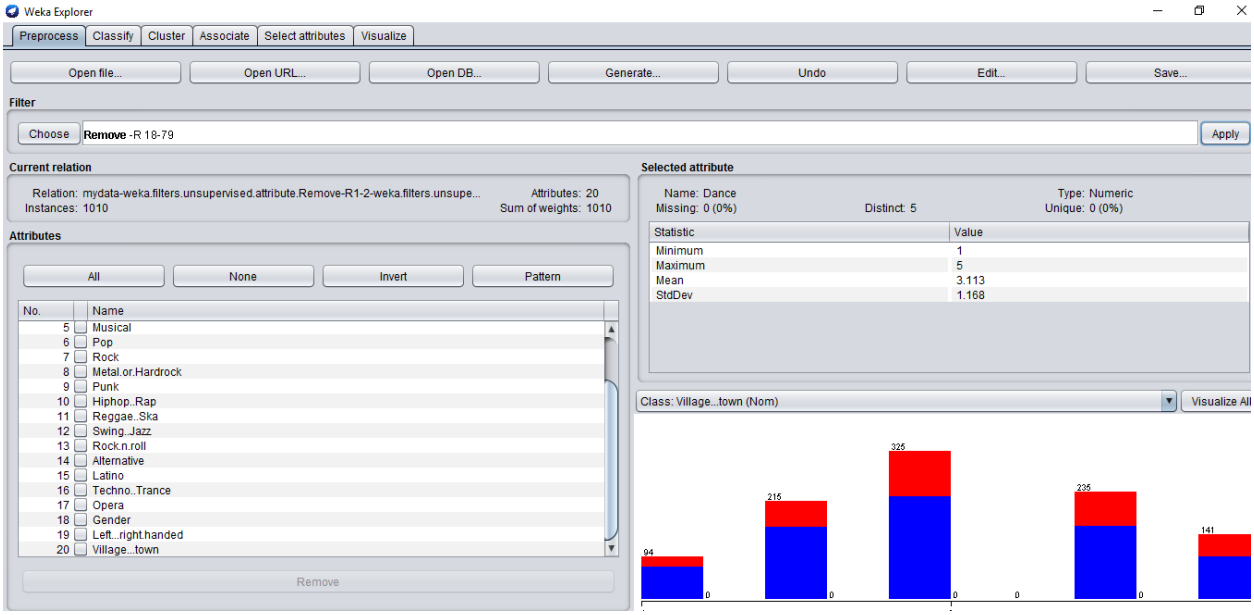
Plotting a ggplot for male and female data to verify what kind of music they like.

- 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.

