CA4010 Data Mining and Data Warehousing
CASE4 Group 15
Report

Emma Duffy & Aoife De Buitlear
12495958 - 12376926

1.Introduction:

The following report is designed to give a broad overview and presentation of our findings from our dataset for the module CA4010.

1.1 Materials:

The accident data to be used in this project was obtained from http://data.gov.uk/dataset/road-accidents-safety-data (Data.Gov.UK, 2013a). This dataset contains the data for road accidents in the North-East and North-West of Great Britain.The statistics within our dataset relate only to personal injury accidents on public roads that are reported to the police, and contained data from 2005 to 2014.

With regards our dataset, we confirmed that our prediction was to “Predict who is most likely to be a fatality in a car accident” based on four years of statistics. Pre-processing, processing and analysis of the data was done using RStudio (Version 0.99.486), Microsoft Excel, python and Java for which we have included the source code in this report.

1.2 Dataset

The data was downloaded in the form of .csv files with separate files containing details related to the accident, the vehicle and the casualties from all accidents entered into STATS19. For the purpose of this investigation the vehicle’s data files were not of interest and only the data related to the accidents and casualties were included in our analysis. There was also an excel file with a breakdown of the csv contents values and what they meant, explaining what each value meant for each attribute e.g. If Sex_of_casualty column contains a 1, the casualty is male.

2. Data Preparation:

2.1 Pre-Processing

When preparing our dataset for our workshop, we realised that we had an abundance of data, and iterating through it was slow and time consuming. We made use of an open source, csv splitting program (CSV file chunker), which enabled us to split our csv file in a more manageable one. Using this program we were able to divide up our sample data (2010-2013) and test data (2014) into two separate csv

files. In order to concentrate on solely the accident and casualties files, we used a Python script to merge both files into one csv file by reference to the common Accident_Index which we called mergedOutput10_13.csv. Our dataset now contained the time period 2010-2013 and contained 791991 rows of data and 46 attributes. Our aim for the project is to predict and then prove 2014's results based on our findings from the four previous years.

The next step in the process was to isolate the variables and attributes of interest and of most relevance to our project. It was decided to retain the attributes contained in the table below which were either of potential interest or attributes that were of referential importance. We stored these attributes within ArrayLists in our java code, in order to check values from different columns against one another and compare values from the different years. Table 1 below displays the attributes and descriptions of our dataset.

It should be noted that the above final dataset contained a number of duplicate values as is evident through the inspection of the Accident_Index. These duplicate occurrences were due to the inclusion of multiple-car accidents as a number of single accident events. Therefore we did not delete duplicates as they each still held relevant and vital information.

Table 1 Accident Attributes and Descriptions

Attribute Name	Description	
Age of casualty	Numeric value between 0 and _	
Sex of casualty	1	Male
	2	Female
Age band of casualty	1	0 - 5
	2	6 - 10
	3	11 - 15
	4	16 - 20
	5	21 - 25
	6	26 - 35
	7	36 - 45
	8	46 - 55
	9	56 - 65
	10	66 - 75
	11	Over 75
Casualty Severity	1	Fatal
	2	Serious

	3	Slight
Day of the week	1	Sunday
	2	Monday
	3	Tuesday
	4	Wednesday
	5	Thursday
	6	Friday
	7	Saturday
Weather conditions	1	Fine no high winds
	2	Raining no high winds
	3	Snowing no high winds
	4	Fine and high winds
	5	Raining and high winds
	6	Snowing and high winds
	7	Fog or mist
	8	Other
	9	unknown
Number of vehicles involved		This contained numeric values for the amount of vehicles involved in the collision. This ranged from 1 to 10.

As discussed previously, when we first began analysing our dataset and working on our workshop (WS3 : Graphical Displays) we strongly believed that working with more attributes would give us a better understanding and insight into our dataset. However upon the beginning of our java code implementation, we realised that many of these attributes would not affect the fact of whether a person would be a fatality or not and could in fact skew our results. For this reason we wanted to ensure that we made our source code as efficient and effective as possible, we decided to concentrate on solely what we believed to be the most crucial attributes : Sex of Casualty, Age band of Casualty and finally Casualty Severity.

