



DATA MINING WRITE UP 2

UWTSD, School of Applied Computing



LETURER, NIK WHITEHEAD PHD

1 TABLE OF CONTENTS

2	Introduction.....	3
2.1	Data Set Description.....	3
2.2	CRIPS-DM Adherences.....	4
3	Attributes Grouping.....	5
4	Pre - Processing of Data	6
5	Data/ Attribute Description.....	7
6	Post – Processing of Data	14
7	Analyses Tables	19
8	JRip – Testing.....	20
8.1	Analyses test for accident severity.....	20
8.1.1	JRip Testing.....	20
8.2	Analyses test for number of vehicles	21
8.2.1	JRip Testing.....	21
8.3	Analyses test for number of casualties	22
8.3.1	JRip Testing.....	22
9	JRip Results.....	23
10	J48 Tree - Testing	24
10.1	Analyses test for accident severity.....	24
10.1.1	J48 Tree Testing.....	24
10.2	Analyses test for number of vehicles	25
10.2.1	J48 Tree Testing.....	25
10.3	Analyses test for number of casualties	26
10.3.1	J48 Tree Testing.....	26
11	J48 Tree Results.....	27
12	RBFN Network - Testing	28
12.1	Analyses test for accident severity.....	28
12.1.1	RBFN Network Testing	28
12.2	Analyses test for number of vehicles	29
12.2.1	RBFN Network Testing	29
12.3	Analyses test for number of casualties	30

12.3.1	RBFN Network Testing	30
13	RBFN Network Results	31
14	Recommendations	32
15	Results of Modelling.....	32
16	Research Note(s)	32
17	Further Modelling Research.....	32
18	Accident Severity Test	33
18.1.1	Accident Severity Confusion Matrix.....	33
18.1.2	Confusion Matrix.....	33
19	Number of Vehicles Test	34
19.1.1	Number of Vehicles Confusion Matrix	34
19.1.2	Confusion Matrix.....	34
20	Number of Casualties Test	35
20.1.1	Number of Vehicles Casualties Matrix	35
20.1.2	Confusion Matrix.....	35
21	resultes of classification(s)	36
22	Possible future Research.....	36
23	References.....	37
24	Appendix	39
24.1	Technology Issues	39
25	Data Set Discretion.....	41

2 INTRODUCTION

This report will outline the data set that will be used. This report will also discuss the processes that this data set will go through: manipulating and sorting of the data, and importing the data set into a data warehouse. Furthermore, this report will then discuss the tools that will be used for the mining of set data and for the visualization aspect of the process (please read appendix to see further notes about the data set and the steps in the processing).

2.1 DATA SET DESCRIPTION

This data set is a subset of the road accidents and safety statistics from the UK Government under the Department of Transport [1], first published on 26th September 2013. This data set has in excess of 1,000,000 records detailing a number of different areas (Please see Appendix for Field Descriptions and Definitions), the attributes that are contained within this data set will be on page 4 under attributes grouping.

The original intention was to process and analyse accident data for South Wales / Wales, the Wales area was not available within this data set for 2010; therefore, this means that the investigation will have to look into different areas, deciding that London is a suitable location as it is a large city with an even larger population, thus meaning it should have a vast amount of data for the relevant year of 2010.

2.2 CRIPS-DM ADHERENCES

The business understanding for this data set is related to the business questions that have been asked in relation to this data set [3].

1. If mapping data is applied to this data set it could be used to reveal a pattern of which major roadways are most prone to collisions leading to the possibility that this information can be used to prevent collisions by distributing the traffic to different areas.
2. This data set can be used to help model what would be an appropriate response by the emergency services.
3. In conjunction with question two, this data could be used to try and forecast the number of fatalities that may happen on a part of the road network.
4. This data could also be used to find out if imposing greater speed limited during harsh weather could mitigate any of the contributing factors to a road crash.

For the understanding section, these attributes have been further grouped into categories. One of the main processed for the government data was to translate numerical data into its text equivalent. When considering preparation, basically ETL processes have been applied as the government data is normally very complete and has few missing records [3]. (Please see figure 1)

During the modelling phase, in which parameters are set, this data set will be cut down to a sample one location and one year to see if these parameters are scalable. This sample will then be applied to the whole data set [3]. Secondly, the data set will go through a process of iterations within all of the processing stages [3]. Lastly, for the evaluation stage, 3 values using 3 different algorithms will be tested, then evaluated to find the mean of the results to see which algorithm it best suited to be used to develop a model that will then be applied to the test data set [3].

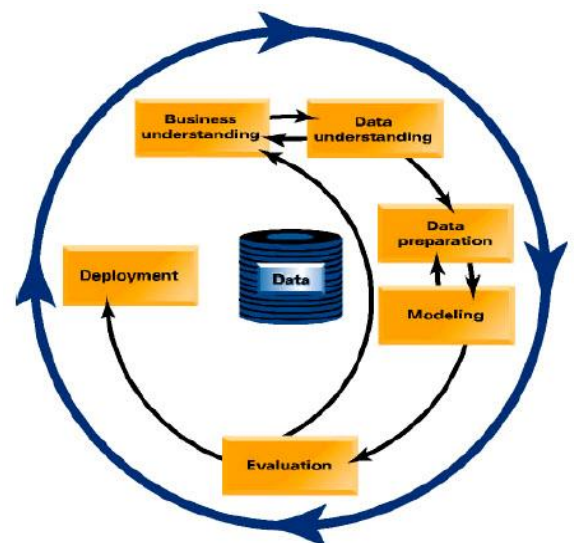


Figure 1 - Crisp-DM Chart [22]

3 ATTRIBUTES GROUPING

These data attributes can be grouped as they fall within a number of categories such as:

- Location
 - Location Easting OGSR
 - Location Northing OGSR
- Time/Date
 - Time
 - Date
 - Day of the Week
- Road Data
 - 1st Road Class
 - 1st Road Number
 - 2nd Road Class
 - 2nd Road Number
 - Junction Details
 - Junction Control
- Facilitates
 - Pedestrian Crossing Human Control
 - Pedestrian Crossing Physical Facilitates
- Conditional Variables
 - Light Conditions
 - Weather Conditions
 - Road Surface Conditions
 - Speed Limit
- Local Authority
 - Local Authority (District)
 - Local Authority (Highway)
 - Urban or Rural Area
- Accident Data
 - Accident Severity
 - Number of Vehicles
 - Number of Casualties
- Indices
 - Accident Index
 - LSOA of Accident Location
- Police Data
 - Police Force
 - Did Police Officer Attend Scene of Accident

4 PRE - PROCESSING OF DATA

This data set went through a number of processes, listed below (these Steps are part of the ETL Process) [4]. The first step of the process was to find the data source; in this case this data was taken from the UK Government Department for Transport of the Road Accidents and Safety Statistical [1]. This data set was in a CSV format and had over a 1,000,000 records; therefore, and an attempt to cut down this data so that only the relevant records were left. To obtain these relevant records a filter option was used within Microsoft Excel [2]. Thereby, allowing data to be sorted by a chosen variable; for this data set the sorting variables were the most recent year of records (which happened to be 2010) and then by the Police Force Attribute (which for this case it was 1 – i.e.: London Metropolitan Police Force).

Once these variables were set, the records dropped to the region 140,000; as a consequence, this means that rather than having the full range of variables for most of the attributes, there is now a sampling of the full range for some of the variables, for example [4]: with the Police Force Attribute, the only variable it has now is (1), whereas if it had not been sorted they range of variables would have been much larger (such as 1, 2,3 62, etc.). This was a result of the data set having any duplicate records removed so that only the unique and relevant records remained [4]. It should also be noted that it is also possible to remove attributes, and for this stage the attributes that were removed were longitude and latitude. The purpose of this removal was that as the overall data was filtered down to one location it was very unnecessary, unless mapping data were to be applied (however, in this case it is not going to be applied) [4].

The second step involved importing the arranged data set into Microsoft Access [5] where it would be used as the data warehouse. This step involved the importing of the data into a number of tables, which fall into two groupings:

1. The Master Table [6]: this tables houses the main recording for all of the records.
2. The Supporting Tables: these tables are used to hold the unique values of the data that were left from the cleaning section of the pre-processing stage.

It should be noted that when it came to the creation of the 'arrf file' [7] the normal process of having to use a text edit program was unnecessary, as this file was of such high quality. This meant that no cleaning was needed and it could be loaded directly into the data mining tool, 'WEKA' [8] and, once loaded, it could be saved out as an 'arrf file' [7].

5 DATA/ ATTRIBUTE DESCRIPTION

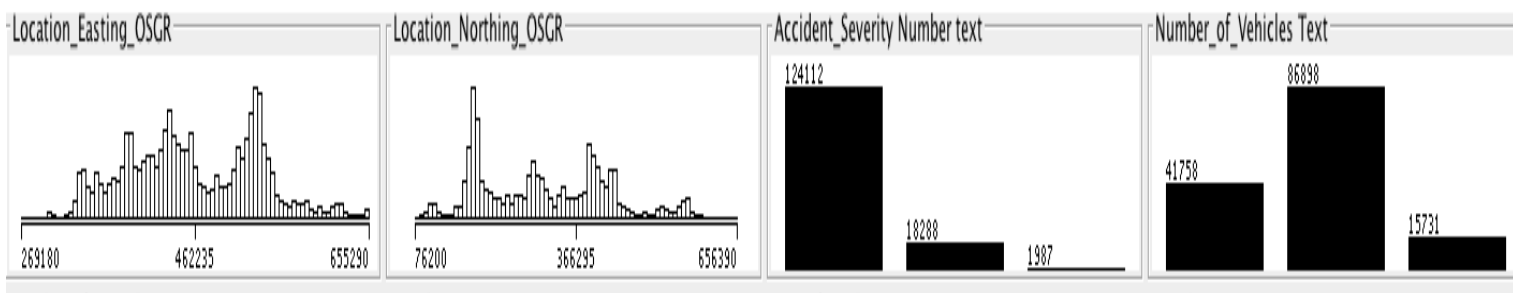


Figure 2- Depiction of Data

-Location Easting OSGR: this is to find a location using the Ordnance Survey Grid Reference with an east heading. [9]

- Location Northing OSGR: this is a way of finding a location using the Ordnance Survey Grid Reference with a north heading. [9]

Notes: with the OSGR what this is that with the Easting there is a major spike in the overall number of accidents the more towards the centre of London

- Accident Severity: this shows how severe an accident is. [9]

Notes: what can be seen with accident severity is that the over severity of an accident will most likely be high, this is mostly due to the fact that London is a very built up area

- Number of Vehicles: this shows the total number of vehicles involved within a single accident. [9]

Notes: what can be seen is that the number of vehicles will normally involve two cars, and, this is most likely due to the fact that London is a build up area with a lot of traffic

