

# 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 respondents 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 country'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
    - Herzegovina
    - 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	2: Generally speaking, how would you rate all. are you	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 you like the traditional cuisine of A
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	N/A	N/A
4	3.30887909...	Novice	Not much	N/A	3	N/A
5	3.30887167...	Novice	Not much	N/A	4	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	N/A	N/A	N/A
...	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	N/A	N/A	N/A
...	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	N/A	N/A	N/A
...	3.30880004...	Novice	Some	N/A	N/A	N/A
...	3.30879944...	Novice	Not at all	N/A	1	N/A
...	3.30879367...	Intermediate	Some	N/A	3	N/A
...	3.30879217E9	Intermediate	Not much	N/A	N/A	N/A
...	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

1: How much you like the traditional cuisine 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 the traditional cuisine of Honduras.
3	N/A	4	4	N/A
4	N/A	4	4	3
4	N/A	Tuples with Missing values for all attributes	N/A	N/A
2	N/A	4	4	N/A
4	N/A	5	4	N/A
N/A	N/A	N/A	N/A	N/A
3	4	3	3	N/A
3	N/A	3	3	3
4	N/A	5	N/A	N/A
5	N/A	5	N/A	N/A
5	N/A	3	N/A	N/A
3	N/A	4	N/A	N/A
3	N/A	3	3	3
3	N/A	N/A	N/A	N/A
5	3	2	N/A	N/A
4	1	5	N/A	N/A
4	N/A	4	3	N/A
N/A	N/A	N/A	N/A	N/A
N/A	N/A	4	N/A	N/A
N/A	N/A	4	N/A	N/A
2	N/A	2	N/A	N/A
3	N/A	4	N/A	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	Some Class Attribute Instances have missing class value.	Male	30-44	\$25,000 - \$49,999	High schoo...	Pacific
3	Will not be able to use for training or test data as we need class value to develop model (training data) and need class value to test model performance (test data)	Female	30-44	\$0,000 - \$99,999	Graduate ...	East North Central
N/A		Male	45-60	\$0 - \$24,999	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	45-60	\$0 - \$24,999	Some colle...	South Atlantic
N/A		Male	18-29	\$25,000 - \$49,999	High schoo...	Middle Atlantic
2		Female	30-44	\$0 - \$24,999	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	45-60	\$0 - \$24,999	Some colle...	Middle Atlantic
N/A		Male	18-29	\$0 - \$24,999	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	45-60	\$100,000 - \$149,...	Graduate ...	East North Central
3		Male	30-44	\$50,000 - \$99,999	Some colle...	Mountain
4		Male	45-60	\$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

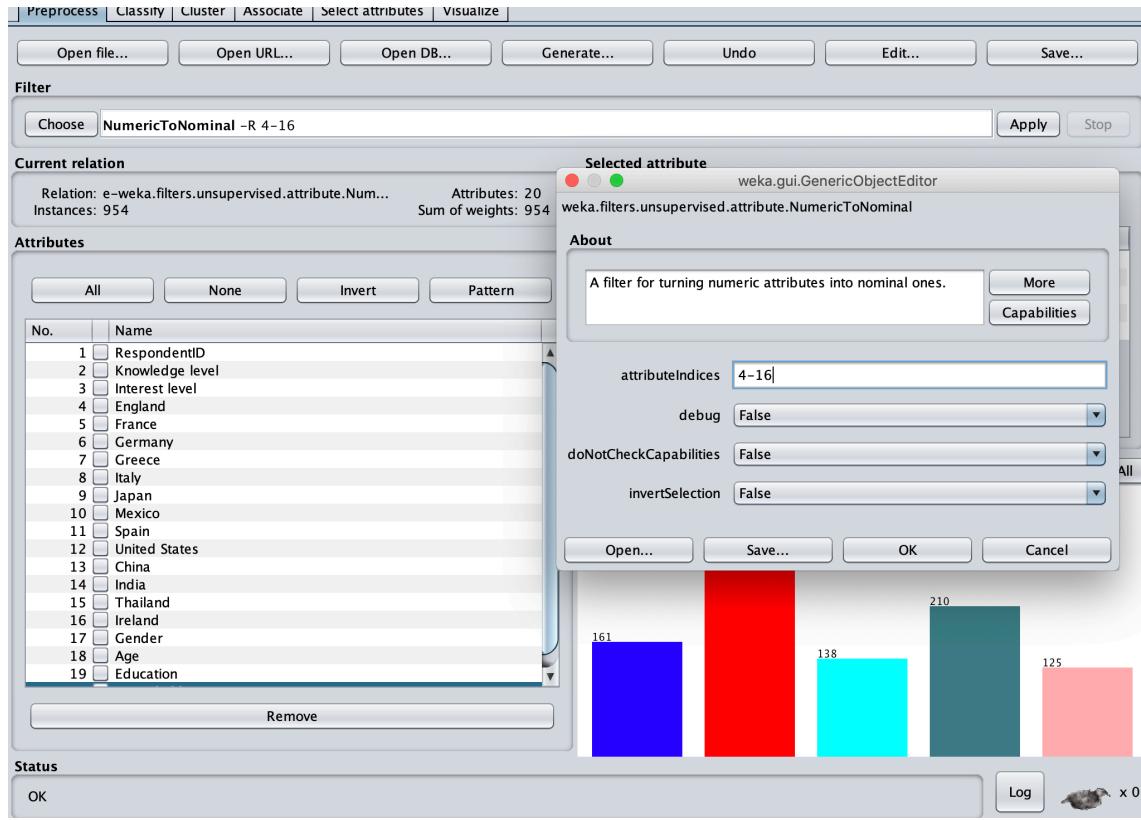


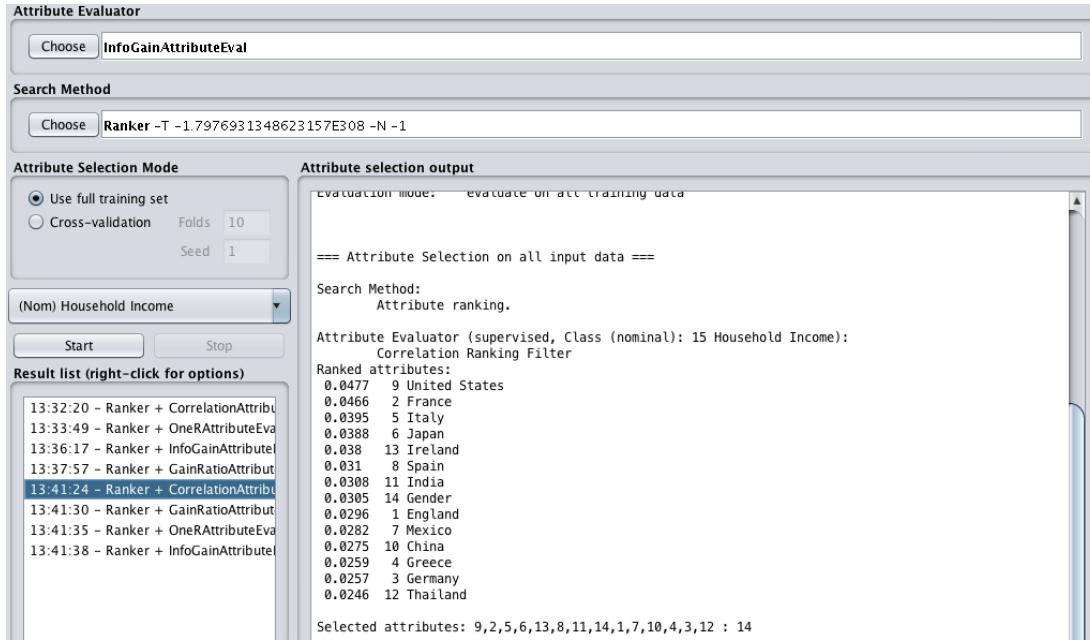
Figure 5: Dataset with Replaced Missing Values

	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 Income
	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	None	4	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	5	3	5	5	3	Male	30-44	Graduate ... \$150,000 - \$199,99	
Novice	Not much	3	3	3	4	4	3	3	4	3	3	3	3	3	3	Male	45-60	Less than ... \$10 - \$24,999	
Novice	Not much	2	3	3	4	4	2	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	2	3	5	5	3	Female	30-44	Graduate ... \$150,000 - \$199,99	
Advanced	A lot	3	5	4	4	5	5	4	4	5	5	4	4	3	4	Male	45-60	Some colle... \$10 - \$24,999	
Novice	Not much	3	4	4	4	4	3	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	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	4	3	Male	16-29	Some colle... \$10 - \$24,999	
Intermediate	Some	5	4	5	5	4	5	4	4	3	3	4	5	5	3	Male	18-29	High schoo... \$25,000 - \$49,999	
Intermediate	Some	3	4	3	4	4	4	4	5	3	3	3	3	3	3	Female	30-44	Graduate ... \$150,000 - \$199,999	
Intermediate	Some	3	4	3	4	4	4	4	5	3	3	3	3	3	3	Female	30-44	Some colle... \$25,000 - \$49,999	
Novice	Not at all	3	4	4	3	4	3	4	4	4	4	4	4	4	3	Male	30-44	Some colle... \$150,000 - \$199,999	
Novice	Not much	3	4	3	3	3	5	4	3	4	4	4	4	4	3	Male	30-44	Some colle... \$10 - \$24,999	
Intermediate	A lot	4	3	4	4	3	4	4	4	4	4	5	4	4	3	Male	18-29	Bachelor d... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	30-44	Some colle... \$25,000 - \$49,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	30-44	Some colle... \$25,000 - \$49,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	30-44	Some colle... \$25,000 - \$49,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	30-44	Some colle... \$25,000 - \$49,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	30-44	Some colle... \$25,000 - \$49,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	30-44	Some colle... \$25,000 - \$49,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	30-44	Some colle... \$25,000 - \$49,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	30-44	Some colle... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	30-44	Some colle... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	30-44	Some colle... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	30-44	Some colle... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	30-44	Some colle... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	30-44	Some colle... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	30-44	Some colle... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	30-44	Some colle... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3	5	2	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	5	3	3	2	4	4	4	4	4	4	4	5	2	Male	18-29	Some colle... \$10 - \$24,999	
Novice	Some	3	4	4	4	4	5	3											

