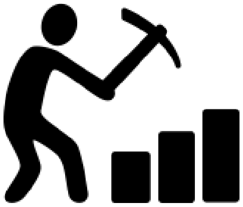
DATA MINING

Final Project



Fariborz Norouzi

METCS699 – Boston University

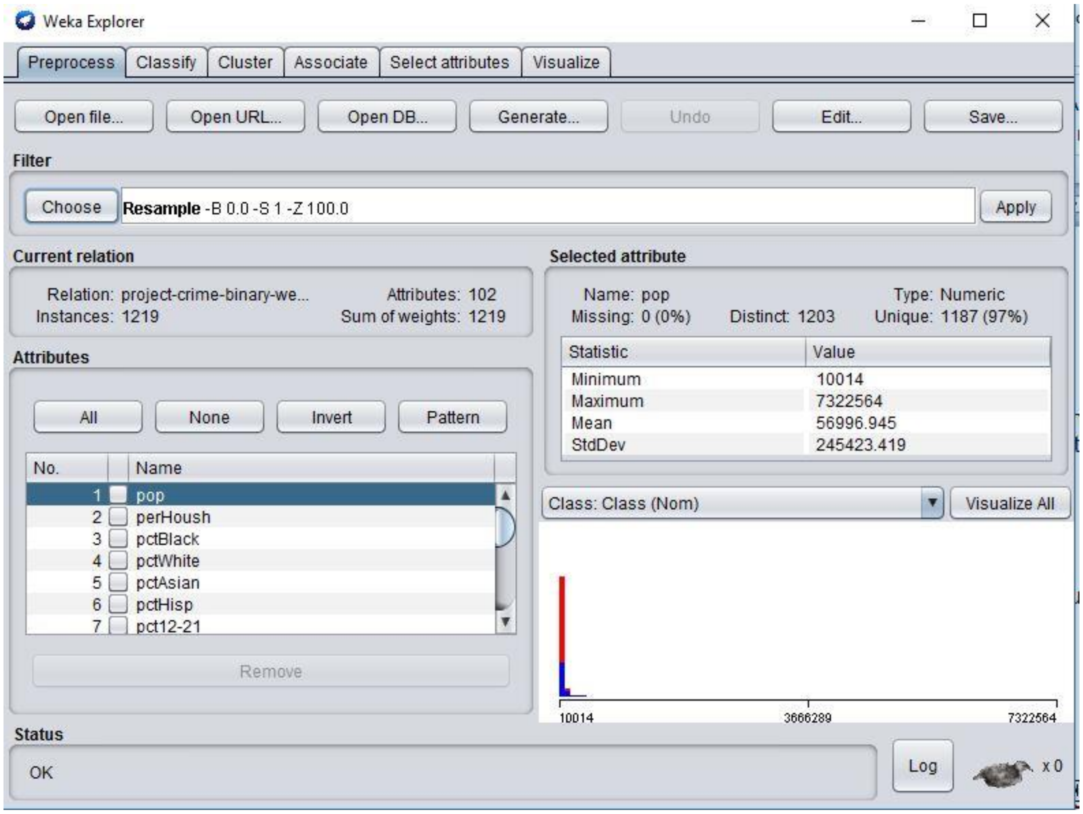
Summer 2017

Dividing a dataset into training and test dataset

For providing dataset into weka and dividing a dataset into training dataset with 1219 instances and test dataset with 600 instances, I did following steps:

1. Import original data and remove the first two attributes and save as modified data

2. Apply by resample filter and save as training set



Filter

Resample -B 0.0 -S 1 -Z 100.0

Current relation

Relation: project-crime-binary-we... Attributes: 102
Instances: 1219 Sum of weights: 1219

Selected attribute

Name: pop Type: Numeric
Missing: 0 (0%) Distinct: 1203 Unique: 1187 (97%)