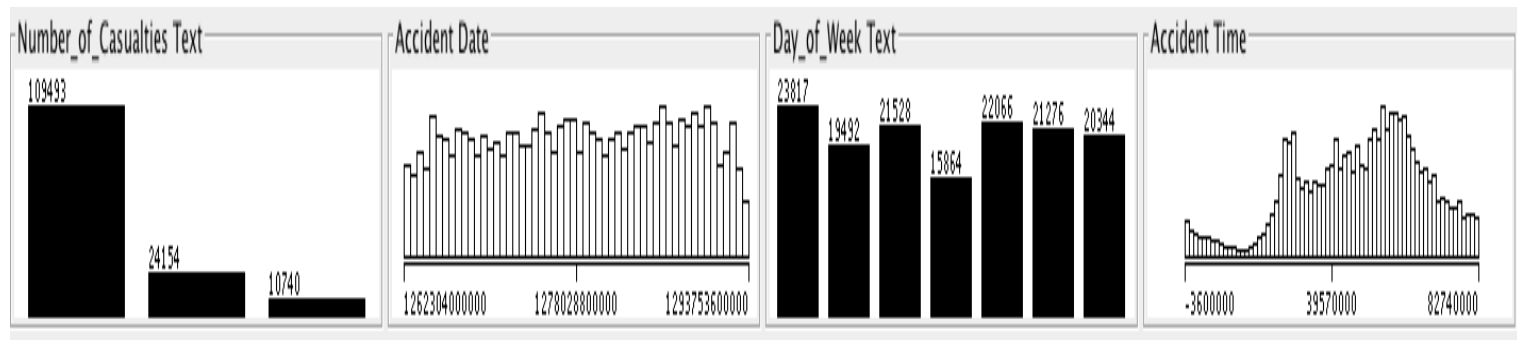


Figure 3 - Depiction of Data

Number of Casualties: this shows the total number of casualties involved in each accident. [9]

Notes: what can be seen with this is the casualties is that the majority of these death normally involve one casualty

Accident Date: this shows the dates that the accidents have taken place for the year of 2010. [9]

Notes: what can be seen with this is that at the start of year there less accidents taking place, with a major drop off just before the end of the year this is mostly likely due to charismas holiday and with an even sharper drop of right at the end, this is mostly likely due to new years day.

Day of the Week: this shows which day of the week the accident happened, shown by a number (EG: Monday will be number 1). [9]

Notes: What this shows is that at the start of the week (Monday) there are less accidents that there is after the start of the week (Tuesday – Saturday) and there will also be a slight drop off at the end of the week (Sunday) this mostly concedes with the working week

Accident Time: this shows the time of the day that the accident occurred. [9]

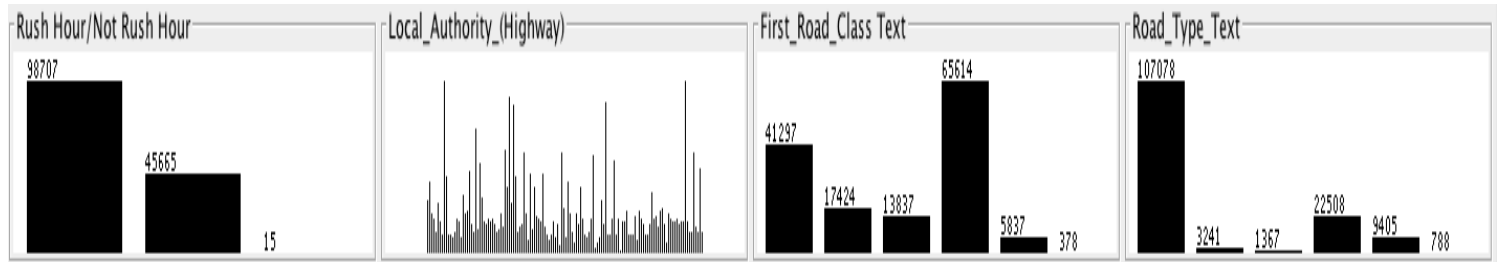


Figure 4 - Depiction of Data

Rush Hour/ Not Rush Hour: this shows that did an accident happen within the rush hour or not rush hour time period.

Notes: what can be seen here is that a large amount of the accidents have taken place during rush hour, and, this is mostly due to fact that London is a very built up area

Local Authority (Highway): this shows in which Local Authority an accident has happened, and on which highway is within that Local Authority. [9]

Notes: what can be seen with this is within the areas of London can vary this is mostly depended due to the fact that London is an urban area with large built up area (sky scrappers, etc) and suburbs (boroughs)

First Road Class: this shows the classification of the road. [9]

Notes: what this shows is that most of these accidents happen on an A Road with least amount of them happening on Motorways and A/M Roads.

Road Type: this numbers the type of features a road may have (EG: it may be a one way street or have a roundabout). [9]

Notes: What this shows is that a large majority of accidents happen on a Single Carriageway's

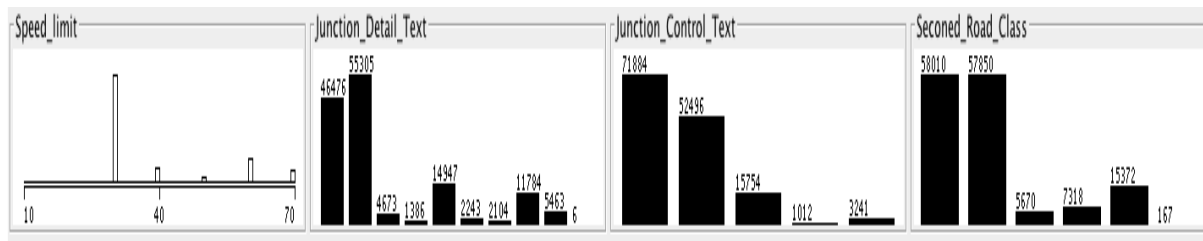


Figure 5 - Depiction of Data

Speed Limit: this shows the overall speed of a vehicle when involved in an accident. [9]

Notes: what this shows most if not all of the accidents happen at 30Mph, this is due to the minimum speed limit that has been imposed by the UK Government.

Junction Detail: this shows the details of junction. [9]

Notes: what this shows is that majority of the accidents have happen at a T/Staggered Junction or that they will not be at or within 50 meters of a mini roundabout

Junction Control: this shows the method of control used at a junction. [9]

Notes: What this shows is that the majority of the accidents will happen at junction with the control method where a person will have to give way, by an authorised person or where it will be automatically triggered

Second Road Class: This shows the classification of road. [9]

Notes: what this shows is that most of these accidents happen on an A Road with least amount of them happening on Motorways, A Roads and roads that have been unclassified.

* It should be noted that 2nd Class Road and Number attributes only apply to a road that has a junction.

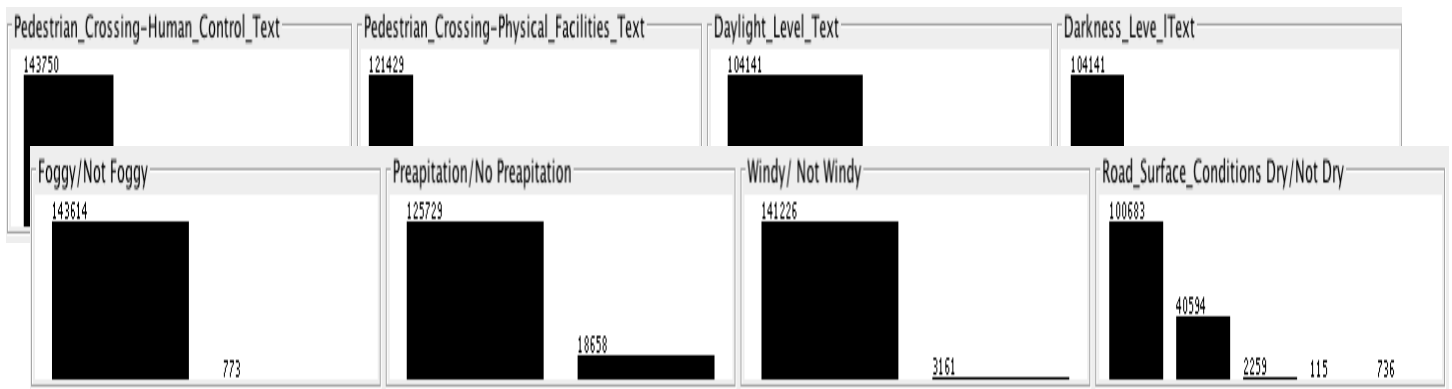


Figure 6 - Depiction of Data

Pedestrian Crossing Human Control: this shows if an accident has taken place within such proximity. [9]

Notes: What this shows is that these accidents have not taken place within 50 meters of a junction

Pedestrian Crossing Physical Facilities: this shows the facilities of the crossing (EG: a crossing may be a zebra, pelican, puffin, toucan or similar non-junction pedestrian light crossing). [9]

Notes: what this shows is that these accidents have taken place with a junction that's been controlled by another person

Daylight Level: this shows the lighting conditions at the scene of an accident, and is used as a variable to see if it contributed to the accident. [9]

Notes: What this shows is that most of these accidents have taken place in daylight

Darkness Level: this shows the darkness conditions at the scene of an accident, and is used as a variable to see if it contributed to the accident.

Notes: What this shows is that most of these accidents have taken place in daylight, however the accidents that happened in the dark and a large amount of these happened when the streetlights were lit

Figure 7 - Depiction of Data

Foggy/Not Foggy: this shows is there was fog or wasn't fog at the scene of an accident and it is used as a variable see if it had an effect on a accident. [9]

Notes: What this shows is that the majority of these accidents happened under foggy conditions

Prepetition/No Prepetition: this shows if there was rain or wasn't rain at the scene of an accident and it is used as a variable see if it had an effect on a accident. [9]

Notes: What this shows is that the majority accidents happened when it was raining

Windy/Not Windy: this shows if there if was windy or not windy at the scene of an accident and it is used as a variable see if it had an effect on a accident. [9]

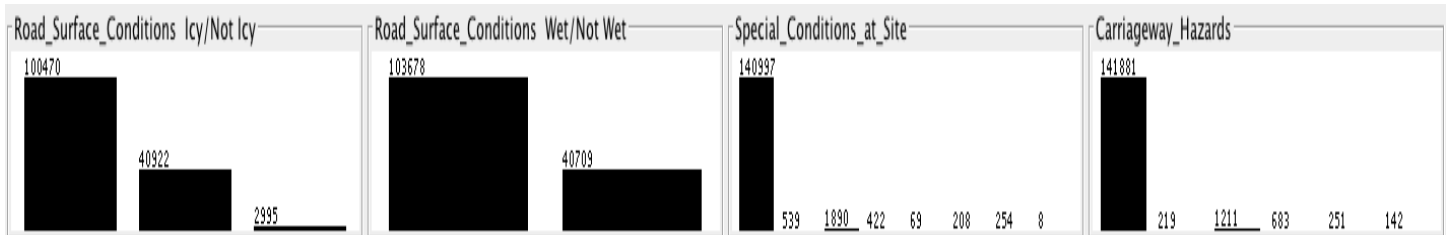
Notes: What this shows is that the majority of accident happened in not windy conditions

Road Surface Condition Dry/Not Dry: this shows if the road surface was dry or not dry at the scene of an accident and it is used as a variable to see if it had an effect on an accident. [9]

Notes: What this shows is the majority of these accident happened when the roads were dry

Figure 8 - Depiction of Data

Road Surface Condition Icy/Not Icy: this shows if the road surface was icy or not icy at the scene of an accident and it is used as a variable to see if it had an effect on an accident. [9]



Notes: What this shows is that the majority of accidents happen when the road conditions are not icy

Road Surface Condition Dry/Not Dry: this shows if the road surface was wet or not wet at the scene of an accident and it is used as a variable to see if it had an effect on an accident. [9]

Notes: what is shown is that the most of these accidents happen when the roads are dry

Special Conditions at Site: this shows if there were any special conditions that had to be made at the site of an accident, used as a variable to see if it had an effect on an accident. [9]

Notes: What this shows is the for most of the accidents that happened in 2010 a majority had no (None) special conditions at the site of an accident while only a semi medium percentage had any conditions at all.