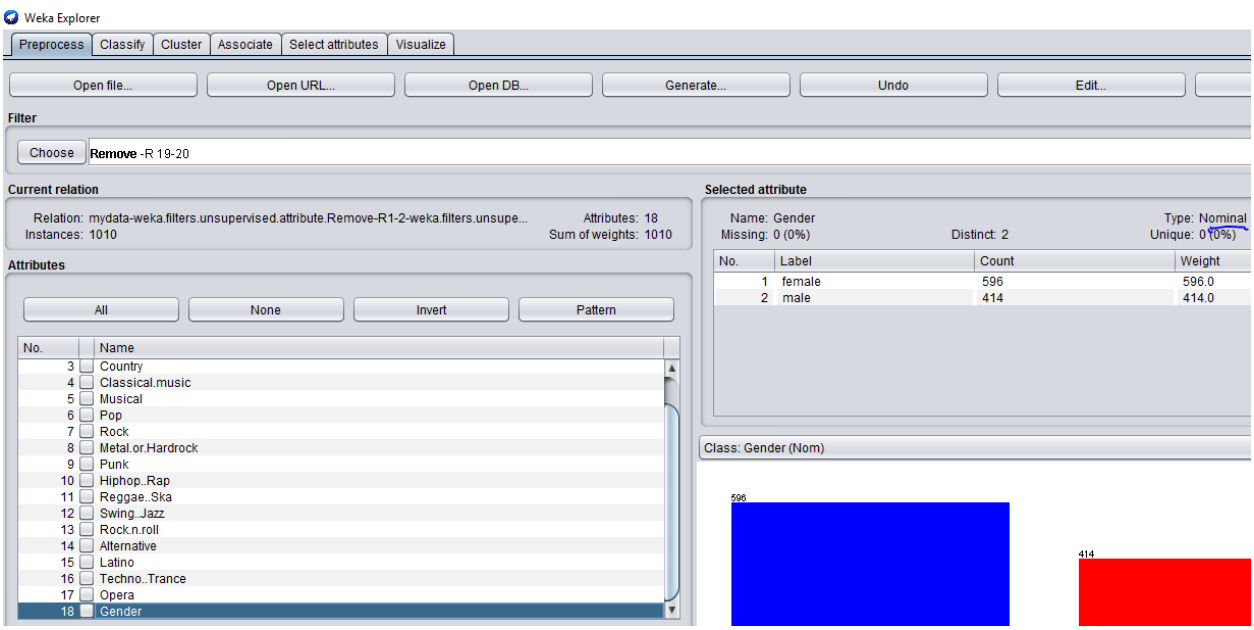


We keep only attributes related to music and keep gender data.

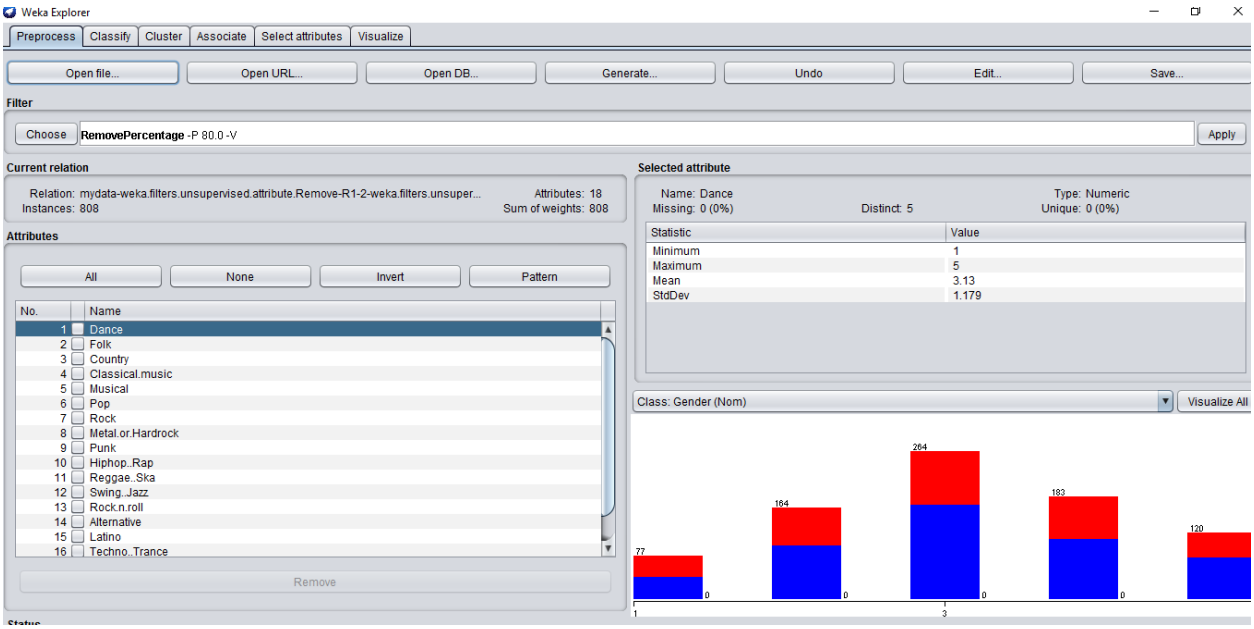
Removing remaining attributes using filter "Remove"