Statistic	Value
Minimum	10014
Maximum	7322564
Mean	56996.945
StdDev	245423.419

Attributes

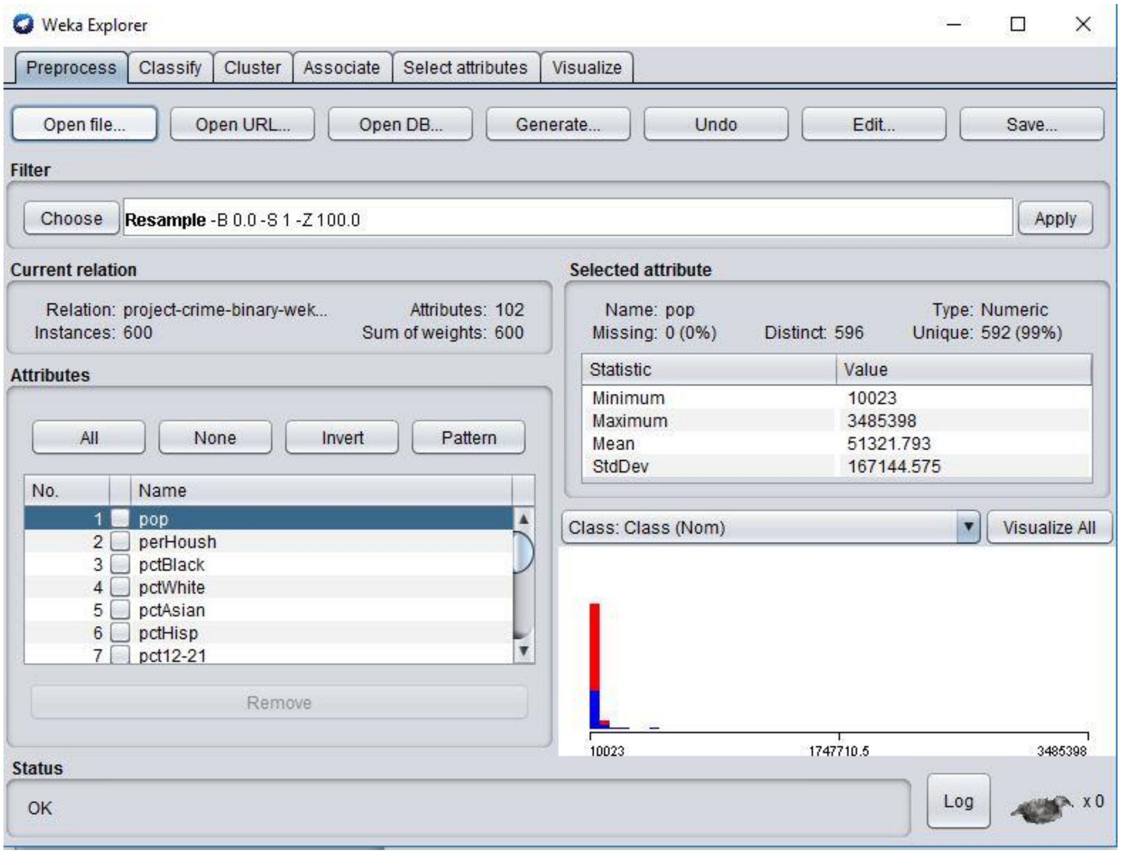
No.		Name
1	□	pop
2	□	perHoush
3	□	pctBlack
4	□	pctWhite
5	□	pctAsian
6	□	pctHisp
7	□	pct12-21

Class: Class (Nom)

Status

OK

3. Undo and apply filter again, but set the invertSelection to True and save as test set.



Filter: **Resample** -B 0.0 -S 1 -Z 100.0

Relation: project-crime-binary-wek... Attributes: 102
Instances: 600 Sum of weights: 600

Name: pop Type: Numeric
Missing: 0 (0%) Distinct: 596 Unique: 592 (99%)

Statistic	Value
Minimum	10023
Maximum	3485398
Mean	51321.793
StdDev	167144.575

No.		Name
1	□	pop
2	□	perHoush
3	□	pctBlack
4	□	pctWhite
5	□	pctAsian
6	□	pctHisp
7	□	pct12-21

Class: Class (Nom)

Status: OK

1- Attribute selection method: CfsSubsetEval

It evaluates the worth of a subset of attributes by process of selecting a subset of relevant features for use in model construction. Subsets of features that are highly correlated with the class while having low inter correlation are preferred.

