Predicting Whether A Car Will Be A "Kick" To Dealerships

Arnav Jain, Vishal Kotha

Final Report

10/11/2022

Table of Contents

Statement/Project Goal

The goal is to be able to predict whether a used car will be a bad buy, and have a serious car issue, such as tampered odometers, mechanical issues the dealer is not able to address, issues with getting the vehicle title from the seller, or some other unforeseen problem. These problems are called “kicks.” The two classes are the car being a bad buy or a kick and it being a normal, issue-free car. The goal of this project is to develop a model to accurately predict whether a car would be a kick, which would aid dealerships in reselling used cars, as well as consumers looking to buy a good car.

Description of Dataset

The datasets contain 72,983 tuples, each containing 32 feature attributes and a class attribute. Each tuple represents a car. The dataset can be found here https://drive.google.com/drive/folders/1ipivR8s1StmyYp0MS5I6AZtGU1tga2pF?usp=sharing. This contains the overall dataset, as well as the intermediate datasets, such as the sampled dataset, and each of the attribute selected datasets. The following list contains a list of the attributes.

- **IsBadBuy(Class Attribute)**
  - Whether the car would become a “kick”/bad buy
- PurchDate
  - When the car was purchased
- Auction
  - Auction at which the vehicle was purchased
- VehYear:
  - The manufacturer's year of the vehicle
- VehicleAge
  - Years since the manufacturer's year
- Make

    - Vehicle Manufacturer
- Model
- Trim
    - Vehicle Trim Level
- SubModel
- Color
- Transmission
    - Vehicles transmission type (Automatic, Manual)
- WheelTypeID
    - The id of the vehicle wheel
- WheelType
    - The vehicle wheel type description (Alloy, Covers)
- VehOdo
    - Vehicles mileage
- Nationality
    - Where the car was made
- Size
    - The size category of the vehicle (Compact, SUV, etc.)
- TopThreeAmericanName
    - Identifies if the manufacturer is one of the top three American manufacturers
- MMRAcquisitionAuctionAveragePrice, MMRAcquisitionAuctionCleanPrice, MMRAcquisitionRetailAveragePrice, MMRAcquisitonRetailCleanPrice, MMRCurrentAuctionAveragePrice, MMRCurrentAuctionCleanPrice, MMRCurrentRetailAveragePrice, MMRCurrentRetailCleanPrice:
    - Price to acquire depending on series of conditions
    - Current - price in current day, Acquisition - price at date of purchase
    - Auction - price in auction, Retail - price in retail
    - Average - average condition price, Clean - above average condition price, Auction - current car condition price
- PRIMEUNIT
    - Is vehicle in higher demand than normal cars
- AcquisitionType
    - How the car was acquired - Auction, Retail buy, etc.
- AUCGUART
    - The guarantee by the manufacturer on the car
- KickDate
    - Date the vehicle was kicked to the auction
- BYRNO
    - Number assigned to the buyer
- VNZIP
    - Zip Code of car purchase
- VNST

    - State of car purchase
- VehBCost
    - Cost to acquire the vehicle
- IsOnlineSale
    - Was the vehicle purchased online
- WarrantyCost
    - Cost of warranty

Pre-processing

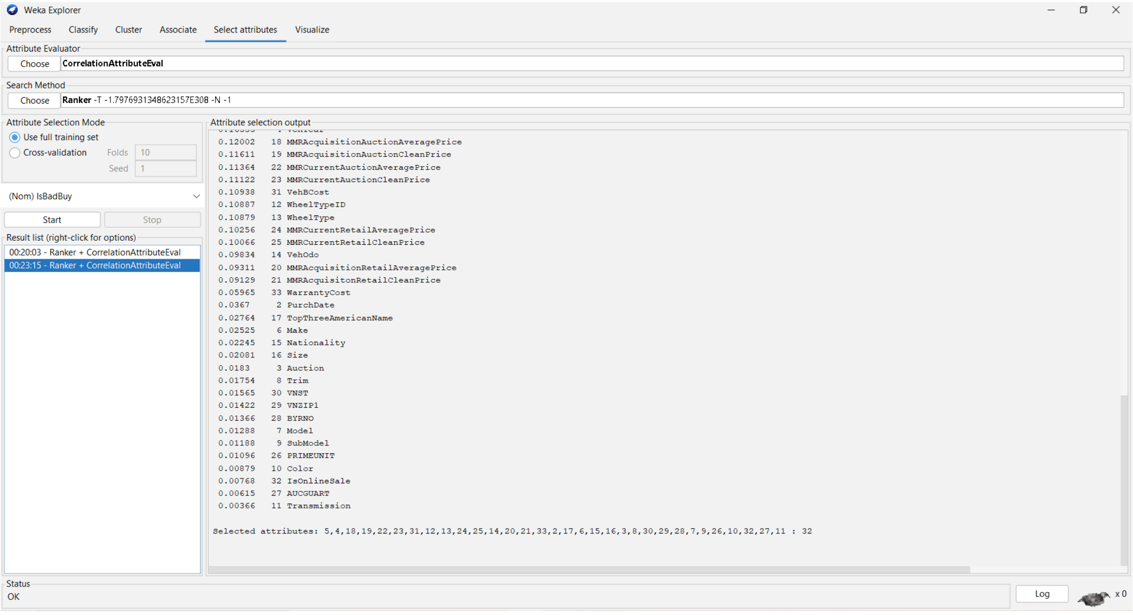
Before we could create and train our models, we had to pre-process data: this included things like sampling our data and splitting the data into a train and test set. We first filled the missing values in the original data using the ReplaceMissingValues function in WEKA. Its methods of imputing are mean for numerical variables and mode for categorical variables. Our original dataset contained 72,893 instances: we attempted to train models with this much data and it took too long. We first did a stratified random sample without resampling, taking only 20% of the original data. Then, we split this sample into train and test sets. The training set contains 11,677 instances (10,241 of class label 0, meaning not a "kick"; 1,436 of class label 1, meaning a "kick"). The test set contains 2,919 instances (2,560 instances of class label 0, meaning not a "kick"; 359 of class label 1, meaning a "kick"). The original dataset, containing 72,983 instances, had 64,007 instances with a class label of 0. This is about 87.701%. The percentage of class label 0 to total dataset of both the training and testing dataset are 87.702% and 87.701% respectively, meaning the datasets are representative of the original dataset.

Attribute Selectors

In order to save time and prevent the curse of dimensionality, we had to use attribute selection to identify which attributes are most helpful for creating a model. We tested four different attribute selection algorithms: correlation attribute selection, reliefF attribute selection, gain ratio attribute selection, and information gain attribute selection. We also tested a baseline with no attribute selection. We list how the four attribute selection works, and the selected attributes below.

Correlation Attribute Selection *- Calculates the Pearson correlation between each attribute and the class label. A threshold, in our case 0.09, is used to select the features.*
reliefF Attribute Selection *- Randomly selected instance, finds the nearest hit (most similar instance with same class label) and nearest miss (most similar instance with other class label) and updates each attribute based on corresponding distance to the hit and miss. We used a threshold of 0.05 to select the features.*
Information Gain Attribute Selection *- Entropy is a measure of how uniform the class labels are. Information Gain is the original entropy minus the weighted sum of the sub-entropies generated by splitting on one of the predictor variables. Repeat for all predictor variables. We used a threshold of 0.007 to select the features.*

***Gain Ratio Attribute Selection** - Since information gain takes a predictor variable, splits it into a bunch of branches, and sums up all the entropies, variables that have more categories will be weighted higher. Gain ratio lessens this bias with a normalizing term called the Intrinsic Information. We used a threshold of 0.005 to select the features.*



Attribute Evaluator: **CorrelationAttributeEval**

Search Method: **Ranker** -T -1.7976931348623157E308 -N -1

Attribute Selection Mode: ● Use full training set ○ Cross-validation Folds 10 Seed 1

(Nom) IsBadBuy

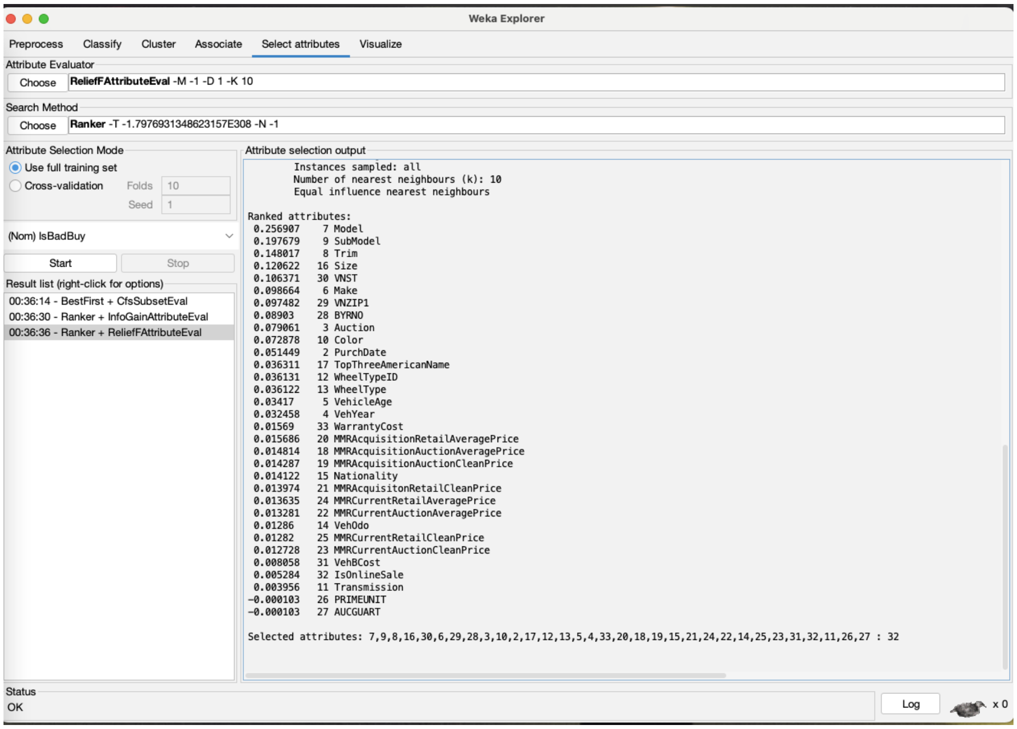
Result list (right-click for options)
00:20:03 - Ranker + CorrelationAttributeEval
00:23:15 - Ranker + CorrelationAttributeEval

Attribute selection output

```
 0.12002   18 MMRAcquisitionAuctionAveragePrice
 0.11611   19 MMRAcquisitionAuctionCleanPrice
 0.11364   22 MMRCurrentAuctionAveragePrice
 0.11122   23 MMRCurrentAuctionCleanPrice
 0.10938   31 VehBCost
 0.10887   12 WheelTypeID
 0.10879   13 WheelType
 0.10256   24 MMRCurrentRetailAveragePrice
 0.10066   25 MMRCurrentRetailCleanPrice
 0.09834   14 VehOdo
 0.09311   20 MMRAcquisitionRetailAveragePrice
 0.09129   21 MMRAcquisitonRetailCleanPrice
 0.05965   33 WarrantyCost
 0.0367     2 PurchDate
 0.02764   17 TopThreeAmericanName
 0.02525    6 Make
 0.02245   15 Nationality
 0.02081   16 Size
 0.0183     3 Auction
 0.01754    8 Trim
 0.01565   30 VNST
 0.01422   29 VNZIP1
 0.01366   28 BYRNO
 0.01288    7 Model
 0.01188    9 SubModel
 0.01096   26 PRIMEUNIT
 0.00879   10 Color
 0.00768   32 IsOnlineSale
 0.00615   27 AUCGUART
 0.00366   11 Transmission

Selected attributes: 5,4,18,19,22,23,31,12,13,24,25,14,20,21,33,2,17,6,15,16,3,8,30,29,28,7,9,26,10,32,27,11 : 32
```

Status
OK

Figure 1: Correlation Attribute Selection



Attribute Evaluator: **ReliefFAttributeEval** -M -1 -D 1 -K 10

Search Method: **Ranker** -T -1.7976931348623157E308 -N -1

Attribute Selection Mode: ● Use full training set ○ Cross-validation Folds 10 Seed 1