### III. Attribute Selection

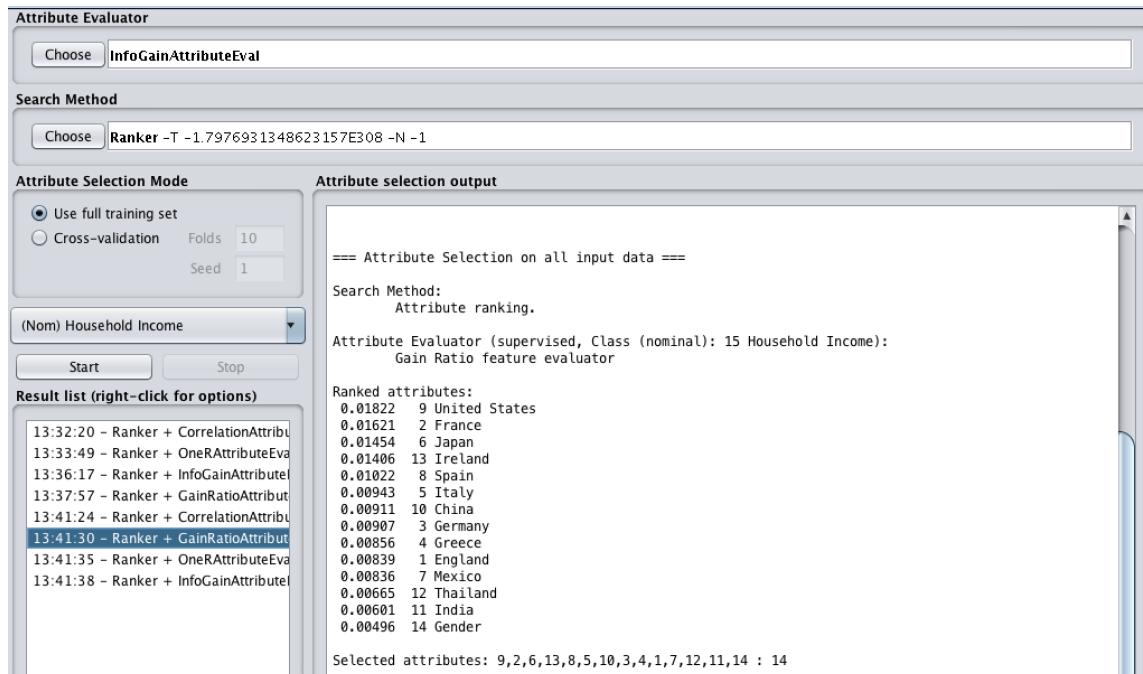
- We ran four different attribute selection algorithms on Weka to gauge the rank of each attribute in relation to the class attribute and selected the top eight attributes for each algorithm. Please find below various attribute selection algorithms and selected attributes we ran in Figures 7 – 10 below:

Figure 7: CorrelationAttributeEval Algorithm



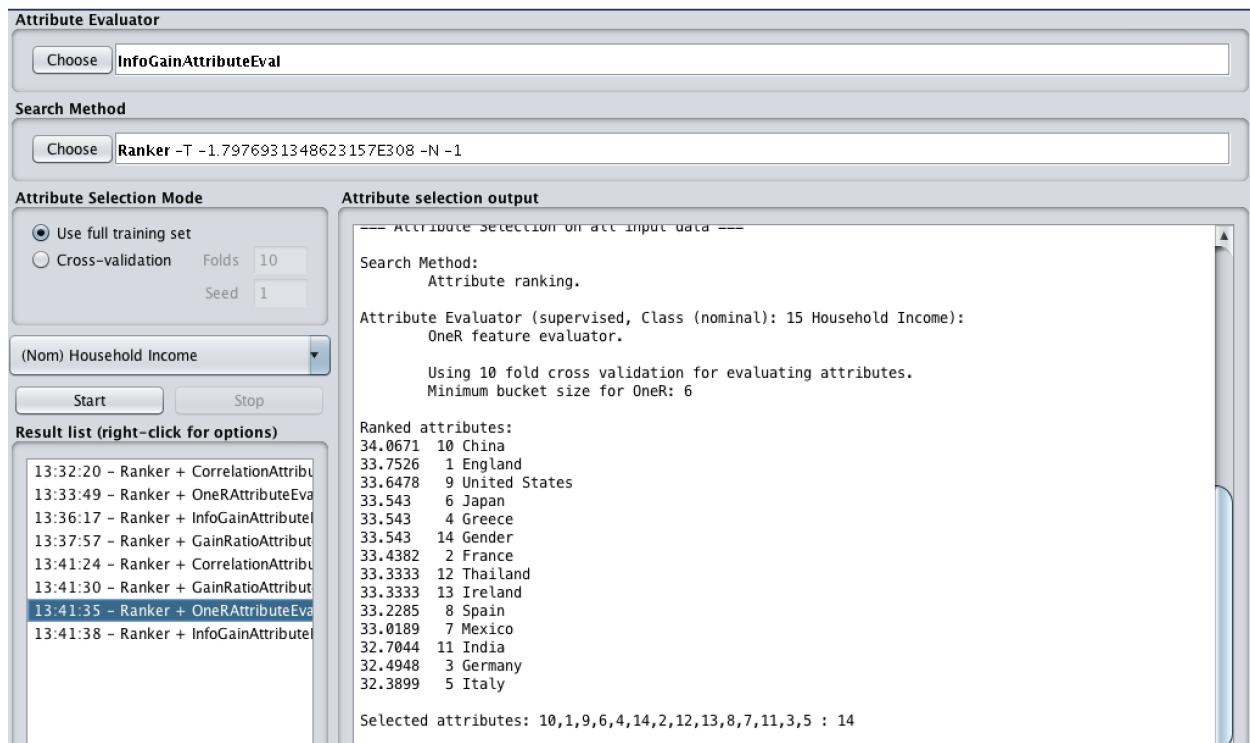
Attributes Selected: United States, France, Italy, Japan, Ireland, Spain, India, Gender.

Figure 8: GainRatioAttributeEval Algorithm

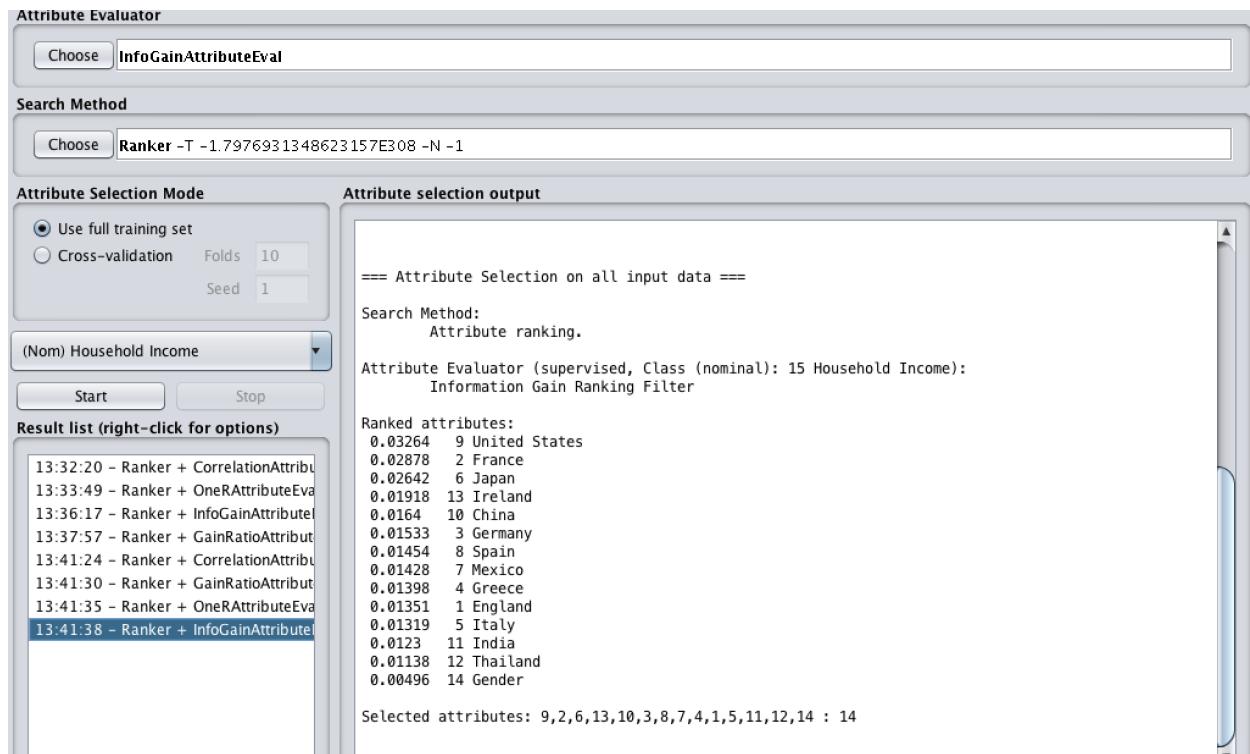


Attributes Selected: United States, France, Japan, Ireland, Spain, Italy, China, Germany.

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:

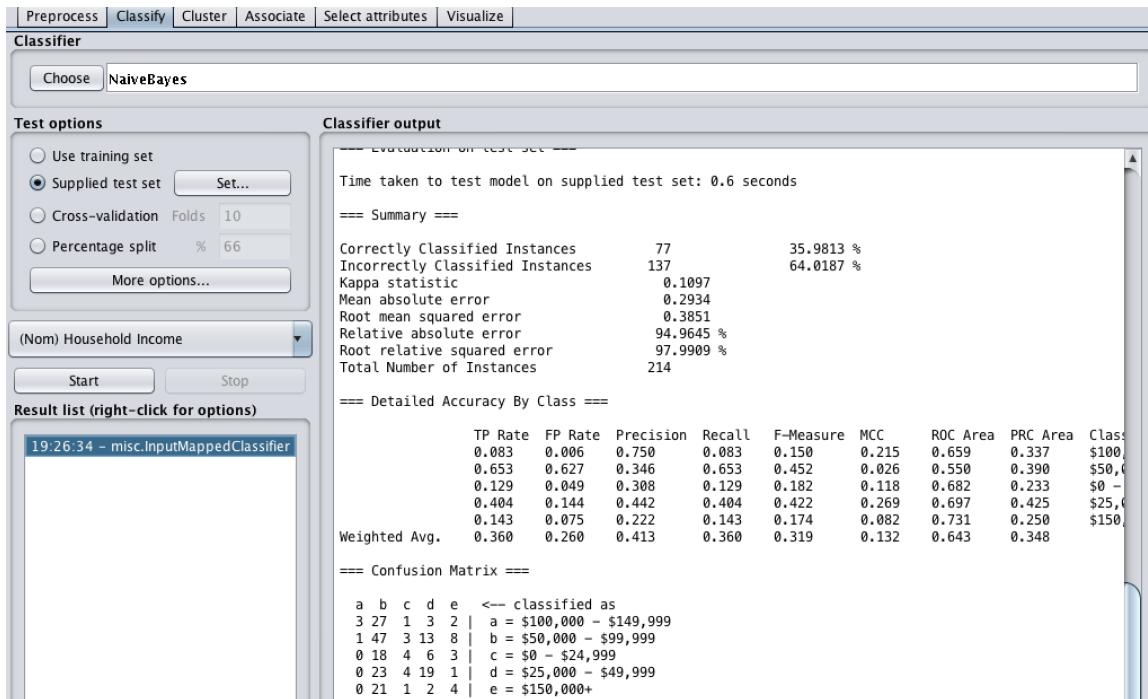
Attribute Set	Attributes Selected
CorrelationAttributeEval	United States, France, Italy, Japan, Ireland, Spain, India, Gender
GainRatioAttributeEval	United States, France, Japan, Ireland, Spain, Italy, China, Germany.
OneRAttributeEval Algorithm	Attributes selected: China, England, United States, Japan, 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

**Classifier**

Choose SimpleLogistic -I 0 -M 500 -H 50 -W 0.0

**Test options**

- Use training set
- Supplied test set [Set...](#)
- Cross-validation Folds 10
- Percentage split % 66

[More options...](#)

(Nom) Household Income

[Start](#) [Stop](#)

