Classification of Emergency Response Incident Results

Team Members: Justin Lee, John Kim

10/21/2024

Table of Contents

Part 1 – Statement/Project Goal

Emergency Medical Services, more commonly known as EMS, is a system that responds to emergencies in need of highly skilled pre-hospital clinicians. The EMS Computer Aided Dispatch System generates data about the incident as it relates to the assignment of resources and the Fire Department's response to the emergency.

We want to predict the disposition code (indicating the final outcome of the incident) of an EMS incident. This will help the medical industry at large and provide insight on what aspects of emergency response are most effective for delivering successful care to patients.

Part 2 – Description of Dataset

We used the EMS Incident Dispatch Data from NYC OpenData[1]. Each instance of the dataset covers the whole process of one specific incident, starting at the time the incident is opened and elapsing until the incident closes. Altogether, the dataset spans the course of 19 years, with 25,984,643 instances. Because of the size of the dataset, we decided to take a subset of it, the most recent 1-month period from December 1st, 2023 to January 1st, 2024.

We were left with 139,499 instances, with each instance having 30 attributes and 1 class attribute.[2]

Through looking at the data, we can see that it is mostly well structured and uniform, especially across time, although it does have missing values in several of the columns. The missing values are as follows.

Attributes	Number of Missing Values
FIRST_ASSIGNMENT_DATETIME	3074
FIRST_ACTIVATION_DATETIME	3335
FIRST_ON_SCENE_DATETIME	8093
INCIDENT_RESPONSE_SECONDS_QY	8111
INCIDENT_TRAVEL_TM_SECONDS_QY	8093
FIRST_TO_HOSP_DATETIME	51569
FIRST_HOSP_ARRIVAL_DATETIME	51883
INCIDENT_CLOSE_DATETIME	12
ZIPCODE	1285

POLICEPRECINCT	1283
CITYCOUNCILDISTRICT	1283
COMMUNITYDISTRICT	1283
COMMUNITYSCHOOLDISTRICT	1360
CONGRESSIONALDISTRICT	1283
INCIDENT_DISPOSITION_CODE	3559

In particular, our class column, INCIDENT_DISPOSITION_CODE, has 3559 instances with missing values, all of which we would remove during preprocessing.

Further, below is our class distribution.[3]

Incident Disposition Code	Number of Instances
82	87996
83	1228
87	3196
90	10820
91	3848
93	23255
94	248
95	11
96	5338
Missing	3559

Our class label column is heavily skewed towards certain class labels, particularly 82, 93, and 90, which we would have to take into account when doing our train/test/validation split.

Part 3 – Preprocessing

Part 3.1 – Enable WEKA to open dataset

Our initial preprocessing steps were to process the CSV file in such a way that WEKA could convert the dataset into the ARFF file format. After debugging, we found most of these issues were because of WEKA's handling of nominal attributes, so we wrote some Python code in VSCode to fix these. In particular, WEKA's discovery algorithm for nominal attributes does not let it explore the whole dataset, and if it encounters a value it did not find originally, it throws the following error.

```
java.io.IOException: Read unknown nominal value ARRESTfor attribute INITIAL_CALL_TYPE (line: 102). Try increasing the size of the memory buffer (-B option) or explicitly specify legal nominal values with the -L option.
```

To address this, we first gathered a list of the nominal attributes that WEKA could not self discover: INITIAL_CALL_TYPE, FINAL_CALL_TYPE, INCIDENT_DISPATCH_AREA, TRANSFER_INDICATOR, STANDBY_INDICATOR, SPECIAL_EVENT_INDICATOR, VALID_DISPATCH_RSPNS_TIME_INDC. We then used python to query the actual unique values for each column.

```
attr_to_values: dict[str, list[str]] = {}
  for attr in ATTRIBUTES_EXPLICIT_NOMINAL:
     attr_to_values[attr] = df[attr].unique().tolist()

  with open(ARGUMENT_FILE, "w+") as f:
     for i, (attr, values) in enumerate(attr_to_values.items()):
        if i != 0:
           f.write(" ")
        f.write(f"-L {attr}:{','.join(values)}")
```

Then, when running WEKA's CSVLoader, we included these attributes as additional arguments:

```
java -cp WEKA.jar weka.core.converters.CSVLoader INPUT_FILE $(cat ARGUMENT_FILE) > OUTPUT_FILE
```

We also encountered another issue with dates, with WEKA treating them as nominal attributes. To circumvent this, we converted them to numerical data, which WEKA was able to parse properly.

```
datetime_columns = [col for col in df.columns if "DATETIME" in col]

# first convert column from strings to datetimes
for col in datetime_columns:
     df[col] = pd.to_datetime(df[col], format="%m/%d/%Y %I:%M:%S %p")

# then convert column from datetimes to ints
```

```
for col in datetime_columns:
    df[col] = df[col].astype(int) // 10**9
```

To do this, we found all of the “datetime” columns, parsed the format provided into a pandas Timestamp object, and then converted them into a Unix timestamp, or the seconds since January 1, 1970, UTC. Fixing these two issues, we were able to successfully convert our CSV to an ARFF file that WEKA would load.

Part 3.2 – Reduce Dimension

We did our preprocessing in Python, and did it on the CSV file before we converted it to an ARFF format. First, we removed 15 attributes that we were confident would be redundant or would have zero correlation with the class. This included one ID attribute, seven attributes describing the date and time (there were better attributes in the dataset that provide the same information but with relative time), and seven attributes describing the area codes where the incident occurred (we kept the one attribute with the zipcode of the incident since its the most specific of the location attributes). This removal halved the dimension of our data, decreasing it from 30 to 15, and made our data more manageable.

The specific attributes that we removed are as follows:
CAD_INCIDENT_ID
INCIDENT_DATETIME
FIRST_ASSIGNMENT_DATETIME
FIRST_ACTIVATION_DATETIME
FIRST_ON_SCENE_DATETIME
FIRST_TO_HOSP_DATETIME
FIRST_HOSP_ARRIVAL_DATETIME
INCIDENT_CLOSE_DATETIME
BOROUGH
INCIDENT_DISPATCH_AREA
POLICEPRECINCT
CITYCOUNCILDISTRICT
COMMUNITYDISTRICT
COMMUNITYSCHOOLDISTRICT
CONGRESSIONALDISTRICT

Part 3.3 – Missing Values

We started by removing all 3559 instances with missing values for our class label (INCIDENT_DISPOSITION_CODE), decreasing the number of instances from 139,499 to 135,940.

Then, we filled out the missing values for INCIDENT_RESPONSE_SECONDS_QY and INCIDENT_TRAVEL_TM_SECONDS_QY with the median of each column, since both of these columns were numeric with outliers.

Although these two columns were the only ones with missing values, our dataset had some hidden missing values. Two attributes, VALID_DISPATCH_RSPNS_TIME_INDC and VALID_INCIDENT_RSPNS_TIME_INDC, indicate whether or not the respective response times were valid. Whenever the indicator attribute value was false, we replaced the default value (0) for the respective attribute, DISPATCH_RESPONSE_SECONDS_QY or INCIDENT_RESPONSE_SECONDS_QY, with the median of that attribute. Since we already used the indicator attributes, we could safely remove both attributes, further decreasing our dimension from 15 to 13.

Part 3.4 – Miscellaneous

Next, we converted three other columns, ZIPCODE, INITIAL_SEVERITY_LEVEL_CODE, and FINAL_SEVERITY_LEVEL_CODE from numeric to nominal, as they represent discrete values.

Finally, we moved the INCIDENT_DISPOSITION_CODE column to the end and converted it from numeric to nominal in order to set it as the class column and prepare it for classification tasks.

