Fall 2016

Identifying the factors that affect the severity of vehicular crashes by driver age

John Dietrich Tollefson
University of Iowa

IDENTIFYING THE FACTORS THAT AFFECT THE SEVERITY OF VEHICULAR CRASHES BY DRIVER AGE

by

John Dietrich Tollefson

A thesis submitted in partial fulfillment
of the requirements for the Master of Science
degree in Electrical and Computer Engineering in the
Graduate College of
The University of Iowa

December 2016

Thesis Supervisor: Assistant Professor Guadalupe Canahuate

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

_____

MASTER'S THESIS

_____

This is to certify that the Master's thesis of

John Dietrich Tollefson

has been approved by the Examining Committee for
the thesis requirement for the Master of Science degree
Electrical and Computer Engineering at the December 2016 graduation.

Thesis Committee: _____
Guadalupe Canahuate, Thesis Supervisor

_____
Jon Kuhl

_____
Gary Christensen

To my parents and all my teachers who helped make this possible

To the University of Iowa Injury Prevention Research Center who funded our project and especially Dr. Corinne Peek-Asa and Tracy Young who gave helpful advice and consulted with us on directions to proceed in.

ABSTRACT

Vehicular crashes are the leading cause of death for young adult drivers, however, very little life course research focuses on drivers in their 20s. Moreover, most data analyses of crash data are limited to simple correlation and regression analysis. This thesis proposes a data-driven approach and usage of machine-learning techniques to further enhance the quality of analysis.

We examine over 10 years of data from the Iowa Department of Transportation by transforming all the data into a format suitable for data analysis. From there, the ages of drivers present in the crash are discretized depending on the ages of drivers present for better analysis. In doing this, we hope to better discover the relationship between driver age and factors present in a given crash.

We use machine learning algorithms to determine important attributes for each age group with the goal of improving predictivity of individual methods. The general format of this thesis follows a Knowledge Discovery workflow, preprocessing and transforming the data into a usable state, from which we perform data mining to discover results and produce knowledge.

We hope to use this knowledge to improve the predictivity of different age groups of drivers with around 60 variables for most sets as well as 10 variables for some. We also explore future directions this data could be analyzed in.

PUBLIC ABSTRACT

This thesis proposes a data-driven approach and usage of machine-learning techniques to further enhance the quality of analysis of car crash data analysis.

This thesis examines car crash data by looking at the different aspects of each crash. We divide the crashes into 6 different groups depending on the ages of drivers involved and attempt to determine important features of each group as a result of this. In doing this, we hope to make clear what factors lead to crashes in different age groups and work to avoid them.

This data could then be potentially used for the benefit of automakers, insurance companies, the trucking industry, and individual consumers. Perhaps having more insight might allow travel to become safer for everyone.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 3 TABLES

CHAPTER 4 TABLES

LIST OF FIGURES

INTRODUCTION

Vehicular crashes are the leading cause of death for young adult drivers, however, very little life course research focuses on drivers in their 20s. Moreover, most data analyses of crash data are limited to simple correlation and regression analysis [15-17]. Working in collaboration with the University of Iowa Injury Prevention Research Center (IPCR), we propose a data-driven approach and the use of machine learning techniques to further improve the quality of the analysis.

The data used in this thesis belongs to the Iowa Department of Transportation. These data are collected in police-reported crashes and include information about the environment, roadway, vehicles, and people involved in the accident. These relational data are organized into three levels: crash, person, and vehicle. All data are tabulated and attributes correspond to numeric or categorical features (with the exception of narrative text that are stored as natural language). Over ten years (2001-2012) of data is currently available for research.

This work focuses on processing these data in order to generate feature vectors that are amenable to data mining and then apply machine learning algorithms to identify the most relevant set of features given an age-group. With this, we hope to be able to better serve different age groups by methods such as tailoring safety information in crashes to demographics who are more likely to deal with crashes affected by those factors. The results of these analyses could then be potentially used for the benefit of automakers, insurance companies, the trucking industry, and individual consumers. Perhaps having more insight might allow travel to become safer for everyone.