The names of attributes that were selected by this method are:

Number of attributes: 11

pctBlack, pctWhite, pctMaleDivorc, pctFemDivorc, pctAllDivorc, pct2Par, pctKids-4w2Par

pct12-17w2Par, kidsBornNevrMarr, pctKidsBornNevrMarr, Class

- **Classifier Algorithm: J48**

Correctly Classified Instances 469 78.1667 %

	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	Class
	0.589	0.121	0.713	0.645	0.494	0.817	0
	0.879	0.411	0.808	0.842	0.494	0.817	1
Weighted Avg.	0.782	0.313	0.776	0.776	0.494	0.817	

Confusion Matrix: J48

```
  a    b   <-- classified as
119  83 |   a = 0
 48 350 |   b = 1
```

- **Classifier Algorithm: Multilayer Perceptron**

Correctly Classified Instances 480 80 %

	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	Class
	0.698	0.148	0.705	0.701	0.551	0.855	0
	0.852	0.302	0.848	0.850	0.551	0.855	1
Weighted Avg.	0.800	0.250	0.800	0.800	0.551	0.855	

Confusion Matrix: Multilayer Perceptron

```
  a    b   <-- classified as
141  61 |   a = 0
 59 339 |   b = 1
```

- **Classifier Algorithm: Random Forest**

Correctly Classified Instances 474 79 %

	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	Class
	0.644	0.136	0.707	0.674	0.521	0.842	0
	0.864	0.356	0.827	0.845	0.521	0.842	1

Weighted Avg.	0.790	0.282	0.786	0.787	0.521	0.842	

Confusion Matrix: Random Forest

```
   a    b   <-- classified as

 130  72 |   a = 0

  54 344 |   b = 1
```

- **Classifier Algorithm: Simple Logistic**

```
Correctly Classified Instances         481              80.1667 %
```

	TP Rate	**FP Rate**	**Precision**	**F-Measure**	**MCC**	**ROC Area**	**Class**
	0.609	0.101	0.755	0.674	0.540	0.852	0
	0.899	0.391	0.819	0.857	0.540	0.852	1
Weighted Avg.	0.802	0.293	0.797	0.796	0.540	0.852	

Confusion Matrix: Simple Logistic

```
   a    b   <-- classified as

 123  79 |   a = 0

  40 358 |   b = 1
```

Best model Performance in **CfsSubset** Attribute selection method is **Multilayer perceptron** because although simple logistic has the highest correctly classified instances which was 481 Correctly Classified Instances (80.167%) but Multilayer perceptron with 480 correct classified instances and better other measure performance such as higher TP class 0 rate, lower average and class one FP rate, higher average Precision, F-Measure, MCC and ROC Area rate.

	CLASS	**J48**	**Multilayer**	**Random Forest**	**Simple logistic**
TP	**0**	0.589	0.698	0.644	0.609
	1	0.879	0.852	0.864	0.899
Ave		0.782	0.800	0.790	0.802

FP	**0**	0.121	0.148	0.136	0.101
	1	0.411	0.302	0.356	0.391
Ave		0.313	0.250	0.282	0.293
Precision	**0**	0.713	0.705	0.707	0.755
	1	0.808	0.848	0.827	0.819
Ave		0.776	0.800	0.786	0.797
F-Measure	**0**	0.645	0.701	0.674	0.674
	1	0.842	0.850	0.845	0.857
Ave		0.776	0.800	0.787	0.796
MCC	**0**	0.494	0.551	0.521	0.540
	1	0.494	0.551	0.521	0.540
Ave		0.494	0.551	0.521	0.540
ROC Area	**0**	0.817	0.855	0.842	0.852
	1	0.817	0.855	0.842	0.852
Ave		0.817	0.855	0.842	0.852

2- Attribute selection method: CorrelationAttributeEval

One of popular filter metrics for classification problems is correlation. This method takes into account the usefulness of individual features for predicting the class label along with the level of intercorrelation among them.