Part 3.4 – Train/Test Split

Given that our distribution of class labels is unbalanced, we used stratified random sampling to get our train/test dataset. We decided to split our dataset in a ratio of 80% for train and 20% for test. The way we accomplished this is with Python and the Pandas library.

```
train_df = df.groupby(CLASS_ATTRIBUTE).sample(frac=TRAIN_SPLIT)
test_df = df.drop(train_df.index)
```

The original dataset has the following class label distribution:

82 87996
93 23255
90 10820
96 5338
91 3848
87 3196
83 1228
94 248
95 11

The train dataset has the following class label distribution:

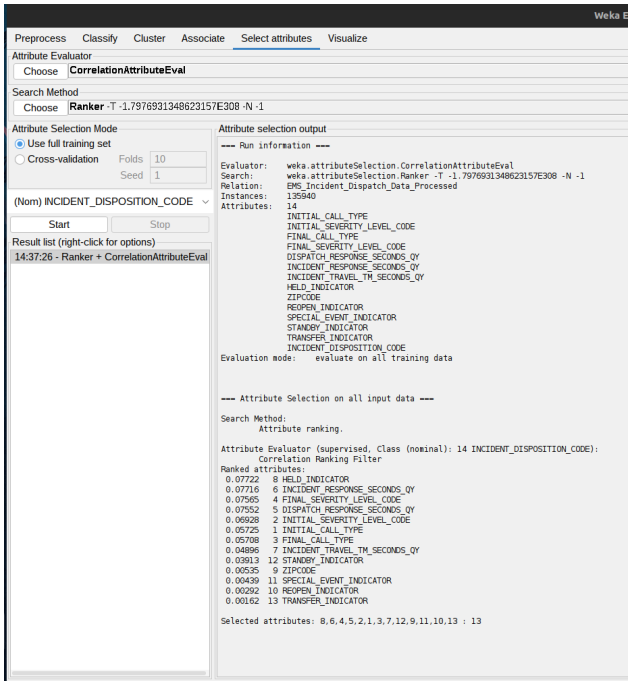
82 70397
93 18604
90 8656
96 4270
91 3078
87 2557

83 982
94 198
95 9

The test dataset has the following class label distribution:
82 17599
93 4651
90 2164
96 1068
91 770
87 639
83 246
94 50
95 2

Using stratified random sampling, we were able to preserve the same distribution of class labels in both our train and test datasets.

Part 4 – Attribute Selection Algorithms & Model Classifiers Used

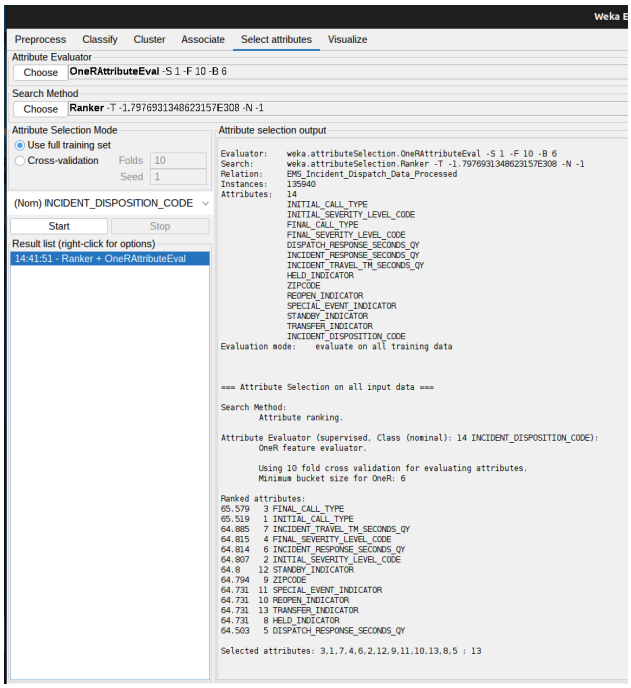


CorrelationAttributeEval

Calculates Pearson's Correlation Coefficient between all attributes and class

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

Cutoff: 0.01
Removed attributes:
ZIPCODE
SPECIAL_EVENT_INDICATOR
REOPEN_INDICATOR
TRANSFER_INDICATOR

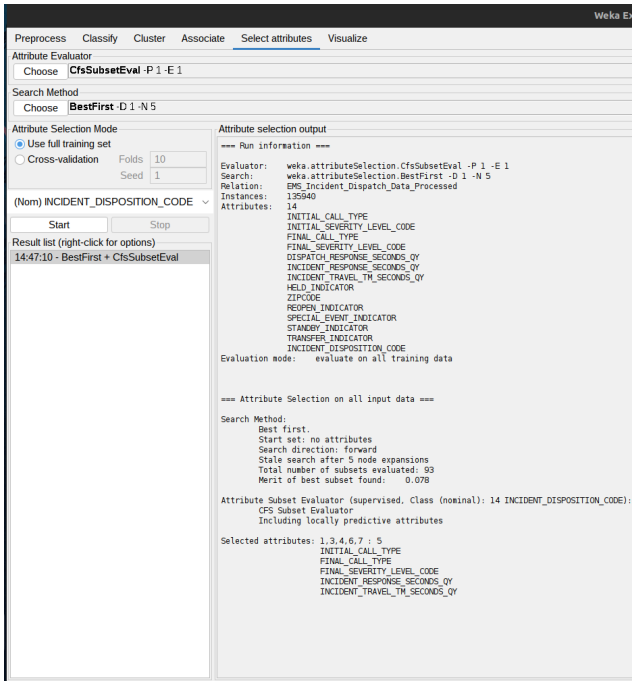


OneRAttributeEval

Evaluates accuracy of each attribute using a OneR classifier

1 for each predictor P
2 for each value V of the predictor, generate rule as
3 find the most frequent class c
4 create a rule if (P = V) then c
5 compute the error rate of the rule
6 select predictor with minimum error rate for its rules

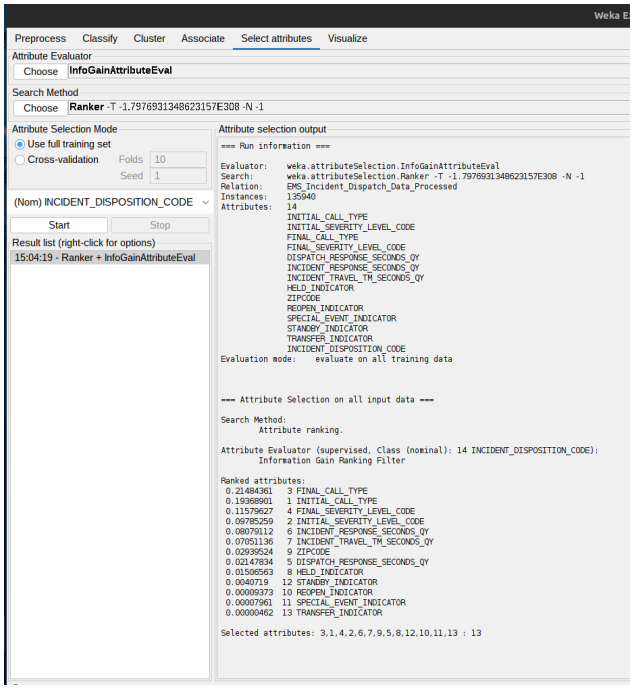
Cutoff: 64.75
Removed attributes:
SPECIAL_EVENT_INDICATOR
REOPEN_INDICATOR
TRANSFER_INDICATOR
HELD_INDICATOR
DISPATCH_RESPONSE_SECONDS_QY



