CS 699 Data Mining Project Report

Analysis & Prediction of Road Crash for Allegheny County

Tzupin Kuo, Samantha Lipsky

Contents

Statement of Data Mining Goal

Data Mining Goal

a. Predict the injury severity into two groups (No injury and Injury).
b. Compare the performance between 20 models.

Detailed Description of The Dataset

Dataset Introduction

Fatalities resulting from vehicle crashes is one of the main causes of death in the United States. Dataset from Western Pennsylvania Regional Data Center contains locations and information about every crash incident reported to the police in Alleghney County (located in the heart of southwestern Pennsylvania, encompasses 130 municipalities. The county seat is Pittsburgh) from 2004 to 2018. Fields include injury severity, fatalities, information about the vehicles involved, location information, and factors that may have contributed to the crash. The purpose of this study is to develop a model to predict the injury severity outcomes for vulnerable road users involving car crashes using different classifier algorithms. The study also attempts to identify factors that are important in making an injury severity difference and to explore the impact of such explanatory variables.

Dataset

Attribute	Type	Codes	Description
CRASH_MONTH	Nominal	▪ Jan ▪ Feb ▪ Mar ▪ Apr ▪ May ▪ Jun ▪ Jul ▪ Aug ▪ Sep ▪ Oct ▪ Nov ▪ Dec	Month when the crash occurred
DAY_OF_WEEK	Nominal	▪ Sunday ▪ Monday ▪ Tuesday ▪ Wednesday ▪ Thursday ▪ Friday	Day of the Week code when crash occurred

		▪ Saturday	
ILLUMINATION	Nominal	▪ Daylight ▪ Dark-no streetlights ▪ Dark-street lights ▪ Dusk ▪ Dawn ▪ Dark-unknown	Code that defines lighting at crash scene
WEATHER	Nominal	▪ No-adverse-conditions ▪ Rain ▪ Sleet (hail) ▪ Snow ▪ Fog	Code for the weather type at time of crash
ROAD_CONDITION	Nominal	▪ Dry ▪ Wet ▪ Sand/mud/dirt/oil/gravel ▪ Snow covered ▪ Slush ▪ Ice ▪ Ice Patches ▪ Water-standing-or-moving	Roadway Surface Condition Code
COLLISION_TYPE	Nominal	▪ Non collision ▪ Rear-end ▪ Head-on ▪ Rear-to-rear (Backing) ▪ Angle ▪ Sideswipe-same-direction ▪ Sideswipe-opposite-direction ▪ Hit fixed object ▪ Hit pedestrian	Collision category that defines the crash
INTERSECT_TYPE	Nominal	▪ Mid-block ▪ Four-way intersection ▪ T-intersection ▪ Y-intersection ▪ Traffic circle ▪ Multi-leg intersection ▪ On-ramp ▪ Off-ramp ▪ Crossover ▪ Railroad-crossing	Code that defines the Intersection Type
LOCAL_ROAD	Nominal	No / Yes	Local Road Indicator
TURNPIKE	Nominal	No / Yes	Turnpike Indicator
WET_ROAD	Nominal	No / Yes	Wet Road Indicator

ICY_ROAD	Nominal	No / Yes	Icy Road Indicator
REAR_END	Nominal	No / Yes	Rear End Collision Indicator
HO_OPPDIR_SDSWP	Nominal	No / Yes	Head on or Side Swipe Indicator
HIT_FIXED_OBJECT	Nominal	No / Yes	Hit Fixed Object Indicator
SV_RUN_OFF_RD	Nominal	No / Yes	Single Vehicle Run Off Road Indicator
WORK_ZONE	Nominal	No / Yes	Work Zone Indicator
PROPERTY_DAMAGE_ONLY	Nominal	No / Yes	Property Damage Only Indicator
FATAL_OR_MAJ_INJ	Nominal	No / Yes	Fatality or Major Injury Indicator
INJURY	Nominal	No / Yes	Injury Indicator
FATAL	Nominal	No / Yes	Fatality Indicator
INTERSECTION	Nominal	No / Yes	Intersection Indicator
UNSIGNALIZED_INT	Nominal	No / Yes	Unsignalized Intersection Indicator
SCHOOL_BUS	Nominal	No / Yes	School Bus Indicator
SCHOOL_ZONE	Nominal	No / Yes	School Zone Indicator
HIT_DEER	Nominal	No / Yes	Hit Deer Indicator
HIT_TREE_SHRUB	Nominal	No / Yes	Hit Tree or Shrub Indicator
HIT_EMBANKMENT	Nominal	No / Yes	Hit Embankment Indicator
HIT_POLE	Nominal	No / Yes	Hit Pole Indicator
HIT_GDRAIL	Nominal	No / Yes	Hit Guide Rail Indicator
HIT_GDRAIL_END	Nominal	No / Yes	Hit Guide Rail End Indicator
HIT_BARRIER	Nominal	No / Yes	Hit Barrier Indicator
HIT_BRIDGE	Nominal	No / Yes	Hit Bridge Indicator
OVERTURNED	Nominal	No / Yes	Overturned Vehicle Indicator
MOTORCYCLE	Nominal	No / Yes	Motorcycle Indicator
BICYCLE	Nominal	No / Yes	Bicycle Indicator

HVY_TRUCK_RELATED	Nominal	No / Yes	Heavy Truck Related Indicator
VEHICLE_FAILURE	Nominal	No / Yes	Vehicle Failure Indicator
TRAIN_TROLLEY	Nominal	No / Yes	Train or Trolley Indicator
PHANTOM_VEHICLE	Nominal	No / Yes	Phantom Vehicle Indicator
ALCOHOL_RELATED	Nominal	No / Yes	Alcohol Related Indicator
DRINKING_DRIVER	Nominal	No / Yes	Drinking Driver Indicator
UNDERAGE_DRNK_DRV	Nominal	No / Yes	Under-age drinking driver Indicator
UNLICENSED	Nominal	No / Yes	Unlicensed Driver Indicator
CELL_PHONE	Nominal	No / Yes	Driver Using Cell Phone Indicator
RUNNING_RED_LT	Nominal	No / Yes	Driver Running Red Light Indicator
TAILGATING	Nominal	No / Yes	Tailgating Indicator
CURVE_DVR_ERROR	Nominal	No / Yes	Curve in Road Driver Error Indicator
SPEEDING	Nominal	No / Yes	Speeding Indicator
SPEEDING_RELATED	Nominal	No / Yes	Speeding Related Indicator
FATIGUE_ASLEEP	Nominal	No / Yes	Fatigue or Asleep Indicator
UNBELTED	Nominal	No / Yes	Anyone in crash unbelted indicator
PEDESTRIAN	Nominal	No / Yes	Pedestrian Indicator
DISTRACTED	Nominal	No / Yes	Distracted Driver Indicator
CURVED_ROAD	Nominal	No / Yes	Curve in Road
MC_DRINKING_DRIVER	Nominal	No / Yes	At least 1 Motorcycle driver has reported or suspected Alcohol Use
INJURY_OR_FATAL	Nominal	No / Yes	At least 1 Person Was Injured or Killed in the Crash
COMM_VEHICLE	Nominal	No / Yes	Crash has at least 1 involved

			Commercial Vehicle
IMPAIRED_DRIVER	Nominal	No / Yes	At least One Driver was Impaired by Drugs or Alcohol
DEER_RELATED	Nominal	No / Yes	Deer Related Indicator
ILLEGAL_DRUG_RELATED	Nominal	No / Yes	At Least 1 Driver or Pedestrian had reported or suspected Illegal Drug Use
ILLUMINATION_DARK	Nominal	No / Yes	Illumination Indicates that the Crash Scene Lighting was Dark
MINOR_INJURY	Nominal	No / Yes	At least 1 Person Sustained a Minor Injury
MODERATE_INJURY	Nominal	No / Yes	At least 1 Person Sustained a Moderate Injury
MAJOR_INJURY	Nominal	No / Yes	At least 1 Person Sustained a Major Injury
NHTSA_AGG_DRIVING	Nominal	No / Yes	The Crash meets the NHTSA definition of Aggressive Driving
RUNNING_STOP_SIGN	Nominal	No / Yes	Driver Running Stop Sign Indicator
TRAIN	Nominal	No / Yes	Train Indicator
TROLLEY	Nominal	No / Yes	Trolley Indicator
MAX_SEVERITY_LEVEL	**Nominal (Class label)**	▪ **Noinjury** ▪ **Injury**	**Injury severity level of the crash**

Class Attribute

"MAX_SEVERITY_LEVEL" attribute

Dataset Description

- Source: Western Pennsylvania Regional Data Center [Link]
  Data is provided by Pennsylvania Department of Transportation (PennDOT).
- The dataset consists of 69 categorical attributes.
- The dataset consists of 11,994 tuples.
- The "MAX_SEVERITY_LEVEL" attribute is used as the class label.

Detailed Description of Mining Tool(s) or Algorithms(s) Used

Data Mining Tools

- Weka

  Weka is a data mining software that uses a collection of machine learning algorithms. These algorithms can be applied directly to the data or called from the Java code. Weka is collection of tools for: Regression, Clustering, Association, Data pre-processing, Classification, Visualization.

- R

  R is a programming language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and is highly extensible.

Data Mining Algorithms

We used 4 classifiers and 5 different feature selection algorithms on the dataset. Finally, we have 20 models with different combination of classifiers and feature selection algorithms)

Feature selection Classifier	**CfsSubset**	**CorrAttr**	**InfoGainAttr**	**GainRatioAttr**	**ClassifierAttrJ48**
IBk	Model 1	Model 2	Model 3	Model 4	Model 5
Naïve Bayes	Model 6	Model 7	Model 8	Model 9	Model 10
Random Forest	Model 11	Model 12	Model 13	Model 14	Model 15
AdaBoostM1+Random Forest	Model 16	Model 17	Model 18	Model 19	Model 20

Classifiers

- K-Nearest Neighbor (IBk)

  In Weka, this algorithm is called IBk (Instance Based Learner). It does not build a model, instead it generates a prediction for a test instance just-in-time. The IBk algorithm uses a distance measure to locate k “close” instances in the training data for each test instance and used those selection instances to make a prediction.

- Naïve Bayes

  Naïve Bayes makes a naive assumption that each feature is independent of other features to make easier for classification.

- Random Forest

  Random Forest is an ensemble learning algorithm that can be used for classification, regression and other tasks. It works by constructing a multitude of decision trees at training time and outputting the predicted class.

- AdaBoostM1 + Random Forest

  AdaBoost is a boosting ensemble model. Boosting model learns from the previous mistakes to improve the accuracy. We used AdaBoostM1 and then run Random Forest algorithm to enhance the ability of prediction.

Feature Selection Algorithms

- CfsSubsetEval (Correlation-based Feature Selection)

  CfsSubsetEval evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.

- CorrelationAttributeEval

  CorrelationAttributeEval evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class.
  Nominal attributes are considered on a value by value basis by treating each value as an indicator. An overall correlation for a nominal attribute is arrived at via a weighted average.

- InfoGainAttributeEval

  InfoGainAttributeEval evaluates the worth of an attribute by measuring the information gain with respect to the class.

- GainRatioAttrEval

  GainRatioAttrEval evaluates the worth of an attribute by measuring the gain ratio with respect to the class.

- ClassifierAttrJ48

  ClassifierAttributeEval evaluates the worth of an attribute by using a user-specified classifier. We used J48 (Decision Tree) as the classifier.

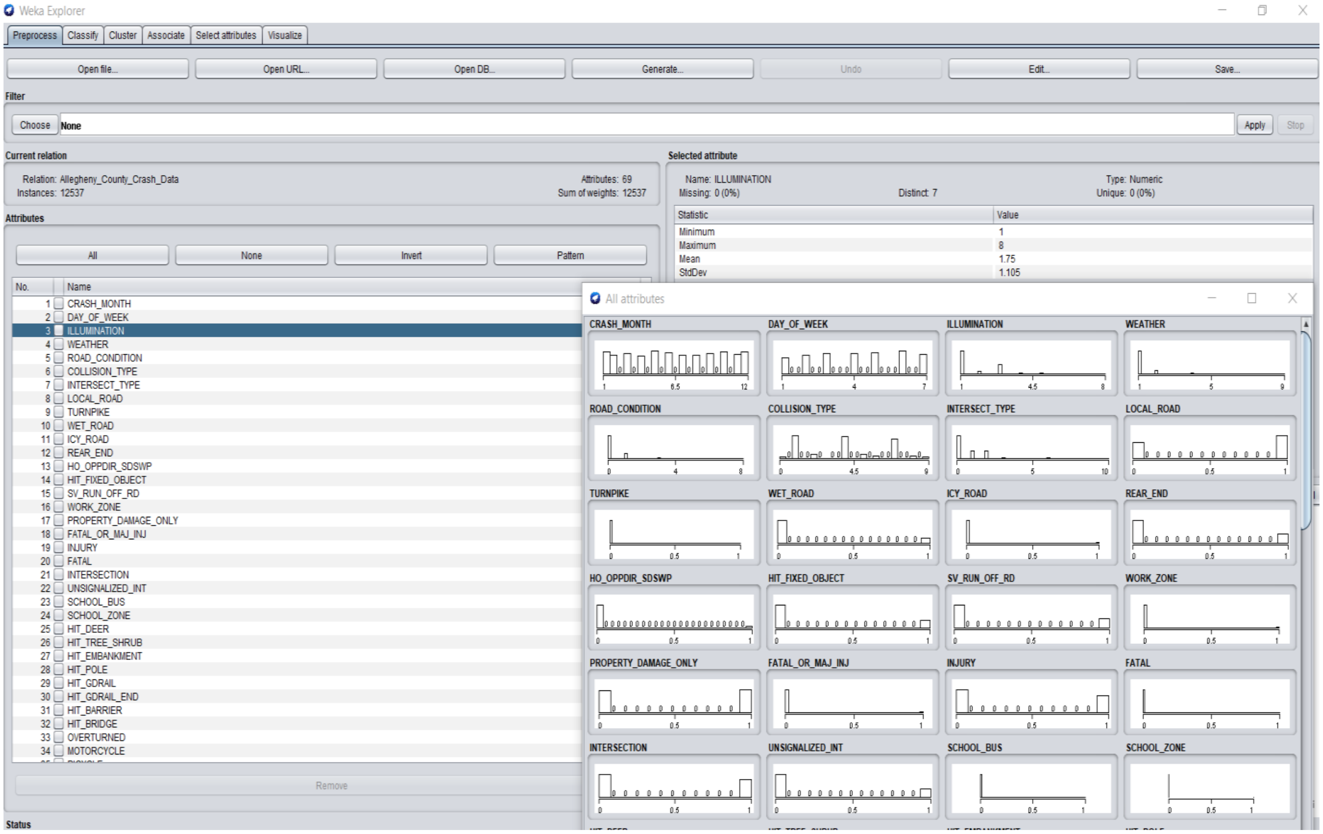
Detailed Description of Data Mining Procedure

Exploratory Data Analysis

We used Weka Explorer to implement exploratory data analysis on the original dataset we got from Western Pennsylvania Regional Data Center.

The original dataset consists 69 attributes, 12,537 instances and a multi-class class attribute. Weka misread the nominal or binominal data to numeric data because of the default setting, so we got a weird result. In order to get the correct information and to prepare for the following classification, we need to do pre-processing first.

File: Allegheny_County_Crash_Data.csv   (Original dataset)



Filter: None

Relation: Allegheny_County_Crash_Data
Instances: 12537
Attributes: 69
Sum of weights: 12537

Name: ILLUMINATION
Missing: 0 (0%)
Distinct: 7
Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	8
Mean	1.75
StdDev	1.105

No.	Name
1	☐ CRASH_MONTH
2	☐ DAY_OF_WEEK
3	☐ ILLUMINATION
4	☐ WEATHER
5	☐ ROAD_CONDITION
6	☐ COLLISION_TYPE
7	☐ INTERSECT_TYPE
8	☐ LOCAL_ROAD
9	☐ TURNPIKE
10	☐ WET_ROAD
11	☐ ICY_ROAD
12	☐ REAR_END
13	☐ HO_OPPDIR_SDSWP
14	☐ HIT_FIXED_OBJECT
15	☐ SV_RUN_OFF_RD
16	☐ WORK_ZONE
17	☐ PROPERTY_DAMAGE_ONLY
18	☐ FATAL_OR_MAJ_INJ
19	☐ INJURY
20	☐ FATAL
21	☐ INTERSECTION
22	☐ UNSIGNALIZED_INT
23	☐ SCHOOL_BUS
24	☐ SCHOOL_ZONE
25	☐ HIT_DEER
26	☐ HIT_TREE_SHRUB
27	☐ HIT_EMBANKMENT
28	☐ HIT_POLE
29	☐ HIT_GDRAIL
30	☐ HIT_GDRAIL_END
31	☐ HIT_BARRIER
32	☐ HIT_BRIDGE
33	☐ OVERTURNED
34	☐ MOTORCYCLE

Figure 1 - The original dataset screenshot in Weka

Pre-Processing

Data Cleaning

- Converting numeric values to nominal values

  We used R to convert all the numeric values to nominal values. For example, we convert 0 to “NO”, 1 to “Yes”

- Missing values

  - Removed Tuples with Illumination = other or unknown (values 8,9)
  - Removed Tuples with Road Conditions = Other
  - Removed Tuples with weather codes = 6,7,8,9 (no information on these codes)
  - Removed Tuples with Collision = (8,9) Other/Unknown
  - Removed Tuples with Intersection = 10 (Other)

Preprocessing with Class Attribute

Class Attribute from the original dataset : Max Severity Level

0 - Not injured
1 - Killed
2 - Major injury
3 - Moderate injury
4 - Minor injury
8 - Injury/ Unknown Severity
9 – Unknown

We deleted unknown classes (8&9) and simplified it to make it a binary class.

Class Attribute from modified dataset : Max Severity Level

- **No injury**
- **Injury** : {Minor injury, Moderate injury, Major injury, Killed}

We compared the original class attribute and the new class attribute below. From Figure 3. We can see the number of “No Injury” (6,256 tuples) is relatively close to that of “Injury” (5,738 tuples), which is a balanced dataset. It is important because balanced dataset won’t need to do further techniques such as over-sampling/under-sampling to deal with biased and inaccurate result.