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

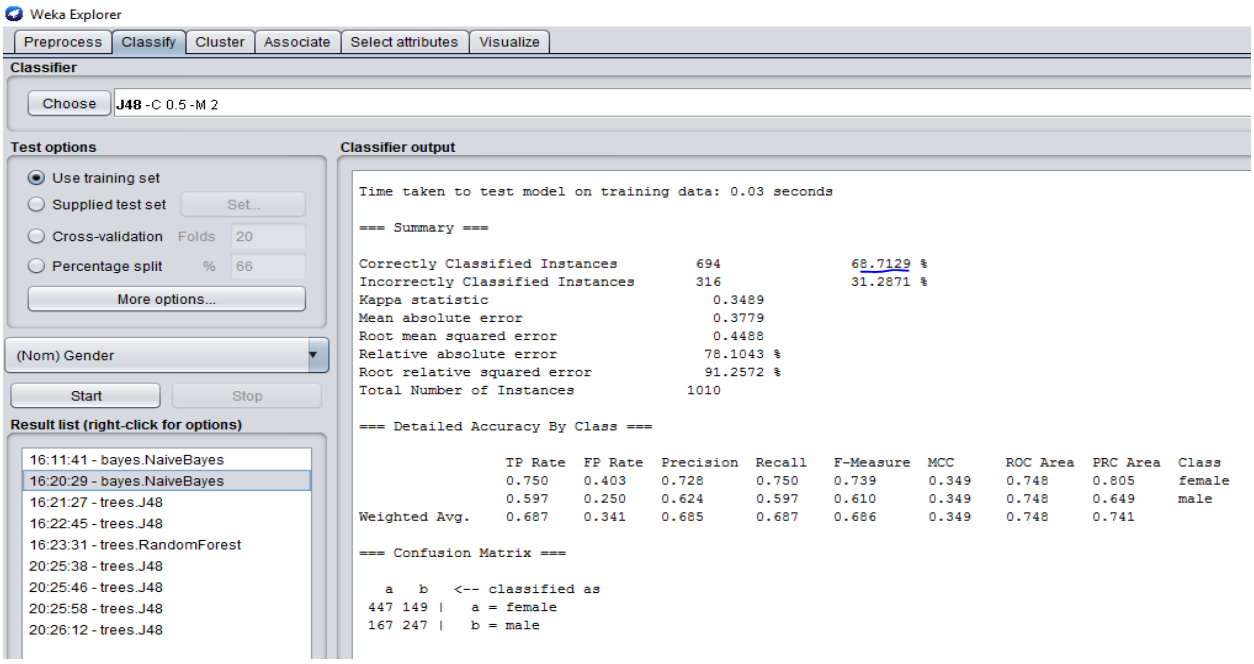


We divide our data into training and testing sets and perform classification.