CfsSubsetEval

Evaluates the importance of a subset of attributes by maximizing individual predictive ability and minimizing redundancy

Removed attributes:
INITIAL_SEVERITY_LEVEL_CODE
DISPATCH_RESPONSE_SECONDS_QY
HELD_INDICATOR
ZIPCODE
REOPEN_INDICATOR
SPECIAL_EVENT_INDICATOR
STANDBY_INDICATOR
TRANSFER_INDICATOR



InfoGainAttributeEval

Evaluates attributes based on information gain of the class attribute
InfoGain(C,A) = H(C) - H(C | A)

Cutoff: 0.01
Removed attributes:
STANDBY_INDICATOR
REOPEN_INDICATOR
SPECIAL_EVENT_INDICATOR
TRANSFER_INDICATOR

Self Selection

Removed “INDICATOR” attributes as their distribution was extremely unbalanced, with virtually all instances having the value of N. Also removed “ZIPCODE” as we thought the geographical information of each scenario would be less important than the time, severity codes, and other features.
Removed attributes: HELD_INDICATOR, ZIPCODE, REOPEN_INDICATOR, SPECIAL_EVENT_INDICATOR, STANDBY_INDICATOR, TRANSFER_INDICATOR

Model Classifiers

NaiveBayes
Performs probabilistic analysis for which class most likely an instance belongs to.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Used Weka Implementation

J48
An open source implementation in Java of a Decision Tree Classifier.
Used Weka Implementation

OneR
Construct a predictor one attribute at a time and selects the one with the lowest error rate.
Pseudo code:

```
1    for each predictor P
2    for each value V of the predictor, generate rule as
3    find the most frequent class c
4    create a rule if (P = V) then c
5    compute the error rate of the rule
6    select predictor with minimum error rate for its rules
```

Used Weka Implementation

FilteredClassifier
Runs an arbitrary classifier on data that was filtered arbitrarily
Used Weka Implementation

Part 5 – Results and Analysis

Part 5.1 – Results

CorrelationAttributeEval

with NaiveBayes

```
=== Summary ===

Correctly Classified Instances       15940               58.6267 %
Incorrectly Classified Instances     11249               41.3733 %
Kappa statistic                          0.2039
Mean absolute error                      0.1029
Root mean squared error                  0.2609
Relative absolute error                 85.3784 %
Root relative squared error            106.2516 %
Total Number of Instances            27189

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  C
                 0.801    0.541    0.731      0.801    0.765      0.275    0.703     0.800     8
                 0.277    0.061    0.283      0.277    0.280      0.218    0.713     0.201     9
                 0.186    0.032    0.124      0.186    0.149      0.127    0.746     0.107     8
                 0.122    0.035    0.124      0.122    0.123      0.087    0.714     0.095     9
                 0.134    0.054    0.338      0.134    0.192      0.120    0.632     0.284     9
                 0.152    0.020    0.179      0.152    0.165      0.143    0.823     0.140     9
                 0.580    0.023    0.044      0.580    0.081      0.155    0.833     0.059     9
                 0.878    0.014    0.363      0.878    0.514      0.559    0.981     0.443     8
                 0.000    0.000    0.000      0.000    0.000      -0.000   0.224     0.000     9
Weighted Avg.    0.586    0.367    0.570      0.586    0.568      0.231    0.699     0.597

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
 14105   810   453   558   788   239   357   285     4 |     a = 82
   925   599   128   149   172    86    71    30     4 |     b = 90
   284    82   119    12    57    34    37    13     1 |     c = 87
   502   136    86   130   104    49    50    11     0 |     d = 96
  3167   404    66   169   625   123    60    37     0 |     e = 93
   272    77   109    30   104   117    59     2     0 |     f = 91
    10     1     1     4     0     4    29     1     0 |     g = 94
    22     7     0     0     1     0     0   216     0 |     h = 83
     2     0     0     0     0     0     0     0     0 |     i = 95
```

with J48

```
=== Summary ===

Correctly Classified Instances       17925               65.9274 %
Incorrectly Classified Instances      9264               34.0726 %
Kappa statistic                          0.1578
Mean absolute error                      0.1064
Root mean squared error                  0.236
Relative absolute error                 88.2482 %
Root relative squared error             96.1332 %
Total Number of Instances            27189

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  C
                 0.960    0.814    0.684      0.960    0.799      0.242    0.716     0.792     8
                 0.103    0.021    0.301      0.103    0.154      0.137    0.711     0.195     9
                 0.443    0.007    0.587      0.443    0.505      0.500    0.766     0.330     8
                 0.017    0.003    0.165      0.017    0.031      0.041    0.670     0.083     9
                 0.068    0.017    0.446      0.068    0.118      0.120    0.636     0.271     9
                 0.058    0.004    0.288      0.058    0.097      0.119    0.763     0.118     9
                 0.000    0.000    ?          0.000    ?          ?        0.800     0.043     9
                 0.593    0.006    0.493      0.593    0.539      0.536    0.939     0.394     8
                 0.000    0.000    ?          0.000    ?          ?        0.576     0.000     9
Weighted Avg.    0.659    0.532    ?          0.659    ?          ?        0.705     0.592

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
 16894   250    67    35   204    35     0   114     0 |     a = 82
  1723   223    54    26    91    33     0    14     0 |     b = 90
   298    21   283     2    20    12     0     3     0 |     c = 87
   916    62    31    18    30     6     0     5     0 |     d = 96
  4097   145    35    21   316    24     0    13     0 |     e = 93
   627    34    11     7    45    45     0     1     0 |     f = 91
    48     0     1     0     0     1     0     0     0 |     g = 94
    93     5     0     0     2     0     0   146     0 |     h = 83
     2     0     0     0     0     0     0     0     0 |     i = 95
```

with OneR

```
=== Summary ===

Correctly Classified Instances       17840               65.6148 %
Incorrectly Classified Instances      9349               34.3852 %
Kappa statistic                          0.0549
Mean absolute error                      0.0764
Root mean squared error                  0.2764
Relative absolute error                 63.3891 %
Root relative squared error            112.5961 %
Total Number of Instances            27189

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  C
                 0.995    0.949    0.658      0.995    0.792      0.152    0.523     0.658     8
                 0.020    0.002    0.483      0.020    0.038      0.085    0.509     0.088     9
                 0.002    0.000    0.200      0.002    0.003      0.016    0.501     0.024     8
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.039     9
                 0.049    0.005    0.659      0.049    0.092      0.148    0.522     0.195     9
                 0.073    0.003    0.421      0.073    0.124      0.166    0.535     0.057     9
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.002     9
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.009     8
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.000     9
Weighted Avg.    0.656    0.616    ?          0.656    ?          ?        0.520     0.470

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
 17510    19     2     0    50    18     0     0     0 |     a = 82
  2061    43     1     0    34    25     0     0     0 |     b = 90
   599     9     1     0    12    18     0     0     0 |     c = 87
  1060     2     0     0     4     2     0     0     0 |     d = 96
  4402     5     1     0   230    13     0     0     0 |     e = 93
   685    11     0     0    18    56     0     0     0 |     f = 91
    49     0     0     0     0     1     0     0     0 |     g = 94
   245     0     0     0     1     0     0     0     0 |     h = 83
     2     0     0     0     0     0     0     0     0 |     i = 95
```