Carriageway Hazards: this shows any hazards at the side of the carriageways. [9]

Notes: What this shows is that there where mostly no (None) hazards in an accident that happened on a carriageway, while only a semi small percentage involved hazards.

Figure 9 - Depiction of Data

Urban or Rural Area: this shows if an accident has happened within an urban or rural area. [9]

Notes: what this shows is that majority of these accidents happen within an urban area, this is due to that that London area is a very built up area



Did Police Officer Attend scene of Accident? - This shows whether or not an officer has attended the scene of an accident. [9]

Notes: What this show is that for the majority of accidents a police officer did attend the site of an accident

6 POST – PROCESSING OF DATA

Nonetheless, it should be noted that this data set has been arranged within the pre-processing stage, meaning that it has been translated to Text Data (this is a normalized [10] attribute rather than using numeric [11] values) meaning the data form an attribute such as Police Force (1 – London Metropolitan Police Force). Demonstrating that some of the attributes can now be collated to one another (i.e.: 'Clustered'). What this involves is the use of location (OSGR Northing/Easting) for the Y

and X-axis, meaning that two variables can be analysed to see if they are a contributing factor to an occurred accident (as can be seen in figure 9).

The first step in this process is the translation of the number values into their text counterparts; an example of this would be the police force number (EG: '1' is equivalent to the London Metropolitan Police Force}. However, to test this the Translated Data Set will cut down to one location to understand if this translated data model works; this has been conducted in-light off some 'trial and error' experimentation, with a cut down un-translated (which still has all of the number values, yet no text values).

As the results, this developed in data set clustering which created a location map of the London road network [12], meaning this formula will be used for any future data mining. It should also be noted that this data set mapped out the large park and green areas of London as well, as these are the same parameters that were used for the first set of experimentation, also this data will be cross correlated with other sources (News Sites) to see if this theory can be validated. (As can be seen by figures 9, 10, 17, 18 & 19)

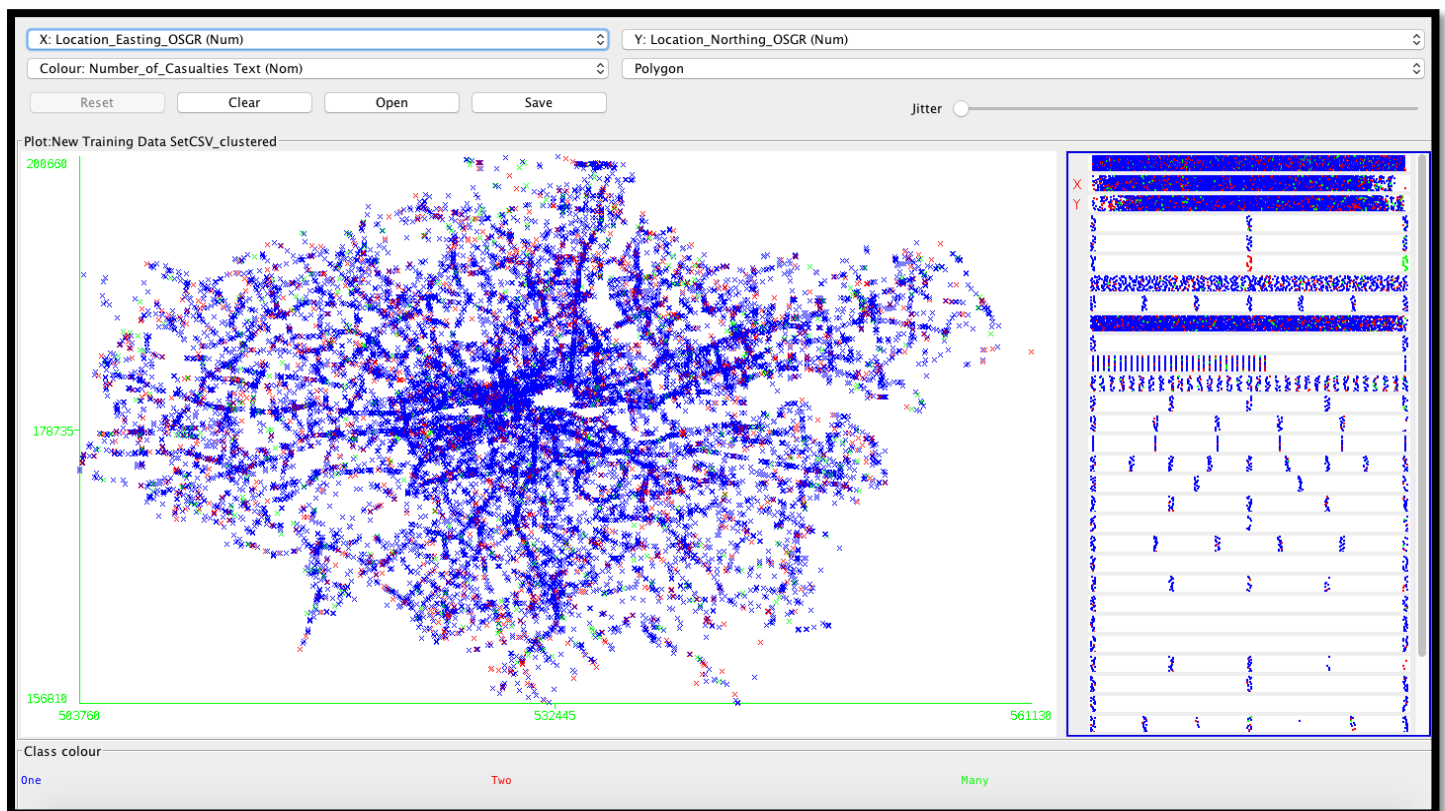


Figure 11 - Test Data Set Cluster 'Mapping London'

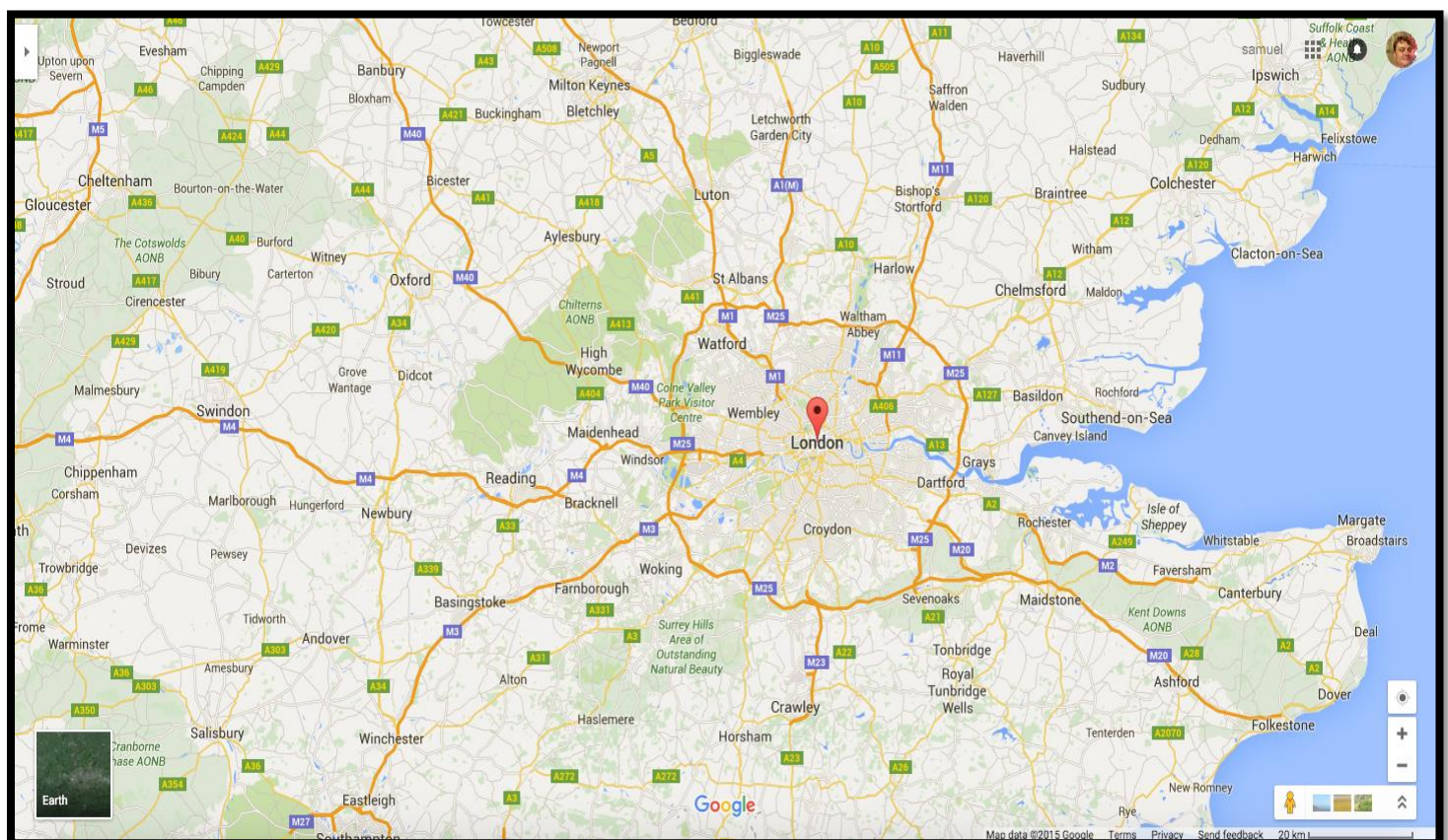


Figure 10 - Google Maps: of London [12]

However these results prompted further research into the ratio of accidents coinciding with the rate of congestion, to see this this was true the data was loaded into an experimental AIP by Google called Fusion Tables [20], which would be compared to a Map featured within an article done by

The City.A.M News [21] it clearly shows that it is possible that the rate of congestion can increases as the rate of accident increases.

