Classification of Emergency Response Incident Results

Justin Lee and John Kim

Introduction

EMS, or Emergency Medical Services, is a system that responds to medical emergencies

Project Goal

We want to predict the INCIDENT_DISPOSITION_CODE (final outcome of incident)

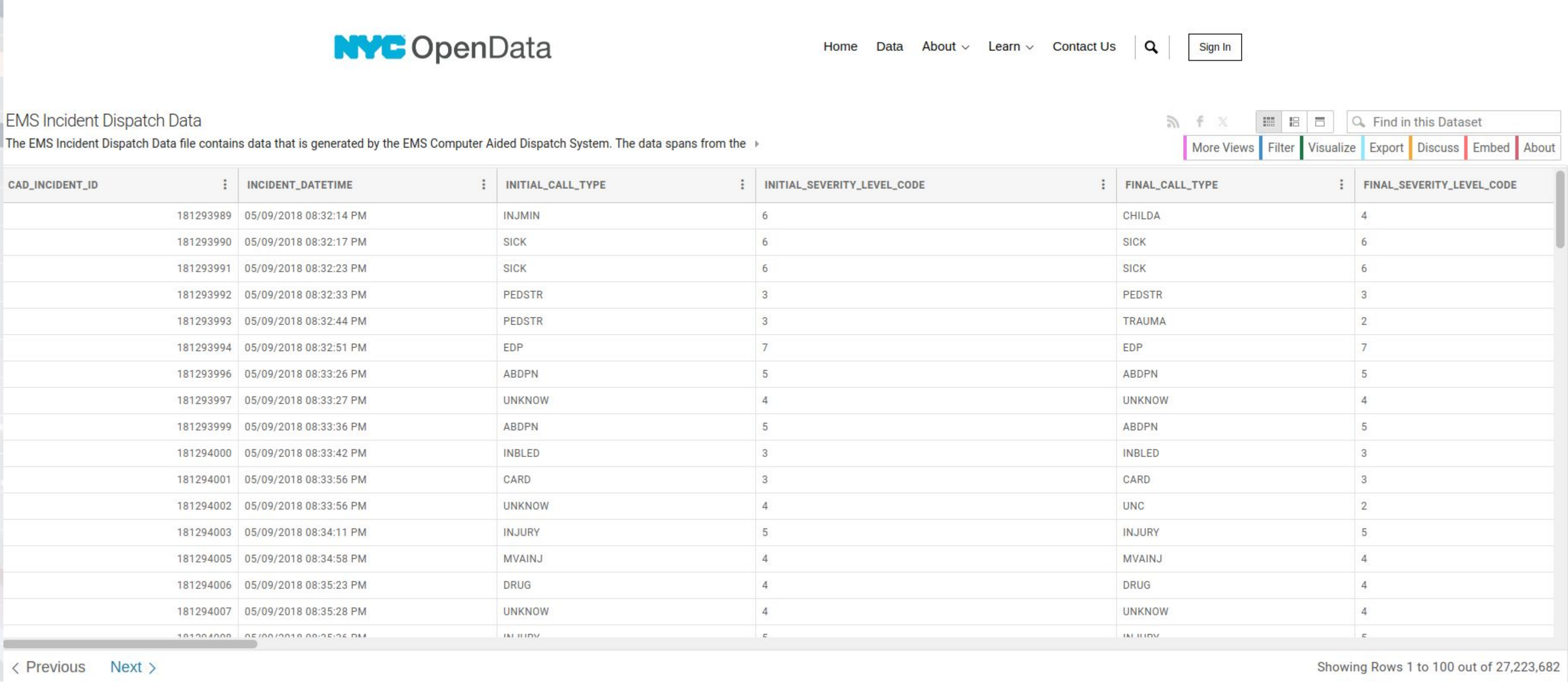
Why do we care?

- Improve EMS Efficiency
- Enhance Patient Care

Dataset Description

NYC OpenData

- Full Dataset spans 19 years
  - 25,984,643 instances
- Our Dataset: December 1st, 2023 to January 1st, 2024
  - 139,499 instances
- 30 attributes and 1 class attribute
  - i.e. INCIDENT_DISPOSITION_CODE, ZIPCODE, INCIDENT_RESPONSE_SECONDS_QY, SPECIAL_EVENT_INDICATOR, etc.