(Nom) IsBadBuy

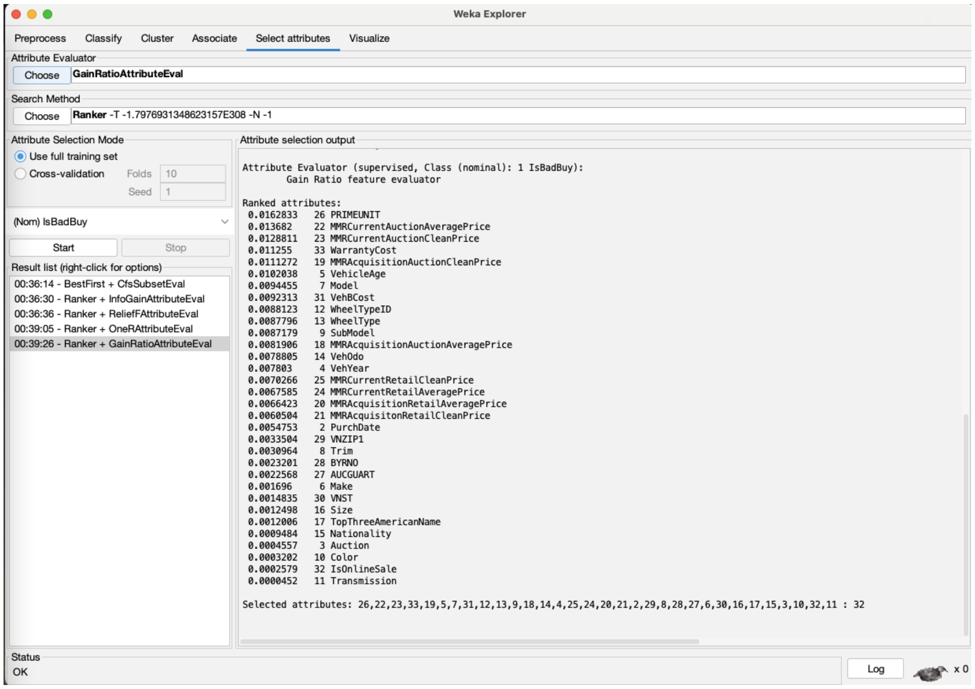
Result list (right-click for options)
00:36:14 - BestFirst + CfsSubsetEval
00:36:30 - Ranker + InfoGainAttributeEval
00:36:36 - Ranker + ReliefFAttributeEval

Attribute selection output

```
        Instances sampled: all
        Number of nearest neighbours (k): 10
        Equal influence nearest neighbours

Ranked attributes:
 0.256907    7 Model
 0.197679    9 SubModel
 0.148017    8 Trim
 0.120622   16 Size
 0.106371   30 VNST
 0.098664    6 Make
 0.097482   29 VNZIP1
 0.08903    28 BYRNO
 0.079061    3 Auction
 0.072878   10 Color
 0.051449    2 PurchDate
 0.036311   17 TopThreeAmericanName
 0.036131   12 WheelTypeID
 0.036122   13 WheelType
 0.03417     5 VehicleAge
 0.032458    4 VehYear
 0.01569    33 WarrantyCost
 0.015686   20 MMRAcquisitionRetailAveragePrice
 0.014814   18 MMRAcquisitionAuctionAveragePrice
 0.014287   19 MMRAcquisitionAuctionCleanPrice
 0.014122   15 Nationality
 0.013974   21 MMRAcquisitonRetailCleanPrice
 0.013635   24 MMRCurrentRetailAveragePrice
 0.013281   22 MMRCurrentAuctionAveragePrice
 0.01286    14 VehOdo
 0.01282    25 MMRCurrentRetailCleanPrice
 0.012728   23 MMRCurrentAuctionCleanPrice
 0.008058   31 VehBCost
 0.005284   32 IsOnlineSale
 0.003956   11 Transmission
-0.000103   26 PRIMEUNIT
-0.000103   27 AUCGUART

Selected attributes: 7,9,8,16,30,6,29,28,3,10,2,17,12,13,5,4,33,20,18,19,15,21,24,22,14,25,23,31,32,11,26,27 : 32
```

Status
OK

Figure 2: ReliefF Attribute Selection



Attribute Evaluator: GainRatioAttributeEval

Search Method: Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode
● Use full training set
○ Cross-validation Folds 10 Seed 1

(Nom) IsBadBuy

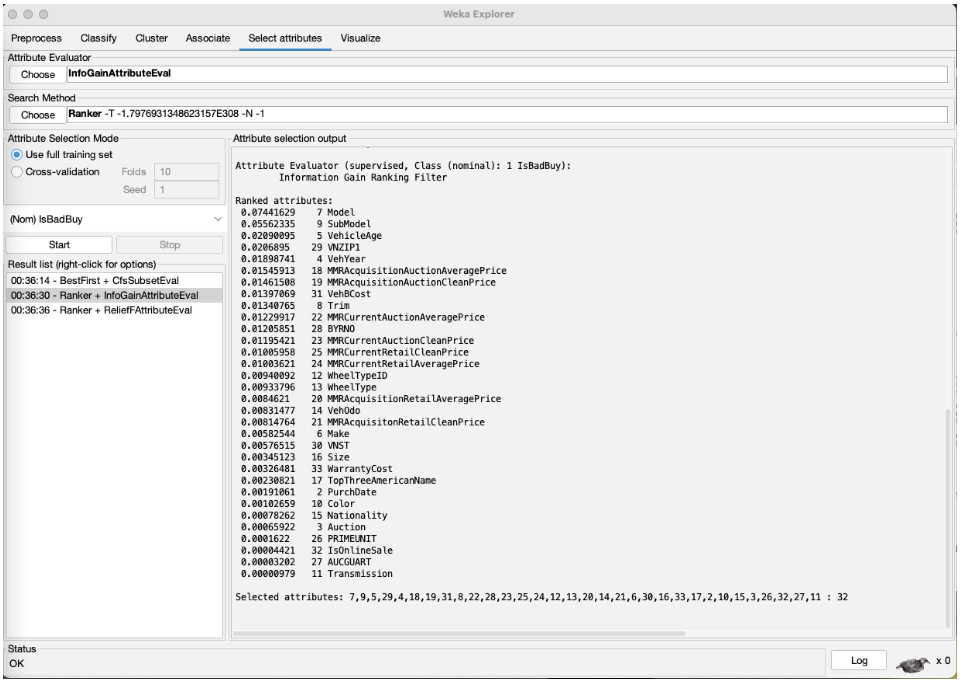
Result list (right-click for options)
00:36:14 - BestFirst + CfsSubsetEval
00:36:30 - Ranker + InfoGainAttributeEval
00:36:36 - Ranker + ReliefFAttributeEval
00:39:05 - Ranker + OneRAttributeEval
00:39:26 - Ranker + GainRatioAttributeEval

Attribute selection output

```
Attribute Evaluator (supervised, Class (nominal): 1 IsBadBuy):
        Gain Ratio feature evaluator

Ranked attributes:
 0.0162833   26 PRIMEUNIT
 0.013682    22 MMRCurrentAuctionAveragePrice
 0.0128811   23 MMRCurrentAuctionCleanPrice
 0.011255    33 WarrantyCost
 0.0111272   19 MMRAcquisitionAuctionCleanPrice
 0.0102038    5 VehicleAge
 0.0094455    7 Model
 0.0092313   31 VehBCost
 0.0088123   12 WheelTypeID
 0.0087796   13 WheelType
 0.0087179    9 SubModel
 0.0081906   18 MMRAcquisitionAuctionAveragePrice
 0.0078805   14 VehOdo
 0.007803     4 VehYear
 0.0070266   25 MMRCurrentRetailCleanPrice
 0.0067585   24 MMRCurrentRetailAveragePrice
 0.0066423   20 MMRAcquisitionRetailAveragePrice
 0.0060504   21 MMRAcquisitonRetailCleanPrice
 0.0054753    2 PurchDate
 0.0033504   29 VNZIP1
 0.0030964    8 Trim
 0.0023201   28 BYRNO
 0.0022568   27 AUCGUART
 0.001696     6 Make
 0.0014835   30 VNST
 0.0012498   16 Size
 0.0012006   17 TopThreeAmericanName
 0.0009484   15 Nationality
 0.0004557    3 Auction
 0.0003202   10 Color
 0.0002579   32 IsOnlineSale
 0.0000452   11 Transmission

Selected attributes: 26,22,23,33,19,5,7,31,12,13,9,18,14,4,25,24,20,21,2,29,8,28,27,6,30,16,17,15,3,10,32,11 : 32
```

Status
OK

Figure 3: Gain Ratio Attribute Selection



Attribute Evaluator: InfoGainAttributeEval

Search Method: Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode
● Use full training set
○ Cross-validation Folds 10 Seed 1

(Nom) IsBadBuy

Result list (right-click for options)
00:36:14 - BestFirst + CfsSubsetEval
00:36:30 - Ranker + InfoGainAttributeEval
00:36:36 - Ranker + ReliefFAttributeEval

Attribute selection output

```
Attribute Evaluator (supervised, Class (nominal): 1 IsBadBuy):
        Information Gain Ranking Filter

Ranked attributes:
 0.07441629    7 Model
 0.05562335    9 SubModel
 0.02090095    5 VehicleAge
 0.0206895    29 VNZIP1
 0.01898741    4 VehYear
 0.01545913   18 MMRAcquisitionAuctionAveragePrice
 0.01461508   19 MMRAcquisitionAuctionCleanPrice
 0.01397069   31 VehBCost
 0.01340765    8 Trim
 0.01229917   22 MMRCurrentAuctionAveragePrice
 0.01205851   28 BYRNO
 0.01195421   23 MMRCurrentAuctionCleanPrice
 0.01005958   25 MMRCurrentRetailCleanPrice
 0.01003621   24 MMRCurrentRetailAveragePrice
 0.00940092   12 WheelTypeID
 0.00933796   13 WheelType
 0.0084621    20 MMRAcquisitionRetailAveragePrice
 0.00831477   14 VehOdo
 0.00814764   21 MMRAcquisitonRetailCleanPrice
 0.00582544    6 Make
 0.00576515   30 VNST
 0.00345123   16 Size
 0.00326481   33 WarrantyCost
 0.00230821   17 TopThreeAmericanName
 0.00191061    2 PurchDate
 0.00102659   10 Color
 0.00078262   15 Nationality
 0.00065922    3 Auction
 0.0001622    26 PRIMEUNIT
 0.00004421   32 IsOnlineSale
 0.00003202   27 AUCGUART
 0.00000979   11 Transmission

Selected attributes: 7,9,5,29,4,18,19,31,8,22,28,23,25,24,12,13,20,14,21,6,30,16,33,17,2,10,15,3,26,32,27,11 : 32
```

Status
OK

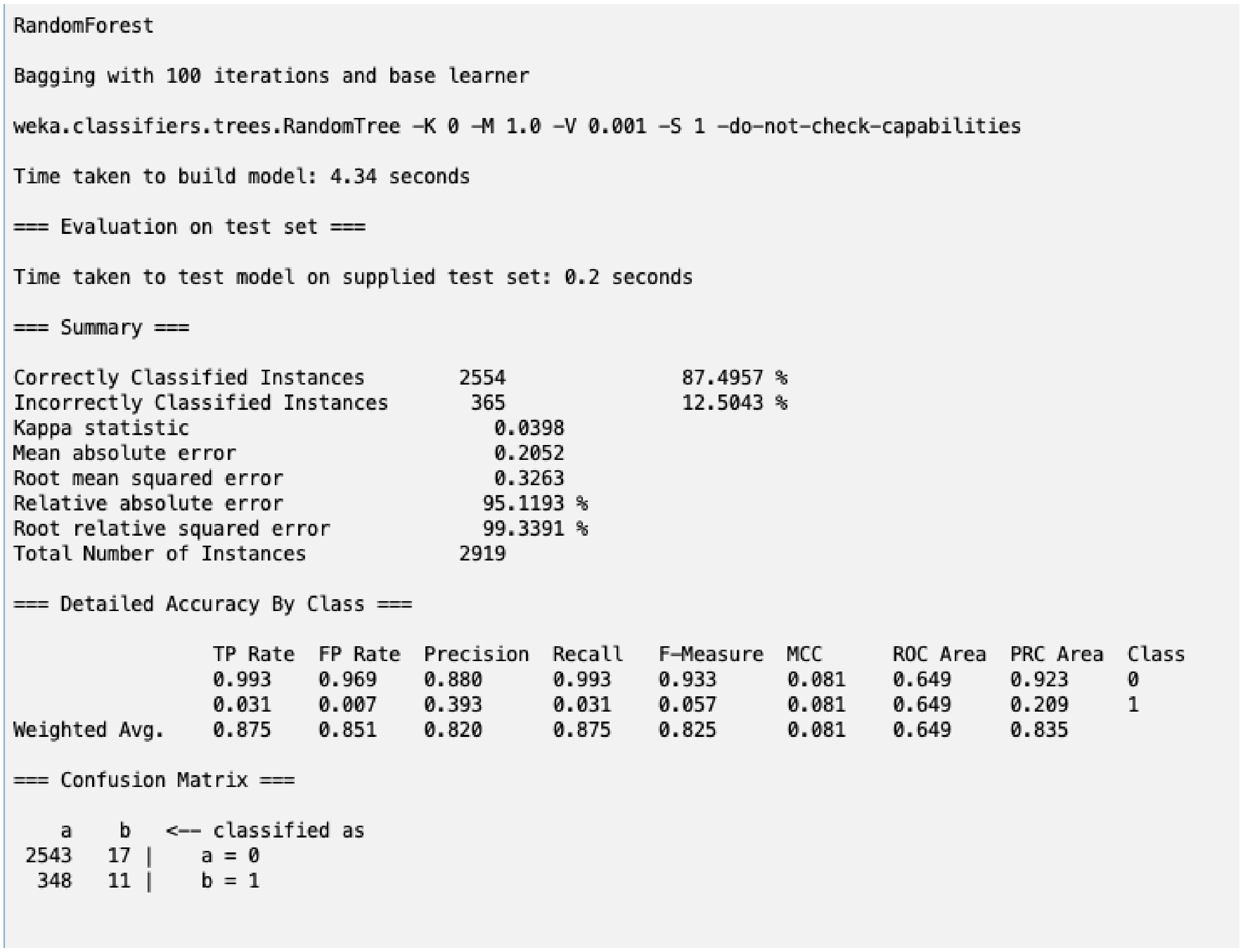
Figure 4: Information Gain Attribute Selection

Models/Results

We tested four different models: Random Forest, Naive Bayes, J48 and Adaboost M1. Below is a brief description of how each model works.