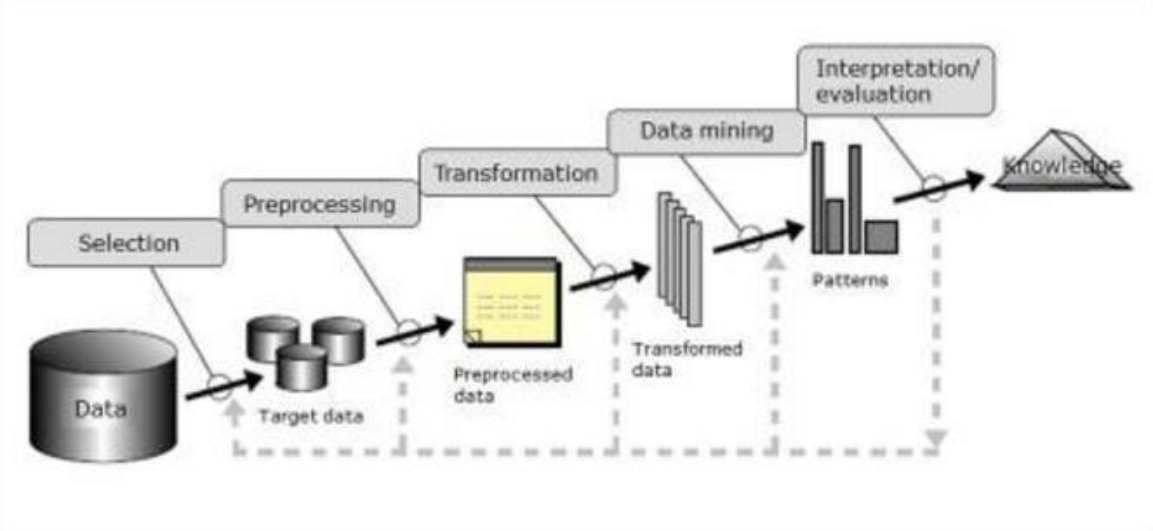
The rest of this thesis is organized as follows. Chapter 1 presents an overview of the methodology followed in this project. Chapter 2 describes the state of the data and how we process and transform it into a state ideal for machine learning processing. In Chapter 3, we will discuss the methods we used in order to extract variables and methods used for classification.

Chapter 4 will detail the results of these methods and some preliminary conclusions. Lastly, we conclude in Chapter 5 and provide directions for future work.

CHAPTER 1 PROJECT OVERVIEW

This project follows the general workflow of a Knowledge Discovery Process depicted in Figure 1.1. As shown in the figure below, the flow of such a problem begins with selection of data to form our target data, followed by preprocessing and transforming the data into a usable state. Then begins the data mining process, from which we extract patterns. Finally, these patterns are interpreted and evaluated, resulting in knowledge.

Figure 1.1: Figure Demonstrating the Knowledge Discovery Process [1]



The data is originally in a state that is not conducive to performing of machine learning algorithms, each bit of data either belonging to the Crash, Person, or Vehicle Levels. To alleviate this problem, we flatten the data onto the Crash Level, as detailed in Chapter 2.

From there, we proceed to use machine learning algorithms to determine important features. These features are classified along a class label (fatality, injury, or no injury), which we derived from the crash severity reported. All algorithms used in this thesis are provided by Weka [12]. Weka is an open source program containing many machine learning algorithms to be used in data analysis. It has its own format, ARFF, which can specify whether the data is nominal, numeric, or the spread of it within the header of the file. One thing to note with this method of

calculation, however, was that often the larger amounts of data, such as with the Age Group containing drivers aged 21-64, would not run the algorithms properly on the 4 GB laptop most data processing was done on. To this end, we moved the data to a university computer with a larger, 16 GB RAM and used it with accessing a remote desktop to compute further results.

Using Weka, we create classification trees for the different driver age groups using all the collected and transformed features from the accident involving the vehicles, the drivers, and the crash characteristics.

Features/values appear in all classification trees indicate common factors that do not differentiate between different age groups. Features/values that appear in few of the trees are candidate factors that deserve further evaluation. Additionally, to evaluate the relevance of the extracted features we applied different metrics such as information gain and chi-square.

In order to determine the accuracy of our training methods, we have set the data from 2012 apart for testing while using the rest of the data to train classifiers. This is important because testing on a different data set than you trained on makes sure that the data has not been overfitted. If it were, it would predict only the training set well and not classify new data as well as it could. As such, we merge the data from 2001 to 2011 into one set.