EMS Incident Dispatch Data

The EMS Incident Dispatch Data file contains data that is generated by the EMS Computer Aided Dispatch System. The data spans from the

CAD_INCIDENT_ID	INCIDENT_DATETIME	INITIAL_CALL_TYPE	INITIAL_SEVERITY_LEVEL_CODE	FINAL_CALL_TYPE	FINAL_SEVERITY_LEVEL_CODE
181293989	05/09/2018 08:32:14 PM	INJMIN	6	CHILDA	4
181293990	05/09/2018 08:32:17 PM	SICK	6	SICK	6
181293991	05/09/2018 08:32:23 PM	SICK	6	SICK	6
181293992	05/09/2018 08:32:33 PM	PEDSTR	3	PEDSTR	3
181293993	05/09/2018 08:32:44 PM	PEDSTR	3	TRAUMA	2
181293994	05/09/2018 08:32:51 PM	EDP	7	EDP	7
181293996	05/09/2018 08:33:26 PM	ABDPN	5	ABDPN	5
181293997	05/09/2018 08:33:27 PM	UNKNOW	4	UNKNOW	4
181293999	05/09/2018 08:33:36 PM	ABDPN	5	ABDPN	5
181294000	05/09/2018 08:33:42 PM	INBLED	3	INBLED	3
181294001	05/09/2018 08:33:56 PM	CARD	3	CARD	3
181294002	05/09/2018 08:33:56 PM	UNKNOW	4	UNC	2
181294003	05/09/2018 08:34:11 PM	INJURY	5	INJURY	5
181294005	05/09/2018 08:34:58 PM	MVAINJ	4	MVAINJ	4
181294006	05/09/2018 08:35:23 PM	DRUG	4	DRUG	4
181294007	05/09/2018 08:35:28 PM	UNKNOW	4	UNKNOW	4

< Previous   Next >

Showing Rows 1 to 100 out of 27,223,682

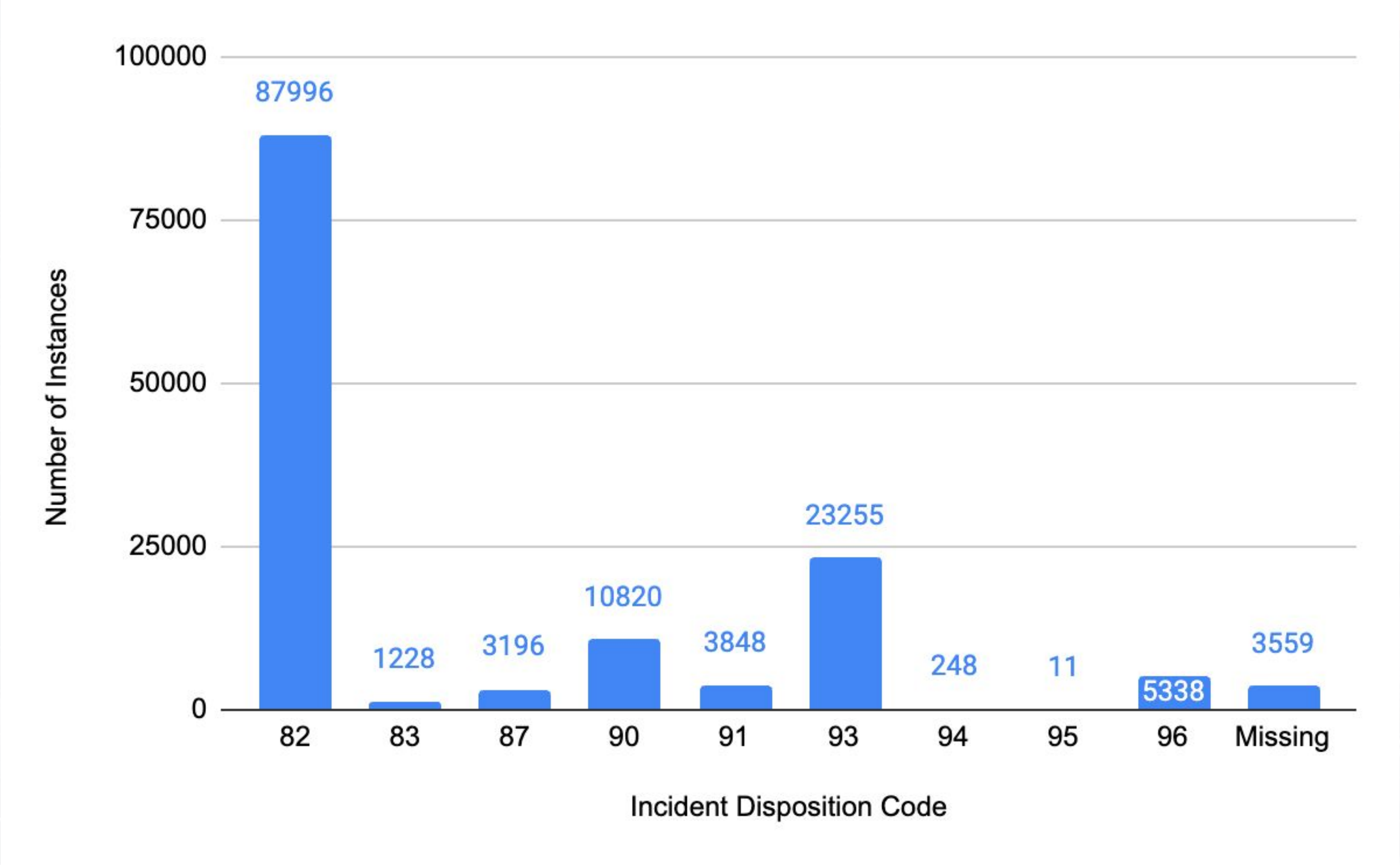
Dataset Description - Attributes

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.
DISPATCH_RESPONSE_SECONDS_QY	The time elapsed in seconds between the incident_datetime and the first_assignment_datetime.
ZIPCODE	The zip code of the incident.
…	…

