

# Career Longevity for NBA Rookies

Data Mining

-By VENKATA VISHNUvardhan ALURI

## Goal of our Data Mining project

The aim of this project is to predict the career longevity for the National Basketball Association (NBA) rookies. This is achieved by performing several data mining tasks on the data set such as data preprocessing, attribute selection, data classifier modelling and performance comparison. The class attribute for prediction is the Target\_5Yrs attribute details of which will be elaborated in the following section of this report.

## Description of Dataset

This Dataset consists of twenty-one attributes with thirteen hundred and forty instances. Each record gives us the information about the player's basketball career such as games and minutes played, field goals made and so forth. All the attributes excluding the Name and the class attribute, are numeric attributes. The class attribute i.e. the Target\_5Yrs is a nominal attribute bearing Boolean values 0 and 1 for prediction. Each attribute description is elaborated in the following table.

	<b>Description</b>
<b>Name</b>	Name
<b>GP</b>	Games Played
<b>MIN</b>	MinutesPlayed
<b>PTS</b>	PointsPerGame
<b>FGM</b>	FieldGoalsMade
<b>FGA</b>	FieldGoalAttempts
<b>FG%</b>	FieldGoalPercent
<b>3P Made</b>	3PointMade
<b>3PA</b>	3PointAttempts
<b>3P%</b>	3PointAttempts
<b>FTM</b>	FreeThrowMade
<b>FTA</b>	FreeThrowAttempts
<b>FT%</b>	FreeThrowPercent
<b>OREB</b>	OffensiveRebounds
<b>DREB</b>	DefensiveRebounds
<b>REB</b>	Rebounds
<b>AST</b>	Assists
<b>STL</b>	Steals
<b>BLK</b>	Blocks
<b>TOV</b>	Turnovers
<b>TARGET_5Yrs</b>	Outcome: 1 if career length >= 5 yrs, 0 if < 5...

## Description of tools involved

### Weka

Waikato Environment for Knowledge Analysis (Weka) is a suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection.

### R Studio

RStudio is a free and open-source integrated development environment for R, a programming language for statistical computing and graphics. RStudio was founded by JJ Allaire, creator of the programming language ColdFusion.

### Microsoft Excel

Microsoft Excel is a spreadsheet developed by Microsoft for Windows, MAC OS, Android and iOS. It features calculation, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications.

## Data Mining Algorithms used

### Information Gain

ID3 uses information gain as its attribute selection measure. This measure is based on pioneering work by Claude Shannon on information theory, which studied the value or “information content” of messages. The attribute with the highest information gain is chosen as the splitting attribute, it minimizes the information needed to classify the tuples and reflects the least randomness.

### Gain Ratio

The information gain measure is biased toward tests with many outcomes. That is, it prefers to select attributes having a large number of values. C4.5, a successor of ID3, uses an extension to information gain known as gain ratio, which attempts to overcome this bias. It applies a kind of normalization to information gain using a “split information” value and represents the potential information generated by splitting the training data set into v partitions, corresponding to the v outcomes of a test on the attribute.

### Correlation Attribute Evaluation

Evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class.

## OneR AttributeEval

Evaluates the worth of an attribute by using the OneR classifier.

## Classifiers

### J48

Classification is the process of building a model of classes from a set of records that contain class labels. Decision Tree Algorithm is to find out the way the attributes-vector behaves for a number of instances. This algorithm generates the rules for the prediction of the target variable. With the help of tree classification algorithm, the critical distribution of the data is easily understandable. J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In the WEKA data-mining tool, J48 is an open source Java implementation of the C4.5 algorithm.

### Naïve Bayes

Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class-conditional independence. It is made to simplify the computations involved and, in this sense, is considered “naive.”

### Adaboost

AdaBoost, short for Adaptive Boosting, is a machine learning meta-algorithm formulated by Yoav Freund and Robert Schapire, who won the 2003 Gödel Prize for their work. It can be used in conjunction with many other types of learning algorithms to improve performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers.

### Bagging

Bootstrap aggregating, also called bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting. Although it is usually applied to decision tree methods, it can be used with any type of method. Bagging is a special case of the model averaging approach.

## Procedure

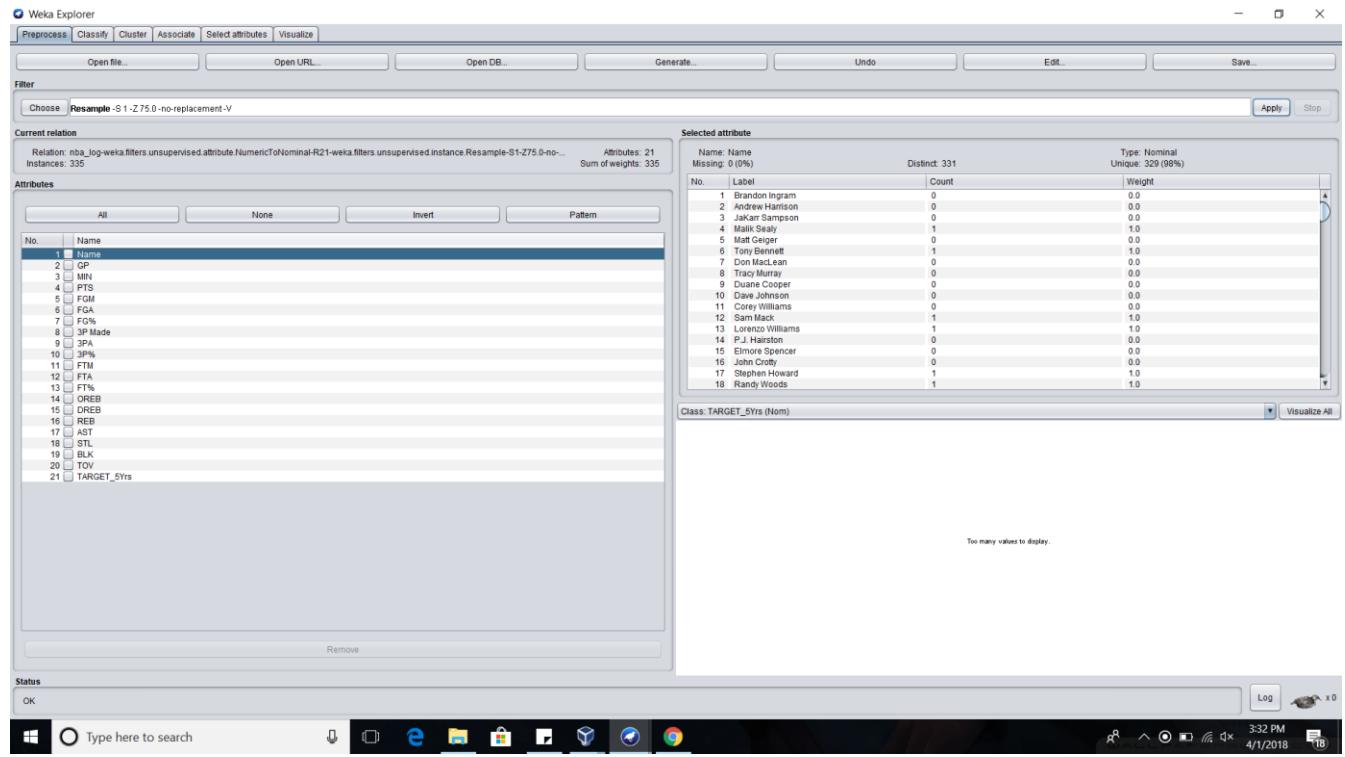
### Preprocessing

Datasets are never ideal. Our dataset has several missing values that should be addressed before classification and prediction in order to achieve accuracy. Data Smoothing helps addressing the issue where missing values are replaced with mean values of respective attributes. This is implemented using R studio where attribute values are fed in R to obtain mean.

Weka implements several filters to preprocess the data set, once the dataset was loaded, it has to be split into training and test set for modelling and testing. Selecting an unsupervised filter, resampling and dividing the set into 75% training and 25 % test by applying ‘True’ to the no replacement option the training set was achieved. The test set was achieved by applying ‘True’ to the invert selection while resampling. Below are screenshots for training and test sets after split. 1340 instances are split into 1005 and 335 respectively.