For our machine learning methods, we decided to use FMeasure as an evaluation metric rather than accuracy. Accuracy and FMeasure are calculated via the following formulas.

Figure 1.2: Formula for Accuracy

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN)$$

Figure 1.3: Formula for FMeasure

$$\text{FMeasure} = 2*(\text{precision}*\text{recall})/(\text{precision}+\text{recall})$$

In the accuracy formula, TP stands for True Positive, TN for True Negative, FP for False Positive, and FN for False Negative. True Positive is the number of values that have a certain attribute value and were classified as having that value for that attribute. True Negative are records that don't have this attribute value and are classified as such. False Positive are classified as having the value when they do not and False Negatives are the inverse, being classified as not having a certain value when they do in reality. Essentially, accuracy measures the number of correct predictions over the number of total predictions. As well, Precision is equivalent to TP/(TP+FP), or the number of positive predictions that were correctly predicted as positive. Recall is equal to TP/(TP+FN), or the number of positive values that were correctly predicted as positive.

The FMeasure was selected in order to account for both false positives and false negatives better rather than just positive predictive accuracy. This was especially important as approximately 70% of crashes in the data involve no injuries and less than 1% involve fatalities. By weighting fatalities as more important, we hope to be able to predict them better.

As we describe these methods further in detail, we hope to show how they can be used to analyze the data better.

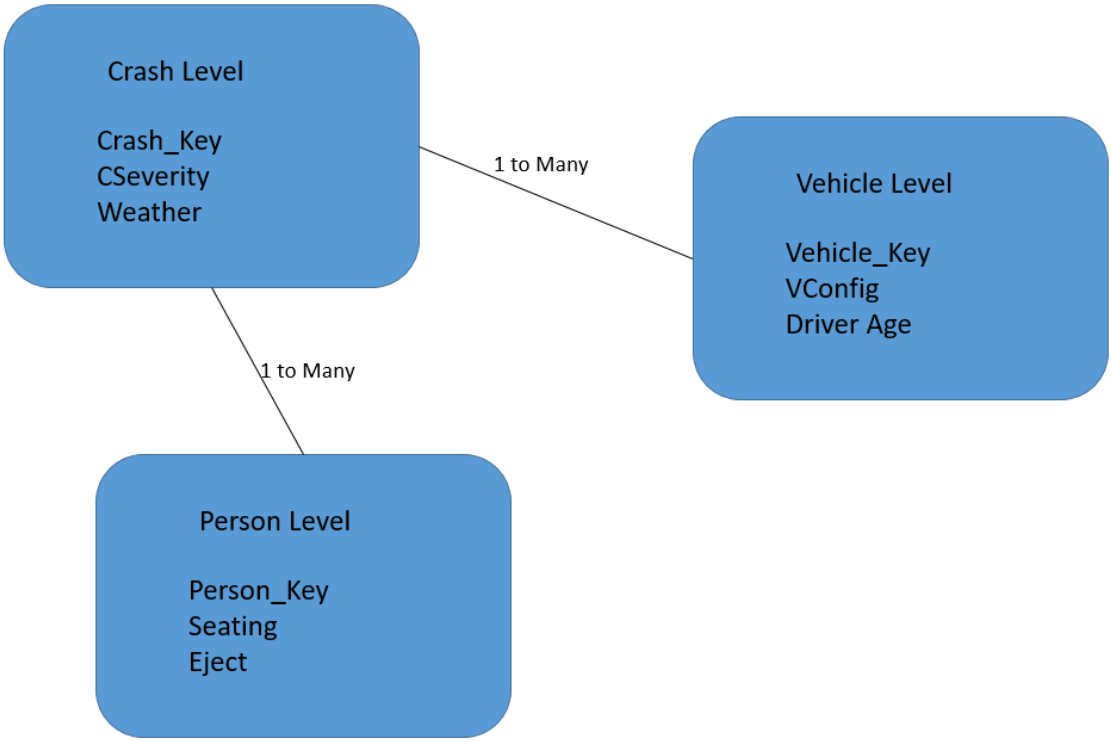
CHAPTER 2 DATA PROCESSING

This chapter will go into why we decided to look at the data in the way we did and how we originally modified the data to accommodate the one to many problem with the data structure and how this made things difficult. We will also describe how we determined which age groups to split the data into. First, we will talk about how we restructured the data and why we did this and then we will discuss how we split the data based on age groups. First, we will discuss the overall state of the data. This will be followed by how we appropriately transformed the data from its original state to one more suited for analysis, then we will discuss how we divided the data into different groups to analyze the values based on ages.