Data Description - Class Labels

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

Dataset Description - Class

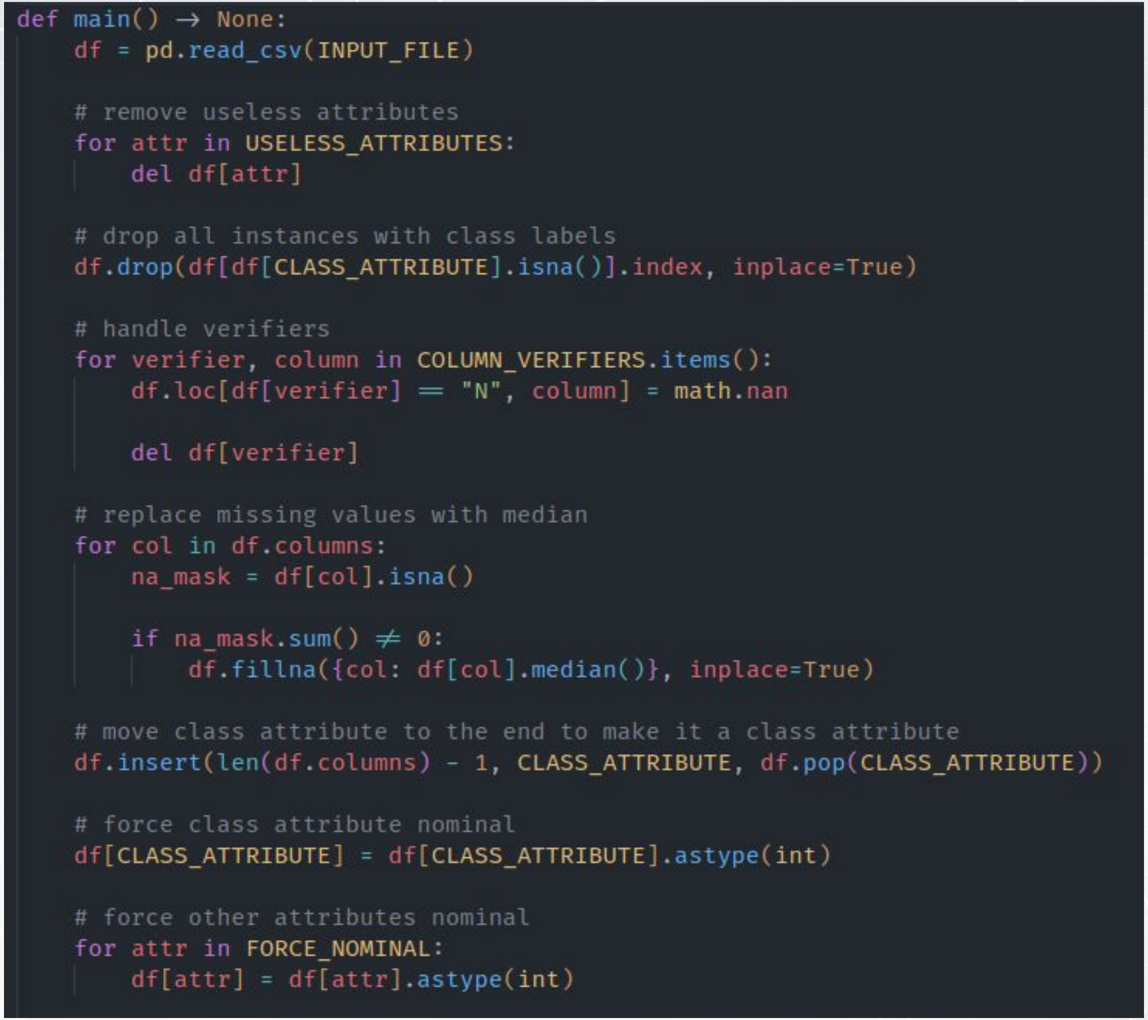


Data Description - Missing Values

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

Preprocessing

- Done in Python with Pandas
- Removed instances with missing class label (3559) (135,940 instances)
- Convert datetimes into ints
- Removed 15 attributes (30 => 15)
    - 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
- Replaced missing values with median for numeric columns
    - INCIDENT_RESPONSE_SECONDS_QY and INCIDENT_TRAVEL_TM_SECONDS_QY
- Converted from numeric to nominal
    - ZIPCODE, INITIAL_SEVERITY_LEVEL_CODE, INCIDENT_DISPOSITION_CODE, FINAL_SEVERITY_LEVEL_CODE
- Made INCIDENT_DISPOSITION_CODE the class