File: Allegheny_county_crash_data_nominal2.csv (Modified dataset)

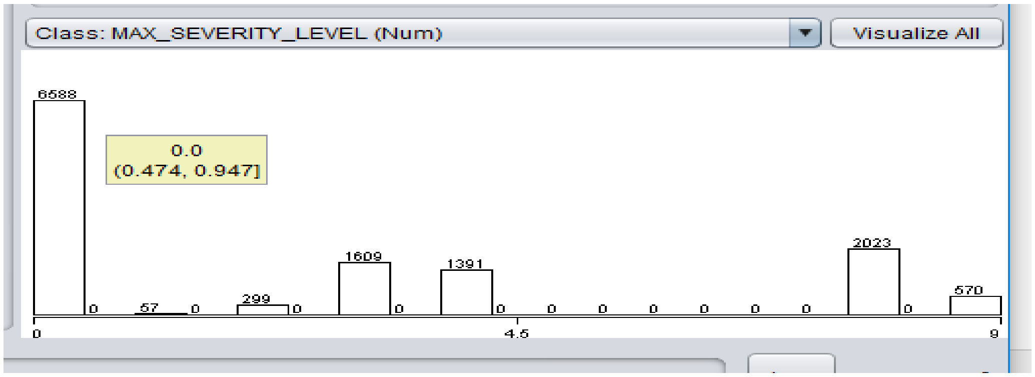


Figure 2 - Original class attribute distribution (from Original dataset)

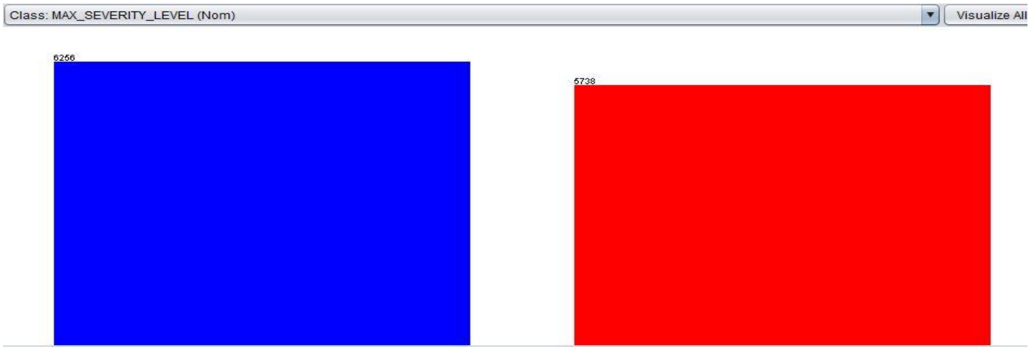


Figure 3 - New class attribute distribution (from modified dataset)

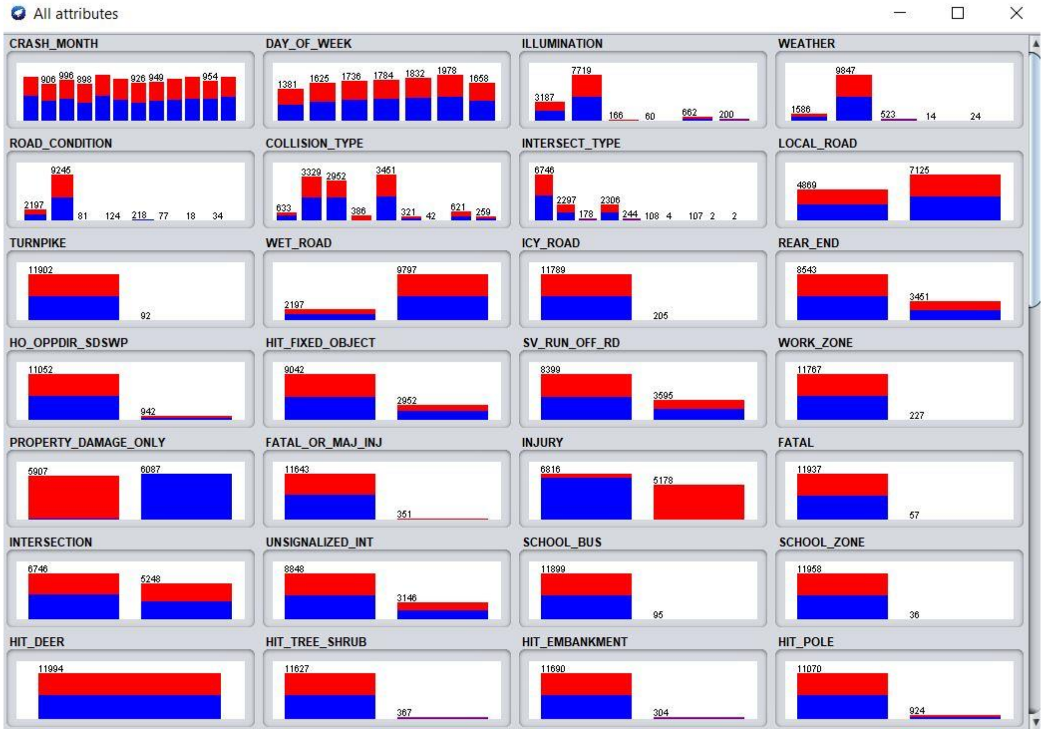


Figure 4 – Part of attribute distribution (from modified dataset)

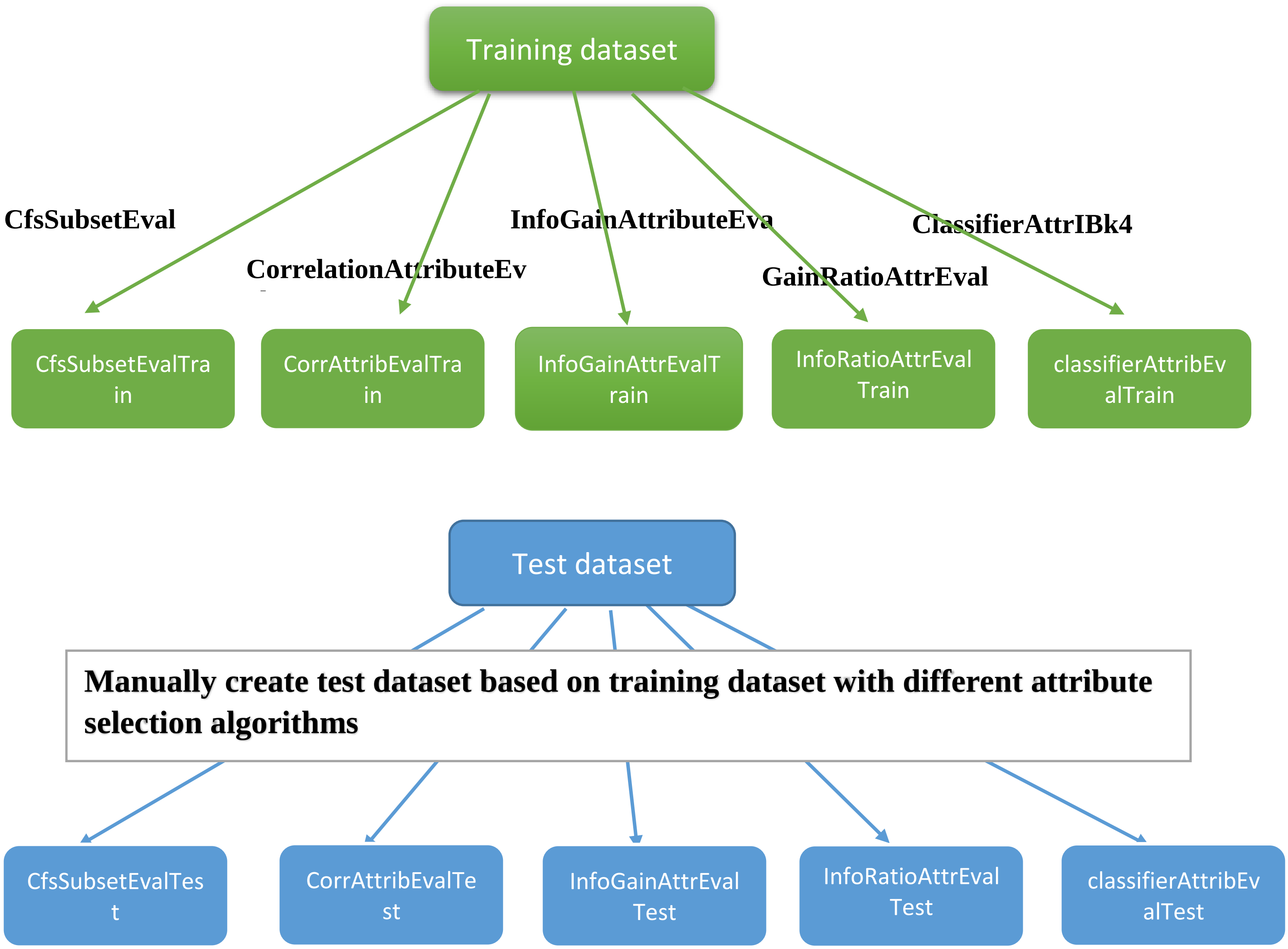
Data Splitting

- We split the new dataset (Allegheny_county_crash_data_nominal2.csv) to training dataset and test dataset.
- Training/Testing stratification based on class followed the Word Document: "how-to-split-dataset-to-training-and-test.docx"
- Training dataset - File: Allegheny_county_crash_data_nominal2_training.arff
- Test dataset - File: Allegheny_county_crash_data_nominal2_test.arff

Feature Selection

We implemented 5 different attribute selection algorithms on training dataset using Weka, and manually selected attributes for test dataset in accordance with the training dataset's result. ***(See Appendix A for list of results of feature selection algorithms on the 1training dataset)***

Workflow



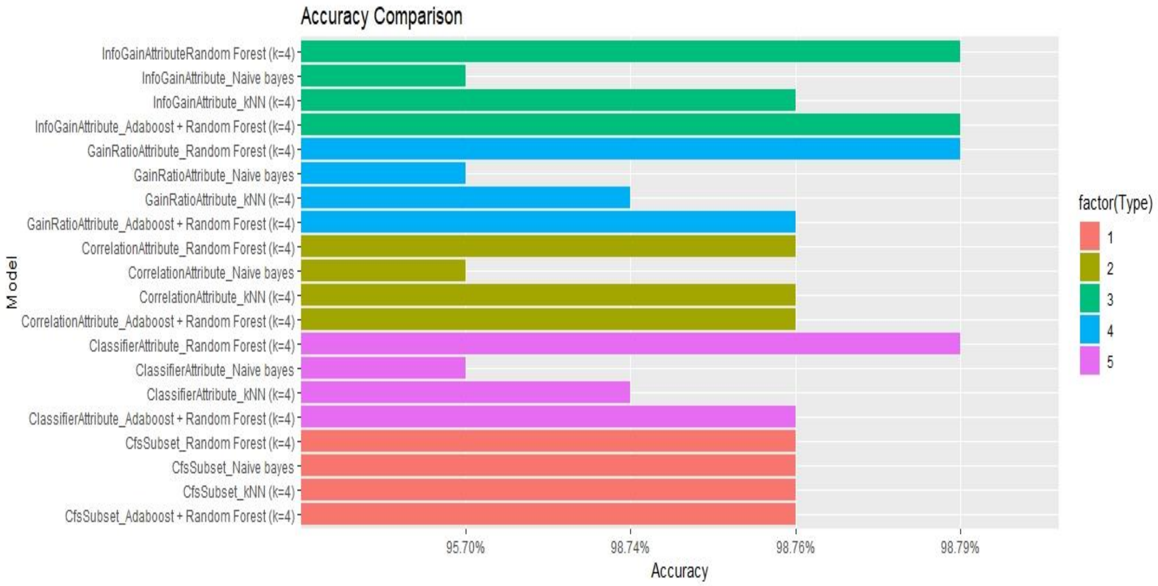
Data Mining Result and Evaluation

We used 5 different test datasets (created with 5 attribute selection algorithms) to validate 20 trained models. ***(See Appendix B for summary data of 20 models)***

Table 1 – Model Performance Measures

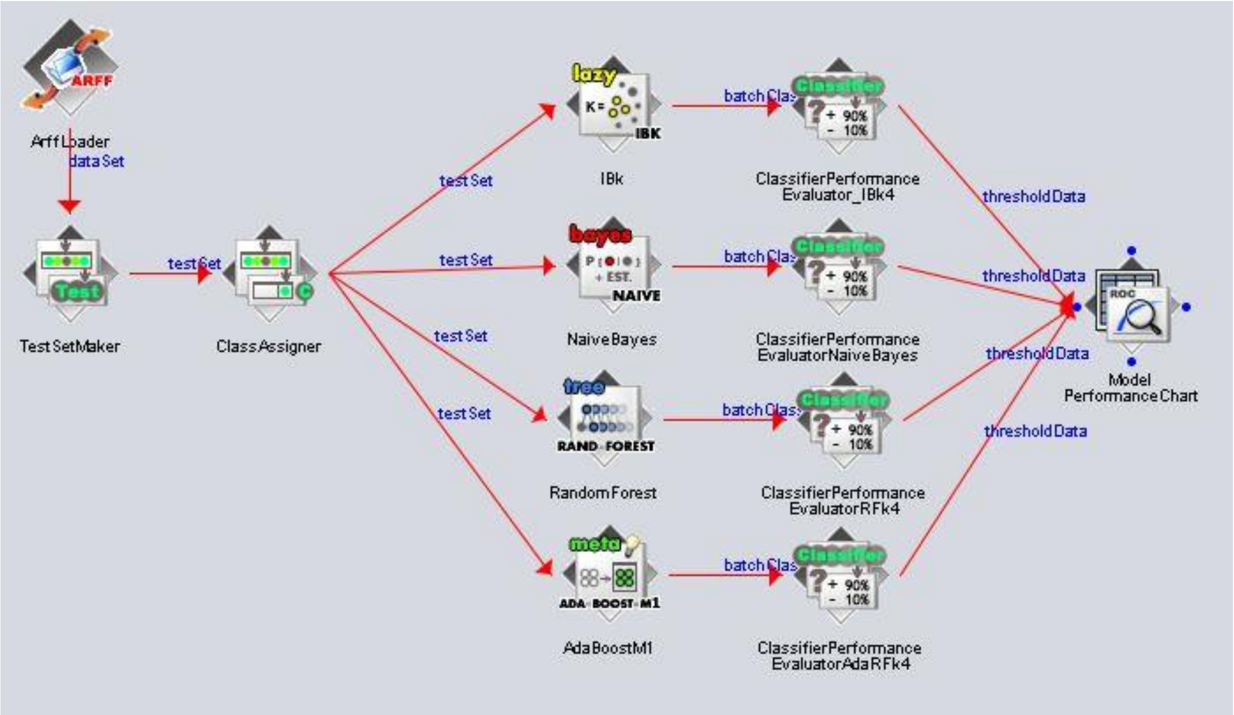
Model	Accuracy	ROC Area	F-Measure	TPR	FPR
CfsSubset_kNN (k=4)	98.76%	0.999	0.988	0.988	0.011
CfsSubset_Naive bayes	98.76%	0.999	0.988	0.988	0.011
CfsSubset_Random Forest (k=4)	98.76%	0.999	0.988	0.988	0.011
CfsSubset_Adaboost + Random Forest (k=4)	98.76%	0.999	0.988	0.988	0.011
CorrelationAttribute_kNN (k=4)	98.76%	0.999	0.988	0.988	0.011
CorrelationAttribute_Naive bayes	95.70%	0.999	0.957	0.957	0.047
CorrelationAttribute_Random Forest (k=4)	98.76%	0.999	0.988	0.988	0.011
CorrelationAttribute_Adaboost + Random Forest (k=4)	98.76%	0.999	0.988	0.988	0.011
InfoGainAttribute_kNN (k=4)	98.76%	0.999	0.988	0.988	0.011
InfoGainAttribute_Naive bayes	95.70%	0.999	0.957	0.957	0.047
InfoGainAttributeRandom Forest (k=4)	98.79%	0.999	0.988	0.988	0.011
InfoGainAttribute_Adaboost + Random Forest (k=4)	98.79%	0.999	0.988	0.988	0.011
GainRatioAttribute_kNN (k=4)	98.74%	0.999	0.987	0.987	0.012
GainRatioAttribute_Naive bayes	95.70%	0.999	0.957	0.957	0.047
GainRatioAttribute_Random Forest (k=4)	98.79%	0.999	0.988	0.988	0.011
GainRatioAttribute_Adaboost + Random Forest (k=4)	98.76%	0.999	0.988	0.988	0.011
ClassifierAttribute_kNN (k=4)	98.74%	0.999	0.987	0.987	0.012
ClassifierAttribute_Naive bayes	95.70%	0.999	0.957	0.957	0.047
ClassifierAttribute_Random Forest (k=4)	98.79%	0.999	0.988	0.988	0.011
ClassifierAttribute_Adaboost + Random Forest (k=4)	98.76%	0.999	0.988	0.988	0.011

Table 2 – Accuracy Comparison



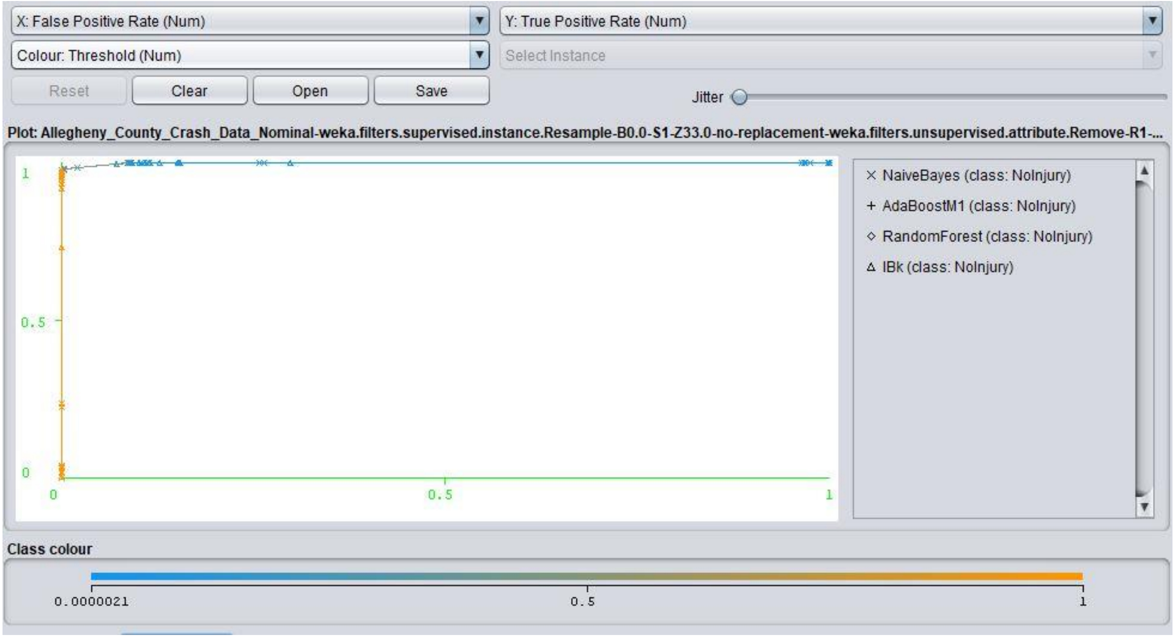
ROC

We used *KnowledgeFlow* in Weka to set up the KF-procedure and generate ROC curves for every model.



Knowledge Flow procedure in Weka

ROC – CfsSubset



X: False Positive Rate (Num)

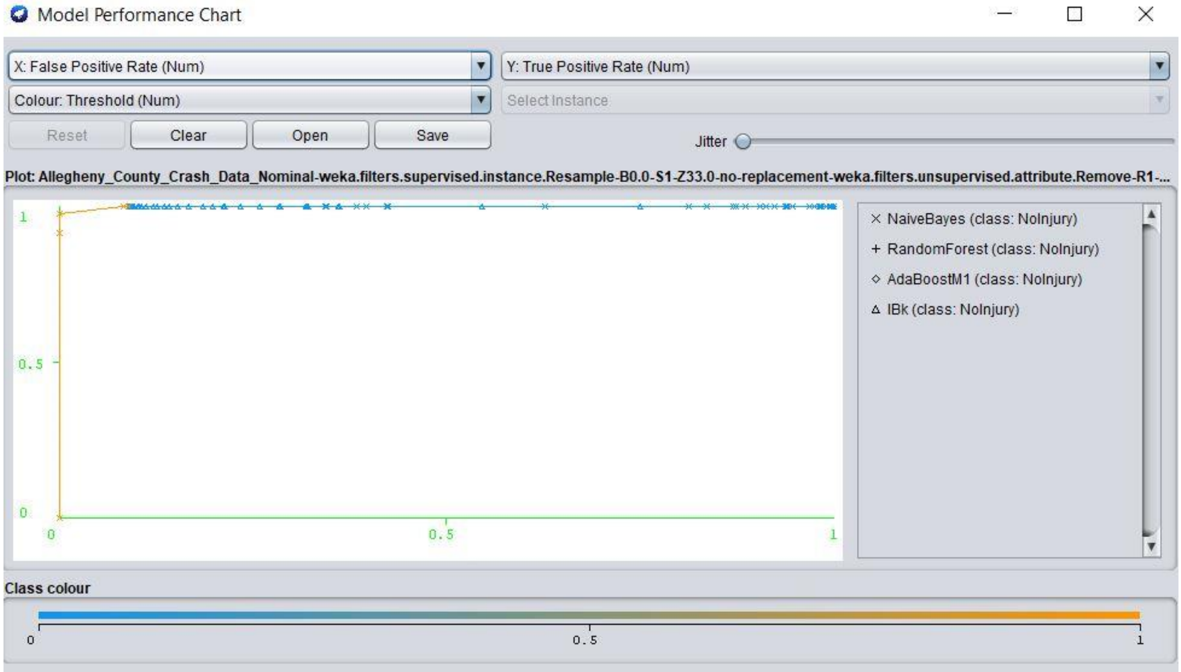
Y: True Positive Rate (Num)

Colour: Threshold (Num)

Plot: Allegheny_County_Crash_Data_Nominal-weka.filters.supervised.instance.Resample-B0.0-S1-Z33.0-no-replacement-weka.filters.unsupervised.attribute.Remove-R1-...

Model 1, Model 6, Model 11, Model 16

ROC – CorrAttribute



Model Performance Chart

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

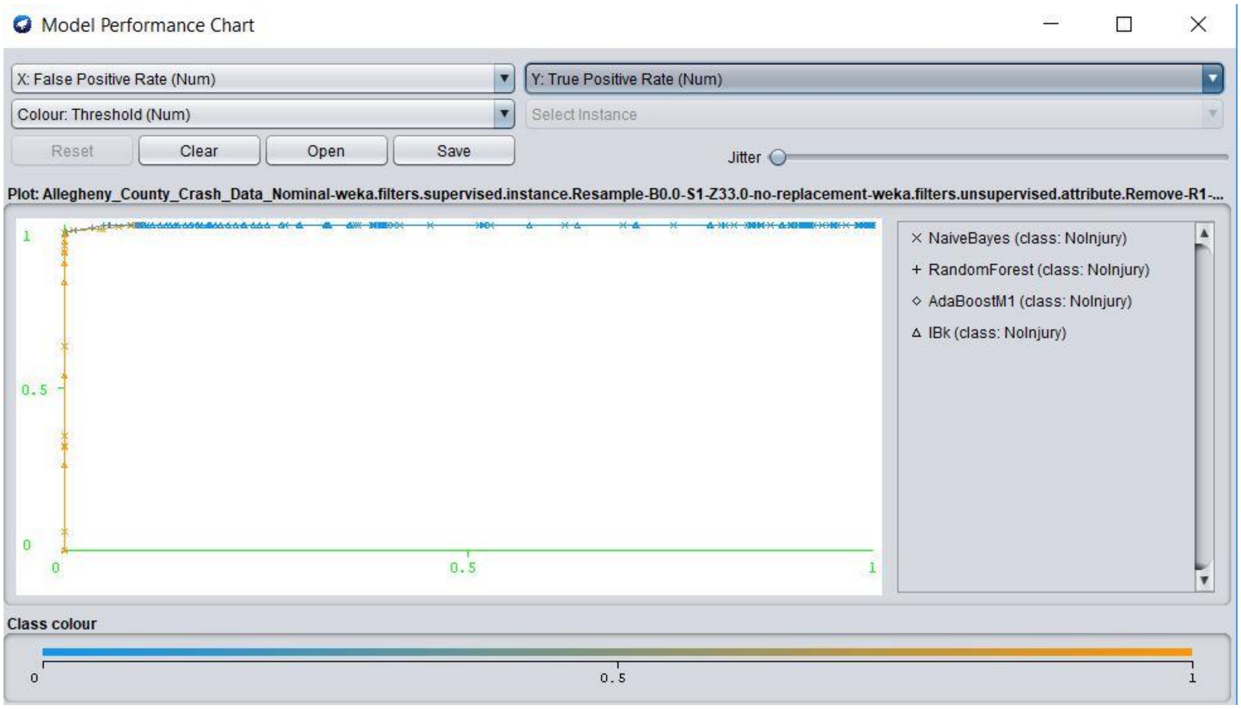
Colour: Threshold (Num)

Plot: Allegheny_County_Crash_Data_Nominal-weka.filters.supervised.instance.Resample-B0.0-S1-Z33.0-no-replacement-weka.filters.unsupervised.attribute.Remove-R1-...