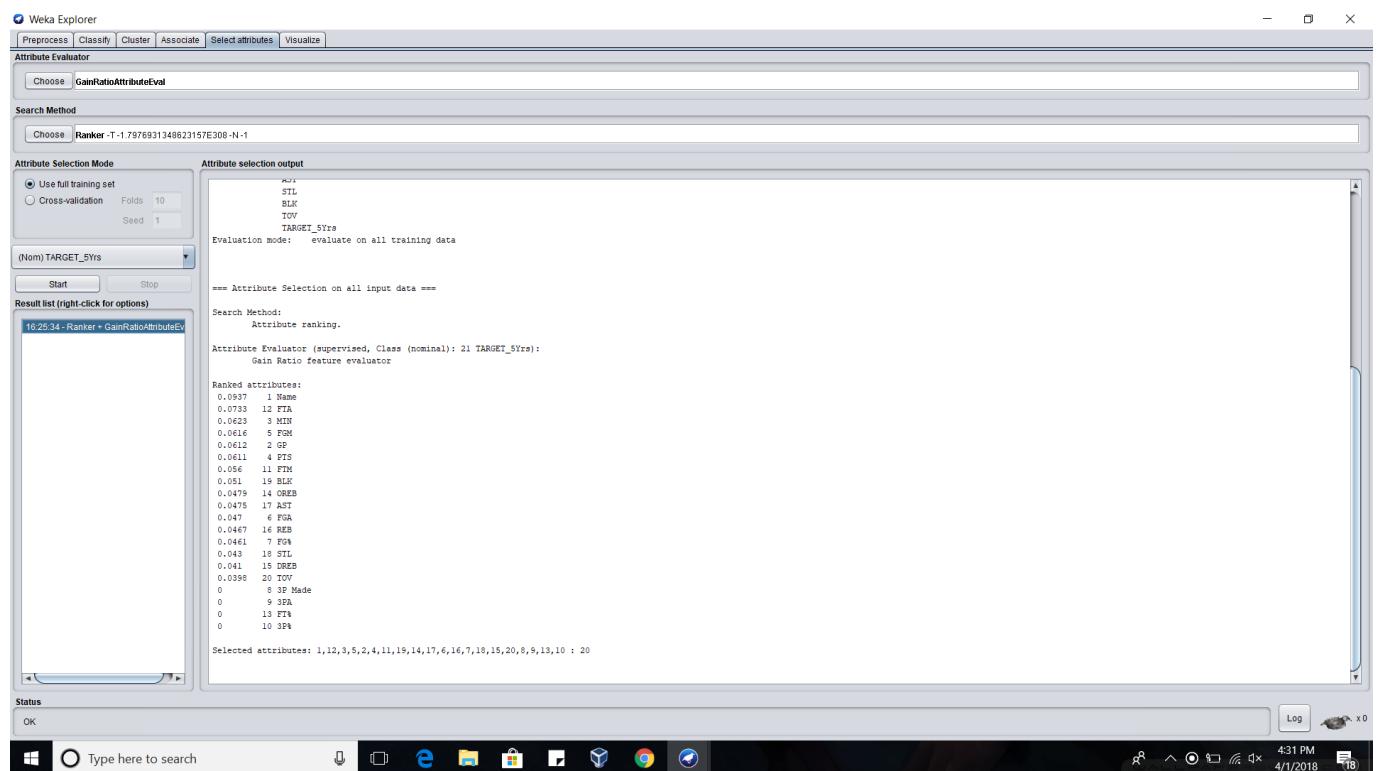
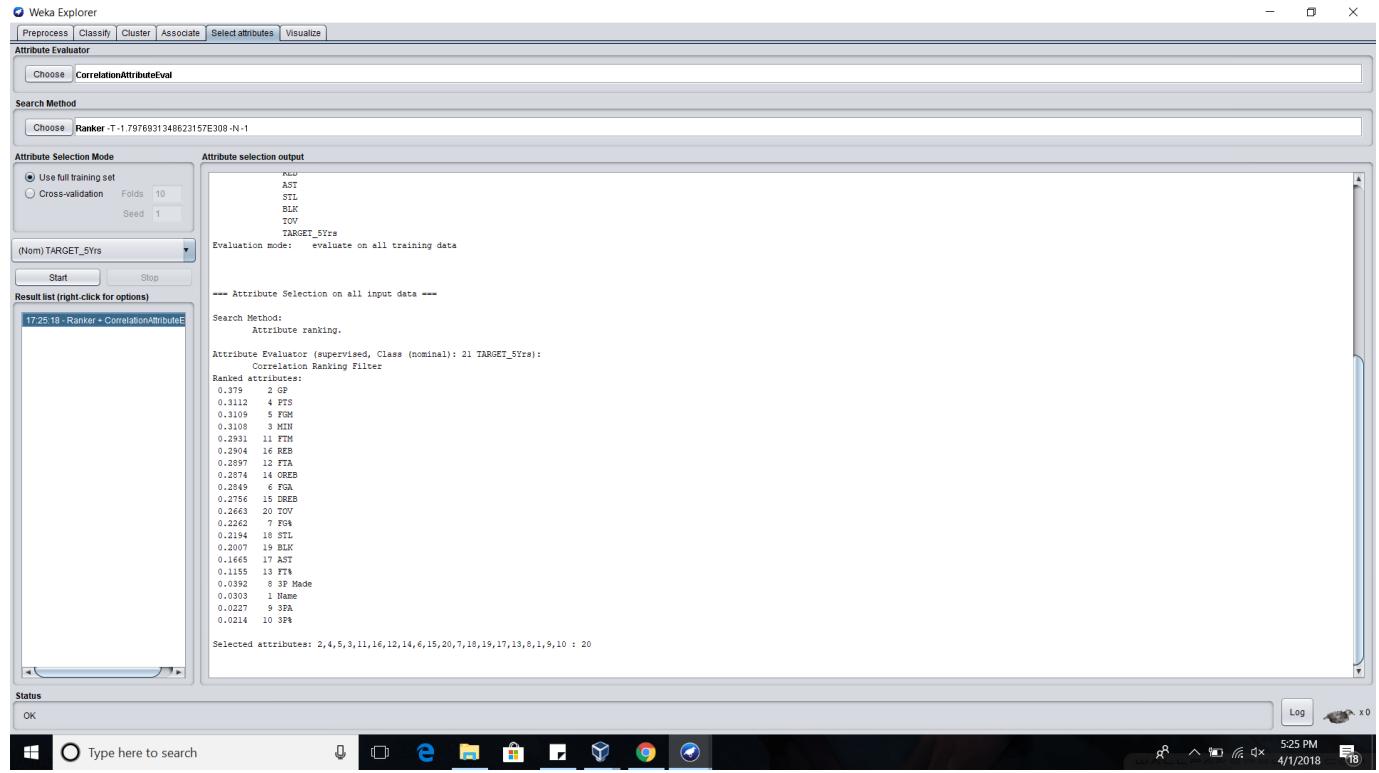
The screenshot shows the Weka Explorer interface with the 'Resample' filter selected. The top menu bar includes 'Preprocess', 'Classify', 'Cluster', 'Associate', 'Select attributes', and 'Visualize'. The main window displays the 'Selected attribute' table for the 'Name' attribute, which is nominal with 956 unique values. The 'Attributes' list on the left shows 21 attributes, with 'Name' being the selected attribute. The bottom status bar shows 'OK' and the system tray indicates the date and time as 4/1/2018 3:29 PM.

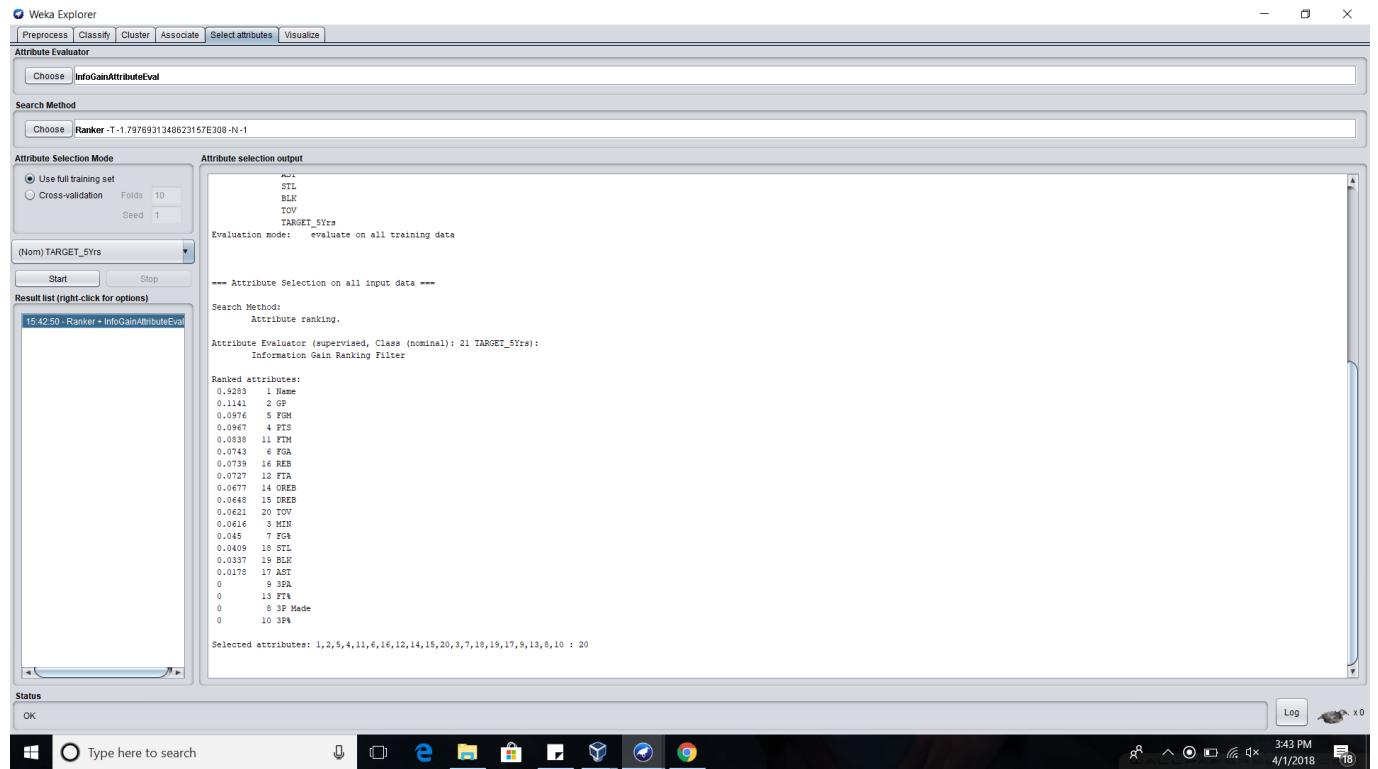
No.	Label	Count	Type: Nominal Unique: 956 (95%)
1	Brandon Ingram	1	1.0
2	Andrew Harrison	1	1.0
3	Jakar Sampson	1	1.0
4	Mike Bibby	0	0.0
5	Matt Geiger	1	1.0
6	Tony Bennett	0	0.0
7	Don MacLean	1	1.0
8	Mike Miller	1	1.0
9	Duane Cooper	1	1.0
10	Dave Johnson	1	1.0
11	Corey Williams	1	1.0
12	Sam Mack	0	0.0
13	Terrence Williams	0	0.0
14	P.J. Hairston	1	1.0
15	Elmore Spencer	1	1.0
16	John Croft	1	1.0
17	Stephen Howard	0	0.0
18	Randy Woods	0	0.0

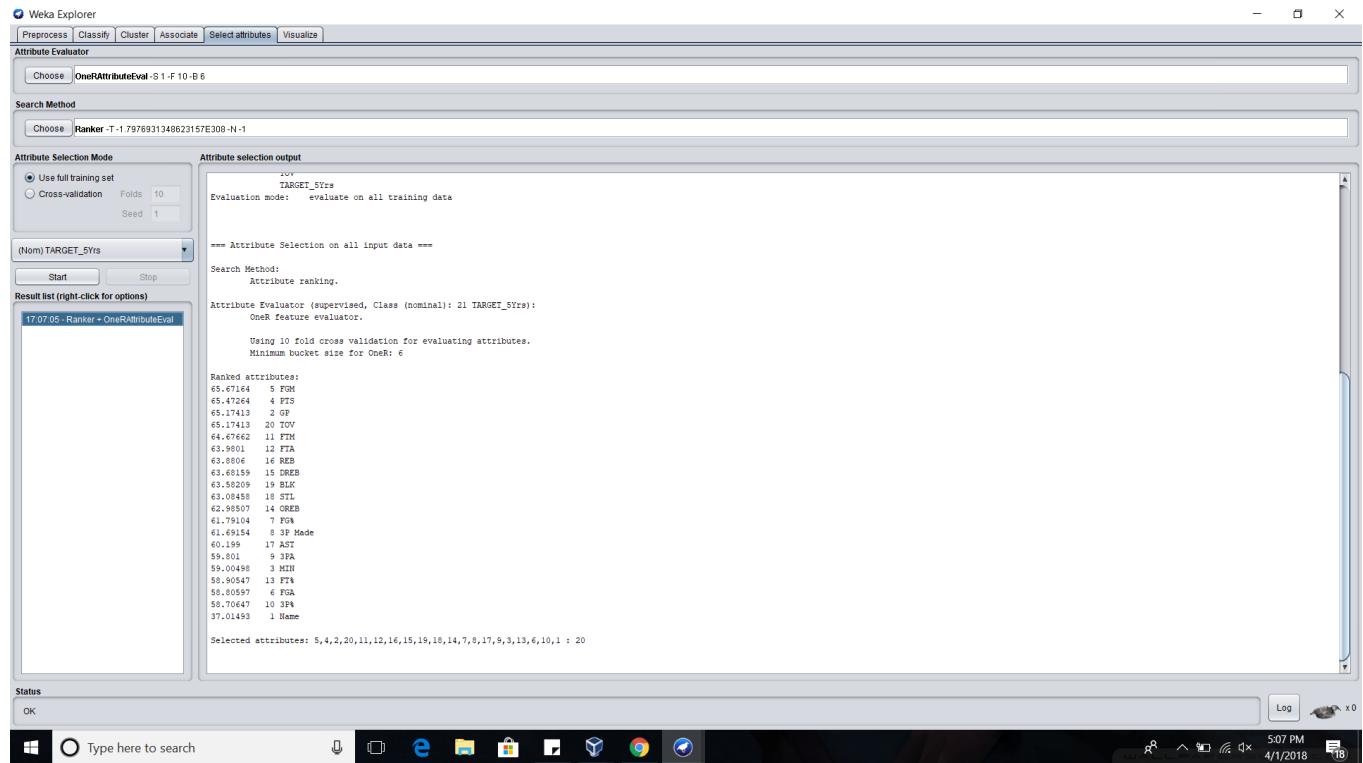


## Attribute Selection

Once the dataset is divided into training and test sets, we click on Select Attributes and run each of the algorithms. For each we select the attributes that are least likely to influence the dataset and remove them from the training set. Below are the screenshots for ranks of attributes after selecting algorithms Correlation Attribute Evaluation-Ranked, Gain Ratio-Ranked, Info Gain-Ranked and OneR – Ranked in order.