with FilteredClassifier

```
=== Summary ===

Correctly Classified Instances       18108               66.6005 %
Incorrectly Classified Instances      9081               33.3995 %
Kappa statistic                          0.1448
Mean absolute error                      0.1066
Root mean squared error                  0.2314
Relative absolute error                 88.4574 %
Root relative squared error             94.2715 %
Total Number of Instances            27189

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  C
                 0.977    0.857    0.677      0.977    0.800      0.232    0.725     0.803     8
                 0.062    0.010    0.351      0.062    0.106      0.120    0.744     0.213     9
                 0.463    0.007    0.605      0.463    0.525      0.520    0.782     0.362     8
                 0.006    0.001    0.231      0.006    0.011      0.030    0.701     0.092     9
                 0.054    0.007    0.627      0.054    0.100      0.148    0.656     0.295     9
                 0.071    0.003    0.396      0.071    0.121      0.159    0.834     0.158     9
                 0.000    0.000    ?          0.000    ?          ?        0.775     0.050     9
                 0.683    0.006    0.508      0.683    0.582      0.584    0.974     0.452     8
                 0.000    0.000    ?          0.000    ?          ?        0.585     0.000     9
Weighted Avg.    0.666    0.557    ?          0.666    ?          ?        0.720     0.608

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
 17196   106    76     7    70    19     0   125     0 |     a = 82
  1894   135    47     5    40    27     0    16     0 |     b = 90
   299    16   296     0    11    16     0     1     0 |     c = 87
   965    48    30     6    10     3     0     6     0 |     d = 96
  4269    62    29     7   252    18     0    14     0 |     e = 93
   668    17    10     1    18    55     0     1     0 |     f = 91
    47     1     1     0     0     1     0     0     0 |     g = 94
    77     0     0     0     1     0     0   168     0 |     h = 83
     2     0     0     0     0     0     0     0     0 |     i = 95
```

OneRAttributeEval

with NaiveBayes

```
=== Summary ===

Correctly Classified Instances       15598               57.3688 %
Incorrectly Classified Instances     11591               42.6312 %
Kappa statistic                          0.2283
Mean absolute error                      0.1054
Root mean squared error                  0.2617
Relative absolute error                 87.4112 %
Root relative squared error            106.5851 %
Total Number of Instances            27189

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
                 0.747    0.453    0.752      0.747    0.750      0.294    0.701     0.797
                 0.324    0.071    0.282      0.324    0.302      0.237    0.728     0.221
                 0.219    0.032    0.140      0.219    0.171      0.150    0.753     0.095
                 0.100    0.024    0.144      0.100    0.118      0.090    0.718     0.097
                 0.225    0.097    0.324      0.225    0.266      0.149    0.636     0.289
                 0.281    0.036    0.185      0.281    0.223      0.200    0.842     0.162
                 0.500    0.019    0.047      0.500    0.085      0.148    0.843     0.087
                 0.846    0.012    0.396      0.846    0.540      0.574    0.969     0.498
                 0.000    0.000    0.000      0.000    0.000      -0.000   0.502     0.000
Weighted Avg.    0.574    0.318    0.582      0.574    0.575      0.253    0.700     0.598

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
 13154  1005   445   374  1600   462   312   244     3 |     a = 82
   755   701   125   109   248   142    53    31     0 |     b = 90
   221    93   140    15    76    59    25     9     1 |     c = 87
   487   150    79   107   118    73    43    11     0 |     d = 96
  2683   446    85   114  1047   210    46    20     0 |     e = 93
   163    73   124    21   138   216    33     2     0 |     f = 91
     9     4     4     3     2     3    25     0     0 |     g = 94
    24    13     0     0     1     0     0   208     0 |     h = 83
     2     0     0     0     0     0     0     0     0 |     i = 95
```

with J48

```
=== Summary ===

Correctly Classified Instances       17888               65.7913 %
Incorrectly Classified Instances      9301               34.2087 %
Kappa statistic                          0.1532
Mean absolute error                      0.108
Root mean squared error                  0.2379
Relative absolute error                 89.6258 %
Root relative squared error             96.9058 %
Total Number of Instances            27189

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
                 0.958    0.817    0.683      0.958    0.797      0.234    0.697     0.777
                 0.110    0.020    0.328      0.110    0.165      0.152    0.681     0.184
                 0.305    0.007    0.517      0.305    0.384      0.386    0.768     0.247
                 0.000    0.001    0.000      0.000    0.000      -0.007   0.665     0.072
                 0.076    0.021    0.431      0.076    0.130      0.122    0.633     0.266
                 0.096    0.005    0.359      0.096    0.152      0.174    0.745     0.128
                 0.000    0.000    ?          0.000    ?          ?        0.816     0.037
                 0.659    0.006    0.503      0.659    0.570      0.571    0.948     0.437
                 0.000    0.000    ?          0.000    ?          ?        0.761     0.000
Weighted Avg.    0.658    0.534    ?          0.658    ?          ?        0.689     0.580

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
 16864   241    76    18   230    52     0   118     0 |     a = 82
  1698   238    39    14   127    29     0    19     0 |     b = 90
   362    35   195     0    18    24     0     5     0 |     c = 87
   978    39    25     0    14     4     0     8     0 |     d = 96
  4100   130    29     4   355    23     0    10     0 |     e = 93
   570    38    12     0    76    74     0     0     0 |     f = 91
    47     2     1     0     0     0     0     0     0 |     g = 94
    78     3     0     0     3     0     0   162     0 |     h = 83
     2     0     0     0     0     0     0     0     0 |     i = 95
```

with OneR

```
=== Summary ===

Correctly Classified Instances       17847               65.6405 %
Incorrectly Classified Instances      9342               34.3595 %
Kappa statistic                          0.0557
Mean absolute error                      0.0764
Root mean squared error                  0.2763
Relative absolute error                 63.3416 %
Root relative squared error            112.554  %
Total Number of Instances            27189

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
                 0.995    0.949    0.658      0.995    0.792      0.153    0.523     0.658
                 0.016    0.002    0.366      0.016    0.030      0.062    0.507     0.084
                 0.005    0.000    0.429      0.005    0.009      0.043    0.502     0.025
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.039
                 0.050    0.005    0.678      0.050    0.094      0.153    0.523     0.197
                 0.082    0.003    0.481      0.082    0.140      0.190    0.540     0.065
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.002
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.009
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.000
Weighted Avg.    0.656    0.615    ?          0.656    ?          ?        0.521     0.470

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
 17513    19     2     0    44    21     0     0     0 |     a = 82
  2074    34     1     0    33    22     0     0     0 |     b = 90
   597    12     3     0     9    18     0     0     0 |     c = 87
  1060     6     0     0     2     0     0     0     0 |     d = 96
  4400     9     1     0   234     7     0     0     0 |     e = 93
   673    11     0     0    23    63     0     0     0 |     f = 91
    48     2     0     0     0     0     0     0     0 |     g = 94
   246     0     0     0     0     0     0     0     0 |     h = 83
     2     0     0     0     0     0     0     0     0 |     i = 95
```

with FilteredClassifier

```
=== Summary ===

Correctly Classified Instances       17847               65.6405 %
Incorrectly Classified Instances      9342               34.3595 %
Kappa statistic                          0.0557
Mean absolute error                      0.0764
Root mean squared error                  0.2763
Relative absolute error                 63.3416 %
Root relative squared error            112.554  %
Total Number of Instances            27189

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
                 0.995    0.949    0.658      0.995    0.792      0.153    0.523     0.658
                 0.016    0.002    0.366      0.016    0.030      0.062    0.507     0.084
                 0.005    0.000    0.429      0.005    0.009      0.043    0.502     0.025
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.039
                 0.050    0.005    0.678      0.050    0.094      0.153    0.523     0.197
                 0.082    0.003    0.481      0.082    0.140      0.190    0.540     0.065
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.002
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.009
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.000
Weighted Avg.    0.656    0.615    ?          0.656    ?          ?        0.521     0.470

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
 17513    19     2     0    44    21     0     0     0 |     a = 82
  2074    34     1     0    33    22     0     0     0 |     b = 90
   597    12     3     0     9    18     0     0     0 |     c = 87
  1060     6     0     0     2     0     0     0     0 |     d = 96
  4400     9     1     0   234     7     0     0     0 |     e = 93
   673    11     0     0    23    63     0     0     0 |     f = 91
    48     2     0     0     0     0     0     0     0 |     g = 94
   246     0     0     0     0     0     0     0     0 |     h = 83
     2     0     0     0     0     0     0     0     0 |     i = 95
```

