Final Project Report

CS699 DATA MINING

Team Members: Alisha Peermohamed | Xi You

SPRING 2019 | MARCH 27 2019

CS699 Final Project Report

Team Members: Alisha Peermohamed, Xi You

I. Project Proposal

Description: Our selected dataset is based on the survey results taken by Americans to gauge their favorite global cuisine. The survey asks respondants questions regarding their preference of various cuisines by asking them to rate the food in preference on a scale of 1 to 5 or N/A. Here, N/A represents: 'I'm unfamiliar with this country's traditional cuisine'. 1 represents: 'I hate this country's traditional cuisine. I think it's one of the worst in the world', 2 represents: 'I dislike this country's traditional cuisine. I think it's considerably below average.', 3 represents: 'I'm OK with this county's traditional cuisine. I think it's about average.', 4 represents: 'I like this country's traditional cuisine. I think it's considerably above average.', and lastly 5 represents: 'I love this country's traditional cuisine. I think it's one of the best in the world.'. Cuisines include Arabic, Greek, Indian, Chinese, Ethiopian and many more (Please see below for a complete list of attributes). In addition, the dataset also asks respondents various demographic questions such as their age, household income, education level, and location. Please see original dataset file: *Original_Project_Data.arff* attached.

Our Dataset: 'America's Favorite Global Cuisine':

- The original dataset has a total of 49 attributes and 1373 tuples.
- Attributes:
    - Respondents preferences on the following countries (1 attribute per country):
        - Algeria
        - Argentine
        - Australia
        - Belgium
        - Herzogovina
        - Brazil
        - Cameroon
        - Chile
        - Colombia
        - Costa Rica
        - Croatia
        - Ecuador
        - England
        - France
        - Germany
        - Ghana
        - Greece
        - Honduras
        - Iran
        - Italy
        - Ivory Coast
        - Japan
        - Mexico
        - Netherlands
        - Nigeria
        - Portugal
        - Russia
        - South Korea
        - Spain
        - Switzerland
        - United States
        - Urugauy
        - China
        - India
        - Thailand
        - Turkey
        - Cuba
        - Ethiopia
        - Vietnam
        - Ireland

- Demographic information about the Respondent:
  - Gender (Male or Female)
  - Age Bracket (18 – 29, 30 – 44, 45 – 60, or >60)
  - Household Income Bracket
    - $0 – $24,999, $25, 000 - $49,999, $50,000 - $99,999, $100,000 - $149,999, or $150,000+
  - Education
  - Location (Census Region)

Project data-mining goal: Predicting a person's household income
Our main goal for this data mining project is to develop an algorithm that will be able to classify a person based on his/her favorite cuisines, age, education, and location and be able to predict their household income level.

Class Attribute: Household Income Bracket

Through our data mining project, we will be using Weka as our main software tool for Data Preprocessing, Attribute Selection, building our classification models, as well as testing our classification models.

II. Data Preprocessing Phase:

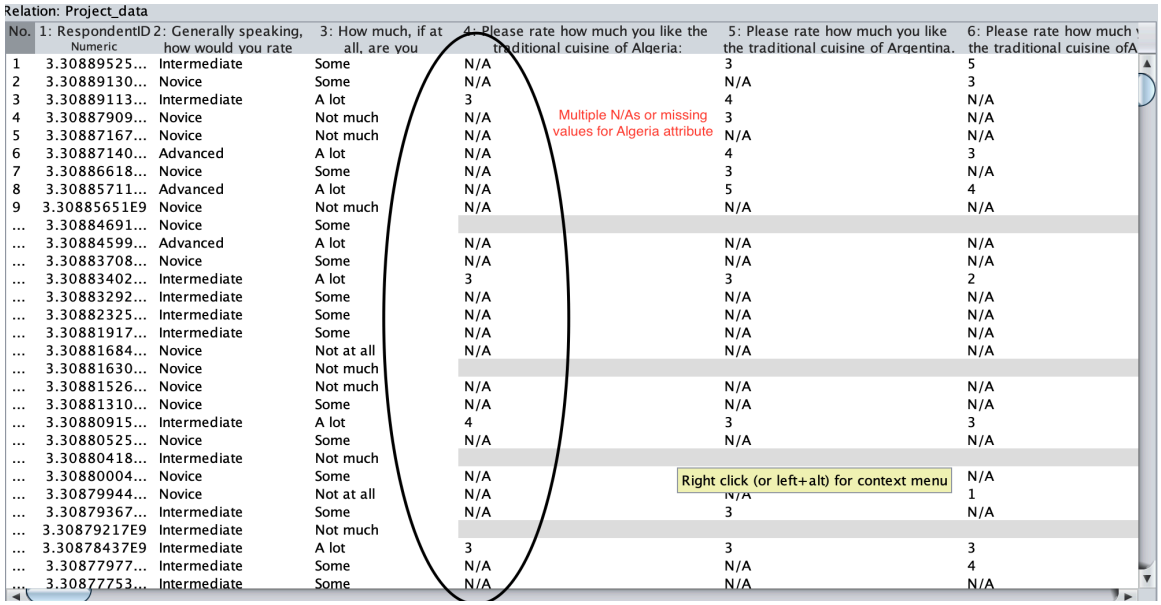
Data Reduction:

- Attribute reduction: We first noticed that some cuisine attributes had several missing values or N/A values. Please see Figure 1 below as an example. Here, the Algeria attribute has several N/A and missing values. These values are not useful to us as they do not indicate an accurate measure of respondents' appeal to a certain cuisine and would distort the model classification. So, we analyzed the number of N/A or missing values for each country attribute and decided to remove all attributes which had 500 or more N/A or missing values. After removal, our dataset had 21 attributes, of these, the dataset includes the following country attributes: England, France, Germany, Greece, Italy, Japan, Mexico, Spain, US, China, India, Thailand, Ireland.

  - In addition, we felt that different location regions had different costs of living and have a significant correlation with household income. We felt this would skew our classification results because we are looking to predict a person's household income based on their cuisine preferences, not their geographical region. For this reason, we decided to remove the location attribute as well.

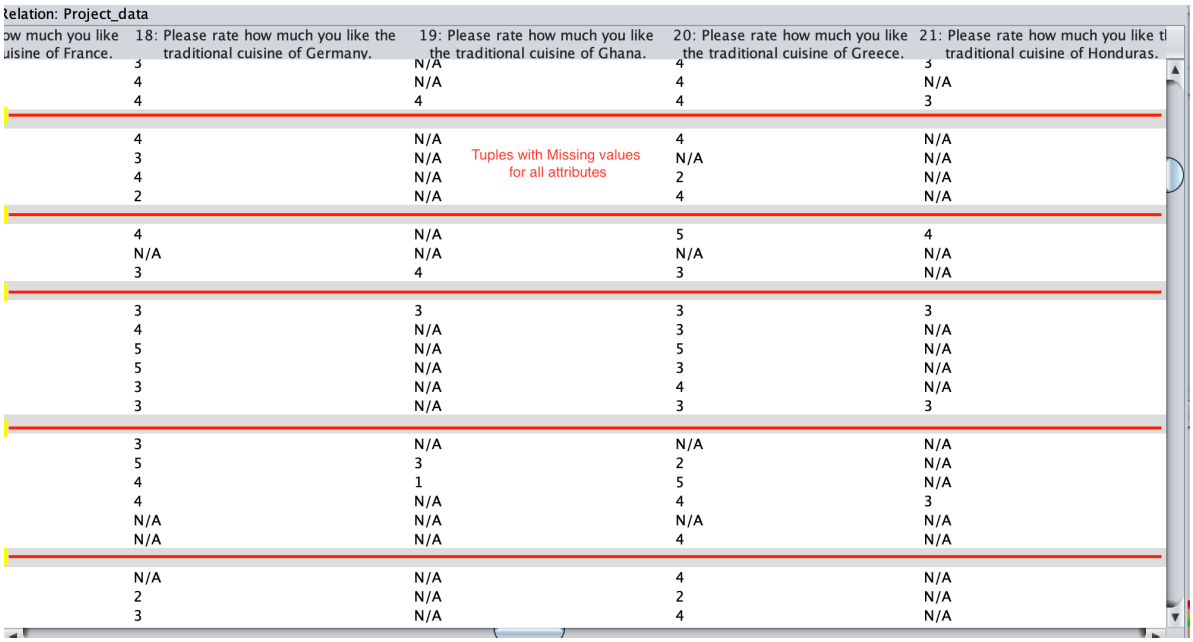
- Tuple Reduction: We also noticed several tuples with all/many missing values, as shown in Figure 2. This may have been because a respondent had started the survey but abandoned it early on in the survey. Again, this is not useful to our classification model as the missing values do not provide any information on the respondent's cuisine preferences. For the purposes of our classification model, we decided to remove these tuples from our dataset. After tuple reduction, our dataset had 1241 tuples and 20 attributes (129 tuples removed).

Figure 1: Attribute Reduction



No.	1: RespondentID Numeric	2: Generally speaking, how would you rate	3: How much, if at all, are you	4: Please rate how much you like the traditional cuisine of Algeria:	5: Please rate how much you like the traditional cuisine of Argentina.	6: Please rate how much the traditional cuisine ofA
1	3.30889525...	Intermediate	Some	N/A	3	5
2	3.30889130...	Novice	Some	N/A	N/A	3
3	3.30889113...	Intermediate	A lot	3	4	N/A
4	3.30887909...	Novice	Not much	N/A	3	N/A
5	3.30887167...	Novice	Not much	N/A	N/A	N/A
6	3.30887140...	Advanced	A lot	N/A	4	3
7	3.30886618...	Novice	Some	N/A	3	N/A
8	3.30885711...	Advanced	A lot	N/A	5	4
9	3.30885651E9	Novice	Not much	N/A	N/A	N/A
...	3.30884691...	Novice	Some			
...	3.30884599...	Advanced	A lot	N/A	N/A	N/A
...	3.30883708...	Novice	Some	N/A	N/A	N/A
...	3.30883402...	Intermediate	A lot	3	3	2
...	3.30883292...	Intermediate	Some	N/A	N/A	N/A
...	3.30882325...	Intermediate	Some	N/A	N/A	N/A
...	3.30881917...	Intermediate	Some	N/A	N/A	N/A
...	3.30881684...	Novice	Not at all	N/A	N/A	N/A
...	3.30881630...	Novice	Not much			
...	3.30881526...	Novice	Not much	N/A	N/A	N/A
...	3.30881310...	Novice	Some	N/A	N/A	N/A
...	3.30880915...	Intermediate	A lot	4	3	3
...	3.30880525...	Novice	Some	N/A	N/A	N/A
...	3.30880418...	Intermediate	Not much			
...	3.30880004...	Novice	Some	N/A		N/A
...	3.30879944...	Novice	Not at all	N/A		1
...	3.30879367...	Intermediate	Some	N/A	3	N/A
...	3.30879217E9	Intermediate	Not much			
...	3.30878437E9	Intermediate	A lot	3	3	3
...	3.30877977...	Intermediate	Some	N/A	N/A	4
...	3.30877753...	Intermediate	Some	N/A	N/A	N/A

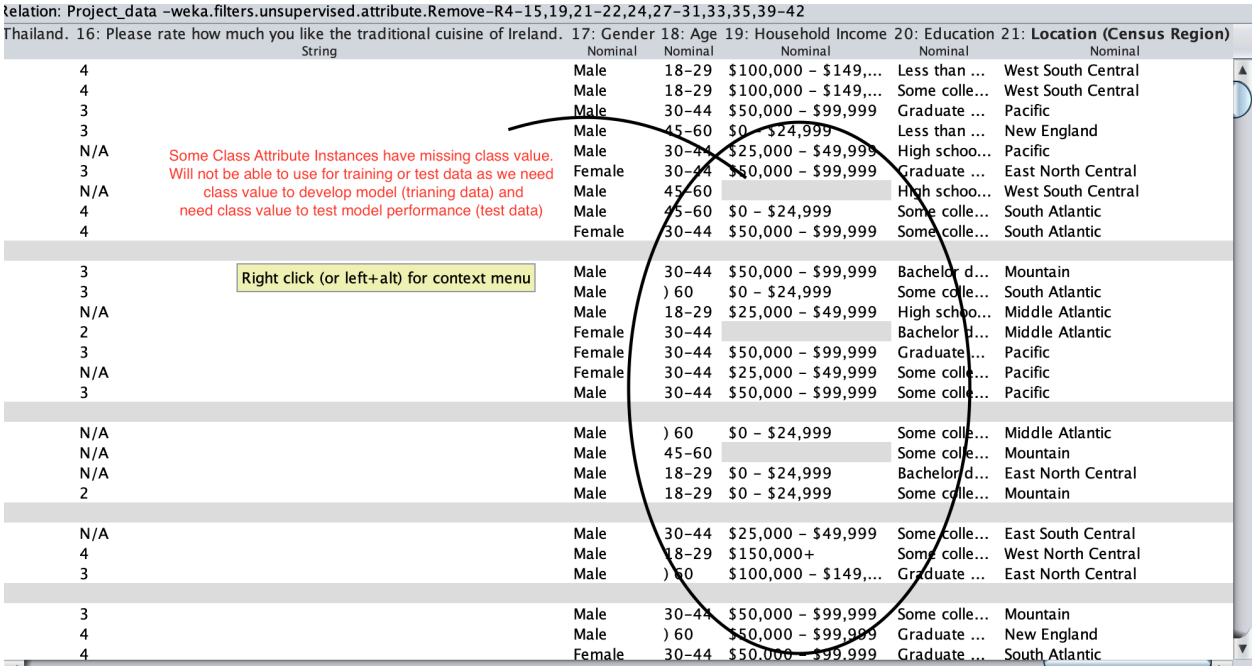
Figure 2: Tuple Reduction



ow much you like uisine of France.	18: Please rate how much you like the traditional cuisine of Germany.	19: Please rate how much you like the traditional cuisine of Ghana.	20: Please rate how much you like the traditional cuisine of Greece.	21: Please rate how much you like tl traditional cuisine of Honduras.
	3	N/A	4	3
	4	N/A	4	N/A
	4	4	4	3
	4	N/A	4	N/A
	3	N/A	N/A	N/A
	4	N/A	2	N/A
	2	N/A	4	N/A
	4	N/A	5	4
	N/A	N/A	N/A	N/A
	3	4	3	N/A
	3	3	3	3
	4	N/A	3	N/A
	5	N/A	5	N/A
	5	N/A	3	N/A
	3	N/A	4	N/A
	3	N/A	3	3
	3	N/A	N/A	N/A
	5	3	2	N/A
	4	1	5	N/A
	4	N/A	4	3
	N/A	N/A	N/A	N/A
	N/A	N/A	4	N/A
	N/A	N/A	4	N/A
	2	N/A	2	N/A
	3	N/A	4	N/A

- Missing Class Value tuple reduction: Our dataset also had several tuples with missing values in the class attribute. Please see Figure 3 for reference. These tuples are not useful to us because we would need the actual class value in the training set to develop the classification model and we would need the actual class value in the test set to evaluate the performance of the dataset. Thus, we decided to remove all tuples with missing values in the class attribute as well. Our final dataset now has 20 attributes and 954 tuples (287 tuples removed).

