Blake Costa  
Matthew Atwell

Brendon Kelly

**Global Terrorism Analysis**

**Introduction:**

References: <http://start.umd.edu/gtd/downloads/Codebook.pdf>

Attributes Analyzed/ Considered Under Review:

**Year(#)** : Occurrence Year between 1970 - 2016

**Month(#)**: Occurrence Month:

**Day(#)**: Occurrence Day

**Region(txt)**: Identifies the region where attack occurred. Regions are divided into the following 12 categories.

1. North America
2. Central America & Caribbean
3. South America
4. East Asia
5. Southeast Asia
6. South Asia
7. Central Asia
8. Western Europe
9. Eastern Europe
10. Middle East & North Africa
11. Sub-Saharan Africa
12. Australia & Oceania

**Casualties(#)** : Nkill + Wounded

1. Nkill: The number of total confirmed fatalities for the incident. This number includes all victims and attackers who died as a direct result of the incident.
2. Wounded**:** The number of confirmed non-fatal injuries to both perpetrators and victims.

**Inclusion Criteria:**

1. crit1 (True/False): Indicates that the violent act’s intention was a political, economic, religious, or social goal. This criteria evaluates to false if the individuals acted out of pure profit or personal motive.
2. crit2 (True/False): Indicates that the violent act was aimed at the intention to coerce, intimidate, or convey another message to a larger audience than the immediate victims.
3. crit3 (True/False): Indicates that the violent act is outside the context of legitimate warfare activities as defined by the 1949 Geneva Convention.

**Doubtterr** (1/0)**:** Indicates if there is uncertainty in determining if an incident meets the inclusion criteria.

1 = Yes - Indicates doubt as to whether the incident is an act of terrorism

0 = No - Indicates little to no doubt that the incident was an act of terrorism

**Target\_type(txt):** Record of the generalized target/victim.

1. 21 possible categories:
   1. Business, Government, Police, Military, Abortion Related, Educational Institution, Food/Water Supply, Journalist/Media, Maritime, Non-Government Organization, Private Citizen/Property, Telecommunications, Non-State Militias, Tourist, Transportation, Utilities, Violent Political Party.

**Nperps(#):** Number of terrorist participating in the incident . If multiple perpetrator groups are involved the total number of perpetrators across groups is recorded.

**Weapontype:** Information on up to four types of weapons used in an attack are recorded for each case, in addition to any information on specific weapon details reported.

1. Biological
2. Chemical
3. Radiological
4. Nuclear
5. Firearm
6. Explosive / Bombs / Dynamite
7. Fake Weapon
8. Incindiary
9. Melee
10. Vehicle
11. Sabotaged Equipment
12. Other
13. Unknown

The elements listed above were the ones that we had considered seriously for our data. And upon further inspection, realized that the best possible ones were continuous variables that helped with casualty rates.

**Preliminary Analysis:**

|  |  |
| --- | --- |
| Table 1.0: Regional Casualties | **Summary:**  During initial analysis of the provided regions, we are able to see areas of the world that suffer greatly from the destructive action of terrorism. As seen in table 1.0 the Middle East produced greatest occurrences of terrorist attacks with a high of 43,123 instances. On the contrary, Australia & Oceania had a significantly lower number of terrorism events with a total count of 256 instances recorded. |
| Table 2.0: 1970 - 2016 | **Summary:** the initial descriptive statistics of the casualties column provided some interesting insights. Analysis of the 154,524 terrorist instances can be seen in table 2.0 unveiling a overall mean of 5.31 casualties for all terrorist attacks. Using this same table our team was able to sum the casualties for all instances worldwide finding a total of **820,882** individuals that have lost their lives in terrorist attacks since 1970. |
| **Timeline Analysis Discoveries**  **Highest Casualties: 2014 Number of Attacks: 14,964.00**  **Number of Casualties: 68,711** | Table 3.0: Frequency Table of Target Type In each Region  **Summary:** Inspecting each regions breakdown by the attack target type allows us to identify common targets within each region. |