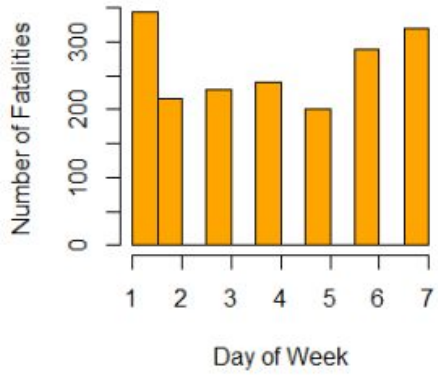
2.3 Workshop findings and results

The first proper insight that we got into our dataset was from the results and findings of our workshop. By using graphical displays such as scatterplots, histograms and quantile plots it made it a lot easier for us to find outliers and understand clusters within our data. Fortunately our data over the years 2010-2013 formed a distinguishable pattern, so this formed the basis of what direction we believed our prediction would go.

Each year there was a clear indication, from the specific displays, what the outcome would be and this included: That the fatality would be a male → this was due to there being 40% more males than females; The age band of the fatality would be in the range from 26-45; The most common weather type was dry weather, which was a surprise to us as we

expected wet weather to be a contributing factor to fatalities, not dry; The most likely day for the accident to occur was on the weekend (Sunday - Friday); and finally, the number of cars to be involved will be two. These were all found by getting the mode of each attribute for each year. The corresponding graphic displays were included in our workshop slideshow.

Below is attached a histogram showing the distribution of fatalities over the days of the week, clearing showing the 3 most common days are Friday, Saturday and Sunday



3.Algorithm Description

3.1 Decision Tree

In order for us to get the best understanding and results from our dataset, we both agreed that using classification would be the best data analysis approach. One of the reasons we decided to use classification was due to the fact that our aim was to create a model or classifier that would be constructed to predict categorical labels.

After much consideration and analysis of our dataset, we agreed that the best suited algorithm and method for constructing a model for our dataset would be a decision tree. A decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision tree consists of an internal node which denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label (a decision or classification). The topmost decision node in a tree which corresponds to the best predictor is called the root node.

With regards to our decision tree, we decided upon the Casualty Severity being our root node as this is the first attribute that will be evaluated. From the values within our dataset, there are three possibilities in the Casualty Severity column, therefore a check is needed to see whether the casualty was a fatality or not. If the casualty was not a fatality it was of no use to us, therefore it was assigned the class label “No Fatality” as seen below. From there we split on the next attribute, which we checked if the fatality was a male or female. If the fatality was a male, we then branched down once again, this time checking if their age fell into the age range which we discovered to be the mode of all male fatalities. If they did fall into this range, they were found to have the highest risk of everyone. The exact same was then checked for females, for whether they fell into the age range which was the

mode age for female. Anyone who did not have these characteristics were thus considered to be a lower risk. Fig.1 attached below is an image of the structure of our decision tree.

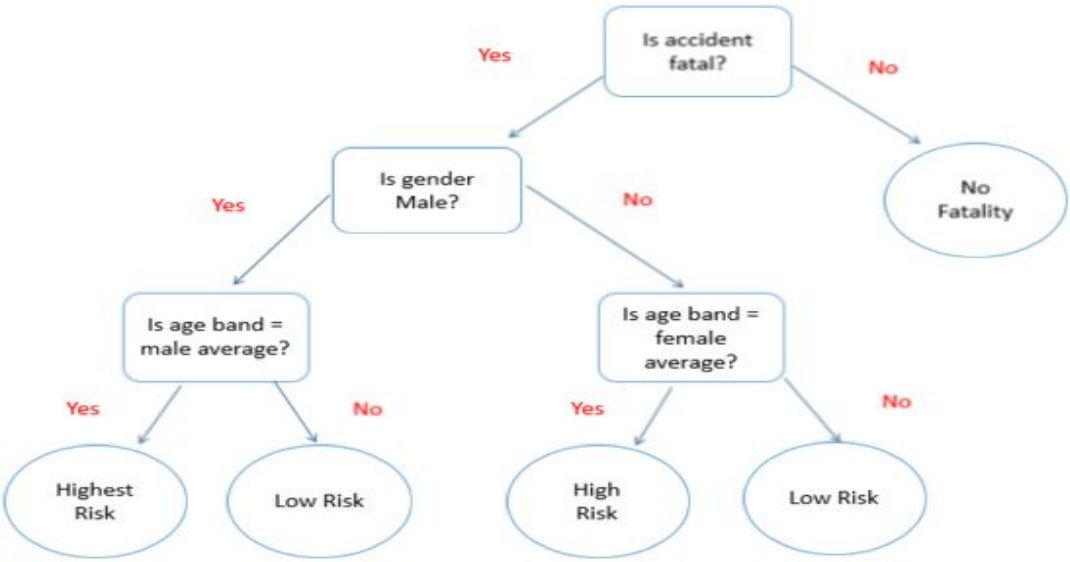


Fig1. Decision Tree

3.2 How it works

As you can see in the picture above our decision tree tries to classify who is most likely to be a fatality by taking into account different attribute scenarios. Once you have a constructed decision tree and want to classify some data sample you simply start at the root node and check if your data has a feature from the node. You do that until you reach a leaf node which already has an answer you were seeking.