CfsSubsetEval

with NaiveBayes

```
=== Summary ===

Correctly Classified Instances       15925               58.5715 %
Incorrectly Classified Instances     11264               41.4285 %
Kappa statistic                          0.2165
Mean absolute error                      0.1047
Root mean squared error                  0.2538
Relative absolute error                 86.8563 %
Root relative squared error            103.3856 %
Total Number of Instances            27189

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.790    0.511    0.739      0.790    0.764      0.288    0.710     0.804     82
               0.286    0.065    0.276      0.286    0.281      0.218    0.730     0.203     90
               0.261    0.040    0.134      0.261    0.178      0.160    0.758     0.125     87
               0.116    0.027    0.148      0.116    0.130      0.100    0.715     0.098     96
               0.154    0.062    0.337      0.154    0.211      0.128    0.633     0.277     93
               0.223    0.035    0.156      0.223    0.184      0.158    0.822     0.141     91
               0.260    0.011    0.042      0.260    0.073      0.101    0.771     0.049     94
               0.862    0.012    0.399      0.862    0.546      0.582    0.977     0.450     83
               0.000    0.000    0.000      0.000    0.000      -0.000   0.357     0.000     95
Weighted Avg.  0.586    0.350    0.575      0.586    0.572      0.244    0.705     0.599

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
 13904   892   563   401   988   435   175   235     6 |     a = 82
   856   618   159   117   193   152    34    35     0 |     b = 90
   289    81   167    11    28    40    16     7     0 |     c = 87
   533   102   100   124    93    76    28    12     0 |     d = 96
  2961   453   100   143   715   224    27    28     0 |     e = 93
   226    76   141    38   102   172    14     1     0 |     f = 91
    13     5    12     2     0     4    13     1     0 |     g = 94
    23    10     0     0     0     1     0   212     0 |     h = 83
     2     0     0     0     0     0     0     0     0 |     i = 95
```

with J48

```
=== Summary ===

Correctly Classified Instances       17998               66.1959 %
Incorrectly Classified Instances      9191               33.8041 %
Kappa statistic                          0.1482
Mean absolute error                      0.1069
Root mean squared error                  0.234
Relative absolute error                 88.7197 %
Root relative squared error             95.3057 %
Total Number of Instances            27189

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.969    0.836    0.680      0.969    0.799      0.238    0.717     0.796     82
               0.072    0.013    0.326      0.072    0.118      0.122    0.728     0.202     90
               0.515    0.009    0.580      0.515    0.546      0.536    0.805     0.399     87
               0.003    0.001    0.079      0.003    0.005      0.008    0.691     0.086     96
               0.060    0.017    0.428      0.060    0.105      0.107    0.642     0.272     93
               0.066    0.003    0.405      0.066    0.114      0.155    0.789     0.132     91
               0.000    0.000    0.000      0.000    0.000      -0.000   0.682     0.023     94
               0.504    0.005    0.484      0.504    0.494      0.489    0.940     0.398     83
               0.000    0.000    ?          0.000    ?          ?        0.800     0.000     95
Weighted Avg.  0.662    0.545    ?          0.662    ?          ?        0.710     0.598

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
 17057   148    89    14   179    16     0    96     0 |     a = 82
  1812   156    54    10    91    27     0    14     0 |     b = 90
   269    12   329     2    12    12     0     3     0 |     c = 87
   951    35    33     3    33     7     0     6     0 |     d = 96
  4222    79    41     5   278    12     1    13     0 |     e = 93
   599    40    19     4    57    51     0     0     0 |     f = 91
    47     1     1     0     0     1     0     0     0 |     g = 94
   114     7     1     0     0     0     0   124     0 |     h = 83
     2     0     0     0     0     0     0     0     0 |     i = 95
```

with OneR

```
=== Summary ===

Correctly Classified Instances       17800               65.4677 %
Incorrectly Classified Instances      9389               34.5323 %
Kappa statistic                          0.0473
Mean absolute error                      0.0767
Root mean squared error                  0.277
Relative absolute error                 63.6603 %
Root relative squared error            112.8367 %
Total Number of Instances            27189

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.995    0.956    0.656      0.995    0.791      0.139    0.520     0.656     82
                 0.018    0.002    0.406      0.018    0.035      0.072    0.508     0.085     90
                 0.008    0.000    0.714      0.008    0.015      0.073    0.504     0.029     87
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.039     96
                 0.040    0.005    0.639      0.040    0.075      0.129    0.517     0.190     93
                 0.071    0.002    0.491      0.071    0.125      0.179    0.535     0.061     91
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.002     94
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.009     83
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.000     95
Weighted Avg.    0.655    0.620    ?          0.655    ?          ?        0.517     0.468

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
 17517    26     2     0    45     9     0     0     0 |     a = 82
  2079    39     0     0    23    23     0     0     0 |     b = 90
   597    10     5     0    11    16     0     0     0 |     c = 87
  1058     3     0     0     4     3     0     0     0 |     d = 96
  4457     5     0     0   184     5     0     0     0 |     e = 93
   682    12     0     0    21    55     0     0     0 |     f = 91
    48     1     0     0     0     1     0     0     0 |     g = 94
   246     0     0     0     0     0     0     0     0 |     h = 83
     2     0     0     0     0     0     0     0     0 |     i = 95
```

with FilteredClassifier

```
=== Summary ===

Correctly Classified Instances       18094               66.549  %
Incorrectly Classified Instances      9095               33.451  %
Kappa statistic                          0.1439
Mean absolute error                      0.107
Root mean squared error                  0.2317
Relative absolute error                 88.7931 %
Root relative squared error             94.3729 %
Total Number of Instances            27189

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
                 0.977    0.856    0.677      0.977    0.800      0.233    0.721     0.801
                 0.058    0.010    0.341      0.058    0.099      0.113    0.751     0.216
                 0.501    0.008    0.600      0.501    0.546      0.538    0.820     0.403
                 0.004    0.001    0.211      0.004    0.007      0.023    0.694     0.090
                 0.050    0.008    0.551      0.050    0.092      0.127    0.650     0.283
                 0.070    0.002    0.505      0.070    0.123      0.180    0.821     0.155
                 0.000    0.000    ?          0.000    ?          ?        0.779     0.039
                 0.671    0.006    0.494      0.671    0.569      0.571    0.972     0.436
                 0.000    0.000    ?          0.000    ?          ?        0.835     0.000
Weighted Avg.    0.665    0.557    ?          0.665    ?          ?        0.718     0.605

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
 17192   115    81     5    80     8     0   118     0 |     a = 82
  1900   126    48     5    43    20     0    22     0 |     b = 90
   279     8   320     0    14    13     0     5     0 |     c = 87
   976    27    33     4    15     5     0     8     0 |     d = 96
  4301    59    32     5   233     6     0    15     0 |     e = 93
   628    32    18     0    38    54     0     0     0 |     f = 91
    46     1     1     0     0     1     0     1     0 |     g = 94
    80     1     0     0     0     0     0   165     0 |     h = 83
     2     0     0     0     0     0     0     0     0 |     i = 95
```