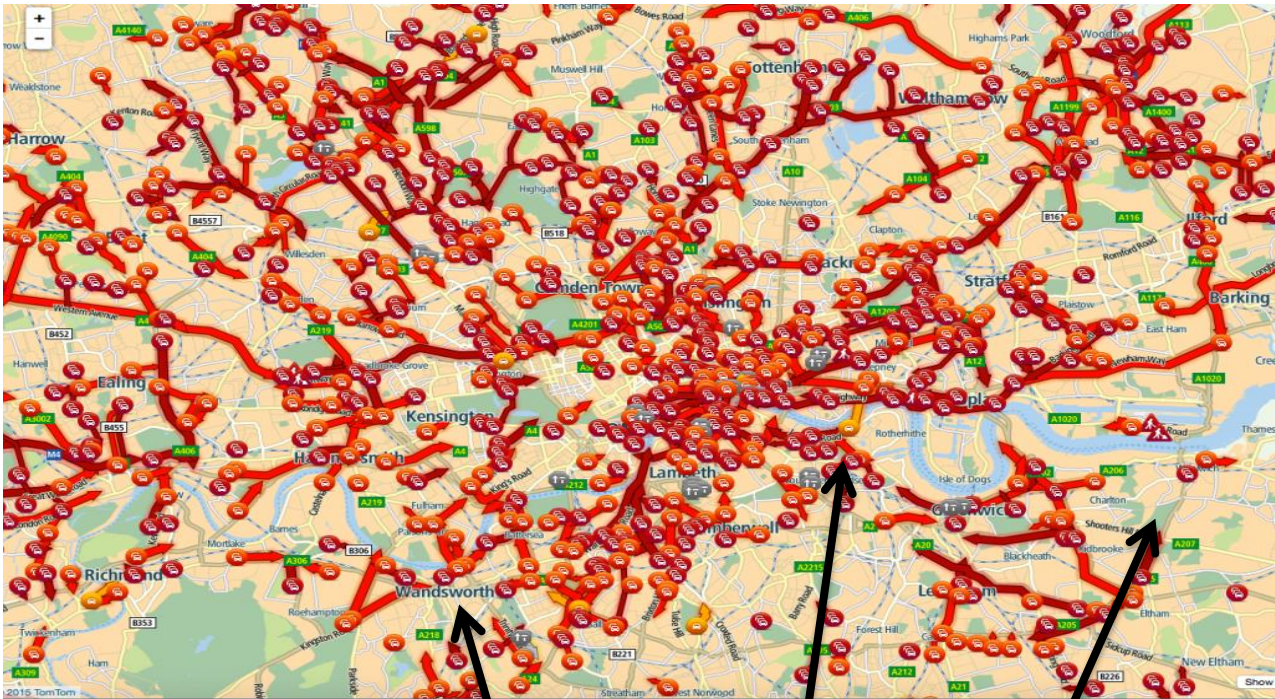
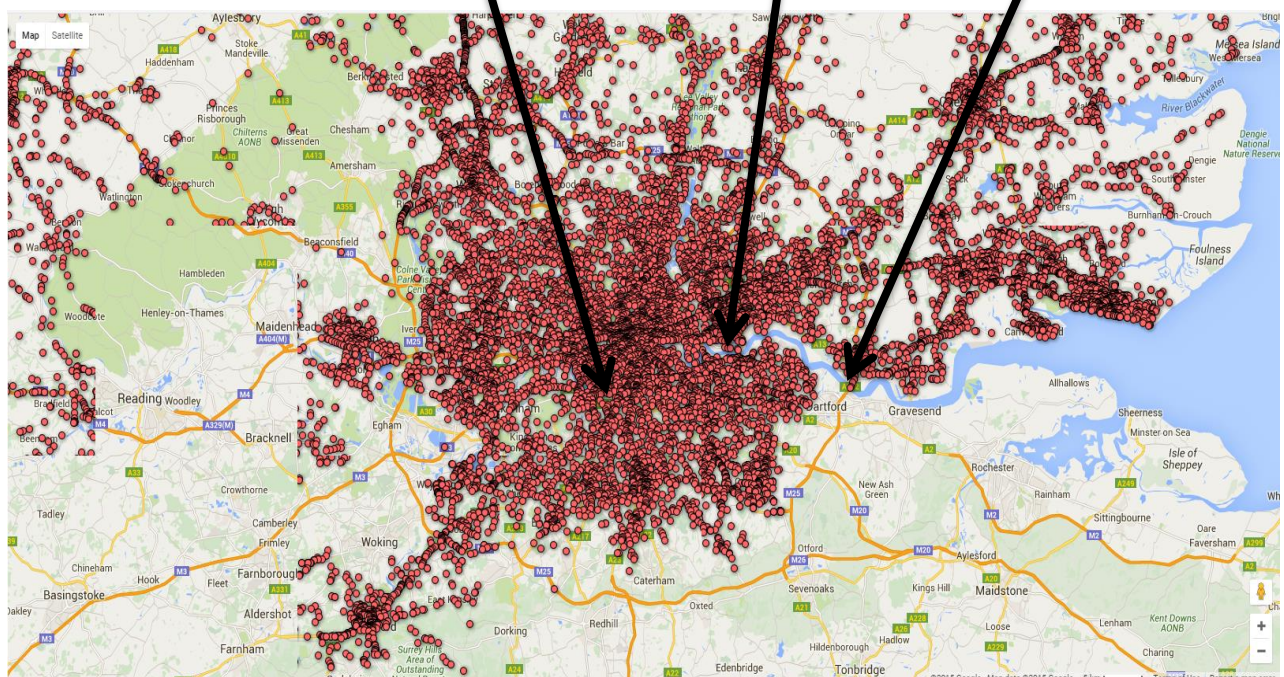


Figure 13 City.A.M Map of Congestion [21]



When these two maps are compared against one another they show that there is a possible correlation between the accident to congestion ratio and what can also be seen is that when these accidents are plotted, predictions can be made, however what is more important is that fact that details can be gained using an API such as this, as is shown in: Figures 19, what this means is that when these kinds of tools are used and then plotted and compared to real world events/actions, more details can be gained from them rather than just using the normal methods such as keeping the data contained within a spreadsheet.



Figure 14 - Google Fusion Table - Centered on Hyde Park with Accident Details

7 ANALYSES TABLES

It should be noted that for this section as the 3 main variables are: accident severity, number of vehicles and number of casualties; these are the results of a classification analyses.

JRip – is a class that implements a propositional rule learner, ‘Repeated Incremental Pruning to Produce Error Reduction’ (RIPPER), which was proposed by William W. Cohen [16].

J48 Tree – is a Class for generating a pruned or unpruned C4.5 decision tree [17].

RBFN Network – is a Class implementing radial basis function networks, trained in a fully supervised manner using WEKA's Optimization class by minimizing squared error with the BFGS method [18].

This section contains the results of the classification analyses test of the 3 key values evaluated (using a very simplistic ‘broad spectrum analysis’ method) [19] and where a decision will be made on what algorithm yielded the best results, then the algorithm in question will then be applied to the whole of the test data set to build a model

8 JRIP – TESTING

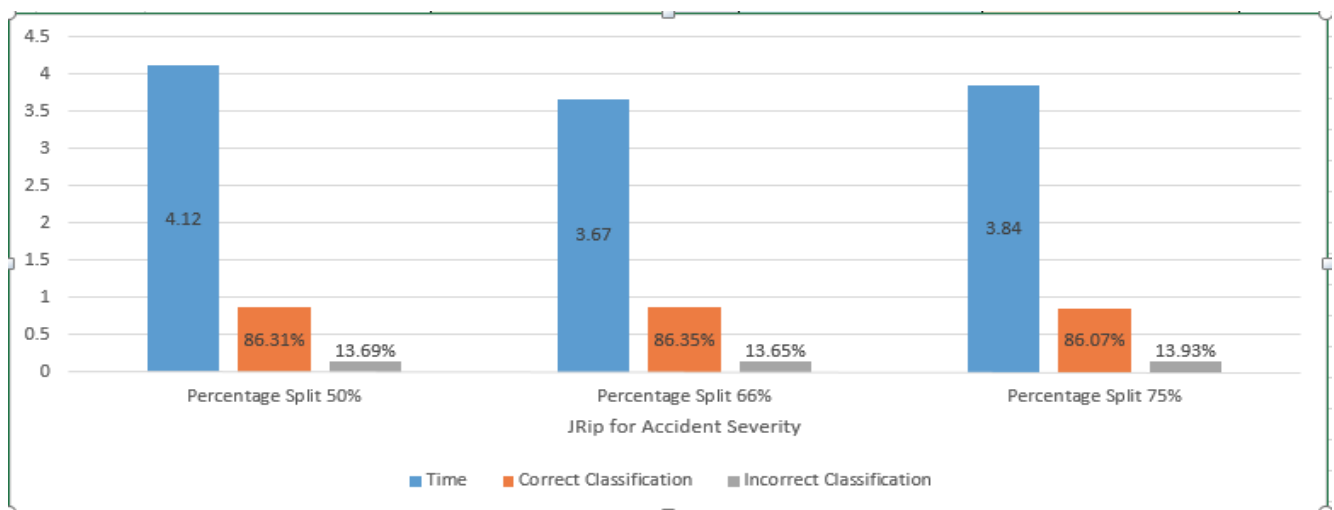
8.1 ANALYSES TEST FOR ACCIDENT SEVERITY

8.1.1 JRip Testing

Table 1 - Accident Severity - JRip Results

Name	Time	Correct Classification	Incorrect Classification	Number of Rules	Percentage Split
JRip	4.12	86.312 %	13.688 %	10	50%
JRip	3.67	86.3478 %	13.6522 %	10	66%
JRip	3.84	86.0739 %	13.9261 %	10	75%

It was found that when using JRip that the best overall results came from using the 66%



JRip for Accident Severity

Percentage Split's	Time	Correct Classification	Incorrect Classification
Percentage Split 50%	4.12	86.31%	13.69%
Percentage Split 66%	3.67	86.35%	13.65%
Percentage Split 75%	3.84	86.07%	13.93%

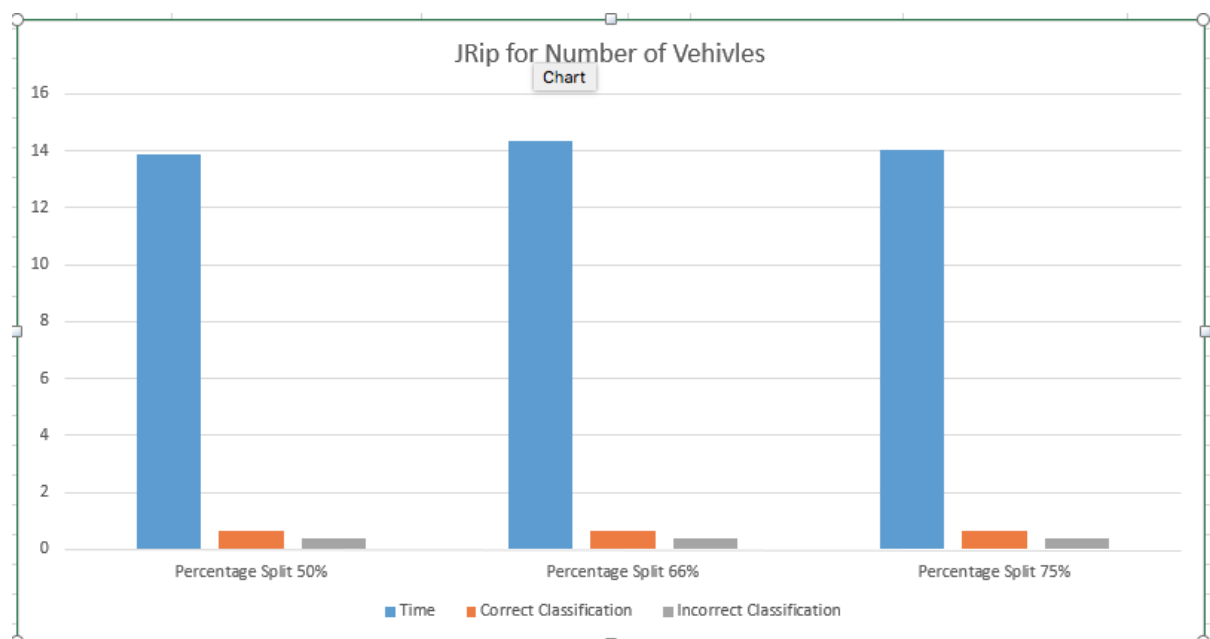
8.2 ANALYSES TEST FOR NUMBER OF VEHICLES