3.3 Entire Process

In order to get the accuracy of our prediction and algorithm, there was a specified process that we had to step through in order to get to this point. The process began with our dataset , which consisted of the years 2010, 2011, 2012 and 2013. As we had chosen classification, the data for these four years was our training data. Training data is analyzed by a classification algorithm and then the learned model or classifier is represented in the form of classification rules. In order to make our prediction for 2014's results, we needed to thoroughly analyze our training data.

Central tendency played a key role in the formation of our prediction. We initially calculated the mode of our relevant attributes for each year individually. There were five main variables that we aimed to calculate and these were: what percentage of all casualty severities were fatalities; what percentage of these fatalities were male; what percentage of these fatalities were female; what percentage of fatal males had the mode age band for males and what percentage of fatal females had the mode age band for females. In the following section, *Results*, we have attached *Table 2* which contains all of the results that we established.

Once we ensured our mode figures were correct, we then combined the 4 years results together and used central tendency again to get the mean of them. Thus, the result

we received was established as our prediction. The results for our prediction is included in Table 3 below. Once our prediction was established, we then needed to construct our decision algorithm.

The next step after constructing our decision tree, was to substitute our test data into our algorithm in order to prove and establish accuracy of our prediction based on our test data results.

(See *Results* section for details on 2014 values and overall accuracy results)

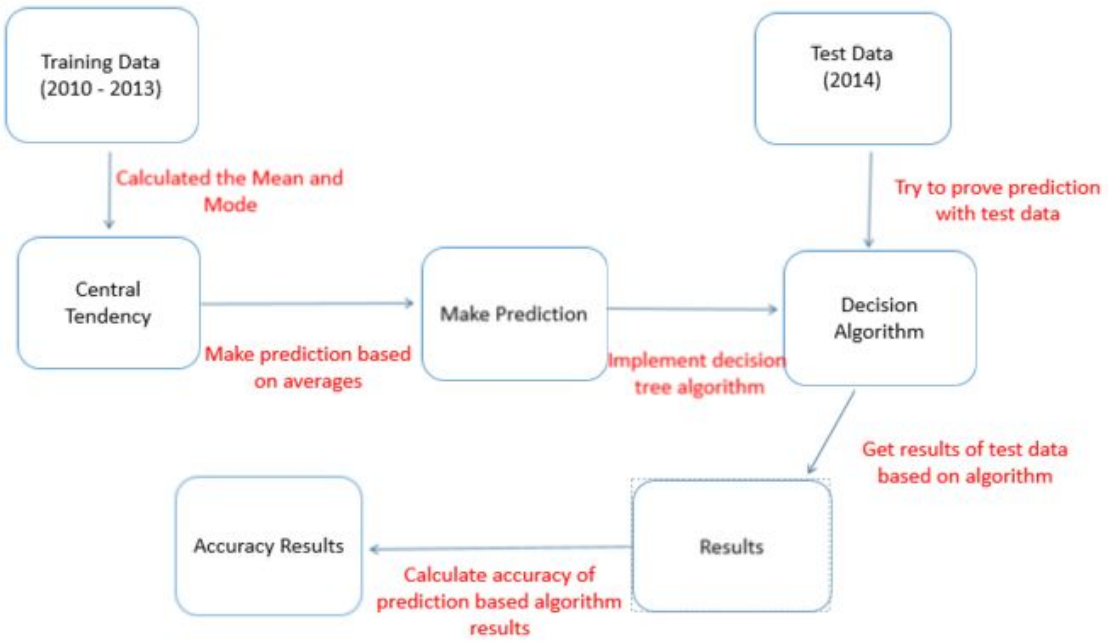


Fig2. Classification Process

4.Results

The following section provides a detailed description of the results and findings with regards to constructing our algorithm and also the results of our algorithm implementation into our test data.