The names of attributes that were selected by this method are:

Attributes: 12

pctKids2Par, pct2Par, pct12-17w2Par, pctAllDivorc, pctKidsBornNevrMarr, pctWhite

pctWdiv, pctHousOwnerOccup, medNumBedrm, pctNotHSgrad, pctWwage, Class

- **Classifier Algorithm: J48**

Correctly Classified Instances 463 77.1667 %

	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	Class
	0.688	0.186	0.653	0.670	0.496	0.776	0
	0.814	0.312	0.837	0.825	0.496	0.776	1
Weighted Avg.	0.772	0.269	0.775	0.773	0.496	0.776	

Confusion Matrix: J48

```
  A    b   <-- classified as
 139  63 |   a = 0
  74 324 |   b = 1
```

- **Classifier Algorithm: Multilayer Perceptron**

Correctly Classified Instances 470 78.3333 %

	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	Class
	0.653	0.151	0.688	0.670	0.509	0.822	0
	0.849	0.347	0.828	0.839	0.509	0.822	1
Weighted Avg.	0.783	0.281	0.781	0.782	0.509	0.822	

Confusion Matrix: Multilayer Perceptron

```
  a    b   <-- classified as
 132  70 |   a = 0
  60 338 |   b = 1
```

- **Classifier Algorithm: Random Forest**

Correctly Classified Instances 474 79 %

	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	Class
	0.629	0.128	0.713	0.668	0.518	0.844	0
	0.872	0.371	0.822	0.846	0.518	0.844	1

Weighted Avg.	0.790	0.289	0.786	0.786	0.518	0.844	

Confusion Matrix: **Random Forest**

```
   a     b   <-- classified as
 127  75 |   a = 0
  51 347 |   b = 1
```

- **Classifier Algorithm: Simple Logistic**

Correctly Classified Instances 476 79.33 %

	TP Rate	**FP Rate**	**Precision**	**F-Measure**	**MCC**	**ROC Area**	**Class**
	0.614	0.116	0.729	0.667	0.523	0.856	0
	0.884	0.386	0.819	0.850	0.523	0.856	1
Weighted Avg.	0.793	0.295	0.789	0.788	0.523	0.856	

Confusion Matrix: Simple Logistic

```
   a     b   <-- classified as
 124  78 |   a = 0
  46 352 |   b = 1
```

Best model Performance in **Correlation Attribute selection** method is Simple Logistic because it had the highest correctly classified instances which was 476 Correctly Classified Instances (79.33%) and also it had better other measure performance that shown with red item in following table.

	CLASS	**J48**	**Multilayer**	**Random Forest**	**Simple logistic**
TP	**0**	0.688	0.653	0.629	0.614
	1	0.814	0.849	0.872	0.884
Ave		0.772	0.783	0.790	0.793

FP	**0**	0.186	0.151	0.128	0.116
	1	0.312	0.347	0.371	0.386
Ave		0.269	0.281	0.289	0.295
Precision	**0**	0.653	0.688	0.713	0.729
	1	0.837	0.828	0.822	0.819
Ave		0.775	0.781	0.786	0.789
F-Measure	**0**	0.670	0.670	0.668	0.667
	1	0.825	0.839	0.846	0.850
Ave		0.773	0.782	0.786	0.788
MCC	**0**	0.496	0.509	0.518	0.523
	1	0.496	0.509	0.518	0.523
Ave		0.496	0.509	0.518	0.523
ROC Area	**0**	0.776	0.822	0.844	0.856
	1	0.776	0.822	0.844	0.856
Ave		0.776	0.822	0.844	0.856

3- Attribute selection method: GainRatioAttributeEval

This method evaluates the worth of an attribute by measuring the gain ratio with respect to the class. Gain Ratio is modification of the information Gain that takes number and size of branches into account when choosing an attribute.

The names of attributes that were selected by this method are:

Attributes: 13

pct2Par, pct12-17w2Par, pctPersOwnOccup, pctAllDivorc, pctBlack, medIncome

pctWdiv, houseVacant, persEmergShelt, persPerOwnOccup, pctFgnImmig-8

pctFgnImmig-5, Class

- **Classifier Algorithm: J48**

Correctly Classified Instances 463 77.1667 %

	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	Class
	0.460	0.070	0.769	0.576	0.459	0.777	0
	0.930	0.540	0.772	0.844	0.459	0.777	1
Weighted Avg.	0.772	0.382	0.771	0.754	0.459	0.777	

Confusion Matrix: J48

```
  a   b   <-- classified as
 93 109 |   a = 0
 28 370 |   b = 1
```