***Data Preprocessing:***

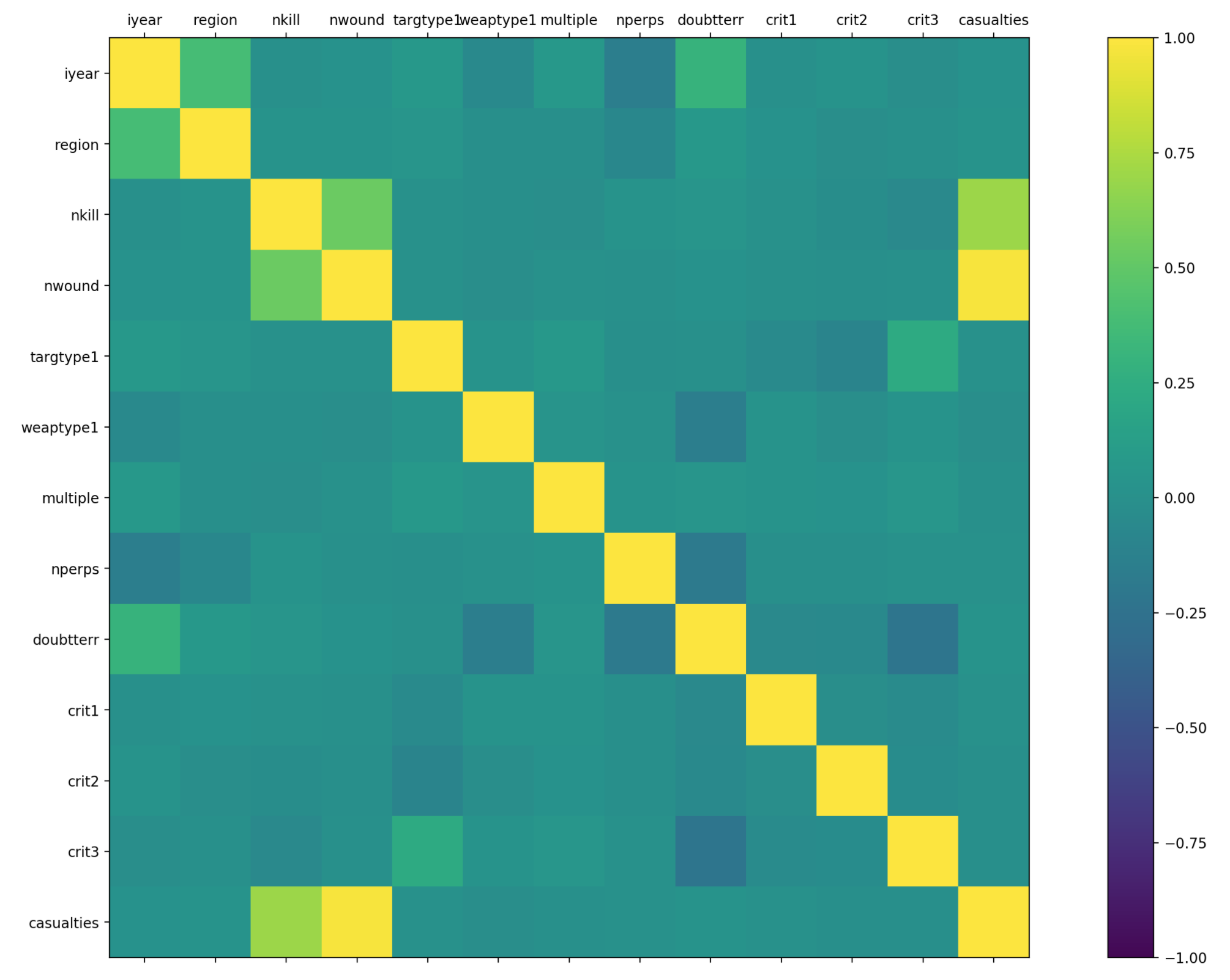
**Techniques Applied To Dataset**

* Handling of Missing or NaN Data
  + NaN = No Specific Entry was recorded
  + Instances with values of NaN were dropped during data preprocessing
    - Many machine learning algorithms misperform when presented with missing values.
    - During our first runs, we received many skewed answers when running normal summary statistics such as mean, median, and IQR. NaN values interfered heavily in this, and it was important to drop them. We also did test runs where certain instances were zero (such as NaN in a nwound row instance), and noticed that thought values and summary statistics did vary, it was not as great as initially as had been anticipated.