## Classification

After the attributes are removed based on the result of the select attribute algorithm, we click on the classify tab of Weka. Here we use several classification algorithms as mentioned above for each select attribute algorithm and report our findings. While classifying, the test set is supplied in the supplied test slot. Screenshots for every classification with respect to each selection attributes are followed.

## Correlation Attribute Evaluation – Naïve Bayes(Training)

**Classifier output:**

	BLK	TOW
mean	0.3242 0.4131	0.957 1.3473
std. dev.	377 628	0.2855 0.451
weight sum	0.1087 0.1087	0.2943 0.5017
precision	0.1444 0.1444	0.1444 0.1444

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

```
17:27:40 - bayesNaiveBayes
Time taken to build model: 0.02 seconds
*** Stratified cross-validation ***
*** Summary ***
Correctly Classified Instances      631          62.7861 %
Incorrectly Classified Instances   374          37.2139 %
Kappa statistic                   0.2588
Mean absolute error               0.1745
Root mean squared error           0.1872
Relative absolute error            75.8215 %
Root relative squared error       121.2737 %
Total Number of Instances         1005

*** Detailed Accuracy By Class ***


|               | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC   | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|-------|
| a             | 0.814   | 0.494   | 0.502     | 0.814  | 0.421     | 0.328 | 0.730    | 0.592    | 0     |
| b             | 0.516   | 0.196   | 0.822     | 0.516  | 0.634     | 0.328 | 0.730    | 0.805    | 1     |
| Weighted Avg. | 0.628   | 0.298   | 0.702     | 0.628  | 0.629     | 0.328 | 0.730    | 0.725    |       |


*** Confusion Matrix ***


|   |  | a   | b   | <-- classified as |
|---|--|-----|-----|-------------------|
|   |  | 0   | 1   |                   |
| 0 |  | 307 | 70  | a = 0             |
| 1 |  | 304 | 324 | b = 1             |


```

## Correlation Attribute Evaluation – Naïve Bayes(Test)

**Classifier output:**

	BLK	TOW
mean	0.2655 0.451	0.957 1.3472
std. dev.	0.2943 0.5017	0.5543 0.7589
weight sum	377 628	377 628
precision	0.1444 0.1444	0.1051 0.1051

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

```
17:28:34 - bayesNaiveBayes
Time taken to build model: 0.01 seconds
*** Evaluation on test set ***
Time taken to test model on supplied test set: 0.01 seconds
*** Summary ***
Correctly Classified Instances      205          61.194 %
Incorrectly Classified Instances   130          38.806 %
Kappa statistic                   0.267
Mean absolute error               0.1857
Root mean squared error           0.6011
Relative absolute error            81.0148 %
Root relative squared error       122.924 %
Total Number of Instances         335

*** Detailed Accuracy By Class ***


|               | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC   | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|-------|
| a             | 0.818   | 0.522   | 0.505     | 0.818  | 0.424     | 0.301 | 0.746    | 0.626    | 0     |
| b             | 0.478   | 0.182   | 0.802     | 0.478  | 0.599     | 0.301 | 0.746    | 0.822    | 1     |
| Weighted Avg. | 0.612   | 0.316   | 0.685     | 0.612  | 0.609     | 0.301 | 0.746    | 0.745    |       |


*** Confusion Matrix ***


|   |  | a   | b  | <-- classified as |
|---|--|-----|----|-------------------|
|   |  | 0   | 1  |                   |
| 0 |  | 108 | 24 | a = 0             |
| 1 |  | 106 | 97 | b = 1             |


```

## Correlation Attribute Evaluation – J48(Training)

Weka Explorer

Classifier output

```

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) TARGET_5Yrs
Start Stop
Result list (right-click for options)
17:29:16 - trees_J48

AST
STL
BLK
TOV
TARGET_5Yrs
Test mode: 10-fold cross-validation

*** Classifier model (full training set) ***

J48 pruned tree
-----
: 1 (1005.0/377.0)

Number of Leaves : 1
Size of the tree : 1

Time taken to build model: 0.04 seconds

*** Stratified cross-validation ***
*** Summary ***

Correctly Classified Instances      628      62.4876 %
Incorrectly Classified Instances   377      37.5124 %
Kappa statistic                   0
Mean absolute error               0.4488
Root mean squared error           0.4442
Relative absolute error            99.9853 %
Root relative squared error       100      %
Total Number of Instances         1005

*** Detailed accuracy By Class ***

      TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.000 0.000 ? 0.000 ? ? 0.496 0.375 0
1.000 1.000 0.625 1.000 0.769 ? 0.496 0.623 1
Weighted Avg. 0.625 0.625 ? 0.625 ? ? 0.496 0.529

*** Confusion Matrix ***

a b <- classified as
0 377 | a = 0
0 628 | b = 1

```

Status

OK

Type here to search

Log

5:29 PM 4/1/2018

## Correlation Attribute Evaluation – J48(Test)

Weka Explorer

Classifier output

```

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) TARGET_5Yrs
Start Stop
Result list (right-click for options)
17:30:17 - trees_J48

TOV
TARGET_5Yrs
Test mode: user supplied test set: size unknown (reading incrementally)

*** Classifier model (full training set) ***

J48 pruned tree
-----
: 1 (1005.0/377.0)

Number of Leaves : 1
Size of the tree : 1

Time taken to build model: 0.01 seconds

*** Evaluation on test set ***

Time taken to test model on supplied test set: 0 seconds

*** Summary ***

Correctly Classified Instances      203      60.597 %
Incorrectly Classified Instances   132      39.403 %
Kappa statistic                   0
Mean absolute error               0.4735
Root mean squared error           0.489
Relative absolute error            99.9889 %
Root relative squared error       100.0019 %
Total Number of Instances         335

*** 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 0.394 0
1.000 1.000 0.606 1.000 0.755 ? 0.500 0.606 1
Weighted Avg. 0.606 0.606 ? 0.606 ? ? 0.500 0.522

*** Confusion Matrix ***

a b <- classified as
0 132 | a = 0
0 203 | b = 1

```

Status

OK

Type here to search

Log

5:30 PM 4/1/2018

## Correlation Attribute Evaluation – Bagging(Training)

Weka Explorer

Classifier

Choose: Bagging -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.REPTree -- -M 2 -V 0.001 -N 3 -S 1 -L 1 -I 0.0

**Test options**

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

(Nom) TARGET\_5Yrs

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

22:00:02 - meta.Bagging

```

Time taken to build model: 1.41 seconds

==== Classifier model (full training set) ====
Bagging with 10 iterations and base learner
weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L 1 -I 0.0

Time taken to build model: 1.41 seconds

==== Stratified cross-validation ====
==== Summary ====

```

Correctly Classified Instances	627	62.3881 %
Incorrectly Classified Instances	378	37.6119 %
Kappa statistic	0.0085	
Mean absolute error	0.4702	
Root mean squared error	0.4886	
Relative absolute error	100.2709 %	
Root relative squared error	100.9181 %	
Total Number of Instances	1005	

```

==== Detailed Accuracy By Class ====

```

TP Rate	FP Rate	Precision	Recall	F-Measure	NCC	ROC Area	FPR Area	Class
0.021	0.014	0.471	0.021	0.441	0.036	0.495	0.377	0
0.986	0.979	0.827	0.986	0.766	0.036	0.495	0.617	1
Weighted Avg.	0.424	0.017	0.569	0.424	0.494	0.026	0.495	0.527

```

==== Confusion Matrix ====

```

a	b	<-- classified as
8 365	1	a = 0
9 619	1	b = 1

Status

OK

Type here to search

Log x0

10:00 PM 4/1/2018

## Correlation Attribute Evaluation – Bagging(Test)

Weka Explorer

Classifier

Choose: Bagging -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.REPTree -- -M 2 -V 0.001 -N 3 -S 1 -L 1 -I 0.0

**Test options**

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

(Nom) TARGET\_5Yrs

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

17:34:44 - meta.Bagging

```

Time taken to build model: 0.97 seconds

==== Evaluation on test set ====

```

Time taken to test model on supplied test set: 0.01 seconds

```

==== Summary ====

```

Correctly Classified Instances	203	60.597 %
Incorrectly Classified Instances	132	39.403 %
Kappa statistic	0.0064	
Mean absolute error	0.473	
Root mean squared error	0.4923	
Relative absolute error	99.1504 %	
Root relative squared error	100.6655 %	
Total Number of Instances	335	

```

==== Detailed Accuracy By Class ====

```

TP Rate	FP Rate	Precision	Recall	F-Measure	NCC	ROC Area	FPR Area	Class
0.015	0.010	0.590	0.015	0.029	0.024	0.508	0.398	0
0.990	0.985	0.607	0.990	0.753	0.024	0.508	0.615	1
Weighted Avg.	0.606	0.601	0.565	0.606	0.468	0.024	0.508	0.530

```

==== Confusion Matrix ====

```

a	b	<-- classified as
2 130	1	a = 0
2 201	1	b = 1

Status

OK

Type here to search

Log x0

5:34 PM 4/1/2018

## Correlation Attribute Evaluation – AdaBoost(Training)

The screenshot shows the Weka Explorer interface with the following details:

- Classifier:** Choose: AdaBoostM1-P 100-S 1-I 10-W weka.classifiers.bayes.NaiveBayes
- Test options:**
  - Use training set
  - Supplied test set Set...
  - Cross-validation Folds 10
  - Percentage split % 66
  - More options...
- Classifier output:**

	mean	0.4212	0.4125
std. dev.	0.4371	0.4965	
weight sum	367.4942	637.5058	
precision	0.1444	0.1444	
TOW			
mean	1.2581	1.5132	
std. dev.	0.825	0.9832	
weight sum	367.4942	637.5058	
precision	0.1051	0.1051	

(Nom) TARGET\_5Yrs

Start Stop

Result list (right-click for options)

```
18:18:14 - meta.AdaBoostM1
Weight: 0.12
Number of performed Iterations: 10
Time taken to build model: 0.31 seconds
--- Stratified cross-validation ---
--- Summary ---

Correctly Classified Instances 669 66.5672 %
Incorrectly Classified Instances 336 33.4328 %
Kappa statistic 0.2784
Mean absolute error 0.3827
Root mean squared error 0.5127
Relative absolute error 81.6124 %
Root relative squared error 105.8866 %
Total Number of Instances 1005

--- Detailed Accuracy By Class ---

IP Rate FP Rate Precision Recall F-Measure MCC ROC Area FRC Area Class
0.525 0.250 0.558 0.525 0.541 0.279 0.652 0.511 0
0.750 0.475 0.725 0.750 0.737 0.279 0.652 0.732 1
Weighted Avg. 0.666 0.390 0.662 0.666 0.664 0.279 0.652 0.649

--- Confusion Matrix ---

a b <-- classified as
198 179 | a = 0
157 471 | b = 1
```

Status: OK

## Correlation Attribute Evaluation – AdaBoost(Test)

The screenshot shows the Weka Explorer interface with the following details:

- Classifier:** Choose: AdaBoostM1-P 100-S 1-I 10-W weka.classifiers.bayes.NaiveBayes
- Test options:**
  - Use training set
  - Supplied test set Set...
  - Cross-validation Folds 10
  - Percentage split % 66
  - More options...
- Classifier output:**

	precision	0.1444	0.1444
TOW			
mean	1.2581	1.5132	
std. dev.	0.825	0.9832	
weight sum	367.4942	637.5058	
precision	0.1051	0.1051	

(Nom) TARGET\_5Yrs

Start Stop

Result list (right-click for options)

```
18:19:33 - meta.AdaBoostM1
Weight: 0.12
Number of performed Iterations: 10
Time taken to build model: 0.33 seconds
--- Evaluation on test set ---
Time taken to test model on supplied test set: 0.05 seconds
--- Summary ---

Correctly Classified Instances 221 65.9701 %
Incorrectly Classified Instances 114 34.0299 %
Kappa statistic 0.2759
Mean absolute error 0.3936
Root mean squared error 0.4132
Relative absolute error 83.0562 %
Root relative squared error 104.9573 %
Total Number of Instances 335

--- Detailed Accuracy By Class ---

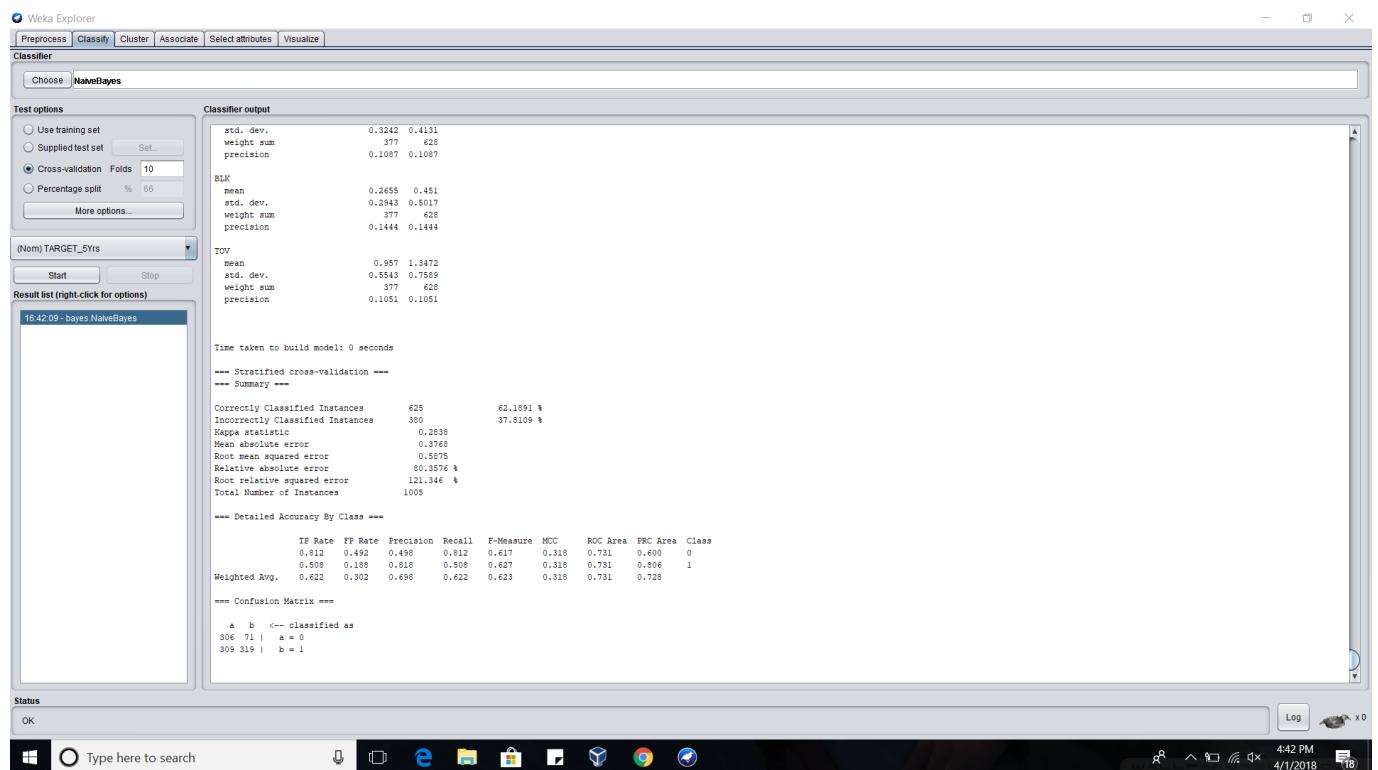
IP Rate FP Rate Precision Recall F-Measure MCC ROC Area FRC Area Class
0.523 0.251 0.575 0.523 0.548 0.277 0.637 0.534 0
0.749 0.477 0.707 0.749 0.727 0.277 0.637 0.687 1
Weighted Avg. 0.660 0.388 0.655 0.660 0.656 0.277 0.637 0.627

--- Confusion Matrix ---

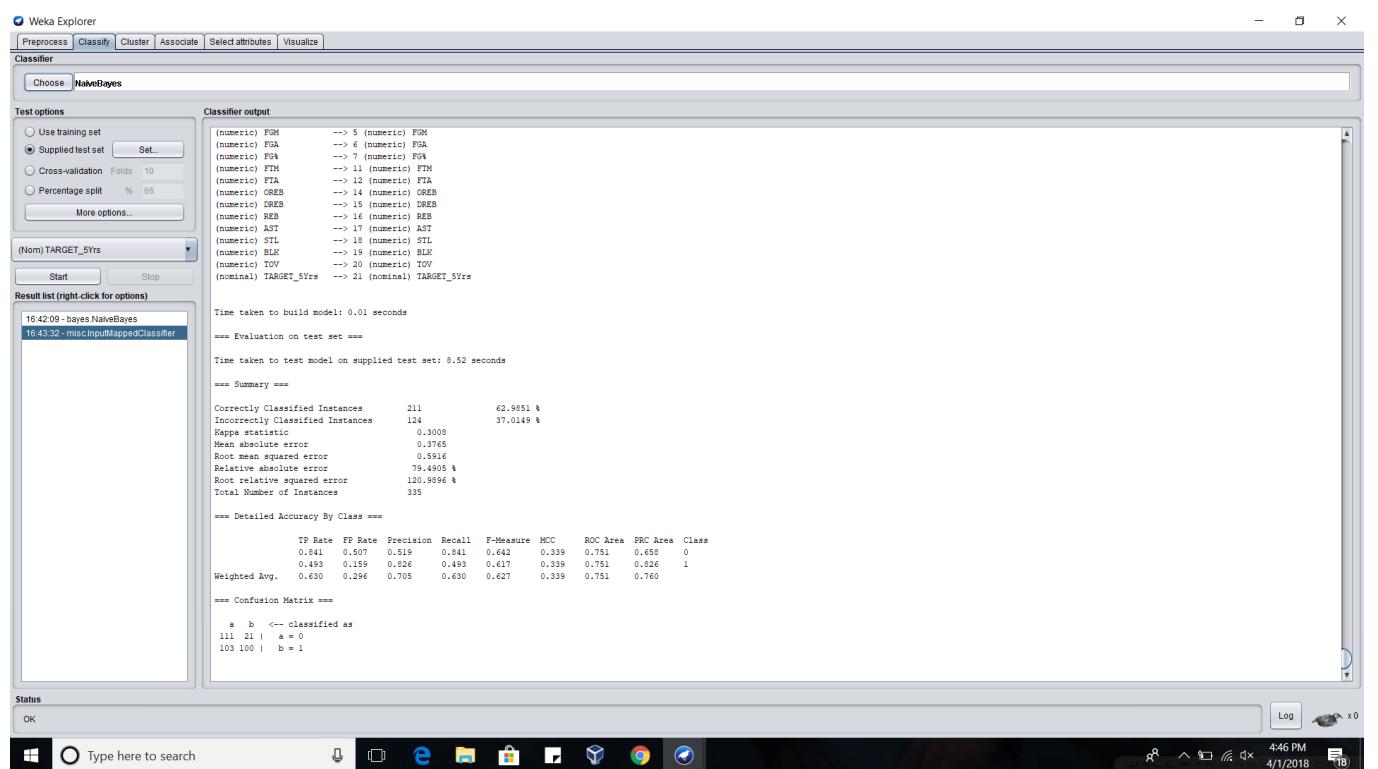
a b <-- classified as
69 63 | a = 0
51 152 | b = 1
```

Status: OK

## Gain Ratio – Naïve Bayes(Training)



## Gain Ratio – Naïve Bayes(Test)



## Gain Ratio – J48(Training)

The screenshot shows the Weka Explorer interface with the 'Classifier' tab selected. The classifier chosen is 'J48 - C 0.25 - M 2'. The 'Test options' panel indicates 'Cross-validation Folds 10'. The 'Classifier output' panel displays the following text:

```

AST
STL
BLK
TOV
TARGET_5Yrs

Test mode: 10-fold cross-validation

*** Classifier model (full training set) ***

J48 pruned tree
-----
: 1 (1005.0/377.0)

Number of Leaves : 1

Size of the tree : 1

Time taken to build model: 0.04 seconds

--- Stratified cross-validation ---
--- Summary ---

Correctly Classified Instances      628          62.4876 %
Incorrectly Classified Instances   377          37.5124 %
Kappa statistic                   0
Mean absolute error               0.4888
Root mean squared error           0.4442
Relative absolute error            99.9853 %
Root relative squared error       100 %
Total Number of Instances         1005

--- Detailed Accuracy By Class ---

      IP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Class
0.000  0.000    ?        0.000    ?        ?        0.486   0.373   0
1.000  1.000    0.625   1.000    0.769    ?        0.496   0.623   1
Weighted Avg.  0.625  0.625    ?        0.625    ?        ?        0.496   0.529

--- Confusion Matrix ---

a b <-- classified as
0 377 | a = 0
0 628 | b = 1

```

The status bar at the bottom shows 'OK'.