8.2.1 JRip Testing

Table 2 - Number of Vehicles - JRip Results

Name	Time	Correct Classification	Incorrect Classification	Number of Rules	Percentage Split
JRip	13.89	63.9781 %	36.0219 %	11	50%
JRip	14.33	64.0668 %	35.9332 %	11	66%
JRip	14.05	63.0925 %	36.9075 %	11	75%

It was found that when using JRip that the best overall results came from using the 66%



JRip for Number of Vehicles			
Percentage Split's	Time	Correct Classification	Incorrect Classification
Percentage Split 50%	13.89	63.98%	36.02%
Percentage Split 66%	14.33	64.07%	35.93%
Percentage Split 75%	14.05	63.09%	36.91%

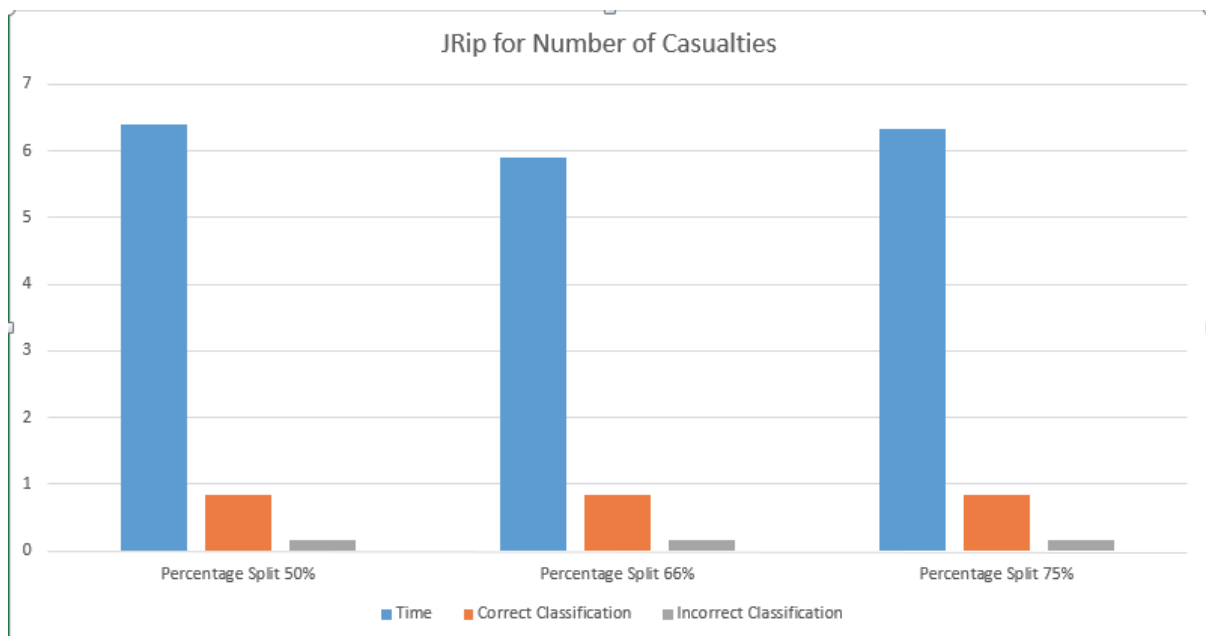
8.3 ANALYSES TEST FOR NUMBER OF CASUALTIES

8.3.1 JRip Testing

Table 3 - Number of Casualties - JRip Results

Name	Time	Correct Classification	Incorrect Classification	Number of Rules	Percentage Split
JRip	6.39	84.7593 %	15.2407 %	2	50%
JRip	5.91	84.7134 %	15.2866 %	2	66%
JRip	6.34	84.472 %	15.528 %	2	75%

It was found that when using JRip that the best overall results came from using the 66%



JRip for Number of Casualties			
Percentage Split's	Time	Correct Classification	Incorrect Classification
Percentage Split 50%	6.39	84.76%	15.24%
Percentage Split 66%	5.91	84.71%	15.29%
Percentage Split 75%	6.34	84.47%	15.53%

9 JRIP RESULTS

With JRip what can be seen is that the overall best results came from using the 66% threshold; any higher than this then the overall results would begin to drop off. What this means is that with this algorithm it has an overall high mean. This can indicate that this maybe a viable algorithm which could be applied to the London Test Data Set, and if successful it could be applied to the 1,000,000 record data set. However, when testing this algorithm it had an overall complexity of 7 and a range of 2 – 11; meaning that this would be a very complex model and pass hinderers may play a part once more.

10J48 TREE - TESTING

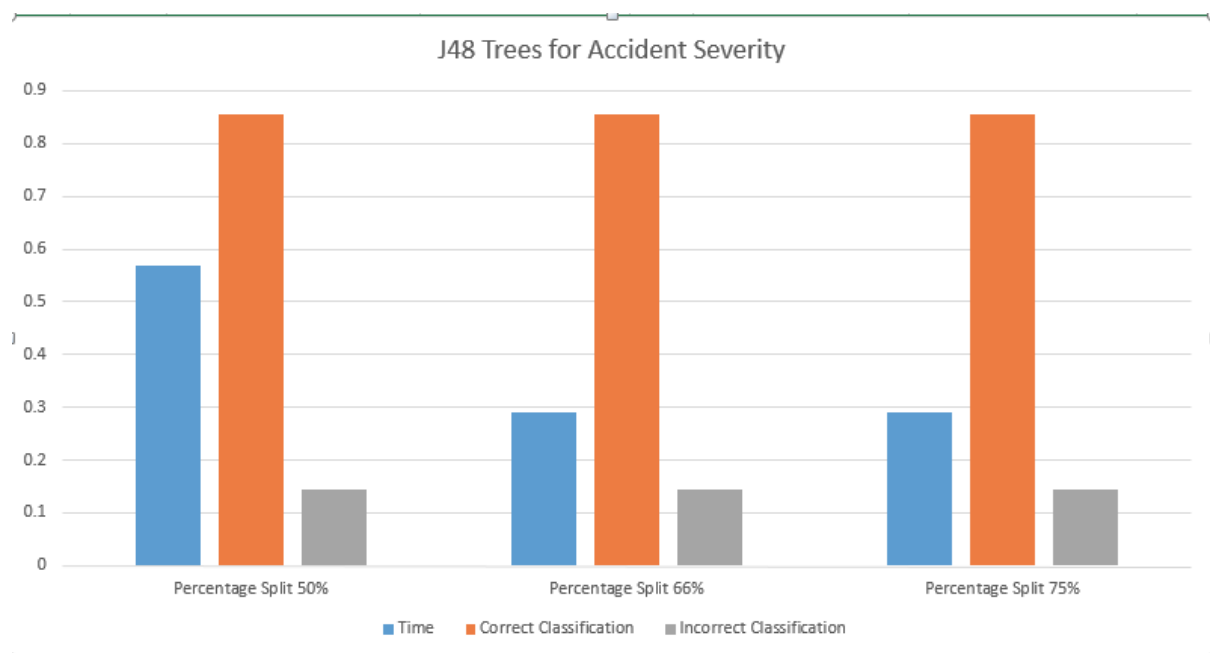
10.1 ANALYSES TEST FOR ACCIDENT SEVERITY

10.1.1 J48 Tree Testing

Table 4 - Accident Severity - J48 Tree Results

Name	Time	Correct Classification	Incorrect Classification	Number of Leaves/Size of Tree	Percentage Split
J48	0.57	85.4621 %	14.5379 %	1/1	50%
J48	0.29	85.5186 %	14.4814 %	1/1	66%
J48	0.29	85.5345 %	14.4655 %	1/1	75%

It was found that when using J48 that the best overall results came from using the 75%



J48 Tree for Accident Severity			
Percentage Split's	Time	Correct Classification	Incorrect Classification
Percentage Split 50%	0.57	85.46%	14.54%
Percentage Split 66%	0.29	85.52%	14.48%
Percentage Split 75%	0.29	85.53%	14.47%

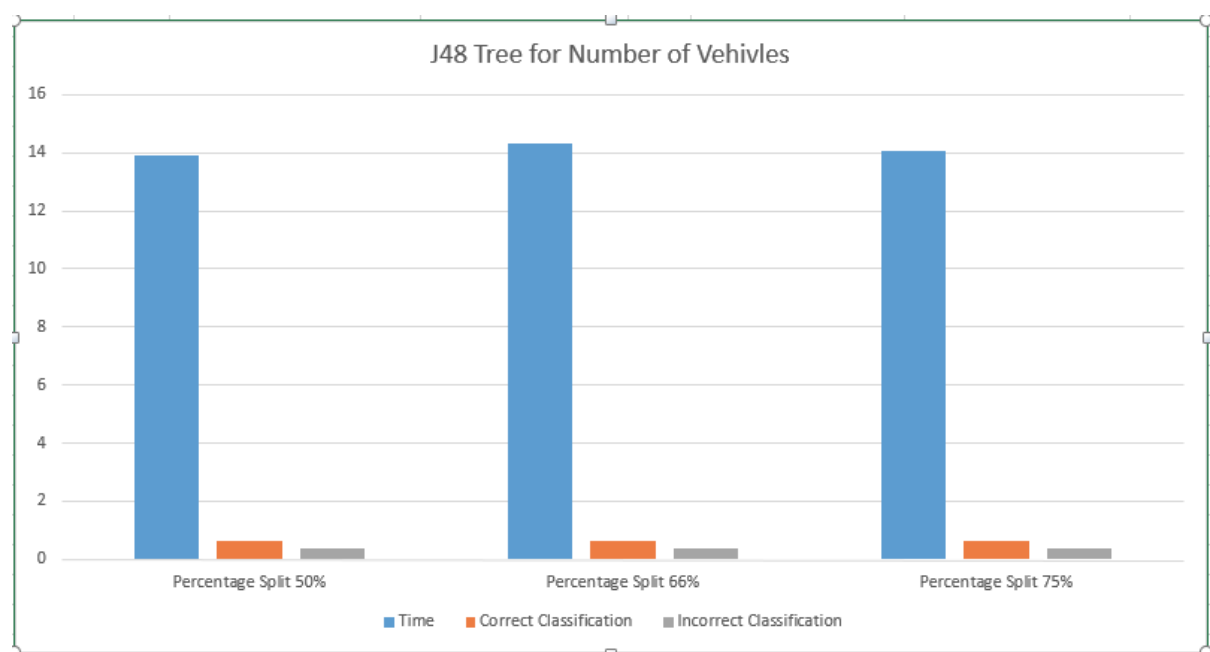
10.2 ANALYSES TEST FOR NUMBER OF VEHICLES