Constructing a prediction was one of the most crucial aspect of the project. As stated above, to do so we used the central tendency to calculate the mean and the mode for each relevant attribute using our java source code. We then did this for each year from 2010-2013. Straight away it is clear to see a pattern forming, as for each year there was at most only a 3% difference. In order to make our prediction we then got the mean of all of these values. However, upon applying this to our java code, we realised there was a much simpler and less time consuming way of developing this comparison between the 4 years. We, instead, parsed our csv file containing the 4 years of test data into an ArrayList and iterated through it while incrementing specific values to gather our result data.

Table 2 below documents the results that we received.

Table 2:

	2010-2013

Total Accidents	791991
Total Fatal Accidents	7218
% of Fatal Accidents	0.91%
Total Males	462640
Total Male Fatalities	5398
% of Fatalities who were Male	74.78%
Total Females	329351
Total Female Fatalities	1820
% Fatalities who were Female	25.21%
Most common male age band for fatalities	6 (26 - 35)
Most common female age band for fatalities	11 (Over 75 years old)
Number of females in a fatal accident with age band = most common female age band	422
Number of males in a fatal accident with age band = most common male age band	882
% of females in a fatal accident with age band = most common female age band	23%
% of males in a fatal accident with age band = most common male age band	15.22%

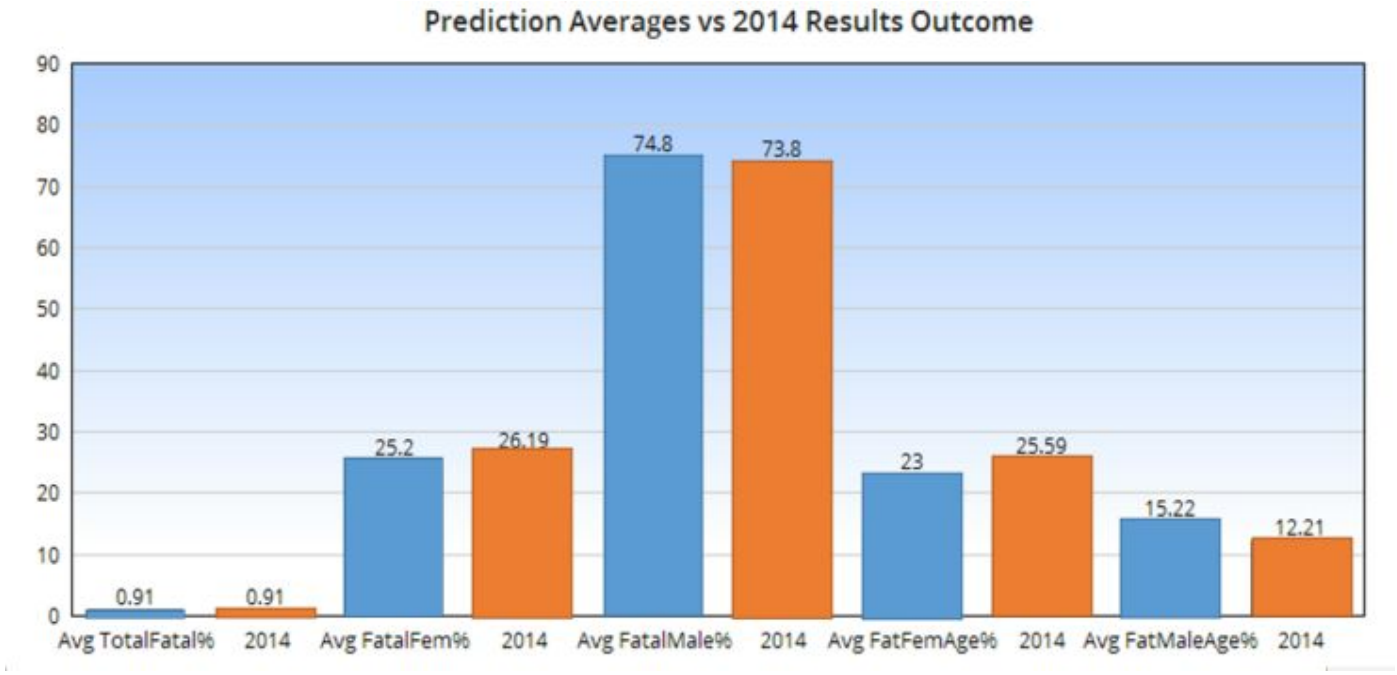
Our prediction data shows us that we would expect fatal accidents to occur near 1% of the time out of all car accidents, that there would be a significant difference between percentage for males and females (almost 3 times the amount), and finally that the average percentage of fatalities whose age is equal to the mode age band is around 23% for females and 15.22% for males. Therefore, combining these values together we constructed the prediction that the person most likely to be in a fatal car accident would be male, between the age of 26-35 (Age band 6).