InfoGainEval

with NaiveBayes

```
=== Summary ===

Correctly Classified Instances       15823               58.1963 %
Incorrectly Classified Instances     11366               41.8037 %
Kappa statistic                          0.2155
Mean absolute error                      0.1028
Root mean squared error                  0.2608
Relative absolute error                 85.3105 %
Root relative squared error            106.213  %
Total Number of Instances            27189

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
                 0.781    0.506    0.739      0.781    0.759      0.283    0.702     0.799
                 0.297    0.066    0.281      0.297    0.289      0.226    0.730     0.209
                 0.207    0.032    0.134      0.207    0.162      0.141    0.753     0.106
                 0.147    0.034    0.151      0.147    0.149      0.115    0.726     0.108
                 0.164    0.067    0.337      0.164    0.221      0.133    0.630     0.283
                 0.208    0.027    0.181      0.208    0.194      0.169    0.833     0.145
                 0.480    0.020    0.042      0.480    0.078      0.138    0.815     0.035
                 0.813    0.013    0.358      0.813    0.497      0.534    0.961     0.486
                 0.000    0.000    0.000      0.000    0.000      -0.000   0.628     0.000
Weighted Avg.    0.582    0.347    0.576      0.582    0.572      0.241    0.701     0.597

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
 13743   892   454   522  1037   365   316   266     4 |     a = 82
   857   643   143   137   188    97    64    33     2 |     b = 90
   250    89   132    15    69    41    28    14     1 |     c = 87
   489   129    88   157    93    60    42    10     0 |     d = 96
  2961   435    74   177   764   155    53    32     0 |     e = 93
   250    80    92    30   115   160    40     3     0 |     f = 91
    10     5     5     1     0     4    24     1     0 |     g = 94
    32    12     0     0     1     1     0   200     0 |     h = 83
     2     0     0     0     0     0     0     0     0 |     i = 95
```

with J48

```
=== Summary ===

Correctly Classified Instances       17631               64.8461 %
Incorrectly Classified Instances      9558               35.1539 %
Kappa statistic                          0.1559
Mean absolute error                      0.108
Root mean squared error                  0.2428
Relative absolute error                 89.5783 %
Root relative squared error             98.9048 %
Total Number of Instances            27189

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
                 0.937    0.784    0.687      0.937    0.793      0.228    0.691     0.769
                 0.135    0.026    0.313      0.135    0.189      0.163    0.675     0.184
                 0.156    0.004    0.485      0.156    0.237      0.266    0.683     0.160
                 0.008    0.004    0.073      0.008    0.015      0.012    0.638     0.068
                 0.111    0.038    0.375      0.111    0.172      0.125    0.615     0.256
                 0.083    0.006    0.277      0.083    0.128      0.139    0.689     0.098
                 0.000    0.000    ?          0.000    ?          ?        0.758     0.019
                 0.654    0.006    0.508      0.654    0.572      0.572    0.939     0.405
                 0.000    0.000    ?          0.000    ?          ?        0.621     0.000
Weighted Avg.    0.648    0.516    ?          0.648    ?          ?        0.677     0.569

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
 16488   330    43    54   510    61     0   113     0 |     a = 82
  1579   292    20    26   182    46     0    19     0 |     b = 90
   420    49   100     2    42    20     0     6     0 |     c = 87
   945    49     9     9    43     5     0     8     0 |     d = 96
  3874   165    22    28   517    35     0    10     0 |     e = 93
   569    44    11     4    78    64     0     0     0 |     f = 91
    45     2     1     0     2     0     0     0     0 |     g = 94
    80     1     0     0     4     0     0   161     0 |     h = 83
     2     0     0     0     0     0     0     0     0 |     i = 95
```

with OneR

```
=== Summary ===

Correctly Classified Instances       17809               65.5008 %
Incorrectly Classified Instances      9380               34.4992 %
Kappa statistic                          0.0508
Mean absolute error                      0.0767
Root mean squared error                  0.2769
Relative absolute error                 63.5993 %
Root relative squared error            112.7826 %
Total Number of Instances            27189

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
                 0.994    0.952    0.657      0.994    0.791      0.143    0.521     0.657
                 0.022    0.002    0.452      0.022    0.041      0.085    0.510     0.088
                 0.003    0.000    0.333      0.003    0.006      0.030    0.501     0.024
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.039
                 0.045    0.005    0.639      0.045    0.084      0.137    0.520     0.192
                 0.064    0.003    0.412      0.064    0.110      0.153    0.530     0.053
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.002
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.009
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.000
Weighted Avg.    0.655    0.617    ?          0.655    ?          ?        0.519     0.469

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
 17502    19     1     0    56    21     0     0     0 |     a = 82
  2052    47     1     0    37    27     0     0     0 |     b = 90
   594    15     2     0    12    16     0     0     0 |     c = 87
  1059     5     0     0     4     0     0     0     0 |     d = 96
  4429     6     1     0   209     6     0     0     0 |     e = 93
   700    11     1     0     9    49     0     0     0 |     f = 91
    49     1     0     0     0     0     0     0     0 |     g = 94
   246     0     0     0     0     0     0     0     0 |     h = 83
     2     0     0     0     0     0     0     0     0 |     i = 95
```

with FilteredClassifier

```
=== Summary ===

Correctly Classified Instances       18013               66.2511 %
Incorrectly Classified Instances      9176               33.7489 %
Kappa statistic                          0.1282
Mean absolute error                      0.1075
Root mean squared error                  0.2324
Relative absolute error                 89.18   %
Root relative squared error             94.6728 %
Total Number of Instances            27189

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
                 0.977    0.873    0.673      0.977    0.797      0.211    0.718     0.799
                 0.046    0.008    0.324      0.046    0.081      0.097    0.747     0.211
                 0.399    0.007    0.565      0.399    0.468      0.465    0.807     0.326
                 0.000    0.001    0.000      0.000    0.000      -0.005   0.691     0.084
                 0.053    0.007    0.603      0.053    0.097      0.142    0.657     0.290
                 0.064    0.003    0.412      0.064    0.110      0.153    0.832     0.147
                 0.000    0.000    ?          0.000    ?          ?        0.778     0.024
                 0.654    0.006    0.511      0.654    0.574      0.574    0.970     0.426
                 0.000    0.000    ?          0.000    ?          ?        0.674     0.001
Weighted Avg.    0.663    0.567    ?          0.663    ?          ?        0.717     0.603

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
 17202    84    81     7    92    21     0   112     0 |     a = 82
  1927   100    46     5    40    27     0    19     0 |     b = 90
   334    16   255     0    12    16     0     6     0 |     c = 87
  1002    33    20     0     5     0     0     8     0 |     d = 96
  4298    55    35     2   246     6     0     9     0 |     e = 93
   674    20    13     1    13    49     0     0     0 |     f = 91
    48     1     1     0     0     0     0     0     0 |     g = 94
    85     0     0     0     0     0     0   161     0 |     h = 83
     2     0     0     0     0     0     0     0     0 |     i = 95
```