***Feature Selection:***

* Dimensionality Reduction:
  + Focusing on the severity of each instance was the primary goal of this analysis, which did not require all the attributes supplied in the dataset. This was because many of the attributes were repetitive or did not correlate to what we needed from the data.
    - Full list of attributes can be view in the GTA/GTD Codebook
* Aggregation:
  + In order to minimize data complexity the following variables were merged together.
    - Casualties : Nkill + Nwounded
    - The purpose behind this is that casualty markers usually are an aggregate/composite between the amount of people injured and killed. And we felt it would be best to represent this as our primary attribute for determining severity.
* Attribute Analysis:
  + **Pearson Correlation Coefficient:**  Identifies linear relationships between attributes. The result will be a value between -1 and 1 where a value below 0 indicates a low correlation or inverse relationship while values between 0 and +1 indicate a high / medium correlation or a positive relationship.
  + In order to aid discovery of relationships between attributes our group organized a Pearson Correlation Graph (Table 3.0) which visualizes attribute strength through changes in coloration. Attributes that are strongly correlated are represented in areas of bright yellow and lime green, while darker areas of green portray weak correlation. The closer the value to 1 (complete correlation) the more yellow the value. The closer the value to -1, the more negative the relationship is (as x increases, y decreases, and vice versa).
    - Strong Relationships:
      * Year & Region
      * Wounded & Killed
      * Criteria\_3 & Target\_Type
      * Doubtterr & Region

**Table 4.0**



**Discretization**

**Establishing Severity Score:**

In taking a predictive modeling approach to our data analysis, we acknowledge that dealing with continuous variables can be burdensome to interpret. This is due to the variety of outcomes that the the record can take on. Instead our group discretized each instance into severity classes of either Low, Medium, or High severity. This is it hopefully present a good metric that determines the ‘danger’ and seriousness of any given event. Severity is based around the attribute “casualties”, which is defined by nkills + nwound (both of which are columns). Thresholds for each feature are established below.

**LOW Severity:**

Anything less than the mean. (5.31 casualties).

**MEDIUM Severity:**

Between the mean and one standard deviation (5.31 and 40.79 casualties).

**HIGH Severity:**

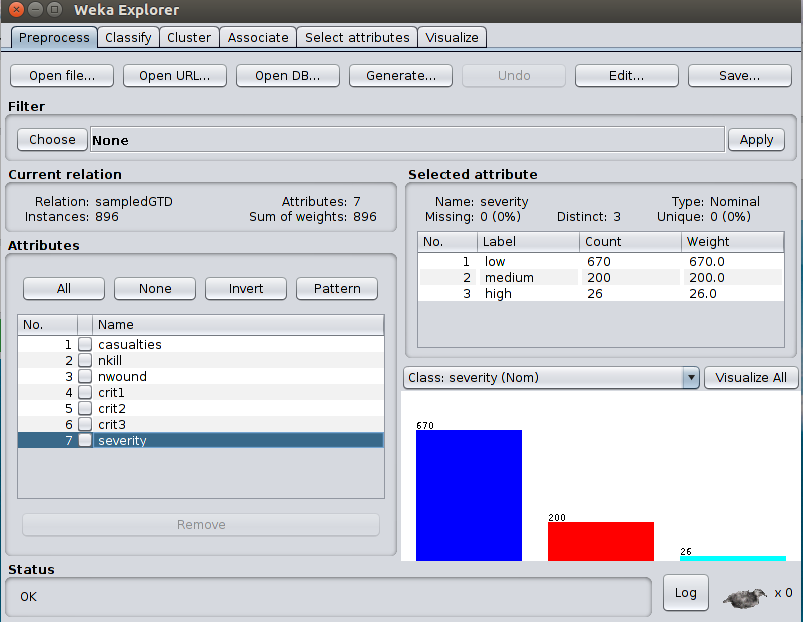
Greater than one standard deviation of data (greater than 40.79 casualties).

Another reason why this metric had been used was because we needed a representative way to categorize our data into class values, in order to run them through the algorithms via weka. Again, as of this stage, the preliminary severity scores are based solely upon the casualty rates due to time constraints.