2.1 Data Overview

Altogether, there are a bit over 625,000 crashes accounted for in the Iowa Department of Transportation data set. This data is organized into several tables, each of which falls into one of three levels. The first level, the Crash Level, focuses on variables related to the overall crash. All these entries are related to each other through a Crash ID. The second level, the Person Level, focuses on the state of people after the crash. These entries are related to each other with a Crash and a Person ID. The third level, the Vehicle Level, focuses on data with the vehicles. These entries are related to each other with Crash and Vehicle IDs. This data is also organized into separate years from 2001 to 2012. Each year has the same overall structure of data, they are merely split for ease of keeping years isolated. Altogether, there are 39 features that we were able to use for data analysis among these.

Figure 2.1: Organization of the Crash Data



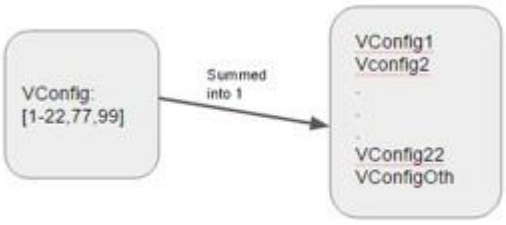
The original state of the data is quite good for organizing it, but it is not the best for analyzing it with data mining. This is because the Person and Vehicle Levels of the crash both have a one to many relationship with the Crash Level, as depicted in Figure 2.1. This means that for each crash ID and its related data, there could be any number of Vehicle IDs or Person IDs with their related data connected with it as well. The way that normal machine learning algorithms work requires the same number of variables for all entries, which could not be done with a variable amount of vehicle and passenger info related to each crash. To achieve a data structure that we could apply normal algorithms upon, we ultimately decided to aggregate the records of the Vehicle and Person levels to collapse everything onto one level of data.

2.2 Feature Generation

As stated previously, the data is organized into multiple tables, each of which falls under one of three overall levels that the data is organized into. The first level, Crash Level, covers data on the crash as a whole. This includes factors such as the overall severity of the crash, time of the crash, weather and road conditions, and whether the road the crash occurred on was in a Rural or Urban area. The second level, Person Level, features data on each of the individual people involved in the crash. This includes information such as where the people were seated, how old they were, and whether they were wearing seatbelts. The third set, Vehicle Level, includes data on the vehicles involved. This includes factors such as type of vehicle, where the vehicle incurred the most damage, and whether the driver of a vehicle was distracted in some way.

We decided that, given the structure, the best way of transform the data was to relate everything on the Crash Level. Given the Crash Level is the highest Level, this would allow us to avoid duplicating information unnecessarily while still keeping the overall information intact. As well, the overall severity of the crash, CSeverity, was what we decided we wanted to predict, so it was a better idea to transform the data from the other two levels to the Crash Level. As a result, each feature that was originally at the person or vehicle level was given a different variable for each potential option for it and each of these variables was then assigned the value equal to the number of occurrences of that type in a given crash. To illustrate this, we will take an example involving VConfig.

Figure 2.2: Transformation of VConfig



The above figure demonstrates the transformation of the data. In this, each instance of VConfig is tallied by its value, then counts of the same value are counted to form the values of new variables. For example, a crash involving 3 vehicles: two passenger cars, indicated by VConfig1, and a train, indicated by VConfig22, would have a VConfig1 value of 2 and a VConfig22 value of 1, all other values being 0. Those crashes which did not have any person or vehicle level information in the dataset were set to 0 by default. In addition, due to the data originally in zinj and zuni only differing in zinj containing injured people and zuni containing non-injured, one query was made to sum their results together to be in the same fields. Altogether, we had 250 features extracted via these methods. A list of the approximately features we generated can be found in Appendix A. The list of queries we applied in order to get these variables can be found in Appendix B.

2.3 Age-Based Splitting

We theorize that each age group will have different variables that predict how severe crashes among it are than other ones. As such, when organizing our data, we split it into different files, one for each Age Group.