SelfSelection

with NaiveBayes

```
=== Summary ===

Correctly Classified Instances       15611               57.4166 %
Incorrectly Classified Instances     11578               42.5834 %
Kappa statistic                          0.2235
Mean absolute error                      0.103
Root mean squared error                  0.2611
Relative absolute error                 85.4645 %
Root relative squared error            106.3667 %
Total Number of Instances            27189

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
                 0.753    0.457    0.752      0.753    0.752      0.297    0.702     0.798
                 0.298    0.064    0.288      0.298    0.293      0.230    0.728     0.208
                 0.232    0.037    0.131      0.232    0.168      0.148    0.747     0.105
                 0.106    0.030    0.125      0.106    0.115      0.082    0.706     0.092
                 0.244    0.107    0.320      0.244    0.277      0.153    0.625     0.280
                 0.127    0.018    0.173      0.127    0.147      0.127    0.827     0.141
                 0.400    0.020    0.035      0.400    0.064      0.113    0.825     0.035
                 0.833    0.015    0.341      0.833    0.483      0.527    0.975     0.438
                 0.000    0.000    0.000      0.000    0.000      -0.000   0.555     0.000
Weighted Avg.    0.574    0.322    0.580      0.574    0.575      0.252    0.698     0.595

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
 13249   848   508   480  1710   201   304   290     9 |     a = 82
   781   645   140   114   283    91    74    35     1 |     b = 90
   221    86   148    11   100    37    18    15     3 |     c = 87
   488   119   107   113   130    47    53    11     0 |     d = 96
  2666   443    67   157  1133    91    52    42     0 |     e = 93
   178    90   144    26   181    98    50     3     0 |     f = 91
    12     1    11     1     3     1    20     1     0 |     g = 94
    31    10     0     0     0     0     0   205     0 |     h = 83
     1     0     1     0     0     0     0     0     0 |     i = 95
```

with J48

```
=== Summary ===

Correctly Classified Instances       17873               65.7361 %
Incorrectly Classified Instances      9316               34.2639 %
Kappa statistic                          0.1623
Mean absolute error                      0.1062
Root mean squared error                  0.2364
Relative absolute error                 88.1144 %
Root relative squared error             96.3002 %
Total Number of Instances            27189

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
                 0.954    0.801    0.686      0.954    0.798      0.243    0.712     0.787
                 0.122    0.027    0.279      0.122    0.170      0.140    0.712     0.194
                 0.452    0.008    0.590      0.452    0.512      0.506    0.759     0.347
                 0.017    0.004    0.157      0.017    0.030      0.039    0.678     0.085
                 0.073    0.018    0.450      0.073    0.125      0.125    0.639     0.272
                 0.057    0.004    0.310      0.057    0.096      0.123    0.743     0.120
                 0.000    0.000    0.000      0.000    0.000      -0.000   0.733     0.033
                 0.528    0.005    0.491      0.528    0.509      0.505    0.935     0.379
                 0.000    0.000    ?          0.000    ?          ?        0.622     0.000
Weighted Avg.    0.657    0.524    ?          0.657    ?          ?        0.702     0.590

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
 16789   330    84    46   212    41     2    95     0 |     a = 82
  1686   265    48    20   115    14     1    15     0 |     b = 90
   285    22   289     4    19    15     0     5     0 |     c = 87
   911    72    29    18    27     5     0     6     0 |     d = 96
  4057   184    19    18   338    21     0    14     0 |     e = 93
   596    66    20     7    37    44     0     0     0 |     f = 91
    40     2     1     2     3     2     0     0     0 |     g = 94
   108     8     0     0     0     0     0   130     0 |     h = 83
     2     0     0     0     0     0     0     0     0 |     i = 95
```

with OneR

```
=== Summary ===

Correctly Classified Instances       17827               65.567  %
Incorrectly Classified Instances      9362               34.433  %
Kappa statistic                          0.0516
Mean absolute error                      0.0765
Root mean squared error                  0.2766
Relative absolute error                 63.4772 %
Root relative squared error            112.6744 %
Total Number of Instances            27189

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
                 0.995    0.954    0.657      0.995    0.791      0.143    0.521     0.657
                 0.021    0.002    0.459      0.021    0.040      0.084    0.509     0.087
                 0.006    0.000    0.364      0.006    0.012      0.045    0.503     0.026
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.039
                 0.043    0.003    0.726      0.043    0.082      0.149    0.520     0.195
                 0.082    0.003    0.450      0.082    0.138      0.183    0.539     0.063
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.002
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.009
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.000
Weighted Avg.    0.656    0.618    ?          0.656    ?          ?        0.519     0.469

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
 17514    30     4     0    30    21     0     0     0 |     a = 82
  2074    45     1     0    23    21     0     0     0 |     b = 90
   601     4     4     0     8    22     0     0     0 |     c = 87
  1060     3     0     0     2     3     0     0     0 |     d = 96
  4433     7     1     0   201     9     0     0     0 |     e = 93
   685     9     1     0    12    63     0     0     0 |     f = 91
    48     0     0     0     1     1     0     0     0 |     g = 94
   246     0     0     0     0     0     0     0     0 |     h = 83
     2     0     0     0     0     0     0     0     0 |     i = 95
```

with FilteredClassifier

```
=== Summary ===

Correctly Classified Instances       18092               66.5416 %
Incorrectly Classified Instances      9097               33.4584 %
Kappa statistic                          0.1375
Mean absolute error                      0.1068
Root mean squared error                  0.2319
Relative absolute error                 88.6265 %
Root relative squared error             94.4449 %
Total Number of Instances            27189

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
                 0.979    0.868    0.674      0.979    0.798      0.222    0.720     0.799
                 0.056    0.008    0.364      0.056    0.097      0.117    0.757     0.221
                 0.468    0.007    0.618      0.468    0.533      0.528    0.794     0.372
                 0.000    0.000    0.000      0.000    0.000      -0.004   0.693     0.084
                 0.051    0.006    0.653      0.051    0.094      0.148    0.654     0.292
                 0.069    0.003    0.424      0.069    0.118      0.162    0.825     0.150
                 0.000    0.000    ?          0.000    ?          ?        0.800     0.039
                 0.659    0.006    0.486      0.659    0.560      0.561    0.974     0.418
                 0.000    0.000    ?          0.000    ?          ?        0.352     0.000
Weighted Avg.    0.665    0.564    ?          0.665    ?          ?        0.717     0.605

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
 17222    95    81     6    53    21     0   121     0 |     a = 82
  1929   121    36     0    40    19     0    19     0 |     b = 90
   298     7   299     0     8    19     0     8     0 |     c = 87
   991    28    27     0    11     4     0     7     0 |     d = 96
  4314    57    19     3   235     8     0    15     0 |     e = 93
   662    22    21     0    12    53     0     0     0 |     f = 91
    45     0     1     1     1     1     0     1     0 |     g = 94
    82     2     0     0     0     0     0   162     0 |     h = 83
     2     0     0     0     0     0     0     0     0 |     i = 95
```

Part 5.2 – Analysis

Just by looking at the "Correctly Classified Instances," we can see that our classifier models aren't particularly great. The best of these 20 models can barely predict ⅔ of unseen data correctly. However, a 66% accuracy is still something statistically significant, so we think our project suggests that there is some viability to properly training on a model on this dataset.

The main source of error likely came from the dataset itself and how the attributes don't have a strong correlation with the class. For example, the attribute HELD_INDICATOR has the highest correlation when we performed CorrelationAttributeEval, but that value was still less than 0.1. The features that we have access to don't seem like enough to capture the nuance of our data, as predicting the outcome of a medical emergency needs to take into account more information, such as a patient's medical background and a more precise metric of their current condition.

Another source of error could have come from the fact that our class distribution was very uneven, meaning that certain class labels got trained more than others. This can be reflected in the huge differences in TP & FP rates between the different class labels.