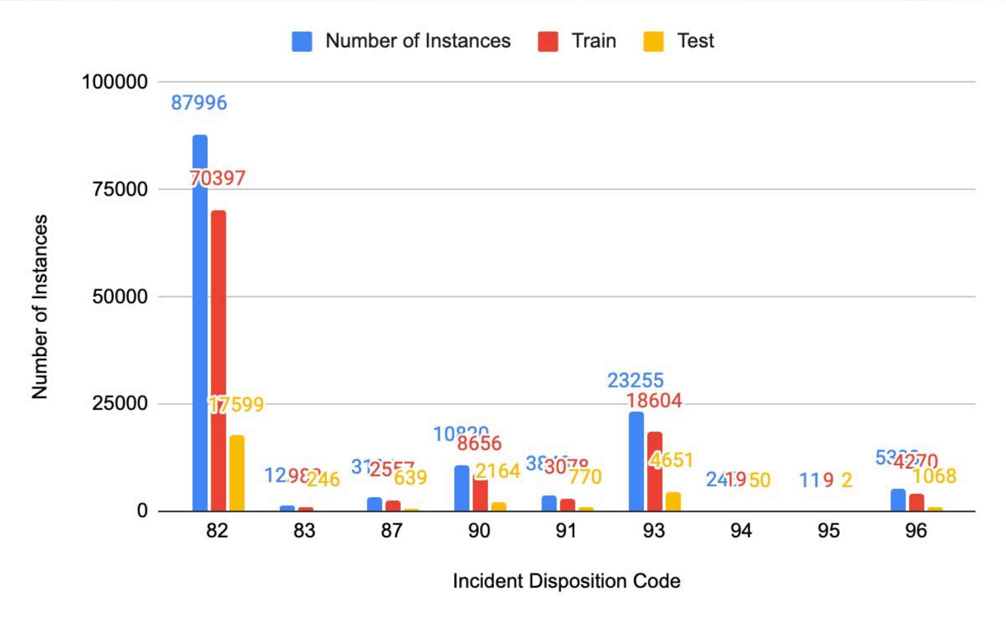


```
def main() → None:
    df = pd.read_csv(INPUT_FILE)

    # remove useless attributes
    for attr in USELESS_ATTRIBUTES:
        del df[attr]

    # drop all instances with class labels
    df.drop(df[df[CLASS_ATTRIBUTE].isna()].index, inplace=True)

    # handle verifiers
    for verifier, column in COLUMN_VERIFIERS.items():
        df.loc[df[verifier] == "N", column] = math.nan

        del df[verifier]

    # replace missing values with median
    for col in df.columns:
        na_mask = df[col].isna()

        if na_mask.sum() ≠ 0:
            df.fillna({col: df[col].median()}, inplace=True)

    # move class attribute to the end to make it a class attribute
    df.insert(len(df.columns) - 1, CLASS_ATTRIBUTE, df.pop(CLASS_ATTRIBUTE))

    # force class attribute nominal
    df[CLASS_ATTRIBUTE] = df[CLASS_ATTRIBUTE].astype(int)

    # force other attributes nominal
    for attr in FORCE_NOMINAL:
        df[attr] = df[attr].astype(int)
```

Train / Test Split

- Train 80%, Test 20%
- Simple stratified random sampling to preserve class label distribution.

Original:		Train:		Test:	
82	87996	82	70397	82	17599
93	23255	93	18604	93	4651
90	10820	90	8656	90	2164
96	5338	96	4270	96	1068
91	3848	91	3078	91	770
87	3196	87	2557	87	639
83	1228	83	982	83	246
94	248	94	198	94	50
95	11	95	9	95	2