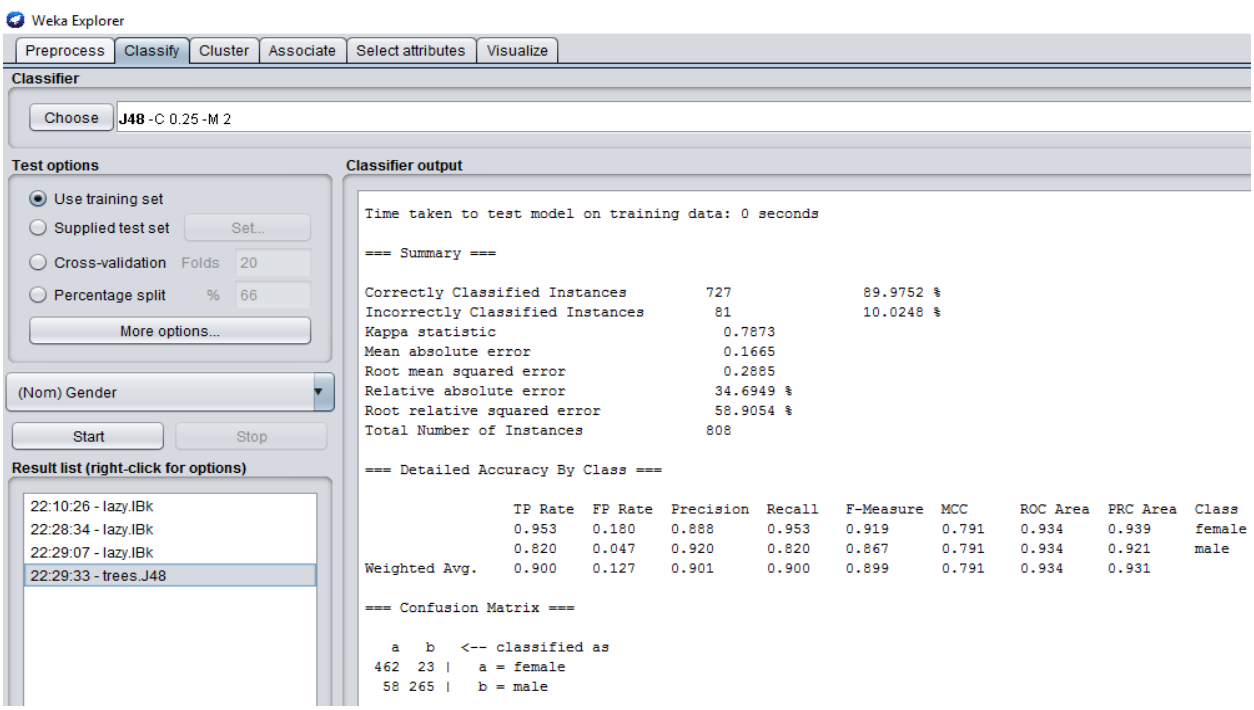


We perform accuracy test on training data set. First, we supply the training data set and check the different algorithm accuracy.

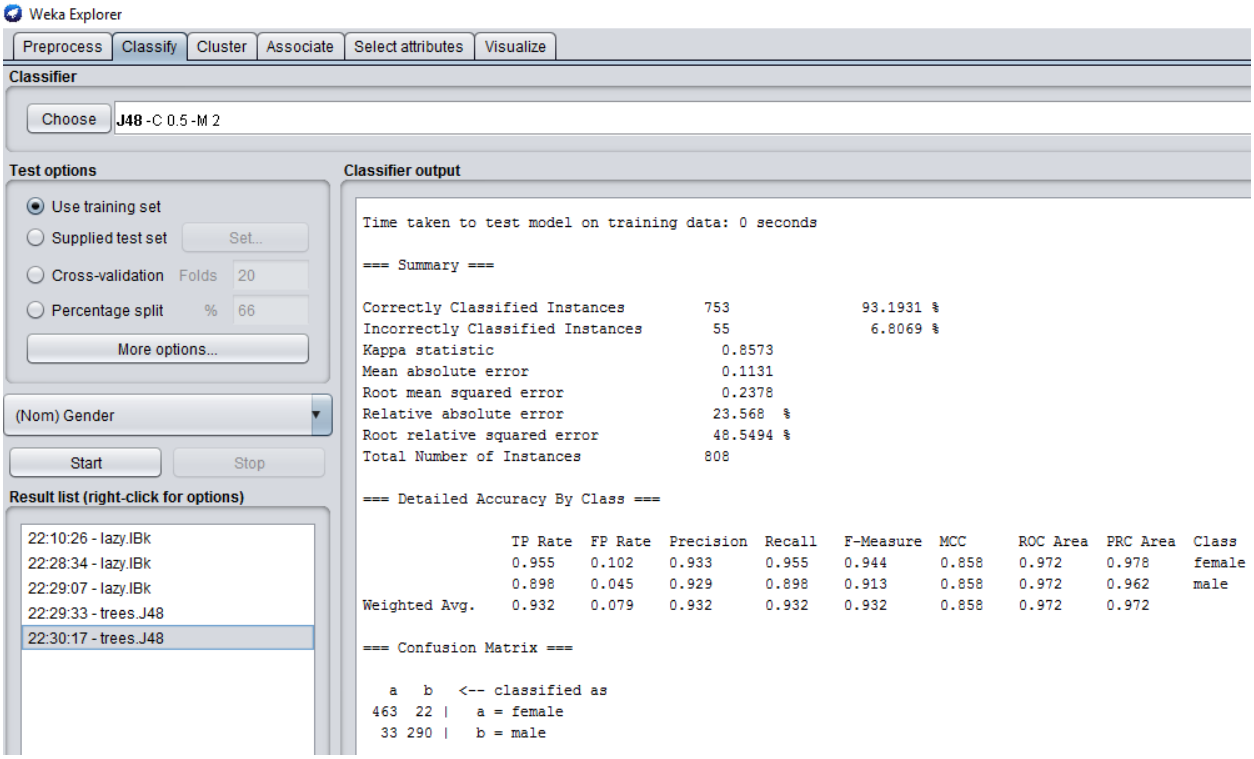
Using NaiveBayes classification we get 69.67% accuracy