10.2.1 J48 Tree Testing

Table 5 - Number of Vehicles - J48 Tree Results

Name	Time	Correct Classification	Incorrect Classification	Number of Leaves/Size of Tree	Percentage Split
J48	2.39	62.2947 %	37.7053 %	25/34	50%
J48	2.33	62.1079 %	37.8921 %	25/34	66%
J48	2.3	62.8473 %	37.1527 %	25/34	75%

It was found that when using J48 that the best overall results came from using the 75%



J48 Tree for Number of Vehicles			
Percentage Split's	Time	Correct Classification	Incorrect Classification
Percentage Split 50%	13.89	63.98%	36.02%
Percentage Split 66%	14.33	64.07%	35.93%
Percentage Split 75%	14.05	63.09%	36.91%

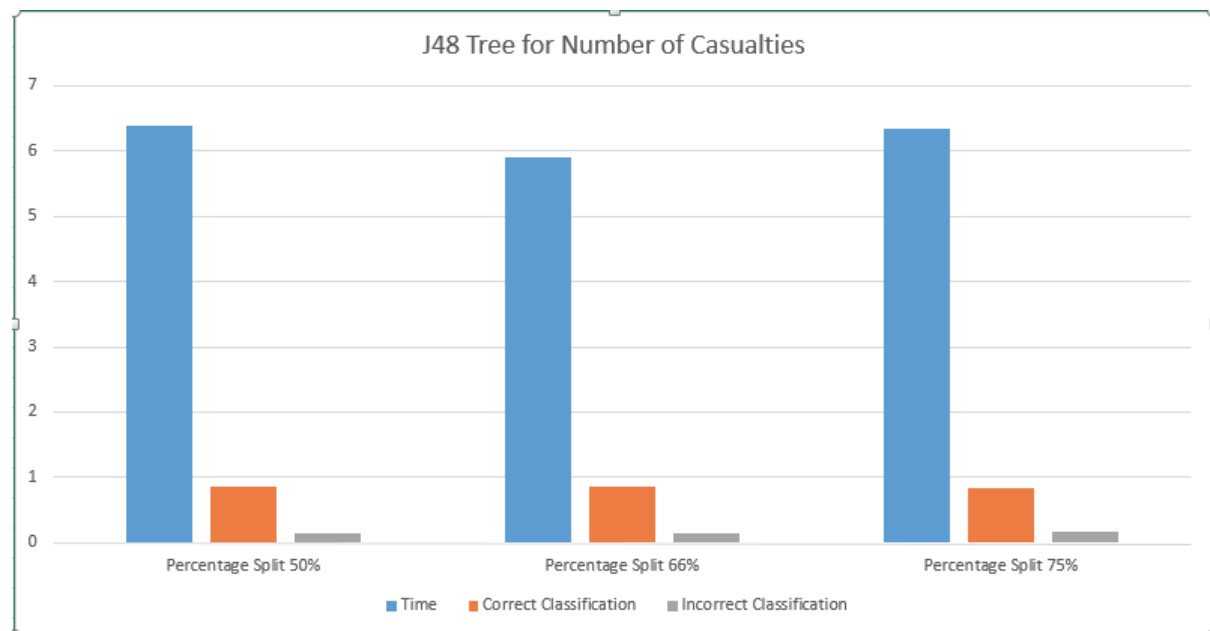
10.3 ANALYSES TEST FOR NUMBER OF CASUALTIES

10.3.1 J48 Tree Testing

Table 6 - Number of Casualties - J48 Tree

Name	Time	Correct Classification	Incorrect Classification	Number of Leaves/Size of Tree	Percentage Split
J48	1.42	84.7593 %	15.2407 %	1/1	50%
J48	1.44	84.7134 %	15.2866 %	1/1	66%
J48	1.44	84.472 %	15.528 %	1/1	75%

It was found that when using J48 that the best overall results came from using the 75%



J48 Tree for Number of Casualties			
Percentage Split's	Time	Correct Classification	Incorrect Classification
Percentage Split 50%	6.39	84.76%	15.24%
Percentage Split 66%	5.91	84.71%	15.29%
Percentage Split 75%	6.34	84.47%	15.53%

11J48 TREE RESULTS

What can be seen with the J48 Tree results is that it has a number of differences when compared to the JRip results: firstly it has a much lower mean build time of 1.386 seconds to 8.06 for JRip, however, when comparing the correct classification rate of 77.52 – 78.20 that means there is a 0.68 variation. Although this should not to prove much of a problem as there has to be a balance between performance and accuracy and as for the complexity issues this would not prove to be a problem as the number of leaves and the size of the tree is only 1.

12RBFN NETWORK - TESTING

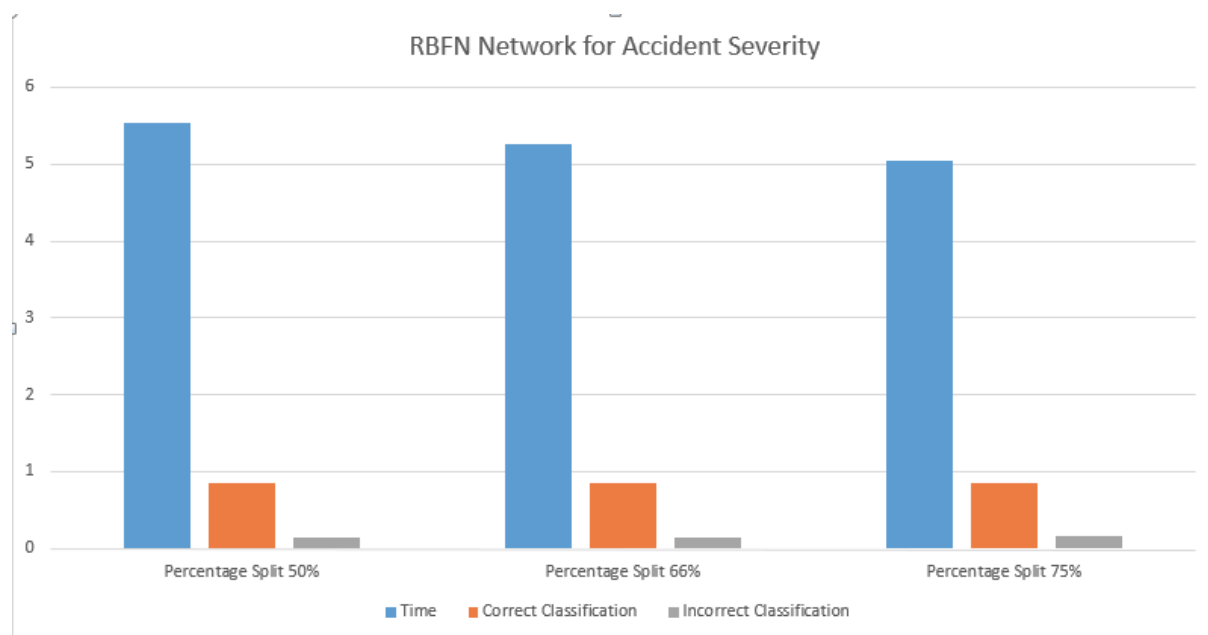
12.1 ANALYSES TEST FOR ACCIDENT SEVERITY

12.1.1 RBFN Network Testing

Table 7 - Accident Severity - RBFN Network Results

Name	Time	Correct Classification	Incorrect Classification	Percentage Split
RBFN Network	5.54	85.1352 %	14.8648 %	50%
RBFN Network	5.27	85.4705 %	14.5295 %	66%
RBFN Network	5.04	84.6355 %	15.3645 %	75%

It was found that when using RBFN Network that the best overall results came from using the 66%



RBFN Network for Accident Severity			
Percentage Split's	Time	Correct Classification	Incorrect Classification
Percentage Split 50%	5.54	85.14%	14.86%
Percentage Split 66%	5.27	85.47%	14.53%
Percentage Split 75%	5.04	84.64%	15.36%

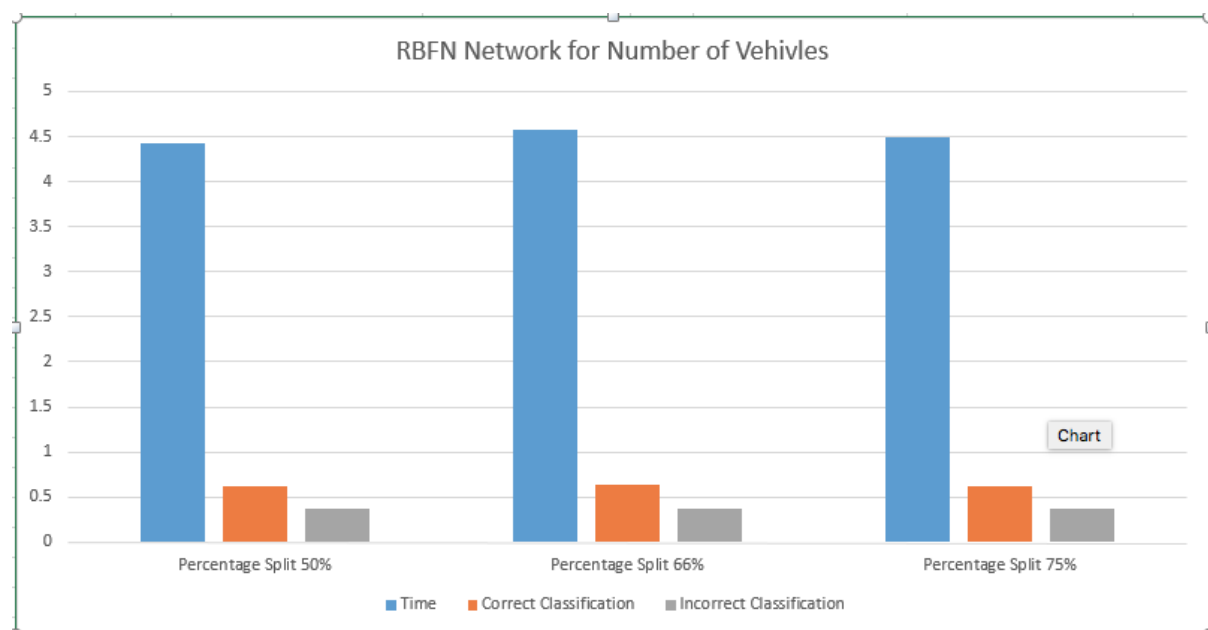
12.2 ANALYSES TEST FOR NUMBER OF VEHICLES

12.2.1 RBFN Network Testing

Table 8 - Number of Vehicles - RBFN Network Results

Name	Time	Correct Classification	Incorrect Classification	Percentage Split
RBFN Network	4.42	62.4826 %	37.5174 %	50%
RBFN Network	4.58	63.2736 %	36.7264 %	66%
RBFN Network	4.5	62.6675 %	37.3325 %	75%

It was found that when using RBFN Network that the best overall results came from using the 66%



RBFN Network for Number of Vehicles			
Percentage Split's	Time	Correct Classification	Incorrect Classification
Percentage Split 50%	4.42	62.48%	37.52%
Percentage Split 66%	4.58	63.27%	36.73%
Percentage Split 75%	4.5	62.67%	37.33%