To determine which ages would be the best to split by, we asked experts from the IPCR who had worked with this data previously, Dr. Corinne Peek-Asa and Tracy Young. From their advice, we have divided the data into 5 age groups. Group A contains crashes with drivers of ages

15 or younger, before they would have a driver's license. Group B contains crashes with drivers aged 16-20, young drivers who cannot legally purchase alcohol. Group C contains crashes with drivers aged 21-64, the largest data set and the biggest group of drivers. Group D contains drivers aged 65-74, for drivers who are older than Group C, but not as old as Group E. Group E contains crashes with drivers aged above 75, the oldest age group. We also maintained a final, where none of the drivers' ages were known. We decided to group these crashes under Group U and analyze them as we do the others.

As we looked at the splits in these values, we discovered various interesting things. For example, despite its low range of ages, over 26.42% of the crashes fall in the group containing drivers aged 16-20. For further comparison, the drivers aged 65-74 make up 8.07% of crashes while the drivers aged 75 and older make up 6.74%. This backs up previous papers' assertions that younger drivers are more likely to be in accidents [13-14]. In terms of comparing injury severity, we also determined that both age groups containing drivers aged 65 and up had an increased number of fatalities in them when compared to the overall average. To compare, .66% of the overall crashes contained a fatal accident while 1.22% of the ones containing a driver aged 75 or older did. This tells us that, while said group might be less likely to have crashes, if they do get into one, it is more likely to be fatal.

CHAPTER 3 ATTRIBUTE SELECTION

After getting our data ready for analysis, we then went through various methods to try and determine the best methods to determine the features most important for each of the six groups. We did this by first determining what measure of a good model we would use. From there, we used various methods such as GainRatioAttributeEval and, later, extracting variables from J48 trees calculated using different methods.

3.1 Variable Selection Across Methods

Originally, we attempted to determine the important features by using multiple methods, finding commonalities between the variables they selected and using that to determine each set of variables. We use a mix of Supervised and Unsupervised Methods for this. Supervised Methods are aware of the Sample Class while Unsupervised Methods do not use class information [10].

The first method we tested was Weka's CFSSubsetEval. This method, as first elaborated on in the paper by M. A. Hall, creates a subset of features that maximizes correlation with each feature to the class variable while minimizing the intercorrelation of the features to each other. [2]

Another method we used was Weka's GainRatioAttributeEval. This method, implemented by M. A. Hall, evaluates an attribute's worth by comparing the gain ratio of adding that attribute to the class. This is calculated via the function GainR(Class, Attribute) = (H(Class) - H(Class | Attribute)) / H(Attribute). [3] In this formula, H(Class) is a measure of the information entropy of the class, which is the expected value of the amount of information in the class. H(Class | Attribute), in turn, means the information entropy of the class given the attributes.

A third method we used was Weka's ChiSquaredAttributeEval.This method, implemented by E. Frank, evaluates the worth of an attribute by computing its Chi Squared Value relative to the class. A Chi Squared Value is a value used to determine if the correlation or difference between two values is significant or not [4].

The Chi Squared Value can be calculated by the formula $D = ((O - E)^2 / E)$ [5]. In this formula, O stands for Observed: The number of values that fit in a given format, such as a given variable being 1 during a crash with a fatality. E stands for Expected, which is the expected value if the values were all evenly distributed. In the above example, would be calculated by multiplying the number of instances of the variable being 1 by the number of fatal crashes and dividing that by the total number of entries. For this method, the resulting D will be calculated for all potential combinations, resulting in the end value.

3.2 Combining of 3 Methods

Once we ran these three methods in Weka for each of our 6 Age Groups with CSeverity as the class, we saved their results and compared them to each other. However, the results are slightly different in that CFSSubsetEval gives a portion of the attributes that fit most while GainRatioAttributeEval and ChiSqaredAttributeEval both give all the attributes ranked by their Gain Ratio or Chi Squared Value, respectively. For these two methods, we considered the top 20 and the top 50 attributes and then combined the three methods by using a voting majority.