Select Instance
Reset
Clear
Open
Save
Jitter
1
0.5
0
0
0.5
1
× NaiveBayes (class: NoInjury)
+ RandomForest (class: NoInjury)
◇ AdaBoostM1 (class: NoInjury)
△ IBk (class: NoInjury)
Class colour
0
0.5
1

Model 2, Model 7, Model 12, Model 17

ROC – InfoGainAttribute



Model Performance Chart

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot: Allegheny_County_Crash_Data_Nominal-weka.filters.supervised.instance.Resample-B0.0-S1-Z33.0-no-replacement-weka.filters.unsupervised.attribute.Remove-R1-...

× NaiveBayes (class: NoInjury)

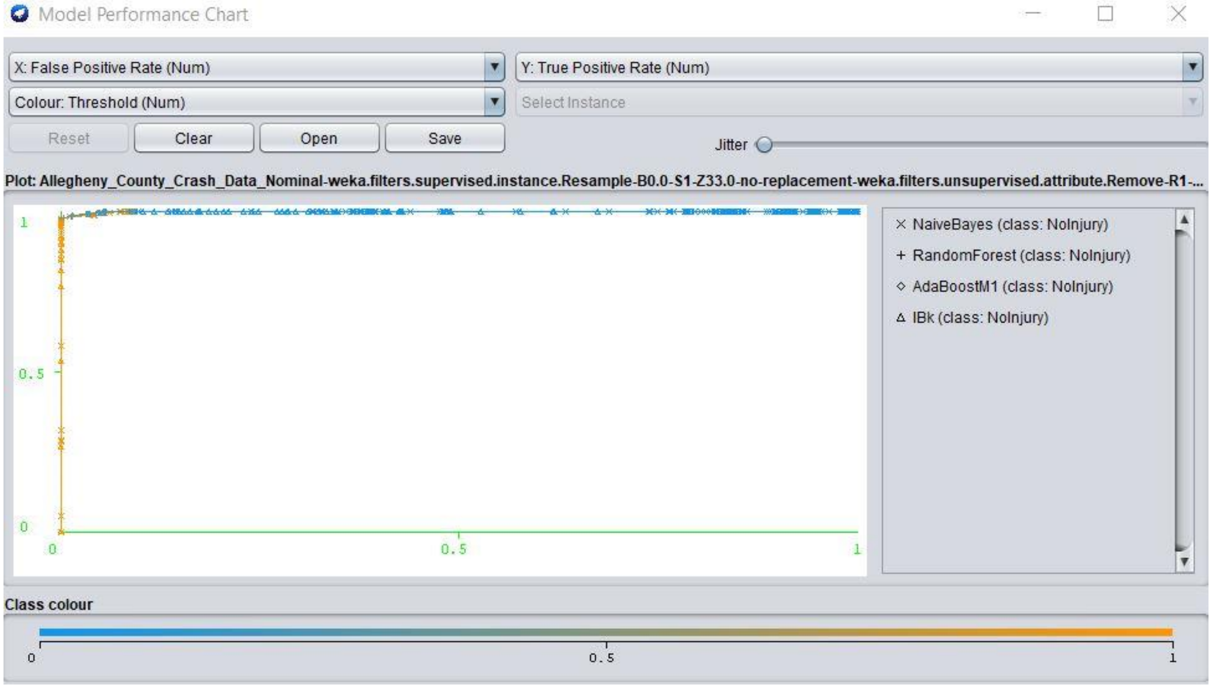
+ RandomForest (class: NoInjury)

◇ AdaBoostM1 (class: NoInjury)

△ IBk (class: NoInjury)

Model 3, Model 8, Model 13, Model 18

ROC – GainRatioAttribute



Model Performance Chart

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot: Allegheny_County_Crash_Data_Nominal-weka.filters.supervised.instance.Resample-B0.0-S1-Z33.0-no-replacement-weka.filters.unsupervised.attribute.Remove-R1-...

× NaiveBayes (class: NoInjury)

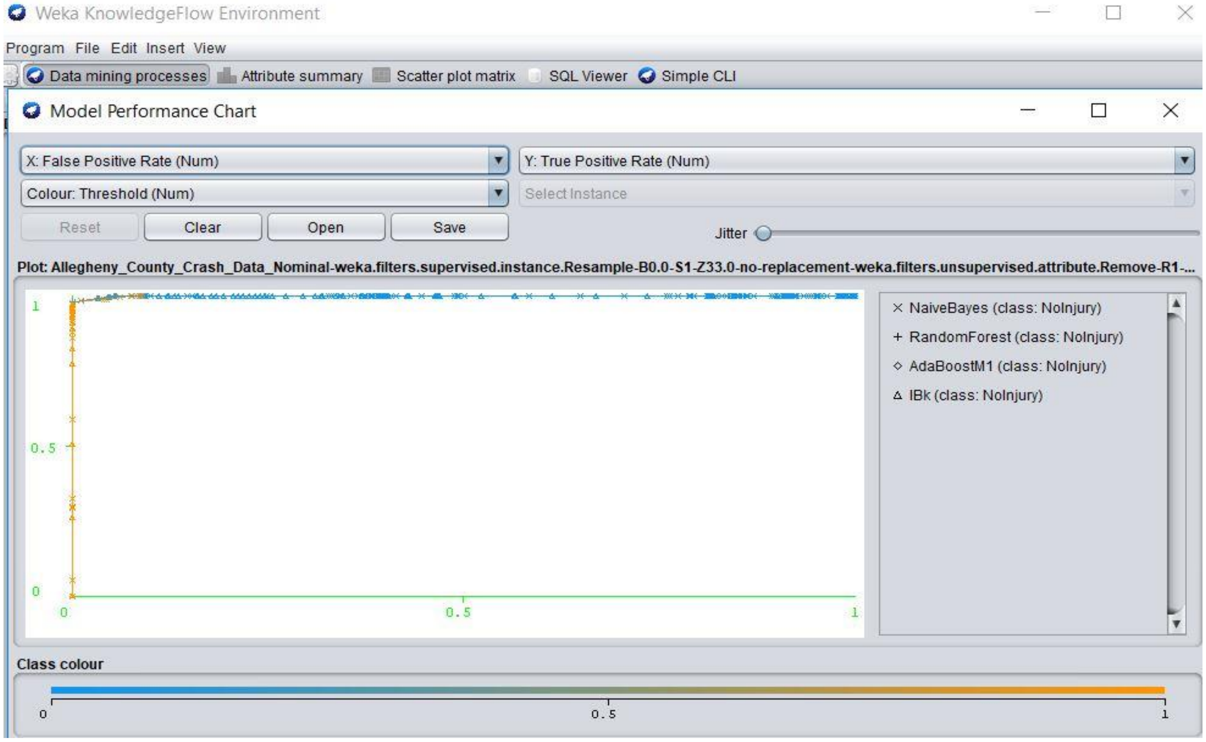
+ RandomForest (class: NoInjury)

◇ AdaBoostM1 (class: NoInjury)

△ IBk (class: NoInjury)

Model 4, Model 9, Model 14, Model 19

ROC – ClassifierAttribute (J48)



X: False Positive Rate (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)

Plot: Allegheny_County_Crash_Data_Nominal-weka.filters.supervised.instance.Resample-B0.0-S1-Z33.0-no-replacement-weka.filters.unsupervised.attribute.Remove-R1-...

× NaiveBayes (class: NoInjury)
+ RandomForest (class: NoInjury)
◇ AdaBoostM1 (class: NoInjury)
△ IBk (class: NoInjury)

Model 5, Model 10, Model 15, Model 20

We decided to focus our comparison of performance on percent accuracy and F-measure (for the weighted average of the classes). These parameters were chosen based upon the binary nature of our model. Likewise, the higher the accuracy, the lower the error rate. Random Forest consistently performed as one of the highest classifiers in terms of percent accuracy and F-measure (the harmonic mean of precision and recall). Random Forest classifiers were also highest when paired with these Selection Attributes: InfoRatio, GainRatio, and ClassifierAttribute in terms of percent accuracy (98.787%) and F-measure (0.988), respectively.

In order to tease out differences in these Attribute Selections for this model, we varied multiple settings in the Random Forest classifier (including k = 0, 2, 3, 4, 16, smaller batch size (n = 25 versus 100), breaks ties randomly, number of iterations from 100 to 1000, data not shown). While none of these changes resulted in a change in value of accuracy or F-measure in the test data sets for InfoRatio, GainRatio, or ClassifierAttribute with Random Forest classification, there were very modest changes in the training data such as 1-3 FPs. Likewise, none of the proposed changes in model parameters resulted in changes in False Negatives (though the models all showed at most 48 FNs out of the total 3957 instances in the Test Set).

Therefore, we evaluated our best model based on the information that would be given with the attributes provided in each selection algorithm. Weka has an output of "Attribute importance based on average impurity decrease (and number of nodes using that attribute)" for

Random Forest. Based on using fewer attributes (10) to achieve the same results for accuracy and F-score, InfoGain, Random Forest, k=4 is our best model for this dataset.

For each selection algorithm, we have the following output:

InfoGain, Random Forest, k = 4

```
weka.classifiers.trees.RandomForest -K 4 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Attribute importance based on average impurity decrease (and number of nodes using that
attribute)

      0.46 (   232)  PROPERTY_DAMAGE_ONLY
      0.33 (    83)  INJURY
      0.3  (   118)  INJURY_OR_FATAL
      0.09 (    32)  MODERATE_INJURY
      0.09 (   104)  COLLISION_TYPE
      0.09 (    22)  MINOR_INJURY
      0.03 (    58)  FATAL_OR_MAJ_INJ
      0.02 (   125)  MOTORCYCLE
      0.01 (    13)  PEDESTRIAN
      0    (     1)  MAJOR_INJURY
```

GainRatio, Random Forest, k = 4

```
weka.classifiers.trees.RandomForest -K 4 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Attribute importance based on average impurity decrease (and number of nodes using that
attribute)

      0.46 (   438)  PROPERTY_DAMAGE_ONLY
      0.37 (   149)  INJURY
      0.37 (   137)  INJURY_OR_FATAL
      0.17 (   165)  COLLISION_TYPE
      0.11 (    54)  MINOR_INJURY
      0.08 (    50)  MODERATE_INJURY
      0.04 (   291)  OVERTURNED
      0.04 (    40)  FATAL
      0.03 (    68)  FATAL_OR_MAJ_INJ
      0.03 (   150)  MOTORCYCLE
      0.02 (    10)  MC_DRINKING_DRIVER
      0.02 (   454)  UNBELTED
      0.01 (    19)  MAJOR_INJURY
      0.01 (    33)  PEDESTRIAN
      0.01 (    23)  BICYCLE
      0    (    13)  TRAIN_TROLLEY
      0    (     5)  TRAIN
```

ClassifierAtt(J48), Random Forest, k =4

```
weka.classifiers.trees.RandomForest -K 4 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Attribute importance based on average impurity decrease (and number of nodes using that
attribute)

      0.47 (   298)  PROPERTY_DAMAGE_ONLY
      0.34 (    91)  INJURY
      0.25 (   140)  INJURY_OR_FATAL
      0.13 (   132)  COLLISION_TYPE
      0.09 (    23)  MINOR_INJURY
      0.06 (    37)  MODERATE_INJURY
      0.03 (    45)  FATAL_OR_MAJ_INJ
      0.02 (   132)  MOTORCYCLE
      0.02 (     9)  PEDESTRIAN
      0.02 (   413)  UNBELTED
      0    (     2)  MAJOR_INJURY
```

Discussion & Conclusion

a) We believe that based on achieving high accuracy and F-score, InfoGain attribute selection with Random Forest classification (k=4) was our best model for classifying No Injury versus Injury. While other models achieved the same results, InfoGain was able to use one less attribute than Classification Attribute (J48). The Classification Attribute showed “UNBELTED” contributing 0.02 to the model; whereas, InfoGain did not have this attribute in its model. One thought is that “UNBELTED” is actually a redundant attribute. Regression analysis could be done to further analyze the two models, if there is numeric representation of this data. One other note, while CfsSubset + Random Forest used fewer attributes (4), the model had lower accuracy and F-score than InfoGain + Random Forest. Therefore, performance alone is not based on the fewest number of attributes, but also the accuracy or error rate (1-accuracy).

b) One of biggest lessons of this project was that class definition can assist in determining how robustly the model classifies the data. If we used more classes, this dataset would have resulted in an unbalanced distribution of classes (skewed right). The trade-off in having a robust model (No Injury versus Injury) was that we lost learning about the granularity within the types of injuries. If we were to do this project again, perhaps we could keep skewed class distribution, and look at other performance measures, such as specificity. Another way of approaching this could be to remove the non-injured class and minor versus major injuries/death. Knowing how to predict major injuries, could result in better preventative actions or local ordinances to reinforce preventative actions.

c) Overall, we were able to predict non-injuries from road accidents from the Allegheny dataset (2004-2018) with high accuracy using InfoGain Attribute Selection and Random Forest. The FPR (FP/N) was also very low (0.011), thus also leading us to conclude that we have a model that will not predict non-injuries as injuries.

d) Clearly state what each team member did

Together: Decided overall performance evaluation criteria and overall strategy.

SL:

- Provided a precursor dataset to the dataset used,
- Cleaned dataset (converted numeric attributes to nominal dataset, removed data that was unknown or other, did the final divide of classes to make binary).
- Split the dataset into training and testing files.
- File: Allegheny_county_crash_data_nominal2_training.arff
  File: Allegheny_county_crash_data_nominal2_test.arff
- Selected Attributes for CfsSubset and CorrelationAttrEval for Models: 1,6,11,16,2,7,12,17.
- Provided R plots to display thresholds for Attributes.
- Performed variation analysis to determine the best Attribute Selection algorithm when three were very similar.
- Wrote Intermediate Report.
- Wrote report from Data Mining Result and Evaluation to Conclusion.

TK:

- Found the dataset for this study.
- Wrote project proposal including statement of data mining, description of dataset.
- Described each attribute from the original dataset.
- Generated KF-Procedure and all ROC diagrams
- Selected Attributes for GainRatio, InfoGain, and ClassifierAttr(J48) for Models: 3,8,13,18,4,9,14,19,5,10,15,20,
- Wrote Preliminary Report,
- Created Performance Criteria plot and table,
- Wrote report up to Data Mining Result and Evaluation.
- Generated Appendices
- Organized dataset files & folders

Appendix A

Feature Selection on Training Dataset using Weka Explorer

- **CfsSubset**

  On Allegheny_county_crash_data_nominal2_training.arff, in Weka used "Select Attributes" tab and set:
  - Attribute Evaluator: CfsSubsetEval -P 1 -E 1
  - Search Method: BestFirst -D 1 -N 5

  File: CfsSubsetEvalTrain.arff (training dataset)
  File: CfsSubsetEvalTest.arff (test dataset)

```
=== Run information ===
Evaluator:    weka.attributeSelection.CfsSubsetEval -P 1 -E 1
Search:       weka.attributeSelection.BestFirst -D 1 -N 5
Relation:     Allegheny_County_Crash_Data_Nominal-weka.filters.supervised.instance.Resample-B0.0-S1-
Z33.0-no-replacement-V
Instances:    8037
Attributes:   69
              CRASH_MONTH
              DAY_OF_WEEK
              ILLUMINATION
              WEATHER
              ROAD_CONDITION
              COLLISION_TYPE
              INTERSECT_TYPE
              LOCAL_ROAD
              TURNPIKE
              WET_ROAD
              ICY_ROAD
              REAR_END
              HO_OPPDIR_SDSWP
              HIT_FIXED_OBJECT
              SV_RUN_OFF_RD
              WORK_ZONE
              PROPERTY_DAMAGE_ONLY
              FATAL_OR_MAJ_INJ
              INJURY
              FATAL
              INTERSECTION
              UNSIGNALIZED_INT
              SCHOOL_BUS
              SCHOOL_ZONE
              HIT_DEER
              HIT_TREE_SHRUB
              HIT_EMBANKMENT
              HIT_POLE
              HIT_GDRAIL
              HIT_GDRAIL_END
```

```
        HIT_BARRIER
        HIT_BRIDGE
        OVERTURNED
        MOTORCYCLE
        BICYCLE
        HVY_TRUCK_RELATED
        VEHICLE_FAILURE
        TRAIN_TROLLEY
        PHANTOM_VEHICLE
        ALCOHOL_RELATED
        DRINKING_DRIVER
        UNDERAGE_DRNK_DRV
        UNLICENSED
        CELL_PHONE
        RUNNING_RED_LT
        TAILGATING
        CURVE_DVR_ERROR
        SPEEDING
        SPEEDING_RELATED
        FATIGUE_ASLEEP
        UNBELTED
        PEDESTRIAN
        DISTRACTED
        CURVED_ROAD
        MC_DRINKING_DRIVER
        INJURY_OR_FATAL
        COMM_VEHICLE
        IMPAIRED_DRIVER
        DEER_RELATED
        ILLEGAL_DRUG_RELATED
        ILLUMINATION_DARK
        MINOR_INJURY
        MODERATE_INJURY
        MAJOR_INJURY
        NHTSA_AGG_DRIVING
        RUNNING_STOP_SIGN
        TRAIN
        TROLLEY
        MAX_SEVERITY_LEVEL
Evaluation mode:    evaluate on all training data
=== Attribute Selection on all input data ===
Search Method:
    Best first.
    Start set: no attributes
    Search direction: forward
    Stale search after 5 node expansions
    Total number of subsets evaluated: 409
    Merit of best subset found:    0.905

Attribute Subset Evaluator (supervised, Class (nominal): 69 MAX_SEVERITY_LEVEL):
    CFS Subset Evaluator
    Including locally predictive attributes

Selected attributes: 5,17,33,56 : 4
            ROAD_CONDITION
            PROPERTY_DAMAGE_ONLY
```

```
          OVERTURNED
          INJURY_OR_FATAL
```

- **CorrelationAttribute**

  On Allegheny_county_crash_data_nominal2_training.arff, in Weka used “Select Attributes” tab and set:

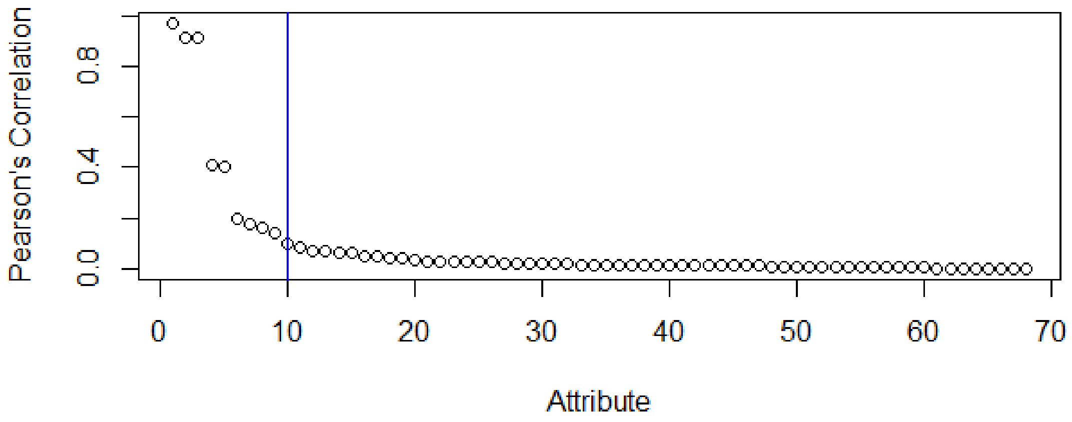
  - Attribute Evaluator: CorrelationAttributeEval
  - Search Method: Ranker -T 0.1 -N -1 (threshold > 0.1)

  File: CorrelAttribEval10Train.arff (training dataset)
  File: CorrelAttribEval10Test.arff (test dataset)

```
=== Run information ===
Evaluator:    weka.attributeSelection.CorrelationAttributeEval
Search:       weka.attributeSelection.Ranker -T 0.1 -N -1
Relation:     Allegheny_County_Crash_Data_Nominal-weka.filters.supervised.instance.Resample-B0.0-S1-
Z33.0-no-replacement-V
Instances:    8037
Attributes:   69
              CRASH_MONTH
              DAY_OF_WEEK
              ILLUMINATION
              WEATHER
              ROAD_CONDITION
              COLLISION_TYPE
              INTERSECT_TYPE
              LOCAL_ROAD
              TURNPIKE
              WET_ROAD
              ICY_ROAD
              REAR_END
              HO_OPPDIR_SDSWP
              HIT_FIXED_OBJECT
              SV_RUN_OFF_RD
              WORK_ZONE
              PROPERTY_DAMAGE_ONLY
              FATAL_OR_MAJ_INJ
              INJURY
              FATAL
              INTERSECTION
              UNSIGNALIZED_INT
              SCHOOL_BUS
              SCHOOL_ZONE
              HIT_DEER
              HIT_TREE_SHRUB
              HIT_EMBANKMENT
              HIT_POLE
              HIT_GDRAIL
              HIT_GDRAIL_END
              HIT_BARRIER
              HIT_BRIDGE
              OVERTURNED
              MOTORCYCLE
              BICYCLE
              HVY_TRUCK_RELATED
              VEHICLE_FAILURE
              TRAIN_TROLLEY
              PHANTOM_VEHICLE
              ALCOHOL_RELATED
              DRINKING_DRIVER
              UNDERAGE_DRNK_DRV
              UNLICENSED
              CELL_PHONE
              RUNNING_RED_LT
              TAILGATING
              CURVE_DVR_ERROR
              SPEEDING
              SPEEDING_RELATED
              FATIGUE_ASLEEP
              UNBELTED
              PEDESTRIAN
              DISTRACTED
              CURVED_ROAD
              MC_DRINKING_DRIVER
              INJURY_OR_FATAL
              COMM_VEHICLE
              IMPAIRED_DRIVER
              DEER_RELATED
              ILLEGAL_DRUG_RELATED
              ILLUMINATION_DARK
              MINOR_INJURY
              MODERATE_INJURY
              MAJOR_INJURY
              NHTSA_AGG_DRIVING
              RUNNING_STOP_SIGN
              TRAIN
              TROLLEY
              MAX_SEVERITY_LEVEL
Evaluation mode:    evaluate on all training data
=== Attribute Selection on all input data ===
Search Method:
              Attribute ranking.
              Threshold for discarding attributes:   0.1

Attribute Evaluator (supervised, Class (nominal): 69 MAX_SEVERITY_LEVEL):
              Correlation Ranking Filter
Ranked attributes:
 0.971    17 PROPERTY_DAMAGE_ONLY
 0.915    56 INJURY_OR_FATAL
 0.909    19 INJURY
 0.409    63 MODERATE_INJURY
 0.403    62 MINOR_INJURY
 0.197    52 PEDESTRIAN
 0.18     18 FATAL_OR_MAJ_INJ
```

```
0.166    64 MAJOR_INJURY
0.139    34 MOTORCYCLE
0.103    51 UNBELTED
```

The threshold for discarding attributes was selected based upon plotting sorted the Pearson Coefficients and looking to see where the inflection point was on the plot (see plot below).