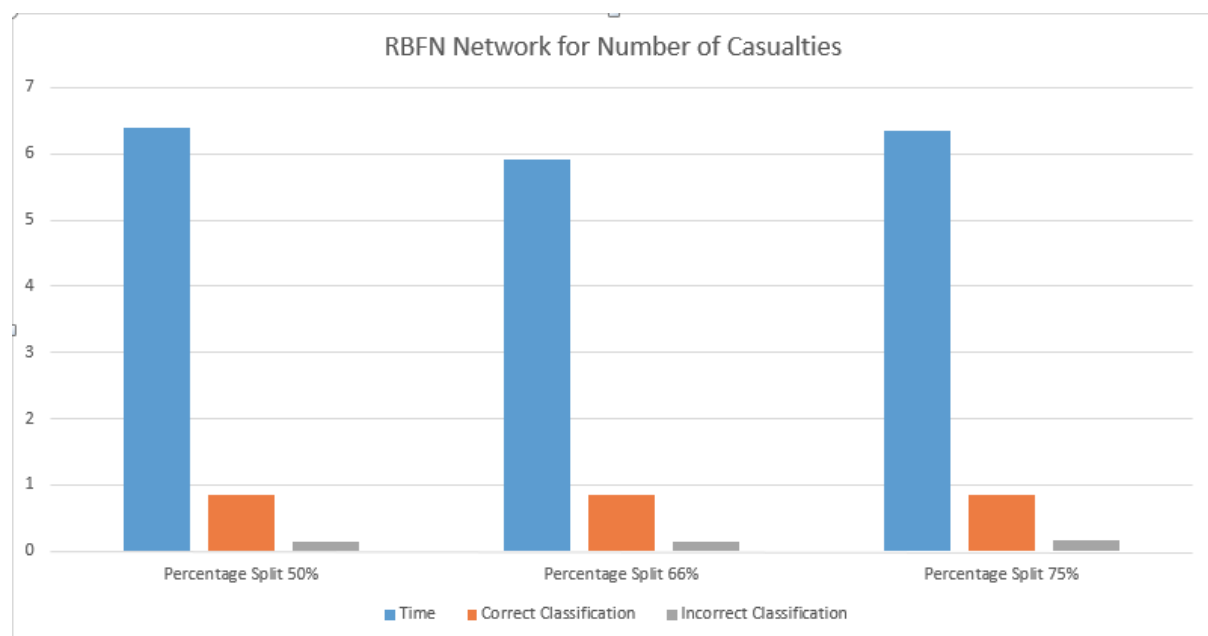
12.3 ANALYSES TEST FOR NUMBER OF CASUALTIES

12.3.1 RBFN Network Testing

Table 9 - Number of Casualties - RBFN Network Results

Name	Time	Correct Classification	Incorrect Classification	Percentage Split
RBFN Network	4.86	84.694 %	15.306 %	50%
RBFN Network	5.26	84.485 %	15.515 %	66%
RBFN Network	4.71	84.4022 %	15.515 %	75%

It was found that when using RBFN Network that the best overall results came from using the 66%



RBFN Network for Number of Casualties			
Percentage Split's	Time	Correct Classification	Incorrect Classification
Percentage Split 50%	6.39	84.76%	15.24%
Percentage Split 66%	5.91	84.71%	15.29%
Percentage Split 75%	6.34	84.47%	15.53%

13RBFN NETWORK RESULTS

With this algorithm it seems to have a varying differing sequence as it seems to lay within the middle range of the readings from the tests. It has a build time of 4.909 seconds whereas JRip has 1.386 seconds and J48 Tree has 8.06, and this middle ground effect can also be observed with the other measures as well. Meaning that this algorithm maybe able to service as an alternative (a Plan B) to the algorithm that is to be picked. If the picked algorithm is not fit for propose.

14 RECOMMENDATIONS

The results from this testing is that J48 Tree using a 66% Split will be used as it has produced the best mean results, in terms of build time, Correct Classification, Incorrect Classification and Number of leaves and size of tree. To sum this up, it has produced the best results and this algorithm will be applied to the main 'Test Data Set' (using 2010 data rather than 2009) to build a read accident model that maybe at a later data could be applied to the entry 'Translated Data Set'. However, in the event that this proves wasted effort then the RBFN Network algorithm will be used as it produced the nest best set of results and it is a very viable alternative.

15 RESULTS OF MODELLING

Within this section will outline the results of the modelling using the 'Training Data Set' and the 'Testing Data Set' when the results of the modelled values and the un-modelled values are compared to one another what can be seen is a real dispersion between the results

16 RESEARCH NOTE(S)

These data set results form the results within section 7 – 13 where not up to spec, meaning that a new algorithm had to be applied through, trial and error it was found that the Random Tree was best suited for this data set.

17 FURTHER MODELLING RESEARCH

It should be noted that these results have been found to these tests of the data set would of proved to be fruitful, meaning that it has been decided that some attributes will need to removed when, and, the first test involved the removal of attributes, then processing it through Weka and using training set to produced a model and then applying it to the data set to produce a best results, and, this set of actions will be applied to the three test variables, what this means is that the location variables will be removed and then the variables that are not related the to test e.g. when testing accident severity the northing, easting, number of vehicles and number of casualties will be removed. Also to improve on the results form within sections 7 – 13 it was decided that the use of RandomTree would be best as 'Tree' algorithm has the best results so far.

18 ACCIDENT SEVERITY TEST

What can be seen when this variable was tested is that the rate of Correctly Classified Instances 144361 = 99.982 % and Incorrectly Classified Instances 26 = 0.018 % and the Confusion Matrix was:

18.1.1 Accident Severity Confusion Matrix

18.1.2 Confusion Matrix			
A	B	C	Classified as
124112	0	0	A = High
12	18265	0	B = Medium
3	0	1984	C = Low

Also when these results are analysis along with the detailed accuracy by class is that True Positive (TP) and False Positive (FP) Rate is for the three values is that:

1. for A = High TP Rate is 1.000 and FP Rate is 0.001
2. for B = Medium TP Rate is 0.999 and FP Rate is 0.000
3. for C = Low TP Rate is 0.998 and FP Rate is 0.000
 - The Kappa statistic is 0.9993
 - The Mean absolute error is 0.0001
 - Root mean squared error is 0.0077
 - Coverage of cases (0.95 level) is 100

Notes: for this test the location attributes were removed along with number of vehicles and number of casualties

19NUMBER OF VEHICLES TEST

What can be seen when this variable was tested is that the rate of Correctly Classified Instances 144321 = 99.9543 % and Incorrectly Classified Instances 66 = 0.0457 % and the Confusion Matrix was:

19.1.1 Number of Vehicles Confusion Matrix

19.1.2 Confusion Matrix			
A	B	C	Classified as
41758	0	0	A = One
34	86863	1	B = Two
6	25	15700	C = Many

Also when these results are analysis along with the detailed accuracy by class is that True Positive (TP) and False Positive (FP) Rate is for the three values is that:

1. for A = One TP Rate is 1.00 and FP Rate is 0.000
2. for B = Two TP Rate is 1.000 and FP Rate is 0.000
3. for C = Many TP Rate 0.998 is and FP Rate is 0.000
 - The Kappa statistic is 0.9992
 - The Mean absolute error is 0.0003
 - The Root mean squared error is 0.0124
 - The Coverage of cases (0.95 level) is 100

Notes: for this test the location attributes were removed along with accident severity and number of casualties.

20NUMBER OF CASUALTIES TEST

What can be seen when this variable was tested is that the rate of Correctly Classified Instances 110756 = 76.7077 % and Incorrectly Classified Instances 33631 = 23.2923 % and the Confusion Matrix was:

20.1.1 Number of Vehicles Casualties Matrix

20.1.2 Confusion Matrix			
A	B	C	Classified as
109493	0	0	A = One
20	24134	0	B = Two
13	8	10719	C = Many

Also when these results are analysis along with the detailed accuracy by class is that True Positive (TP) and False Positive (FP) Rate is for the three values is that:

1. for A = One TP Rate is 1.000 and FP Rate is 0.001
 2. for B = Two TP Rate 0.999 and FP Rate is 0.000
 3. for C = Many TP Rate 0.998 is and FP Rate is 0.000
- The Kappa statistic is 0.9993
 - The Mean absolute error is 0.0002
 - The Root mean squared error is 0.0098
 - The Coverage of cases (0.95 level) is 100

Notes: for this test the location attributes were removed along with accident severity and number of vehicles.

21 RESULTES OF CLASSIFICATION (s)

What can be seen within these results is that the overall range of Correctly and Incorrectly Classified Instances is within the same ranges as one another, this is most likely due to the fact that when using other algorithm such as the ones within the 'Results of Training Data Set' section of this report they had a varying degree of success, however with the RandomTree algorithm is did not prove to be the case, and, what this means is that these values have been mapped out with a rate of error that's 0.000 – 0.001 this means is that if the incorrect values where to be removed this means that the rate of success would be 100%.

22 POSSIBLE FUTURE RESEARCH

What this means is that the business question 1 has been answered with the clustering analyses on pages 10 – 14 and as for the rest these where the wrong questions to ask, although this data can still be used to measure:

1. How severity and accident could be
2. How many vehicles would be involved
3. How casualties can be expected to be involved
 - a. However with should be noted that with these three values, alone could be used to created an accident index in which to see if there would be a correlation between them
 - i. Also Speed would of played a major valuable in these accident the this should have been explored as a viable measure form the beginning

23 REFERENCES

- [1] UK Government, Department of Transport, Road Accident and Safety Tables Index, 26th September 2013 Available at <https://www.gov.uk/government/statistical-data-sets/road-accidents-and-safety-statistical-tables-index>
- [2] Business Dictionary, Microsoft Excel Available at <http://www.businessdictionary.com/definition/Microsoft-Excel.html>
- [3] The Modelling Agency, SPSS, CRIPS – DM 1.0 Step by step data minging guild by Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler) Available at <https://the-modeling-agency.com/crisp-dm.pdf>
- [4] ETL Process, Data Warehouse Available at <http://datawarehouse4u.info/ETL-process.html>
- [5] PC.net, Glossary: Access, Available at <http://pc.net/glossary/definition/access>
- [6] Search Windows Server, Master File Table Definition, by Margaret Rouse Available at <http://searchwindowsserver.techtarget.com/definition/master-file-table>
- [7] CS Waikato, Attribute-Relation File Format (ARFF), April 1st 2002 Available at <http://www.cs.waikato.ac.nz/ml/weka/arff.html>
- [8] Weka, The University of Waikato, Weka 3: Data Mining Software in Java Available at <http://www.cs.waikato.ac.nz/ml/weka/>
- [9] Department for Transport, STATS 20, Instructions for the Completion of Road Accident Reports from non-Crash Sources, September 2011
- [10] Source Forge, Weka, Class Normalize Available at <http://weka.sourceforge.net/doc.dev/weka/filters/unsupervised/attribute/Normalize.html>
- [11] Source Forge, Weka, Class Attribute Available at <http://weka.sourceforge.net/doc.dev/weka/core/Attribute.html>
- [12] Google Map of the City of London Available at <https://www.google.co.uk/maps/@51.528308,-0.3817765,10z>
- [13] SDSC UC San Diego, PAGE Predictive Analytics Centre of Excellence, Disadvantages of Weka Available at http://pace.sdsc.edu/sites/pace.sdsc.edu/bootcamp/201402/media/PACE_Bootcamp_TS2_WEKA_Intro.pdf
- [14] Data Mining Techniques for Data Cleaning, Proceedings of the 4th World Congress on Engineering Asset Management, Athens, Greece 28 – 30 September 2009, by Kalaviany Natarajan, Jiuyong Li and Andy Koronios