**Result list (right-click for options)**

- 14:48:41 - misc.InputMappedClassifier
- 14:49:33 - misc.InputMappedClassifier
- 14:49:52 - misc.InputMappedClassifier
- 14:50:10 - misc.InputMappedClassifier
- 14:50:20 - misc.InputMappedClassifier

**Classifier output**

Time taken to test model on supplied test set: 0.74 seconds

==== Summary ===

	Correctly Classified Instances	92	42.9907 %
Incorrectly Classified Instances	122	57.0093 %	
Kappa statistic	0.173		
Mean absolute error	0.2698		
Root mean squared error	0.3666		
Relative absolute error	87.3163 %		
Root relative squared error	93.2749 %		
Total Number of Instances	214		

==== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.222	0.022	0.667	0.222	0.333	0.325	0.696	0.408	\$100,000+	
0.944	0.796	0.376	0.944	0.538	0.195	0.643	0.480	\$50,000+	
0.032	0.000	1.000	0.032	0.063	0.166	0.645	0.244	\$0 - \$24,999	
0.170	0.006	0.889	0.170	0.286	0.339	0.662	0.397	\$25,000+	
0.250	0.022	0.636	0.250	0.359	0.349	0.694	0.383	\$150,000+	
Weighted Avg.	0.430	0.276	0.662	0.430	0.356	0.264	0.663	0.403	

==== Confusion Matrix ===

a b c d e	<-- classified as
8 26 0 0 2	a = \$100,000 - \$149,999
3 68 0 0 1	b = \$50,000 - \$99,999
1 28 1 1 0	c = \$0 - \$24,999
0 38 0 8 1	d = \$25,000 - \$49,999
0 21 0 0 7	e = \$150,000+

## c) Random Forest Classification Model

**Classifier**

Choose SimpleLogistic -I 0 -M 500 -H 50 -W 0.0

**Test options**

- Use training set
- Supplied test set [Set...](#)
- Cross-validation Folds 10
- Percentage split % 66

[More options...](#)

(Nom) Household Income

[Start](#) [Stop](#)

**Result list (right-click for options)**

- 14:48:41 - misc.InputMappedClassifier
- 14:49:33 - misc.InputMappedClassifier
- 14:49:52 - misc.InputMappedClassifier
- 14:50:10 - misc.InputMappedClassifier
- 14:50:20 - misc.InputMappedClassifier

**Classifier output**

Time taken to test model on supplied test set: 0.52 seconds

==== Summary ===

	Correctly Classified Instances	187	87.3832 %
Incorrectly Classified Instances	27	12.6168 %	
Kappa statistic	0.8351		
Mean absolute error	0.1355		
Root mean squared error	0.2039		
Relative absolute error	43.8744 %		
Root relative squared error	51.8848 %		
Total Number of Instances	214		

==== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.722	0.017	0.897	0.722	0.800	0.771	0.989	0.952	\$100,000+	
0.931	0.085	0.848	0.931	0.887	0.828	0.990	0.981	\$50,000+	
0.871	0.011	0.931	0.871	0.900	0.884	0.998	0.988	\$0 - \$24,999	
0.894	0.054	0.824	0.894	0.857	0.816	0.995	0.982	\$25,000+	
0.893	0.005	0.962	0.893	0.926	0.916	0.995	0.975	\$150,000+	
Weighted Avg.	0.874	0.045	0.878	0.874	0.873	0.836	0.993	0.977	

==== Confusion Matrix ===

a b c d e	<-- classified as
26 4 1 4 1	a = \$100,000 - \$149,999
1 67 1 3 0	b = \$50,000 - \$99,999
0 3 27 1 0	c = \$0 - \$24,999
2 3 0 42 0	d = \$25,000 - \$49,999
0 2 0 1 25	e = \$150,000+

#### d) OneR Classification Model

**Classifier**

Choose SimpleLogistic -I 0 -M 500 -H 50 -W 0.0

**Test options**

- Use training set
- Supplied test set [Set...](#)
- Cross-validation Folds 10
- Percentage split % 66
- [More options...](#)

(Nom) Household Income

[Start](#) [Stop](#)

**Result list (right-click for options)**

- 14:48:41 - misc.InputMappedClassifier
- 14:49:33 - misc.InputMappedClassifier
- 14:49:52 - misc.InputMappedClassifier
- 14:50:10 - misc.InputMappedClassifier**
- 14:50:20 - misc.InputMappedClassifier

**Classifier output**

Time taken to test model on supplied test set: 0.71 seconds

== Summary ==

	Correctly Classified Instances	71	33.1776 %
Incorrectly Classified Instances	143	66.8224 %	
Kappa statistic	0.0101		
Mean absolute error	0.2673		
Root mean squared error	0.517		
Relative absolute error	86.5171 %		
Root relative squared error	131.559 %		
Total Number of Instances	214		

== Detailed Accuracy By Class ==

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	?	0.500	0.168	\$100	
0.903	0.901	0.337	0.903	0.491	0.002	0.501	0.337	\$50,0	
0.000	0.000	?	0.000	?	?	0.500	0.145	\$0 -	
0.128	0.090	0.286	0.128	0.176	0.053	0.519	0.228	\$25,0	
0.000	0.000	?	0.000	?	?	0.500	0.131	\$150,	
Weighted Avg.	0.332	0.323	?	0.332	?	0.504	0.230		

== Confusion Matrix ==

a b c d e	<-- classified as
0 34 0 2 0   a = \$100,000 - \$149,999	
0 65 0 7 0   b = \$50,000 - \$99,999	
0 26 0 5 0   c = \$0 - \$24,999	
0 41 0 6 0   d = \$25,000 - \$49,999	
0 27 0 1 0   e = \$150,000+	

#### e) Simple Logistic Classification Model

**Classifier**

Choose SimpleLogistic -I 0 -M 500 -H 50 -W 0.0

**Test options**

- Use training set
- Supplied test set [Set...](#)
- Cross-validation Folds 10
- Percentage split % 66
- [More options...](#)

(Nom) Household Income

[Start](#) [Stop](#)

**Result list (right-click for options)**

- 14:48:41 - misc.InputMappedClassifier
- 14:49:33 - misc.InputMappedClassifier
- 14:49:52 - misc.InputMappedClassifier
- 14:50:10 - misc.InputMappedClassifier
- 14:50:20 - misc.InputMappedClassifier**

**Classifier output**

Time taken to test model on supplied test set: 0.3 seconds

== Summary ==

	Correctly Classified Instances	71	33.1776 %
Incorrectly Classified Instances	143	66.8224 %	
Kappa statistic	0.0249		
Mean absolute error	0.3025		
Root mean squared error	0.3881		
Relative absolute error	97.9264 %		
Root relative squared error	98.7641 %		
Total Number of Instances	214		

== Detailed Accuracy By Class ==

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	?	0.584	0.210	\$100	
0.792	0.831	0.326	0.792	0.462	-0.048	0.536	0.368	\$50,0	
0.000	0.000	?	0.000	?	?	0.644	0.243	\$0 -	
0.298	0.144	0.368	0.298	0.329	0.167	0.682	0.371	\$25,0	
0.000	0.005	0.000	0.000	0.000	-0.027	0.698	0.207	\$150,	
Weighted Avg.	0.332	0.312	?	0.332	?	0.613	0.303		

== Confusion Matrix ==

a b c d e	<-- classified as
0 31 0 5 0   a = \$100,000 - \$149,999	
0 57 0 15 0   b = \$50,000 - \$99,999	
0 27 0 3 1   c = \$0 - \$24,999	
0 33 0 14 0   d = \$25,000 - \$49,999	
0 27 0 1 0   e = \$150,000+	

## OneRAttributeEval Set:

### a) Naïve Bayes Model

**Classifier**

Choose J48 -C 0.25 -M 2

**Test options**

- Use training set
- Supplied test set [Set...](#)
- Cross-validation Folds 10
- Percentage split % 66 [More options...](#)

(Nom) Household Income

[Start](#) [Stop](#)

**Result list (right-click for options)**

- 14:48:41 - misc.InputMappedClassifier
- 14:49:33 - misc.InputMappedClassifier
- 14:49:52 - misc.InputMappedClassifier
- 14:50:10 - misc.InputMappedClassifier
- 14:50:20 - misc.InputMappedClassifier
- 14:53:41 - misc.InputMappedClassifier

**Classifier output**

Time taken to test model on supplied test set: 0.3 seconds

==== Summary ===

	Correctly Classified Instances	71	33.1776 %
Incorrectly Classified Instances	143	66.8224 %	
Kappa statistic	0.0249		
Mean absolute error	0.3025		
Root mean squared error	0.3881		
Relative absolute error	97.9264 %		
Root relative squared error	98.7641 %		
Total Number of Instances	214		

==== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	?	0.584	0.210	\$100,000+	
0.792	0.831	0.326	0.792	0.462	-0.048	0.536	0.368	\$50,000+	
0.000	0.000	?	0.000	?	?	0.644	0.243	\$0 - \$24,999	
0.298	0.144	0.368	0.298	0.329	0.167	0.682	0.371	\$25,000+	
0.000	0.005	0.000	0.000	0.000	-0.027	0.698	0.207	\$150,000+	
Weighted Avg.	0.332	0.312	?	0.332	?	0.613	0.303		

==== Confusion Matrix ===

	a	b	c	d	e	<-- classified as
0	31	0	5	0	1	a = \$100,000 - \$149,999
0	57	0	15	0	1	b = \$50,000 - \$99,999
0	27	0	3	1	1	c = \$0 - \$24,999
0	33	0	14	0	1	d = \$25,000 - \$49,999
0	27	0	1	0	1	e = \$150,000+

### b) J48

**Preprocess** **Classify** **Cluster** **Associate** **Select attributes** **Visualize**

**Classifier**

Choose SimpleLogistic -I 0 -M 500 -H 50 -W 0.0

**Test options**

- Use training set
- Supplied test set [Set...](#)
- Cross-validation Folds 10
- Percentage split % 66 [More options...](#)

(Nom) Household Income

[Start](#) [Stop](#)

**Result list (right-click for options)**

- 14:48:41 - misc.InputMappedClassifier
- 14:49:33 - misc.InputMappedClassifier
- 14:49:52 - misc.InputMappedClassifier
- 14:50:10 - misc.InputMappedClassifier
- 14:50:20 - misc.InputMappedClassifier
- 14:56:30 - misc.InputMappedClassifier
- 14:56:47 - misc.InputMappedClassifier
- 14:56:54 - misc.InputMappedClassifier
- 14:57:02 - misc.InputMappedClassifier
- 14:57:09 - misc.InputMappedClassifier

**Classifier output**

Time taken to test model on supplied test set: 0.12 seconds

==== Summary ===

	Correctly Classified Instances	120	56.0748 %
Incorrectly Classified Instances	94	43.9252 %	
Kappa statistic	0.4195		
Mean absolute error	0.2348		
Root mean squared error	0.3415		
Relative absolute error	76.0131 %		
Root relative squared error	86.8921 %		
Total Number of Instances	214		

==== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.500	0.073	0.581	0.500	0.537	0.454	0.791	0.534	\$100,000+	
0.694	0.261	0.575	0.694	0.629	0.417	0.796	0.676	\$50,000+	
0.387	0.060	0.522	0.387	0.444	0.372	0.840	0.449	\$0 - \$24,999	
0.596	0.150	0.528	0.596	0.560	0.428	0.819	0.535	\$25,000+	
0.429	0.043	0.600	0.429	0.500	0.447	0.841	0.467	\$150,000+	
Weighted Avg.	0.561	0.147	0.561	0.561	0.555	0.812	0.561		