- **InfoGainAttribute**

On Allegheny_county_crash_data_nominal2_training.arff, in Weka used "Select Attributes" tab and set:

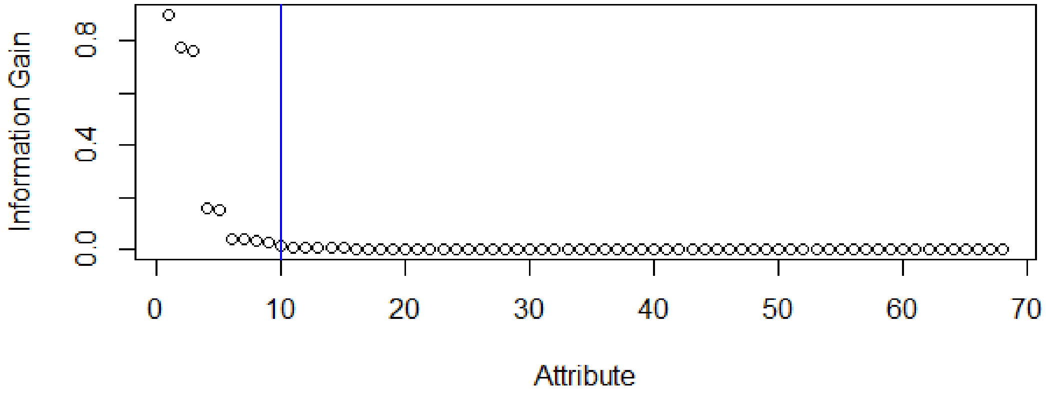
  - Attribute Evaluator: InfoGainAttributeEval
  - Search Method: Ranker -T 0.01 -N -1 (threshold > 0.01)

File: InfoGainAttribEvalTrain.arff (training dataset)
File: InfoGainAttribEvalTest.arff (test dataset)

```
=== Run information ===
Evaluator:    weka.attributeSelection.InfoGainAttributeEval
Search:       weka.attributeSelection.Ranker -T 0.01 -N -1
Relation:     Allegheny_County_Crash_Data_Nominal-weka.filters.supervised.instance.Resample-B0.0-S1-
Z33.0-no-replacement-V
Instances:    8037
Attributes:   69
              CRASH_MONTH
              DAY_OF_WEEK
              ILLUMINATION
              WEATHER
              ROAD_CONDITION
              COLLISION_TYPE
              INTERSECT_TYPE
              LOCAL_ROAD
```

TURNPIKE
WET_ROAD
ICY_ROAD
REAR_END
HO_OPPDIR_SDSWP
HIT_FIXED_OBJECT
SV_RUN_OFF_RD
WORK_ZONE
PROPERTY_DAMAGE_ONLY
FATAL_OR_MAJ_INJ
INJURY
FATAL
INTERSECTION
UNSIGNALIZED_INT
SCHOOL_BUS
SCHOOL_ZONE
HIT_DEER
HIT_TREE_SHRUB
HIT_EMBANKMENT
HIT_POLE
HIT_GDRAIL
HIT_GDRAIL_END
HIT_BARRIER
HIT_BRIDGE
OVERTURNED
MOTORCYCLE
BICYCLE
HVY_TRUCK_RELATED
VEHICLE_FAILURE
TRAIN_TROLLEY
PHANTOM_VEHICLE
ALCOHOL_RELATED
DRINKING_DRIVER
UNDERAGE_DRNK_DRV
UNLICENSED
CELL_PHONE
RUNNING_RED_LT
TAILGATING
CURVE_DVR_ERROR
SPEEDING
SPEEDING_RELATED
FATIGUE_ASLEEP
UNBELTED
PEDESTRIAN
DISTRACTED
CURVED_ROAD
MC_DRINKING_DRIVER
INJURY_OR_FATAL
COMM_VEHICLE
IMPAIRED_DRIVER
DEER_RELATED
ILLEGAL_DRUG_RELATED
ILLUMINATION_DARK
MINOR_INJURY
MODERATE_INJURY
MAJOR_INJURY

```
              NHTSA_AGG_DRIVING
              RUNNING_STOP_SIGN
              TRAIN
              TROLLEY
              MAX_SEVERITY_LEVEL
Evaluation mode:    evaluate on all training data
=== Attribute Selection on all input data ===
Search Method:
              Attribute ranking.
              Threshold for discarding attributes:   0.01

Attribute Evaluator (supervised, Class (nominal): 69 MAX_SEVERITY_LEVEL):
              Information Gain Ranking Filter

Ranked attributes:
 0.9021    17 PROPERTY_DAMAGE_ONLY
 0.7754    56 INJURY_OR_FATAL
 0.7646    19 INJURY
 0.1575    63 MODERATE_INJURY
 0.1536    62 MINOR_INJURY
 0.0415     6 COLLISION_TYPE
 0.037     52 PEDESTRIAN
 0.0314    18 FATAL_OR_MAJ_INJ
 0.0266    64 MAJOR_INJURY
 0.0167    34 MOTORCYCLE

Selected attributes: 17,56,19,63,62,6,52,18,64,34 : 10
```



- **GainRatioAttribute**

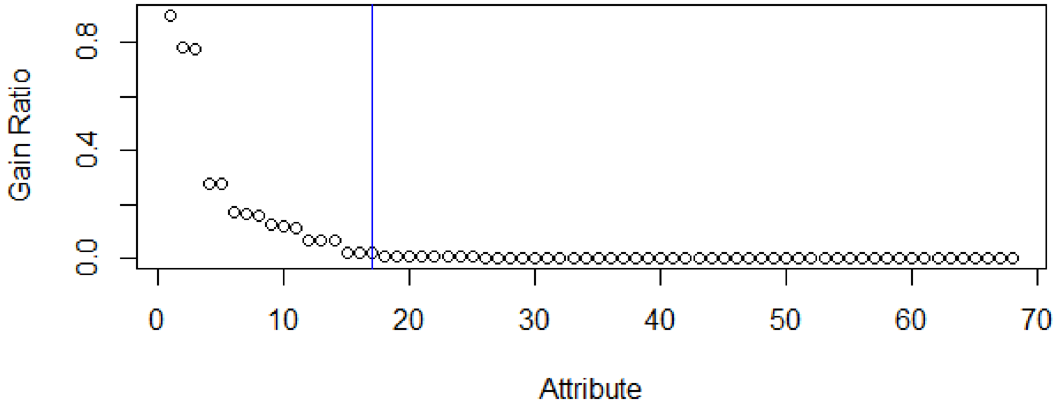
On Allegheny_county_crash_data_nominal2_training.arff, in Weka used “Select Attributes” tab and set:

  - Attribute Evaluator: GainRatioAttributeEval
  - Search Method: Ranker -T 0.01 -N -1 (threshold > 0.01)

File: GainRatioAttribEvalTrain.arff (training dataset)
File: GainRatioAttribEvalTest.arff (test dataset)

```
=== Run information ===
Evaluator:    weka.attributeSelection.GainRatioAttributeEval
Search:       weka.attributeSelection.Ranker -T 0.01 -N -1
Relation:     Allegheny_County_Crash_Data_Nominal-weka.filters.supervised.instance.Resample-B0.0-S1-
Z33.0-no-replacement-V
Instances:    8037
Attributes:   69
              CRASH_MONTH
              DAY_OF_WEEK
              ILLUMINATION
              WEATHER
              ROAD_CONDITION
              COLLISION_TYPE
              INTERSECT_TYPE
              LOCAL_ROAD
              TURNPIKE
              WET_ROAD
              ICY_ROAD
              REAR_END
              HO_OPPDIR_SDSWP
              HIT_FIXED_OBJECT
              SV_RUN_OFF_RD
              WORK_ZONE
              PROPERTY_DAMAGE_ONLY
              FATAL_OR_MAJ_INJ
              INJURY
              FATAL
              INTERSECTION
              UNSIGNALIZED_INT
              SCHOOL_BUS
              SCHOOL_ZONE
              HIT_DEER
              HIT_TREE_SHRUB
              HIT_EMBANKMENT
              HIT_POLE
              HIT_GDRAIL
              HIT_GDRAIL_END
              HIT_BARRIER
              HIT_BRIDGE
              OVERTURNED
              MOTORCYCLE
              BICYCLE
              HVY_TRUCK_RELATED
              VEHICLE_FAILURE
              TRAIN_TROLLEY
              PHANTOM_VEHICLE
              ALCOHOL_RELATED
              DRINKING_DRIVER
              UNDERAGE_DRNK_DRV
              UNLICENSED
              CELL_PHONE
              RUNNING_RED_LT
              TAILGATING
              CURVE_DVR_ERROR
              SPEEDING
              SPEEDING_RELATED
              FATIGUE_ASLEEP
              UNBELTED
              PEDESTRIAN
              DISTRACTED
              CURVED_ROAD
              MC_DRINKING_DRIVER
              INJURY_OR_FATAL
              COMM_VEHICLE
              IMPAIRED_DRIVER
              DEER_RELATED
              ILLEGAL_DRUG_RELATED
              ILLUMINATION_DARK
              MINOR_INJURY
              MODERATE_INJURY
              MAJOR_INJURY
              NHTSA_AGG_DRIVING
              RUNNING_STOP_SIGN
              TRAIN
              TROLLEY
              MAX_SEVERITY_LEVEL
Evaluation mode:    evaluate on all training data

=== Attribute Selection on all input data ===
Search Method:
              Attribute ranking.
              Threshold for discarding attributes:   0.01

Attribute Evaluator (supervised, Class (nominal): 69 MAX_SEVERITY_LEVEL):
              Gain Ratio feature evaluator

Ranked attributes:
 0.9022    17 PROPERTY_DAMAGE_ONLY
 0.7852    56 INJURY_OR_FATAL
 0.7751    19 INJURY
 0.2786    63 MODERATE_INJURY
 0.2756    62 MINOR_INJURY
 0.169     52 PEDESTRIAN
 0.1663    18 FATAL_OR_MAJ_INJ
 0.1599    64 MAJOR_INJURY
 0.1227    35 BICYCLE
 0.1179    20 FATAL
 0.1115    34 MOTORCYCLE
 0.0691    55 MC_DRINKING_DRIVER
```

```
0.0651    38 TRAIN_TROLLEY
0.0651    67 TRAIN
0.0197    33 OVERTURNED
0.0173    51 UNBELTED
0.0171     6 COLLISION_TYPE

Selected attributes: 17,56,19,63,62,52,18,64,35,20,34,55,38,67,33,51,6 : 17
```



- **ClassifierAttribute**

On Allegheny_county_crash_data_nominal2_training.arff, in Weka used "Select Attributes" tab and set:

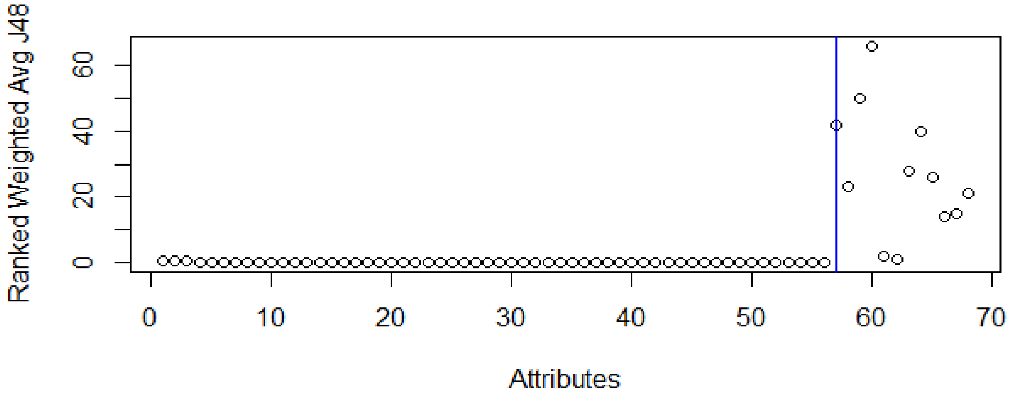
- Attribute Evaluator: ClassifierAttributeEval J48 -C 0.25 -M 2 (classifier: J48 decision tree)
- Search Method: Ranker -T 0.01 -N -1 (threshold > 0.01)

File: ClassifierAttribEvalTrain.arff (training dataset)
File: ClassifierAttribEvalTest.arff (test dataset)

```
=== Run information ===
Evaluator:    weka.attributeSelection.ClassifierAttributeEval -execution-slots 1 -B weka.classifiers.trees.J48
-F 5 -T 0.01 -R 1 -E DEFAULT -- -C 0.25 -M 2
Search:       weka.attributeSelection.Ranker -T 0.01 -N -1
Relation:     Allegheny_County_Crash_Data_Nominal-weka.filters.supervised.instance.Resample-B0.0-S1-
Z33.0-no-replacement-V
Instances:    8037
Attributes:   69
              CRASH_MONTH
              DAY_OF_WEEK
              ILLUMINATION
              WEATHER
```

ROAD_CONDITION
COLLISION_TYPE
INTERSECT_TYPE
LOCAL_ROAD
TURNPIKE
WET_ROAD
ICY_ROAD
REAR_END
HO_OPPDIR_SDSWP
HIT_FIXED_OBJECT
SV_RUN_OFF_RD
WORK_ZONE
PROPERTY_DAMAGE_ONLY
FATAL_OR_MAJ_INJ
INJURY
FATAL
INTERSECTION
UNSIGNALIZED_INT
SCHOOL_BUS
SCHOOL_ZONE
HIT_DEER
HIT_TREE_SHRUB
HIT_EMBANKMENT
HIT_POLE
HIT_GDRAIL
HIT_GDRAIL_END
HIT_BARRIER
HIT_BRIDGE
OVERTURNED
MOTORCYCLE
BICYCLE
HVY_TRUCK_RELATED
VEHICLE_FAILURE
TRAIN_TROLLEY
PHANTOM_VEHICLE
ALCOHOL_RELATED
DRINKING_DRIVER
UNDERAGE_DRNK_DRV
UNLICENSED
CELL_PHONE
RUNNING_RED_LT
TAILGATING
CURVE_DVR_ERROR
SPEEDING
SPEEDING_RELATED
FATIGUE_ASLEEP
UNBELTED
PEDESTRIAN
DISTRACTED
CURVED_ROAD
MC_DRINKING_DRIVER
INJURY_OR_FATAL
COMM_VEHICLE
IMPAIRED_DRIVER
DEER_RELATED
ILLEGAL_DRUG_RELATED

```
              ILLUMINATION_DARK
              MINOR_INJURY
              MODERATE_INJURY
              MAJOR_INJURY
              NHTSA_AGG_DRIVING
              RUNNING_STOP_SIGN
              TRAIN
              TROLLEY
              MAX_SEVERITY_LEVEL
Evaluation mode:    evaluate on all training data

=== Attribute Selection on all input data ===
Search Method:
              Attribute ranking.
              Threshold for discarding attributes:   0.01

Attribute Evaluator (supervised, Class (nominal): 69 MAX_SEVERITY_LEVEL):
              Classifier feature evaluator

              Using     Wrapper Subset Evaluator
              Learning scheme: weka.classifiers.trees.J48
              Scheme options: -C 0.25 -M 2
              Subset evaluation: classification accuracy
              Number of folds for accuracy estimation: 5

Ranked attributes:
 0.4635    17 PROPERTY_DAMAGE_ONLY
 0.4344    56 INJURY_OR_FATAL
 0.4314    19 INJURY
 0.1329    63 MODERATE_INJURY
 0.1299    62 MINOR_INJURY
 0.0383     6 COLLISION_TYPE
 0.0347    52 PEDESTRIAN
 0.0289    18 FATAL_OR_MAJ_INJ
 0.0256    51 UNBELTED
 0.0245    64 MAJOR_INJURY
 0.0193    34 MOTORCYCLE

Selected attributes: 17,56,19,63,62,6,52,18,51,64,34 : 11
```