## Gain Ratio – J48(Test)

The screenshot shows the Weka Explorer interface with the 'Classifier' tab selected. The classifier chosen is 'J48 - C 0.25 - M 2'. The 'Test options' panel indicates 'Supplied test set'. The 'Classifier output' panel displays the following text:

```

(Fnom) FGM      --> 5 (numeric) FGM
(Fnom) FGA      --> 6 (numeric) FGA
(Fnom) FG9      --> 7 (numeric) FG9
(Fnom) FTH      --> 11 (numeric) FTH
(Fnom) ITA      --> 12 (numeric) ITA
(Fnom) GREB     --> 14 (numeric) GREB
(Fnom) DREB     --> 15 (numeric) DREB
(Fnom) REB      --> 16 (numeric) REB
(Fnom) AST      --> 17 (numeric) AST
(Fnom) STL      --> 18 (numeric) STL
(Fnom) BLK      --> 19 (numeric) BLK
(Fnom) TOV      --> 20 (numeric) TOV
(nominal) TARGET_5Yrs --> 21 (nominal) TARGET_5Yrs

Time taken to build model: 0.01 seconds

*** Evaluation on test set ***

Time taken to test model on supplied test set: 0.25 seconds

*** Summary ---

Correctly Classified Instances      203          60.597 %
Incorrectly Classified Instances   132          39.403 %
Kappa statistic                   0
Mean absolute error               0.4735
Root mean squared error           0.489
Relative absolute error            99.9889 %
Root relative squared error       100.0019 %
Total Number of Instances         335

--- Detailed Accuracy By Class ---

      IP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Class
0.000  0.000    ?        0.000    ?        ?        0.500   0.394   0
1.000  1.000    0.606   1.000    0.755    ?        0.500   0.606   1
Weighted Avg.  0.606  0.606    ?        0.606    ?        ?        0.500   0.522

--- Confusion Matrix ---

a b <-- classified as
0 132 | a = 0
0 203 | b = 1

```

The status bar at the bottom shows 'OK'.

## Gain Ratio – Bagging(Training)

Weka Explorer

Classifier output

```
Choose: Bagging -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.REPTree -- -M 2 -V 0.001 -N 3 -S 1 -L 1 -I 0.0
```

Test options

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

(Nom) TARGET\_5Yrs

Start Stop

Result list (right-click for options)

16:51:18 - meta.Bagging

```
Bagging with 10 iterations and base learner
weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L 1 -I 0.0

Time taken to build model: 0.75 seconds

*** Classifier model (full training set) ***

Bagging with 10 iterations and base learner
weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L 1 -I 0.0

Time taken to build model: 0.75 seconds

*** Stratified cross-validation ***
*** Summary ***

Correctly Classified Instances      627           62.3801 %
Incorrectly Classified Instances   378           37.6119 %
Kappa statistic                   0.0085
Mean absolute error               0.4702
Root mean squared error           0.4886
Relative absolute error            100.2709 %
Root relative squared error       100.9181 %
Total Number of Instances         1005

*** Detailed Accuracy By Class ***

     TP Rate   FP Rate  Precision  Recall   F-Measure  MCC   ROC Area  FPR Area  Class
0       0.021    0.014   0.471    0.021   0.041    0.026   0.495    0.377    0
1       0.986    0.979   0.427    0.986   0.766    0.026   0.495    0.617    1
Weighted Avg.   0.624    0.617   0.568    0.624   0.494    0.026   0.495    0.527

*** Confusion Matrix ***

a   b   <- classified as
0   627  |   a = 0
9   378  |   b = 1
```

Status

OK

Type here to search

Log x0

451 PM 4/1/2018

## Gain Ratio – Bagging(Test)

Weka Explorer

Classifier output

```
Choose: Bagging -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.REPTree -- -M 2 -V 0.001 -N 3 -S 1 -L 1 -I 0.0
```

Test options

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

(Nom) TARGET\_5Yrs

Start Stop

Result list (right-click for options)

16:51:18 - meta.Bagging

16:53:12 - misc.InputMappedClassifier

```
Time taken to build model: 0.71 seconds

*** Evaluation on test set ***

Time taken to test model on supplied test set: 0.13 seconds

*** Summary ***

Correctly Classified Instances      203           60.597 %
Incorrectly Classified Instances   132           39.403 %
Kappa statistic                   0.0064
Mean absolute error               0.473
Root mean squared error           0.4923
Relative absolute error            100.8804 %
Root relative squared error       100.6658 %
Total Number of Instances         335

*** Detailed Accuracy By Class ***

     TP Rate   FP Rate  Precision  Recall   F-Measure  MCC   ROC Area  FPR Area  Class
0       0.015    0.010   0.500    0.015   0.029    0.024   0.508    0.398    0
1       0.990    0.985   0.607    0.990   0.753    0.024   0.508    0.615    1
Weighted Avg.   0.606    0.601   0.565    0.606   0.468    0.024   0.508    0.530

*** Confusion Matrix ***

a   b   <- classified as
2   130  |   a = 0
2   201  |   b = 1
```

Status

OK

Type here to search

Log x0

453 PM 4/1/2018

## Gain Ratio – AdaBoost(Training)

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose AdaBoostM1-P 100-S 1 -I10 -W weka.classifiers.bayes.NaiveBayes

Test options

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

(Nom) TARGET\_5Yrs

Start Stop

Result list (right-click for options)

18:14:20 - meta AdaBoostM1

```

Time taken to build model: 0.28 seconds
--- Stratified cross-validation ---
--- Summary ---

Correctly Classified Instances      646      64.2786 %
Incorrectly Classified Instances   359      35.7214 %
Kappa statistic                   0.222
Mean absolute error               0.3954
Root mean squared error          0.5215
Relative absolute error           84.3209 %
Root relative squared error     107.7189 %
Total Number of Instances        1005

--- Detailed Accuracy By Class ---
            TP Rate FP Rate Precision Recall  F-Measure MCC ROC Area PR Area Class
0          0.472   0.255    0.527   0.472   0.498   0.223   0.636   0.452   0
1          0.745   0.528    0.702   0.745   0.733   0.223   0.636   0.728   1
Weighted Avg.  0.643   0.425    0.636   0.643   0.639   0.223   0.636   0.636   1

--- Confusion Matrix ---
      a   b   <- classified as
  178 159 | a = 0
  160 468 | b = 1

```

Status

OK

Type here to search

Log x0

6:14 PM 4/1/2018

## Gain Ratio – AdaBoost(Test)

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose AdaBoostM1-P 100-S 1 -I10 -W weka.classifiers.bayes.NaiveBayes

Test options

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

(Nom) TARGET\_5Yrs

Start Stop

Result list (right-click for options)

18:15:39 - misc.InputMappedClassifier

```

Time taken to build model: 0.26 seconds
--- Evaluation on test set ---
Time taken to test model on supplied test set: 9.08 seconds

--- Summary ---

Correctly Classified Instances      234      69.8507 %
Incorrectly Classified Instances   101      30.1493 %
Kappa statistic                   0.3695
Mean absolute error               0.356
Root mean squared error          0.5137
Relative absolute error           76.1176 %
Root relative squared error     105.0424 %
Total Number of Instances        335

--- Detailed Accuracy By Class ---
            TP Rate FP Rate Precision Recall  F-Measure MCC ROC Area PR Area Class
0          0.621   0.231    0.637   0.621   0.619   0.369   0.672   0.645   0
1          0.749   0.379    0.752   0.749   0.751   0.369   0.672   0.705   1
Weighted Avg.  0.699   0.329    0.699   0.699   0.699   0.369   0.672   0.649

--- Confusion Matrix ---
      a   b   <- classified as
  62 50 | a = 0
  51 152 | b = 1

```

Status

OK

Type here to search

Log x0

6:15 PM 4/1/2018

## Info Gain – Naïve Bayes(Training)

Weka Explorer

Classifier

Choose NaiveBayes

**Test options**

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

(Nom) TARGET\_5Yrs

Start Stop

Result list (right-click for options)

```
16:31:25 - bayes.NaiveBayes
15:47:02 - bayes.NaiveBayes
16:04:28 - bayes.NaiveBayes
```

Time taken to build model: 0.03 seconds

\*\*\* Stratified cross-validation \*\*\*

\*\*\* Summary \*\*\*

	Correctly Classified Instances	625	62.1591 %
Incorrectly Classified Instances	380	37.8109 %	
Kappa statistic	0.2838		
Mean absolute error	0.3768		
Root mean squared error	0.3776		
Relative absolute error	86.3576 %		
Root relative squared error	121.346 %		
Total Number of Instances	1005		

\*\*\* Detailed Accuracy By Class \*\*\*

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.812	0.492	0.495	0.812	0.617	0.318	0.731	0.600	1	
0.508	0.188	0.818	0.508	0.427	0.318	0.731	0.806	0	
Weighted Avg.	0.622	0.302	0.698	0.622	0.423	0.318	0.728		

\*\*\* Confusion Matrix \*\*\*

```
a b <- classified as
306 71 | 0
309 319 | 1
```

Status

OK

Type here to search

Log

4:04 PM 4/1/2018

## Info Gain – Naïve Bayes(Test)

Weka Explorer

Classifier

Choose NaiveBayes

**Test options**

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

(Nom) TARGET\_5Yrs

Start Stop

Result list (right-click for options)

```
16:31:25 - bayes.NaiveBayes
15:47:02 - bayes.NaiveBayes
16:04:28 - bayes.NaiveBayes
16:05:46 - misc.InputMappedClassifier
```

Time taken to build model: 0.01 seconds

\*\*\* Evaluation on test set \*\*\*

Time taken to test model on supplied test set: 0.6 seconds

\*\*\* Summary \*\*\*

	Correctly Classified Instances	211	62.9851 %
Incorrectly Classified Instances	124	37.0149 %	
Kappa statistic	0.3008		
Mean absolute error	0.3755		
Root mean squared error	0.5516		
Relative absolute error	79.4905 %		
Root relative squared error	120.9896 %		
Total Number of Instances	335		

\*\*\* Detailed Accuracy By Class \*\*\*

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.841	0.507	0.519	0.841	0.642	0.339	0.751	0.658	0	
0.493	0.159	0.826	0.493	0.617	0.339	0.751	0.826	1	
Weighted Avg.	0.630	0.296	0.705	0.630	0.627	0.339	0.760		