==== Confusion Matrix ===

	a	b	c	d	e	<-- classified as
18	11	2	4	1	1	a = \$100,000 - \$149,999
5	50	3	11	3	1	b = \$50,000 - \$99,999
4	8	12	6	1	1	c = \$0 - \$24,999
4	10	2	28	3	1	d = \$25,000 - \$49,999
0	8	4	4	12	1	e = \$150,000+

### c) Random Forest

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

**Classifier**

Choose SimpleLogistic -I 0 -M 500 -H 50 -W 0.0

Test options		Classifier output	
<input type="radio"/> Use training set		Evaluation on test set	
<input checked="" type="radio"/> Supplied test set	Set...	Time taken to test model on supplied test set: 0.12 seconds	
<input type="radio"/> Cross-validation	Folds 10		
<input type="radio"/> Percentage split	% 66		
<a href="#">More options...</a>			
(Nom) Household Income			
<a href="#">Start</a>	<a href="#">Stop</a>		
Result list (right-click for options)			
<pre>14:48:41 - misc.InputMappedClassifier 14:49:33 - misc.InputMappedClassifier 14:49:52 - misc.InputMappedClassifier 14:50:10 - misc.InputMappedClassifier 14:50:20 - misc.InputMappedClassifier 14:56:30 - misc.InputMappedClassifier 14:56:47 - misc.InputMappedClassifier 14:56:54 - misc.InputMappedClassifier 14:57:02 - misc.InputMappedClassifier 14:57:09 - misc.InputMappedClassifier</pre>			

**Classifier output**

Time taken to test model on supplied test set: 0.12 seconds

== Summary ==

Correctly Classified Instances	195	91.1215 %
Incorrectly Classified Instances	19	8.8785 %
Kappa statistic	0.8846	
Mean absolute error	0.1362	
Root mean squared error	0.2047	
Relative absolute error	44.0927 %	
Root relative squared error	52.0999 %	
Total Number of Instances	214	

== Detailed Accuracy By Class ==

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.889	0.017	0.914	0.889	0.901	0.882	0.988	0.958	\$100,000+	
0.958	0.035	0.932	0.958	0.945	0.917	0.993	0.986	\$50,000+	
0.871	0.033	0.818	0.871	0.844	0.817	0.990	0.944	\$0 - \$24,999	
0.936	0.030	0.898	0.936	0.917	0.893	0.991	0.976	\$25,000+	
0.821	0.000	1.000	0.821	0.902	0.894	0.993	0.968	\$150,000+	
Weighted Avg.	0.911	0.026	0.914	0.911	0.888	0.991	0.971		

== Confusion Matrix ==

a b c d e	<-- classified as
32 2 1 1 0	a = \$100,000 - \$149,999
1 69 2 0 0	b = \$50,000 - \$99,999
0 3 27 1 0	c = \$0 - \$24,999
2 0 1 44 0	d = \$25,000 - \$49,999
0 0 2 3 23	e = \$150,000+

### d) OneR

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

**Classifier**

Choose SimpleLogistic -I 0 -M 500 -H 50 -W 0.0

Test options		Classifier output	
<input type="radio"/> Use training set		Evaluation on test set	
<input checked="" type="radio"/> Supplied test set	Set...	Time taken to test model on supplied test set: 0.1 seconds	
<input type="radio"/> Cross-validation	Folds 10		
<input type="radio"/> Percentage split	% 66		
<a href="#">More options...</a>			
(Nom) Household Income			
<a href="#">Start</a>	<a href="#">Stop</a>		
Result list (right-click for options)			
<pre>14:48:41 - misc.InputMappedClassifier 14:49:33 - misc.InputMappedClassifier 14:49:52 - misc.InputMappedClassifier 14:50:10 - misc.InputMappedClassifier 14:50:20 - misc.InputMappedClassifier 14:56:30 - misc.InputMappedClassifier 14:56:47 - misc.InputMappedClassifier 14:56:54 - misc.InputMappedClassifier 14:57:02 - misc.InputMappedClassifier 14:57:09 - misc.InputMappedClassifier</pre>			

**Classifier output**

Time taken to test model on supplied test set: 0.1 seconds

== Summary ==

Correctly Classified Instances	73	34.1121 %
Incorrectly Classified Instances	141	65.8879 %
Kappa statistic	0.0095	
Mean absolute error	0.2636	
Root mean squared error	0.5134	
Relative absolute error	85.3071 %	
Root relative squared error	130.6358 %	
Total Number of Instances	214	

== Detailed Accuracy By Class ==

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	?	0.500	0.168	\$100,000+	
1.000	0.979	0.341	1.000	0.509	0.085	0.511	0.341	\$50,000+	
0.000	0.000	?	0.000	?	?	0.500	0.145	\$0 - \$24,999	
0.021	0.012	0.333	0.021	0.040	0.033	0.505	0.222	\$25,000+	
0.000	0.000	?	0.000	?	?	0.500	0.131	\$150,000+	
Weighted Avg.	0.341	0.332	?	0.341	?	?	0.505	0.230	

== Confusion Matrix ==

a b c d e	<-- classified as
0 36 0 0 0	a = \$100,000 - \$149,999
0 72 0 0 0	b = \$50,000 - \$99,999
0 30 0 1 0	c = \$0 - \$24,999
0 46 0 1 0	d = \$25,000 - \$49,999
0 27 0 1 0	e = \$150,000+

### e) Simple Logistic

Preprocess Classify Cluster Associate Select attributes Visualize

**Classifier**

Choose SimpleLogistic -I 0 -M 500 -H 50 -W 0.0

**Test options**

- Use training set
- Supplied test set Set...
- Cross-validation Folds 10
- Percentage split % 66
- [More options...](#)

(Nom) Household Income

Start Stop

**Result list (right-click for options)**

- 14:48:41 - misc.InputMappedClassifier
- 14:49:33 - misc.InputMappedClassifier
- 14:49:52 - misc.InputMappedClassifier
- 14:50:10 - misc.InputMappedClassifier
- 14:50:20 - misc.InputMappedClassifier
- 14:56:30 - misc.InputMappedClassifier
- 14:56:47 - misc.InputMappedClassifier
- 14:56:54 - misc.InputMappedClassifier
- 14:57:02 - misc.InputMappedClassifier
- 14:57:09 - misc.InputMappedClassifier

**Classifier output**

Time taken to test model on supplied test set: 0.11 seconds

== Summary ==

	Correctly Classified Instances	73	34.1121 %
Incorrectly Classified Instances	141	65.8879 %	
Kappa statistic	0.0109		
Mean absolute error	0.3064		
Root mean squared error	0.3908		
Relative absolute error	99.1744 %		
Root relative squared error	99.4536 %		
Total Number of Instances	214		

== Detailed Accuracy By Class ==

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	?	0.598	0.218	\$100+	
1.000	0.972	0.343	1.000	0.511	0.098	0.516	0.346	\$50,-	
0.000	0.000	?	0.000	?	?	0.541	0.188	\$0,-	
0.021	0.012	0.333	0.021	0.040	0.033	0.612	0.294	\$25,-	
0.000	0.005	0.000	0.000	0.000	-0.027	0.640	0.196	\$150+	
Weighted Avg.	0.341	0.330	?	0.341	?	0.571	0.271		

== Confusion Matrix ==

a b c d e	<-- classified as
0 36 0 0 0	a = \$100,000 - \$149,999
0 72 0 0 0	b = \$50,000 - \$99,999
0 29 0 1 1	c = \$0 - \$24,999
0 46 0 1 0	d = \$25,000 - \$49,999
0 27 0 1 0	e = \$150,000+

### GainRatioAttributeEval Set:

#### a) Naïve Bayes Classification Model

Preprocess Classify Cluster Associate Select attributes Visualize

**Classifier**

Choose SimpleLogistic -I 0 -M 500 -H 50 -W 0.0

**Test options**

- Use training set
- Supplied test set Set...
- Cross-validation Folds 10
- Percentage split % 66
- [More options...](#)

(Nom) Household Income

Start Stop

**Result list (right-click for options)**

- 14:48:41 - misc.InputMappedClassifier
- 14:49:33 - misc.InputMappedClassifier
- 14:49:52 - misc.InputMappedClassifier
- 14:50:10 - misc.InputMappedClassifier
- 14:50:20 - misc.InputMappedClassifier
- 14:56:30 - misc.InputMappedClassifier
- 14:56:47 - misc.InputMappedClassifier
- 14:56:54 - misc.InputMappedClassifier
- 14:57:02 - misc.InputMappedClassifier
- 14:57:09 - misc.InputMappedClassifier
- 14:59:04 - misc.InputMappedClassifier
- 14:59:39 - misc.InputMappedClassifier
- 14:59:54 - misc.InputMappedClassifier

**Classifier output**

Time taken to test model on supplied test set: 0.15 seconds

== Summary ==

	Correctly Classified Instances	73	34.1121 %
Incorrectly Classified Instances	141	65.8879 %	
Kappa statistic	0.0109		
Mean absolute error	0.3069		
Root mean squared error	0.3914		
Relative absolute error	99.3388 %		
Root relative squared error	99.6034 %		
Total Number of Instances	214		

== Detailed Accuracy By Class ==

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	?	0.536	0.184	\$100+	
1.000	0.972	0.343	1.000	0.511	0.098	0.509	0.338	\$50,-	
0.000	0.000	?	0.000	?	?	0.536	0.182	\$0,-	
0.021	0.012	0.333	0.021	0.040	0.033	0.596	0.293	\$25,-	
0.000	0.005	0.000	0.000	0.000	-0.027	0.610	0.182	\$150+	
Weighted Avg.	0.341	0.330	?	0.341	?	0.550	0.259		

== Confusion Matrix ==

a b c d e	<-- classified as
0 36 0 0 0	a = \$100,000 - \$149,999
0 72 0 0 0	b = \$50,000 - \$99,999
0 29 0 1 1	c = \$0 - \$24,999
0 46 0 1 0	d = \$25,000 - \$49,999
0 27 0 1 0	e = \$150,000+

## b) J48 Classification Model

**Classifier**

Choose **J48 -C 0.25 -M 2**

**Test options**

- Use training set
- Supplied test set **Set...**
- Cross-validation Folds **10**
- Percentage split % **66**

**More options...**

(Nom) Household Income

**Start Stop**

**Result list (right-click for options)**

- 14:49:52 - misc.InputMappedClassifier
- 14:50:10 - misc.InputMappedClassifier
- 14:50:20 - misc.InputMappedClassifier
- 14:56:30 - misc.InputMappedClassifier
- 14:56:47 - misc.InputMappedClassifier
- 14:56:54 - misc.InputMappedClassifier
- 14:57:02 - misc.InputMappedClassifier
- 14:57:09 - misc.InputMappedClassifier
- 14:59:04 - misc.InputMappedClassifier
- 14:59:39 - misc.InputMappedClassifier
- 14:59:54 - misc.InputMappedClassifier
- 15:00:13 - misc.InputMappedClassifier
- 15:00:35 - misc.InputMappedClassifier