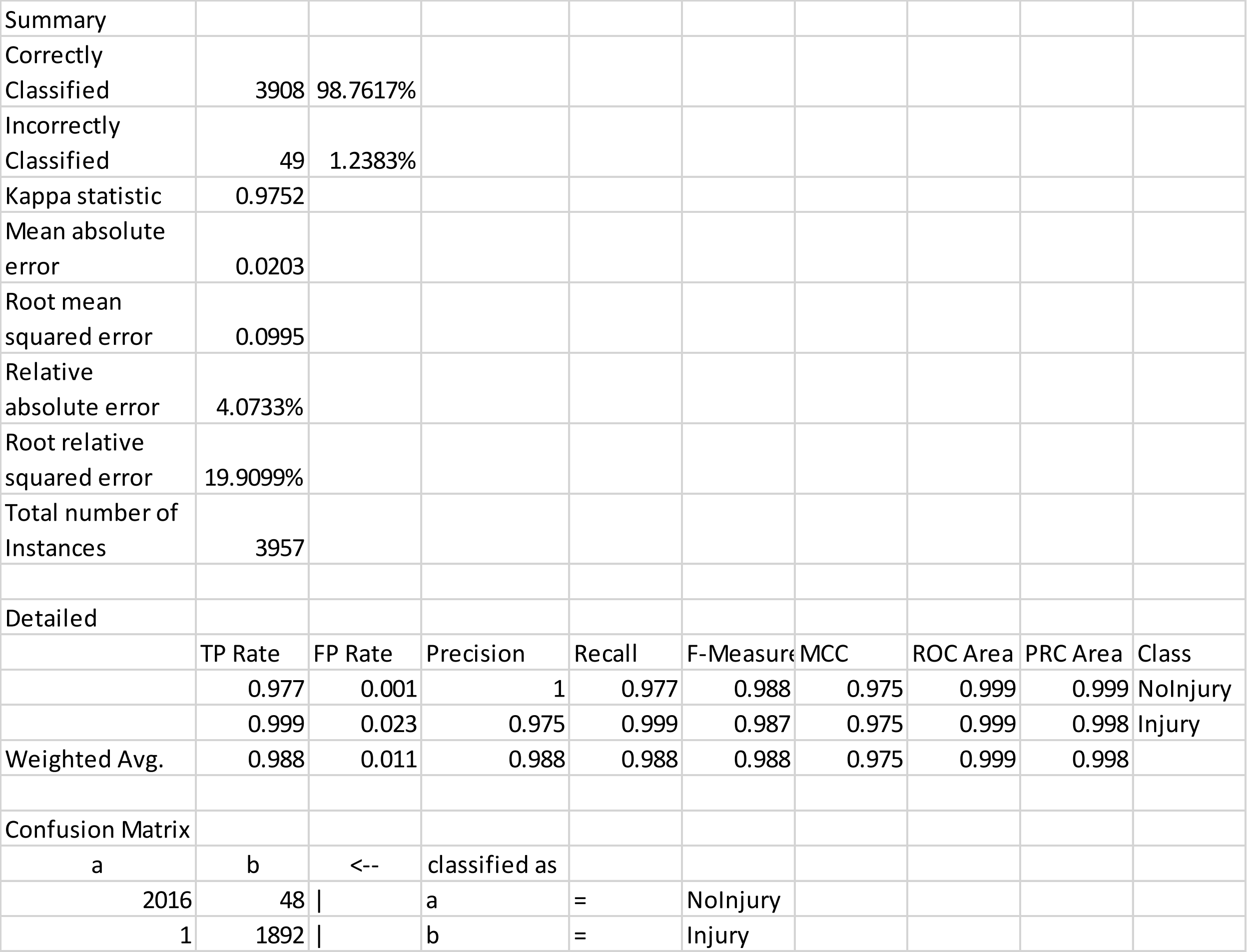


Appendix B

Model's Summary Data

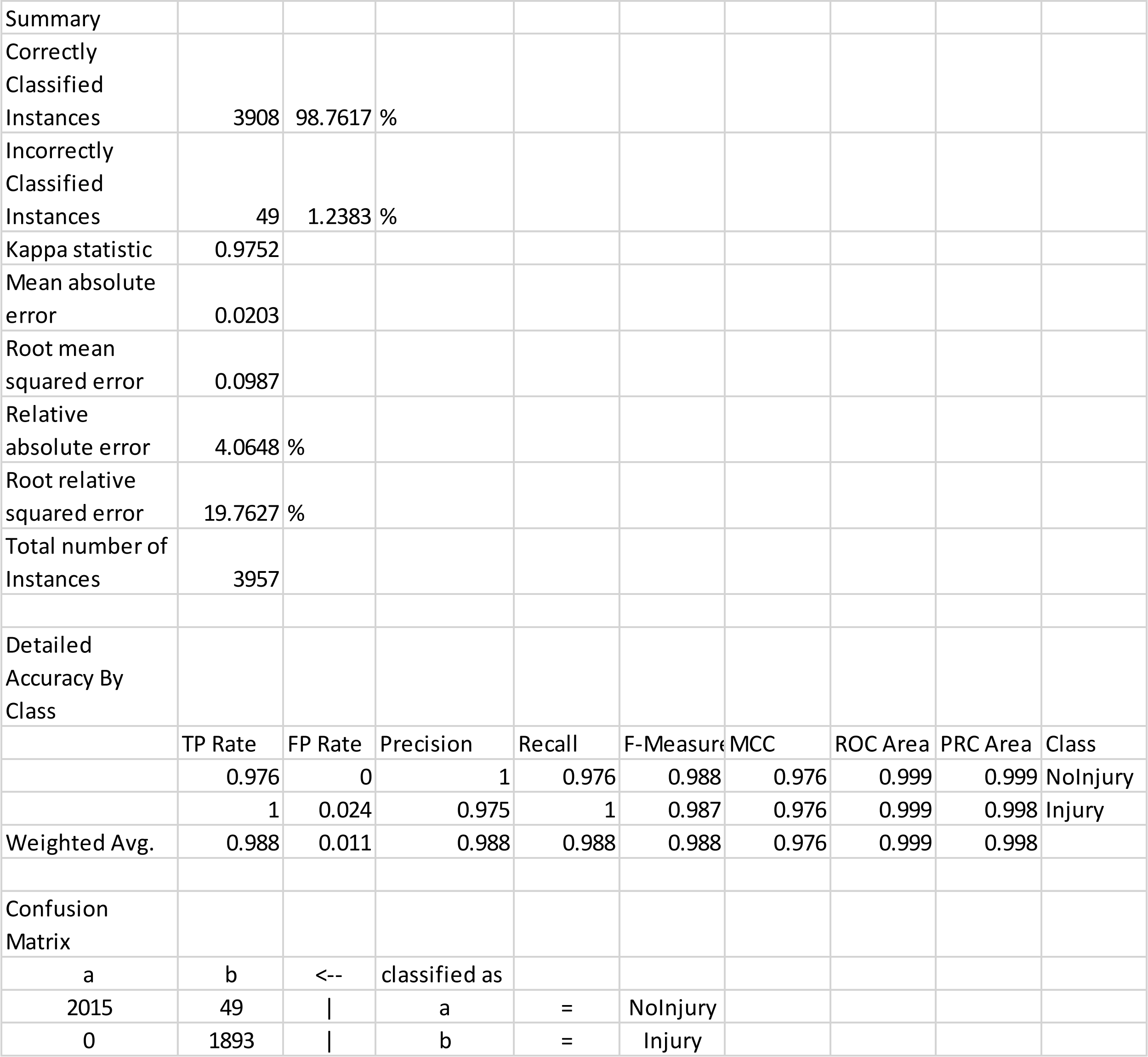
All performance measures including confusion table for 20 models can be seen in the tables below.

- **CfSubset IBk k=4 (Model 1)**



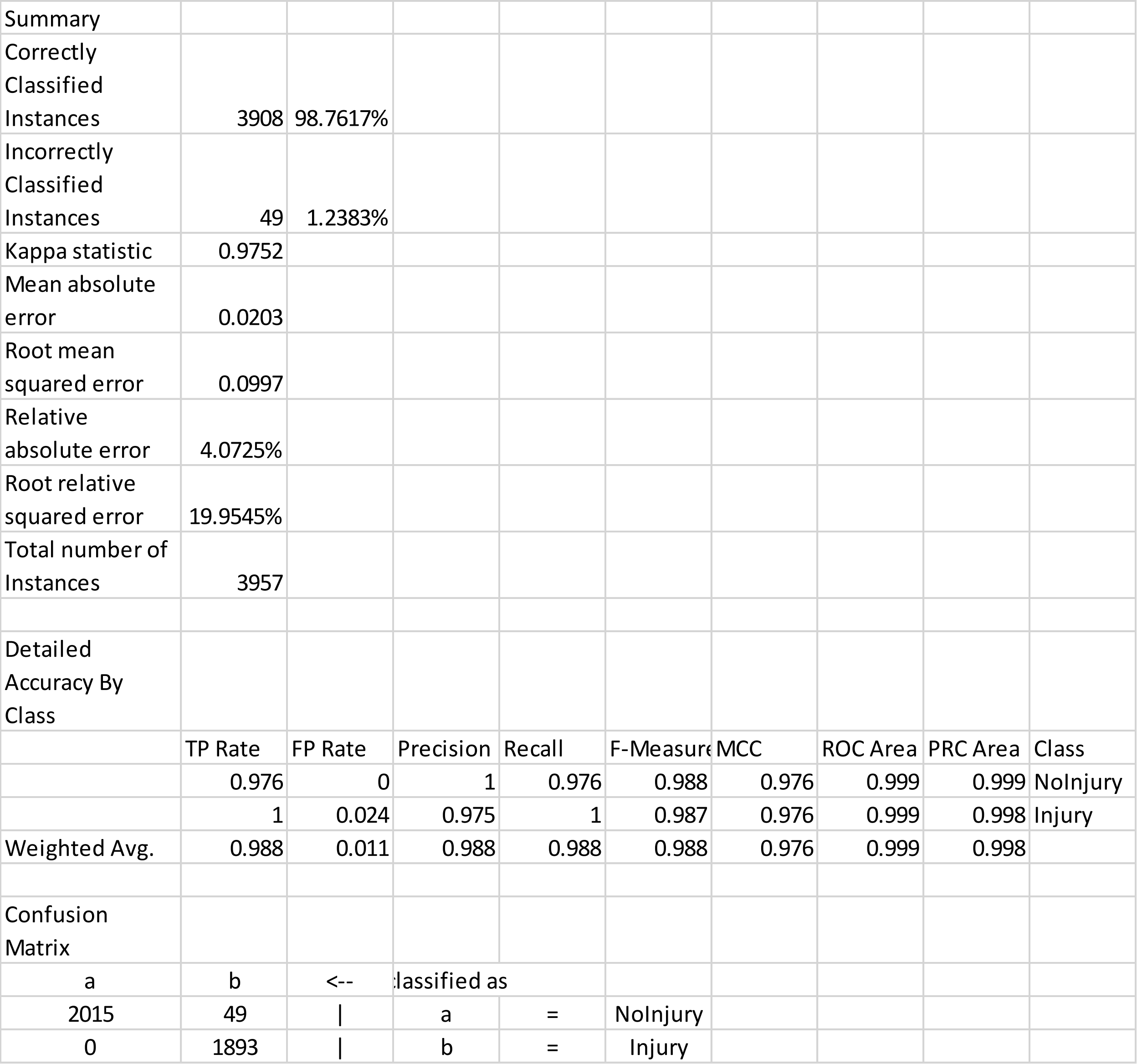
Summary									
Correctly Classified	3908	98.7617%							
Incorrectly Classified	49	1.2383%							
Kappa statistic	0.9752								
Mean absolute error	0.0203								
Root mean squared error	0.0995								
Relative absolute error	4.0733%								
Root relative squared error	19.9099%								
Total number of Instances	3957								
Detailed									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.977	0.001	1	0.977	0.988	0.975	0.999	0.999	NoInjury
	0.999	0.023	0.975	0.999	0.987	0.975	0.999	0.998	Injury
Weighted Avg.	0.988	0.011	0.988	0.988	0.988	0.975	0.999	0.998	
Confusion Matrix									
a	b	<--	classified as						
2016	48	|	a	=	NoInjury				
1	1892	|	b	=	Injury				

- **CfSubset Naïve Bayes (Model 6)**



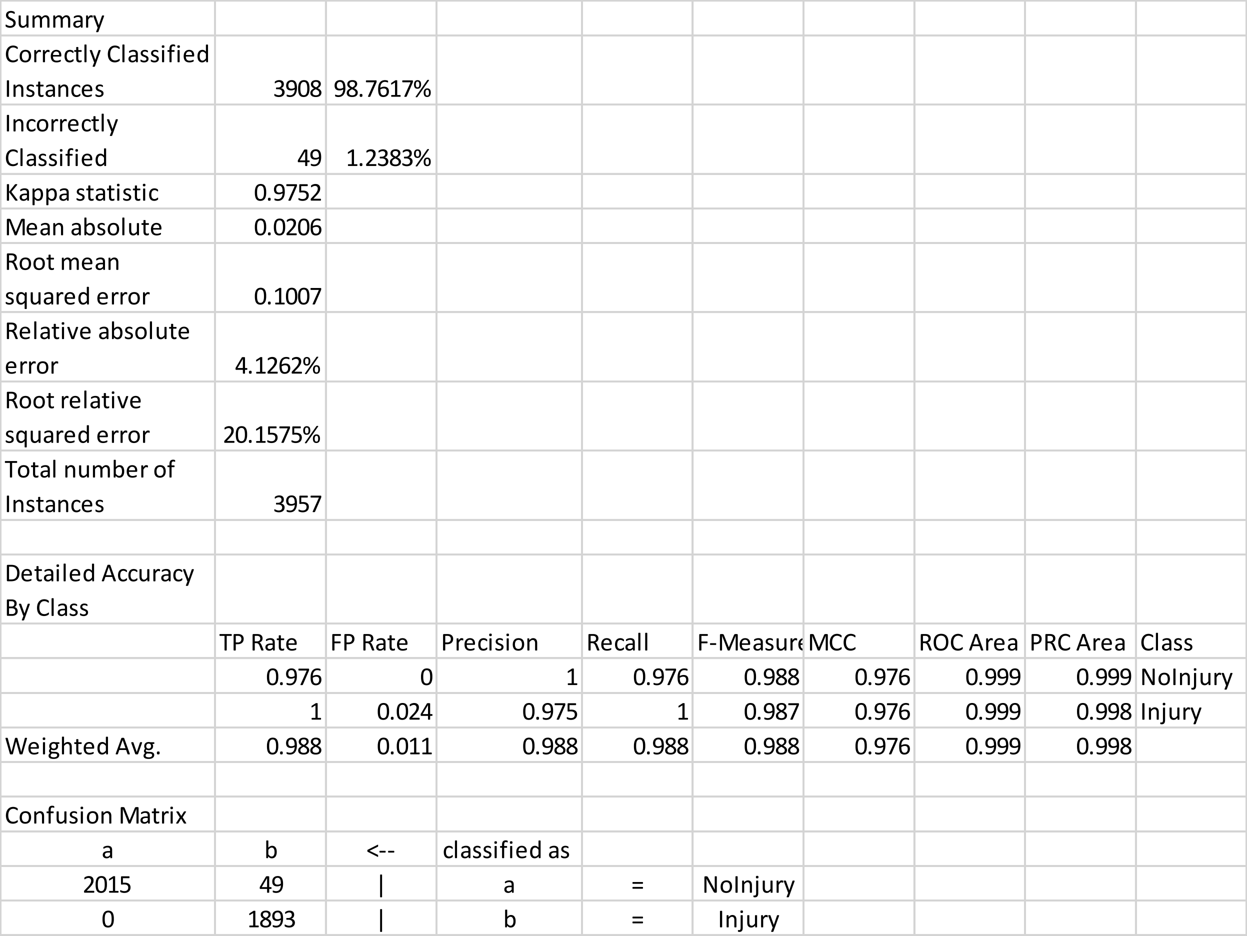
Summary									
Correctly Classified Instances	3908	98.7617	%						
Incorrectly Classified Instances	49	1.2383	%						
Kappa statistic	0.9752								
Mean absolute error	0.0203								
Root mean squared error	0.0987								
Relative absolute error	4.0648	%							
Root relative squared error	19.7627	%							
Total number of Instances	3957								
Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.976	0	1	0.976	0.988	0.976	0.999	0.999	NoInjury
	1	0.024	0.975	1	0.987	0.976	0.999	0.998	Injury
Weighted Avg.	0.988	0.011	0.988	0.988	0.988	0.976	0.999	0.998	
Confusion Matrix									
a	b	<--	classified as						
2015	49	|	a	=	NoInjury				
0	1893	|	b	=	Injury				

- **CfSubset Random Forest k =4 (Model 11)**



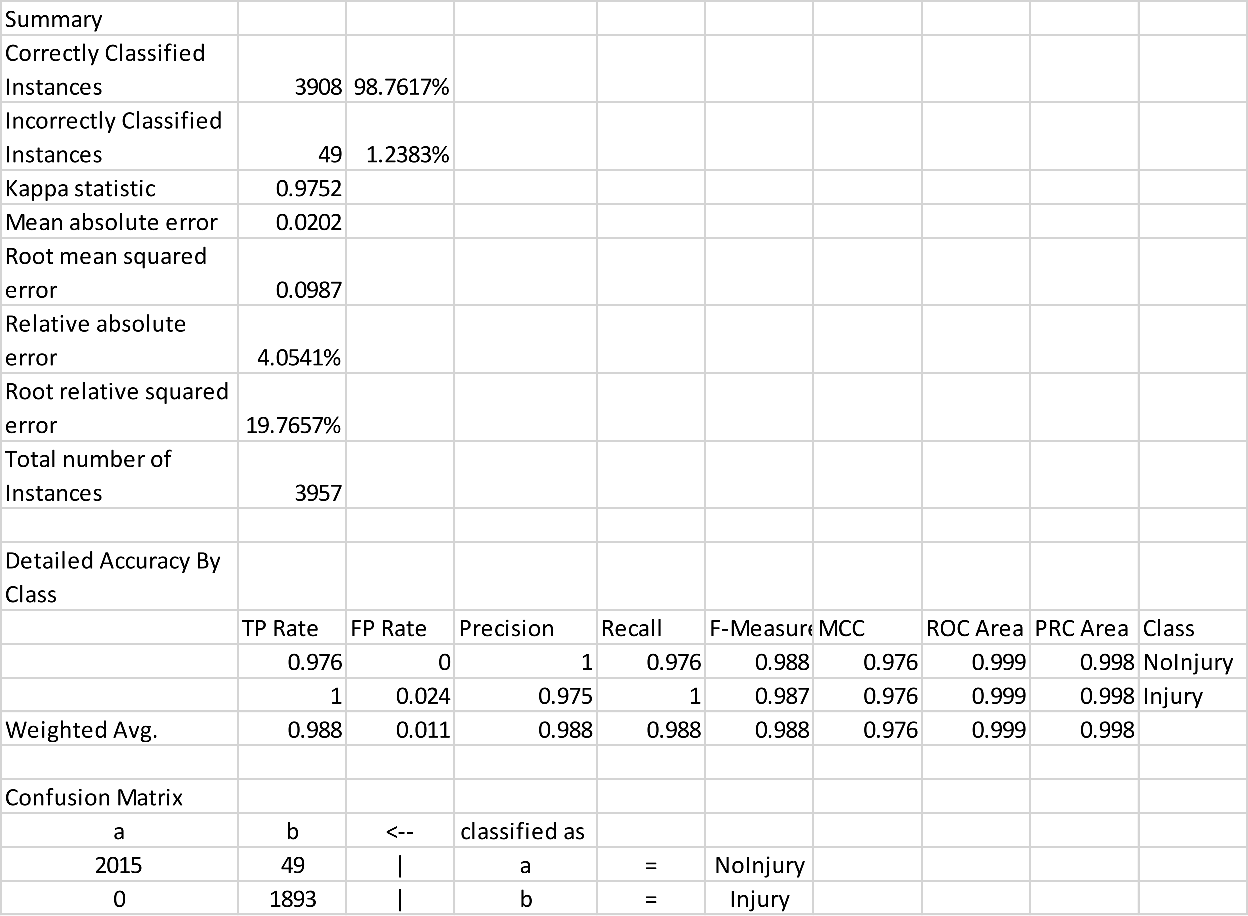
Summary									
Correctly Classified Instances	3908	98.7617%							
Incorrectly Classified Instances	49	1.2383%							
Kappa statistic	0.9752								
Mean absolute error	0.0203								
Root mean squared error	0.0997								
Relative absolute error	4.0725%								
Root relative squared error	19.9545%								
Total number of Instances	3957								
Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.976	0	1	0.976	0.988	0.976	0.999	0.999	NoInjury
	1	0.024	0.975	1	0.987	0.976	0.999	0.998	Injury
Weighted Avg.	0.988	0.011	0.988	0.988	0.988	0.976	0.999	0.998	
Confusion Matrix									
a	b	<--	lassified as						
2015	49	|	a	=	NoInjury				
0	1893	|	b	=	Injury				

- **CfSubset Ada Boost M1 + Random Forest k = 4 (Model 16)**



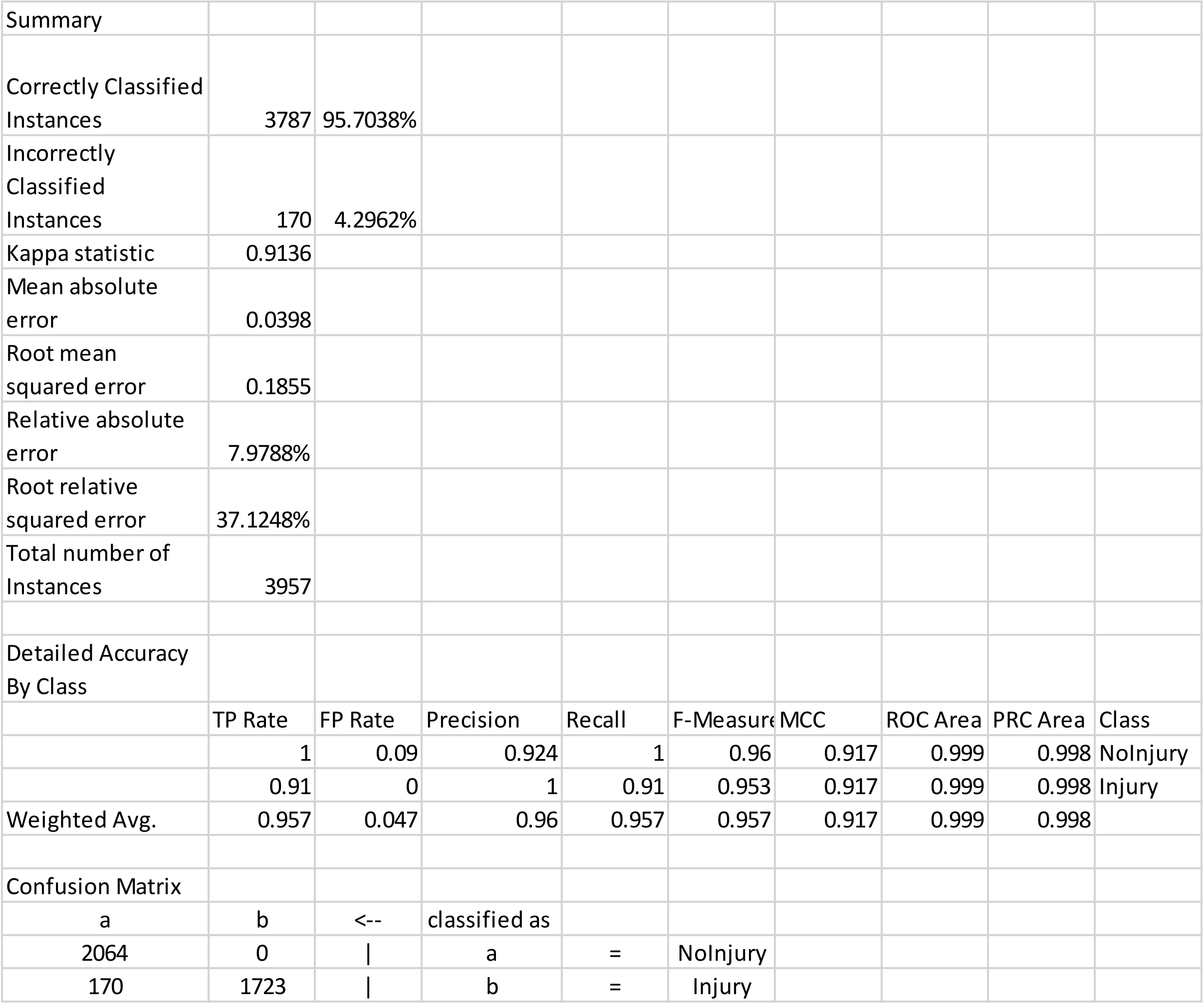
Summary									
Correctly Classified Instances	3908	98.7617%							
Incorrectly Classified	49	1.2383%							
Kappa statistic	0.9752								
Mean absolute	0.0206								
Root mean squared error	0.1007								
Relative absolute error	4.1262%								
Root relative squared error	20.1575%								
Total number of Instances	3957								
Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.976	0	1	0.976	0.988	0.976	0.999	0.999	NoInjury
	1	0.024	0.975	1	0.987	0.976	0.999	0.998	Injury
Weighted Avg.	0.988	0.011	0.988	0.988	0.988	0.976	0.999	0.998	
Confusion Matrix									
a	b	<--	classified as						
2015	49	|	a	=	NoInjury				
0	1893	|	b	=	Injury				