- **Classifier Algorithm: Multilayer Perceptron**

Correctly Classified Instances 495 82.5 %

	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	Class
	0.693	0.108	0.765	0.727	0.600	0.848	0
	0.892	0.307	0.851	0.871	0.600	0.848	1
Weighted Avg.	0.825	0.240	0.822	0.823	0.600	0.848	

Confusion Matrix : Multilayer Perceptron

```
  a   b   <-- classified as
140  62 |   a = 0
 43 355 |   b = 1
```

- **Classifier Algorithm: Random Forest**

Correctly Classified Instances 479 79.83%

	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	Class
	0.609	0.106	0.745	0.670	0.533	0.856	0
	0.894	0.391	0.818	0.855	0.533	0.856	1
Weighted	0.798	0.295	0.794	0.793	0.533	0.856	

Avg.							

Confusion Matrix: Random Forest

```
   a   b   <-- classified as
 123  79 |   a = 0
  42 356 |   b = 1
```

- **Classifier Algorithm: Simple Logistic**

Correctly Classified Instances 488 81.3333 %

	TP Rate	**FP Rate**	**Precision**	**F-Measure**	**MCC**	**ROC Area**	**Class**
	0.619	0.088	0.781	0.691	0.567	0.863	0
	0.912	0.381	0.825	0.866	0.567	0.863	1
Weighted Avg.	0.813	0.282	0.810	0.807	0.567	0.863	

Confusion Matrix: Simple Logistic

```
   a   b   <-- classified as
 125  77 |   a = 0
  35 363 |   b = 1
```

Best model Performance in **Gain Ratio Attribute selection** method is Multilayer Perceptron because it had the highest correctly classified instance which was 495 Correctly Classified Instances (82.5%) and also it had better measure performance than other classified models.

	CLASS	**J48**	**Multilayer**	**Random Forest**	**Simple logistic**
TP	**0**	0.460	0.693	0.609	0.619
	1	0.930	0.892	0.894	0.912
Ave		0.772	0.825	0.798	0.813
FP	**0**	0.070	0.108	0.106	0.088

	1	0.540	0.307	0.391	0.381
Ave		0.382	0.240	0.295	0.282
Precision	**0**	0.769	0.765	0.745	0.781
	1	0.772	0.851	0.818	0.825
Ave		0.771	0.822	0.794	0.810
F-Measure	**0**	0.576	0.727	0.670	0.691
	1	0.844	0.871	0.855	0.866
Ave		0.754	0.823	0.793	0.807
MCC	**0**	0.459	0.600	0.533	0.567
	1	0.459	0.600	0.533	0.567
Ave		0.459	0.600	0.533	0.567
ROC Area	**0**	0.777	0.848	0.856	0.863
	1	0.777	0.848	0.856	0.863
Ave		0.777	0.848	0.856	0.863

4- Attribute selection method: InfoGainAttributeEval

This method evaluates the worth of an attribute by measuring the information gain with respect to the class. It is measuring how each feature contributes in decreasing the overall entropy.

The names of attributes that were selected by this method are:

Attributes: 9

pct2Par, pctAllDivorc, persPoverty, pctWhite, pctPersOwnOccup, pctWdiv,

persHomeless, blackPerCap, Class

- **Classifier Algorithm: J48**

Correctly Classified Instances 483 80.5 %

	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	Class
	0.629	0.106	0.751	0.685	0.550	0.839	0
	0.894	0.371	0.826	0.859	0.550	0.839	1
Weighted Avg.	0.805	0.282	0.801	0.800	0.550	0.839	

Confusion Matrix: J48

```
  a   b   <-- classified as
127  75 |   a = 0
 42 356 |   b = 1
```

- **Classifier Algorithm: Multilayer Perceptron**

Correctly Classified Instances 476 79.3333 %

	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	Class
	0.649	0.133	0.712	0.679	0.528	0.842	0
	0.867	0.351	0.829	0.848	0.528	0.842	1
Weighted Avg.	0.793	0.278	0.790	0.791	0.528	0.842	

Confusion Matrix: Multilayer Perceptron

```
  a   b   <-- classified as
131  71 |   a = 0
 53 345 |   b = 1
```

- **Classifier Algorithm: Random Forest**

Correctly Classified Instances 490 81.6667 %

	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	Class
	0.673	0.111	0.756	0.712	0.580	0.853	0
	0.889	0.327	0.843	0.866	0.580	0.853	1
Weighted Avg.	0.817	0.254	0.813	0.814	0.580	0.853	