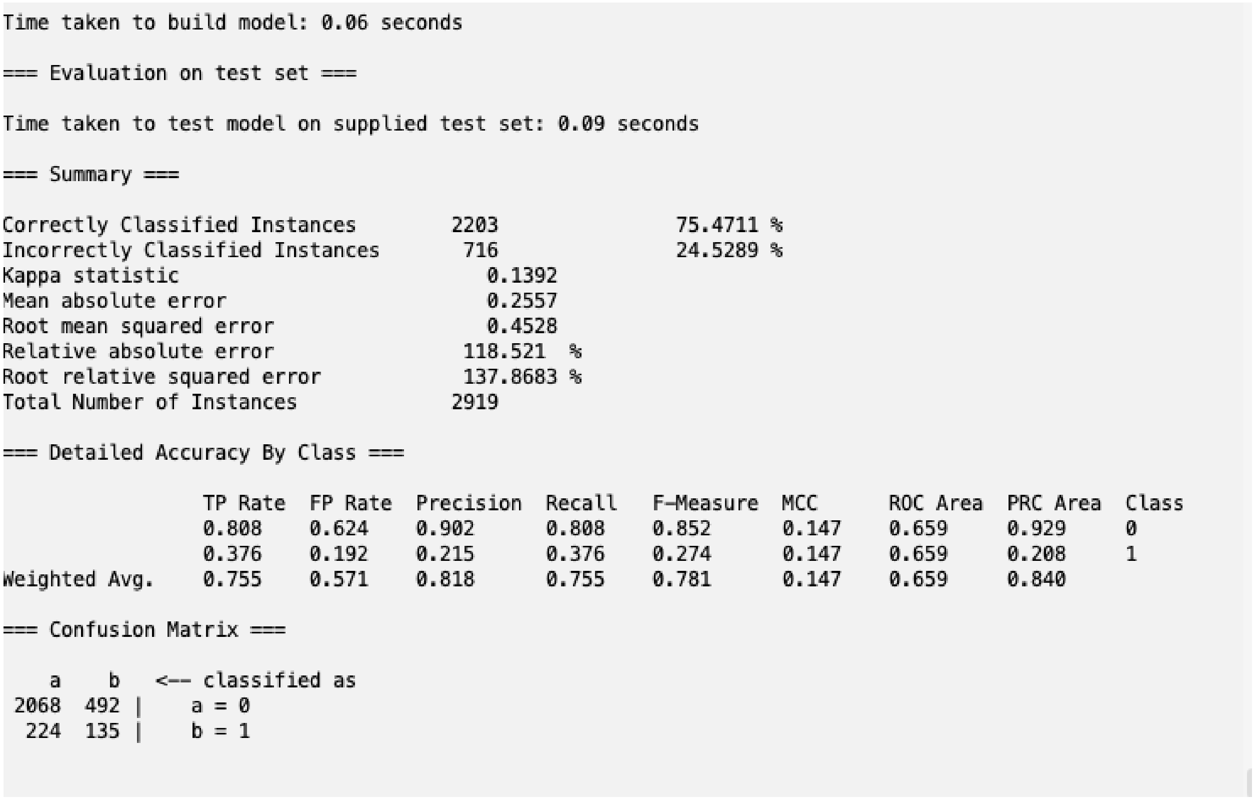
***Random Forest** - A random forest creates a bunch of decision trees. It picks the majority output of all these decision trees as its final classification (an ensemble learner).*
***Naive Bayes** - A probabilistic machine learning model that predicts the class. It's naive because it assumes that the predictors/features are independent.*
***J48** - A tree based classification method that utilizes a top down approach.*
***Adaboost M1** - A boosted decision tree classifier, specifically oriented for binary classification problems.*

Below are the screenshots outlining the results of the four models with the four attribute selection algorithms, along with one without attribute selection.



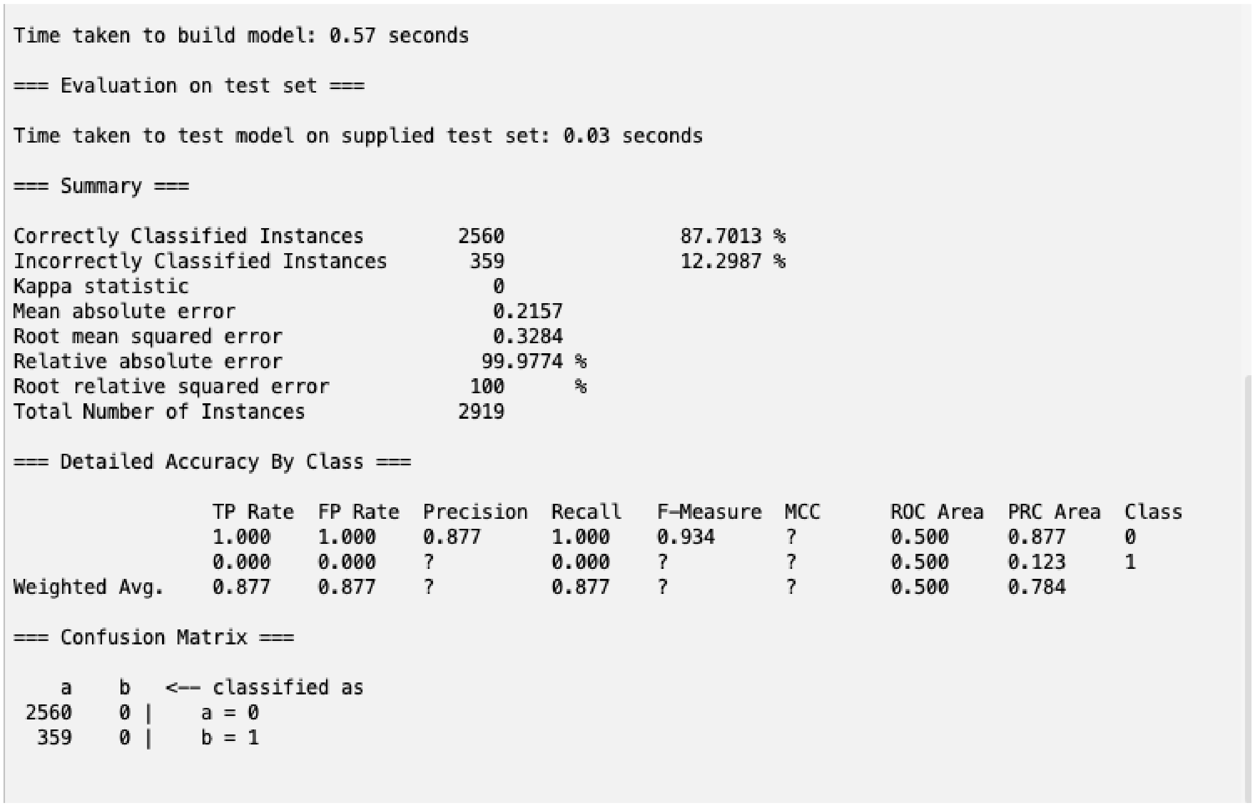
```
RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 4.34 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.2 seconds

=== Summary ===

Correctly Classified Instances        2554               87.4957 %
Incorrectly Classified Instances       365               12.5043 %
Kappa statistic                          0.0398
Mean absolute error                      0.2052
Root mean squared error                  0.3263
Relative absolute error                 95.1193 %
Root relative squared error             99.3391 %
Total Number of Instances             2919

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.993    0.969    0.880      0.993    0.933      0.081    0.649     0.923     0
                 0.031    0.007    0.393      0.031    0.057      0.081    0.649     0.209     1
Weighted Avg.    0.875    0.851    0.820      0.875    0.825      0.081    0.649     0.835

=== Confusion Matrix ===

    a    b   <-- classified as
 2543   17 |    a = 0
  348   11 |    b = 1
```

Random Forest performance on Normal Dataset



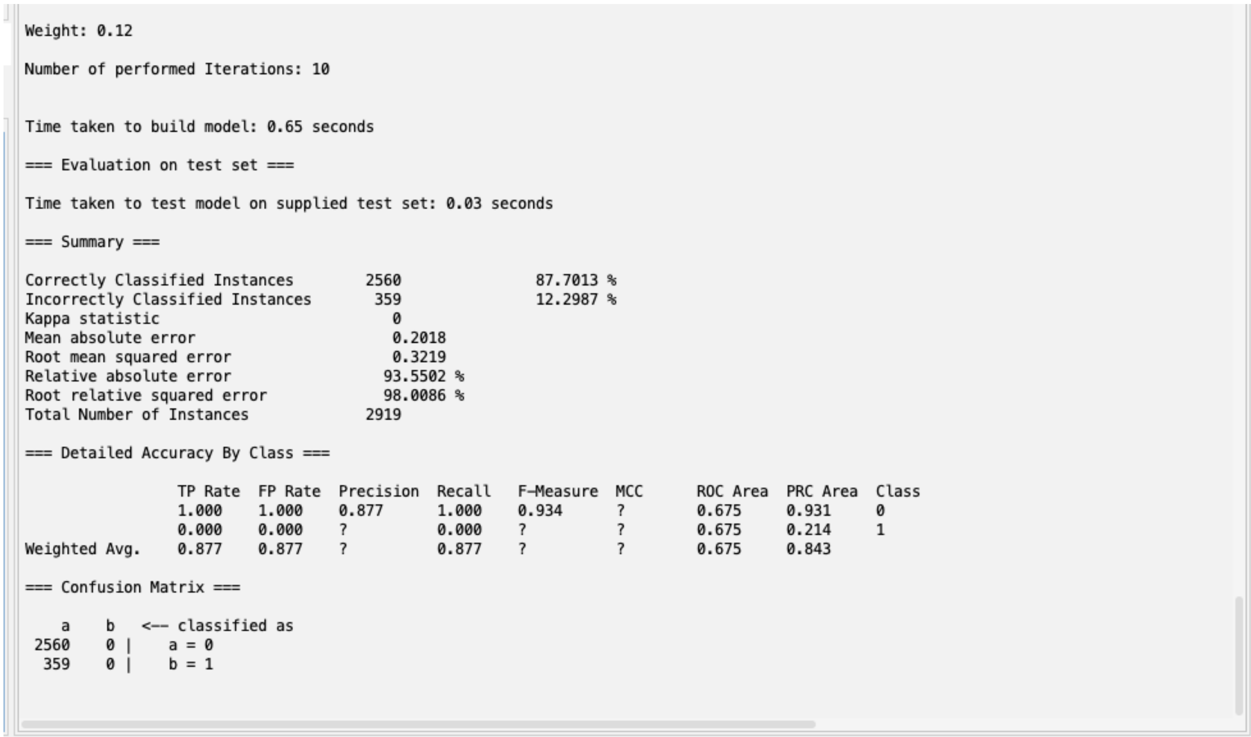
```
Time taken to build model: 0.06 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.09 seconds

=== Summary ===

Correctly Classified Instances        2203               75.4711 %
Incorrectly Classified Instances       716               24.5289 %
Kappa statistic                          0.1392
Mean absolute error                      0.2557
Root mean squared error                  0.4528
Relative absolute error                118.521  %
Root relative squared error            137.8683 %
Total Number of Instances             2919

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.808    0.624    0.902      0.808    0.852      0.147    0.659     0.929     0
                 0.376    0.192    0.215      0.376    0.274      0.147    0.659     0.208     1
Weighted Avg.    0.755    0.571    0.818      0.755    0.781      0.147    0.659     0.840

=== Confusion Matrix ===

    a    b   <-- classified as
 2068  492 |    a = 0
  224  135 |    b = 1
```