- **CorrAttr IBk k = 4 (Model 2)**



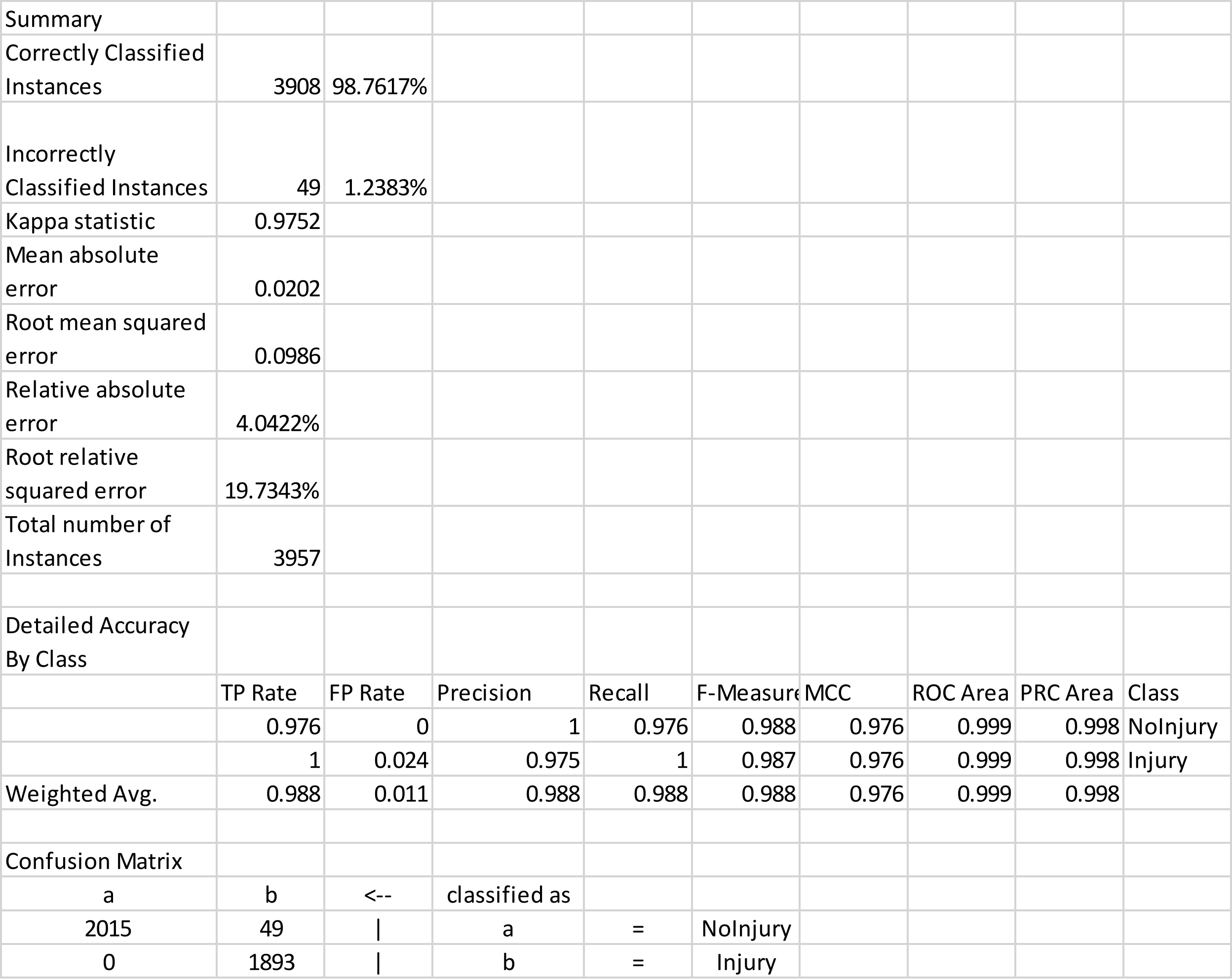
Summary									
Correctly Classified Instances	3908	98.7617%							
Incorrectly Classified Instances	49	1.2383%							
Kappa statistic	0.9752								
Mean absolute error	0.0202								
Root mean squared error	0.0987								
Relative absolute error	4.0541%								
Root relative squared error	19.7657%								
Total number of Instances	3957								
Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.976	0	1	0.976	0.988	0.976	0.999	0.998	NoInjury
	1	0.024	0.975	1	0.987	0.976	0.999	0.998	Injury
Weighted Avg.	0.988	0.011	0.988	0.988	0.988	0.976	0.999	0.998	
Confusion Matrix									
a	b	<--	classified as						
2015	49	|	a	=	NoInjury				
0	1893	|	b	=	Injury				

- **CorrAttr Naïve Bayes (Model 7)**



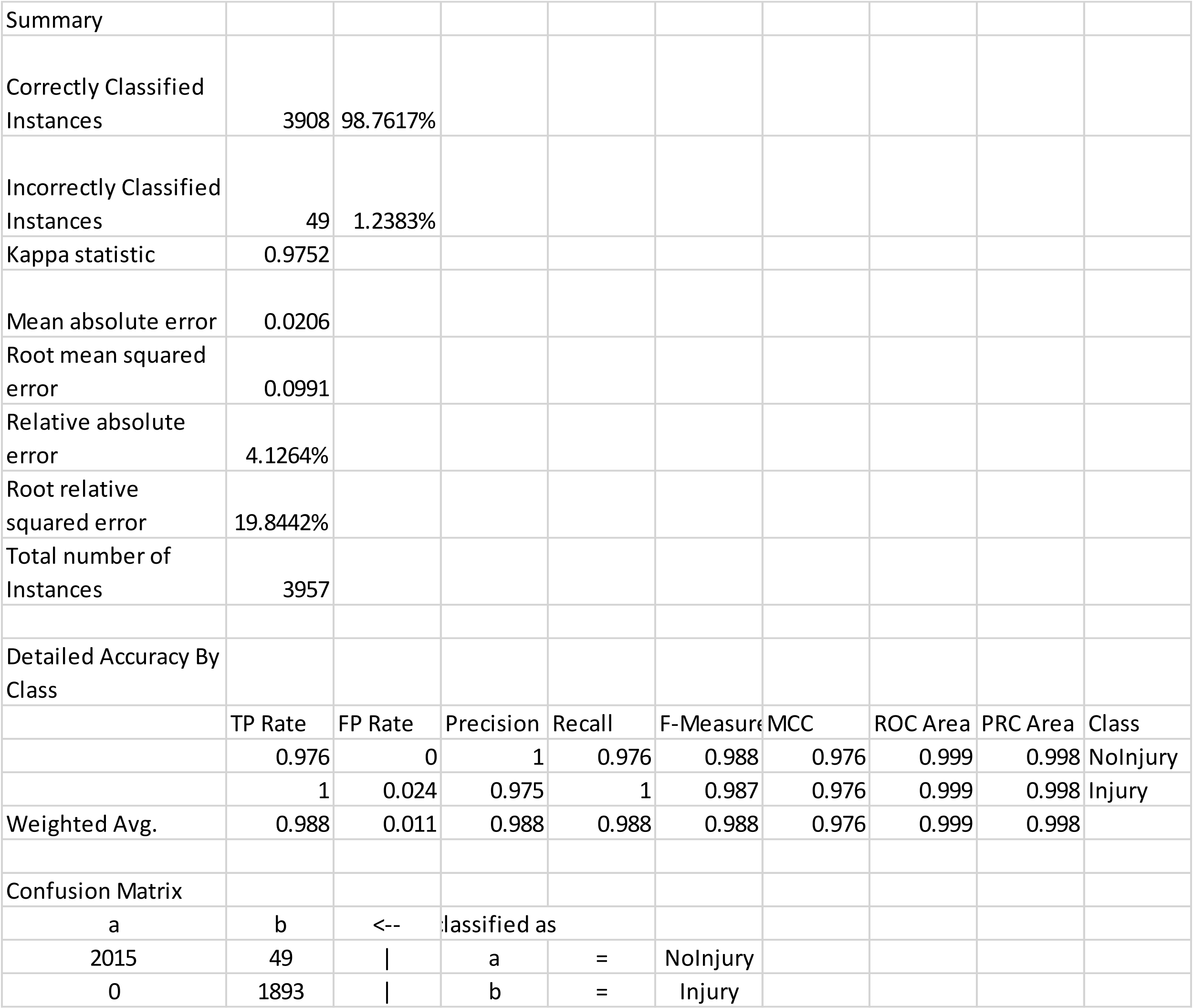
Summary									
Correctly Classified Instances	3787	95.7038%							
Incorrectly Classified Instances	170	4.2962%							
Kappa statistic	0.9136								
Mean absolute error	0.0398								
Root mean squared error	0.1855								
Relative absolute error	7.9788%								
Root relative squared error	37.1248%								
Total number of Instances	3957								
Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measur	MCC	ROC Area	PRC Area	Class
	1	0.09	0.924	1	0.96	0.917	0.999	0.998	NoInjury
	0.91	0	1	0.91	0.953	0.917	0.999	0.998	Injury
Weighted Avg.	0.957	0.047	0.96	0.957	0.957	0.917	0.999	0.998	
Confusion Matrix									
a	b	<--	classified as						
2064	0	|	a	=	NoInjury				
170	1723	|	b	=	Injury				

- **CorrAttr Random Forest k = 4 (Model 12)**



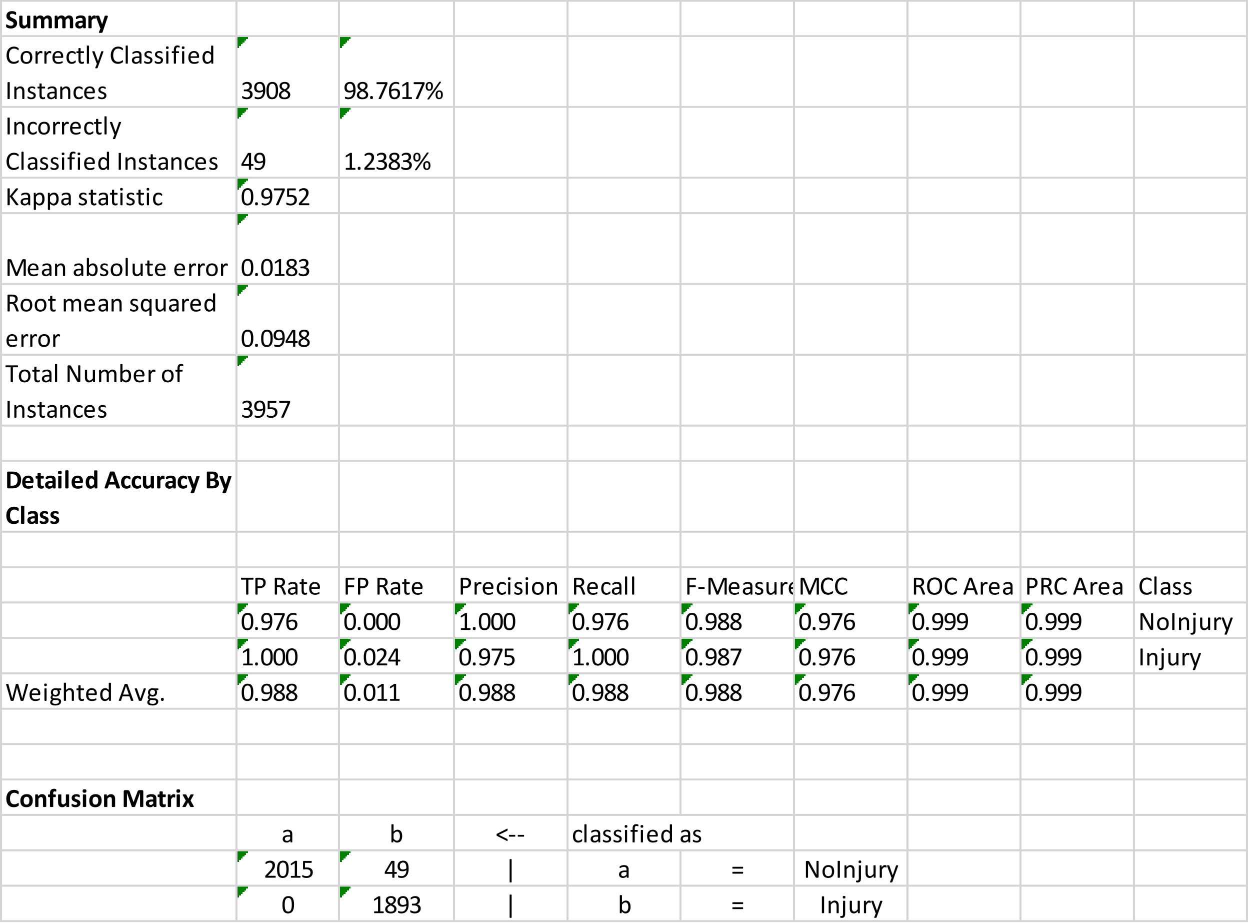
Summary									
Correctly Classified Instances	3908	98.7617%							
Incorrectly Classified Instances	49	1.2383%							
Kappa statistic	0.9752								
Mean absolute error	0.0202								
Root mean squared error	0.0986								
Relative absolute error	4.0422%								
Root relative squared error	19.7343%								
Total number of Instances	3957								
Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measur	MCC	ROC Area	PRC Area	Class
	0.976	0	1	0.976	0.988	0.976	0.999	0.998	NoInjury
	1	0.024	0.975	1	0.987	0.976	0.999	0.998	Injury
Weighted Avg.	0.988	0.011	0.988	0.988	0.988	0.976	0.999	0.998	
Confusion Matrix									
a	b	<--	classified as						
2015	49	|	a	=	NoInjury				
0	1893	|	b	=	Injury				

- **CorrAttr Ada Boost M1 + Random Forest (Model 17)**



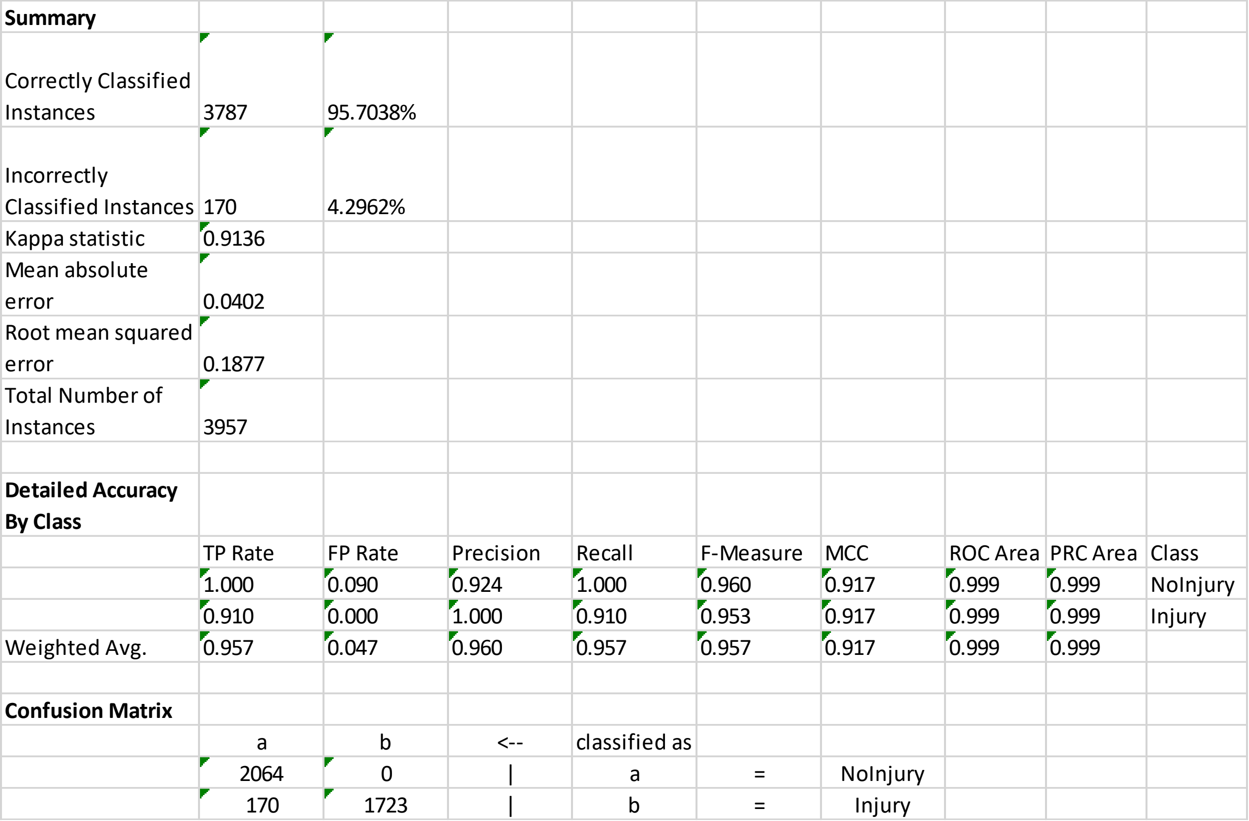
Summary									
Correctly Classified Instances	3908	98.7617%							
Incorrectly Classified Instances	49	1.2383%							
Kappa statistic	0.9752								
Mean absolute error	0.0206								
Root mean squared error	0.0991								
Relative absolute error	4.1264%								
Root relative squared error	19.8442%								
Total number of Instances	3957								
Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.976	0	1	0.976	0.988	0.976	0.999	0.998	NoInjury
	1	0.024	0.975	1	0.987	0.976	0.999	0.998	Injury
Weighted Avg.	0.988	0.011	0.988	0.988	0.988	0.976	0.999	0.998	
Confusion Matrix									
a	b	<--	classified as						
2015	49	|	a	=	NoInjury				
0	1893	|	b	=	Injury				