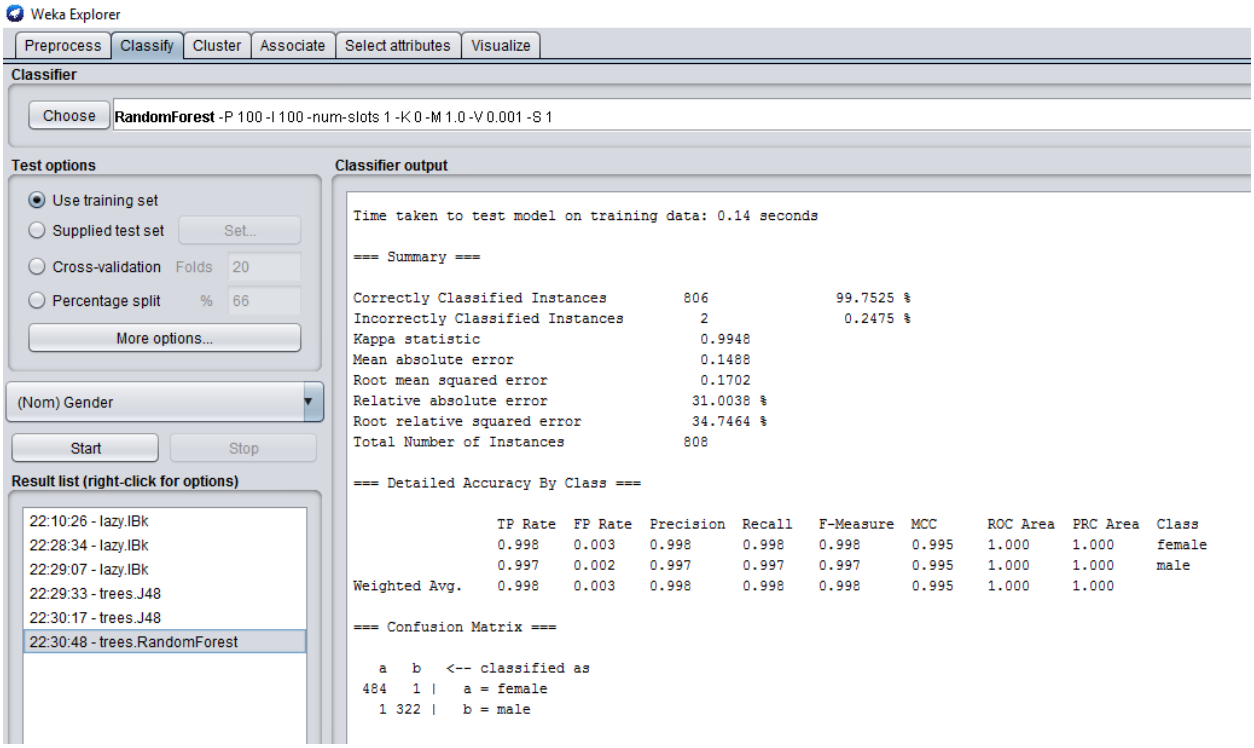
Now we use classification algorithm J48 decision tree to check accuracy which comes 89.97% with confidence factor=0.25.