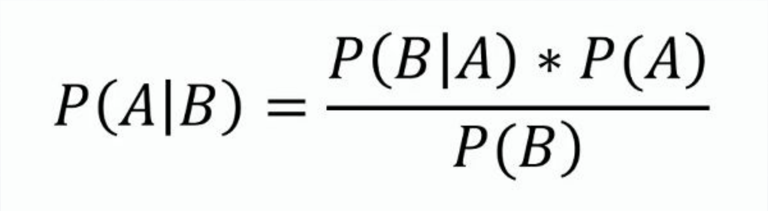


Attribute Selection Methods

Attribute Selection Algorithm	Removed Attributes
CorrelationAttributeEval (0.01)	ZIPCODE, SPECIAL_EVENT_INDICATOR, REOPEN_INDICATOR, TRANSFER_INDICATOR
OneRAttributeEval (64.75)	SPECIAL_EVENT_INDICATOR, REOPEN_INDICATOR, TRANSFER_INDICATOR, HELD_INDICATOR, DISPATCH_RESPONSE_SECONDS_QY
CfsSubsetEval	INITIAL_SEVERITY_LEVEL_CODE, DISPATCH_RESPONSE_SECONDS_QY, HELD_INDICATOR, ZIPCODE, REOPEN_INDICATOR, SPECIAL_EVENT_INDICATOR, STANDBY_INDICATOR, TRANSFER_INDICATOR
InfoGainEval (0.01)	STANDBY_INDICATOR, REOPEN_INDICATOR, SPECIAL_EVENT_INDICATOR, TRANSFER_INDICATOR
Self Eval	HELD_INDICATOR, ZIPCODE, REOPEN_INDICATOR, SPECIAL_EVENT_INDICATOR, STANDBY_INDICATOR, TRANSFER_INDICATOR

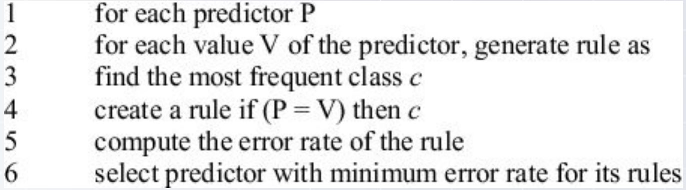
Classifier Models

- NaiveBayes



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

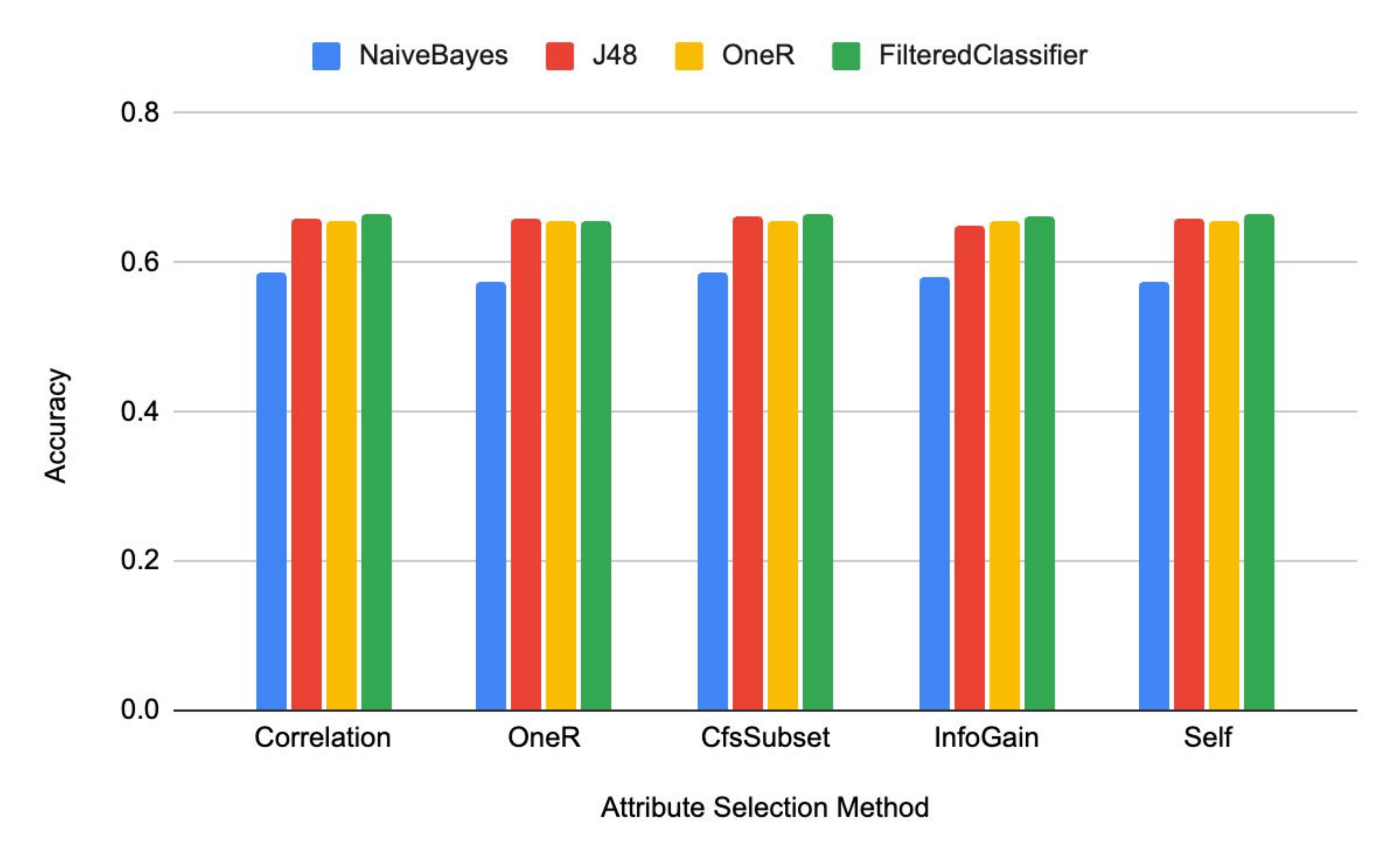
- J48
  - Implementation in Java of a Decision Tree Classifier
- OneR