**Classifier output**

Time taken to test model on supplied test set: 0.17 seconds

== Summary ==

	Correctly Classified Instances	121	56.5421 %
Incorrectly Classified Instances	93	43.4579 %	
Kappa statistic	0.408		
Mean absolute error	0.2265		
Root mean squared error	0.3371		
Relative absolute error	73.313 %		
Root relative squared error	85.7837 %		
Total Number of Instances	214		

== Detailed Accuracy By Class ==

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.361	0.056	0.565	0.361	0.441	0.368	0.808	0.494	\$100,000 - \$149,999	
0.847	0.408	0.513	0.847	0.639	0.417	0.792	0.628	\$50,000 - \$99,999	
0.323	0.027	0.667	0.323	0.435	0.407	0.867	0.565	\$0 - \$24,999	
0.532	0.078	0.658	0.532	0.588	0.492	0.856	0.638	\$25,000 - \$49,999	
0.429	0.038	0.632	0.429	0.511	0.463	0.842	0.546	\$150,000+	
Weighted Avg.	0.565	0.173	0.591	0.565	0.548	0.430	0.826	0.588	

== Confusion Matrix ==

a b c d e	<-- classified as
13 18 0 3 2	a = \$100,000 - \$149,999
1 61 2 6 2	b = \$50,000 - \$99,999
4 14 10 2 1	c = \$0 - \$24,999
2 16 2 25 2	d = \$25,000 - \$49,999
3 10 1 2 12	e = \$150,000+

## c) Random Forest Classification Model

**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 **10**
- Percentage split % **66**

**More options...**

(Nom) Household Income

**Start Stop**

**Result list (right-click for options)**

- 14:50:10 - misc.InputMappedClassifier
- 14:50:20 - misc.InputMappedClassifier
- 14:56:30 - misc.InputMappedClassifier
- 14:56:47 - misc.InputMappedClassifier
- 14:56:54 - misc.InputMappedClassifier
- 14:57:02 - misc.InputMappedClassifier
- 14:57:09 - misc.InputMappedClassifier
- 14:59:04 - misc.InputMappedClassifier
- 14:59:39 - misc.InputMappedClassifier
- 14:59:54 - misc.InputMappedClassifier
- 15:00:13 - misc.InputMappedClassifier
- 15:00:35 - misc.InputMappedClassifier
- 15:02:49 - misc.InputMappedClassifier
- 15:03:23 - misc.InputMappedClassifier
- 15:03:49 - misc.InputMappedClassifier

**Classifier output**

Time taken to test model on supplied test set: 0.15 seconds

== Summary ==

	Correctly Classified Instances	183	85.514 %
Incorrectly Classified Instances	31	14.486 %	
Kappa statistic	0.8112		
Mean absolute error	0.1416		
Root mean squared error	0.2167		
Relative absolute error	45.842 %		
Root relative squared error	55.1547 %		
Total Number of Instances	214		

== Detailed Accuracy By Class ==

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.778	0.028	0.848	0.778	0.812	0.777	0.981	0.937	\$100,000 - \$149,999	
0.903	0.085	0.844	0.903	0.872	0.806	0.980	0.961	\$50,000 - \$99,999	
0.935	0.049	0.763	0.935	0.841	0.816	0.993	0.960	\$0 - \$24,999	
0.894	0.030	0.894	0.894	0.894	0.864	0.992	0.976	\$25,000 - \$49,999	
0.679	0.000	1.000	0.679	0.809	0.805	0.996	0.978	\$150,000+	
Weighted Avg.	0.855	0.047	0.864	0.855	0.854	0.815	0.987	0.962	

== Confusion Matrix ==

a b c d e	<-- classified as
28 2 2 4 0	a = \$100,000 - \$149,999
2 65 4 1 0	b = \$50,000 - \$99,999
0 2 29 0 0	c = \$0 - \$24,999
1 2 2 42 0	d = \$25,000 - \$49,999
2 6 1 0 19	e = \$150,000+

#### d) OneR Classification Model

**Classifier**

Choose **OneR -B 6**

**Test options**

- Use training set
- Supplied test set [Set...](#)
- Cross-validation Folds 10
- Percentage split % 66

[More options...](#)

(Nom) Household Income

[Start](#) [Stop](#)

**Result list (right-click for options)**

- 14:50:20 - misc.InputMappedClassifier
- 14:56:30 - misc.InputMappedClassifier
- 14:56:47 - misc.InputMappedClassifier
- 14:56:54 - misc.InputMappedClassifier
- 14:57:02 - misc.InputMappedClassifier
- 14:57:09 - misc.InputMappedClassifier
- 14:59:04 - misc.InputMappedClassifier
- 14:59:39 - misc.InputMappedClassifier
- 14:59:54 - misc.InputMappedClassifier
- 15:00:13 - misc.InputMappedClassifier
- 15:00:35 - misc.InputMappedClassifier
- 15:02:49 - misc.InputMappedClassifier
- 15:03:23 - misc.InputMappedClassifier

**Classifier output**

```
Time taken to test model on supplied test set: 0.17 seconds
```

**Summary**

	Correctly Classified Instances	71	33.1776 %
Incorrectly Classified Instances	143	66.8224 %	
Kappa statistic	0.0101		
Mean absolute error	0.2673		
Root mean squared error	0.517		
Relative absolute error	86.5171 %		
Root relative squared error	131.559 %		
Total Number of Instances	214		

**Detailed Accuracy By Class**

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	?	0.002	0.500	0.168	\$100,000
0.903	0.901	0.337	0.903	0.491	0.002	0.501	0.337	0.510	\$149,999
0.000	0.000	?	0.000	?	?	0.500	0.145	0.000	\$50,000
0.128	0.090	0.286	0.128	0.176	0.053	0.519	0.228	0.212	\$24,999
0.000	0.000	?	0.000	?	?	0.500	0.131	0.000	\$25,000
Weighted Avg.	0.332	0.323	?	0.332	?	?	0.504	0.230	

**Confusion Matrix**

a	b	c	d	e	classified as
0	34	0	2	0	a = \$100,000 - \$149,999
0	65	0	7	0	b = \$50,000 - \$99,999
0	26	0	5	0	c = \$0 - \$24,999
0	41	0	6	0	d = \$25,000 - \$49,999
0	27	0	1	0	e = \$150,000+

#### e) Simple Logistic Classification Model

**Classifier**

Choose **SimpleLogistic -I 0 -M 500 -H 50 -W 0.0**

**Test options**

- Use training set
- Supplied test set [Set...](#)
- Cross-validation Folds 10
- Percentage split % 66

[More options...](#)

(Nom) Household Income

[Start](#) [Stop](#)

**Result list (right-click for options)**

- 14:56:30 - misc.InputMappedClassifier
- 14:56:47 - misc.InputMappedClassifier
- 14:56:54 - misc.InputMappedClassifier
- 14:57:02 - misc.InputMappedClassifier
- 14:57:09 - misc.InputMappedClassifier
- 14:59:04 - misc.InputMappedClassifier
- 14:59:39 - misc.InputMappedClassifier
- 14:59:54 - misc.InputMappedClassifier
- 15:00:13 - misc.InputMappedClassifier
- 15:00:35 - misc.InputMappedClassifier
- 15:02:49 - misc.InputMappedClassifier
- 15:03:23 - misc.InputMappedClassifier
- 15:03:49 - misc.InputMappedClassifier

**Classifier output**

```
Time taken to test model on supplied test set: 0.16 seconds
```

**Summary**

	Correctly Classified Instances	76	35.514 %
Incorrectly Classified Instances	138	64.486 %	
Kappa statistic	0.0579		
Mean absolute error	0.3044		
Root mean squared error	0.3899		
Relative absolute error	98.5398 %		
Root relative squared error	99.2278 %		
Total Number of Instances	214		

**Detailed Accuracy By Class**

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	?	0.038	0.549	0.189	\$100,000
0.847	0.817	0.345	0.847	0.490	0.038	0.515	0.354	0.100	\$149,999
0.032	0.000	1.000	0.032	0.063	0.166	0.549	0.195	0.000	\$50,000
0.298	0.126	0.400	0.298	0.341	0.193	0.644	0.366	0.212	\$24,999
0.000	0.005	0.000	0.000	0.000	-0.027	0.661	0.270	0.000	\$25,000
Weighted Avg.	0.355	0.303	?	0.355	?	?	0.573	0.295	

**Confusion Matrix**

a	b	c	d	e	classified as
0	32	0	4	0	a = \$100,000 - \$149,999
0	61	0	11	0	b = \$50,000 - \$99,999
0	26	1	3	1	c = \$0 - \$24,999
0	33	0	14	0	d = \$25,000 - \$49,999
0	25	0	3	0	e = \$150,000+