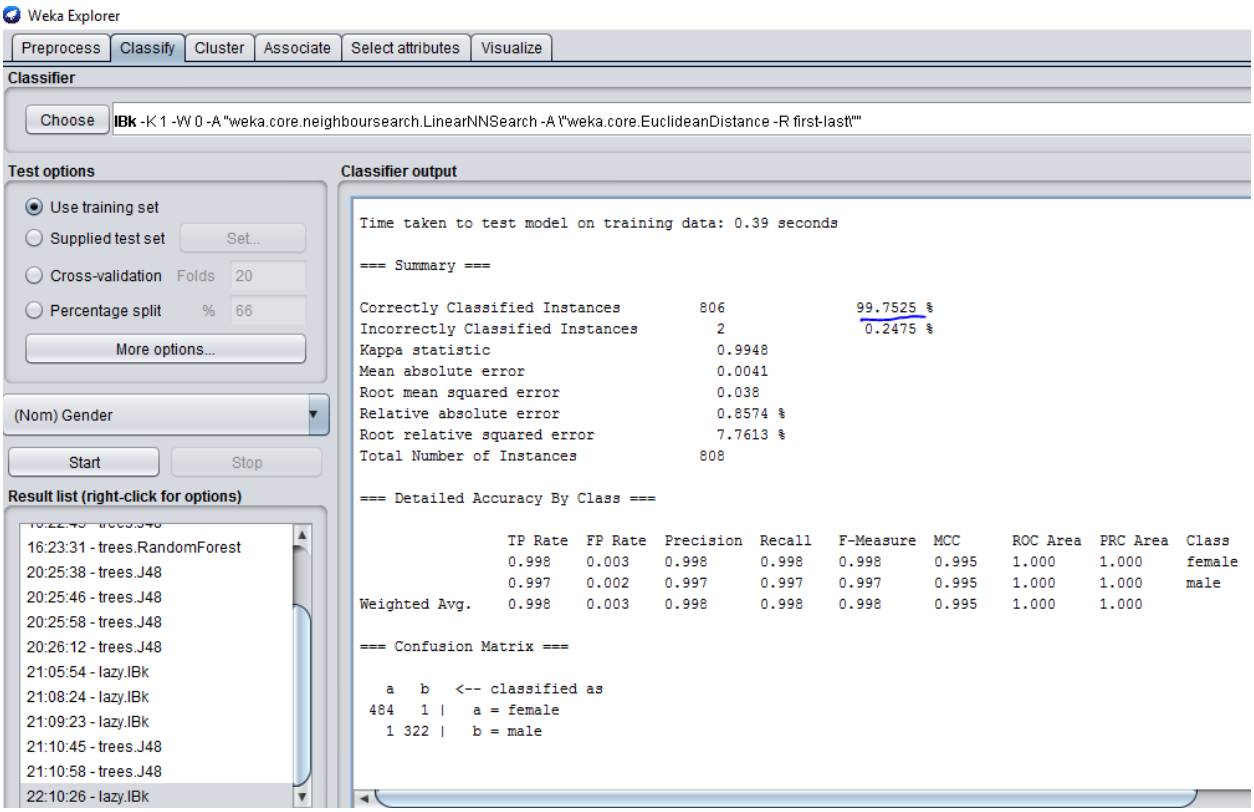
For J48 decision tree classification with confidence factor =0.5 we get accuracy 93.19%.



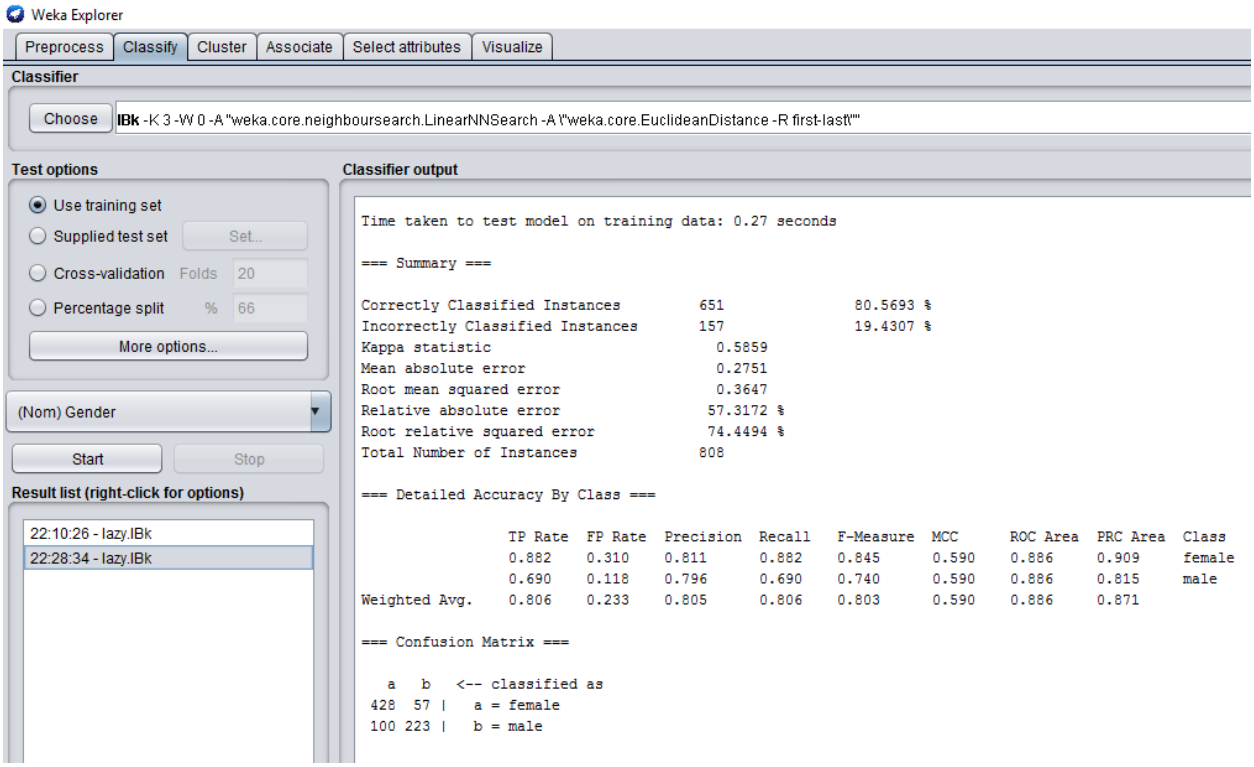
Using Random Forest algorithm, we get accuracy 99.75%