Naive Bayes performance on Normal Dataset



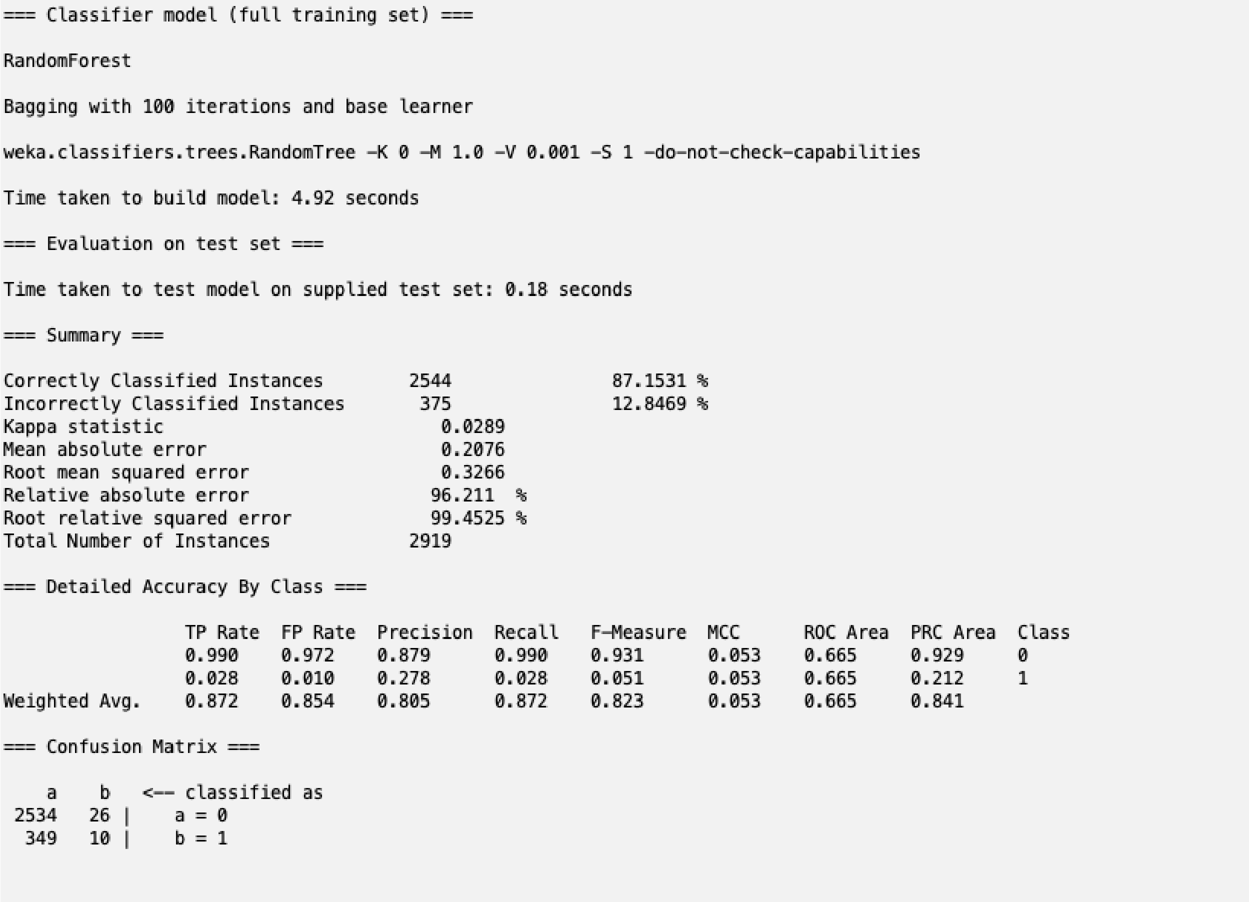
```
Time taken to build model: 0.57 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.03 seconds

=== Summary ===

Correctly Classified Instances        2560               87.7013 %
Incorrectly Classified Instances       359               12.2987 %
Kappa statistic                          0
Mean absolute error                      0.2157
Root mean squared error                  0.3284
Relative absolute error                 99.9774 %
Root relative squared error            100      %
Total Number of Instances             2919

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.877      1.000    0.934      ?        0.500     0.877     0
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.123     1
Weighted Avg.    0.877    0.877    ?          0.877    ?          ?        0.500     0.784

=== Confusion Matrix ===

    a    b   <-- classified as
 2560    0 |    a = 0
  359    0 |    b = 1
```

J48 performance on Normal Dataset



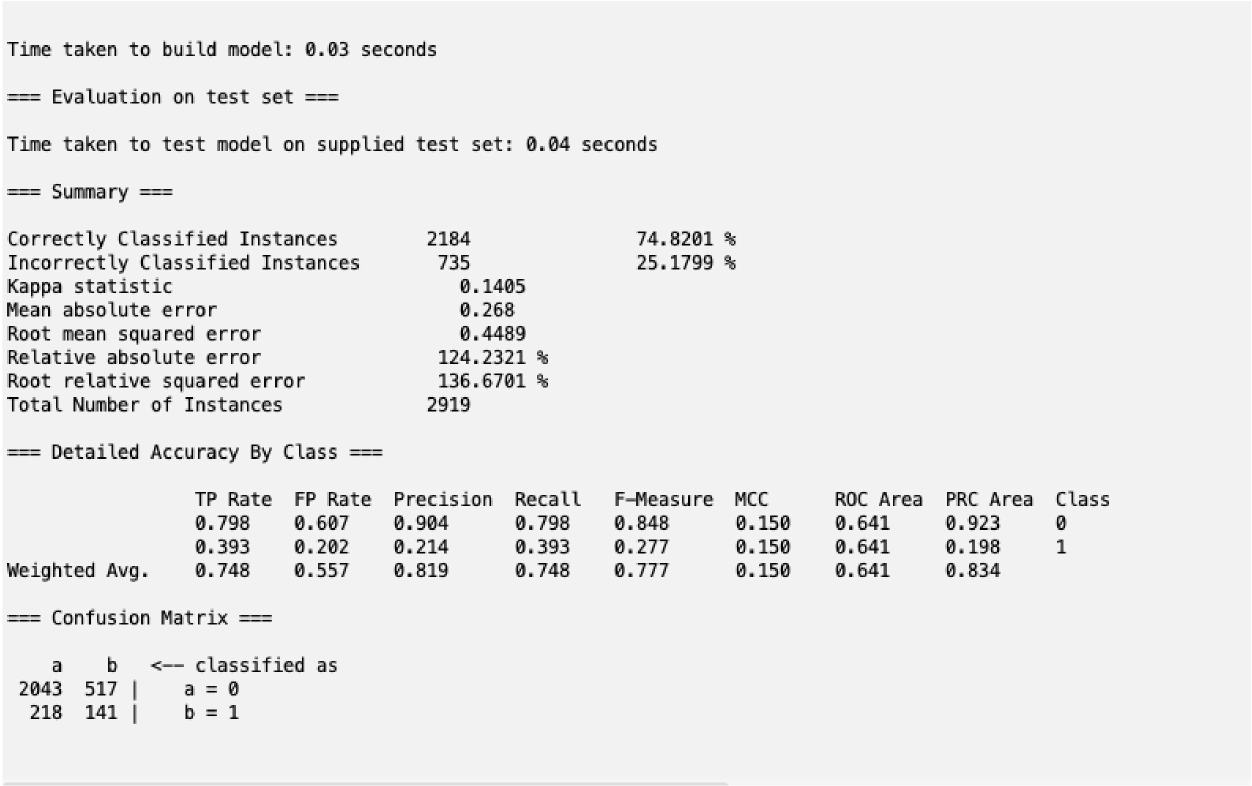
```
Weight: 0.12

Number of performed Iterations: 10


Time taken to build model: 0.65 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.03 seconds

=== Summary ===

Correctly Classified Instances        2560               87.7013 %
Incorrectly Classified Instances       359               12.2987 %
Kappa statistic                          0
Mean absolute error                      0.2018
Root mean squared error                  0.3219
Relative absolute error                 93.5502 %
Root relative squared error             98.0086 %
Total Number of Instances             2919

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.877      1.000    0.934      ?        0.675     0.931     0
                 0.000    0.000    ?          0.000    ?          ?        0.675     0.214     1
Weighted Avg.    0.877    0.877    ?          0.877    ?          ?        0.675     0.843

=== Confusion Matrix ===

    a    b   <-- classified as
 2560    0 |    a = 0
  359    0 |    b = 1
```

Adaboost M1 performance on Normal Dataset



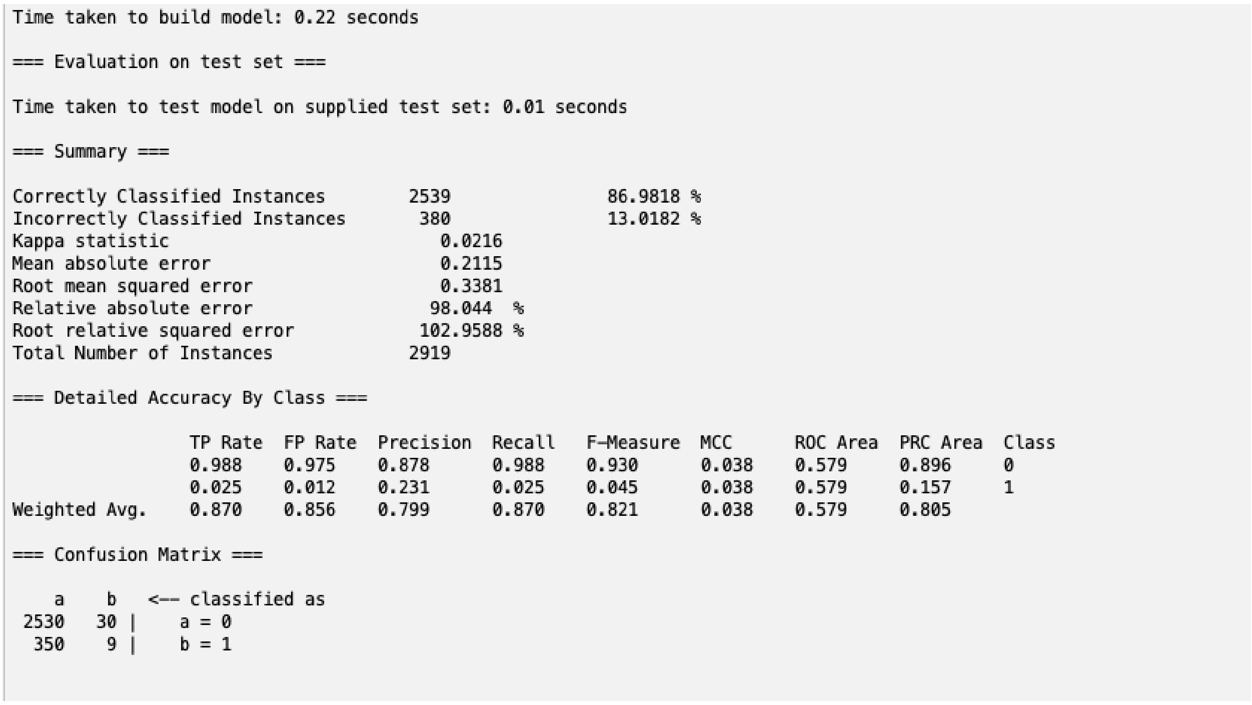
```
=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 4.92 seconds

=== Evaluation on test set ===

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

=== Summary ===

Correctly Classified Instances        2544               87.1531 %
Incorrectly Classified Instances       375               12.8469 %
Kappa statistic                          0.0289
Mean absolute error                      0.2076
Root mean squared error                  0.3266
Relative absolute error                 96.211  %
Root relative squared error             99.4525 %
Total Number of Instances             2919

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.990    0.972    0.879      0.990    0.931      0.053    0.665     0.929     0
                 0.028    0.010    0.278      0.028    0.051      0.053    0.665     0.212     1
Weighted Avg.    0.872    0.854    0.805      0.872    0.823      0.053    0.665     0.841

=== Confusion Matrix ===

    a    b   <-- classified as
 2534   26 |    a = 0
  349   10 |    b = 1
```

Random Forest performance with Correlation Attribute Selection



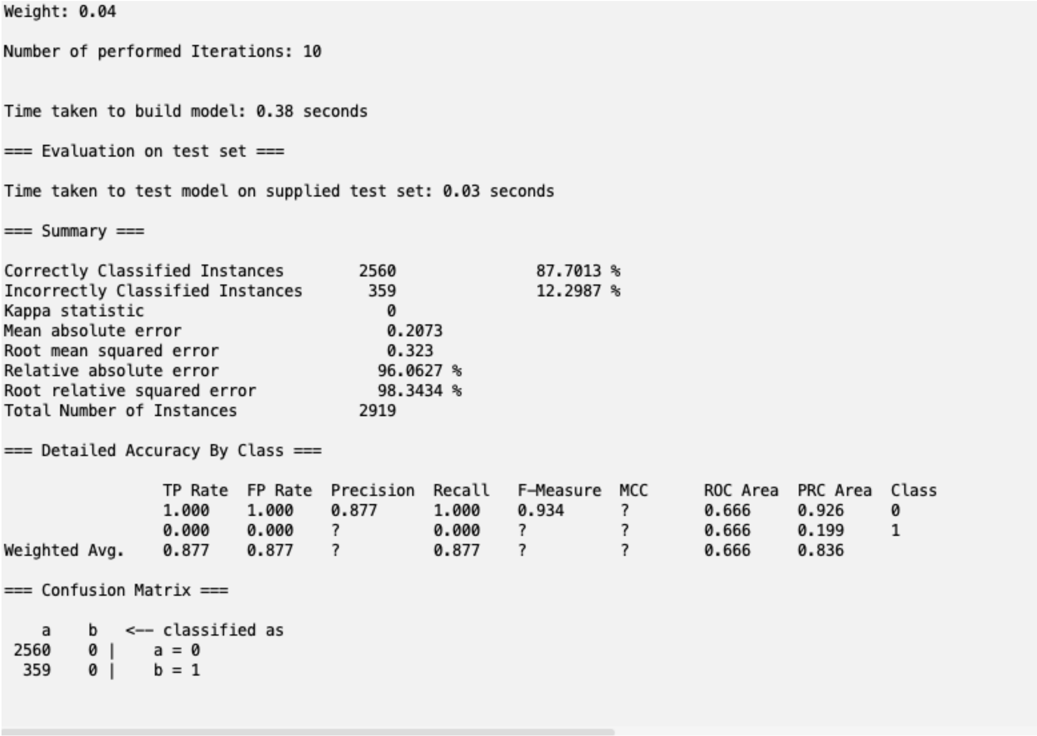
```
Time taken to build model: 0.03 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.04 seconds

=== Summary ===

Correctly Classified Instances        2184               74.8201 %
Incorrectly Classified Instances       735               25.1799 %
Kappa statistic                          0.1405
Mean absolute error                      0.268 
Root mean squared error                  0.4489
Relative absolute error                124.2321 %
Root relative squared error            136.6701 %
Total Number of Instances             2919     

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.798    0.607    0.904      0.798    0.848      0.150    0.641     0.923     0
                 0.393    0.202    0.214      0.393    0.277      0.150    0.641     0.198     1
Weighted Avg.    0.748    0.557    0.819      0.748    0.777      0.150    0.641     0.834     

=== Confusion Matrix ===

    a    b   <-- classified as
 2043  517 |    a = 0
  218  141 |    b = 1
```