## InfoGainAttributeEval Set:

### a) Naïve Bayes Classification Model

**Classifier**

Choose **NaiveBayes**

Test options		Classifier output																																																																																																					
<input type="radio"/> Use training set		<b>Evaluation on test set</b>																																																																																																					
<input checked="" type="radio"/> Supplied test set	<b>Set...</b>	Time taken to test model on supplied test set: 0.1 seconds																																																																																																					
<input type="radio"/> Cross-validation	Folds 10																																																																																																						
<input type="radio"/> Percentage split	% 66																																																																																																						
<b>(Nom) Household Income</b>																																																																																																							
<b>Start</b>		<b>Stop</b>																																																																																																					
<b>Result list (right-click for options)</b>																																																																																																							
<pre>14:48:41 - misc.InputMappedClassifier 14:49:33 - misc.InputMappedClassifier 14:49:52 - misc.InputMappedClassifier 14:50:10 - misc.InputMappedClassifier 14:50:20 - misc.InputMappedClassifier 14:56:30 - misc.InputMappedClassifier 14:56:47 - misc.InputMappedClassifier 14:56:54 - misc.InputMappedClassifier 14:57:02 - misc.InputMappedClassifier 14:57:09 - misc.InputMappedClassifier 14:59:04 - misc.InputMappedClassifier</pre>																																																																																																							
<b>Summary</b> <table border="1"> <thead> <tr> <th>Correctly Classified Instances</th> <th>70</th> <th>32.7103 %</th> </tr> </thead> <tbody> <tr> <td>Incorrectly Classified Instances</td> <td>144</td> <td>67.2897 %</td> </tr> <tr> <td>Kappa statistic</td> <td>0.0685</td> <td></td> </tr> <tr> <td>Mean absolute error</td> <td>0.2934</td> <td></td> </tr> <tr> <td>Root mean squared error</td> <td>0.3864</td> <td></td> </tr> <tr> <td>Relative absolute error</td> <td>94.9584 %</td> <td></td> </tr> <tr> <td>Root relative squared error</td> <td>98.3183 %</td> <td></td> </tr> <tr> <td>Total Number of Instances</td> <td>214</td> <td></td> </tr> </tbody> </table> <b>Detailed Accuracy By Class</b> <table border="1"> <thead> <tr> <th></th> <th>TP Rate</th> <th>FP Rate</th> <th>Precision</th> <th>Recall</th> <th>F-Measure</th> <th>MCC</th> <th>ROC Area</th> <th>PRC Area</th> <th>Class</th> </tr> </thead> <tbody> <tr> <td>0.056</td> <td>0.034</td> <td>0.250</td> <td>0.056</td> <td>0.091</td> <td>0.043</td> <td>0.615</td> <td>0.286</td> <td>\$100</td> </tr> <tr> <td>0.625</td> <td>0.613</td> <td>0.341</td> <td>0.625</td> <td>0.441</td> <td>0.012</td> <td>0.572</td> <td>0.420</td> <td>\$50,-</td> </tr> <tr> <td>0.065</td> <td>0.049</td> <td>0.182</td> <td>0.065</td> <td>0.095</td> <td>0.024</td> <td>0.662</td> <td>0.220</td> <td>\$0,-</td> </tr> <tr> <td>0.362</td> <td>0.156</td> <td>0.395</td> <td>0.362</td> <td>0.378</td> <td>0.213</td> <td>0.701</td> <td>0.431</td> <td>\$25,-</td> </tr> <tr> <td>0.143</td> <td>0.086</td> <td>0.200</td> <td>0.143</td> <td>0.167</td> <td>0.066</td> <td>0.676</td> <td>0.268</td> <td>\$150,-</td> </tr> <tr> <td>Weighted Avg.</td> <td>0.327</td> <td>0.264</td> <td>0.296</td> <td>0.327</td> <td>0.282</td> <td>0.070</td> <td>0.634</td> <td>0.351</td> </tr> </tbody> </table> <b>Confusion Matrix</b> <table border="1"> <thead> <tr> <th>a b c d e</th> <th>&lt;-- classified as</th> </tr> </thead> <tbody> <tr> <td>2 25 2 4 3</td> <td>a = \$100,000 - \$149,999</td> </tr> <tr> <td>4 45 4 13 6</td> <td>b = \$50,000 - \$99,999</td> </tr> <tr> <td>0 21 2 5 3</td> <td>c = \$0 - \$24,999</td> </tr> <tr> <td>2 22 2 17 4</td> <td>d = \$25,000 - \$49,999</td> </tr> <tr> <td>0 19 1 4 4</td> <td>e = \$150,000+</td> </tr> </tbody> </table>				Correctly Classified Instances	70	32.7103 %	Incorrectly Classified Instances	144	67.2897 %	Kappa statistic	0.0685		Mean absolute error	0.2934		Root mean squared error	0.3864		Relative absolute error	94.9584 %		Root relative squared error	98.3183 %		Total Number of Instances	214			TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	0.056	0.034	0.250	0.056	0.091	0.043	0.615	0.286	\$100	0.625	0.613	0.341	0.625	0.441	0.012	0.572	0.420	\$50,-	0.065	0.049	0.182	0.065	0.095	0.024	0.662	0.220	\$0,-	0.362	0.156	0.395	0.362	0.378	0.213	0.701	0.431	\$25,-	0.143	0.086	0.200	0.143	0.167	0.066	0.676	0.268	\$150,-	Weighted Avg.	0.327	0.264	0.296	0.327	0.282	0.070	0.634	0.351	a b c d e	<-- classified as	2 25 2 4 3	a = \$100,000 - \$149,999	4 45 4 13 6	b = \$50,000 - \$99,999	0 21 2 5 3	c = \$0 - \$24,999	2 22 2 17 4	d = \$25,000 - \$49,999	0 19 1 4 4	e = \$150,000+
Correctly Classified Instances	70	32.7103 %																																																																																																					
Incorrectly Classified Instances	144	67.2897 %																																																																																																					
Kappa statistic	0.0685																																																																																																						
Mean absolute error	0.2934																																																																																																						
Root mean squared error	0.3864																																																																																																						
Relative absolute error	94.9584 %																																																																																																						
Root relative squared error	98.3183 %																																																																																																						
Total Number of Instances	214																																																																																																						
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class																																																																																														
0.056	0.034	0.250	0.056	0.091	0.043	0.615	0.286	\$100																																																																																															
0.625	0.613	0.341	0.625	0.441	0.012	0.572	0.420	\$50,-																																																																																															
0.065	0.049	0.182	0.065	0.095	0.024	0.662	0.220	\$0,-																																																																																															
0.362	0.156	0.395	0.362	0.378	0.213	0.701	0.431	\$25,-																																																																																															
0.143	0.086	0.200	0.143	0.167	0.066	0.676	0.268	\$150,-																																																																																															
Weighted Avg.	0.327	0.264	0.296	0.327	0.282	0.070	0.634	0.351																																																																																															
a b c d e	<-- classified as																																																																																																						
2 25 2 4 3	a = \$100,000 - \$149,999																																																																																																						
4 45 4 13 6	b = \$50,000 - \$99,999																																																																																																						
0 21 2 5 3	c = \$0 - \$24,999																																																																																																						
2 22 2 17 4	d = \$25,000 - \$49,999																																																																																																						
0 19 1 4 4	e = \$150,000+																																																																																																						

### b) J48 Classification Model

**Classifier**

Choose **J48 -C 0.25 -M 2**

Test options		Classifier output																																																																																																					
<input type="radio"/> Use training set		<b>Evaluation on test set</b>																																																																																																					
<input checked="" type="radio"/> Supplied test set	<b>Set...</b>	Time taken to test model on supplied test set: 0.14 seconds																																																																																																					
<input type="radio"/> Cross-validation	Folds 10																																																																																																						
<input type="radio"/> Percentage split	% 66																																																																																																						
<b>(Nom) Household Income</b>																																																																																																							
<b>Start</b>		<b>Stop</b>																																																																																																					
<b>Result list (right-click for options)</b>																																																																																																							
<pre>14:48:41 - misc.InputMappedClassifier 14:49:33 - misc.InputMappedClassifier 14:49:52 - misc.InputMappedClassifier 14:50:10 - misc.InputMappedClassifier 14:50:20 - misc.InputMappedClassifier 14:56:30 - misc.InputMappedClassifier 14:56:47 - misc.InputMappedClassifier 14:56:54 - misc.InputMappedClassifier 14:57:02 - misc.InputMappedClassifier 14:57:09 - misc.InputMappedClassifier 14:59:04 - misc.InputMappedClassifier 14:59:39 - misc.InputMappedClassifier</pre>																																																																																																							
<b>Summary</b> <table border="1"> <thead> <tr> <th>Correctly Classified Instances</th> <th>125</th> <th>58.4112 %</th> </tr> </thead> <tbody> <tr> <td>Incorrectly Classified Instances</td> <td>89</td> <td>41.5888 %</td> </tr> <tr> <td>Kappa statistic</td> <td>0.4314</td> <td></td> </tr> <tr> <td>Mean absolute error</td> <td>0.2211</td> <td></td> </tr> <tr> <td>Root mean squared error</td> <td>0.3332</td> <td></td> </tr> <tr> <td>Relative absolute error</td> <td>71.5752 %</td> <td></td> </tr> <tr> <td>Root relative squared error</td> <td>84.7939 %</td> <td></td> </tr> <tr> <td>Total Number of Instances</td> <td>214</td> <td></td> </tr> </tbody> </table> <b>Detailed Accuracy By Class</b> <table border="1"> <thead> <tr> <th></th> <th>TP Rate</th> <th>FP Rate</th> <th>Precision</th> <th>Recall</th> <th>F-Measure</th> <th>MCC</th> <th>ROC Area</th> <th>PRC Area</th> <th>Class</th> </tr> </thead> <tbody> <tr> <td>0.472</td> <td>0.084</td> <td>0.531</td> <td>0.472</td> <td>0.500</td> <td>0.407</td> <td>0.840</td> <td>0.568</td> <td>\$100</td> </tr> <tr> <td>0.889</td> <td>0.430</td> <td>0.512</td> <td>0.889</td> <td>0.650</td> <td>0.440</td> <td>0.779</td> <td>0.593</td> <td>\$50,-</td> </tr> <tr> <td>0.323</td> <td>0.016</td> <td>0.769</td> <td>0.323</td> <td>0.455</td> <td>0.451</td> <td>0.848</td> <td>0.545</td> <td>\$0,-</td> </tr> <tr> <td>0.489</td> <td>0.024</td> <td>0.852</td> <td>0.489</td> <td>0.622</td> <td>0.580</td> <td>0.864</td> <td>0.694</td> <td>\$25,-</td> </tr> <tr> <td>0.393</td> <td>0.032</td> <td>0.647</td> <td>0.393</td> <td>0.489</td> <td>0.450</td> <td>0.838</td> <td>0.532</td> <td>\$150,-</td> </tr> <tr> <td>Weighted Avg.</td> <td>0.584</td> <td>0.171</td> <td>0.645</td> <td>0.584</td> <td>0.569</td> <td>0.468</td> <td>0.826</td> <td>0.596</td> </tr> </tbody> </table> <b>Confusion Matrix</b> <table border="1"> <thead> <tr> <th>a b c d e</th> <th>&lt;-- classified as</th> </tr> </thead> <tbody> <tr> <td>17 16 0 1 2</td> <td>a = \$100,000 - \$149,999</td> </tr> <tr> <td>5 64 1 1 1</td> <td>b = \$50,000 - \$99,999</td> </tr> <tr> <td>2 16 10 2 1</td> <td>c = \$0 - \$24,999</td> </tr> <tr> <td>2 19 1 23 2</td> <td>d = \$25,000 - \$49,999</td> </tr> <tr> <td>6 10 1 0 11</td> <td>e = \$150,000+</td> </tr> </tbody> </table>				Correctly Classified Instances	125	58.4112 %	Incorrectly Classified Instances	89	41.5888 %	Kappa statistic	0.4314		Mean absolute error	0.2211		Root mean squared error	0.3332		Relative absolute error	71.5752 %		Root relative squared error	84.7939 %		Total Number of Instances	214			TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	0.472	0.084	0.531	0.472	0.500	0.407	0.840	0.568	\$100	0.889	0.430	0.512	0.889	0.650	0.440	0.779	0.593	\$50,-	0.323	0.016	0.769	0.323	0.455	0.451	0.848	0.545	\$0,-	0.489	0.024	0.852	0.489	0.622	0.580	0.864	0.694	\$25,-	0.393	0.032	0.647	0.393	0.489	0.450	0.838	0.532	\$150,-	Weighted Avg.	0.584	0.171	0.645	0.584	0.569	0.468	0.826	0.596	a b c d e	<-- classified as	17 16 0 1 2	a = \$100,000 - \$149,999	5 64 1 1 1	b = \$50,000 - \$99,999	2 16 10 2 1	c = \$0 - \$24,999	2 19 1 23 2	d = \$25,000 - \$49,999	6 10 1 0 11	e = \$150,000+
Correctly Classified Instances	125	58.4112 %																																																																																																					
Incorrectly Classified Instances	89	41.5888 %																																																																																																					
Kappa statistic	0.4314																																																																																																						
Mean absolute error	0.2211																																																																																																						
Root mean squared error	0.3332																																																																																																						
Relative absolute error	71.5752 %																																																																																																						
Root relative squared error	84.7939 %																																																																																																						
Total Number of Instances	214																																																																																																						
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class																																																																																														
0.472	0.084	0.531	0.472	0.500	0.407	0.840	0.568	\$100																																																																																															
0.889	0.430	0.512	0.889	0.650	0.440	0.779	0.593	\$50,-																																																																																															
0.323	0.016	0.769	0.323	0.455	0.451	0.848	0.545	\$0,-																																																																																															
0.489	0.024	0.852	0.489	0.622	0.580	0.864	0.694	\$25,-																																																																																															
0.393	0.032	0.647	0.393	0.489	0.450	0.838	0.532	\$150,-																																																																																															
Weighted Avg.	0.584	0.171	0.645	0.584	0.569	0.468	0.826	0.596																																																																																															
a b c d e	<-- classified as																																																																																																						
17 16 0 1 2	a = \$100,000 - \$149,999																																																																																																						
5 64 1 1 1	b = \$50,000 - \$99,999																																																																																																						
2 16 10 2 1	c = \$0 - \$24,999																																																																																																						
2 19 1 23 2	d = \$25,000 - \$49,999																																																																																																						
6 10 1 0 11	e = \$150,000+																																																																																																						