Figure 3: Missing value in Class Attribute



Relation: Project_data -weka.filters.unsupervised.attribute.Remove-R4-15,19,21-22,24,27-31,33,35,39-42

Thailand.	16: Please rate how much you like the traditional cuisine of Ireland.	17: Gender	18: Age	19: Household Income	20: Education	21: **Location (Census Region)**
	String	Nominal	Nominal	Nominal	Nominal	Nominal
4		Male	18-29	$100,000 - $149,...	Less than ...	West South Central
4		Male	18-29	$100,000 - $149,...	Some colle...	West South Central
3		Male	30-44	$50,000 - $99,999	Graduate ...	Pacific
3		Male	45-60	$0 - $24,999	Less than ...	New England
N/A		Male	30-44	$25,000 - $49,999	High schoo...	Pacific
3		Female	30-44	$50,000 - $99,999	Graduate ...	East North Central
N/A		Male	45-60		High schoo...	West South Central
4		Male	45-60	$0 - $24,999	Some colle...	South Atlantic
4		Female	30-44	$50,000 - $99,999	Some colle...	South Atlantic
3		Male	30-44	$50,000 - $99,999	Bachelor d...	Mountain
3		Male	) 60	$0 - $24,999	Some colle...	South Atlantic
N/A		Male	18-29	$25,000 - $49,999	High schoo...	Middle Atlantic
2		Female	30-44		Bachelor d...	Middle Atlantic
3		Female	30-44	$50,000 - $99,999	Graduate ...	Pacific
N/A		Female	30-44	$25,000 - $49,999	Some colle...	Pacific
3		Male	30-44	$50,000 - $99,999	Some colle...	Pacific
N/A		Male	) 60	$0 - $24,999	Some colle...	Middle Atlantic
N/A		Male	45-60		Some colle...	Mountain
N/A		Male	18-29	$0 - $24,999	Bachelor d...	East North Central
2		Male	18-29	$0 - $24,999	Some colle...	Mountain
N/A		Male	30-44	$25,000 - $49,999	Some colle...	East South Central
4		Male	18-29	$150,000+	Some colle...	West North Central
3		Male	) 60	$100,000 - $149,...	Graduate ...	East North Central
3		Male	30-44	$50,000 - $99,999	Some colle...	Mountain
4		Male	) 60	$50,000 - $99,999	Graduate ...	New England
4		Female	30-44	$50,000 - $99,999	Graduate ...	South Atlantic

Data Cleaning:

- Attribute name formatting: For the sake of simplicity, we also named each of our country attributes from 'Please rate how much you like the traditional cuisine of <country>.' to that country's name. This way, it is easier and faster to analyze the data.

- Formatting attribute types: We also found that various attributes were incorrectly categorized on Weka. For example, the dataset showed values in Numeric form for each country attribute. However, when it comes to replacing missing values for each attribute, we wanted to replace with that attribute's mode, not its mean. We did not want to be replacing with float values as respondents could not respond in float values during the survey. Thus, we decided to convert all country attributes to Nominal using the NumerictoNominal filter. Please see Figure 4 for reference.

- Replace Missing Values for attributes: While the data reductions described above eliminated several missing values, we decided to replace the remaining missing values and N/A's with that attribute's mode. Please see Figure 5 for the latest version of the dataset. Our complete latest dataset file: *Latest_Project_Dataset.arff* is also attached.

Lastly, we divided our complete dataset into training and test sets using a 66/33 ratio split. We used the 'Resample' filter to preserve the class distribution. Please refer to the following files as the training set and the test set, respectively: *Latest_Project_Dataset_training.arff* (training set), *Latest_Project_Dataset_Test.arff* (test set).

Figure 4: NumerictoNominal Attribute Filter

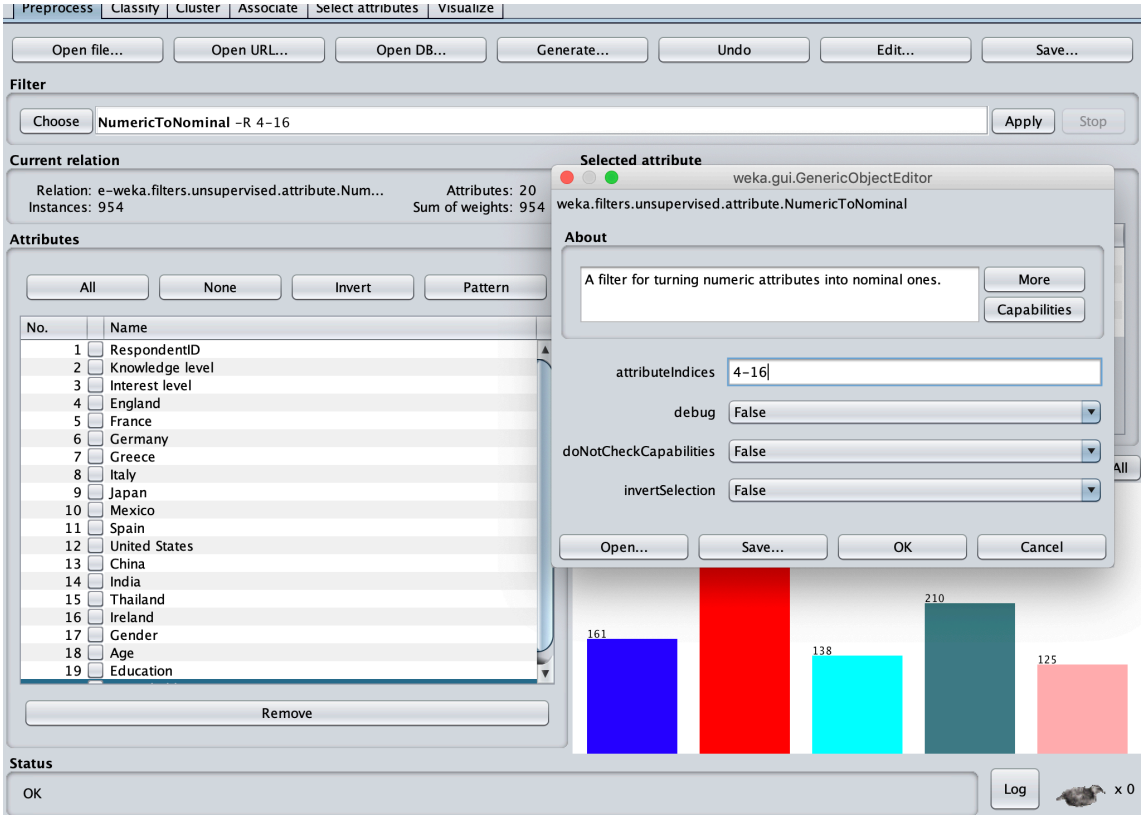


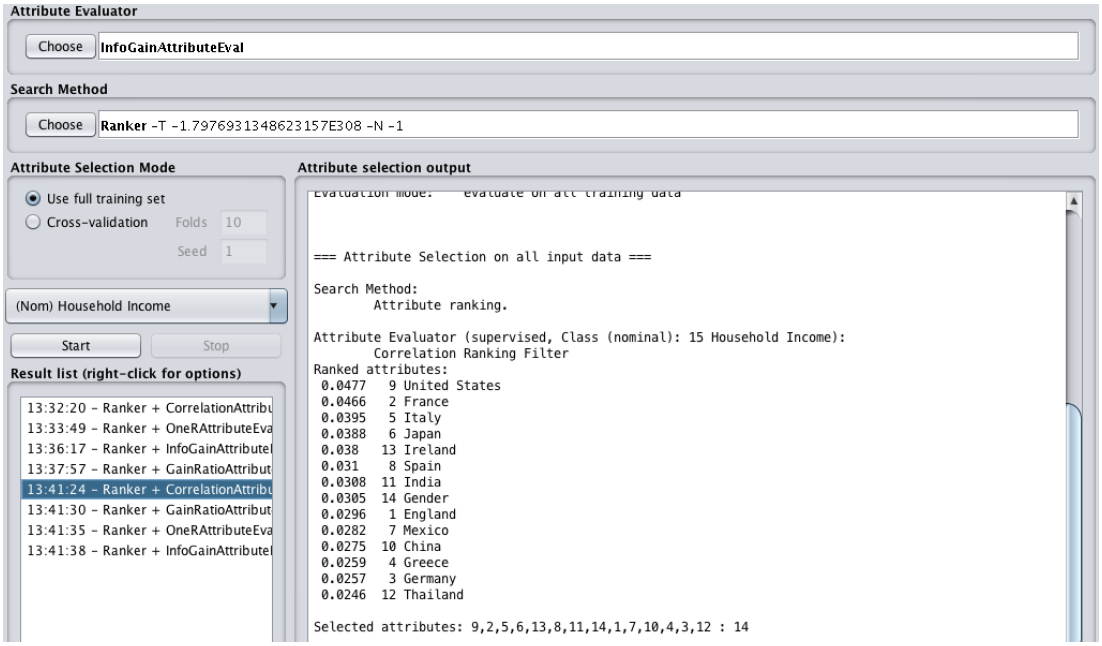
Figure 5: Dataset with Replaced Missing Values

2: Knowledge level	3: Interest level	4: England	5: France	6: Germany	7: Greece	8: Italy	9: Japan	10: Mexico	11: Spain	12: United States	13: China	14: India	15: Thailand	16: Ireland	17: Gender	18: Age	19: Education	20: **Household Incom**
Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
Intermediate	Some	4	3	5	4	5	1	5	4	5	4	5	4	4	Male	18-29	Less than ...	$100,000 - $149,99
Novice	Some	4	4	4	4	5	3	5	3	4	2	2	4	4	Male	18-29	Some colle...	$100,000 - $149,99
Intermediate	A lot	3	4	2	4	5	3	4	4	3	3	5	5	3	Male	30-44	Graduate ...	$50,000 - $99,999
Novice	Not much	3	3	3	4	3	3	4	3	3	4	3	2	3	Male	45-60	Less than ...	$0 - $24,999
Novice	Not much	2	3	3	4	4	2	4	3	3	3	3	3	3	Male	30-44	High schoo...	$25,000 - $49,999
Advanced	A lot	5	5	5	5	5	5	5	5	2	3	5	5	3	Female	30-44	Graduate ...	$50,000 - $99,999
Advanced	A lot	3	5	5	4	5	4	4	5	5	4	4	3	4	Male	45-60	Some colle...	$0 - $24,999
Novice	Not much	3	4	4	4	4	3	4	4	4	3	4	3	4	Female	30-44	Some colle...	$50,000 - $99,999
Advanced	A lot	2	5	2	4	5	5	4	3	5	4	4	4	3	Male	30-44	Bachelor d...	$50,000 - $99,999
Novice	Some	2	4	3	4	3	3	3	3	4	3	2	3	3	Male	) 60	Some colle...	$0 - $24,999
Intermediate	A lot	5	4	4	5	4	5	4	4	3	4	5	5	3	Male	18-29	High schoo...	$25,000 - $49,999
Intermediate	Some	3	4	3	4	4	4	4	3	3	3	4	4	3	Female	30-44	Graduate ...	$50,000 - $99,999
Intermediate	Some	3	4	3	4	4	4	5	4	3	4	3	5	3	Female	30-44	Some colle...	$25,000 - $49,999
Novice	Not at all	3	4	4	3	4	4	4	4	4	4	4	4	3	Male	30-44	Some colle...	$50,000 - $99,999
Novice	Not much	3	4	3	3	5	4	3	4	4	4	4	4	3	Male	) 60	Some colle...	$0 - $24,999
Intermediate	A lot	4	3	4	4	4	5	5	4	4	5	4	4	3	Male	18-29	Bachelor d...	$0 - $24,999
Novice	Some	3	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle...	$0 - $24,999
Novice	Some	3	3	5	3	2	4	4	4	4	4	1	4	3	Male	30-44	Some colle...	$25,000 - $49,999
Novice	Not at all	3	4	5	4	5	4	4	4	3	4	3	5	4	Male	18-29	Some colle...	$150,000+
Intermediate	Some	3	3	4	2	5	4	5	3	5	5	2	1	3	Male	) 60	Graduate ...	$100,000 - $149,99
Intermediate	A lot	3	4	3	4	3	2	4	3	3	2	4	5	3	Male	30-44	Some colle...	$50,000 - $99,999
Intermediate	Some	4	4	4	5	5	4	2	4	4	3	4	3	4	Male	) 60	Graduate ...	$50,000 - $99,999
Intermediate	Some	2	3	2	2	4	3	4	4	4	2	2	5	4	Female	30-44	Graduate ...	$50,000 - $99,999
Intermediate	Some	3	3	5	3	5	5	5	5	4	5	4	4	4	Female	30-44	Some colle...	$25,000 - $49,999
Advanced	A lot	3	5	4	3	5	4	4	3	5	4	4	4	3	Male	) 60	Some colle...	$150,000+
Novice	Not much	3	3	3	4	4	4	3	4	3	3	2	5	3	Male	18-29	Some colle...	$25,000 - $49,999
Novice	Some	3	4	3	3	3	3	3	4	5	4	2	4	3	Male	30-44	High schoo...	$25,000 - $49,999
Intermediate	Some	3	3	4	3	3	4	4	4	4	4	3	3	3	Male	45-60	Bachelor d...	$25,000 - $49,999
Intermediate	Not much	3	4	5	4	4	4	2	4	5	4	2	4	3	Female	45-60	Bachelor d...	$0 - $24,999
Intermediate	Not much	4	5	4	2	5	5	5	4	4	3	3	4	4	Male	18-29	Some colle...	$150,000+
Intermediate	Not much	2	1	1	4	4	4	4	4	3	5	3	4	2	Male	45-60	Bachelor d...	$50,000 - $99,999
Intermediate	A lot	2	5	3	4	4	4	3	4	3	3	4	3	1	Female	45-60	Some colle...	$150,000+
Intermediate	Not much	3	4	3	4	4	4	5	4	3	2	5	3	3	Male	30-44	Graduate ...	$100,000 - $149,99
Novice	Not at all	3	4	3	4	3	3	4	4	4	3	4	4	3	Male	) 60	Some colle...	$0 - $24,999
Novice	Some	3	4	3	4	4	4	5	4	4	3	4	4	3	Female	30-44	Some colle...	$50,000 - $99,999
Advanced	A lot	4	4	4	5	5	5	5	5	5	4	4	4	4	Male	) 60	High schoo...	$150,000+
Intermediate	Some	4	4	4	3	5	4	4	4	3	3	2	4	3	Male	18-29	Bachelor d...	$0 - $24,999
Novice	Some	3	3	3	3	3	4	4	4	3	4	4	4	3	Male	45-60	Graduate ...	$50,000 - $99,999
Novice	Some	4	5	3	4	5	2	5	5	5	4	4	4	3	Female	18-29	Some colle...	$50,000 - $99,999
Novice	Not at all	1	3	3	5	5	4	3	4	5	4	4	3	3	Male	45-60	High schoo...	$50,000 - $99,999
Advanced	A lot	3	3	2	5	5	5	5	5	3	4	5	5	3	Male	45-60	Bachelor d...	$25,000 - $49,999
Novice	Some	3	4	3	4	5	4	4	4	3	4	4	4	5	Female	45-60	Some colle...	$0 - $24,999
Novice	A lot	3	4	2	4	3	3	5	4	3	3	4	3	3	Female	18-29	Some colle...	$50,000 - $99,999

III. Attribute Selection

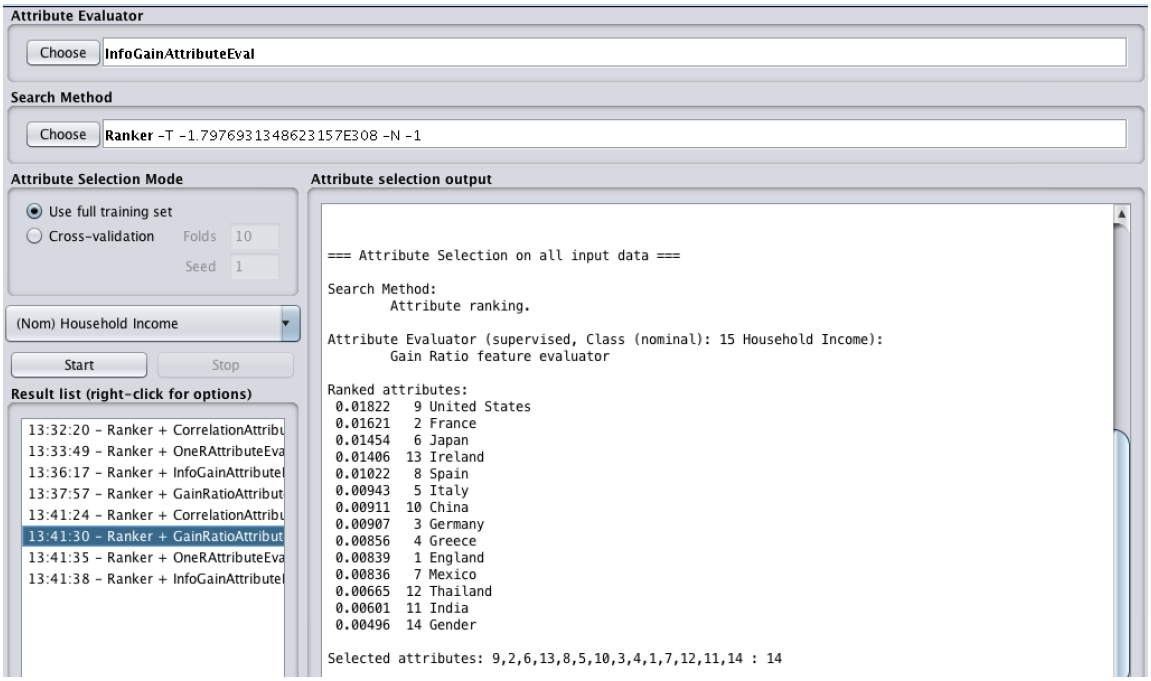
- We ran four different attribute selection algorithms on Weka to gauge the rank of each attribute in relation to the class attribute and selected the top eight attributes for each algorithm. Please find below various attribute selection algorithms and selected attributes we ran in Figures 7 – 10 below:

Figure 7: CorrelationAttributeEval Algorithm



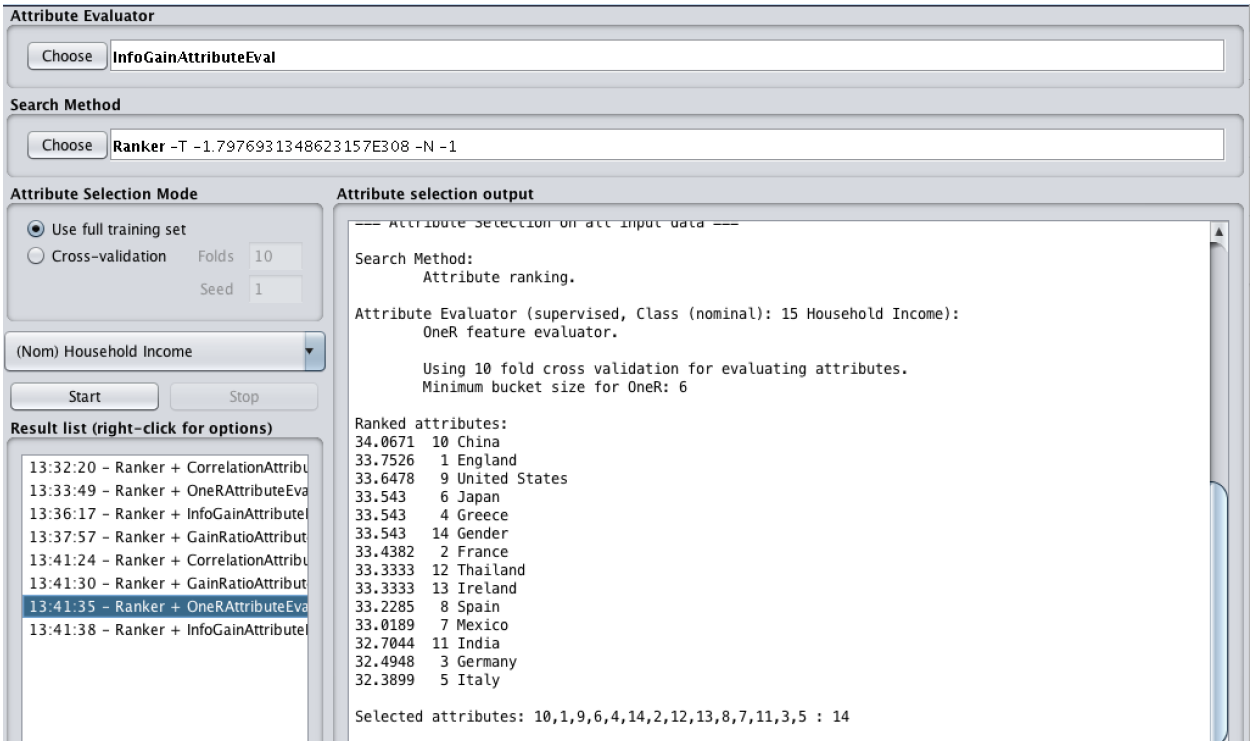
Attributes Selected: United States, France, Italy, Japan, Ireland, Spain, India, Gender.

Figure 8: GainRatioAttributeEval Algorithm



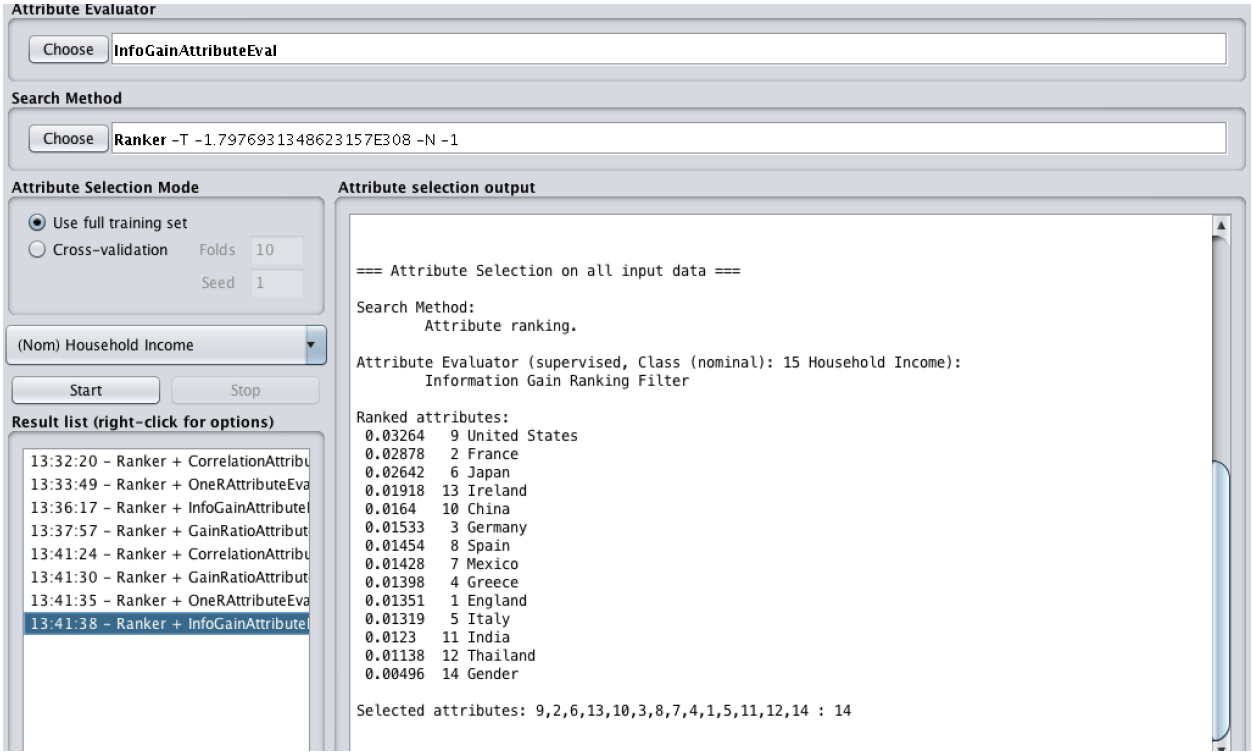
Attributes Selected: United States, France, Japan, Ireland, Spain, Italy, China, Germany.

Figure 9: OneRAttributeEval Algorithm



Attributes selected: China, England, United States, Japan, Greece, Gender, France, Thailand.

Figure 10: InfoGain Attribute Selection Algorithm



Attributes selected: United States, France, Japan, Ireland, China, Germany, Spain, Mexico

- Lastly, we also chose and developed our own set of attributes based on the most frequently occurring attributes amongst the previous four algorithms. Resulting in our fifth attribute set, shown below.
    - Our own Attribute Set: US, France, Spain, Japan, Germany, Gender, Age, and Education.

So, to consolidate, the table below shows our 5 different attribute sets. We will be using these to build our classification models:

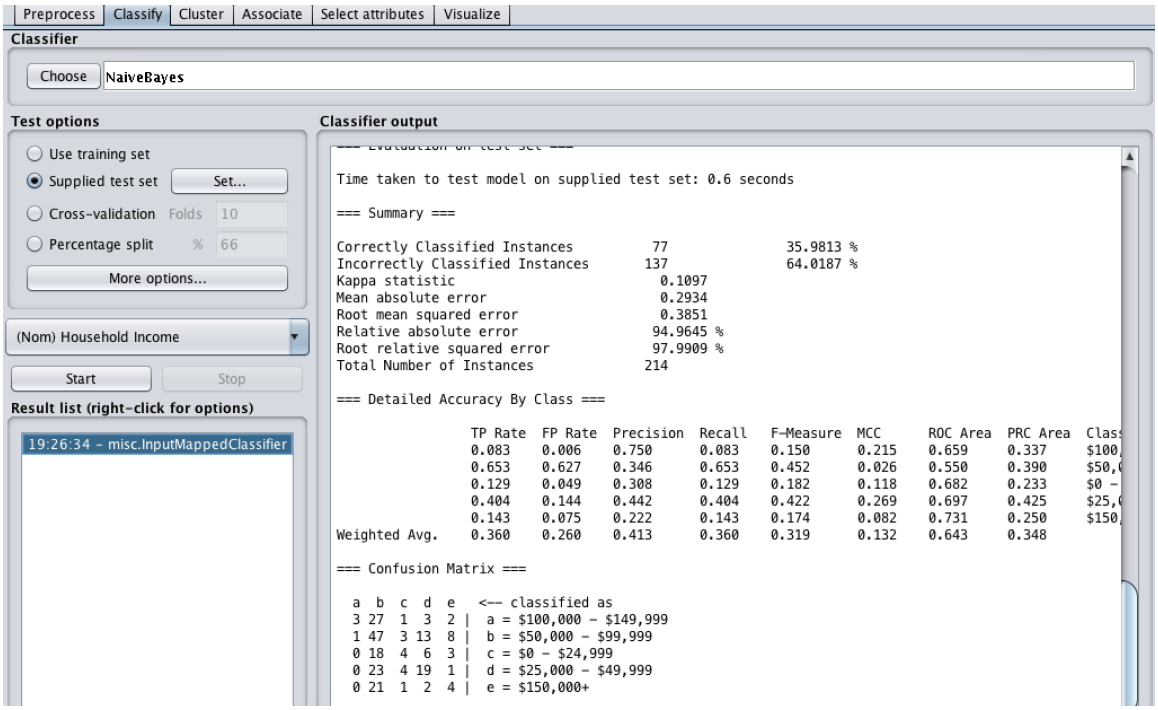
Attribute Set	Attributes Selected
CorrelationAttributeEval	United States, France, Italy, Japan, Ireland, Spain, India, Gender
GainRatioAttributeEval	United States, France, Japan, Ireland, Spain, Italy, China, Germany.
OneRAttributeEval Algorithm	Attributes selected: China, England, United States, Japan, Greece, Gender, France, Thailand
InfoGainAttributeEval Algorithm	United States, France, Japan, Ireland, China, Germany, Spain, Mexico
Our Own Attribute Set	US, France, Spain, Japan, Germany, Gender, Age, and Education

IV. Classification Models:

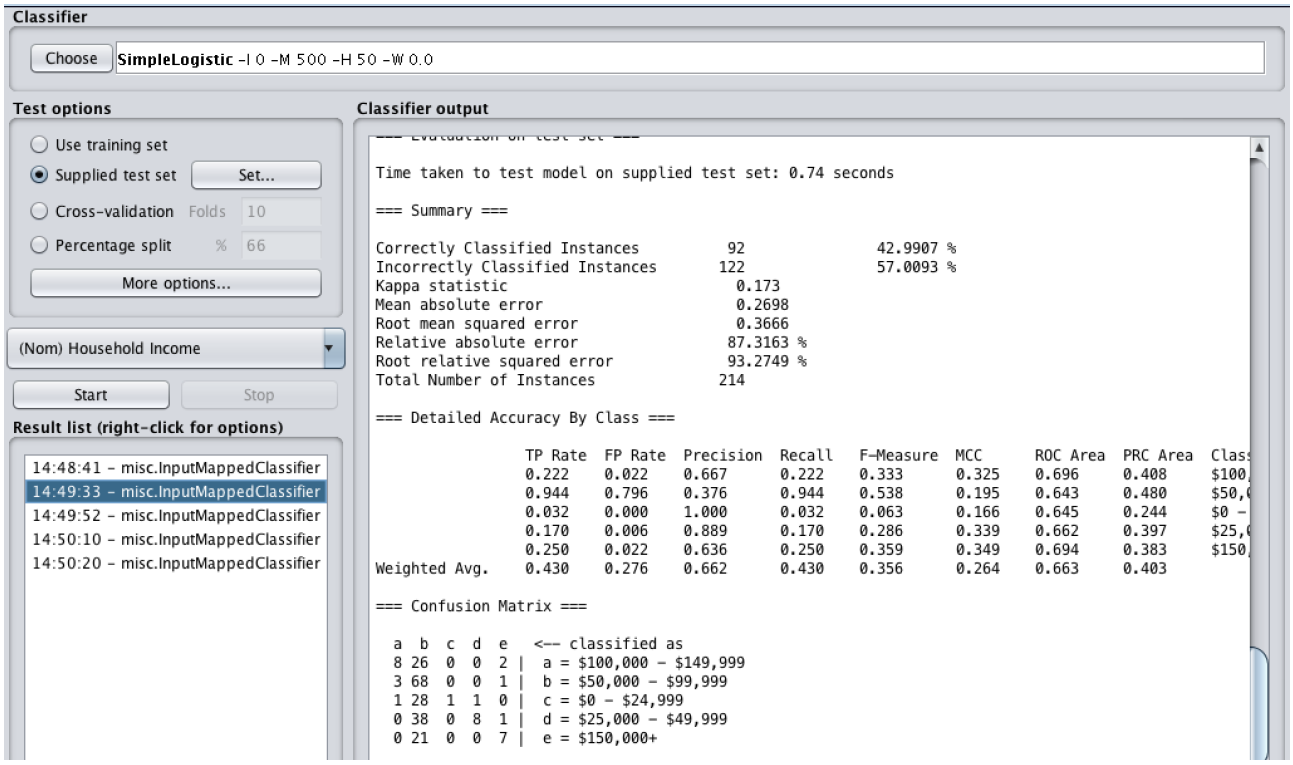
We then used our training set data and built 5 different classification models for each of our 5 attribute sets. We used five different classification algorithms: Naïve Bayes, J48, Simple Logistic, OneR, & RandomForest. Please see screenshots of the various models below organized by each attribute set.