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

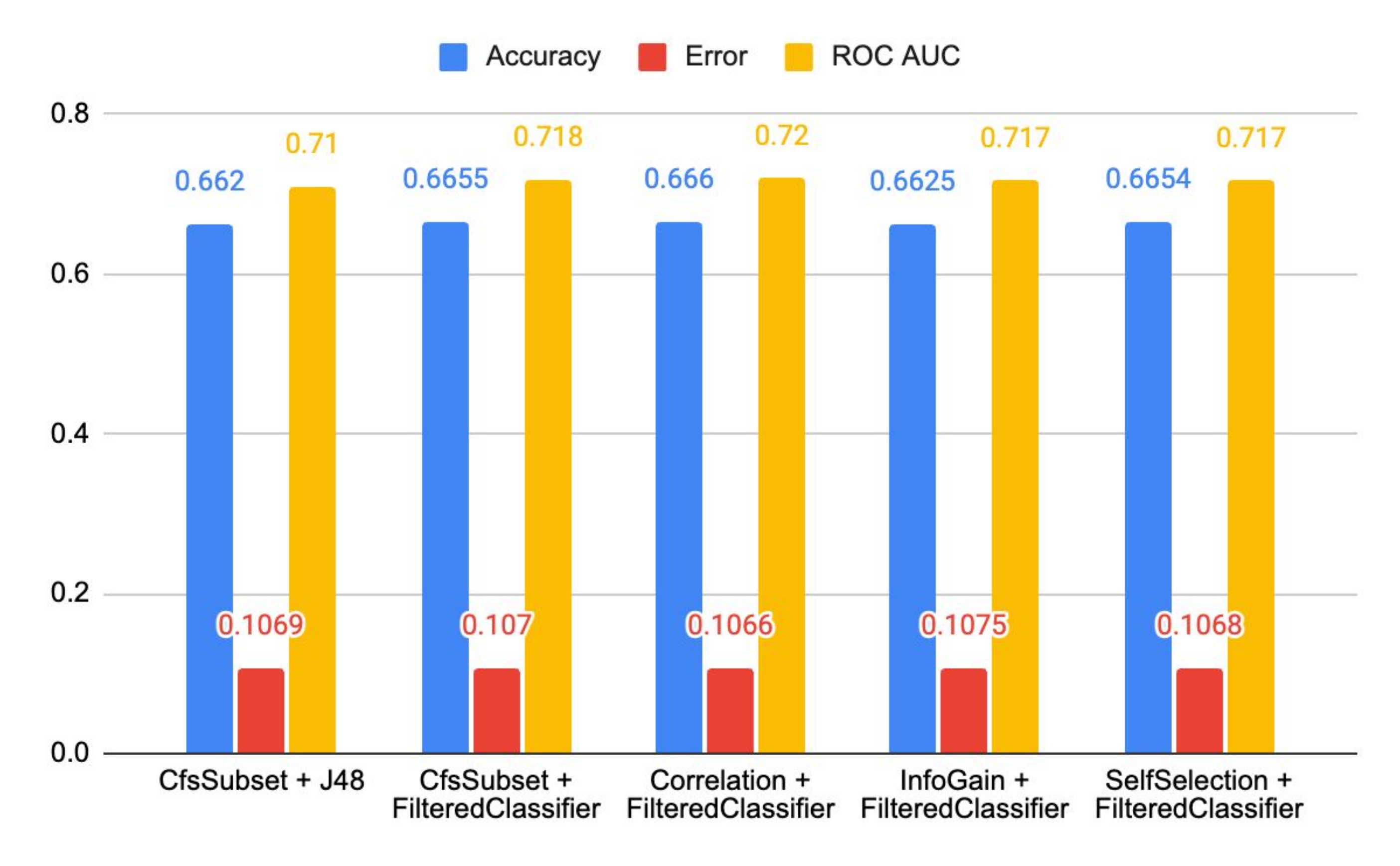
- FilteredClassifier
  - Runs an arbitrary classifier on data that was filtered arbitrarily

Results



NaiveBayes
J48
OneR
FilteredClassifier
0.8
0.6
0.4
0.2
0.0
Accuracy
Correlation
OneR
CfsSubset
InfoGain
Self
Attribute Selection Method