Naive Bayes performance with Correlation Attribute Selection



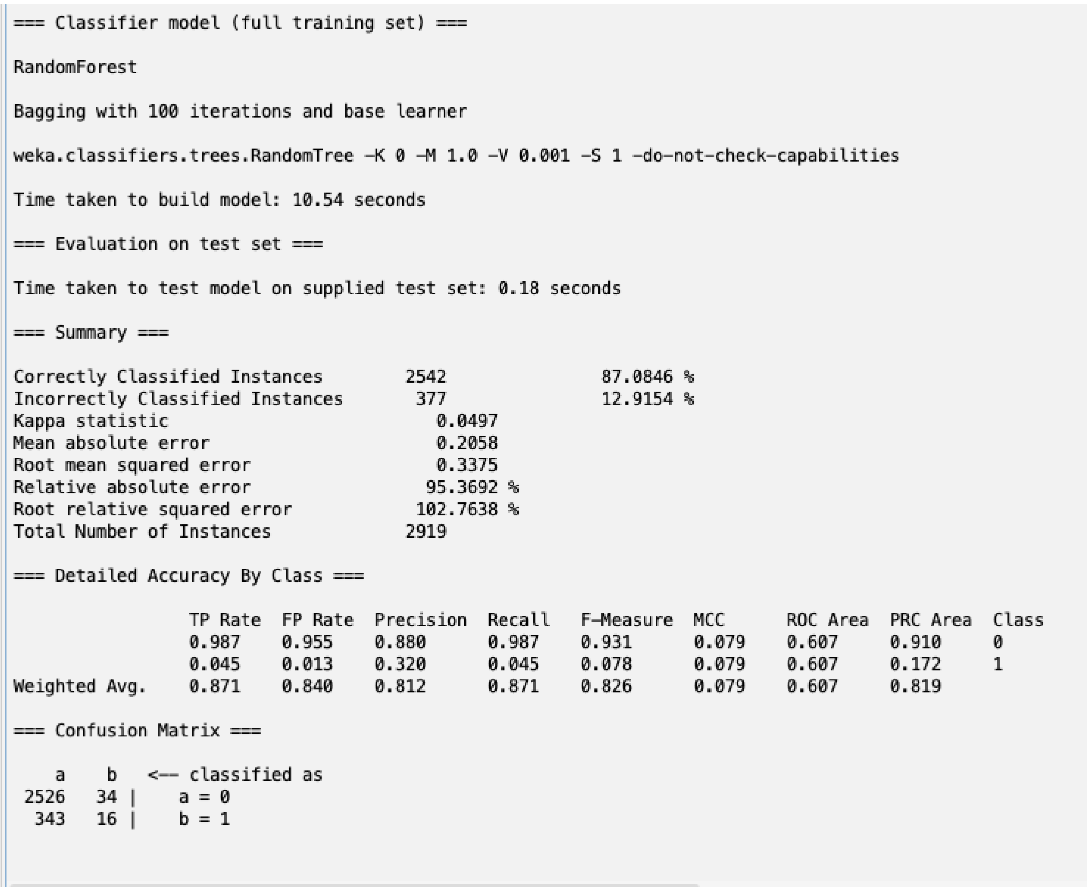
```
Time taken to build model: 0.22 seconds

=== Evaluation on test set ===

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

=== Summary ===

Correctly Classified Instances        2539               86.9818 %
Incorrectly Classified Instances       380               13.0182 %
Kappa statistic                          0.0216
Mean absolute error                      0.2115
Root mean squared error                  0.3381
Relative absolute error                 98.044  %
Root relative squared error            102.9588 %
Total Number of Instances             2919     

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.988    0.975    0.878      0.988    0.930      0.038    0.579     0.896     0
                 0.025    0.012    0.231      0.025    0.045      0.038    0.579     0.157     1
Weighted Avg.    0.870    0.856    0.799      0.870    0.821      0.038    0.579     0.805     

=== Confusion Matrix ===

    a    b   <-- classified as
 2530   30 |    a = 0
  350    9 |    b = 1
```

J48 performance with Correlation Attribute Selection



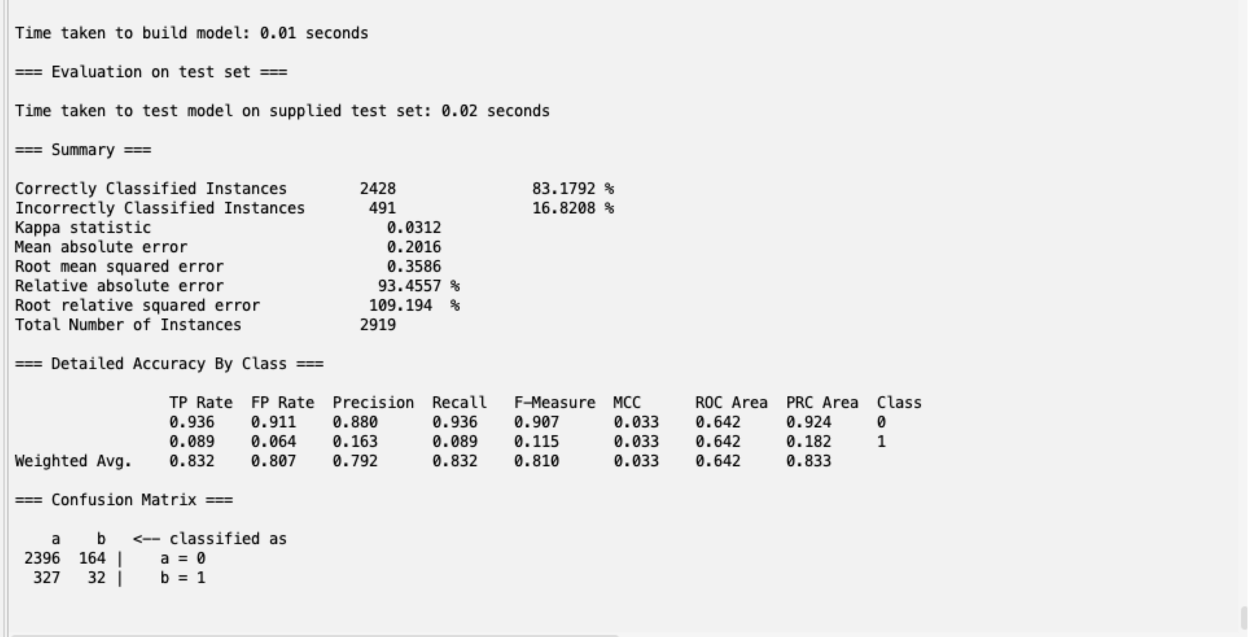
```
Weight: 0.04

Number of performed Iterations: 10


Time taken to build model: 0.38 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.03 seconds

=== Summary ===

Correctly Classified Instances        2560               87.7013 %
Incorrectly Classified Instances       359               12.2987 %
Kappa statistic                          0
Mean absolute error                      0.2073
Root mean squared error                  0.323
Relative absolute error                 96.0627 %
Root relative squared error             98.3434 %
Total Number of Instances             2919

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.877      1.000    0.934      ?        0.666     0.926     0
                 0.000    0.000    ?          0.000    ?          ?        0.666     0.199     1
Weighted Avg.    0.877    0.877    ?          0.877    ?          ?        0.666     0.836

=== Confusion Matrix ===

    a    b   <-- classified as
 2560    0 |    a = 0
  359    0 |    b = 1
```

Adaboost M1 performance with Correlation Attribute Selection



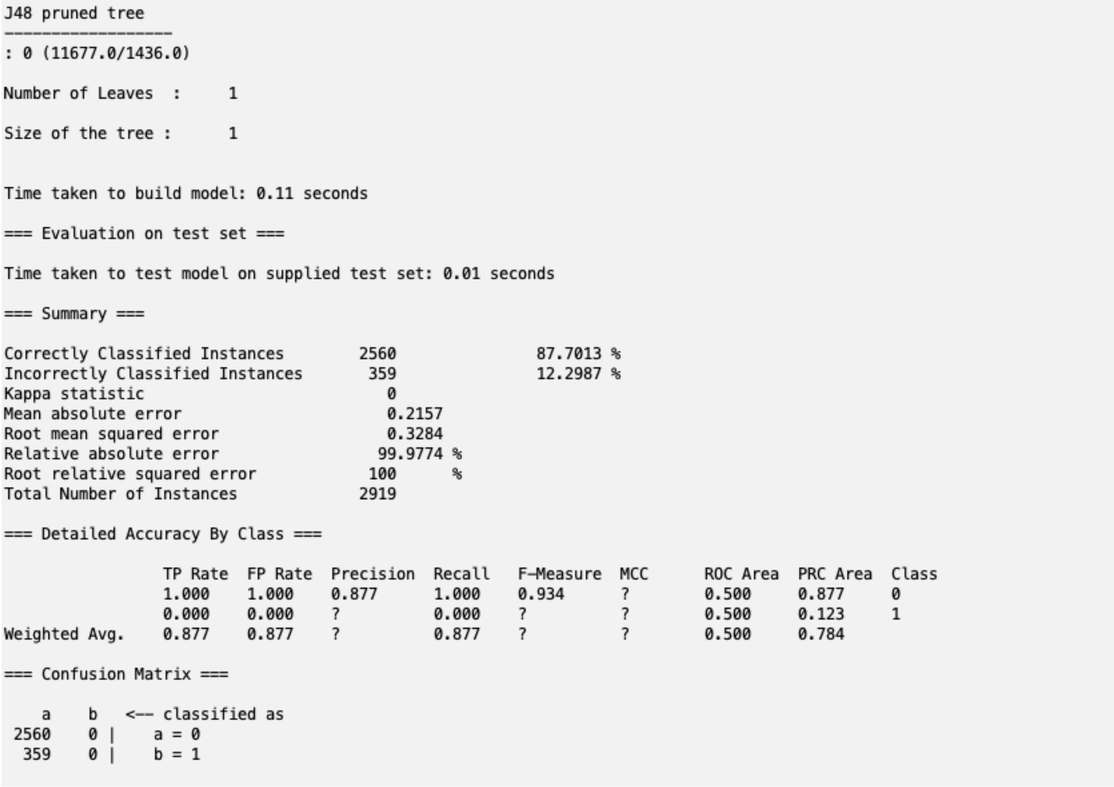
```
=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 10.54 seconds

=== Evaluation on test set ===

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

=== Summary ===

Correctly Classified Instances        2542               87.0846 %
Incorrectly Classified Instances       377               12.9154 %
Kappa statistic                          0.0497
Mean absolute error                      0.2058
Root mean squared error                  0.3375
Relative absolute error                 95.3692 %
Root relative squared error            102.7638 %
Total Number of Instances             2919

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.987    0.955    0.880      0.987    0.931      0.079    0.607     0.910     0
                 0.045    0.013    0.320      0.045    0.078      0.079    0.607     0.172     1
Weighted Avg.    0.871    0.840    0.812      0.871    0.826      0.079    0.607     0.819

=== Confusion Matrix ===

    a    b   <-- classified as
 2526   34 |    a = 0
  343   16 |    b = 1
```