### c) Random Forest Classification Model

**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 10
- Percentage split % 66

[More options...](#)

(Nom) Household Income [▼](#)

[Start](#) [Stop](#)

**Result list (right-click for options)**

```
14:48:41 - misc.InputMappedClassifier
14:49:33 - misc.InputMappedClassifier
14:49:52 - misc.InputMappedClassifier
14:50:10 - misc.InputMappedClassifier
14:50:20 - misc.InputMappedClassifier
14:56:30 - misc.InputMappedClassifier
14:56:47 - misc.InputMappedClassifier
14:56:54 - misc.InputMappedClassifier
14:57:02 - misc.InputMappedClassifier
14:57:09 - misc.InputMappedClassifier
14:59:04 - misc.InputMappedClassifier
14:59:39 - misc.InputMappedClassifier
14:59:54 - misc.InputMappedClassifier
```

**Classifier output**

Time taken to test model on supplied test set: 0.15 seconds

==== Summary ===

	Correctly Classified Instances	188	87.8505 %
Incorrectly Classified Instances	26	12.1495 %	
Kappa statistic	0.841		
Mean absolute error	0.1364		
Root mean squared error	0.2054		
Relative absolute error	44.1587 %		
Root relative squared error	52.2684 %		
Total Number of Instances	214		

==== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class:
0.778	0.017	0.903	0.778	0.836	0.809	0.975	0.929	\$100+	
0.944	0.085	0.850	0.944	0.895	0.840	0.990	0.982	\$50,-	
0.806	0.027	0.833	0.806	0.820	0.790	0.993	0.961	\$0 -	
0.936	0.036	0.880	0.936	0.907	0.881	0.996	0.985	\$25,-	
0.821	0.000	1.000	0.821	0.902	0.894	0.996	0.980	\$150+	
Weighted Avg.	0.879	0.043	0.883	0.879	0.878	0.843	0.990	0.971	

==== Confusion Matrix ===

a b c d e	<-- classified as
28 3 3 2 0	a = \$100,000 - \$149,999
1 68 1 2 0	b = \$50,000 - \$99,999
0 6 25 0 0	c = \$0 - \$24,999
0 2 1 44 0	d = \$25,000 - \$49,999
2 1 0 2 23	e = \$150,000+

### d) OneR Classification Model

**Classifier**

Choose **OneR -B 6**

**Test options**

- Use training set
- Supplied test set [Set...](#)
- Cross-validation Folds 10
- Percentage split % 66

[More options...](#)

(Nom) Household Income [▼](#)

[Start](#) [Stop](#)

**Result list (right-click for options)**

```
14:48:41 - misc.InputMappedClassifier
14:49:33 - misc.InputMappedClassifier
14:49:52 - misc.InputMappedClassifier
14:50:10 - misc.InputMappedClassifier
14:50:20 - misc.InputMappedClassifier
14:56:30 - misc.InputMappedClassifier
14:56:47 - misc.InputMappedClassifier
14:56:54 - misc.InputMappedClassifier
14:57:02 - misc.InputMappedClassifier
14:57:09 - misc.InputMappedClassifier
14:59:04 - misc.InputMappedClassifier
14:59:39 - misc.InputMappedClassifier
14:59:54 - misc.InputMappedClassifier
```

**Classifier output**

Time taken to test model on supplied test set: 0.14 seconds

==== Summary ===

	Correctly Classified Instances	73	34.1121 %
Incorrectly Classified Instances	141	65.8879 %	
Kappa statistic	0.0095		
Mean absolute error	0.2636		
Root mean squared error	0.5134		
Relative absolute error	85.3071 %		
Root relative squared error	130.6358 %		
Total Number of Instances	214		

==== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class:
0.000	0.000	?	0.000	?	?	0.500	0.168	\$100+	
1.000	0.979	0.341	1.000	0.509	0.085	0.511	0.341	\$50,-	
0.000	0.000	?	0.000	?	?	0.500	0.145	\$0 -	
0.021	0.012	0.333	0.021	0.040	0.033	0.505	0.222	\$25,-	
0.000	0.000	?	0.000	?	?	0.500	0.131	\$150+	
Weighted Avg.	0.341	0.332	?	0.341	?	0.505	0.230		

==== Confusion Matrix ===

a b c d e	<-- classified as
0 36 0 0 0	a = \$100,000 - \$149,999
0 72 0 0 0	b = \$50,000 - \$99,999
0 30 0 1 0	c = \$0 - \$24,999
0 46 0 1 0	d = \$25,000 - \$49,999
0 27 0 1 0	e = \$150,000+

### e) Simple Logistic Classification Model

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

**Classifier**

Choose SimpleLogistic -I 0 -M 500 -H 50 -W 0.0

**Test options**

- Use training set
- Supplied test set Set...
- Cross-validation Folds 10
- Percentage split % 66
- [More options...](#)

(Nom) Household Income

Start Stop

**Result list (right-click for options)**

```
14:48:41 - misc.InputMappedClassifier
14:49:33 - misc.InputMappedClassifier
14:49:52 - misc.InputMappedClassifier
14:50:10 - misc.InputMappedClassifier
14:50:20 - misc.InputMappedClassifier
14:56:30 - misc.InputMappedClassifier
14:56:47 - misc.InputMappedClassifier
14:56:54 - misc.InputMappedClassifier
14:57:02 - misc.InputMappedClassifier
14:57:09 - misc.InputMappedClassifier
14:59:04 - misc.InputMappedClassifier
14:59:39 - misc.InputMappedClassifier
14:59:54 - misc.InputMappedClassifier
```

**Classifier output**

Time taken to test model on supplied test set: 0.15 seconds

== Summary ==

	Correctly Classified Instances	73	34.1121 %
Incorrectly Classified Instances	141	65.8879 %	
Kappa statistic	0.0109		
Mean absolute error	0.3069		
Root mean squared error	0.3914		
Relative absolute error	99.3388 %		
Root relative squared error	99.6034 %		
Total Number of Instances	214		

== Detailed Accuracy By Class ==

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	?	0.536	0.184	\$100,000 - \$149,999	
1.000	0.972	0.343	1.000	0.511	0.098	0.509	0.338	\$50,000 - \$99,999	
0.000	0.000	?	0.000	?	?	0.536	0.182	\$0 - \$24,999	
0.021	0.012	0.333	0.021	0.040	0.033	0.596	0.293	\$25,000 - \$49,999	
0.000	0.005	0.000	0.000	0.000	-0.027	0.610	0.182	\$150,000+	
Weighted Avg.	0.341	0.330	?	0.341	?	0.550	0.259		

== Confusion Matrix ==

a b c d e	<-- classified as
0 36 0 0 0	a = \$100,000 - \$149,999
0 72 0 0 0	b = \$50,000 - \$99,999
0 29 0 1 1	c = \$0 - \$24,999
0 46 0 1 0	d = \$25,000 - \$49,999
0 27 0 1 0	e = \$150,000+

### Our Own Attribute Set:

#### a) Naïve Bayes Classification Model

**Classifier**

Choose SimpleLogistic -I 0 -M 500 -H 50 -W 0.0

**Test options**

- Use training set
- Supplied test set Set...
- Cross-validation Folds 10
- Percentage split % 66
- [More options...](#)

(Nom) Household Income

Start Stop

**Result list (right-click for options)**

```
14:59:04 - misc.InputMappedClassifier
14:59:39 - misc.InputMappedClassifier
14:59:54 - misc.InputMappedClassifier
15:00:13 - misc.InputMappedClassifier
15:00:35 - misc.InputMappedClassifier
15:02:49 - misc.InputMappedClassifier
15:03:23 - misc.InputMappedClassifier
15:03:49 - misc.InputMappedClassifier
15:04:05 - misc.InputMappedClassifier
15:04:22 - misc.InputMappedClassifier
15:07:29 - misc.InputMappedClassifier
15:08:14 - misc.InputMappedClassifier
15:08:34 - misc.InputMappedClassifier
```

**Classifier output**

Time taken to test model on supplied test set: 0.14 seconds

== Summary ==

	Correctly Classified Instances	76	35.514 %
Incorrectly Classified Instances	138	64.486 %	
Kappa statistic	0.119		
Mean absolute error	0.2844		
Root mean squared error	0.3791		
Relative absolute error	92.0559 %		
Root relative squared error	96.4767 %		
Total Number of Instances	214		

== Detailed Accuracy By Class ==

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.083	0.034	0.333	0.083	0.133	0.092	0.708	0.329	\$100,000 - \$149,999	
0.556	0.556	0.336	0.556	0.419	-0.001	0.575	0.443	\$50,000 - \$99,999	
0.290	0.066	0.429	0.290	0.346	0.266	0.764	0.401	\$0 - \$24,999	
0.404	0.174	0.396	0.404	0.400	0.229	0.723	0.463	\$25,000 - \$49,999	
0.179	0.065	0.294	0.179	0.222	0.142	0.722	0.292	\$150,000+	
Weighted Avg.	0.355	0.249	0.357	0.355	0.330	0.123	0.677	0.403	

== Confusion Matrix ==

a b c d e	<-- classified as
3 26 3 1 3	a = \$100,000 - \$149,999
3 40 4 17 8	b = \$50,000 - \$99,999
0 15 9 6 1	c = \$0 - \$24,999
1 24 3 19 0	d = \$25,000 - \$49,999
2 14 2 5 5	e = \$150,000+

## b) J48 Classification Model

Preprocess Classify Cluster Associate Select attributes Visualize

**Classifier**

Choose SimpleLogistic -I 0 -M 500 -H 50 -W 0.0

**Test options**

- Use training set
- Supplied test set Set...
- Cross-validation Folds 10
- Percentage split % 66
- More options...