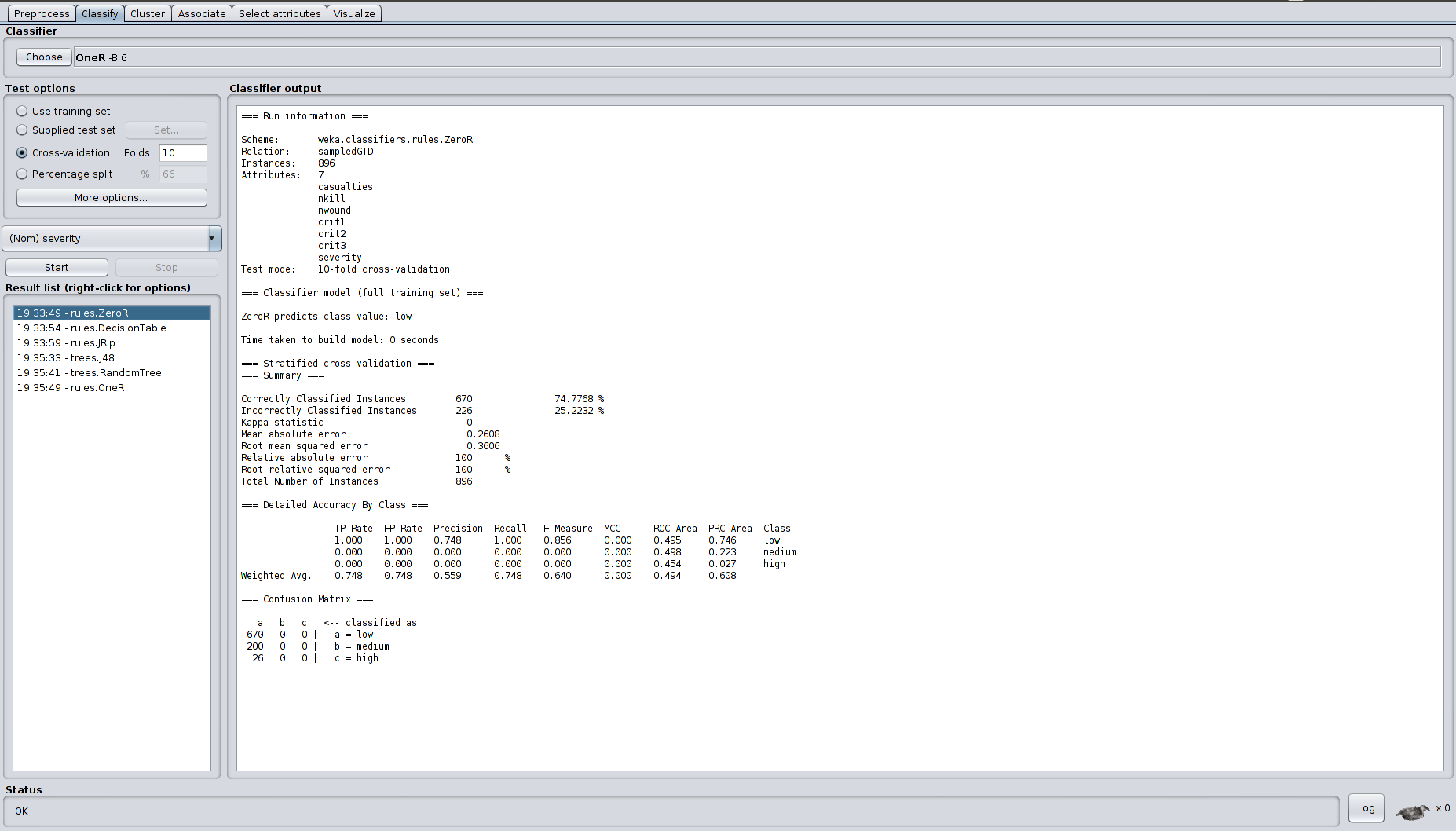
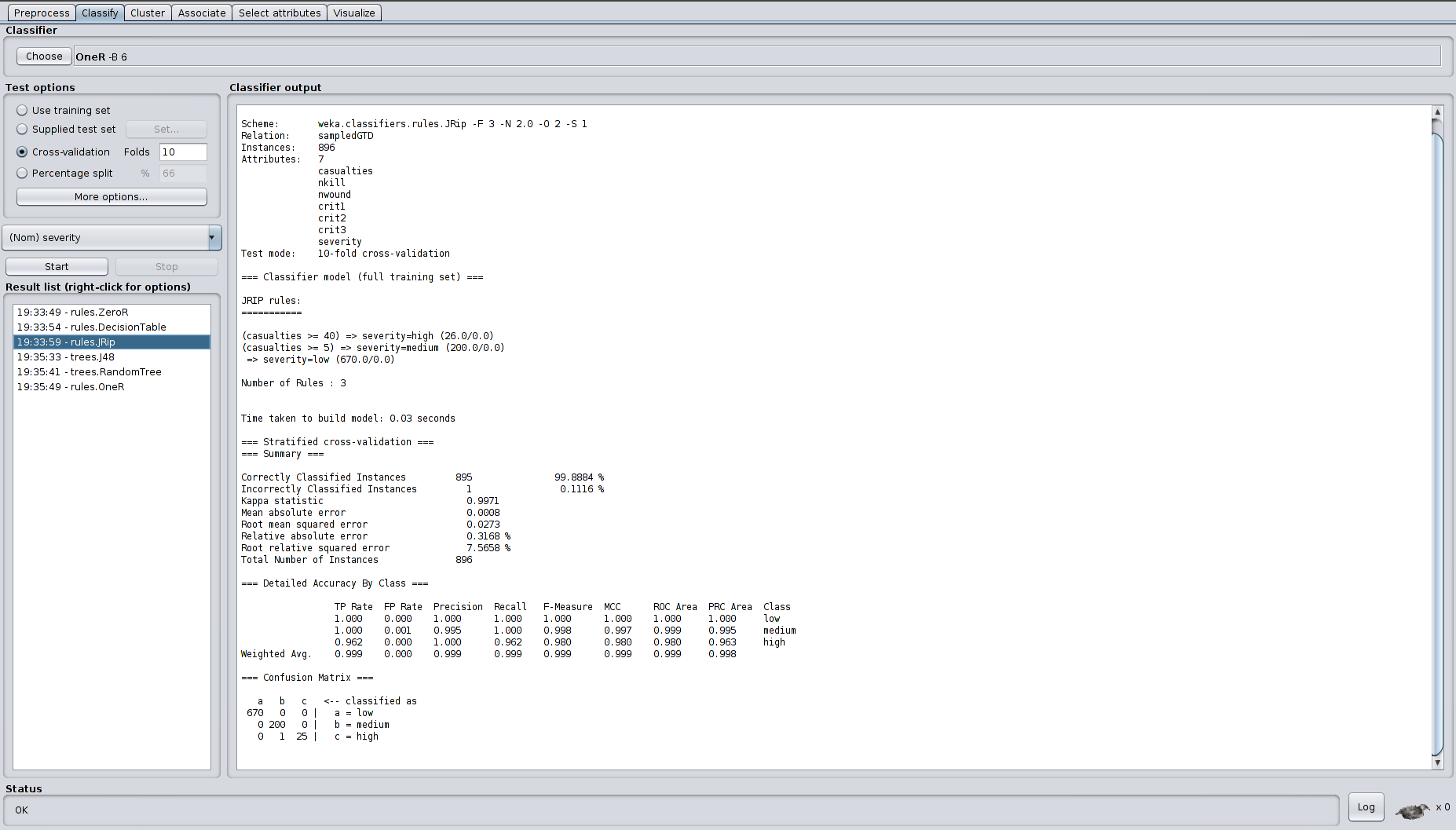
Future goals with the final presentation is to include more attributes, specifically in the usage for classification of the severity. The hope is to see how other types of data affect the outcome. The ones used, were primarily used because of the ability to be able to work with the data well, as well as avoiding some pesky Python issues. During our preliminary analysis we noticed that (the discretization was originally made in pure reflection of the number of casualties) casualties had the highest correlation in predicting the severity of a given attack. Other things that will be factored in in the future to give a more realistic metric as to the severity of an attack (as there are more factors besides casualties), will be weights, targtype, etc. This will hopefully paint a more realistic picture with the data, and what should be qualified as a ‘high’ risk incident. Though not required, assigning weighted values to nominal data (such as targtype), and feeding that through another learning algorithm such as a Neural Network will hopefully give more meaningful results. For example, taking an attribute such as ‘weapontype’and assigning a weighted value for a given instance could help give more representative features (for example, a handgun and a dirty bomb should not be on the same level).

The relevance of the algorithms listed below were being used as a baseline to determine severity, and though many of them did extremely well (95% and above, even OneR), we don’t entirely feel it’s truly representative of accurate classification of severity. ZeroR’s accuracy was at close to 75% with solely casualties (due to the fact that casualties is just a composite of nkill/nwound), but this is to be expected. Though in the future, it’s important to note better results would be obtained by factoring other things initially besides number of wounded/killed, as this raw metric doesn’t always truly show what potentially dangerous situations and trends really are. A high number of sophisticated attacks with lower casualty rates should be taken more seriously than frantic events taken on my a single individual. Though they are both equally tragic, higher levels of sophistication with technology, weapons, etc., can cause very worrying trends. This is because the potential for disarray can be much higher, as it represents an entity/NSE (Non-State Entity), rather than an individual that is targeting people. And a group that can plan politically motivated attacks, rather than crimes of passion, can cause much greater issues in long term stability than any single individual can do.

Appendix of Preliminarily Ran Values (will be updating with)



This was the default screen after importing to weka.

-