Random Forest performance on ReliefF Attribute Selection



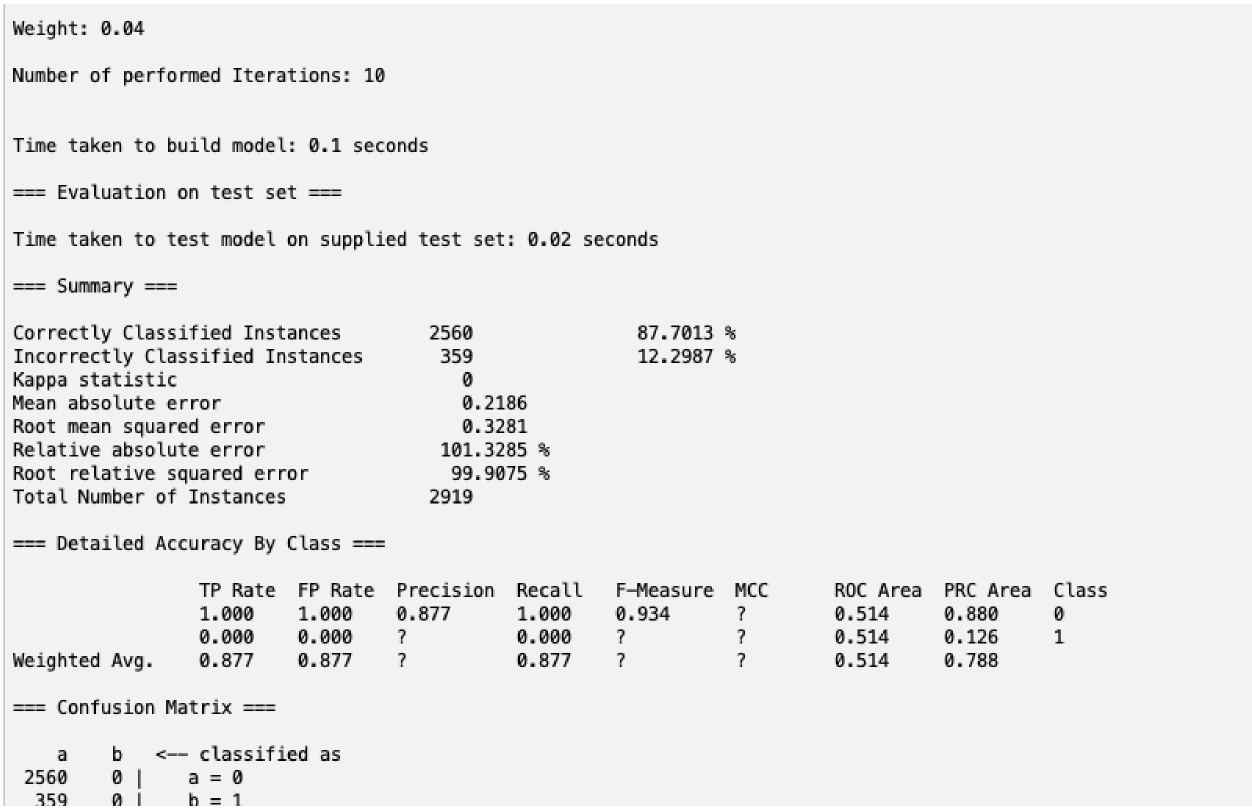
```
Time taken to build model: 0.01 seconds

=== Evaluation on test set ===

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

=== Summary ===

Correctly Classified Instances        2428               83.1792 %
Incorrectly Classified Instances       491               16.8208 %
Kappa statistic                          0.0312
Mean absolute error                      0.2016
Root mean squared error                  0.3586
Relative absolute error                 93.4557 %
Root relative squared error            109.194  %
Total Number of Instances             2919

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.936    0.911    0.880      0.936    0.907      0.033    0.642     0.924     0
                 0.089    0.064    0.163      0.089    0.115      0.033    0.642     0.182     1
Weighted Avg.    0.832    0.807    0.792      0.832    0.810      0.033    0.642     0.833

=== Confusion Matrix ===

    a    b   <-- classified as
 2396  164 |    a = 0
  327   32 |    b = 1
```

Naive Bayes performance on ReliefF Attribute Selection



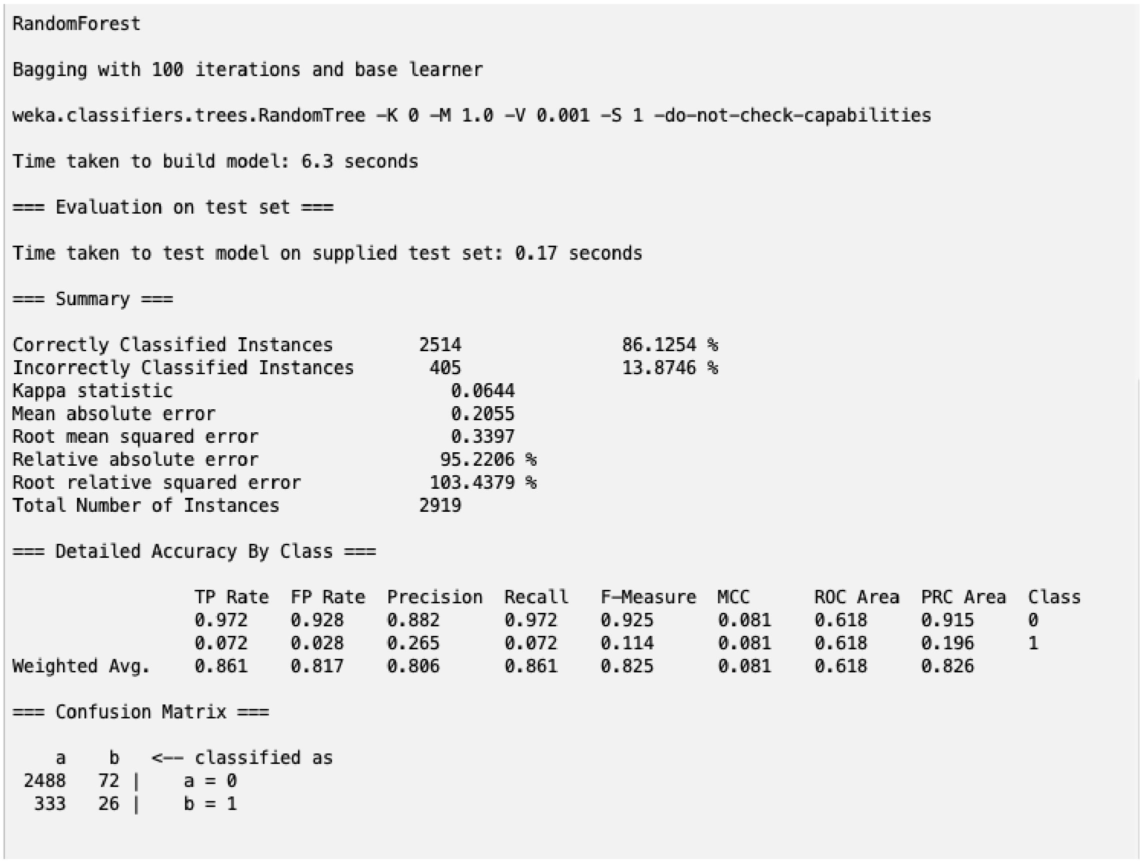
```
J48 pruned tree
------------------
: 0 (11677.0/1436.0)

Number of Leaves  :     1

Size of the tree :      1


Time taken to build model: 0.11 seconds

=== Evaluation on test set ===

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

=== Summary ===

Correctly Classified Instances        2560               87.7013 %
Incorrectly Classified Instances       359               12.2987 %
Kappa statistic                          0
Mean absolute error                      0.2157
Root mean squared error                  0.3284
Relative absolute error                 99.9774 %
Root relative squared error            100      %
Total Number of Instances             2919

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.877      1.000    0.934      ?        0.500     0.877     0
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.123     1
Weighted Avg.    0.877    0.877    ?          0.877    ?          ?        0.500     0.784

=== Confusion Matrix ===

    a    b   <-- classified as
 2560    0 |    a = 0
  359    0 |    b = 1
```

J48 performance on ReliefF Attribute Selection



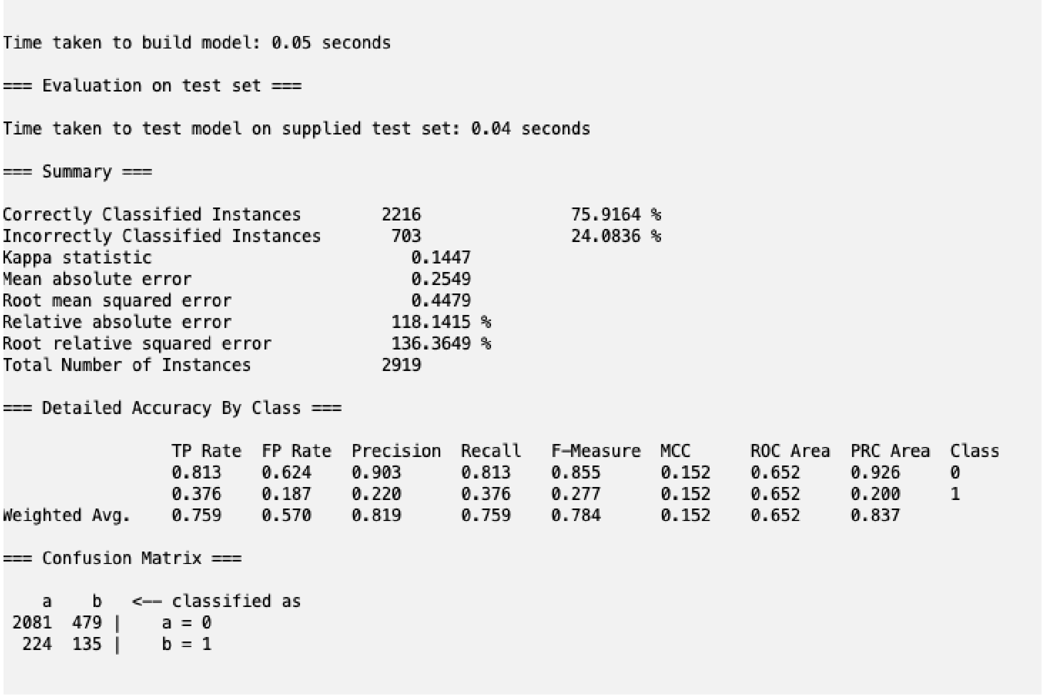
```
Weight: 0.04

Number of performed Iterations: 10


Time taken to build model: 0.1 seconds

=== Evaluation on test set ===

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

=== Summary ===

Correctly Classified Instances        2560               87.7013 %
Incorrectly Classified Instances       359               12.2987 %
Kappa statistic                          0
Mean absolute error                      0.2186
Root mean squared error                  0.3281
Relative absolute error                101.3285 %
Root relative squared error             99.9075 %
Total Number of Instances             2919

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.877      1.000    0.934      ?        0.514     0.880     0
                 0.000    0.000    ?          0.000    ?          ?        0.514     0.126     1
Weighted Avg.    0.877    0.877    ?          0.877    ?          ?        0.514     0.788

=== Confusion Matrix ===

    a    b   <-- classified as
 2560    0 |    a = 0
  359    0 |    b = 1
```

Adaboost Performance with ReliefF Attribute Selection



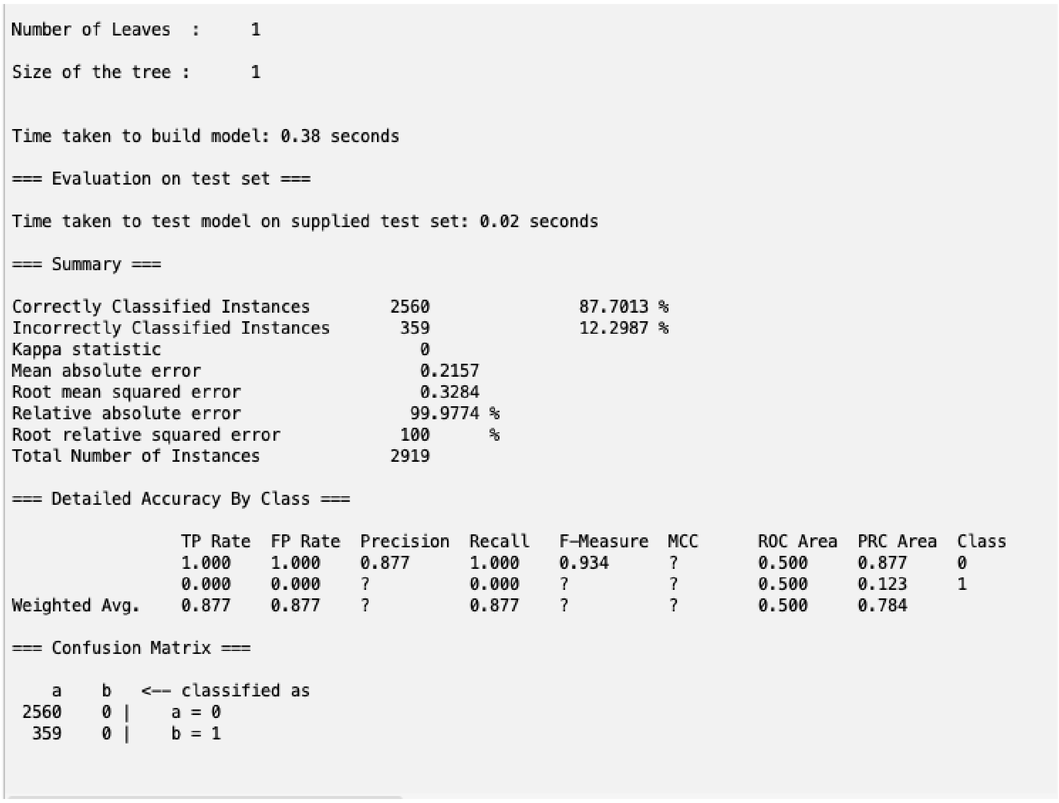
```
RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 6.3 seconds

=== Evaluation on test set ===

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

=== Summary ===

Correctly Classified Instances        2514               86.1254 %
Incorrectly Classified Instances       405               13.8746 %
Kappa statistic                          0.0644
Mean absolute error                      0.2055
Root mean squared error                  0.3397
Relative absolute error                 95.2206 %
Root relative squared error            103.4379 %
Total Number of Instances             2919

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.972    0.928    0.882      0.972    0.925      0.081    0.618     0.915     0
                 0.072    0.028    0.265      0.072    0.114      0.081    0.618     0.196     1
Weighted Avg.    0.861    0.817    0.806      0.861    0.825      0.081    0.618     0.826

=== Confusion Matrix ===

    a    b   <-- classified as
 2488   72 |    a = 0
  333   26 |    b = 1
```