Results

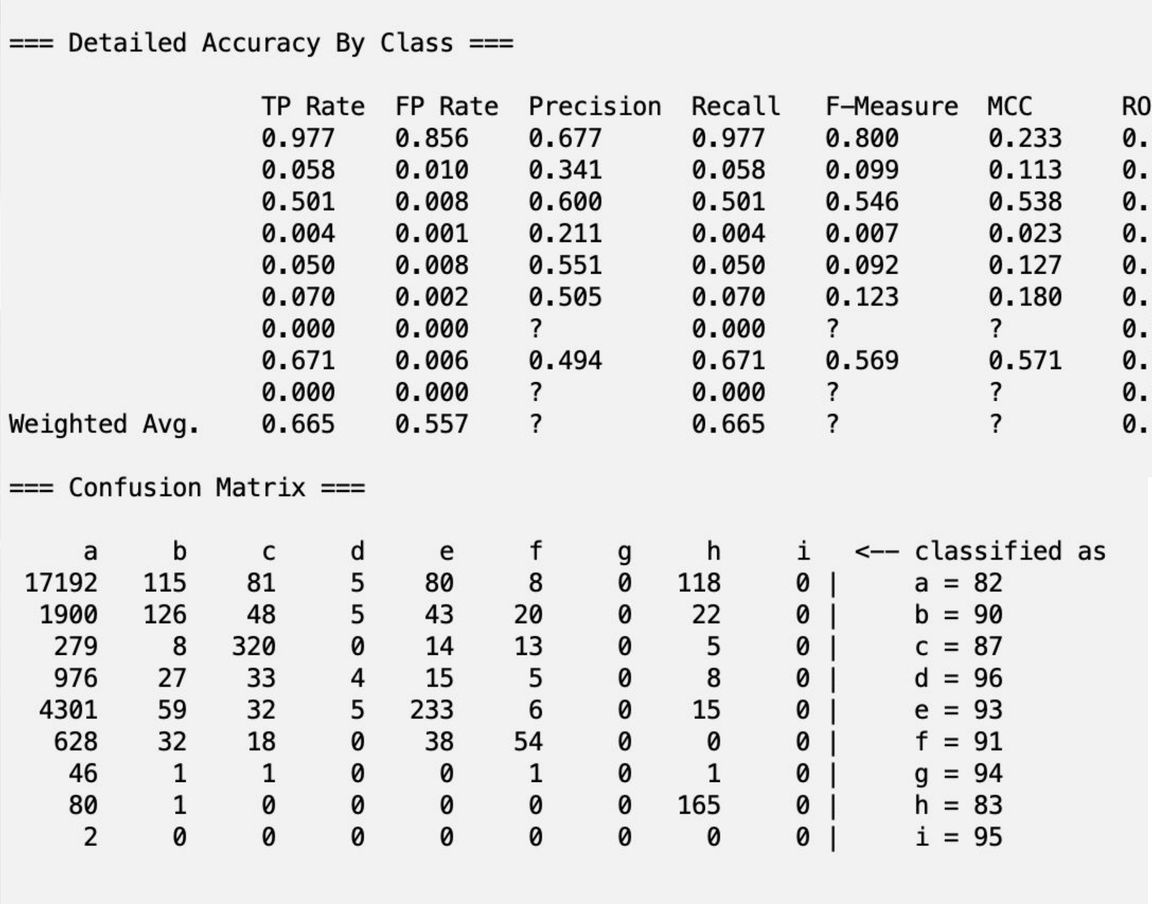


Accuracy
Error
ROC AUC
0.8
0.6
0.4
0.2
0.0
0.662
0.71
0.1069
0.6655
0.718
0.107
0.666
0.72
0.1066
0.6625
0.717
0.1075
0.6654
0.717
0.1068
CfsSubset + J48
CfsSubset + FilteredClassifier
Correlation + FilteredClassifier
InfoGain + FilteredClassifier
SelfSelection + FilteredClassifier

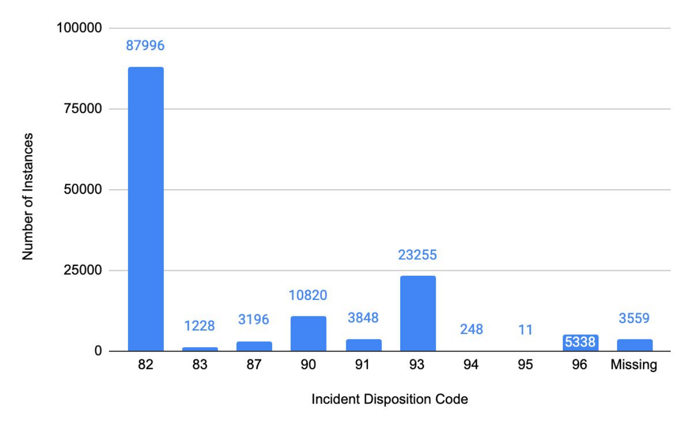
Analysis

- Best Model + Attribute Selection Algorithm
  - CorrelationAttributeEval + FilteredClassifier
  - Accuracy: 66.60%
  - Mean Absolute Error: 0.1066
  - ROC Area Under Curve: 0.72