- [15] Microsoft, Developer Network, Training and Testing Data Sets Available at <https://msdn.microsoft.com/en-us/library/bb895173.aspx>
- [16] Source Forge, Weka, Class JRip Available at <http://weka.sourceforge.net/doc.dev/weka/classifiers/rules/JRip.html>
- [17] Source Forge, Weka, Class J48 Available at <http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/J48.html>
- [18] Source Forge, Weka, Class RBFC Classifier Available at <http://weka.sourceforge.net/doc.packages/RBFNetwork/weka/classifiers/functions/RBFClassifier.html>
- [19] Kybernetes, Information & Knowledge Management, Modelling & Analysis of Uncertain Systems (3), By Mian-yun Chen, Yi Lin, Hejing Xiong, 1972, ISSN: 0368-492X
- [20] Google Support Page, About Fusion Tables Available at <https://support.google.com/fusiontables/answer/2571232>
- [21] City.A.M TfL Tube strike 2015: This map shows the worst roads in London for congestion and traffic jams, Emma Haslett, 9th July 2015, Map of 'All 1,450-odd congestion hotspots (Source: TomTom)'
- [22] IBM Knowledge Center, SPSS Modeler 15.0.0, CRISP-DM Help, Available at http://www-01.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.crispdm.help/crisp_overview.htm?lang=en

24 APPENDIX

24.1 TECHNOLOGY ISSUES

One technological issue found was with the Translated Data Set (Full Data Set with 1,000,000 Translated Records) as they were over 300MB. Therefore a large amount of time would pass before it could load into WEKA; this was a result of hardware deficiencies (i.e.: computers not being to a specific specification this is normally known as Memory Bound) [13].

The only way to address this problem is to cut down the data set again by either 20% or 50%, or reverting back to the original parameters of using one location and one year's worth of data. As a result of the use of the original parameters, the location used is London and the year is 2010. The cut-down version [14] of the translated data [14] set is a little over 50MB and, as a result, will not give a full picture of the accident rate within the UK (It should be noted that this is due to the hardware and software limitations) [13]. However, this is still useful as a test case to see if this style of model has any use, although to test to see if these original parameters have any bearing on the translated data set, a training data set [15] will be created from the translated data set using 2009, data in place of 2010, and the results of this testing are below.

Another Technology issues that had an effect on this project was a problem with the building of the model (getting the Training and Testing data to be compatible), and, these issues where:

1. Headers of the Arff files being the same
 - a. An example of this was with this the training arff file the title header was 'cleaned translated data set' when it should have been 'training data set'
2. Another issues was that the attributes headers needed to be exactly the same in both files.
3. Another issues that also relates to the attributes headers was that the format that they used needed to be changed, as in the 'Date' attribute was set to as a 'set' (meaning all of the values where in an order) rather than it being set as a string value,
4. This same issues also applied to the time attribute and it also underwent the same process

5. Also the issues for these arff files was that some of the data was either misspelt or wrong – meaning that some of the misspelt values where changes and for the wrong vales they where either changed or deleted
 - a. This was done through a process of opening both files in 'Weka' as it would identify the problem
 - b. Then once the problem was found the files where opened in a text editor (notepad plus) then the problem values where corrected
 - c. This process was then repeated until the problem was solved

However it should be noted that to understand this problem extensive research was required:

25 DATA SET DISCRETION

Data Set for this Assignment the data will be sorted by the year and location, in that the year this will be sorted by in 2010 as this is the most up to date data and as for the location it will have to be London as data on more relevant locations such as South Wales (Swansea, Neath) are not available within this data set for 2010.

Data Set Analyses with this data set as it is government data it is a very complete form most data sets and wont need much cleaning (processing) and it is in a CSV file format so that means most if not all of the work is complete all that is needed is to select the records that are needed and convert the file into an arff file, that will make it work with weka and tableua is means that that normal stages within the CRIPS-DM framework maybe bypassed to save time in the processing stages, this means that the data mining may happen ahead of time leading to an answer to the business questions earlier than expected ,

Please note that all of the following discretions came from the STATS 20 Instruction for Completion of Road Accident Reports from non-CRASH Sources manual issued by the Department for Transport.

Accident Index - this is the index number that is used to record the incident that has taken place

Location Easting OSGR – this refers to the Ordnance Survey Grid Reference using the eastern heading this is used to interact a section of an area where an incident that has taken place

Location Northing OSGR - this refers to the Ordnance Survey Grid Reference using the northern heading that is used to interact a section of an area where an incident that has taken place

Longitude - is a geographic coordinate that specifies the east-west position of a point on the Earth's surface

Latitude - the angular distance of a place north or south of the earth's equator, or of the equator of a celestial object, usually expressed in degrees and minutes.

Police Force – this indicates that which police force was in attendance at the scene of an incident for example of this 01 – Metropolitan Police, 03 – Cumbria and 06 – Greater Manchester and 62 – South Wales

Accident severity – indicates that the severity of accident for example a person who moves quickly to avoid being involved in an accident, is successful in that, but in doing so incurs an injury (eg. twists an ankle). Also includes occupant of vehicle, which manoeuvres or brakes suddenly to avoid an impact, but in so doing sustains an injury, a pedestrian who injures himself on a parked vehicle

Number of vehicles – this indicates the number of vehicles that have been involved with the accident

Number of casualties - this indicates the number of casualties involved with the accident i.e number of deaths

Date – this shows the date that the accident has taken place it follows the format of Day/Month/Year

Day of the week – this shows what day of the week an accident has taken place and this is shown as a number for example a Monday will be shown as a number 1 and a Sunday will be shown as a number 7

Time - this shows at what time of day an accident would of taken place this uses the 24 hour system

Local authority (district) – this shows in what local authority an accident has taken place and this is shown using the Code Number that is associated with a location for example the greater London area uses E09000001 - E09000033, EHEATHROW

Local authority (highway) – this shows in which local authority is responsible for the handling of any accident that may happen in it's jurisdiction

1st road class – this defines the type of road and they are: 1- Motorway, 2 - A(M), 3 – A, 4 – B, 5 – C and 6 - Unclassified

1st road number – this shows the number that is given to any one road for example A46, also if the road number contains less than four digits, leading zeros should be inserted to the left of the number (eg. A46 is coded 0046).

Road type – this defines the road type for when an accident might occur such as: 1- Roundabout, 2 - One way street, 3 - Dual carriageway, 6 - Single carriageway, 7 - Slip Road and 9 - Unknown

Speed limit - this applies to the given speed limit that's applicable to the road where an accident has occurred

Junction detail – this will define any given details about a junction and these details are:

- 00. Not at or within 20 metres of junction
- 01. Roundabout
- 02. Mini roundabout
- 03. T or staggered junction
- 05. Slip road
- 06. Crossroads
- 07. Junction – more than 4 arms (not a roundabout)
- 08. Using private drive or entrance
- 09. Other junction

Junction control – this shows the method that controls the operation of a junction and they are:

- 1. Authorised person
- 2. Automatic traffic signal
- 3. Stop sign
- 4. Give way or uncontrolled

2nd road class – this is only applicable to junctions but it still follows the same format for the 1st class of roads that are 1- Motorway, 2 - A(M), 3 - A, 4 - B, 5 - C and 6 - Unclassified

2nd road number - this shows the number that is given to any one road for example A46 and this example is applicable to both classes of roads, also if the road number contains less than four digits, leading zeros should be inserted to the left of the number (eg. A46 is coded 0046).

Pedestrian crossing human control – this section is to show if that in the event of an accident and they follow this format:

- 0.None within 50 metres
- 1.Control by school crossing patrol
- 2.Control by other authorised person

Pedestrian crossing physical facilities – this decides the physical facilities of a pedestrian crossing and they are:

- 0.No physical crossing facility within 50 metres
- 1.Zebra crossing
- 4.Pelican, puffin, toucan or similar non-junction pedestrian light crossing
- 5.Pedestrian phase at traffic signal junction
- 6.Footbridge or subway
- 8.Central refuge - no other controls

However it should be noted that with pedestrian crossing there is only a need to record when any of two following things happen:

- If a pedestrian is involved in an accident on,
- The pedestrian crossing facility was within 50 metres of the accident, along the route travelled or intended to be travelled, by vehicle(s) involved in the accident, and was considered to be a factor in the accident, regardless of whether there was direct pedestrian involvement.

Light conditions – this described the conditions of the light at an accident and they would only be recorded whether in daylight or in darkness (only half an hour after sunset or half an hour before sunrise)

- 1.Daylight

- 4.Darkness: street lights present and lit
- 5.Darkness: street lights present but unlit
- 6.Darkness: no street lighting
- 7.Darkness: street lighting unknown

Weather conditions - This refers to weather conditions at the time and location of the accident they are:

- 1.Fine without high winds – ('Fine without high winds' means any weather condition which does not have an adverse effect on driving)
- 2.Raining without high winds - ('Raining' includes drizzle, hail and sleet not tending to build up a deposit)
- 3.Snowing without high winds – ('Snowing' includes sleet building up a deposit)
- 4.Fine with high winds
- 5.Raining with high winds ('Raining' includes drizzle, hail and sleet not tending to build up a deposit)
- 6.Snowing with high winds
- 7.Fog or mist - if hazard - ('Fog' does not include light mists which did not constitute a driving hazard on the road where the accident occurred.)
- 8.Other - ('Other' should be used to indicate any other adverse weather condition which is not separately specified by the codes above. This code should be avoided if at all possible.)
- 9.Unknown

. Road surface conditions - this refers to the road surface condition at the time of the accident:

- 1.Dry
- 2.Wet/Damp
- 3.Snow
- 4.Frost/Ice
- 5.Flood (surface water over 3cm deep)

Special considerations at site – this section discards any special considerations that have to be made and they are:

- 0.None
- 1Automatic traffic signal out
- 2Automatic traffic signal partially defective
- 3Permanent road signing or marking defective or obscured
- 4.Roadworks
- 5.Road surface defective
- 6.Oil or diesel
- 7.Mud

Carriageway hazards – this section is outlines of any hazards that may have contributed to an accident and they are:

- 0.None
- 1.Dislodged vehicle load in carriageway
- 2.Other object in carriageway
- 3.Involvement with previous accident
- 6.Pedestrian in carriageway - not injured
- 7.Any animal in carriageway (except ridden horse)

Urban or rural – this section indicates that if the accident has taken place in either of the two environments

Did police officer attend scene of accident – this is to show if an officer attended a scene of an accident

- . 1.Yes – is meant to be used where a police officer attended the scene of the accident and obtained the details for a report. It may not be necessary for a reporting officer to have witnessed the actual accident for code 1 to be used.
- . 2.No - is to be used in all other cases (eg. accident reported "over the counter" at a police station) where details can be recorded by a police officer or member of civilian staff.
- . 3.No – Accident details completed by member of the public using “self reporting” form – is only meant to be used for public records details of the accident themselves using a dedicated “self reporting”

LSOA of accident location – this refers to LSOA meaning (*Leaving Scene of an Accident*)