Random Forest with Gain Ratio Attribute Selection



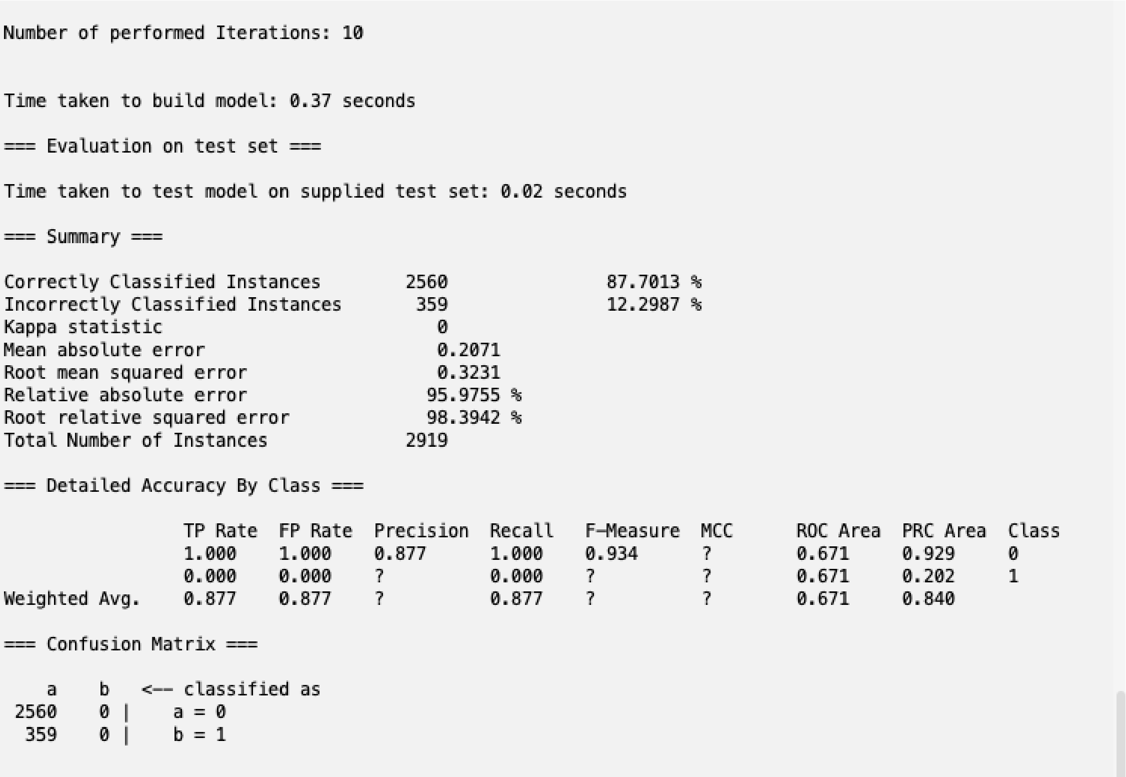
```
Time taken to build model: 0.05 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.04 seconds

=== Summary ===

Correctly Classified Instances        2216               75.9164 %
Incorrectly Classified Instances       703               24.0836 %
Kappa statistic                          0.1447
Mean absolute error                      0.2549
Root mean squared error                  0.4479
Relative absolute error                118.1415 %
Root relative squared error            136.3649 %
Total Number of Instances             2919

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.813    0.624    0.903      0.813    0.855      0.152    0.652     0.926     0
                 0.376    0.187    0.220      0.376    0.277      0.152    0.652     0.200     1
Weighted Avg.    0.759    0.570    0.819      0.759    0.784      0.152    0.652     0.837

=== Confusion Matrix ===

    a    b   <-- classified as
 2081  479 |    a = 0
  224  135 |    b = 1
```

Naive Bayes with Gain Ratio Attribute Selection



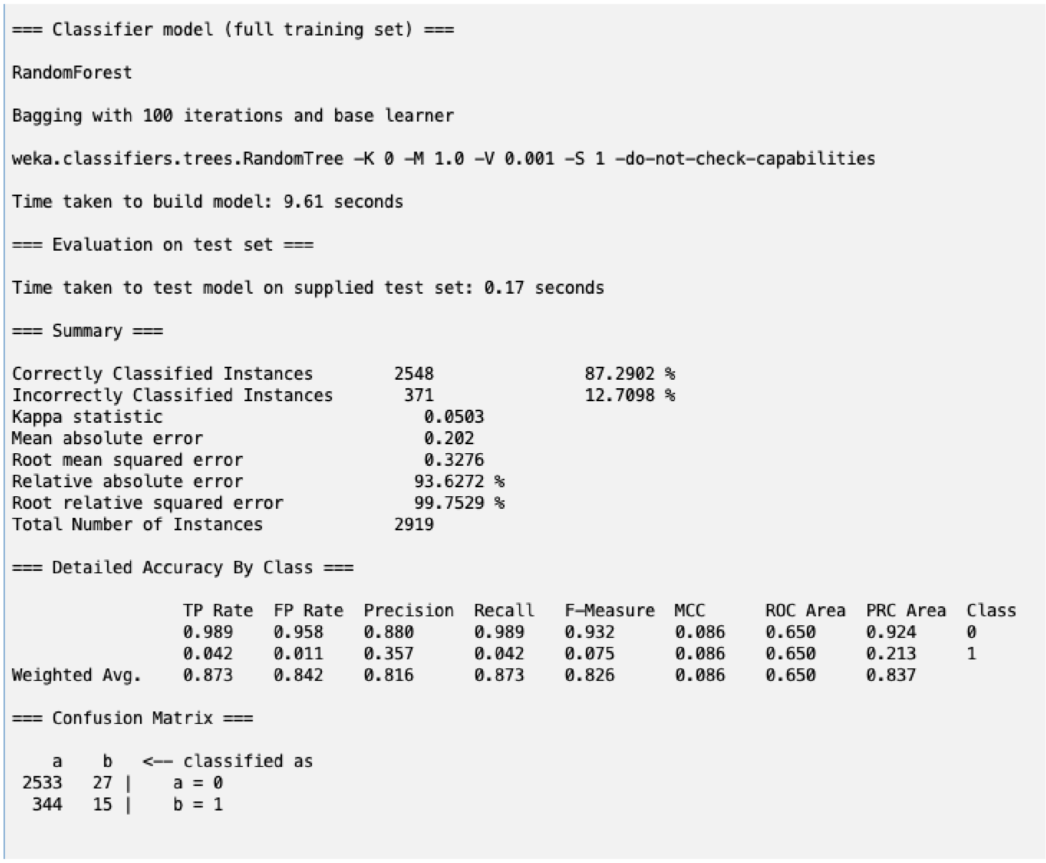
```
Number of Leaves  :     1

Size of the tree :      1


Time taken to build model: 0.38 seconds

=== Evaluation on test set ===

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

=== Summary ===

Correctly Classified Instances        2560               87.7013 %
Incorrectly Classified Instances       359               12.2987 %
Kappa statistic                          0
Mean absolute error                      0.2157
Root mean squared error                  0.3284
Relative absolute error                 99.9774 %
Root relative squared error            100      %
Total Number of Instances             2919

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.877      1.000    0.934      ?        0.500     0.877     0
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.123     1
Weighted Avg.    0.877    0.877    ?          0.877    ?          ?        0.500     0.784

=== Confusion Matrix ===

    a    b   <-- classified as
 2560    0 |    a = 0
  359    0 |    b = 1
```

J48 Performance on Gain Ratio Attribute Selection



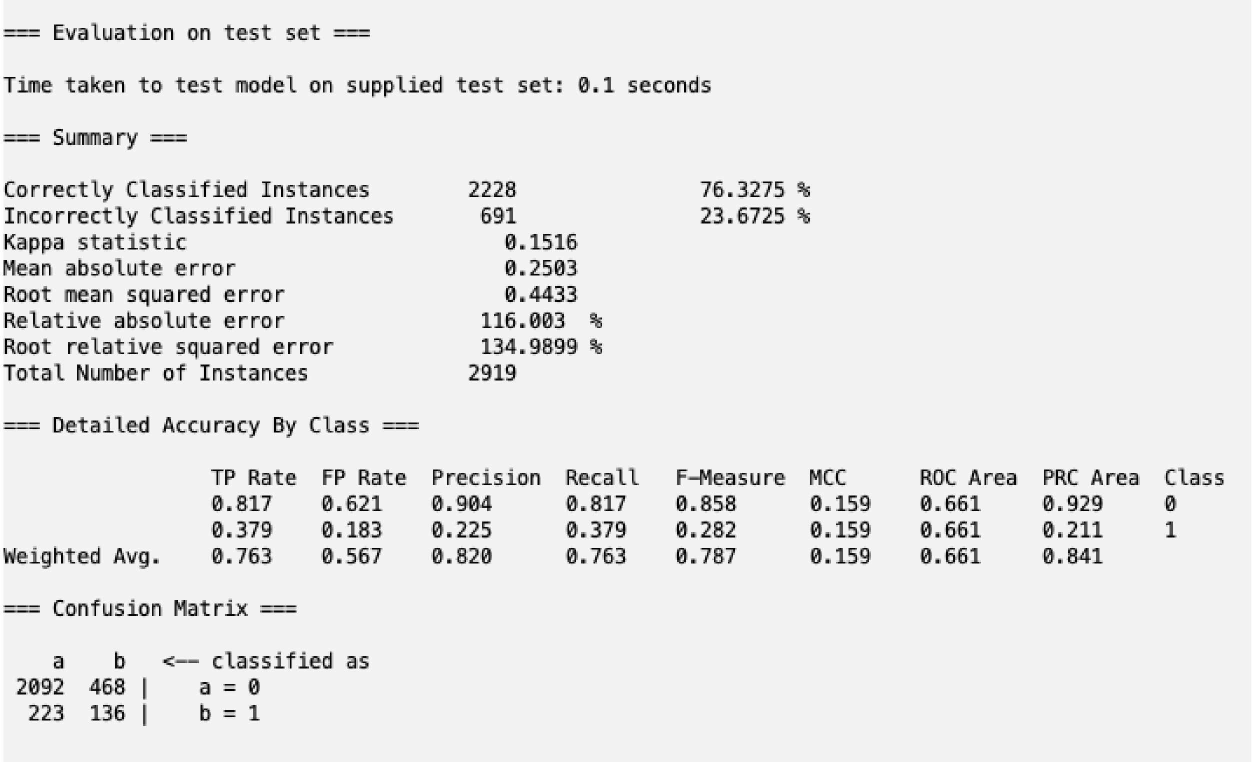
```
Number of performed Iterations: 10


Time taken to build model: 0.37 seconds

=== Evaluation on test set ===

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

=== Summary ===

Correctly Classified Instances        2560               87.7013 %
Incorrectly Classified Instances       359               12.2987 %
Kappa statistic                          0
Mean absolute error                      0.2071
Root mean squared error                  0.3231
Relative absolute error                 95.9755 %
Root relative squared error             98.3942 %
Total Number of Instances             2919

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.877      1.000    0.934      ?        0.671     0.929     0
                 0.000    0.000    ?          0.000    ?          ?        0.671     0.202     1
Weighted Avg.    0.877    0.877    ?          0.877    ?          ?        0.671     0.840

=== Confusion Matrix ===

    a    b   <-- classified as
 2560    0 |    a = 0
  359    0 |    b = 1
```

AdaBoost with Gain Ratio Attribute Selection



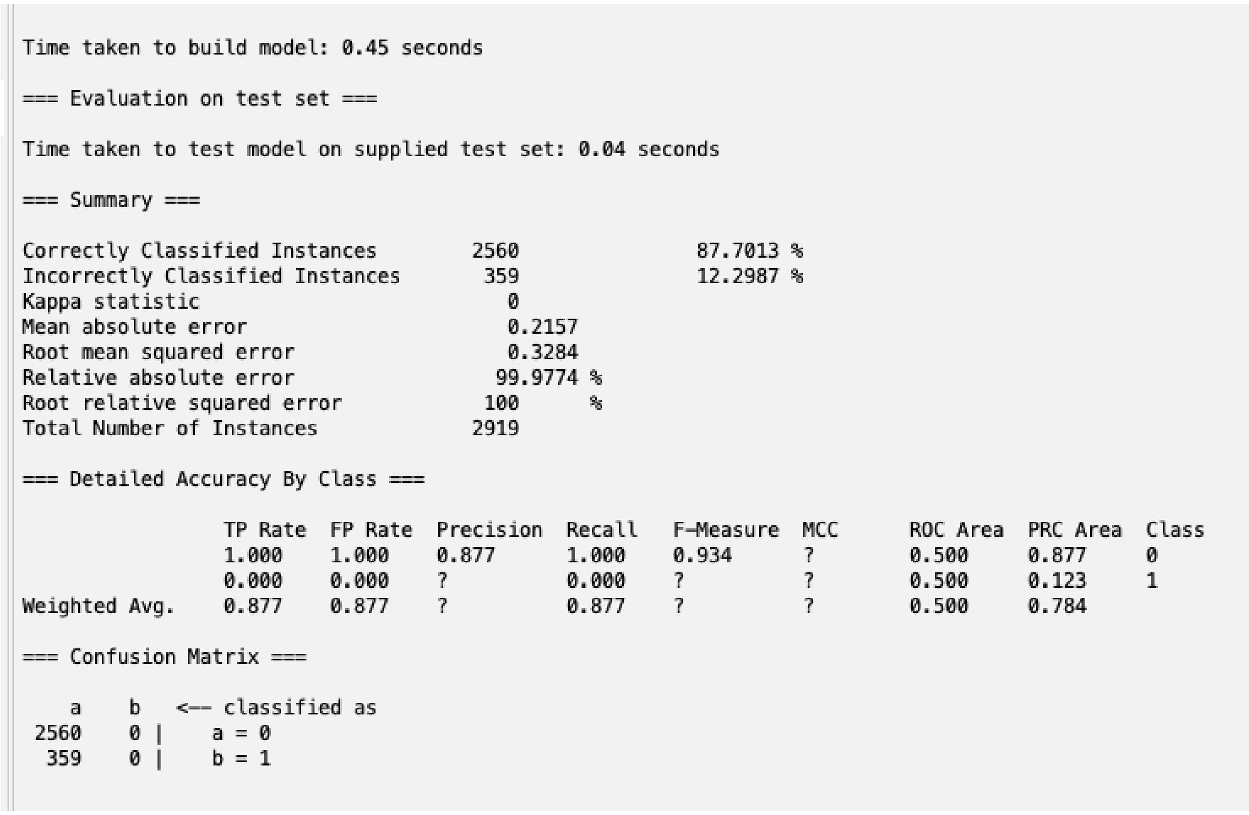
```
=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 9.61 seconds

=== Evaluation on test set ===

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

=== Summary ===

Correctly Classified Instances        2548               87.2902 %
Incorrectly Classified Instances       371               12.7098 %
Kappa statistic                          0.0503
Mean absolute error                      0.202
Root mean squared error                  0.3276
Relative absolute error                 93.6272 %
Root relative squared error             99.7529 %
Total Number of Instances             2919

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.989    0.958    0.880      0.989    0.932      0.086    0.650     0.924     0
                 0.042    0.011    0.357      0.042    0.075      0.086    0.650     0.213     1
Weighted Avg.    0.873    0.842    0.816      0.873    0.826      0.086    0.650     0.837

=== Confusion Matrix ===

    a    b   <-- classified as
 2533   27 |    a = 0
  344   15 |    b = 1
```