After running 4 models on 5 datasets that were each created by a different attribute selection algorithm, the following results had 66% or greater accuracy for the corresponding test dataset:

CfsSubsetEval with J48 (66.20%)

- mean absolute error: 0.1069
- root mean squared error: 0.234
- relative absolute error: 88.72%
- root relative squared error: 95.31%

CfsSubsetEval with FilteredClassifier (66.55%)

- mean absolute error: 0.107
- root mean squared error: 0.2317
- relative absolute error: 88.79%
- root relative squared error: 94.37%

CorrelationAttributeEval with FilteredClassifier (66.60%)

- mean absolute error: 0.1066
- root mean squared error: 0.2314
- relative absolute error: 88.48%
- root relative squared error: 94.27%

InfoGainEval with FilteredClassifier (66.25%)

- mean absolute error: 0.1075
- root mean squared error: 0.2324
- relative absolute error: 89.18%

- root relative squared error: 94.67%

SelfSelection with FilteredClassifier (66.54%)
- mean absolute error: 0.1068
- root mean squared error: 0.2319
- relative absolute error: 88.63%
- root relative squared error: 94.44%

Although the 5 models above had similar accuracy and error, the model using the CorrelationAttributeEval dataset with the FilteredClassifier had the highest accuracy and least amount of error (across the four available error scores). For this reason, we picked the *CorrelationAttributeEval with FilteredClassifier* model as our final model and attribute selection algorithm combination that performs best on our dataset.

Looking at the *CorrelationAttributeEval* attribute selection results, we can see that it was tied for the fewest number of attributes removed: 4. The attributes it chose to remove were remarkably similar to our *SelfSelection*, except it decided to keep the HELD_INDICATOR and SPECIAL_EVENT_INDICATOR. It decided to remove the geographical information, ZIPCODE, along with ⅗ of the INDICATOR attributes. This suggests that some of the relationships in our data are more complicated, and that more features are necessary to fully capture that complexity.

Part 6 – Conclusion and How to Reproduce Our Model

The *CorrelationAttributeEval with FilteredClassifier* model had the best results of the 20 models for this project.

Steps to Reproduce Our Model: CorrelationAttributeEval with FilteredClassifier:
1. Download the `*Files*` folder next to this report
2. Download a `*weka.jar*` file in order to run weka scripts from the command line
3. Install dependencies within Python files if not already installed, `*pip install pandas*`
4. Run the `*run_data_pipeline.sh*` script within the `*Preprocessing_Scripts*` directory
5. This will produce the processed datasets within `*processed_data/preprocessing*` and the train / test splits of the attribute selection datasets within `*processed_data/attribute_selection_arff*`
6. Open WEKA explorer
7. Load the `*EMS_Incident_Dispatch_FilteredClassifier_Train.arff*` file in the Preprocess tab
8. Click on the Classify tab
9. Click “Choose” and select FilteredClassifier under meta
10. Click “Supplied test set” under Test Options, load the `*EMS_Incident_Dispatch_FilteredClassifier_Test.arff*` and click “Close”
11. Select INCIDENT_DISPOSITION_CODE as the class
12. Click Start

13. Once the model finishes running, right-click on the run under “Results List” > “Save Model” > enter file name and select directory to save in > click “Save”

Part 7 – Team Members and Tasks Performed

Finding the Data & Building Proposal: Both
Preprocessing Initial Attempt: Justin
Preprocessing & Project Update: Both
Attribute Selection Algorithms: Justin
Classifiers: John
Results Output: John
Results Analysis: Both
Building Final Report: Both

Part 8 – Appendix and Sources

Data Source Website
https://data.cityofnewyork.us/Public-Safety/EMS-Incident-Dispatch-Data/76xm-jjuj/data

Files Attached with Report

- Initial_Data/EMS_Incident_Dispatch_Data.csv – Data downloaded from the NYC OpenData website
- Processed_Data/EMS_Incident_Dispatch_Data_Processed.arff – Data after preprocessing
- Processed_Data/EMS_Incident_Dispatch_Data_Train.arff – Train data after split
- Processed_Data/EMS_Incident_Dispatch_Data_Test.arff – Test data after split
- Attribute_Selection_Data folder – contains 10 arff files, one train and one test dataset for each type of attribute selection
- Preprocessing_Scripts - scripts used for preprocessing. To run all the preprocessing steps, run *`run_data_pipeline.sh`*
- FilteredClassifier_Model.model – Chosen model for the project

Data Attribute Descriptions

Attribute	Description
CAD_INCIDENT_ID	An incident identifier comprising the julian date and a 4 character sequence number starting at 1 each day.
INCIDENT_DATETIME	The date and time the incident was created in the dispatch system
INITIAL_CALL_TYPE *	The call type assigned at the time of incident creation.

INITIAL_SEVERITY_LEVEL_CODE	The segment(priority) assigned at the time of incident creation.
FINAL_CALL_TYPE *	The call type at the time the incident closes.
FINAL_SEVERITY_LEVEL_CODE	The segment(priority) assigned at the time the incident closes.
FIRST_ASSIGNMENT_DATETIME	The date and time the first unit is assigned.
VALID_DISPATCH_RSPNS_TIME_INDC	Indicates that the components comprising the calculation of the DISPATCH_RESPONSE_SECONDS_QY are valid.
DISPATCH_RESPONSE_SECONDS_QY	The time elapsed in seconds between the incident_datetime and the first_assignment_datetime.
FIRST_ACTIVATION_DATETIME	The date and time the first unit gives the signal that it is enroute to the location of the incident.
FIRST_ON_SCENE_DATETIME	The date and time the first unit signals that it has arrived at the location of the incident.
VALID_INCIDENT_RSPNS_TIME_INDC	Indicates that the components comprising the calculation of theINCIDENT_RESPONSE_SECONDS_QY are valid.
INCIDENT_RESPONSE_SECONDS_QY	The time elapsed in seconds between the incident_datetime and the first_on_scene_datetime.
INCIDENT_TRAVEL_TM_SECONDS_QY	The time elapsed in seconds between the first_assignment_datetime and the first_on_scene_datetime.
FIRST_TO_HOSP_DATETIME	The date and time the first unit gives the signal that it is enroute to the hospital.
FIRST_HOSP_ARRIVAL_DATETIME	The date and time the first unit signals that it has arrived at the hospital.
INCIDENT_CLOSE_DATETIME	The date and time the incident closes in the dispatch system.
HELD_INDICATOR	Indicates that for some reason a unit could not be assigned immediately
BOROUGH	The borough of the incident location.
INCIDENT_DISPATCH_AREA	The dispatch area of the incident.
ZIPCODE	The zip code of the incident.
POLICEPRECINCT	The police precinct of the incident.
CITYCOUNCILDISTRICT	The city council district.
COMMUNITYDISTRICT	The community district.
COMMUNITYSCHOOLDISTRICT	The community school district.

CONGRESSIONALDISTRICT	The congressional district.
REOPEN_INDICATOR	Indicates that at some point the incident was closed but then reopened.
SPECIAL_EVENT_INDICATOR	Indicates that the incident was a special event such as the NYC Marathon.
STANDBY_INDICATOR	Indicates that the units were assigned to stand by incase they were needed.
TRANSFER_INDICATOR	Indicates that the incident was created for the transportation of a patient from one facility (ie a hospital or nursing home) to another.
INCIDENT_DISPOSITION_CODE	A code indicating the final outcome of the incident. See incident dispositions.

Data Incident Dispositions

INCIDENT_DISPOSITION_CODE	Description
82	transporting patient
83	patient pronounced dead
87	cancelled
90	unfounded
91	condition corrected
92	treated not transported
93	refused medical aid
94	treated and transported
95	triaged at scene no transport
96	patient gone on arrival
CANCEL	cancelled
DUP	duplicate incident
NOTSNT	unit not sent
ZZZZZZ	no disposition