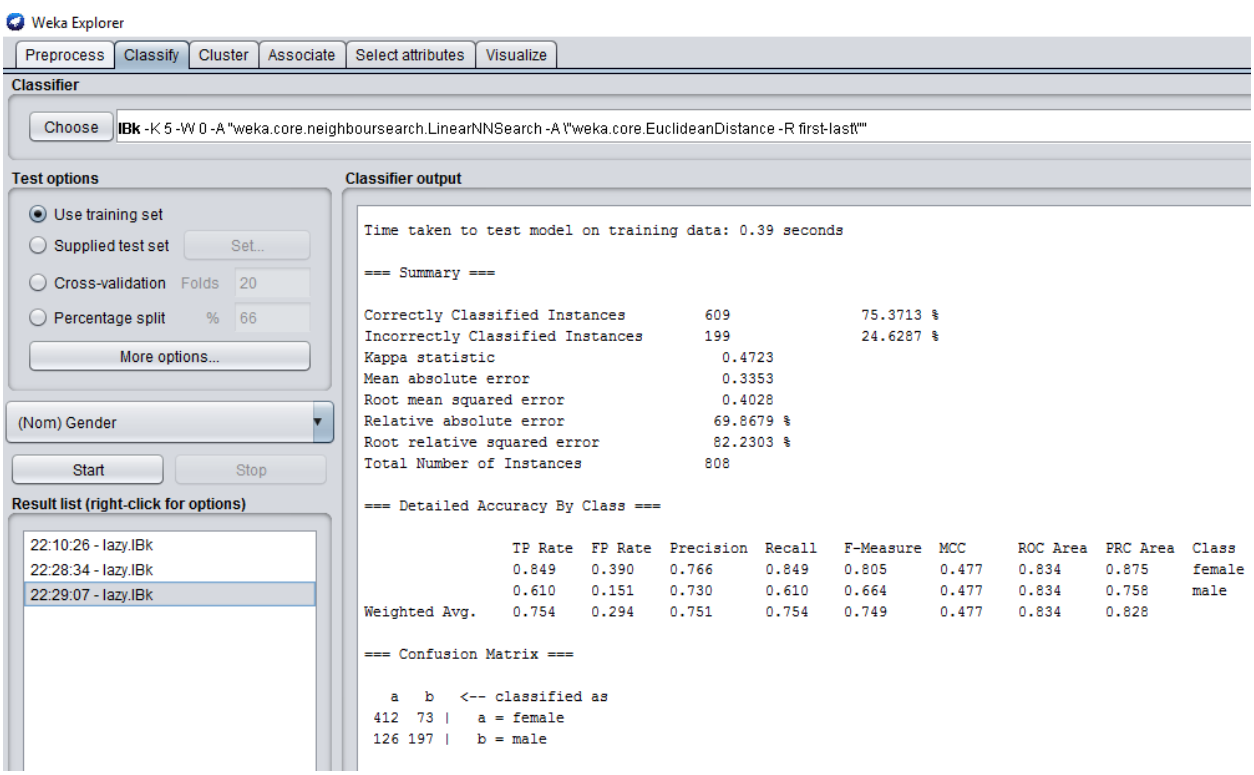
For KNN Algorithm using Key=1 we get accuracy 99.75%



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



Changing key=5 we get accuracy 76.13%.



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 ca 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 you 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