Random Forest with Information Gain Attribute Selection



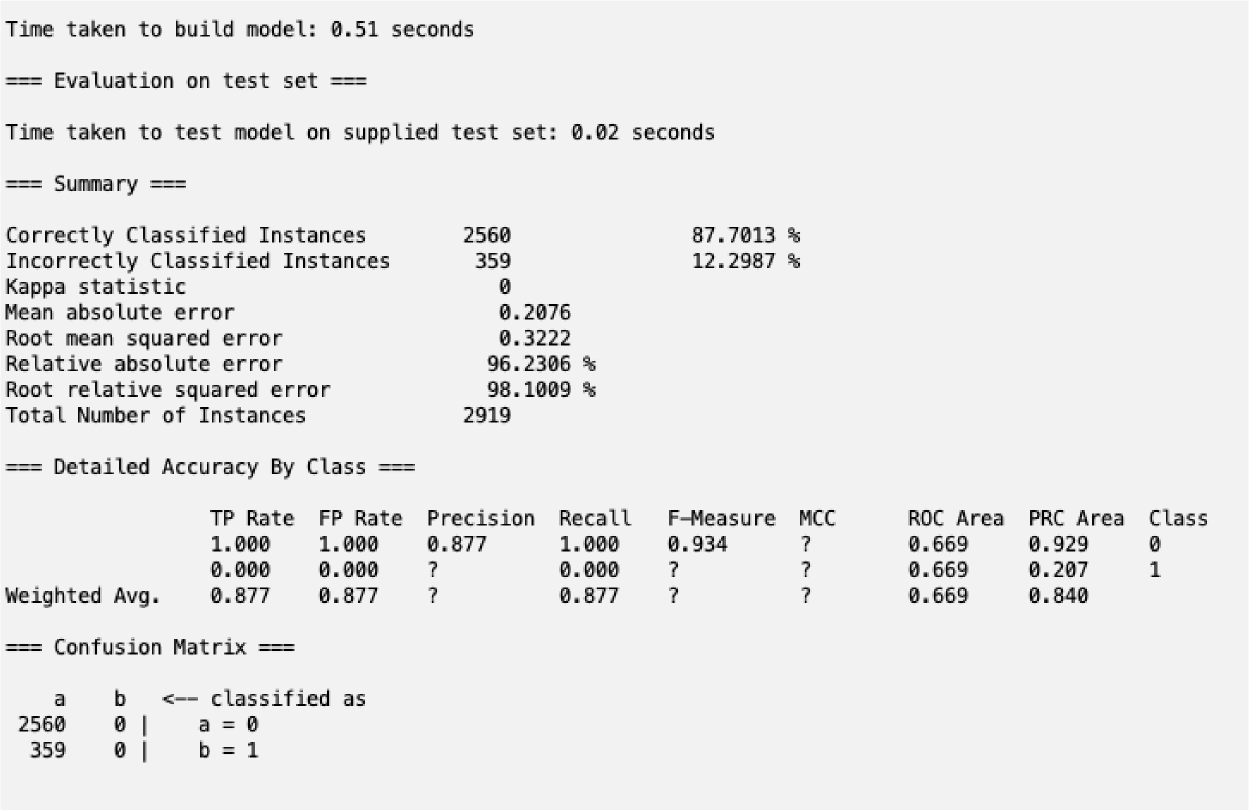
```
=== Evaluation on test set ===

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

=== Summary ===

Correctly Classified Instances        2228               76.3275 %
Incorrectly Classified Instances       691               23.6725 %
Kappa statistic                          0.1516
Mean absolute error                      0.2503
Root mean squared error                  0.4433
Relative absolute error                116.003  %
Root relative squared error            134.9899 %
Total Number of Instances             2919

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.817    0.621    0.904      0.817    0.858      0.159    0.661     0.929     0
                 0.379    0.183    0.225      0.379    0.282      0.159    0.661     0.211     1
Weighted Avg.    0.763    0.567    0.820      0.763    0.787      0.159    0.661     0.841

=== Confusion Matrix ===

    a    b   <-- classified as
 2092  468 |    a = 0
  223  136 |    b = 1
```

Naive Bayes with Information Gain Attribute Selection



```
Time taken to build model: 0.45 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.04 seconds

=== Summary ===

Correctly Classified Instances        2560               87.7013 %
Incorrectly Classified Instances       359               12.2987 %
Kappa statistic                          0
Mean absolute error                      0.2157
Root mean squared error                  0.3284
Relative absolute error                 99.9774 %
Root relative squared error            100      %
Total Number of Instances             2919

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.877      1.000    0.934      ?        0.500     0.877     0
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.123     1
Weighted Avg.    0.877    0.877    ?          0.877    ?          ?        0.500     0.784

=== Confusion Matrix ===

    a    b   <-- classified as
 2560    0 |    a = 0
  359    0 |    b = 1
```

J48 Performance on Information Gain Attribute Selection



```
Time taken to build model: 0.51 seconds

=== Evaluation on test set ===

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

=== Summary ===

Correctly Classified Instances        2560               87.7013 %
Incorrectly Classified Instances       359               12.2987 %
Kappa statistic                          0
Mean absolute error                      0.2076
Root mean squared error                  0.3222
Relative absolute error                 96.2306 %
Root relative squared error             98.1009 %
Total Number of Instances             2919

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.877      1.000    0.934      ?        0.669     0.929     0
                 0.000    0.000    ?          0.000    ?          ?        0.669     0.207     1
Weighted Avg.    0.877    0.877    ?          0.877    ?          ?        0.669     0.840

=== Confusion Matrix ===

    a    b   <-- classified as
 2560    0 |    a = 0
  359    0 |    b = 1
```

AdaBoost with Information Gain Attribute Selection

Analysis/Conclusion

Accuracy (%)				
	Random Forests	Naive Bayes	J48	Adaboost
Normal Dataset	87.4957	75.4711	87.7013	87.7013
Correlation	87.1531	74.8201	86.9818	87.7013
ReliefF	87.0846	83.1792	87.7013	87.7013
Gain Ratio	86.1254	75.9164	87.7013	87.7013
Info Gain	87.2902	76.3275	87.7013	87.7013

Table 1: Accuracy for Various Models

TP Rate				
	Random Forests	Naive Bayes	J48	Adaboost
Normal Dataset	0.875	0.755	0.877	0.877

Correlation	0.872	0.748	0.87	0.877
ReliefF	0.871	0.832	0.877	0.877
Gain Ratio	0.861	0.759	0.877	0.877
Information Gain	0.873	0.763	0.877	0.877

Table 2: Weighted True Positive Rate for Various Models

	FP Rate			
	Random Forests	Naive Bayes	J48	Adaboost
Normal Dataset	0.851	0.571	0.877	0.877
Correlation	0.854	0.557	0.856	0.877
ReliefF	0.84	0.807	0.877	0.877
Gain Ratio	0.817	0.57	0.877	0.877
Information Gain	0.842	0.567	0.877	0.877

Table 3: Weighted False Positive Rate for Various models

Due to the heavy class imbalance towards the negative(isn't a bad buy) class, accuracy, true positives, and false positives, aren't the best metrics to view its performance. As seen in Table 1, J48 and Adaboost seem to have high accuracies, but instead just predicts the "isn't a bad buy" class nearly 100% of the time, yielding high accuracies, but not a very useful model. That's why it is better to look at alternate metrics, such as recall, which measures how well a model performs on the positive class, which is the "bad buy" class.

	Recall			
	Random Forests	Naive Bayes	J48	Adaboost
Normal Dataset	0.031	0.376	0	0
Correlation	0.028	0.393	0.025	0
ReliefF	0.045	0.089	0	0
Gain Ratio	0.072	0.376	0	0
Information Gain	0.042	0.379	0	0

Table 4: Recall Scores of Various Models

As seen in Table 4, all models seemed to struggle to have high accuracies on the IsBadBuy positive target class. In this case, recall is a much more important metric, because the goal of the project was to identify cars with the potential to be at risk to be a bad buy. If a model predicts every car to be a good buy, then it fails to accomplish its task. J48 and Adaboost don't adapt well to heavily imbalanced datasets, with its recall scores of 0. Random Forests can handle them, but do so poorly, seen with the low recall scores. Naive Bayes has a higher recall rate, so it can correctly identify bad buys a higher percentage of the time, when compared to the other models. Since this best achieves the goal of the project, the Naive Bayes, although having the lower accuracy, is the best model for the task. The correlation attribute selector method also seemed to have the highest recall for the Naive Bayes model. However, there is no general pattern across all models demonstrating one attribute selection superiority over the others. The combination of correlation attribute selection and Naive bayes model will be the best model for this project.

In this project, we learned about how to utilize WEKA's software to create models to accomplish the task of determining whether a car would be a bad buy. We needed to deal with real world, messy data and experiment with various models and attribute selection algorithms to find the best approach to the dataset. Furthermore, we identified the flaws that arise from the heavily imbalanced dataset as the one we had here. We learned about the need to look at metrics other than just the accuracy, such as investigating the confusion matrix or recall. In the future, we hope to experiment with alternate methods to deal with the class imbalance, such as generating artificial data, as well as try different models such as neural networks.

How to Reproduce Our Model

We provided all the datasets necessary in the google drive link. The original dataset is kick.arff; the original dataset with all missing values filled is kick_missing_filled.arff. The training and testing datasets are kick_balanced_train.arff and kick_balanced_test.arff respectively. The training and testing datasets for the correlation attribute selection with a threshold of 0.09 are kick_balanced_correlation_train.arff and kick_balanced_correlation_test.arff respectively. The training and testing datasets for the reliefF attribute selection with a threshold of 0.05 are kick_balanced_relieff_train.arff and kick_balanced_reliff_test.arff respectively.The training and testing datasets for the gain ratio attribute selection with a threshold of 0.007 are kick_balanced_train_GR.arff and kick_balanced_test_GR.arff respectively. The training and testing datasets for the information gain attribute selection with a threshold of 0.005 are kick_balanced_train_IG.arff and kick_balanced_test_IG.arff respectively. All of these datasets are located in the folder containing their attribute selection algorithm's name. Using these five datasets provided (the four pairs of datasets for the four attribute selection algorithms in addition to the original dataset split into train and test, kick_balanced_train.arff and kick_balanced_test.arff) and the four models (Random Forest, Naive Bayes, J48, Adaboost M1) on WEKA allows for reproducibility. For our attribute selection algorithms, we used the full training set. We trained all our models on the training set before setting the test set to the corresponding testing dataset and evaluating the model's performance.

Team Members and Tasks Performed

Vishal Kotha

- Initial Sampling
- Correlation Coefficient Attribute Selection + Model Training
- ReliefF Attribute Selection + Model Training

Arnav Jain

- Missing Value Imputation
- Train-Test Split
- Gain Ratio Attribute Selection + Model Training
- Info Gain Attribute Selection + Model Training

Together

- No Attribute Selection Model Training
- Writing of Reports
- Presentation

References

Data Source:
https://www.openml.org/search?type=data&sort=runs&id=41162&status=active

Models:
https://weka.sourceforge.io/doc.dev/weka/classifiers/trees/RandomForest.html
https://weka.sourceforge.io/doc.dev/weka/classifiers/bayes/NaiveBayes.html
https://weka.sourceforge.io/doc.dev/weka/classifiers/trees/J48.html
https://weka.sourceforge.io/doc.packages/realAdaBoost/

Attribute Selectors:
https://weka.sourceforge.io/doc.dev/weka/attributeSelection/CorrelationAttributeEval.html
https://weka.sourceforge.io/doc.dev/weka/attributeSelection/ReliefFAttributeEval.html
https://weka.sourceforge.io/doc.dev/weka/attributeSelection/GainRatioAttributeEval.html
https://weka.sourceforge.io/doc.dev/weka/attributeSelection/InfoGainAttributeEval.html