\*\*\* Confusion Matrix \*\*\*

```
a b <- classified as
111 21 | 0
103 100 | 1
```

Status

OK

Type here to search

Log

4:06 PM 4/1/2018

## Info Gain – J48(Training)

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose J48 -C 0.25-M 2

Test options

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

(Nom) TARGET\_5Yrs

Start Stop

Result list (right-click for options)

- 16:31:25 - bayes NaiveBayes
- 15:47:02 - bayes NaiveBayes
- 16:04:28 - bayes NaiveBayes
- 16:00:46 - misc.InputMappedClassifier
- 16:08:53 - trees J48

Classifier output

```

ASI
STL
BLK
TOV
TARGET_5Yrs
Test mode: 10-fold cross-validation

*** Classifier model (full training set) ***

J48 pruned tree
=====
: 1 (1005.0/377.0)

Number of Leaves : 1

Size of the tree : 1

Time taken to build model: 0.06 seconds

*** Stratified cross-validation ***

*** Summary ***

Correctly Classified Instances      628          62.4876 %
Incorrectly Classified Instances   377          37.5124 %
Kappa statistic                   0
Mean absolute error               0.4488
Root mean squared error           0.4542
Relative absolute error            99.9853 %
Root relative squared error       100 %
Total Number of Instances         1005

*** Detailed Accuracy By Class ***

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  FRC Area  Class
0.000    0.000     ?        0.000     ?        ?        0.496   0.373     0
1.000    1.000     0.625   1.000     0.769     ?        0.496   0.623     1
Weighted Avg.                      0.625     ?        0.625     ?        ?        0.496   0.529

*** Confusion Matrix ***

a  b  <- classified as
0 377 | a = 0
0 628 | b = 1

```

Status

OK

## Info Gain – J48(test)

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose J48 -C 0.25-M 2

Test options

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

(Nom) TARGET\_5Yrs

Start Stop

Result list (right-click for options)

- 16:31:25 - bayes NaiveBayes
- 15:47:02 - bayes NaiveBayes
- 16:04:28 - bayes NaiveBayes
- 16:00:46 - misc.InputMappedClassifier
- 16:10:59 - misc.InputMappedClassifier
- 16:10:59 - misc.InputMappedClassifier

Classifier output

```

(FGM) FGK --> 5 (numeric) FGK
(FGM) FGK --> 1 (nominal) FGK
(FGM) FGK --> 7 (numeric) FGK
(FGM) FGK --> 11 (numeric) FGK
(FGM) FGK --> 12 (nominal) FGK
(FGM) FGK --> 14 (numeric) FGK
(FGM) FGK --> 15 (nominal) FGK
(FGM) FGK --> 17 (nominal) FGK
(FGM) FGK --> 18 (nominal) FGK
(FGM) FGK --> 19 (nominal) FGK
(FGM) FGK --> 20 (nominal) FGK
(nominal) TARGET_5Yrs --> 21 (nominal) TARGET_5Yrs

Time taken to build model: 0.01 seconds

*** Evaluation on test set ***

Time taken to test model on supplied test set: 0.63 seconds

*** Summary ***

Correctly Classified Instances      203          60.597 %
Incorrectly Classified Instances   132          39.403 %
Kappa statistic                   0
Mean absolute error               0.4735
Root mean squared error           0.489
Relative absolute error            99.9889 %
Root relative squared error       100.0019 %
Total Number of Instances         335

*** Detailed Accuracy By Class ***

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  FRC Area  Class
0.000    0.000     ?        0.000     ?        ?        0.394   0.394     0
1.000    1.000     0.606   1.000     0.755     ?        0.500   0.606     1
Weighted Avg.                      0.606     ?        0.606     ?        ?        0.500   0.522

*** Confusion Matrix ***

a  b  <- classified as
0 132 | a = 0
0 203 | b = 1

```

Status

OK

## Info Gain – Bagging(Training)

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose: Bagging -P 100 -S 1 -num-slots 1 -I 10 -V weka.classifiers.trees.REPTree -- -M 2 -V 0.001 -N 3 -S 1 -L 1 -I 0.0

Test options

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

(Nom) TARGET\_5Yrs

Start Stop

Result list (right-click for options)

```
16:18:30 - meta.Bagging
Bagging with 10 iterations and base learner
weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L 1 -I 0.0
Time taken to build model: 0.26 seconds
--- Classifier model (full training set) ---
Bagging with 10 iterations and base learner
weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L 1 -I 0.0
Time taken to build model: 0.26 seconds
--- Stratified cross-validation ---
--- Summary ---

Correctly Classified Instances      203          60.557 %
Incorrectly Classified Instances    132          39.403 %
Kappa statistic                   0.0064
Mean absolute error               0.479
Root mean squared error           0.489
Relative absolute error            100.052 %
Root relative squared error       100.0764 %
Total Number of Instances         335

--- Detailed Accuracy By Class ---

      TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
0       0.015   0.010     0.500   0.015     0.029   0.024   0.494     0.399     0
1       0.990   0.985     0.607   0.990     0.753   0.024   0.494     0.602     1
Weighted Avg.   0.606   0.601     0.565   0.606     0.468   0.024   0.494     0.522

--- Confusion Matrix ---

a   b   <- classified as
2 130 |   a = 0
2 201 |   b = 1
```

Status

OK

Type here to search

Log

4:18 PM 4/1/2018

## Info Gain – Bagging(test)

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose: Bagging -P 100 -S 1 -num-slots 1 -I 10 -V weka.classifiers.trees.REPTree -- -M 2 -V 0.001 -N 3 -S 1 -L 1 -I 0.0

Test options

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

(Nom) TARGET\_5Yrs

Start Stop

Result list (right-click for options)

```
22:30:37 - meta.Bagging
Bagging with 10 iterations and base learner
weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L 1 -I 0.0
Time taken to build model: 1.14 seconds
--- Evaluation on test set ---
Time taken to test model on supplied test set: 0.02 seconds
--- Summary ---

Correctly Classified Instances      203          60.557 %
Incorrectly Classified Instances    132          39.403 %
Kappa statistic                   0.0064
Mean absolute error               0.473
Root mean squared error           0.4923
Relative absolute error            99.8804 %
Root relative squared error       100.6658 %
Total Number of Instances         335

--- Detailed Accuracy By Class ---

      TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
0       0.015   0.010     0.500   0.015     0.029   0.024   0.398     0.398     0
1       0.990   0.985     0.607   0.990     0.753   0.024   0.508     0.615     1
Weighted Avg.   0.606   0.601     0.565   0.606     0.468   0.024   0.508     0.530

--- Confusion Matrix ---

a   b   <- classified as
2 130 |   a = 0
2 201 |   b = 1
```

Status

OK

Type here to search

Log

10:31 PM 4/1/2018

## Info Gain – AdaBoost(Training)

Weka Explorer

Classifier output

```

mean 0.5044 0.3914
std. dev. 0.4493 0.3951
weight sum 357.3161 647.6839
precision 0.1444 0.1444

T0V
mean 1.3258 1.7876
std. dev. 0.7548 1.0335
weight sum 357.3161 647.6839
precision 0.1051 0.1051

```

Weight: 0.47

Number of performed Iterations: 10

Time taken to build model: 0.32 seconds

\*\*\* Stratified cross-validation \*\*\*

\*\*\* Summary \*\*\*

	Correctly Classified Instances	646	64.2786 %
Incorrectly Classified Instances	359	35.7214 %	
Kappa statistic	0.255		
Mean absolute error	0.5294		
Root mean squared error	0.5315		
Relative absolute error	84.3209 %		
Root relative squared error	107.7189 %		
Total Number of Instances	1005		

\*\*\* Detailed Accuracy By Class \*\*\*

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0	0.472	0.255	0.527	0.472	0.496	0.223	0.436	0.492	0
1	0.745	0.528	0.702	0.745	0.723	0.223	0.636	0.728	1
Weighted Avg.	0.643	0.425	0.636	0.643	0.638	0.223	0.636	0.636	

\*\*\* Confusion Matrix \*\*\*

a b	a	b
178 199	178	199
160 469	160	469

Type here to search

## Info Gain – AdaBoost(Test)

Weka Explorer

Classifier output

```

| (numeric) FGM --> 5 (numeric) FGM
| (numeric) FGA --> 6 (numeric) FGA
| (numeric) FGt --> 7 (numeric) FGt
| (numeric) FHN --> 11 (numeric) FHN
| (numeric) STA --> 12 (numeric) STA
| (numeric) GReB --> 14 (numeric) GReB
| (numeric) GReB --> 15 (numeric) GReB
| (numeric) REB --> 16 (numeric) REB
| (numeric) AST --> 17 (numeric) AST
| (numeric) STL --> 18 (numeric) STL
| (numeric) BIM --> 19 (numeric) BIM
| (numeric) T0V --> 20 (numeric) T0V
| (nominal) TARGET_5Yrs --> 21 (nominal) TARGET_5Yrs

```

Time taken to build model: 0.28 seconds

\*\*\* Evaluation on test set \*\*\*

Time taken to test model on supplied test set: 0.05 seconds

\*\*\* Summary \*\*\*

	Correctly Classified Instances	234	69.8507 %
Incorrectly Classified Instances	101	30.1493 %	
Kappa statistic	0.3695		
Mean absolute error	0.356		
Root mean squared error	0.5137		
Relative absolute error	75.1176 %		
Root relative squared error	105.0424 %		
Total Number of Instances	335		

\*\*\* Detailed Accuracy By Class \*\*\*

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0	0.621	0.251	0.617	0.621	0.619	0.369	0.672	0.565	0
1	0.749	0.379	0.752	0.749	0.751	0.369	0.672	0.705	1
Weighted Avg.	0.699	0.329	0.699	0.699	0.699	0.369	0.672	0.649	

\*\*\* Confusion Matrix \*\*\*

a b	a	b
82 50	82	50
51 152	51	152

Type here to search

## OneR – Naïve Bayes(Training)

The screenshot shows the Weka Explorer interface with the NaiveBayes classifier selected. The 'Test options' section indicates 'Cross-validation Folds 10'. The 'Classifier output' pane displays detailed statistics for the 'BLK' and 'TOV' classes, including mean, std. dev., weight sum, and precision values. The 'Result list' pane shows the command used: '17:20\$ -bayes.NaiveBayes'. The status bar at the bottom right shows the date and time: 4/1/2018 5:12 PM.

## OneR – Naïve Bayes(Test)

The screenshot shows the Weka Explorer interface with the NaiveBayes classifier selected. The 'Test options' section indicates 'Supplied test set'. The 'Classifier output' pane displays detailed statistics for the 'BLK' and 'TOV' classes, including mean, std. dev., weight sum, and precision values. The 'Result list' pane shows the command used: '17:13:41 -bayes.NaiveBayes'. The status bar at the bottom right shows the date and time: 4/1/2018 5:13 PM.