Figure 3.1: Selected Attributes when k=20

	Whole	A (<=15)	B (16-20)	C (21-64)	D (65-74)	E (>=75)	U (?)
CauseAnimal	X			X			
CauseOversizedLoad			X				
MinorPassengers							X
18-70Passengers	X			X			X
FSeat	X	X	X	X	X	X	X
BSeat	X	X	X	X	X		X
FPassenger	X	X	X	X	X	X	X
Pedestrian	X		X	X	X	X	X
Pedacyclist	X		X	X	X	X	X
OccProcNone	X	X	X	X	X	X	X
OccProcOth							X
EjectNot	X		X	X	X	X	
EjectFull	X	X	X	X			X
FrontEjected							X
SideEjected	X	X	X	X	X		X
FrontAirbagDeploy	X		X	X	X	X	
TrapNonMech					X	X	
TrapMech	X	X	X	X	X	X	
TrapNot							X
DriverGenderUnknown							X
Motorcycle	X			X			
Moped/ATV		X					
Totaled	X	X	X	X	X	X	

Figure 3.2: Set 2 Selected Values

	Whole	A (<=15)	B (16-20)	C (21-64)	D (65-74)	E (>=75)	U (?)
Vehicles							X
Drug	X	X	X	X	X	X	
DANone	X		X	X			
NonColl		X	X				
CauseAnimal	X			X	X		
CauseOversizeLoad			X				
CauseCargoLoss		X					
CauseROWPedestrian							X
CauseFatigued							X
Rural		X					
Urban		X					
UnpavedRoad		X					
WeatherOth	X			X	X		
LightOth	X			X	X		
WaterOnRoad		X					
CSurfCondOth	X			X	X		
MinorPassengers		X		X	X		X
18-70Passengers	X			X		X	X
SeniorPassengers						X	X
FSeat	X	X	X	X	X	X	X
BSeat	X	X	X	X	X	X	X
FDriver		X				X	X
FPassenger	X	X	X	X	X	X	X
RDriverSide	X	X	X	X	X	X	X
RPassengerSide		X	X				
3MiddleSeat							X
3PassengerSide		X	X	X			X
Pedestrian	X	X	X	X	X	X	X
Pedacyclist	X	X	X	X	X	X	X
SeatingOth							X
OccProcNone	X	X	X	X	X	X	X

Figure 3.2 cont: Set 2 Selected Values

	Whole	A (<=15)	B (16-20)	C (21-64)	D (65-74)	E (>=75)	U (?)
ShoulderLapBelt			X		X	X	
Helmet		X	X				
OccProcOth							X
EjectNot	X	X	X	X	X	X	X
EjectFull	X	X	X	X	X	X	X
FrontEjected	X		X	X	X	X	X
SideEjected	X	X	X	X	X	X	X
FrontAirbagDeploy	X	X	X	X	X	X	X
FrontSideAirbagDeploy					X	X	
AirbagNotDep							X
AirDepOth							X
TrapNot	X	X	X	X	X	X	X
TrapNonMech	X	X	X	X	X	X	
TrapMech	X	X	X	X	X	X	X
EmerOth							X
CargoOth	X						
AtStoplight							X
DAgeBin3							X
DriverGenderUnknown							X
MobileHome					X		
Motorcycle	X	X	X	X	X	X	
Moped/ATV		X					
MostDamageTop	X		X		X		
NoDamage					X		
DisablingDamage					X	X	
Totaled	X	X	X	X	X	X	

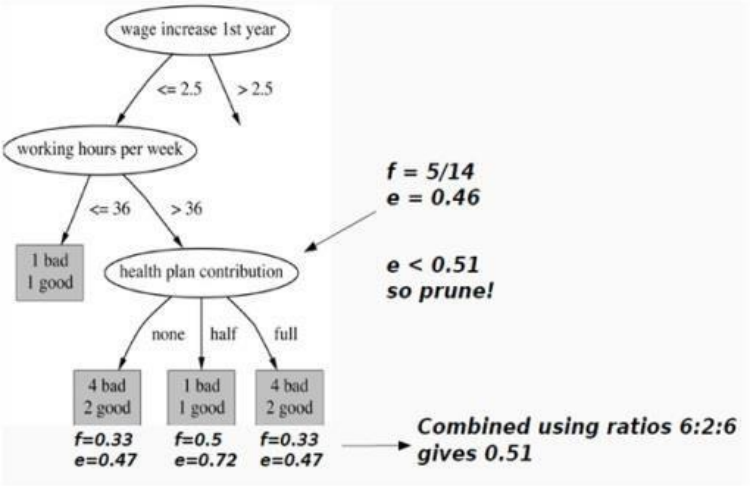
3.3 Classifiers

The First Classifier we use is Naive Bayes. This Classifier analyzes the training data and generates probabilities for whether a certain value of a variable is associated with a class or not. It then compares probability given the test point and assigns the value to a class based on that.