- **InfoGainAttr IBk k = 4 (Model 3)**



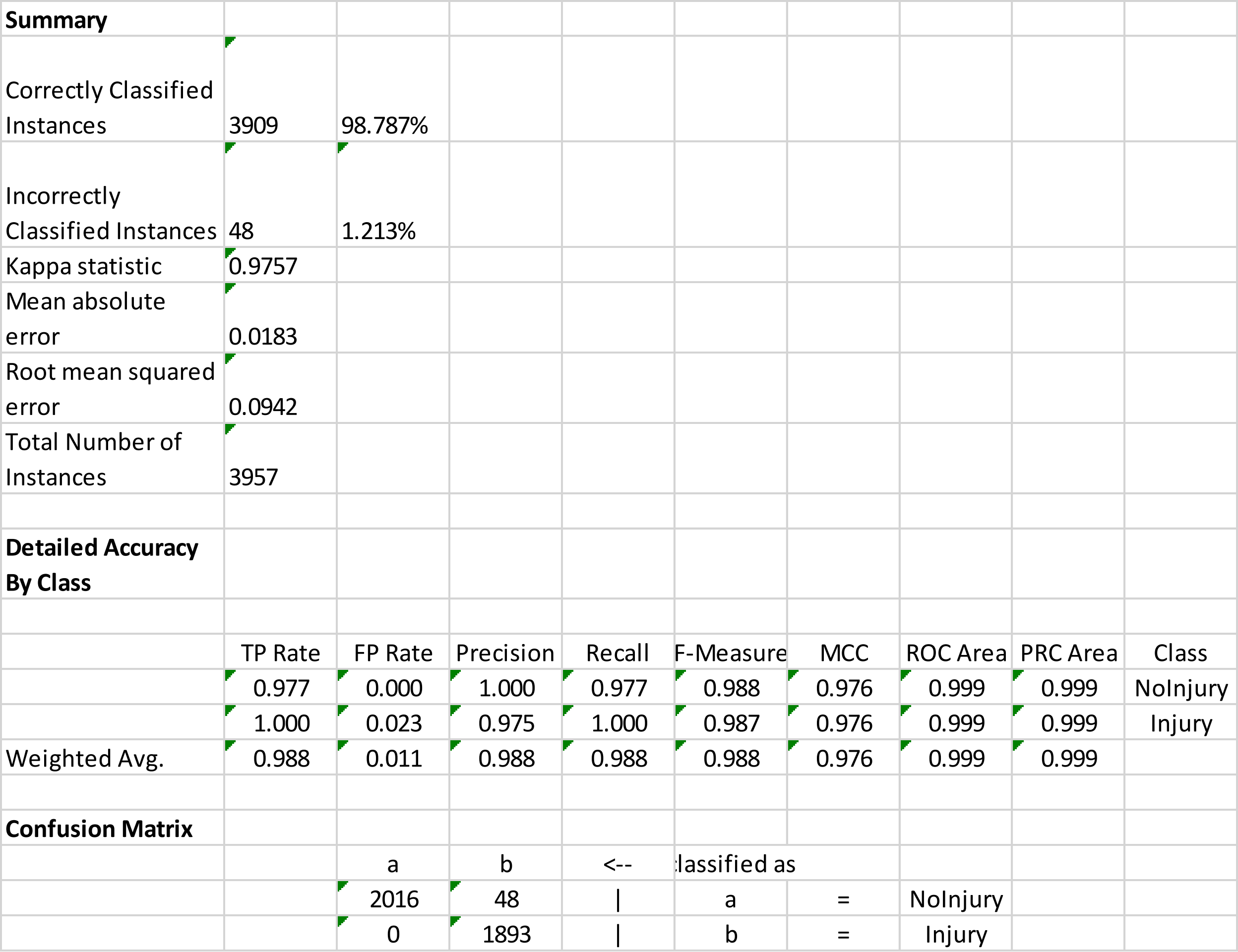
Summary									
Correctly Classified Instances	3908	98.7617%							
Incorrectly Classified Instances	49	1.2383%							
Kappa statistic	0.9752								
Mean absolute error	0.0183								
Root mean squared error	0.0948								
Total Number of Instances	3957								
Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.976	0.000	1.000	0.976	0.988	0.976	0.999	0.999	NoInjury
	1.000	0.024	0.975	1.000	0.987	0.976	0.999	0.999	Injury
Weighted Avg.	0.988	0.011	0.988	0.988	0.988	0.976	0.999	0.999	
Confusion Matrix									
	a	b	<--	classified as					
	2015	49	|	a	=	NoInjury			
	0	1893	|	b	=	Injury			

- **InfoGainAttr Naïve Bayes (Model 8)**



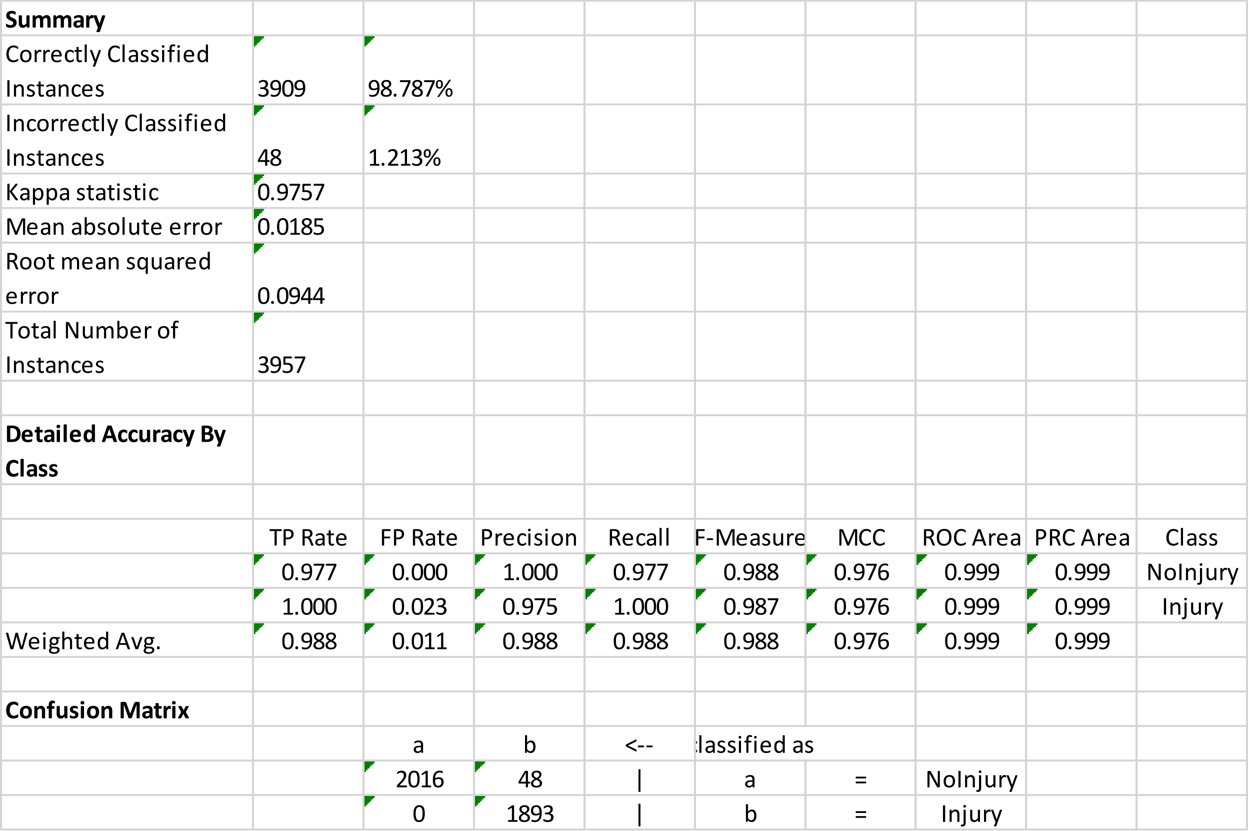
Summary									
Correctly Classified Instances	3787	95.7038%							
Incorrectly Classified Instances	170	4.2962%							
Kappa statistic	0.9136								
Mean absolute error	0.0402								
Root mean squared error	0.1877								
Total Number of Instances	3957								
Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.090	0.924	1.000	0.960	0.917	0.999	0.999	NoInjury
	0.910	0.000	1.000	0.910	0.953	0.917	0.999	0.999	Injury
Weighted Avg.	0.957	0.047	0.960	0.957	0.957	0.917	0.999	0.999	
Confusion Matrix									
	a	b	<--	classified as					
	2064	0	|	a	=	NoInjury			
	170	1723	|	b	=	Injury			

- **InfoGainAttr Random Forest k = 4 (Model 13)**



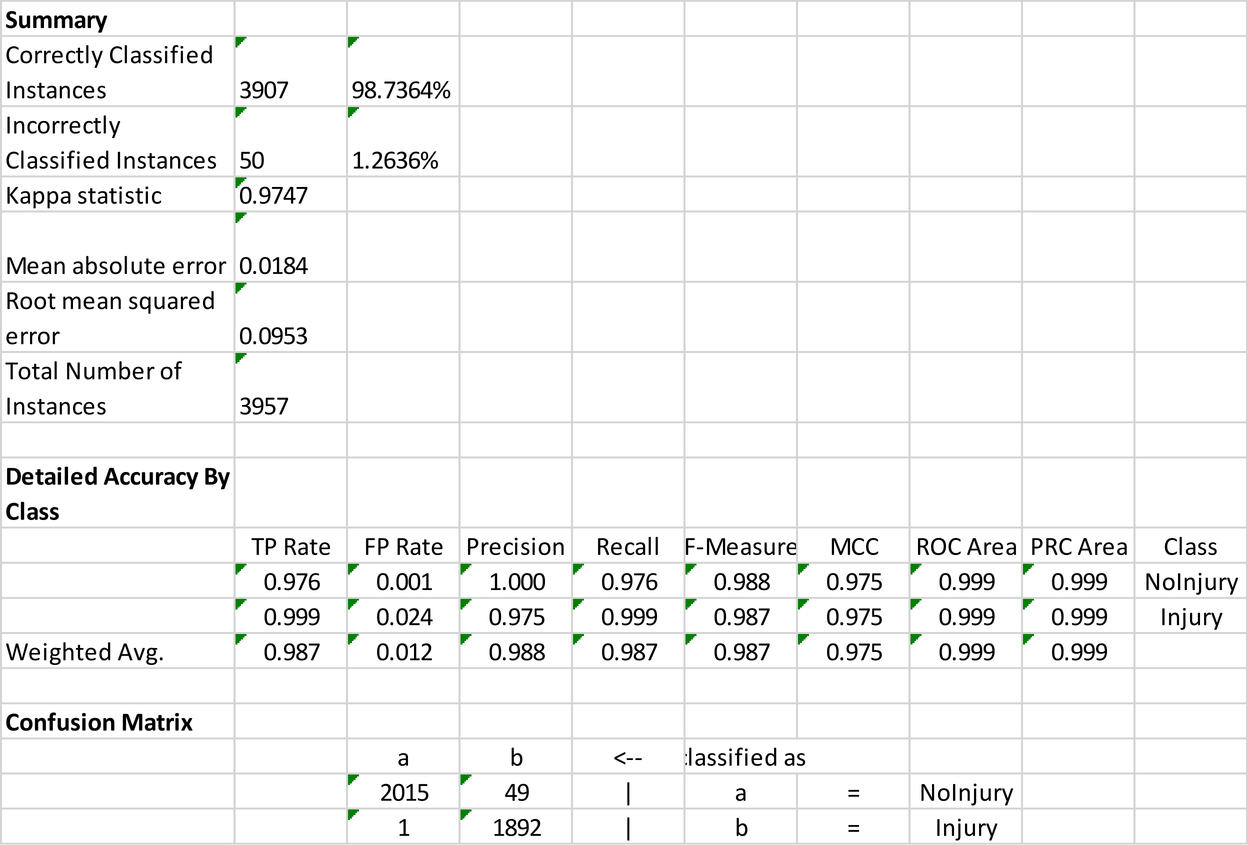
Summary									
Correctly Classified Instances	3909	98.787%							
Incorrectly Classified Instances	48	1.213%							
Kappa statistic	0.9757								
Mean absolute error	0.0183								
Root mean squared error	0.0942								
Total Number of Instances	3957								
Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.977	0.000	1.000	0.977	0.988	0.976	0.999	0.999	NoInjury
	1.000	0.023	0.975	1.000	0.987	0.976	0.999	0.999	Injury
Weighted Avg.	0.988	0.011	0.988	0.988	0.988	0.976	0.999	0.999	
Confusion Matrix									
		a	b	<--	lassified as				
		2016	48	|	a	=	NoInjury		
		0	1893	|	b	=	Injury		

- **InfoGainAttr Ada Boost M1 + Random Forest k = 4 (Model 18)**



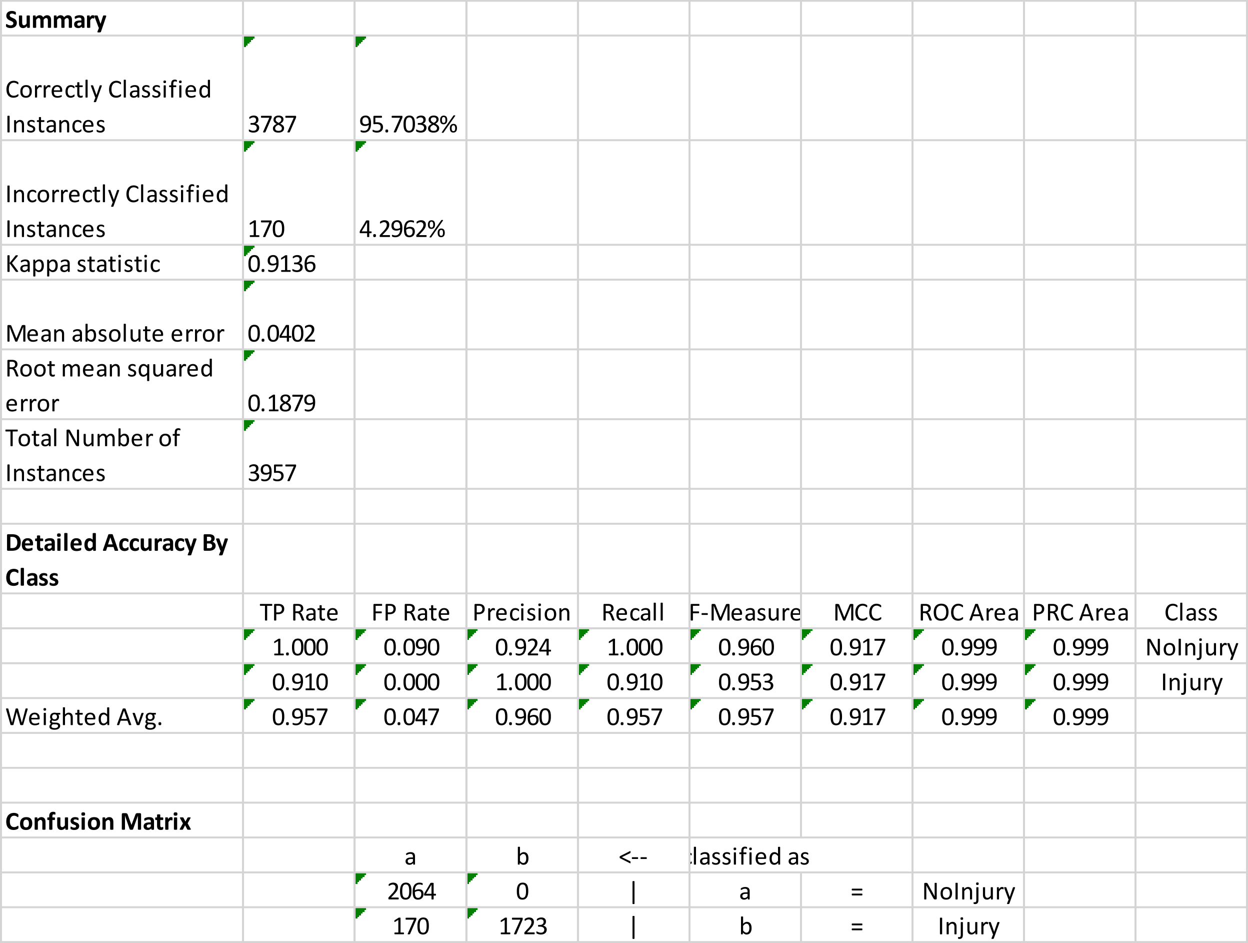
Summary									
Correctly Classified Instances	3909	98.787%							
Incorrectly Classified Instances	48	1.213%							
Kappa statistic	0.9757								
Mean absolute error	0.0185								
Root mean squared error	0.0944								
Total Number of Instances	3957								
Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.977	0.000	1.000	0.977	0.988	0.976	0.999	0.999	NoInjury
	1.000	0.023	0.975	1.000	0.987	0.976	0.999	0.999	Injury
Weighted Avg.	0.988	0.011	0.988	0.988	0.988	0.976	0.999	0.999	
Confusion Matrix									
		a	b	<--	lassified as				
		2016	48	|	a	=	NoInjury		
		0	1893	|	b	=	Injury		

- **GainRatioAttr IBk k = 4 (Model 4)**



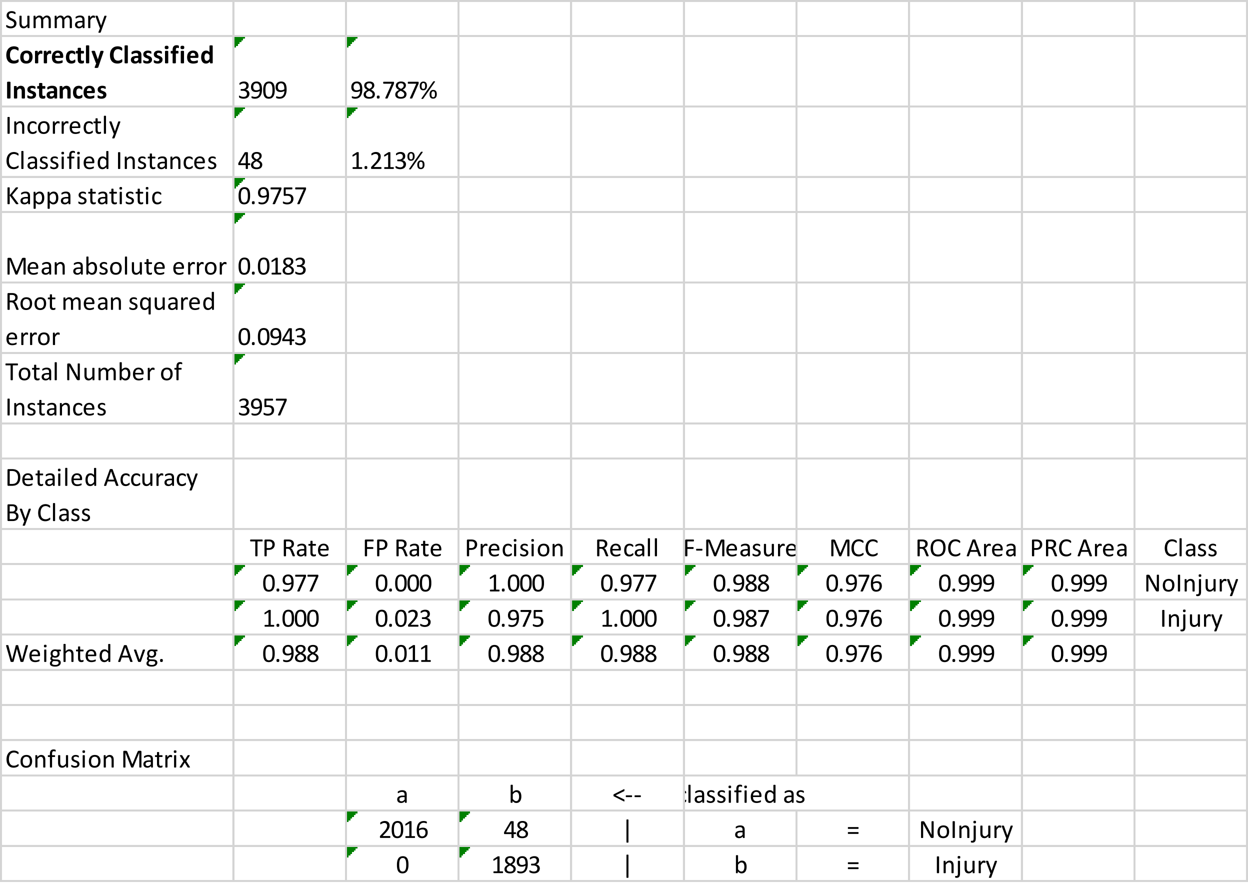
Summary									
Correctly Classified Instances	3907	98.7364%							
Incorrectly Classified Instances	50	1.2636%							
Kappa statistic	0.9747								
Mean absolute error	0.0184								
Root mean squared error	0.0953								
Total Number of Instances	3957								
Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.976	0.001	1.000	0.976	0.988	0.975	0.999	0.999	NoInjury
	0.999	0.024	0.975	0.999	0.987	0.975	0.999	0.999	Injury
Weighted Avg.	0.987	0.012	0.988	0.987	0.987	0.975	0.999	0.999	
Confusion Matrix									
		a	b	<--	lassified as				
		2015	49	|	a	=	NoInjury		
		1	1892	|	b	=	Injury		