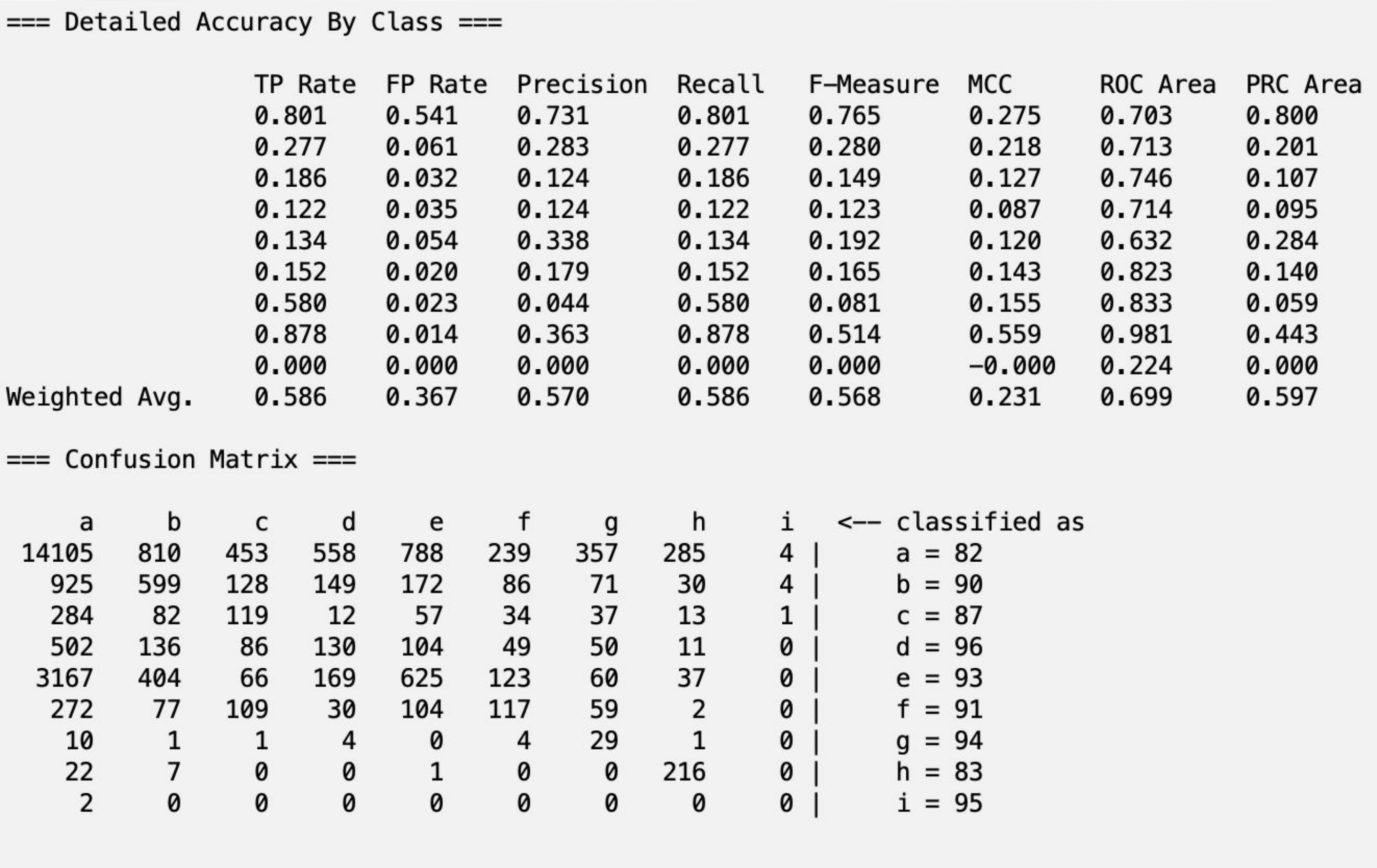
Analysis - Our Best Model



```
=== 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
```



Analysis - Another Model Comparison



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

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

Analysis Cont.

- Models fits data partially but not completely
- > ⅔ accuracy shows some potential for fitting the data
- Sources of error:
  - Not complicated enough data
  - Needs more features, i.e. a patient's medical background and a more precise metric of their current condition
  - Unbalanced classes

Reproduce

1. Download dataset and relevant files
2. Run the `*run_data_pipeline.sh*` script within the `*Preprocessing_Scripts*` directory
3. 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*`
4. Open WEKA explorer
5. Load the `*EMS_Incident_Dispatch_FilteredClassifier_Train.arff*` file in the Preprocess tab
6. Classify using `*FilteredClassifier*` and supply test dataset as `*EMS_Incident_Dispatch_FilteredClassifier_Test.arff*`
7. Select INCIDENT_DISPOSITION_CODE as the class
8. Save Model and model training / testing results

Thanks!

Any Questions?