(Nom) Household Income

Start Stop

**Result list (right-click for options)**

- 14:59:04 - misc.InputMappedClassifier
- 14:59:39 - misc.InputMappedClassifier
- 14:59:54 - misc.InputMappedClassifier
- 15:00:13 - misc.InputMappedClassifier
- 15:00:35 - misc.InputMappedClassifier
- 15:02:49 - misc.InputMappedClassifier
- 15:03:23 - misc.InputMappedClassifier
- 15:03:49 - misc.InputMappedClassifier
- 15:04:05 - misc.InputMappedClassifier
- 15:04:22 - misc.InputMappedClassifier
- 15:07:29 - misc.InputMappedClassifier
- 15:08:14 - misc.InputMappedClassifier
- 15:08:34 - misc.InputMappedClassifier

**Classifier output**

Time taken to test model on supplied test set: 0.15 seconds

==== Summary ===

	Correctly Classified Instances	136	63.5514 %
Incorrectly Classified Instances	78	36.4486 %	
Kappa statistic	0.5137		
Mean absolute error	0.1922		
Root mean squared error	0.3072		
Relative absolute error	62.2059 %		
Root relative squared error	78.1823 %		
Total Number of Instances	214		

==== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.528	0.051	0.679	0.528	0.594	0.529	0.930	0.728	\$100k+	
0.819	0.289	0.590	0.819	0.686	0.503	0.850	0.721	\$50k+	
0.645	0.066	0.625	0.645	0.635	0.572	0.932	0.706	\$0	
0.596	0.072	0.700	0.596	0.644	0.556	0.891	0.721	\$25k+	
0.357	0.022	0.714	0.357	0.476	0.458	0.903	0.605	\$150k+	
Weighted Avg.	0.636	0.134	0.650	0.636	0.626	0.523	0.891	0.705	

==== Confusion Matrix ===

a b c d e	<-- classified as
19 12 1 0 4	a = \$100,000 - \$149,999
3 59 4 6 0	b = \$50,000 - \$99,999
1 8 20 2 0	c = \$0 - \$24,999
2 12 5 28 0	d = \$25,000 - \$49,999
3 9 2 4 10	e = \$150,000+

## c) Random Forest Classification Model

Preprocess Classify Cluster Associate Select attributes Visualize

**Classifier**

Choose SimpleLogistic -I 0 -M 500 -H 50 -W 0.0

**Test options**

- Use training set
- Supplied test set Set...
- Cross-validation Folds 10
- Percentage split % 66
- More options...

(Nom) Household Income

Start Stop

**Result list (right-click for options)**

- 14:59:04 - misc.InputMappedClassifier
- 14:59:39 - misc.InputMappedClassifier
- 14:59:54 - misc.InputMappedClassifier
- 15:00:13 - misc.InputMappedClassifier
- 15:00:35 - misc.InputMappedClassifier
- 15:02:49 - misc.InputMappedClassifier
- 15:03:23 - misc.InputMappedClassifier
- 15:03:49 - misc.InputMappedClassifier
- 15:04:05 - misc.InputMappedClassifier
- 15:04:22 - misc.InputMappedClassifier
- 15:07:29 - misc.InputMappedClassifier
- 15:08:14 - misc.InputMappedClassifier
- 15:08:34 - misc.InputMappedClassifier

**Classifier output**

Time taken to test model on supplied test set: 0.19 seconds

==== Summary ===

	Correctly Classified Instances	199	92.9907 %
Incorrectly Classified Instances	15	7.0093 %	
Kappa statistic	0.9091		
Mean absolute error	0.1176		
Root mean squared error	0.1758		
Relative absolute error	38.0732 %		
Root relative squared error	44.7277 %		
Total Number of Instances	214		

==== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.917	0.006	0.971	0.917	0.943	0.932	0.999	0.995	\$100k+	
0.944	0.035	0.932	0.944	0.938	0.906	0.997	0.995	\$50k+	
0.903	0.022	0.875	0.903	0.889	0.870	0.996	0.974	\$0	
0.957	0.012	0.957	0.957	0.957	0.945	0.999	0.997	\$25k+	
0.893	0.016	0.893	0.893	0.893	0.877	0.997	0.982	\$150k+	
Weighted Avg.	0.930	0.021	0.931	0.930	0.930	0.910	0.998	0.991	

==== Confusion Matrix ===

a b c d e	<-- classified as
33 1 1 0 1	a = \$100,000 - \$149,999
0 68 3 1 0	b = \$50,000 - \$99,999
0 2 28 0 1	c = \$0 - \$24,999
0 1 0 45 1	d = \$25,000 - \$49,999
1 1 0 1 25	e = \$150,000+

#### d) OneR Classification Model

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

Choose SimpleLogistic -I 0 -M 500 -H 50 -W 0.0

**Test options**

- Use training set
- Supplied test set [Set...](#)
- Cross-validation Folds 10
- Percentage split % 66 [More options...](#)

(Nom) Household Income

[Start](#) [Stop](#)

**Result list (right-click for options)**

```
14:59:04 - misc.InputMappedClassifier
14:59:39 - misc.InputMappedClassifier
14:59:54 - misc.InputMappedClassifier
15:00:13 - misc.InputMappedClassifier
15:00:35 - misc.InputMappedClassifier
15:02:49 - misc.InputMappedClassifier
15:03:23 - misc.InputMappedClassifier
15:03:49 - misc.InputMappedClassifier
15:04:05 - misc.InputMappedClassifier
15:04:22 - misc.InputMappedClassifier
15:07:29 - misc.InputMappedClassifier
15:08:14 - misc.InputMappedClassifier
15:08:34 - misc.InputMappedClassifier
```

**Classifier output**

Time taken to test model on supplied test set: 0.16 seconds

==== Summary ===

	Correctly Classified Instances	68	31.7757 %
Incorrectly Classified Instances	146	68.2243 %	
Kappa statistic	0.014		
Mean absolute error	0.2729		
Root mean squared error	0.5224		
Relative absolute error	88.3321 %		
Root relative squared error	132.9318 %		
Total Number of Instances	214		

==== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	?	0.500	0.168	\$100K+	
0.792	0.739	0.352	0.792	0.487	0.058	0.526	0.349	\$149K+	
0.000	0.000	?	0.000	?	?	0.500	0.145	\$50K+	
0.234	0.246	0.212	0.234	0.222	-0.011	0.494	0.218	\$24K+	
0.000	0.000	?	0.000	?	?	0.500	0.131	\$0K+	
Weighted Avg.	0.318	0.303	?	0.318	?	?	0.508	0.232	

==== Confusion Matrix ===

a	b	c	d	e	<-- classified as
0	31	0	5	0	a = \$100,000 - \$149,999
0	57	0	15	0	b = \$50,000 - \$99,999
0	15	0	16	0	c = \$0 - \$24,999
0	36	0	11	0	d = \$25,000 - \$49,999
0	23	0	5	0	e = \$150,000+

#### e) Simple Logistic Classification Model

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

Choose SimpleLogistic -I 0 -M 500 -H 50 -W 0.0

**Test options**

- Use training set
- Supplied test set [Set...](#)
- Cross-validation Folds 10
- Percentage split % 66 [More options...](#)

(Nom) Household Income

[Start](#) [Stop](#)

**Result list (right-click for options)**

```
14:59:04 - misc.InputMappedClassifier
14:59:39 - misc.InputMappedClassifier
14:59:54 - misc.InputMappedClassifier
15:00:13 - misc.InputMappedClassifier
15:00:35 - misc.InputMappedClassifier
15:02:49 - misc.InputMappedClassifier
15:03:23 - misc.InputMappedClassifier
15:03:49 - misc.InputMappedClassifier
15:04:05 - misc.InputMappedClassifier
15:04:22 - misc.InputMappedClassifier
15:07:29 - misc.InputMappedClassifier
15:08:14 - misc.InputMappedClassifier
15:08:34 - misc.InputMappedClassifier
```

**Classifier output**

Time taken to test model on supplied test set: 0.18 seconds

==== Summary ===

	Correctly Classified Instances	73	34.1121 %
Incorrectly Classified Instances	141	65.8879 %	
Kappa statistic	0.0721		
Mean absolute error	0.3041		
Root mean squared error	0.389		
Relative absolute error	98.4288 %		
Root relative squared error	98.9842 %		
Total Number of Instances	214		

==== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	?	0.557	0.205	\$100K+	
0.792	0.739	0.352	0.792	0.487	0.058	0.542	0.361	\$149K+	
0.516	0.197	0.308	0.516	0.386	0.262	0.710	0.249	\$50K+	
0.000	0.000	?	0.000	?	?	0.552	0.239	\$24K+	
0.000	0.000	?	0.000	?	?	0.622	0.189	\$0K+	
Weighted Avg.	0.341	0.277	?	0.341	?	?	0.581	0.269	

==== Confusion Matrix ===

a	b	c	d	e	<-- classified as
0	31	5	0	0	a = \$100,000 - \$149,999
0	57	15	0	0	b = \$50,000 - \$99,999
0	15	16	0	0	c = \$0 - \$24,999
0	36	11	0	0	d = \$25,000 - \$49,999
0	23	5	0	0	e = \$150,000+

## 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:

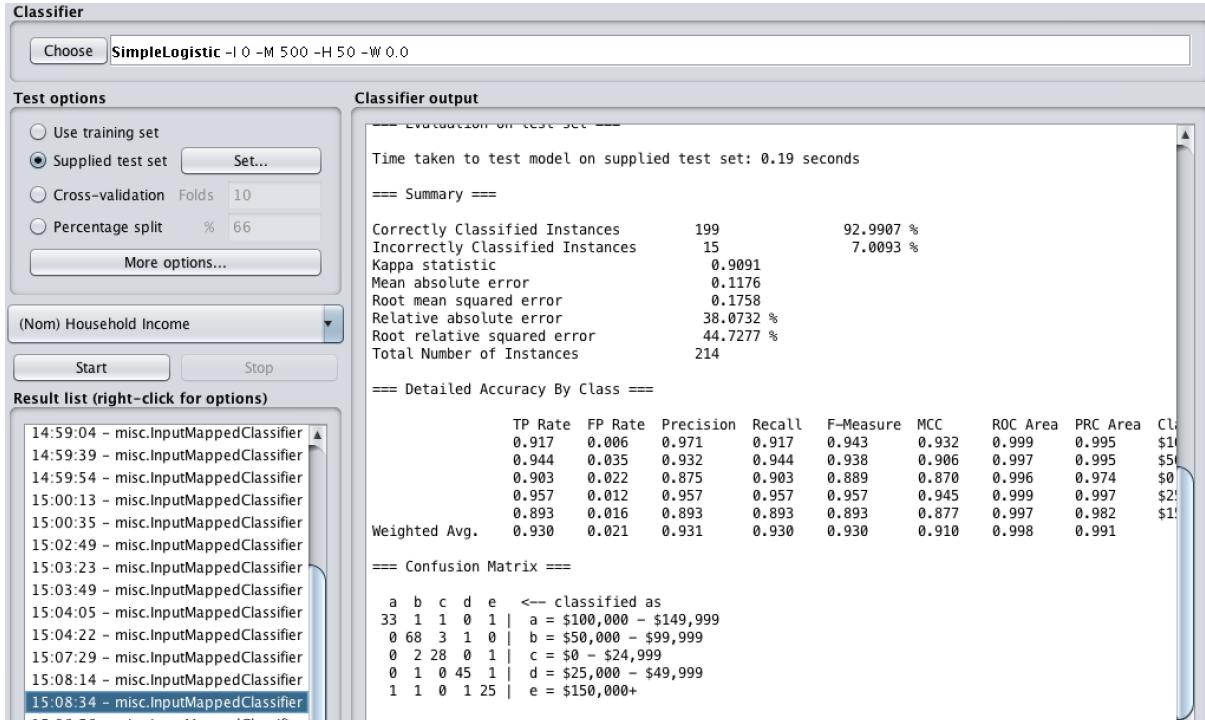
Attribute Set	TP Rate	FP Rate
CorrelationAttributeEval	0.874	0.045
OneRAttributeEval Set	0.911	0.026
GainRatioAttributeEval Set	0.855	0.047
InfoGainAttributeEval Set	0.879	0.043
Our Own Attribute Set	0.930	0.021

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.