- **GainRatioAttr Naïve Bayes (Model 9)**



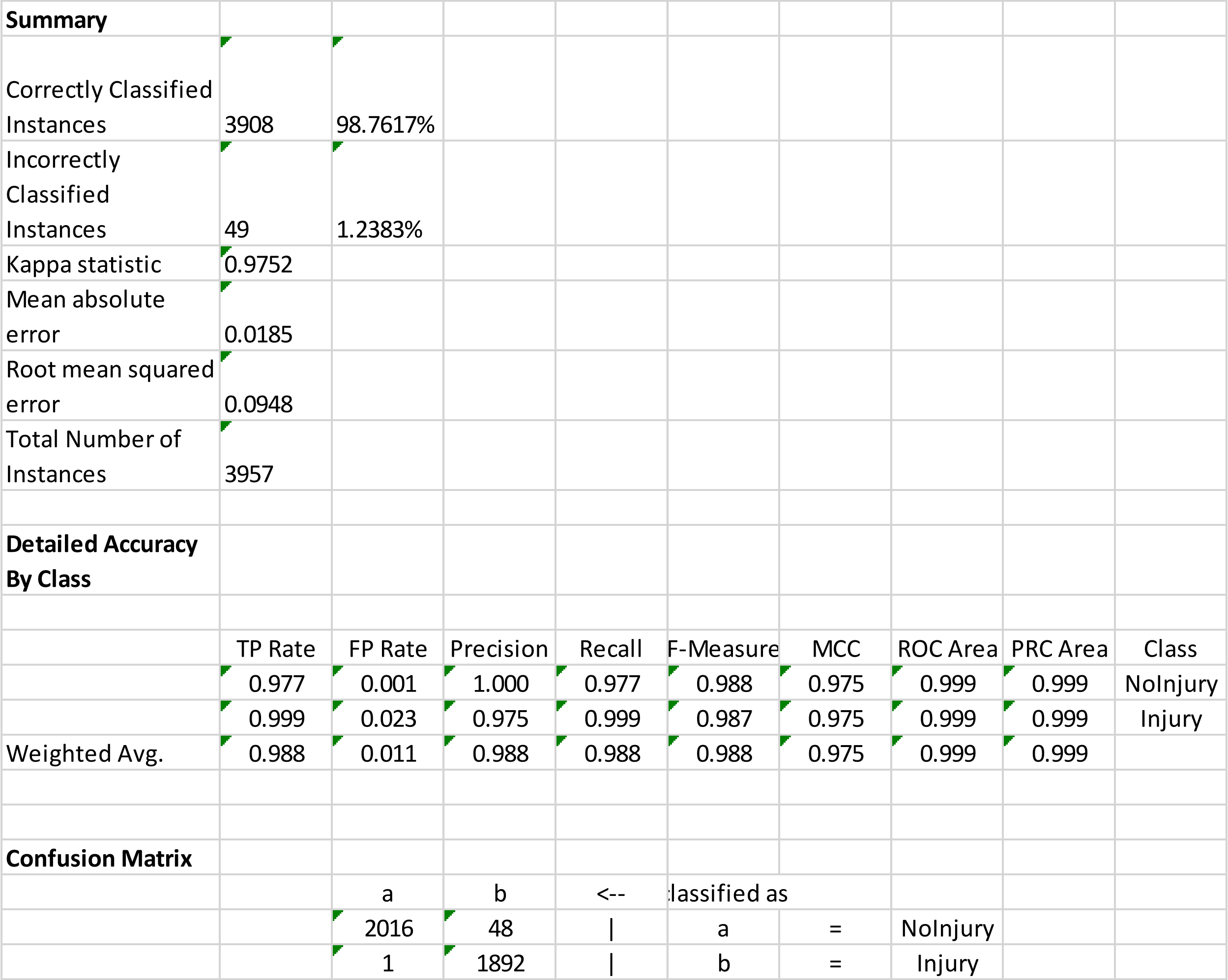
Summary									
Correctly Classified Instances	3787	95.7038%							
Incorrectly Classified Instances	170	4.2962%							
Kappa statistic	0.9136								
Mean absolute error	0.0402								
Root mean squared error	0.1879								
Total Number of Instances	3957								
Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.090	0.924	1.000	0.960	0.917	0.999	0.999	NoInjury
	0.910	0.000	1.000	0.910	0.953	0.917	0.999	0.999	Injury
Weighted Avg.	0.957	0.047	0.960	0.957	0.957	0.917	0.999	0.999	
Confusion Matrix									
		a	b	<--	lassified as				
		2064	0	|	a	=	NoInjury		
		170	1723	|	b	=	Injury		

- **GainRatioAttr Random Forest k = 4 (Model 14)**



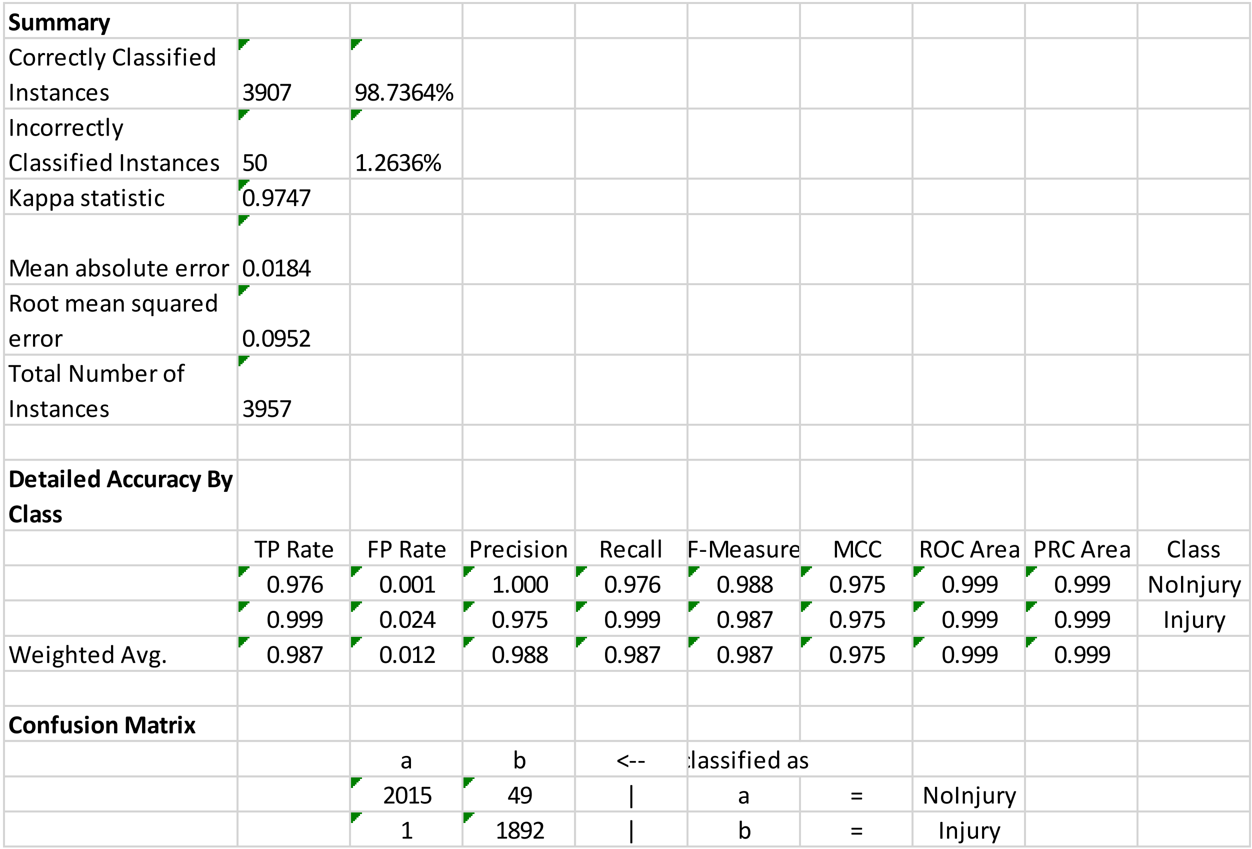
Summary									
Correctly Classified Instances	3909	98.787%							
Incorrectly Classified Instances	48	1.213%							
Kappa statistic	0.9757								
Mean absolute error	0.0183								
Root mean squared error	0.0943								
Total Number of Instances	3957								
Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.977	0.000	1.000	0.977	0.988	0.976	0.999	0.999	NoInjury
	1.000	0.023	0.975	1.000	0.987	0.976	0.999	0.999	Injury
Weighted Avg.	0.988	0.011	0.988	0.988	0.988	0.976	0.999	0.999	
Confusion Matrix									
		a	b	<--	lassified as				
		2016	48	|	a	=	NoInjury		
		0	1893	|	b	=	Injury		

- **GainRatioAttr Ada Boost M1 + Random Forest k = 4 (Model 19)**



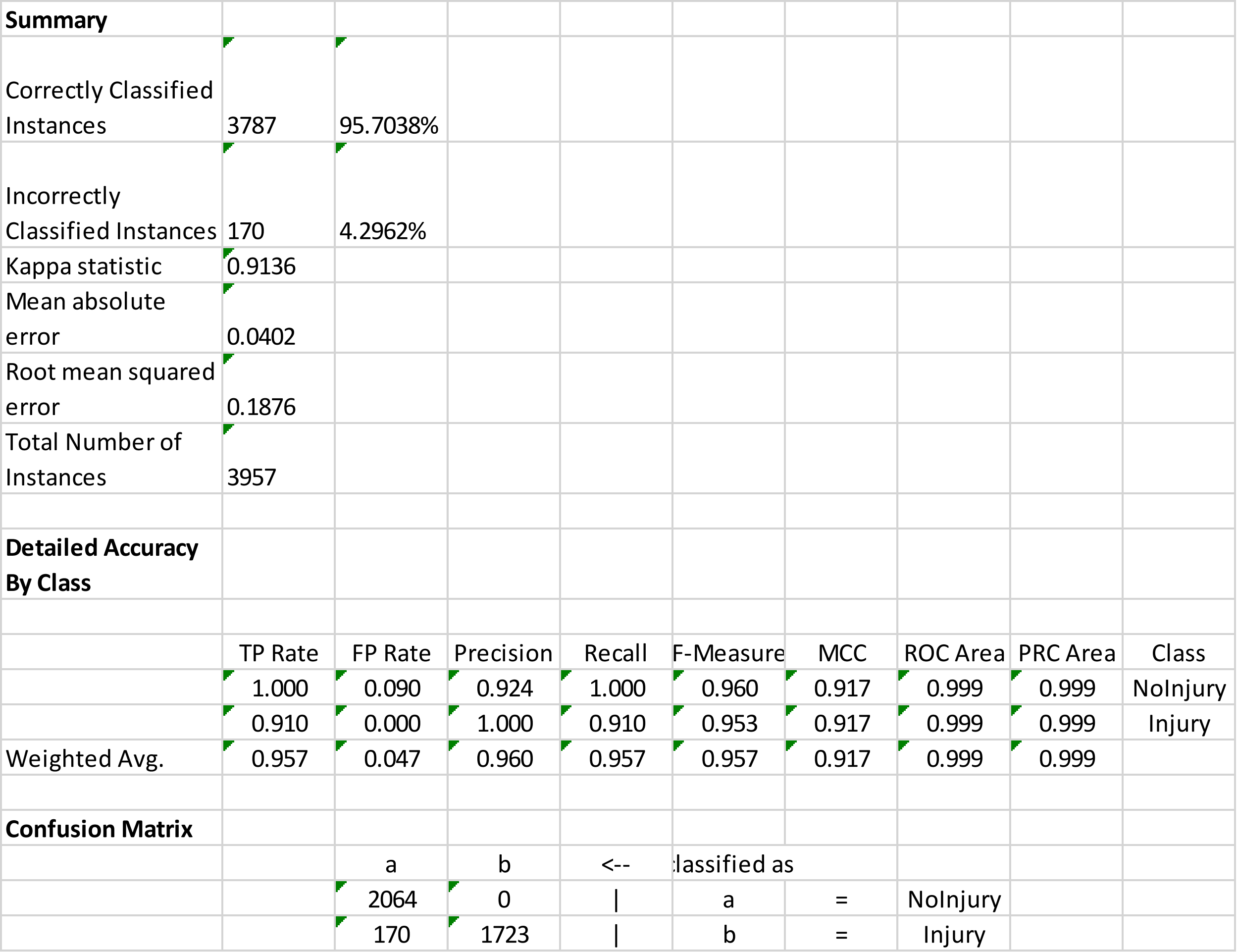
Summary									
Correctly Classified Instances	3908	98.7617%							
Incorrectly Classified Instances	49	1.2383%							
Kappa statistic	0.9752								
Mean absolute error	0.0185								
Root mean squared error	0.0948								
Total Number of Instances	3957								
Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.977	0.001	1.000	0.977	0.988	0.975	0.999	0.999	NoInjury
	0.999	0.023	0.975	0.999	0.987	0.975	0.999	0.999	Injury
Weighted Avg.	0.988	0.011	0.988	0.988	0.988	0.975	0.999	0.999	
Confusion Matrix									
		a	b	<--	lassified as				
		2016	48	|	a	=	NoInjury		
		1	1892	|	b	=	Injury		

- **ClassifierAttr J48 + IBk k =4 (Model 5)**



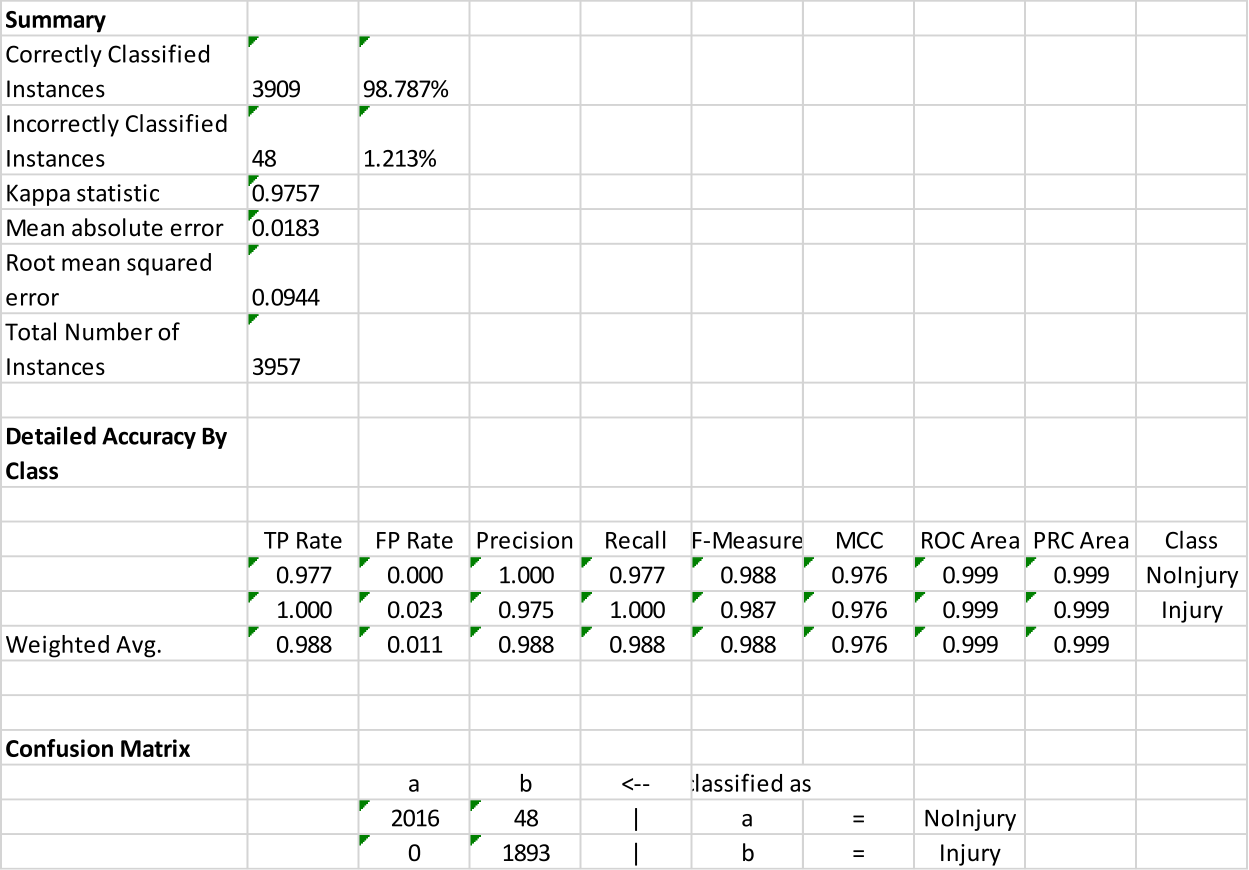
Summary									
Correctly Classified Instances	3907	98.7364%							
Incorrectly Classified Instances	50	1.2636%							
Kappa statistic	0.9747								
Mean absolute error	0.0184								
Root mean squared error	0.0952								
Total Number of Instances	3957								
Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.976	0.001	1.000	0.976	0.988	0.975	0.999	0.999	NoInjury
	0.999	0.024	0.975	0.999	0.987	0.975	0.999	0.999	Injury
Weighted Avg.	0.987	0.012	0.988	0.987	0.987	0.975	0.999	0.999	
Confusion Matrix									
		a	b	<--	lassified as				
		2015	49	|	a	=	NoInjury		
		1	1892	|	b	=	Injury		

- **ClassifierAttr J48 + Naïve Bayes (Model 10)**



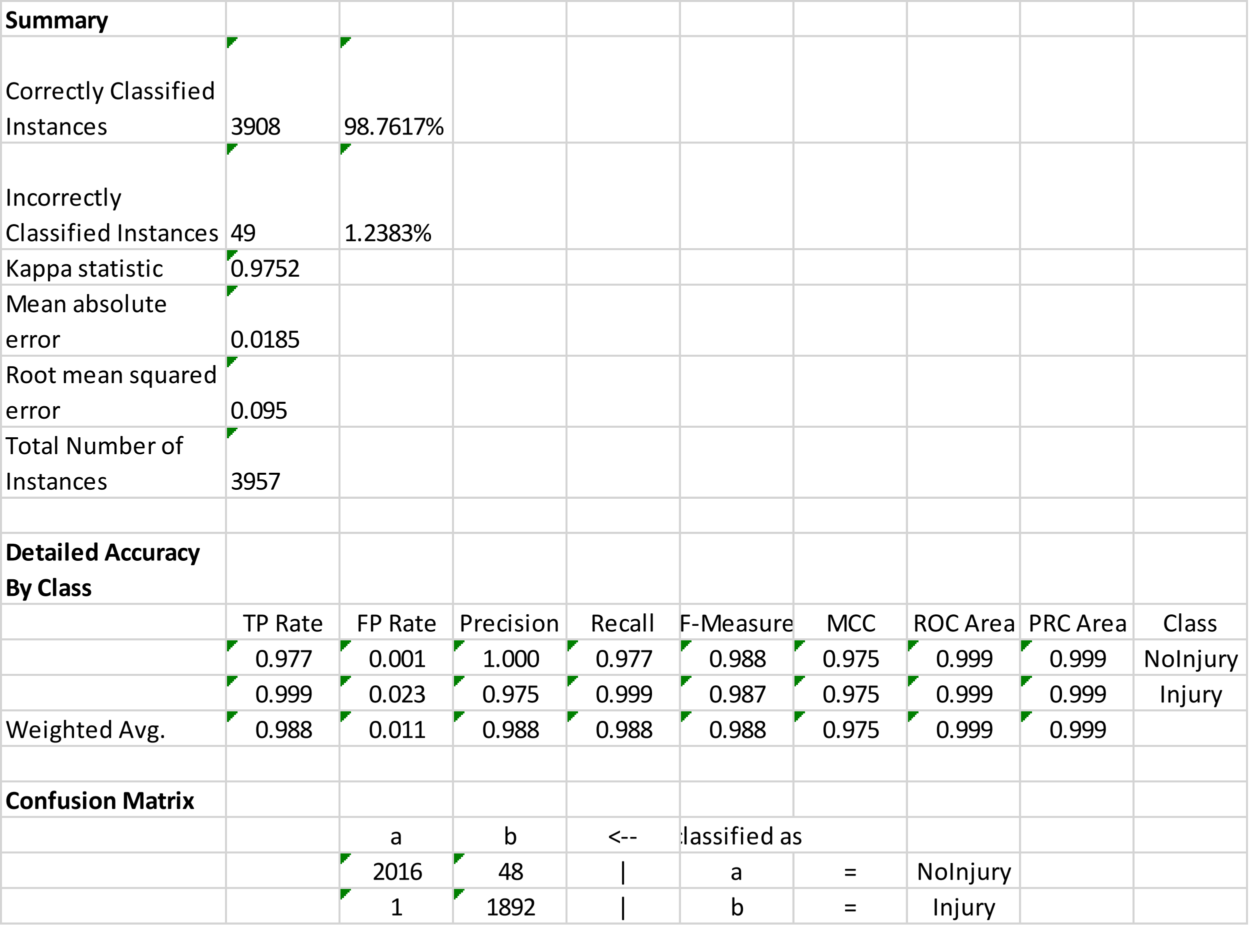
Summary									
Correctly Classified Instances	3787	95.7038%							
Incorrectly Classified Instances	170	4.2962%							
Kappa statistic	0.9136								
Mean absolute error	0.0402								
Root mean squared error	0.1876								
Total Number of Instances	3957								
Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.090	0.924	1.000	0.960	0.917	0.999	0.999	NoInjury
	0.910	0.000	1.000	0.910	0.953	0.917	0.999	0.999	Injury
Weighted Avg.	0.957	0.047	0.960	0.957	0.957	0.917	0.999	0.999	
Confusion Matrix									
		a	b	<--	lassified as				
		2064	0	|	a	=	NoInjury		
		170	1723	|	b	=	Injury		

- **ClassifierAttr J48 + Random Forest k = 4 (Model 15)**



Summary									
Correctly Classified Instances	3909	98.787%							
Incorrectly Classified Instances	48	1.213%							
Kappa statistic	0.9757								
Mean absolute error	0.0183								
Root mean squared error	0.0944								
Total Number of Instances	3957								
Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.977	0.000	1.000	0.977	0.988	0.976	0.999	0.999	NoInjury
	1.000	0.023	0.975	1.000	0.987	0.976	0.999	0.999	Injury
Weighted Avg.	0.988	0.011	0.988	0.988	0.988	0.976	0.999	0.999	
Confusion Matrix									
		a	b	<--	classified as				
		2016	48	|	a	=	NoInjury		
		0	1893	|	b	=	Injury		

- **ClassifierAttr J48 + Ada Boost M1 + Random Forest (Model 20)**



Summary									
Correctly Classified Instances	3908	98.7617%							
Incorrectly Classified Instances	49	1.2383%							
Kappa statistic	0.9752								
Mean absolute error	0.0185								
Root mean squared error	0.095								
Total Number of Instances	3957								
Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.977	0.001	1.000	0.977	0.988	0.975	0.999	0.999	NoInjury
	0.999	0.023	0.975	0.999	0.987	0.975	0.999	0.999	Injury
Weighted Avg.	0.988	0.011	0.988	0.988	0.988	0.975	0.999	0.999	
Confusion Matrix									
		a	b	<--	classified as				
		2016	48	|	a	=	NoInjury		
		1	1892	|	b	=	Injury		