## OneR – J48(Training)

Weka Explorer

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) TARGET\_SYrs

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

Result list (right-click for options)

17:14:43-trees-J48

```

AST
STL
BLK
TOV
TARGET_SYrs
Test mode: 10-fold cross-validation

--- Classifier model (full training set) ---

J48 pruned tree
-----
: 1 (1005.0/377.0)

Number of Leaves : 1
Size of the tree : 1

Time taken to build model: 0.04 seconds

--- Stratified cross-validation ---
--- Summary ---

Correctly Classified Instances 628 62.4076 %
Incorrectly Classified Instances 377 37.5124 %
Kappa statistic 0
Mean absolute error 0.4488
Root mean squared error 0.4842
Relative absolute error 95.9853 %
Root relative squared error 100 %
Total Number of Instances 1005

--- Detailed Accuracy By Class ---

      TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.000 0.000 ? 0.000 ? ? 0.496 0.373 0
1.000 1.000 0.625 1.000 0.769 ? 0.496 0.623 1
Weighted Avg. 0.625 0.625 ? 0.625 ? ? 0.496 0.529

--- Confusion Matrix ---

a b <- classified as
0 377 | a = 0
0 628 | b = 1

```

Status

OK

## OneR – J48(Test)

Weka Explorer

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) TARGET\_SYrs

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

Result list (right-click for options)

17:15:30-trees-J48

```

TOV
TARGET_SYrs
Test mode: user supplied test set: size unknown (reading incrementally)

--- Classifier model (full training set) ---

J48 pruned tree
-----
: 1 (1005.0/377.0)

Number of Leaves : 1
Size of the tree : 1

Time taken to build model: 0.02 seconds

--- Evaluation on test set ---

Time taken to test model on supplied test set: 0.02 seconds

--- Summary ---

Correctly Classified Instances 203 60.597 %
Incorrectly Classified Instances 132 39.403 %
Kappa statistic 0
Mean absolute error 0.4735
Root mean squared error 0.489
Relative absolute error 95.9889 %
Root relative squared error 100.0019 %
Total Number of Instances 335

--- 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 0.394 0
1.000 1.000 0.606 1.000 0.755 ? 0.500 0.606 1
Weighted Avg. 0.606 0.606 ? 0.606 ? ? 0.500 0.522

--- Confusion Matrix ---

a b <- classified as
0 132 | a = 0
0 203 | b = 1

```

Status

OK

## OneR – Bagging(Training)

The screenshot shows the Weka Explorer interface with the 'Classifier' tab selected. The 'Choose' dropdown is set to 'Bagging -P 100-S 1-num-slots 1-I 10-W weka.classifiers.trees.REPTree-- -M 2-V 0.001-N 3-S 1-L 1-I 0'. The 'Test options' panel shows 'Cross-validation Folds 10' selected. The 'Classifier output' panel displays the following text:

```

FIA
FT%
OREB
DREB
REB
AST
STL
BLK
TOV
TARGET_5Yrs
Test mode: 10-fold cross-validation

*** Classifier model (full training set) ***

Bagging with 10 iterations and base learner
weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L 1 -I 0

Time taken to build model: 1.37 seconds

*** Stratified cross-validation ***
*** Summary ***

Correctly Classified Instances      627      62.3881 %
Incorrectly Classified Instances   378      37.6119 %
Kappa statistic                      0.0085
Mean absolute error                  0.4702
Root mean squared error              0.4886
Relative absolute error               100.2706 %
Root relative squared error          100.5181 %
Total Number of Instances           1005

*** Detailed Accuracy By Class ***

    TP Rate   FP Rate  Precision  Recall   F-Measure  MCC   ROC Area  FPR Area  Class
    0.021     0.014    0.174    0.021    0.041    0.026   0.485    0.379    0
    0.986     0.979    0.627    0.986    0.766    0.026   0.485    0.617    1
Weighted Avg.       0.624     0.617    0.568    0.624    0.494    0.026   0.485    0.527

*** Confusion Matrix ***

    a | b <- classified as
    0 | 369 | 0
    9 | 619 | 1
  
```

## OneR – Bagging(Test)

The screenshot shows the Weka Explorer interface with the 'Classifier' tab selected. The 'Choose' dropdown is set to 'Bagging -P 100-S 1-num-slots 1-I 10-W weka.classifiers.trees.REPTree-- -M 2-V 0.001-N 3-S 1-L 1-I 0'. The 'Test options' panel shows 'Supplied test set' selected. The 'Classifier output' panel displays the following text:

```

DREB
REB
AST
STL
BLK
TOV
TARGET_5Yrs
Test mode: user supplied test set: size unknown (reading incrementally)

*** Classifier model (full training set) ***

Bagging with 10 iterations and base learner
weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L 1 -I 0

Time taken to build model: 1.03 seconds

*** Evaluation on test set ***

Time taken to test model on supplied test set: 0.02 seconds

*** Summary ***

Correctly Classified Instances      203      60.597 %
Incorrectly Classified Instances   132      39.403 %
Kappa statistic                      0.0064
Mean absolute error                  0.473
Root mean squared error              0.4823
Relative absolute error               100.4658 %
Root relative squared error          100.8804 %
Total Number of Instances           335

*** Detailed Accuracy By Class ***

    TP Rate   FP Rate  Precision  Recall   F-Measure  MCC   ROC Area  FPR Area  Class
    0.015     0.010    0.500    0.015    0.029    0.024   0.508    0.398    0
    0.990     0.985    0.607    0.990    0.753    0.024   0.508    0.615    1
Weighted Avg.       0.606     0.601    0.565    0.606    0.468    0.024   0.508    0.530

*** Confusion Matrix ***

    a | b <- classified as
    0 | 130 | 0
    2 | 201 | 1
  
```

## OneR – AdaBoost(Training)

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose AdaBoostM1-P 100-S 1-I 10-W weka.classifiers.bayes.NaiveBayes

Test options

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

(Nom) TARGET\_5Yrs

Start Stop

Result list (right-click for options)

```
18:20:32 - meta AdaBoostM1
18:20:32 - meta AdaBoostM1
```

Weight: 0.12

Number of performed Iterations: 10

Time taken to build model: 0.31 seconds

\*\*\* Stratified cross-validation \*\*\*

\*\*\* Summary \*\*\*

	Correctly Classified Instances	669	66.5672 %
Incorrectly Classified Instances	336	33.4328 %	
Kappa statistic	0.2784		
Mean absolute error	0.5297		
Root mean squared error	0.5127		
Relative absolute error	11.6124 %		
Root relative squared error	105.8966 %		
Total Number of Instances	1005		

\*\*\* Detailed Accuracy By Class \*\*\*

	IP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.525	0.250	0.558	0.525	0.541	0.279	0.452	0.511	0	0
0.750	0.475	0.725	0.750	0.737	0.279	0.652	0.732	1	1
Weighted Avg.	0.666	0.390	0.662	0.666	0.279	0.652	0.649		

\*\*\* Confusion Matrix \*\*\*

a	b	<- classified as
198	179	a = 0
157	471	b = 1

Status OK

## OneR – AdaBoost(Test)

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose AdaBoostM1-P 100-S 1-I 10-W weka.classifiers.bayes.NaiveBayes

Test options

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

(Nom) TARGET\_5Yrs

Start Stop

Result list (right-click for options)

```
18:20:32 - meta AdaBoostM1
18:21:21 - meta AdaBoostM1
```

Weight: 0.12

Number of performed Iterations: 10

Time taken to build model: 0.3 seconds

\*\*\* Evaluation on test set \*\*\*

Time taken to test model on supplied test set: 0.06 seconds

\*\*\* Summary \*\*\*

	Correctly Classified Instances	201	65.9701 %
Incorrectly Classified Instances	114	34.0299 %	
Kappa statistic	0.2759		
Mean absolute error	0.3935		
Root mean squared error	0.5132		
Relative absolute error	83.0962 %		
Root relative squared error	104.9573 %		
Total Number of Instances	335		

\*\*\* Detailed Accuracy By Class \*\*\*

	IP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.523	0.251	0.575	0.523	0.548	0.277	0.637	0.534	0	0
0.749	0.477	0.707	0.749	0.727	0.277	0.637	0.687	1	1
Weighted Avg.	0.660	0.388	0.655	0.660	0.277	0.637	0.627		

\*\*\* Confusion Matrix \*\*\*

a	b	<- classified as
69	63	a = 0
51	152	b = 1

Status OK

## Manual – Naïve Bayes(Training)

The screenshot shows the Weka Explorer interface with the NaiveBayes classifier selected. The 'Test options' section is set to 'Cross-validation Folds 10'. The 'Classifier output' pane displays statistical results for three classes: BLK, FGK, and TOV. The 'Result list' pane shows the command used to run the classifier.

```

std. dev.      1.123 1.5897
weight sum     377   628
precision     0.1493 0.1493

BLK
mean          0.2455 0.451
std. dev.     0.2943 0.4017
weight sum    377   628
precision     0.1444 0.1444

TOV
mean          0.957 1.3472
std. dev.     0.5543 0.7865
weight sum    377   628
precision     0.1051 0.1051

Time taken to build model: 0.01 seconds

*** Stratified cross-validation ***
*** Summary ***

Correctly Classified Instances      633           62.9551 %
Incorrectly Classified Instances   372           37.0149 %
Kappa statistic                   0.2996
Mean absolute error               0.1752
Root mean squared error           0.1846
Relative absolute error           80.0271 %
Root relative squared error      120.7241 %
Total Number of Instances        1005

*** Detailed Accuracy By Class ***

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  FRC Area  Class
      0.828   0.487   0.504   0.825   0.436   0.336   0.734   0.603   1
      0.513   0.175   0.830   0.513   0.634   0.336   0.734   0.808   0
Weighted Avg.   0.630   0.292   0.708   0.650   0.631   0.336   0.734   0.731   1

*** Confusion Matrix ***

a b <- classified as
311 66 | a = 0
306 322 | b = 1

```

## Manual – Naïve Bayes(Test)

The screenshot shows the Weka Explorer interface with the NaiveBayes classifier selected. The 'Test options' section is set to 'Supplied test set'. The 'Classifier output' pane displays detailed classification rules for each class. The 'Result list' pane shows the command used to run the classifier.

```

| (numeric) PTS      --> 4 (numeric) PTS
| (numeric) FGK      --> 5 (numeric) FGK
| (numeric) FGA      --> 6 (numeric) FGA
| (numeric) FG4      --> 7 (numeric) FG4
| (numeric) FN       --> 10 (numeric) FN
| (numeric) FTA      --> 12 (numeric) FTA
| (numeric) OREB      --> 14 (numeric) OREB
| (numeric) DRKB      --> 15 (numeric) DRKB
| (numeric) REB       --> 16 (numeric) REB
| (numeric) AST       --> 17 (numeric) AST
| (numeric) BLK       --> 19 (numeric) BLK
| (numeric) TOV       --> 20 (numeric) TOV
(nominal) TARGET_5Yrs --> 21 (nominal) TARGET_5Yrs

Time taken to build model: 0.02 seconds

*** Evaluation on test set ***

Time taken to test model on supplied test set: 0.54 seconds

*** Summary ***

Correctly Classified Instances      209           62.3881 %
Incorrectly Classified Instances   126           37.6119 %
Kappa statistic                   0.2866
Mean absolute error               0.1760
Root mean squared error           0.5886
Relative absolute error           75.554 %
Root relative squared error      120.4213 %
Total Number of Instances        335

*** Detailed Accuracy By Class ***

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  FRC Area  Class
      0.833   0.512   0.514   0.833   0.436   0.327   0.751   0.660   0
      0.498   0.167   0.818   0.498   0.611   0.327   0.751   0.825   1
Weighted Avg.   0.624   0.303   0.698   0.624   0.621   0.327   0.751   0.760

*** Confusion Matrix ***

a b <- classified as
110 22 | a = 0
104 99 | b = 1