In order to prove the accuracy of our prediction, we needed to prove that the corresponding values for 2014 resembled our average values. Below you can see that we calculated the exact same variables as we had done so for 2010,2011,2012 and 2013. Immediately patterns began to form in similarity with the two sets of values. All outcomes were as we had hoped and expected, that they would vary in a range of 1-3% of a difference.

Table 3:

	2014
Percent of fatalities	0.91%
Percent Male Fatalities	73.80%
Percent Female Fatal	26.19%
% of males in a fatal accident with age band = most common male age band	12.21.%
% of females in a fatal accident with age band = most common female age band	25.59%

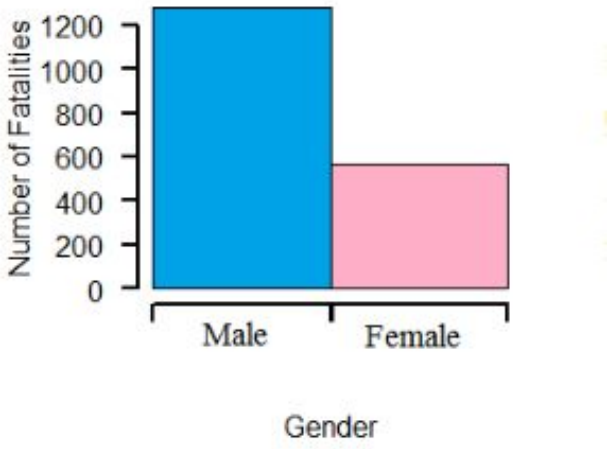
In order to give a clearer representation of the relationship between our prediction and our 2014 results, below we have included a chart to show the similarities in values. The bars in blue represent the percentage averages of our prediction and the bars in orange represents our 2014 results. By displaying it visually it is easier to see the main difference is with regards Percent Fatal, Female and mode age band and the 2010-2013 average.



Our results taught us a lot about our dataset. Initially at the beginning of our project we felt that we already had a basic knowledge and estimation of what attributes the person most likely to be a fatality would contain. However, throughout our investigation we encountered some data that surprised us and some that was different to what we had first expected to find.

One of the main unexpected results that we received from our findings was the huge difference in numbers with regards male fatalities and female fatalities. The most immediate pattern formed each year was the huge amount of male fatalities in comparison to females.

We investigated further into our dataset and found that on average each year there is about 70-75% males and 25%-30% females.



Another example of unexpected results was the mode age band for females. Originally we had the assumption that both male and females that were fatalities would have have very similar age bands, perhaps more younger males than females. This in fact was the case, however we never could have foreseen the extreme difference in ages. For males, the mode age band of a fatality was the age band 6, which is the age range 26-35, whereas the female mode age band is 11 , which is the age range over 75.

5.Conclusion

In conclusion, we were very happy with the success rate of our prediction initially, due to the fact it gained a accuracy rate of greater than 90%. Once we added the age band attribute we were able to deliver a much more defined prediction at an accuracy rate of 61% which we were satisfied with. When we added even more to the prediction however, we noticed a great decrease in the accuracy. We had planned on adding the attribute of weather to our prediction also but due to this resulting in very low percentage values (Percentage of females with all attributes matching prediction: 3.267605633802817% .Percentage of males with all attributes matching prediction: 7.774647887323943% - all attributes being gender, age band and weather) we felt that it did not contribute to our prediction. Additionally we feel that we have learned and gained a lot of knowledge about our dataset, data mining tools and techniques, pattern matching and logical thinking, as well as the importance of data mining for a software engineer.

While the data in this project did provide a clear indication of where problems lie within specific groups, further research and analysis with additional attributes would be of huge benefit. If time was not a constraint, further investigation into human contributing factors and also non-human factors would have been of interest to us.

In reflection of our workshop, we realise that Graphical Displays was a good choice for us, as it helped us visualise the data and aided in our understanding of it - which is what we had hoped for. The information we discovered from our workshop preparation was extremely valuable to us when deciding upon our algorithm. Some figures changed slightly when we implemented our java code, but for the majority of cases, our prediction was as expected and quite accurate.