Confusion Matrix: Random Forest

```
   a     b   <-- classified as
 136  66 |   a = 0
  44 354 |   b = 1
```

- **Classifier Algorithm: Simple Logistic**

Correctly Classified Instances   485   80.8333 %

	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	Class
	0.599	0.085	0.781	0.678	0.554	0.860	0
	0.915	0.401	0.818	0.864	0.554	0.860	1
Weighted Avg.	0.808	0.295	0.805	0.801	0.554	0.860	

Confusion Matrix: Simple Logistic

```
   a     b   <-- classified as
 121  81 |   a = 0
  34 364 |   b = 1
```

Best model Performance in **Information Gain Attribute selection** method is Random Forest because it had the highest correctly classified instance which was 490 Correctly Classified Instances (81.67%) and also it had better measure performance than other classified models.

	CLASS	J48	Multilayer	Random Forest	Simple logistic
TP	**0**	0.629	0.649	0.673	0.599
	1	0.894	0.867	0.889	0.915
Ave		0.805	0.793	0.817	0.808
FP	**0**	0.106	0.133	0.111	0.085
	1	0.371	0.351	0.327	0.401
Ave		0.282	0.278	0.254	0.295
Precision	**0**	0.751	0.712	0.756	0.781

	1	0.826	0.829	0.843	0.818
Ave		0.801	0.790	0.813	0.805
F-Measure	**0**	0.685	0.679	0.712	0.678
	1	0.859	0.848	0.866	0.864
Ave		0.800	0.791	0.814	0.801
MCC	**0**	0.550	0.528	0.580	0.554
	1	0.550	0.528	0.580	0.554
Ave		0.550	0.528	0.580	0.551
ROC Area	**0**	0.839	0.842	0.853	0.860
	1	0.839	0.842	0.853	0.860
Ave		0.839	0.842	0.853	0.860

Discussion:

With an overall view of the all process and selected four models for final evaluation, we have:

CfsSubsetEval & Multilayer perceptron => 481 Correctly Classified Instances (80.167%)

CorrelationAttributeEval & Simple logistic => 476 Correctly Classified Instances (79.33%)

GainRatioAttributeEval & Multilayer perceptron => 495 Correctly Classified Instances (82.5%)

InfoGainAttributeEval & Random Forest => 490 Correctly Classified Instances (81.67%)

Attribute Selection Method		**CfsSubset**	**Correlation**	**Gain Ratio**	**Info Gain**
Model	**CLASS**	**Multilayer**	**Simple logistic**	**Multilayer**	**Random Forest**
TP	**0**	0.698	0.614	0.693	0.673
	1	0.852	0.884	0.892	0.889

Ave		0.800	0.793	0.825	0.817
FP	**0**	0.148	0.116	0.108	0.111
	1	0.302	0.386	0.307	0.327
Ave		0.250	0.295	0.240	0.254
Precision	**0**	0.705	0.729	0.765	0.756
	1	0.848	0.819	0.851	0.843
Ave		0.800	0.789	0.822	0.813
F-Measure	**0**	0.701	0.667	0.727	0.712
	1	0.850	0.850	0.871	0.866
Ave		0.800	0.788	0.823	0.814
MCC	**0**	0.551	0.523	0.600	0.580
	1	0.551	0.523	0.600	0.580
Ave		0.551	0.523	0.600	0.580
ROC Area	**0**	0.855	0.856	0.848	0.853
	1	0.855	0.856	0.848	0.853
Ave		0.855	0.856	0.848	0.853

By considering the above table and comparison of accuracies, we can say that Gain Ratio Attribute selection method with multilayer perceptron is the best model because it has the highest correctly classified instance which is 495 Correctly Classified Instances (82.5%) and it has the highest TP rate, lowest FP rate, highest Precision, F-Measure and MCC than other models, therefore, I chose this model as the best classification model.

What you learned from this project.

I learned how to use WEKA for building and testing classification models, and also how to evaluate measure performance for choosing the best model.

Any other observations from this project.

After doing this project I am interested to learn more concepts about data mining and applying it practically on huge data sets in real world applications.