```

## Manual – J48(Training)

Weka Explorer

Classifier

Choose: J48 - C 0.25 - M 2

**Test options**

- Use training set
- Supplied test set
- Cross-validation Folds: 10
- Percentage split %: 66
- 

**Classifier output**

```

REB
AST
BLK
TOV
TARGET_5Yrs
Test mode: 10-fold cross-validation

*** Classifier model (full training set) ***

J48 pruned tree
=====
: 1 (1005./377.0)

Number of Leaves : 1
Size of the tree : 1

Time taken to build model: 0.03 seconds

*** Stratified cross-validation ***
*** Summary ***

Correctly Classified Instances      628          62.4876 %
Incorrectly Classified Instances   377          37.5124 %
Kappa statistic                   0
Mean absolute error               0.4486
Root mean squared error           0.4442
Relative absolute error            99.9853 %
Root relative squared error       100 %
Total Number of Instances         1005

*** Detailed Accuracy By Class ***

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Class
0.000    0.000     ?        0.000    ?        ?        0.496    0.373    0
1.000    1.000     0.625   1.000    0.769    ?        0.496    0.623    1
Weighted Avg. 0.625  0.625     ?        0.625    ?        ?        0.496    0.529

*** Confusion Matrix ***

a b <- classified as
0 377 | a = 0
0 628 | b = 1

```

Status

OK

## Manual – J48(Test)

Weka Explorer

Classifier

Choose: J48 - C 0.25 - M 2

**Test options**

- Use training set
- Supplied test set
- Cross-validation Folds: 10
- Percentage split %: 66
- 

**Classifier output**

```

(F) PTS      --> 4 (numeric) PTS
(numeric) FGM    --> 5 (numeric) FGM
(numeric) TSL    --> 6 (numeric) TSL
(numeric) FGA    --> 7 (numeric) FGA
(numeric) FTM    --> 11 (numeric) FTM
(numeric) FTA    --> 12 (numeric) FTA
(numeric) OREB   --> 14 (numeric) OREB
(numeric) DREB   --> 15 (numeric) DREB
(numeric) REB    --> 16 (numeric) REB
(numeric) AST    --> 17 (numeric) AST
(numeric) BLK    --> 19 (numeric) BLK
(numeric) TOV    --> 20 (numeric) TOV
(nominal) TARGET_5Yrs --> 21 (nominal) TARGET_5Yrs

Time taken to build model: 0.01 seconds
Evaluation on test set
Time taken to test model on supplied test set: 10.03 seconds
Summary

Correctly Classified Instances      203          60.597 %
Incorrectly Classified Instances   132          39.403 %
Kappa statistic                   0
Mean absolute error               0.4735
Root mean squared error           0.4859
Relative absolute error            99.9889 %
Root relative squared error       100.0019 %
Total Number of Instances         335

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    0.394    0
1.000    1.000     0.606   1.000    0.755    ?        0.500    0.606    1
Weighted Avg. 0.606  0.606     ?        0.606    ?        ?        0.500    0.522

Confusion Matrix

a b <- classified as
0 132 | a = 0
0 203 | b = 1

```

Status

OK

## Manual – Bagging(Training)

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

**Classifier**

Choose: Bagging -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.REPTree -- -M 2 -V 0.001 -N 3 -S 1 -L 1 -I 0.0

**Test options**

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

(Nom) TARGET\_5Yrs

Start Stop

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

17.46.30 - misc.Bagging

```

Bagging with 10 iterations and base learner
weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L 1 -I 0.0

Time taken to build model: 0.87 seconds

*** Stratified cross-validation ***
*** Summary **

Correctly Classified Instances      637           62.3881 %
Incorrectly Classified Instances    375           37.6119 %
Kappa statistic                      0.0085
Mean absolute error                  0.4702
Root mean squared error              0.4986
Relative absolute error              100.2709 %
Root relative squared error         100.9181 %
Total Number of Instances           1005

*** Detailed Accuracy By Class ***

      TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   FRC Area   Class
  0.021     0.014     0.471     0.021     0.041     0.026     0.495     0.377     0
  0.986     0.979     0.627     0.986     0.766     0.026     0.495     0.617     1
Weighted Avg.     0.424     0.617     0.568     0.624     0.494     0.026     0.495     0.527

*** Confusion Matrix ***

  a   b   <- classified as
  8 369 |   a = 0
  9 419 |   b = 1

```

**Status**

OK

## Manual – Bagging(Test)

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

**Classifier**

Choose: Bagging -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.REPTree -- -M 2 -V 0.001 -N 3 -S 1 -L 1 -I 0.0

**Test options**

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

(Nom) TARGET\_5Yrs

Start Stop

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

17.46.33 - misc.InputMappedClassifier

```

Time taken to build model: 0.79 seconds

*** Evaluation on test set **

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

*** Summary **

Correctly Classified Instances      203           60.597 %
Incorrectly Classified Instances    132           39.403 %
Kappa statistic                      0.0064
Mean absolute error                  0.4713
Root mean squared error              0.4923
Relative absolute error              99.8804 %
Root relative squared error         100.6668 %
Total Number of Instances           335

*** Detailed Accuracy By Class ***

      TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   FRC Area   Class
  0.015     0.010     0.500     0.015     0.029     0.024     0.508     0.398     0
  0.990     0.985     0.607     0.990     0.753     0.024     0.508     0.615     1
Weighted Avg.     0.406     0.401     0.565     0.606     0.465     0.024     0.505     0.530

*** Confusion Matrix ***

  a   b   <- classified as
  2 130 |   a = 0
  2 201 |   b = 1

```

**Status**

OK

## Manual – AdaBoost(Training)

Weka Explorer

Preprocess Classify Cluster Associate Selected attributes Visualize

Classifier

Choose AdaBoostM1-P 100-S 1-I 10-W weka.classifiers.bayes.NaiveBayes

Test options

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

(Nom) TARGET\_5Yrs

Start Stop

Result list (right-click for options)

18:22:38 - metaAdaBoostM1

```

Weight: 1.13
Number of performed Iterations: 10
Time taken to build model: 0.24 seconds
--- Stratified cross-validation ---
--- Summary ---

Correctly Classified Instances      654      65.0744 %
Incorrectly Classified Instances   351      34.9254 %
Kappa statistic                   0.2716
Mean absolute error               0.3928
Root mean squared error           0.5100
Relative absolute error            83.7887 %
Root relative squared error       105.4406 %
Total Number of Instances         1005

--- Detailed Accuracy By Class ---

    TP Rate  FP Rate  Precision  Recall   F-Measure  MCC   ROC Area  PRC Area  Class
0       0.598   0.314     0.531   0.592     0.560   0.273   0.448   0.514   0
1       0.696   0.408     0.737   0.696     0.711   0.273   0.448   0.756   1

Weighted Avg.      0.651   0.373     0.660   0.651     0.654   0.273   0.448   0.653

--- Confusion Matrix ---

    a   b <- Classified as
223 154 |  a = 0
197 431 |  b = 1

```

Status

OK

Type here to search

Log x 0

6:23 PM 4/1/2018

## Manual – AdaBoost(Test)

Weka Explorer

Preprocess Classify Cluster Associate Selected attributes Visualize

Classifier

Choose AdaBoostM1-P 100-S 1-I 10-W weka.classifiers.bayes.NaiveBayes

Test options

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

(Nom) TARGET\_5Yrs

Start Stop

Result list (right-click for options)

18:23:59 - misc.InputMappedClassifier

```

Time taken to build model: 0.27 seconds
--- Evaluation on test set ---

Time taken to test model on supplied test set: 9.04 seconds

--- Summary ---

Correctly Classified Instances      222      66.2687 %
Incorrectly Classified Instances   113      33.7313 %
Kappa statistic                   0.2572
Mean absolute error               0.3771
Root mean squared error           0.5108
Relative absolute error            79.4152 %
Root relative squared error       104.4605 %
Total Number of Instances         335

--- Detailed Accuracy By Class ---

    TP Rate  FP Rate  Precision  Recall   F-Measure  MCC   ROC Area  PRC Area  Class
0       0.412   0.157     0.600   0.452     0.502   0.265   0.450   0.538   0
1       0.813   0.568     0.688   0.813     0.745   0.265   0.450   0.708   1

Weighted Avg.      0.663   0.418     0.653   0.663     0.649   0.265   0.450   0.641

--- Confusion Matrix ---

    a   b <- classified as
57 75 |  a = 0
38 165 |  b = 1

```

Status

OK

Type here to search

Log x 0

6:24 PM 4/1/2018

## Performance Comparison

Comparing all the classifiers with respective attribute selection algorithms, AdaBoost classifier with selection algorithms Gain Ratio and Info Ratio, yielded better with 69.85% accuracy while most of the four classifiers did poorly with J48.

### TP Rate

Is also called sensitivity or True Positive Rate, it is the proportion that tested positive to, we observe that TP Rate for Info Gain, Adaboost performed highest with 0.69.

### FP Rate

Is the probability of falsely rejecting the null hypothesis for a particular test. In the classification done here, we see that it is higher for Bagging over the rest, indicating that in this classifier the probability of falsely rejecting is more and thus a disadvantage over the rest.

### ROC Curve

It is a curve where true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points of a parameter, and higher the value or closer to 1, the better it is. In the classifications done we see that Naïve Bayes yields 0.74 being the highest and AdaBoost recording 0.69 and J48 with 0.5 being least.

Based on the observations, AdaBoostM1 is the best-fit classifier for this dataset as it yields better TP Rate, FP Rate and Accuracy compared to other models.

## Project Responsibility and Learning

In the course of this project, Venkata Vishnuvardhan Aluri and Aakash T.C split the work based on each phase of the project. Vishnu Aluri does classifier modelling and Dataset retrieval, while Aakash performed pre-processing. Documentation was performed together. There are several major takeaways from the project. Mainly, we get to realize the scope of Data Mining in solving various real world problems. In Addition to that, we got the hands on experience with WEKA and applying various machine-learning algorithms learnt.

## References

The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011