CorrelationAttributeEval Attribute Set:

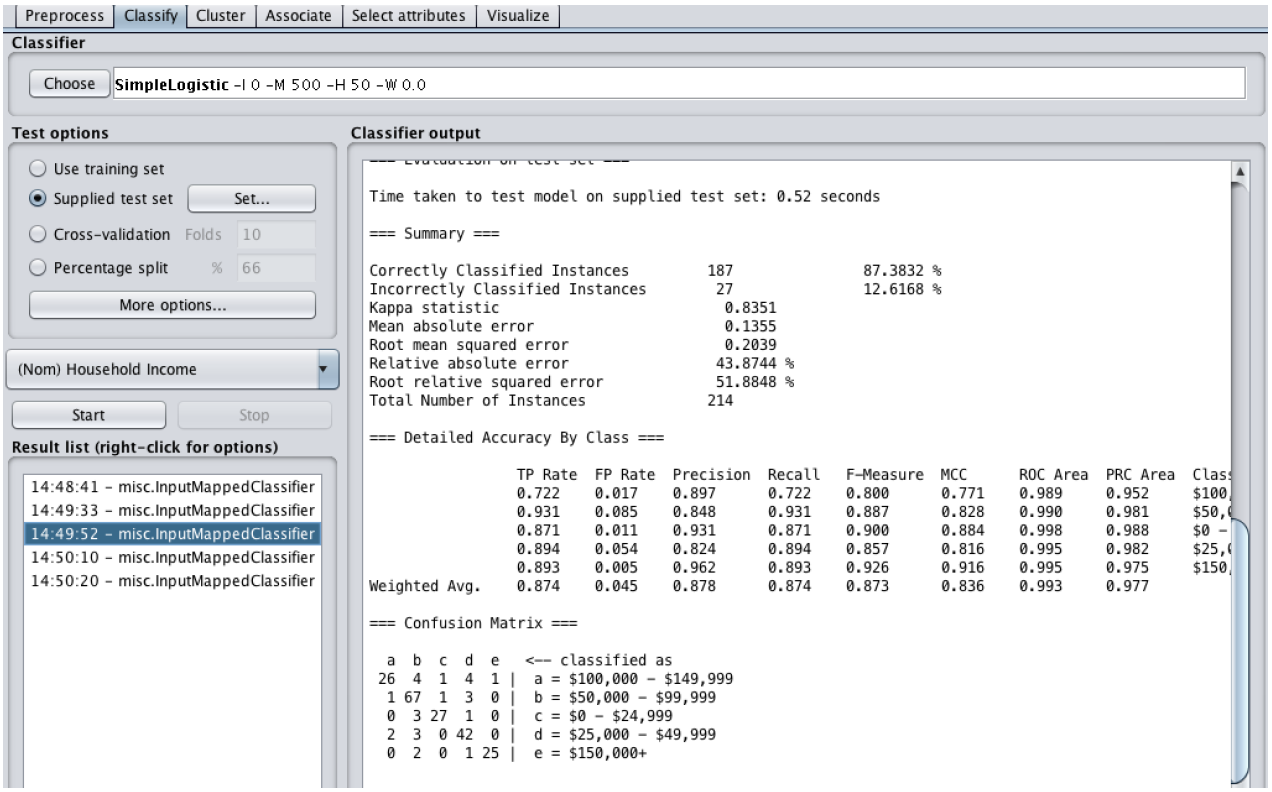
a) Naïve Bayes Classification Model



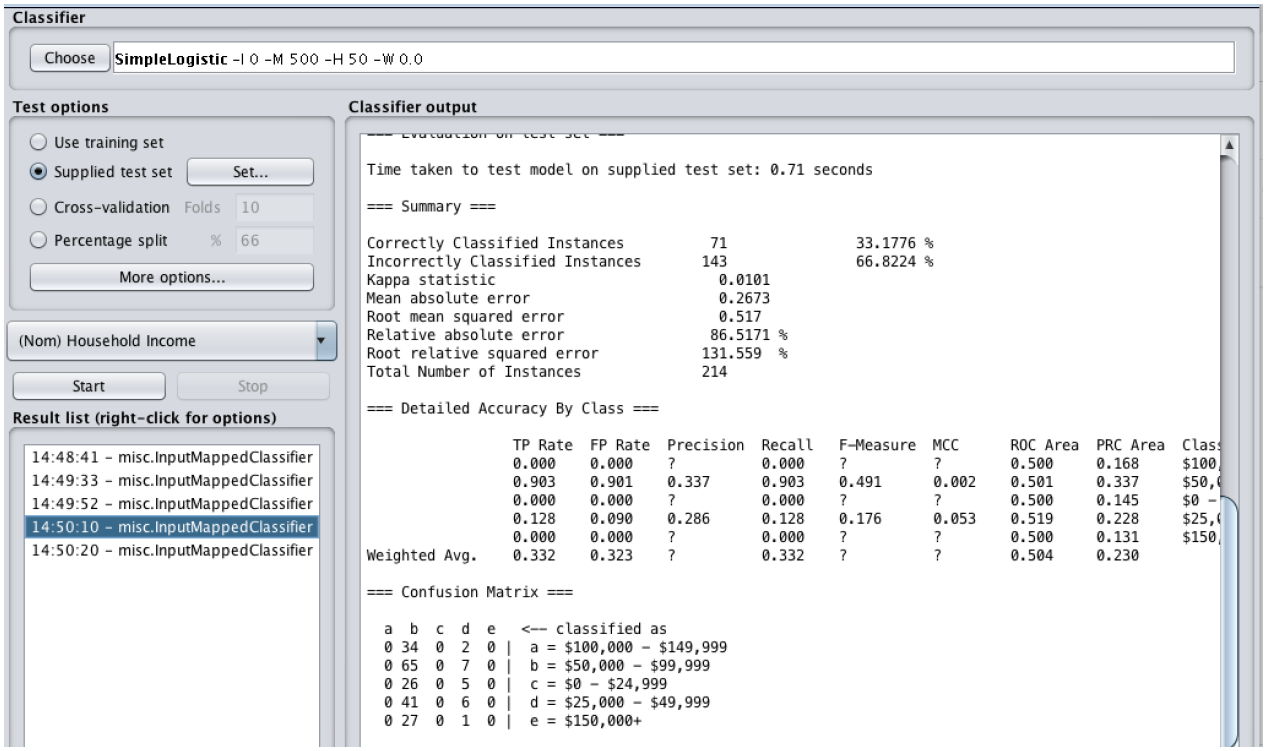
b) J48 Classification Model



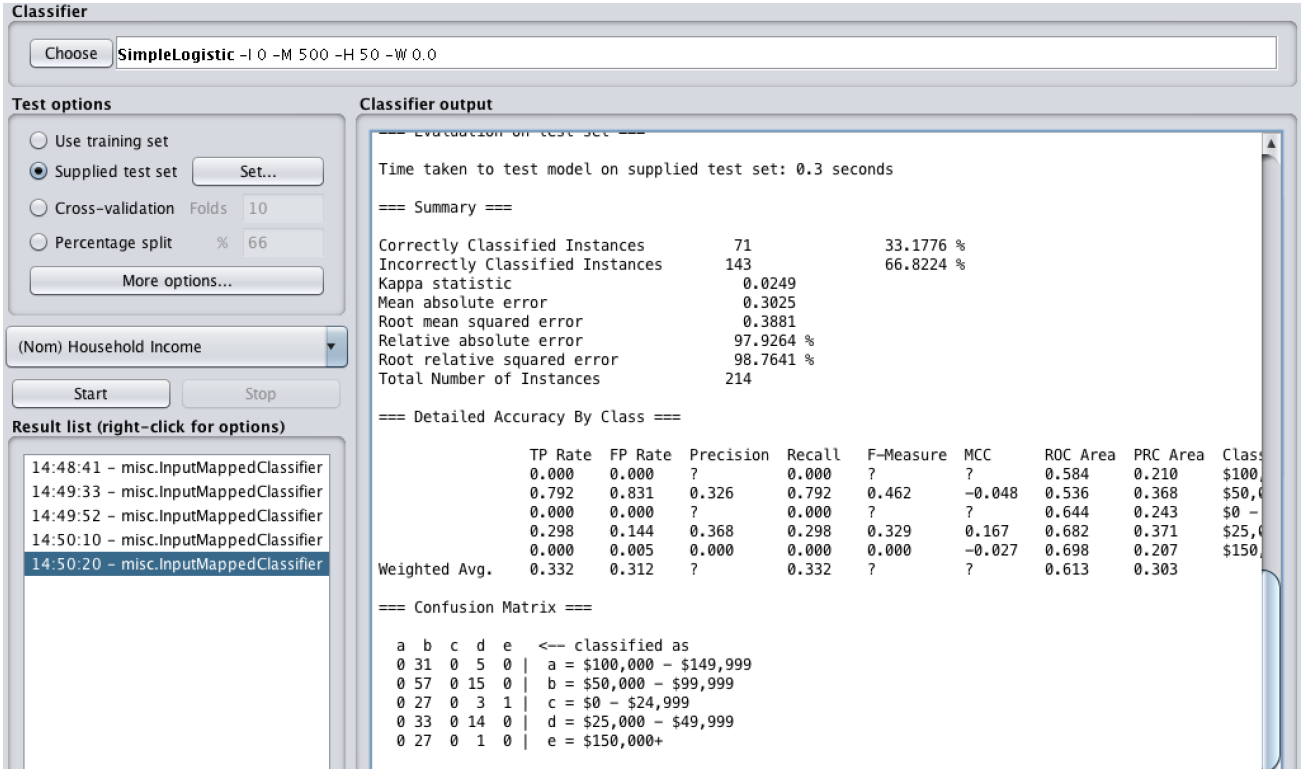
c) Random Forest Classification Model



d) OneR Classification Model

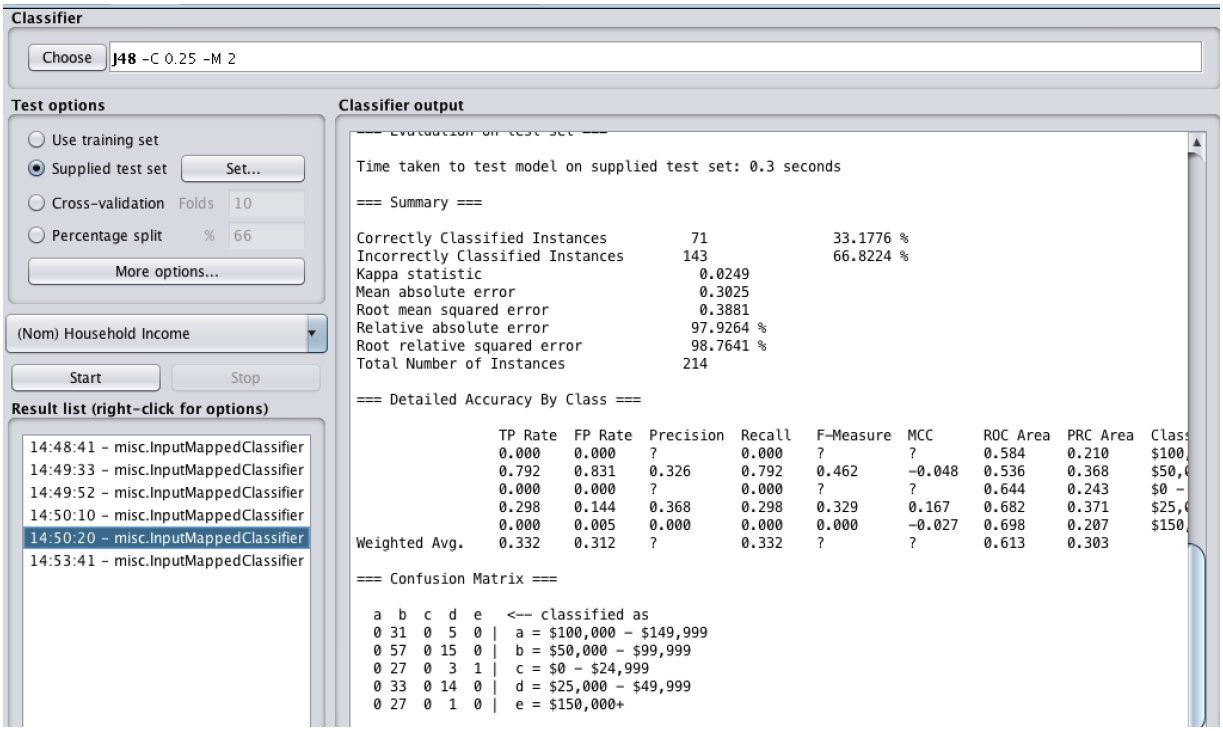


e) Simple Logistic Classification Model

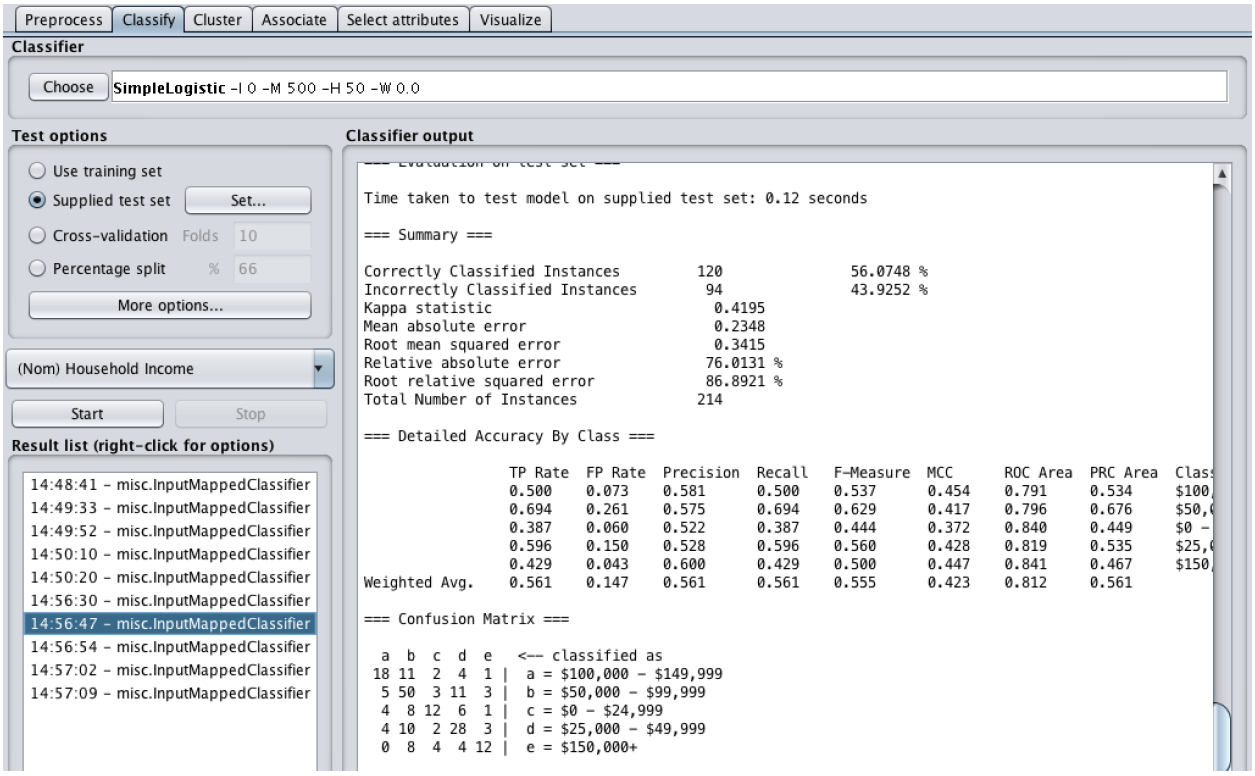


OneRAttributeEval Set:

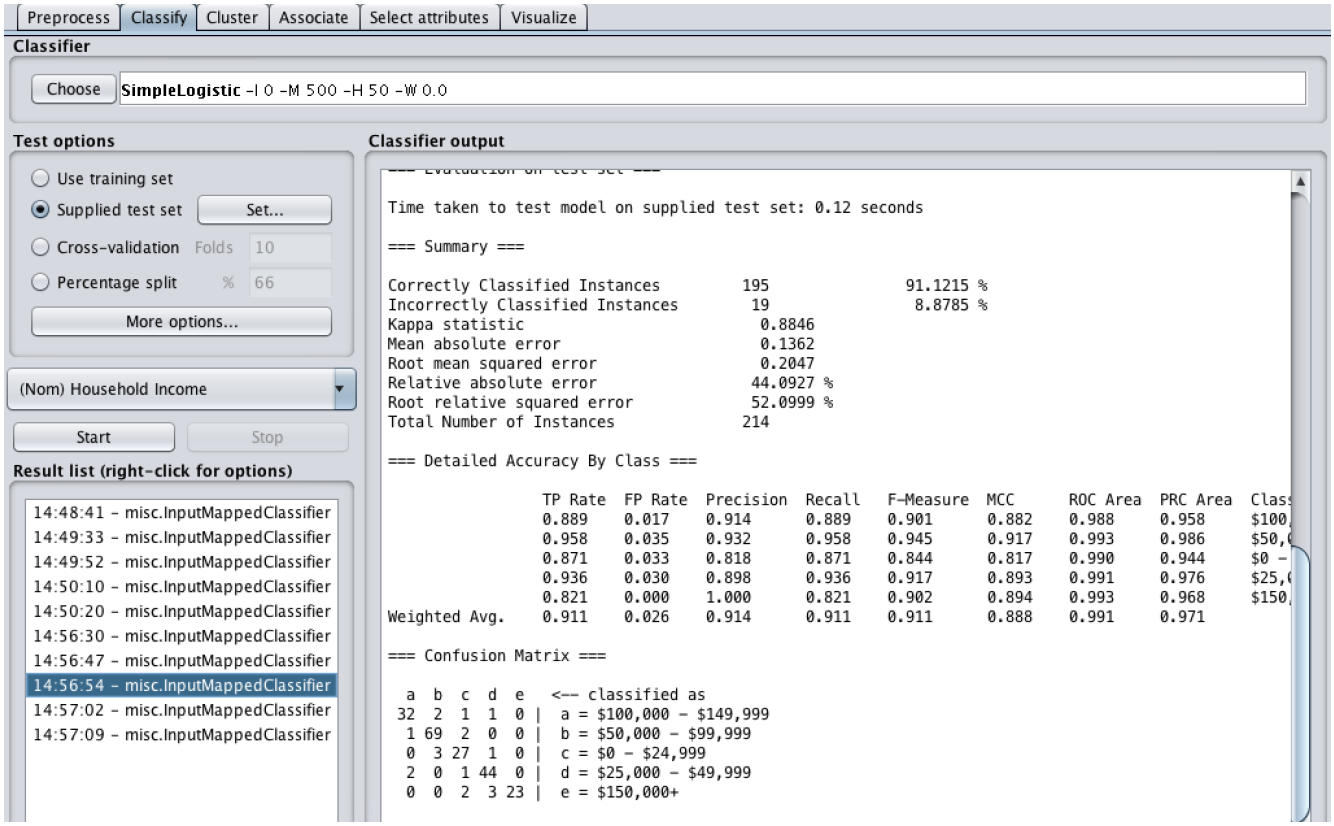
a) Naïve Bayes Model



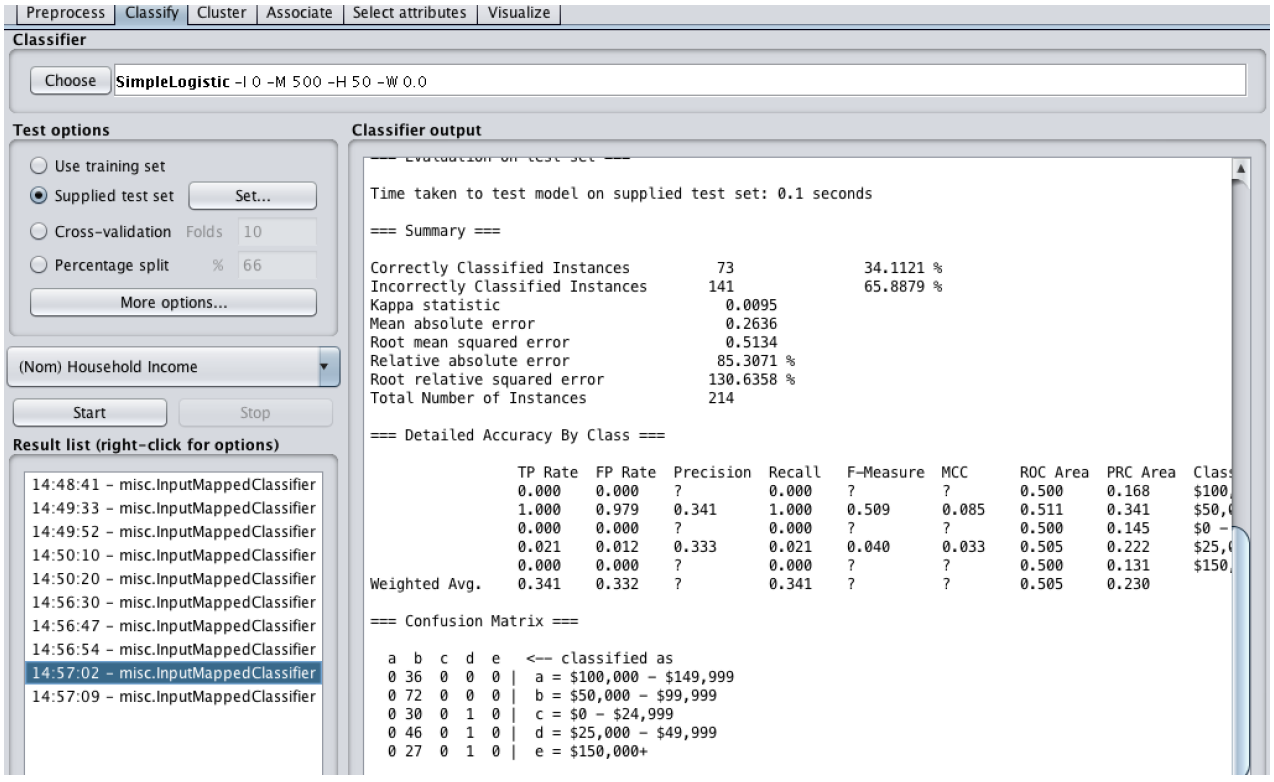
b) J48



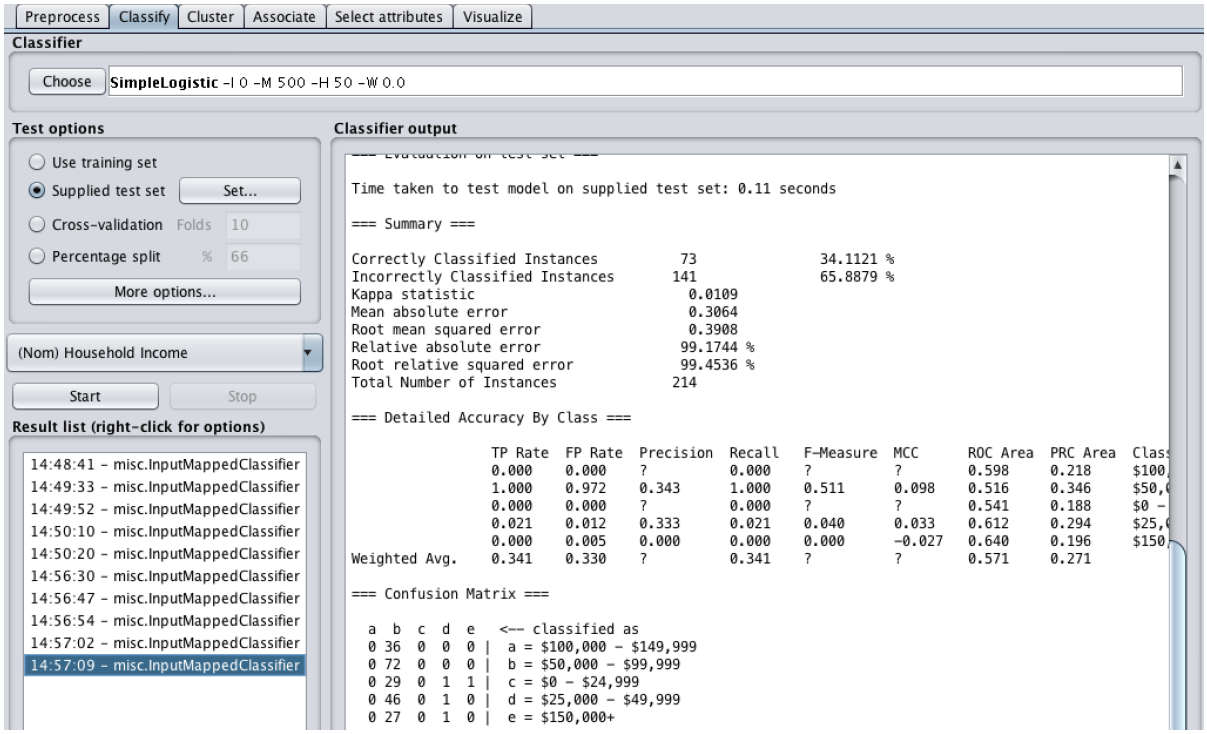
c) Random Forest



d) OneR

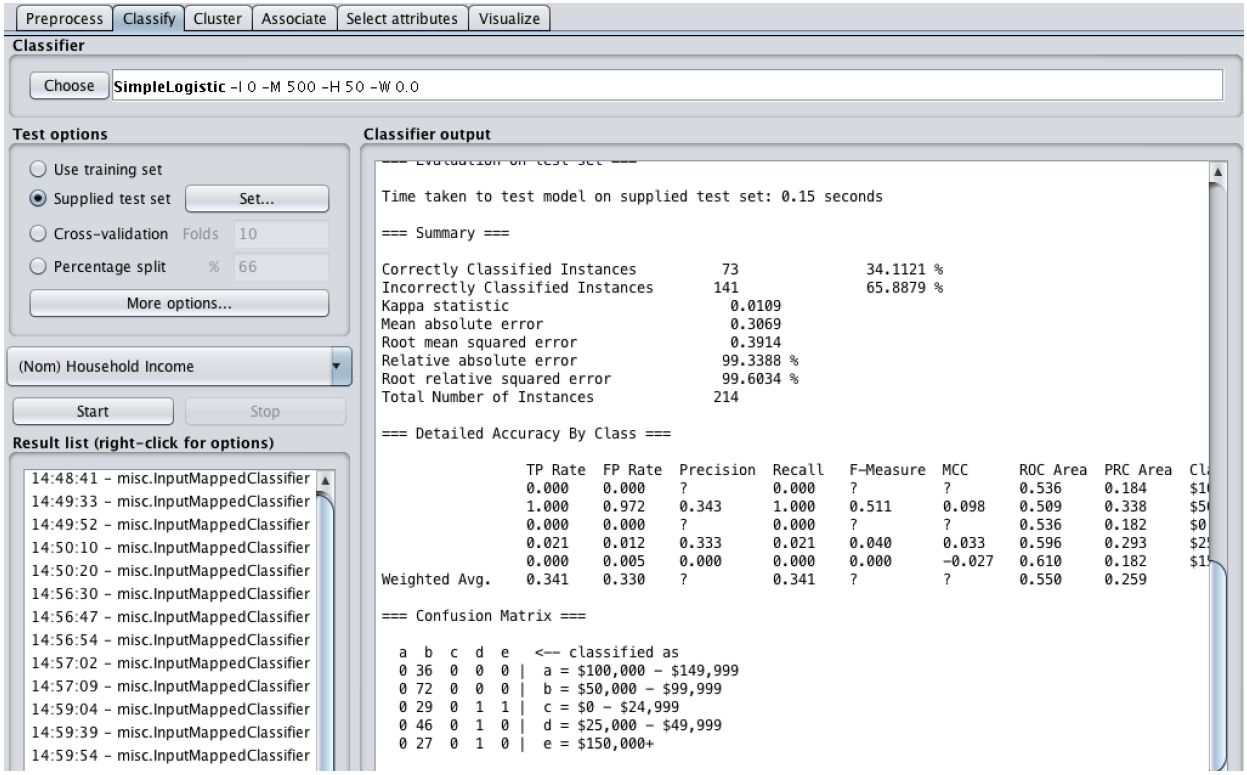


e) Simple Logistic

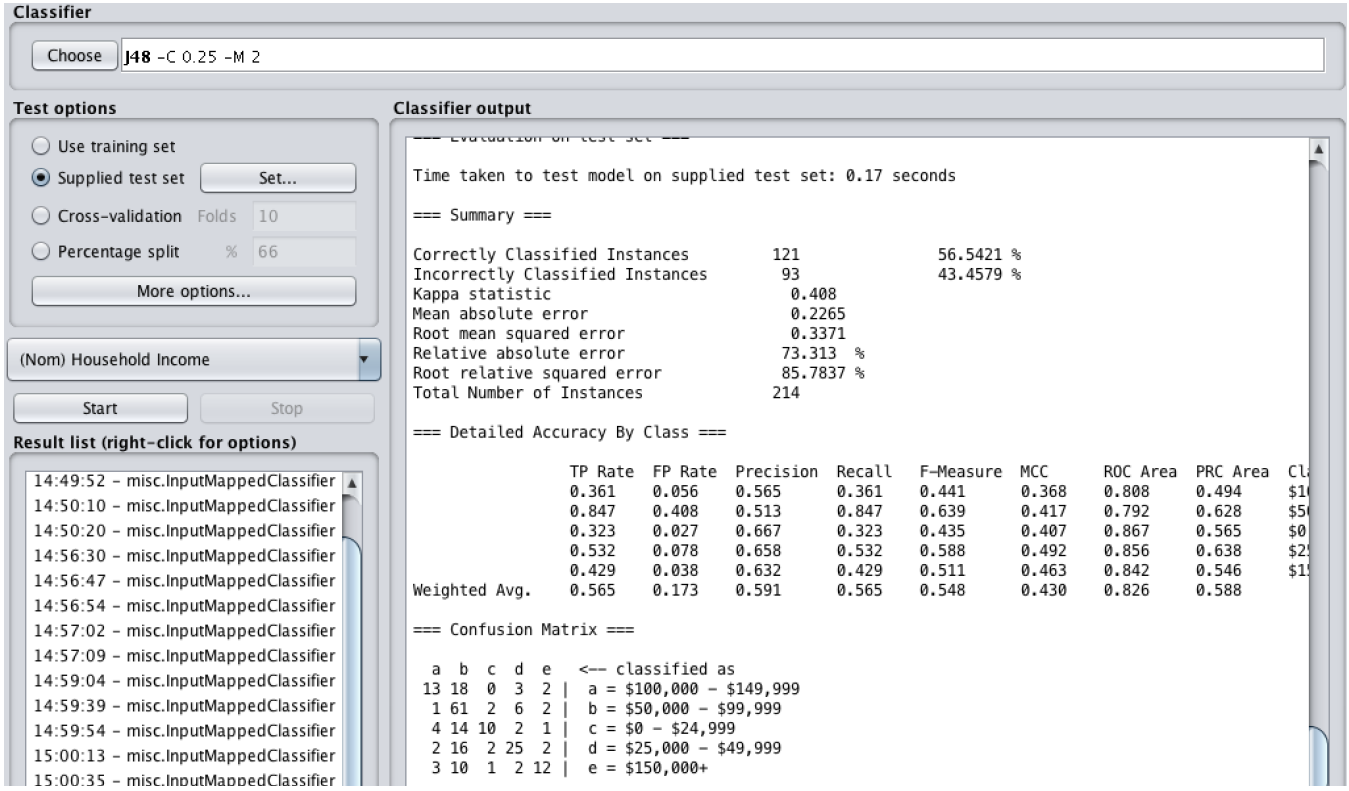


GainRatioAttributeEval Set:

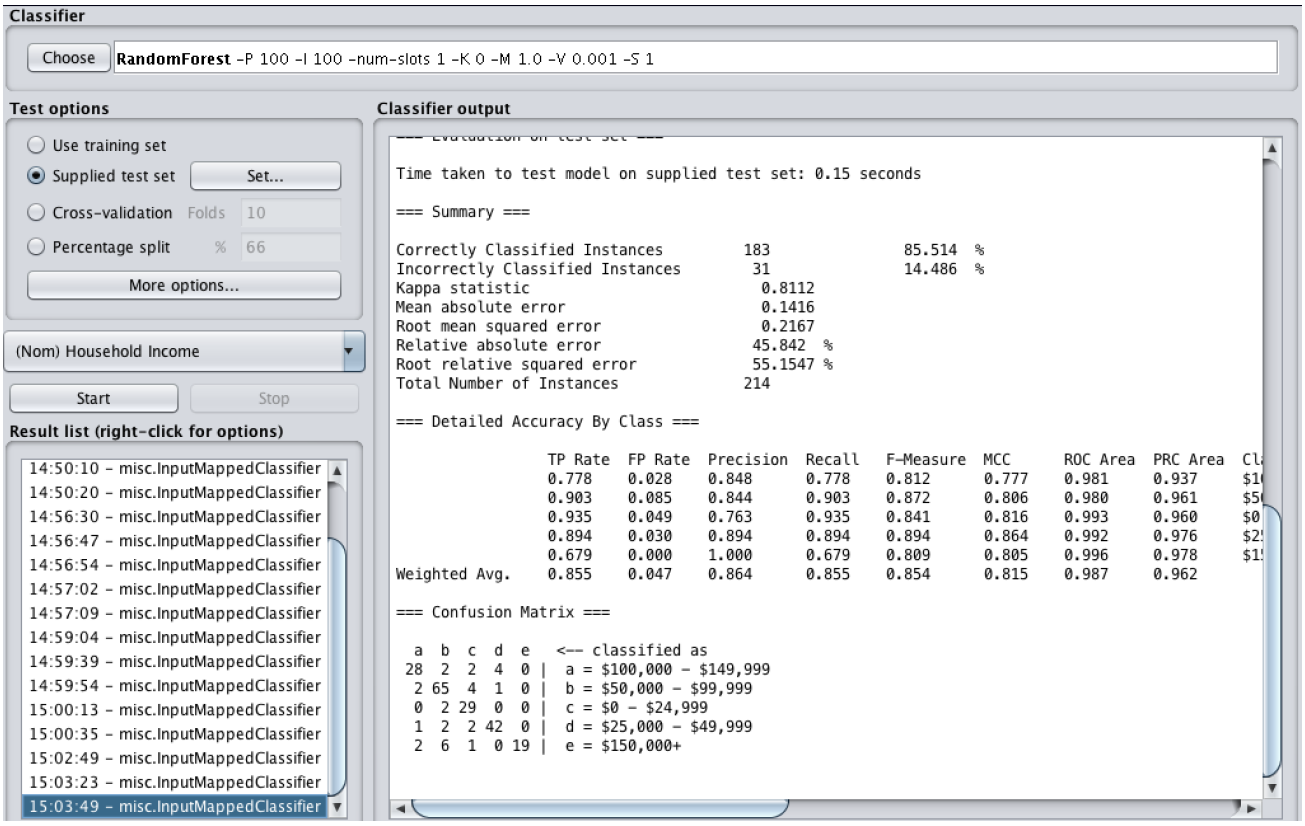
a) Naïve Bayes Classification Model



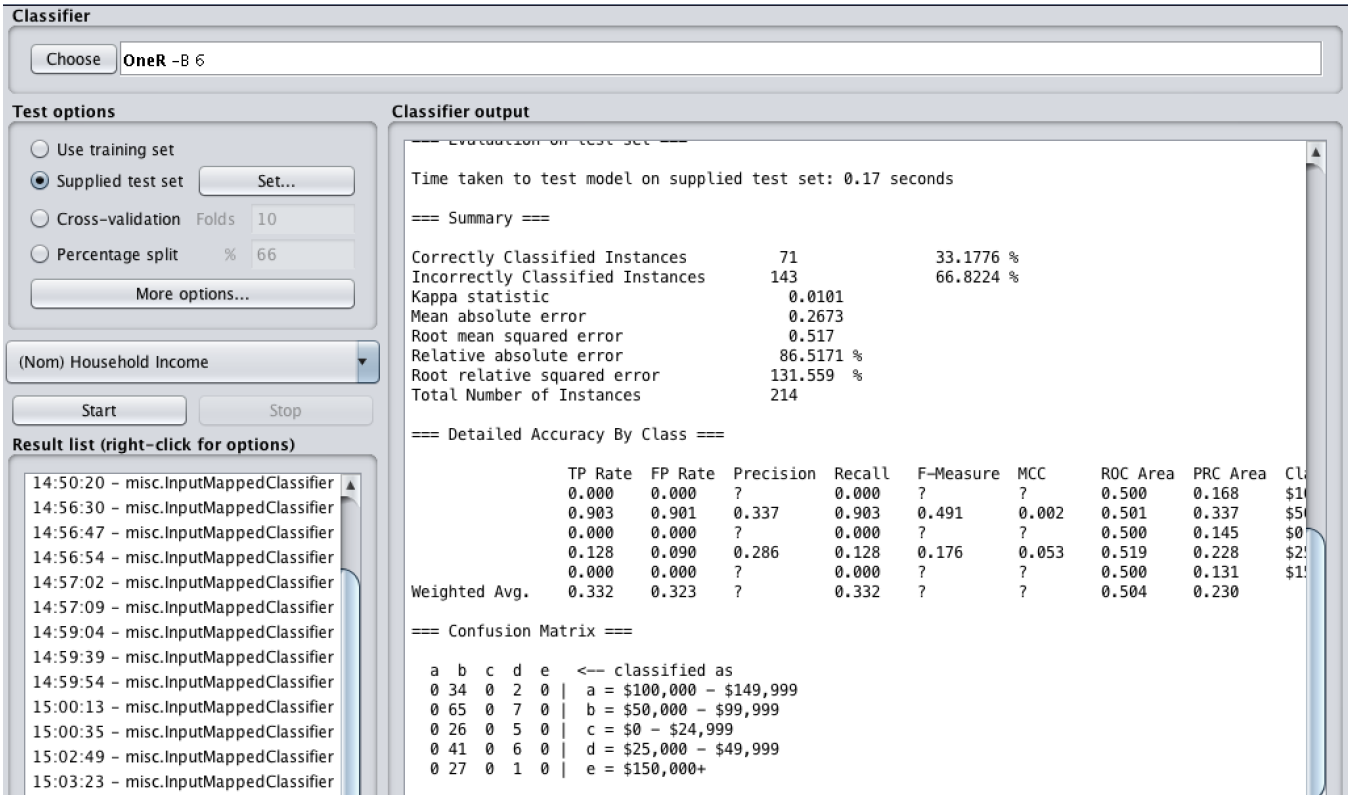
b) J48 Classification Model



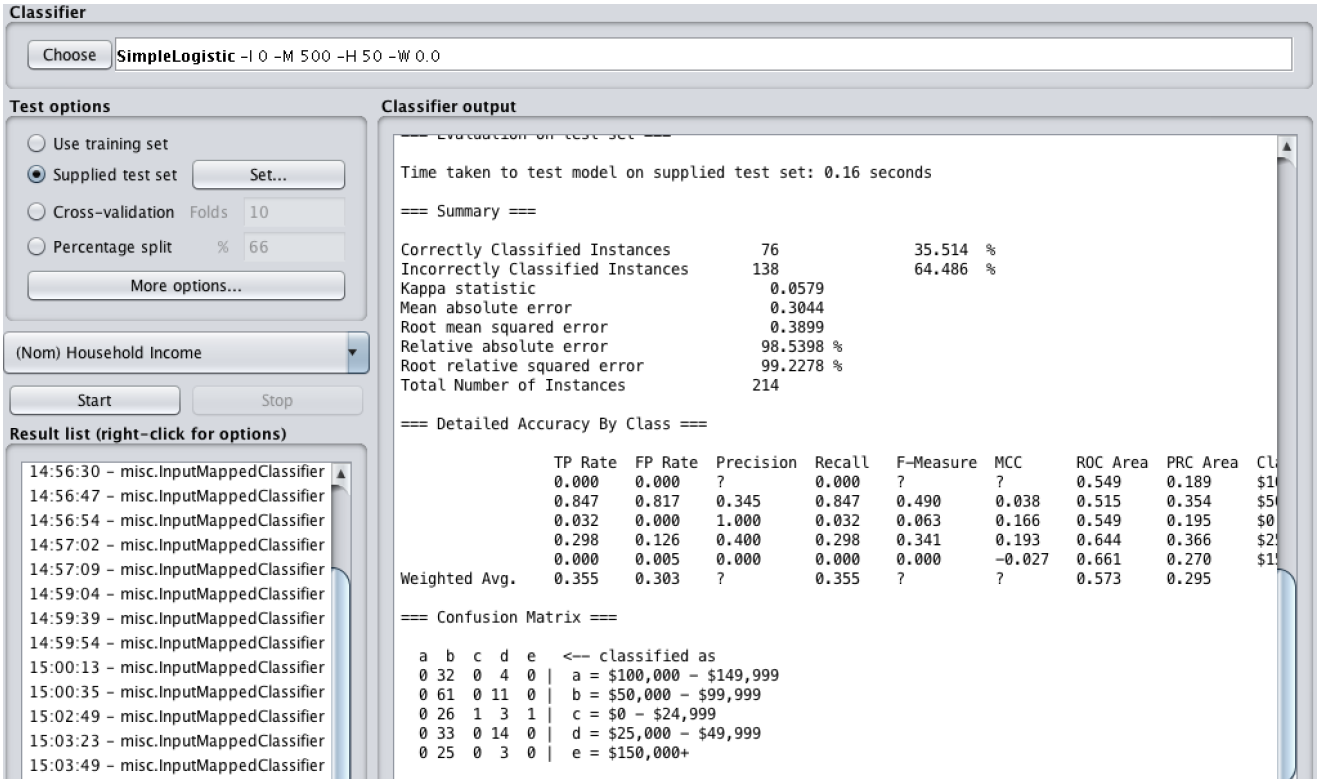
c) Random Forest Classification Model



d) OneR Classification Model

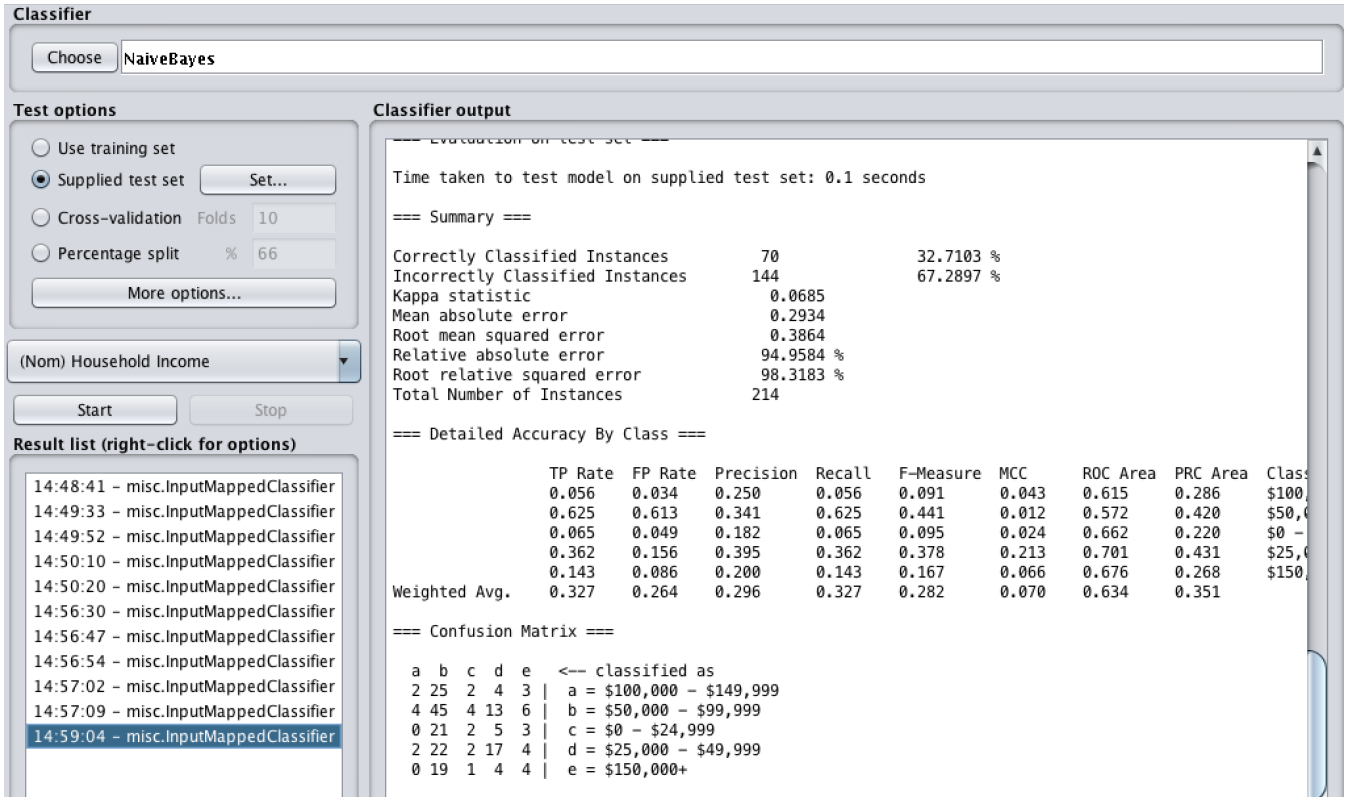


e) Simple Logistic Classification Model

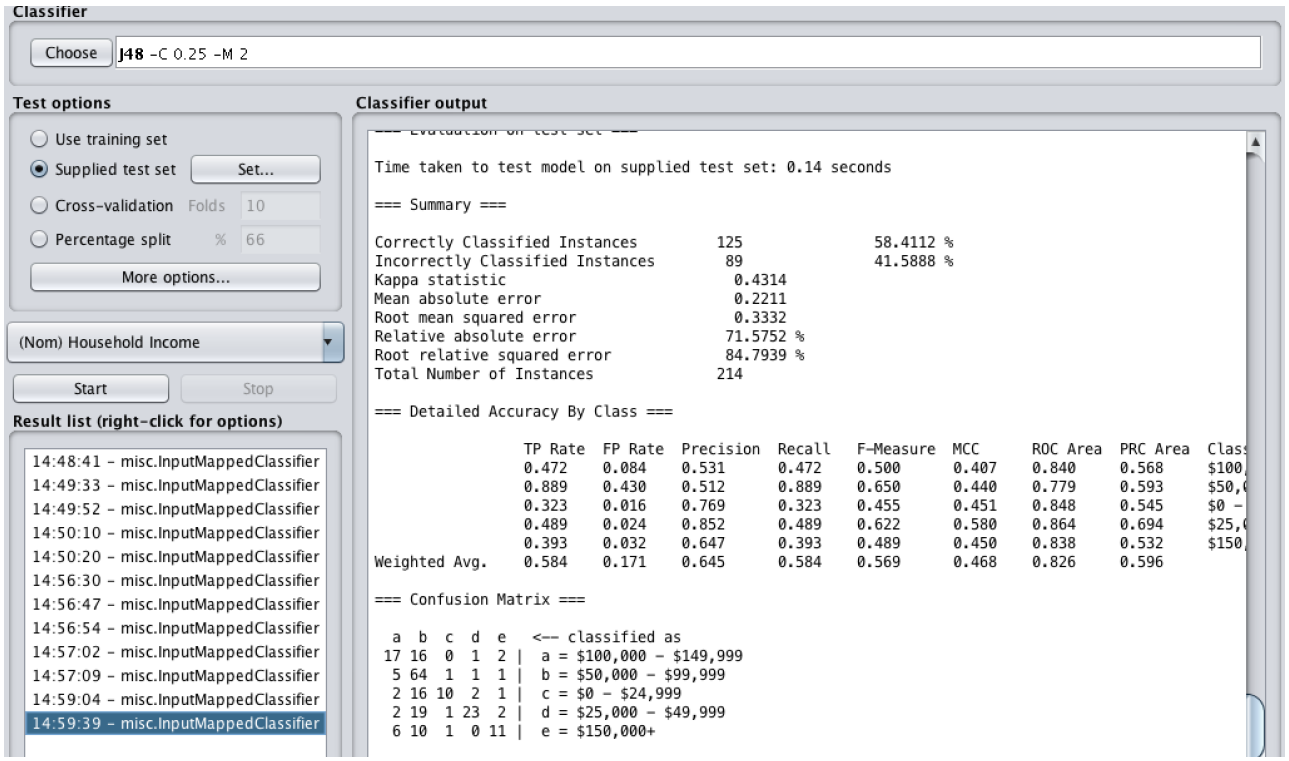


InfoGainAttributeEval Set:

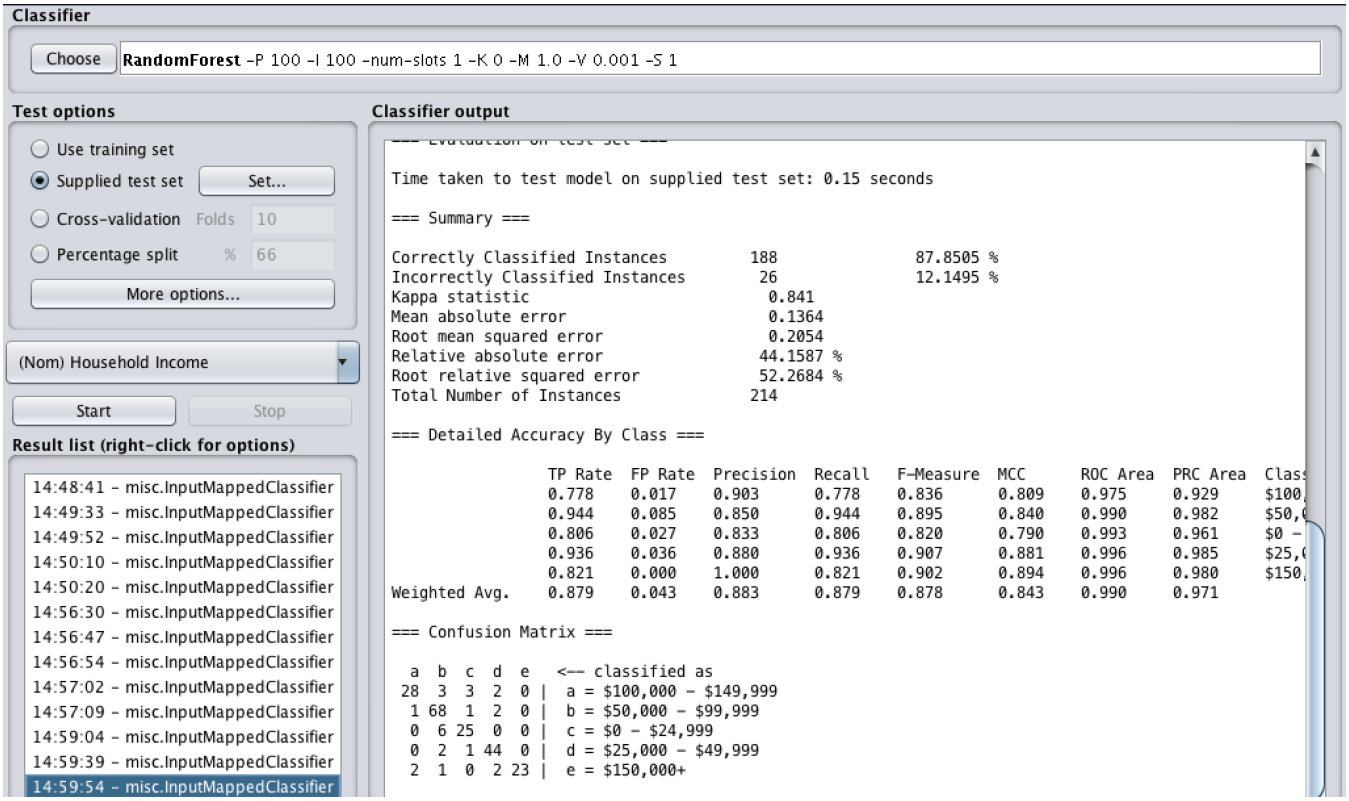
a) Naïve Bayes Classification Model



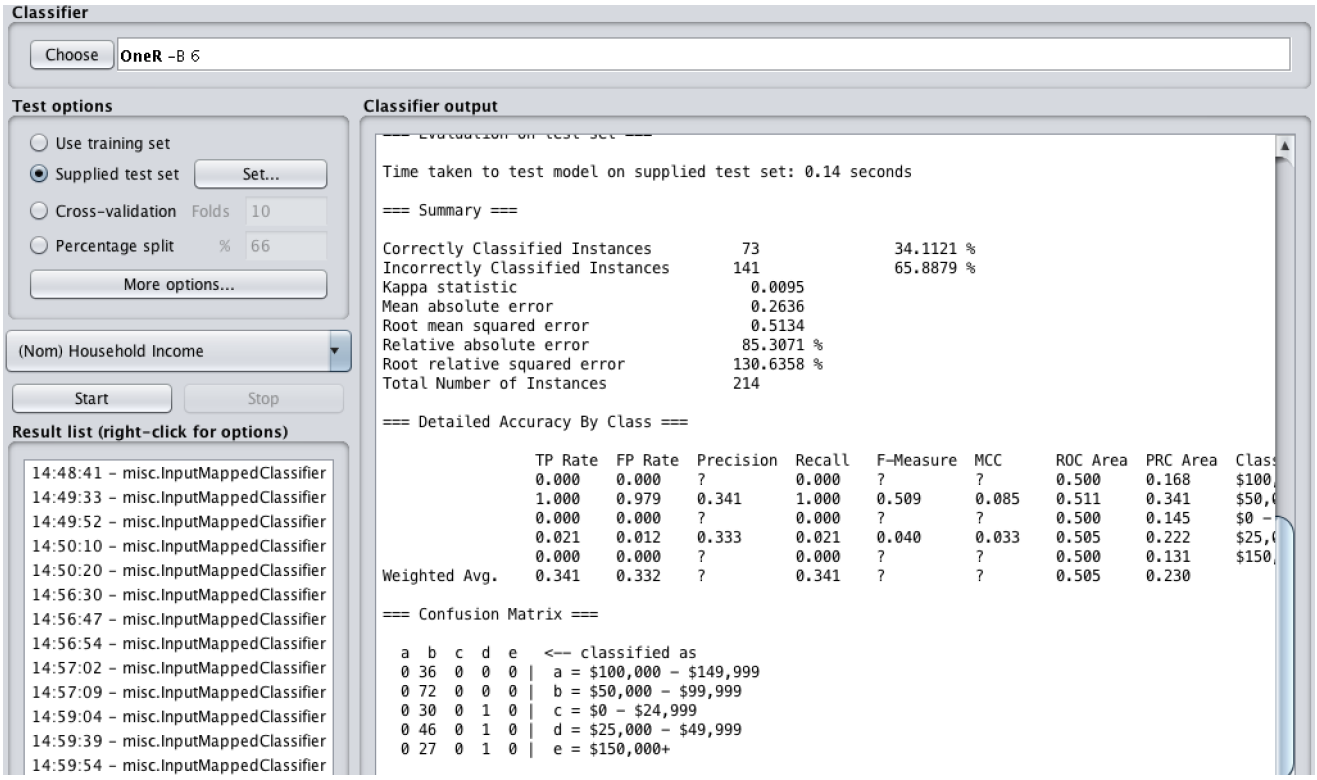
b) J48 Classification Model



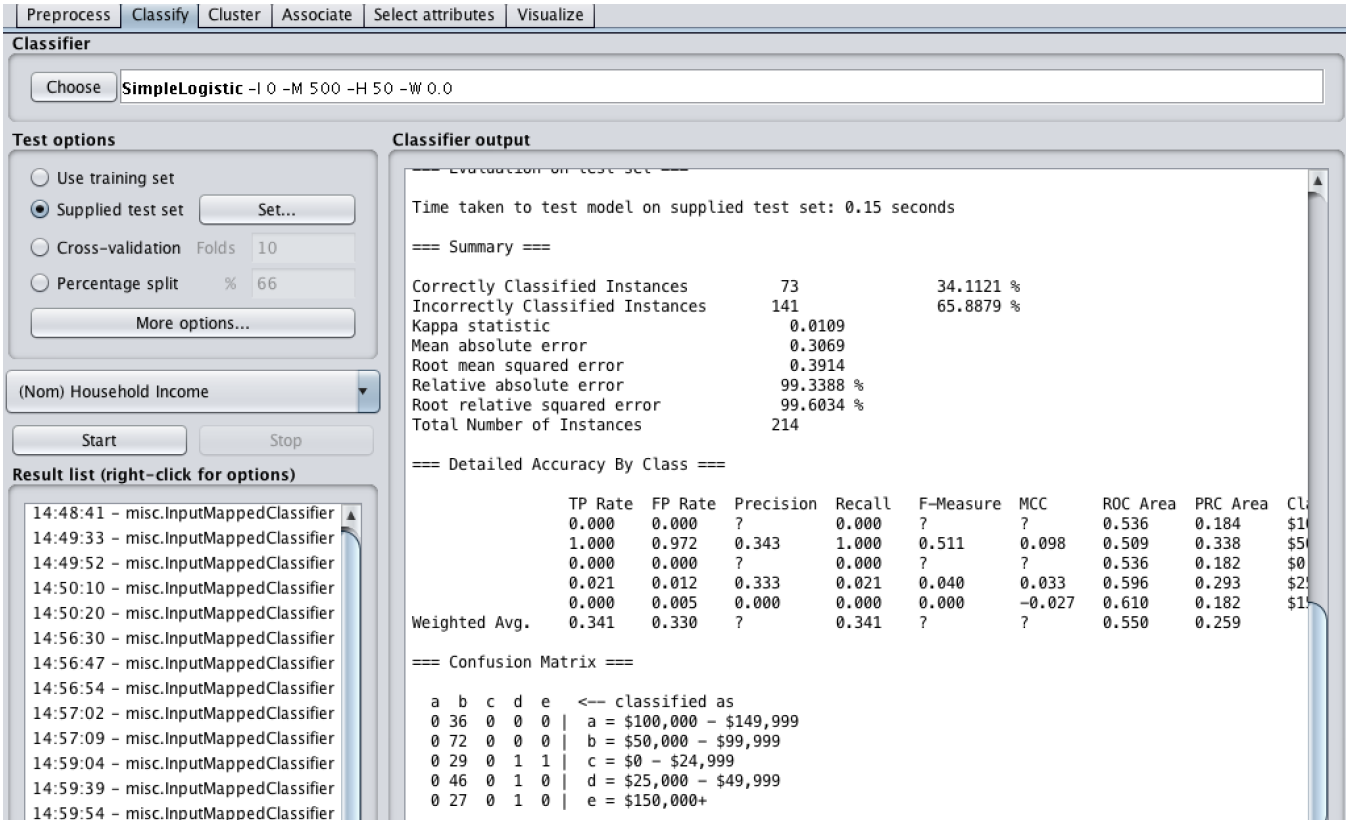
c) Random Forest Classification Model



d) OneR Classification Model

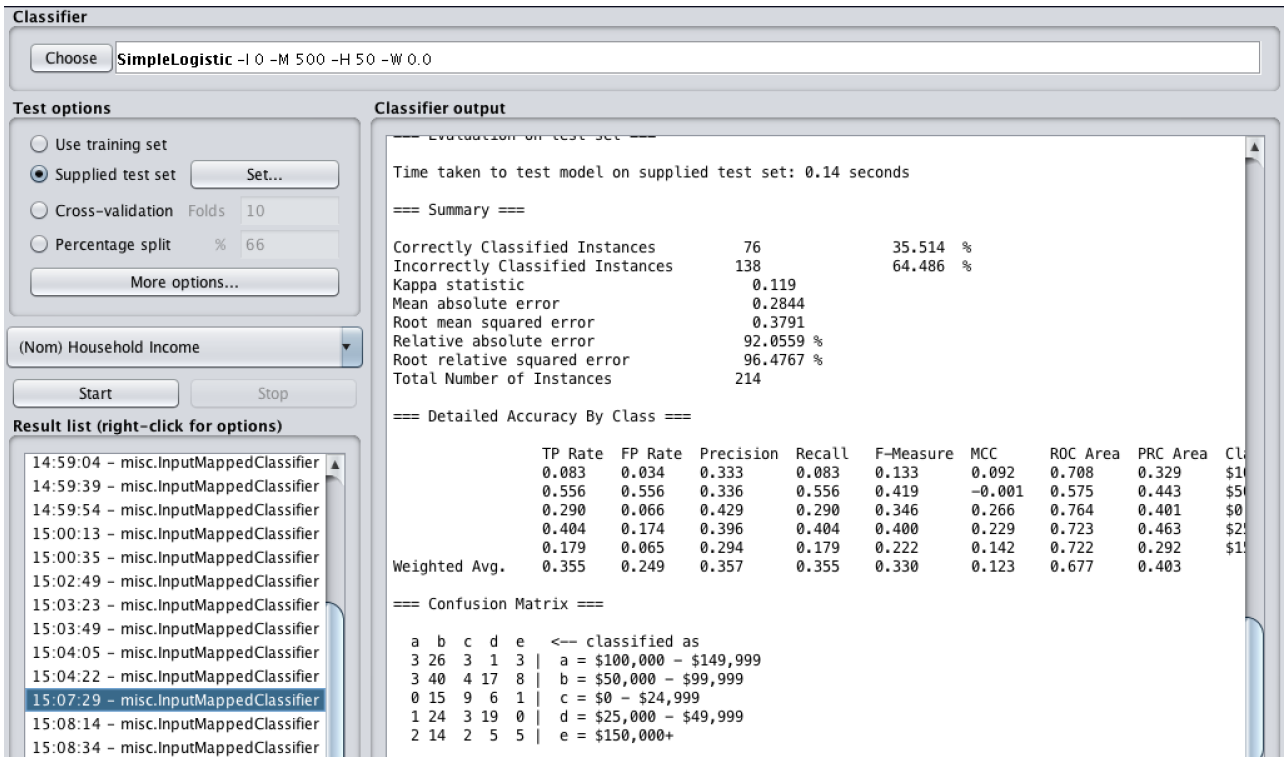


e) Simple Logistic Classification Model

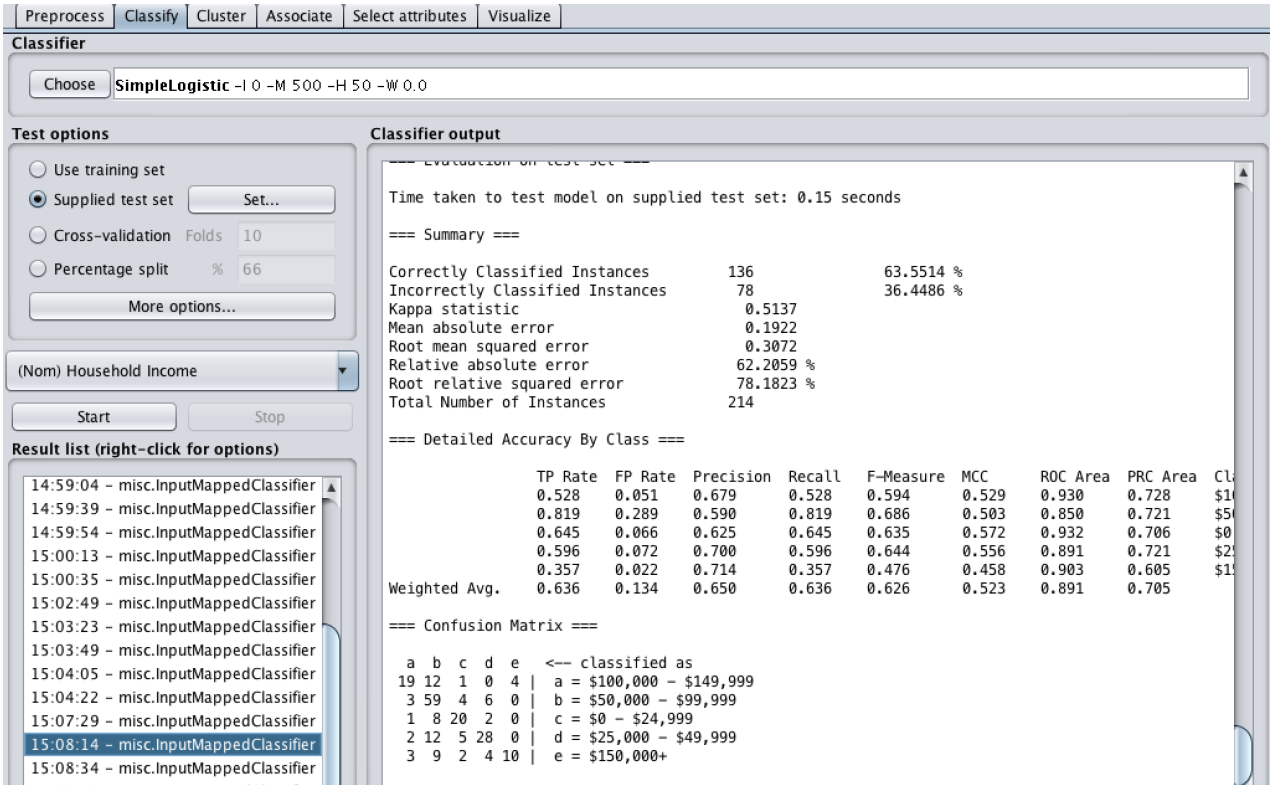


Our Own Attribute Set:

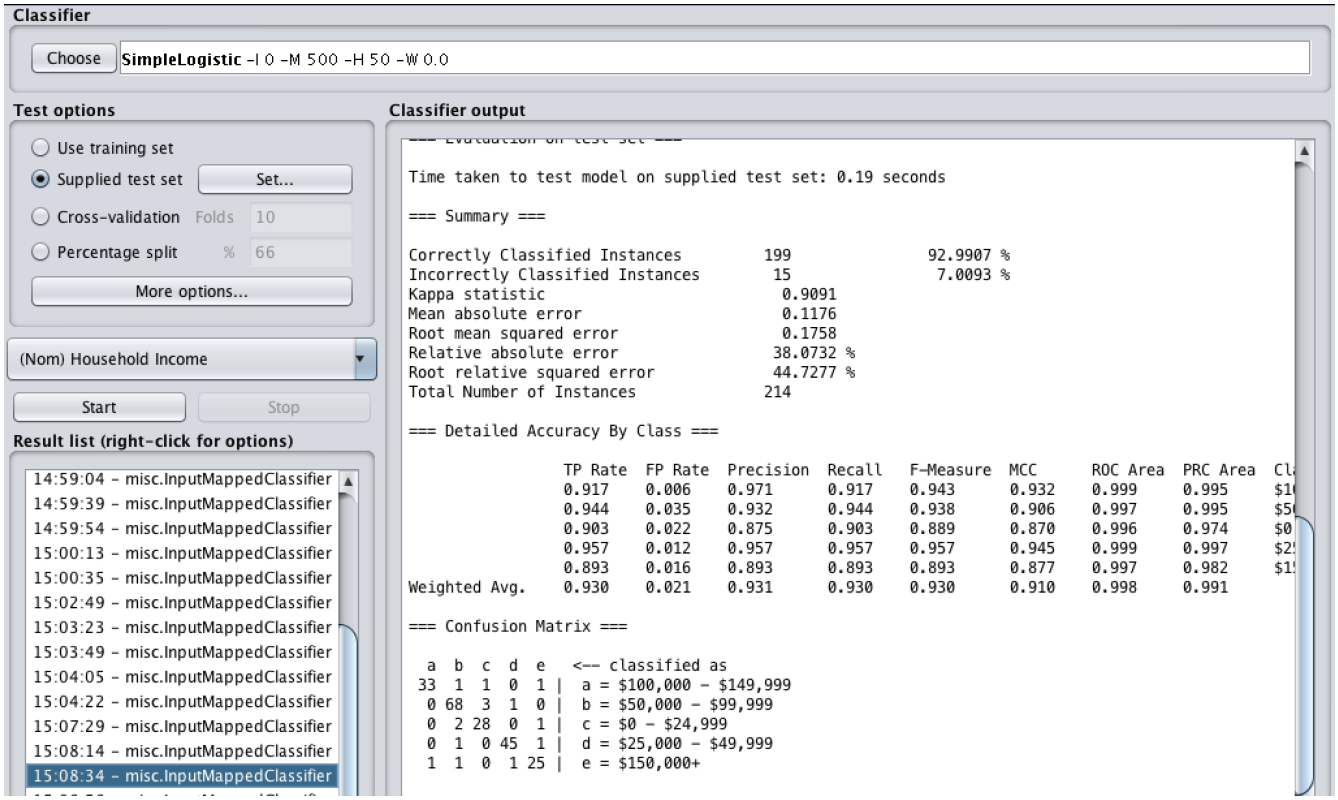
a) Naïve Bayes Classification Model



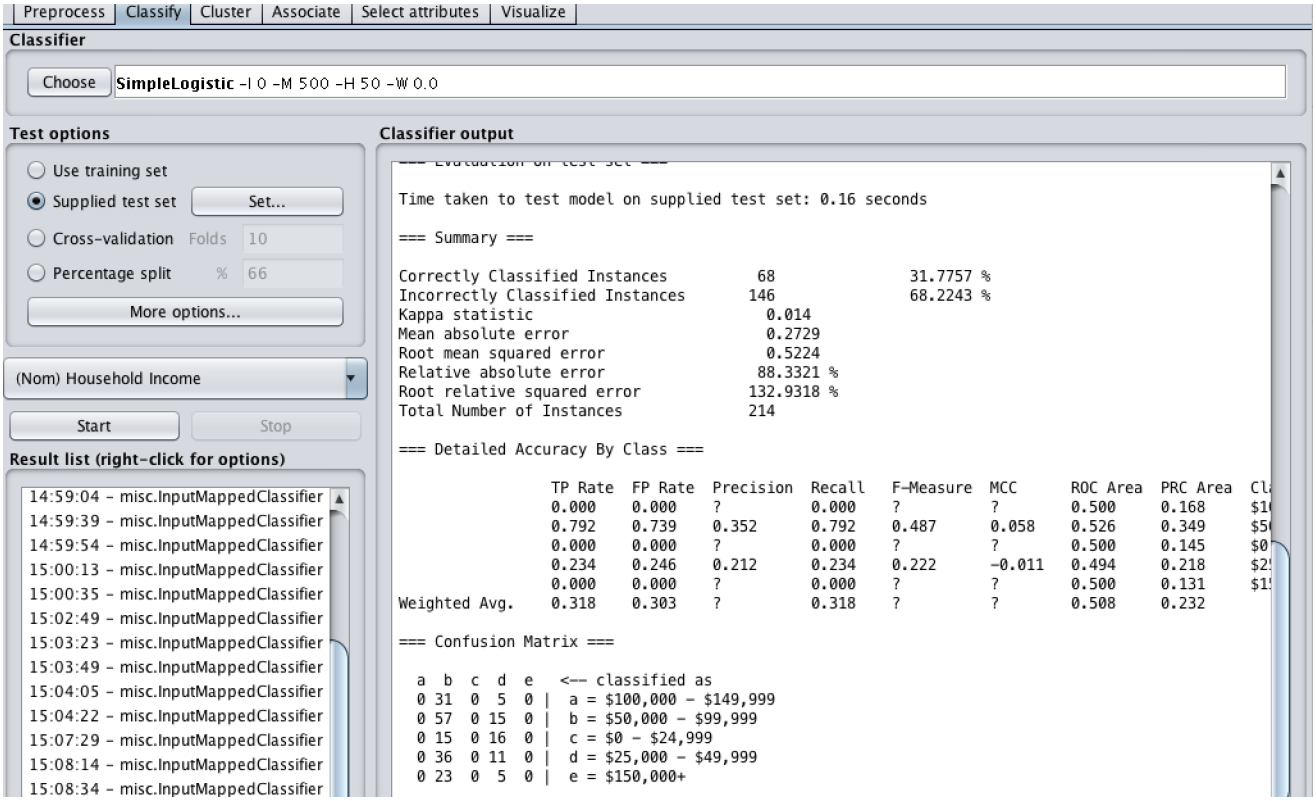
b) J48 Classification Model



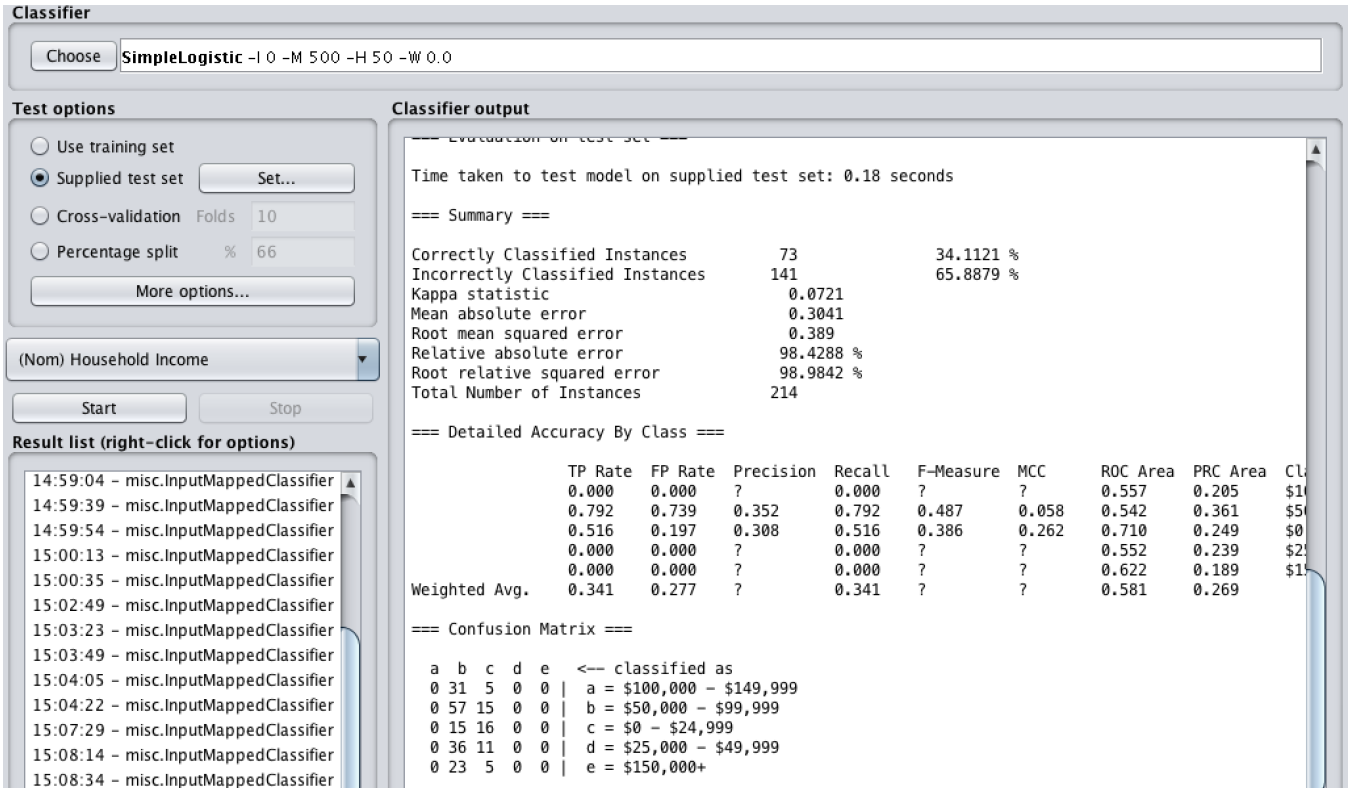
c) Random Forest Classification Model



d) OneR Classification Model



e) Simple Logistic Classification Model



V. Model Performance Evaluation

As shown in each of the 25 classification models, we used our training set data (*Latest_Project_Dataset_training.arff)* to build our classification models and used our test set data (*Latest_Project_Dataset_Test.arff)* to test the performance of each model using the 'Supplied test set' option in the 'Test Options' section.

Comparing the performance of all Classification Models:

We first decided to examine each of our attribute sets and chose one classification model each attribute that had the highest model performance based on accuracy rates:

Attribute Set	Best Performance Model	Accuracy
CorrelationAttributeEval	Random Forest	87.38%
OneRAttributeEval Set	Random Forest	91.12%
GainRatioAttributeEval Set	Random Forest	85.51%
InfoGainAttributeEval Set	Random Forest	87.85%
Our Own Attribute Set	Random Forest	92.99%

From our analysis, we can see that the Random Forest models result in substantially higher performance levels across all attribute sets. Not only does Random Forest exhibit higher accuracy rates but the area under the ROC curve for all the Random Forest classification models is at least 0.987. Since the closer the area under the ROC curve is to 1, the better the predictions made by that classification model, we can say that the Random Forest models predict far more True Positives than they do False Positives. This is also supported by the models' consistently high True Positive Rates and consistently low False positive rates across all attribute sets. Please see the table below for a consolidation of these statistics:

True Positive and False Positive Rates for the Random Forest models across all Attribute Sets:

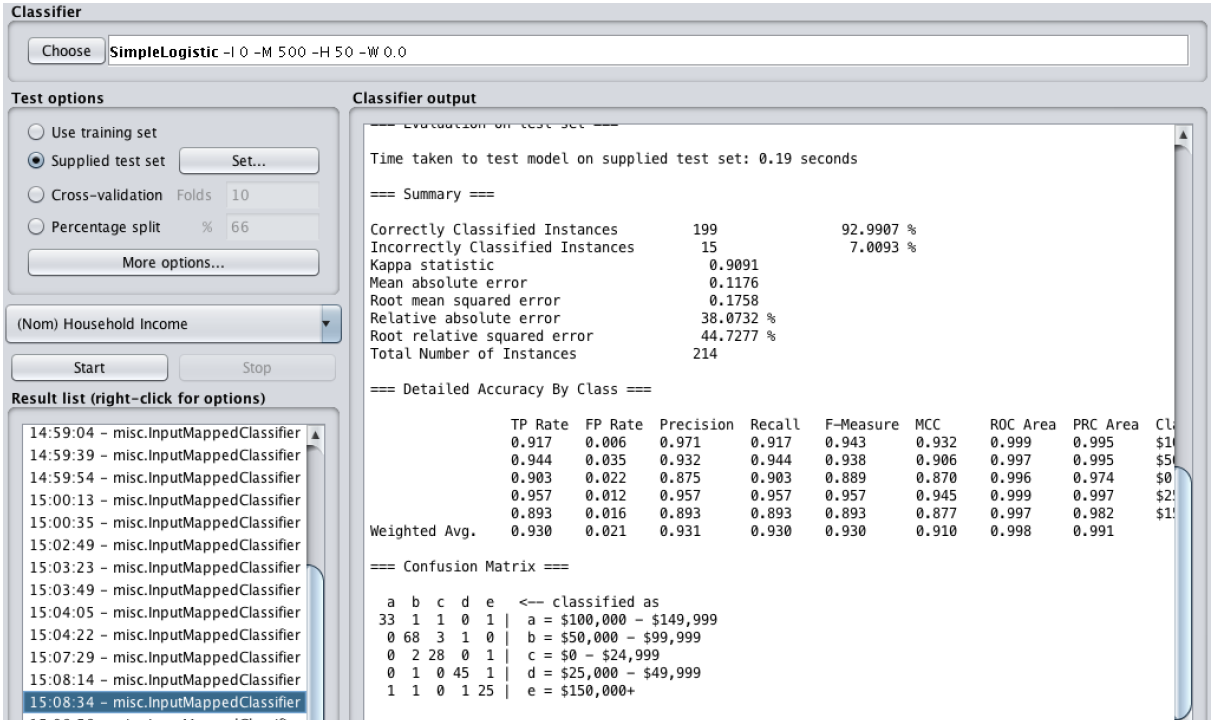
Attribute Set	TP Rate	FP Rate
CorrelationAttributeEval	0.874	0.045
OneRAttributeEval Set	0.911	0.026
GainRatioAttributeEval Set	0.855	0.047
InfoGainAttributeEval Set	0.879	0.043
Our Own Attribute Set	0.930	0.021

In addition, these rates are significantly superior when compared to those of the various other classification models. We found that the TP Rates across all other classification models fell between 0.3 and 0.5, which is significantly lower than those for the Random Forest classification models. The FP rates across all other classification models fell between 0.15 and 0.35, which is significantly higher than those of the Random Forest models. These statistics indicate that the Random Forest models correctly classified far greater tuples than any of the other classification models.

Selecting One Model for Our Data Mining Goal:

Taking a closer look at the various attribute sets used with each of the Random Forest models, we identified **the Random Forest classification model using our own selected attribute set as the best performance model for our data mining goal.** This is because this model offers the highest performance

at almost 98% accuracy, an ROC Area of 0.998, and a TP Rate of 0.93. Please see a screenshot of this classification model performance evaluation below:



VI. Conclusion

At the start of this data mining project, we set out to build a classification model that will accurately predict the household income bracket of a person based on his or her preferences in various international cuisines. Through different combinations of various attribute selection algorithms combined with different classification algorithms, we have found that the Random Forest classification model using the following attribute set to give us best prediction results: US, France, Spain, Japan, Germany, Gender, Age, and Education.

- This attribute set gave us the highest performance levels possibly because we as a team had chosen these attributes together and had based our choices on the most frequently occurring attributes across all other attribute selection algorithms, giving us the strongest attributes from all the attribute selection algorithms.

Lessons learned from this project:

- Throughout this project, we got a practical understanding of what it is like to work with actual data in the data mining process. We realize that not all data is going to be preprocessed, cleansed, or 'pretty' and may include several missing attributes or class labels that first need to be handled before we can begin building our classification models.
- In addition, not all decisions are made cut and dry in a rule based fashion. We as data miners need to make executive decisions regarding the removal or selection of certain attributes. For instance, while the selection algorithms determine the rank of each attribute, what determines how *many* attributes we should include when building our classification models? Certain logical decisions

need to be made by us as data miners in order to make sure we get the best results without distorting the integrity of the results.

Overall, this project gave us a thorough understanding of how we can use various classification algorithms on real-world datasets. Our best model gives us indication of a person's household income based on their preferences on international cuisines at 93% accuracy. In terms of practicality, this can be quite useful to cities, retail centers, or communities that are trying to plan what types of restaurants they should include in their malls or cities based on the type of customer demographic they are trying to attract. If say, a mall is trying to attract wealthier visitors, it may consider using this classification model to determine which cuisines to include in their mall such that wealthier visitors will be most compelled to visit.

Distribution of Project Work throughout group project:

Xi You and Alisha both worked closely together throughout the entire project process to make sound decisions on data cleansing, reduction, attribute selection, model development, and performance evaluation. I (Alisha) later wrote the write up for the report with detailed description, screenshots, and explanations of our decisions.