The second classifier we used was J48, a Weka Method which attempts to create a tree to split based on values and predict the class from this. J48 works via a C4.5 Algorithm, which is an extenstion of the ID3, or Iterative Dichotomizer 3, Algorithm. This algorithm works via a method known as information gain. Namely, it iterates through each potential variable split and determines which split results in the most information gain. It then splits by this variable and repeats the process until it has reached a point where it can classify. [11]

Another important feature of J48 is pruning. Once the tree has been built, it iterates through the branches and determine how confident it is with each one. If its confidence in the branch is lower than the confidence factor, then the branch is pruned. This is used to improve generalization of the classification and avoid overfitting. An example of how this would work is shown in the figure below.

Figure 3.3: Pruning Example [15]



3.4 SevGroup

To reduce the sparsity of the samples, our collaborators from IPCR, Dr. Corinne Peek-Asa and Tracy Young, usually lump CSeverity possibilities together to have fewer classes. Our original class variable, CSeverity, has 5 different options, Fatality (1), Major Injury (2), Minor Injury (3), Possible Injury (4), and No Injury (5). There were two ways we thought to do this. The first was to have it grouped by Fatality (1), Non-Fatal Injury (2,3,4), and No Injury (5). The second was to have it grouped by Major Injury (1,2), Minor Injury (3, 4) and No Injury (5). To determine accuracy, we recompiled the data with two additional classes, SevGroup1 and SevGroup2, representing the first and second option of grouping, respectively. We tried using them as the class variable with J48 and compared the FMeasures resulting with the original CSeverity.

3.5 Fatality Groups

Another method we experimented with to offset the low number of fatalities was to oversample the fatalities. We merged all the data from 2011, with the fatalities from the years 2001 to 2010 to use as our training data. We labeled this data as 'Fatality'.
We then compared the results of this method to that with methods trained using the data from 2001 to 2011, labeled as 'Full', as shown below.

Table 3.1: Full vs Fatality Group, No Cost Matrix, SevGroup1, FMeasure

	A (<=15)		B (16-20)		C (21-64)		D (65-74)		E (>=75)		U(?)	
	Full	Fatality	Full	Fatality	Full	Fatality	Full	Fatality	Full	Fatality	Full	Fatality
Fatal	0.25	0.067	0.286	0.198	0.339	0.3	0.222	0.255	0.192	0.159	0	0
Injury	0.761	0.639	0.652	0.584	0.647	0.587	0.638	0.58	0.667	0.573	0.963	0.935
No Injury	0.88	0.871	0.857	0.853	0.877	0.873	0.872	0.865	0.871	0.865	0.996	0.996
Weighted	0.834	0.784	0.791	0.767	0.809	0.789	0.8	0.779	0.802	0.769	0.99	0.987
Average	0.63	0.525	0.598	0.545	0.621	0.587	0.577	0.567	0.577	0.532	0.653	0.644

In order to get a better picture of the FMeasure across age groups, we used two methods of average: Weighted Average and Average. Weighted Average is the result automatically calculated by Weka, which averages each of the 3 FMeasures weighted by the percentage of entries that actually are that class. This gives a good measurement of how well this predicts for the set overall. Average, however, calculates the average of the three using equal weights. This method is better for determining how much differences between specific classes affected things, as approximately 70% of the crashes did not involve an injury. Together, these can both paint a better picture of how good the different methods are for prediction, ultimately aiming for a higher Weighted Average while keeping Average high as well.

Table 3.2: Full vs Fatality Weighted and Average

	Weighted		Average	
	Full	Fatality	Full	Fatality
A (<=15)	0.834	0.784	0.63	0.525
B (16-20)	0.791	0.767	0.598	0.545
C (21-64)	0.809	0.789	0.621	0.587
D (65-74	0.8	0.779	0.577	0.567
E (>=75)	0.802	0.769	0.577	0.532
U (?)	0.99	0.987	0.653	0.644
Average	0.838	0.813	0.609	0.567

Unfortunately, as we can see from the results, this method did not pan out like we hoped it had. The oversampling of fatalities and lack of other points led to more points getting classified incorrectly overall.

3.6 Variable Extraction from Trees

Another supervised variable extraction method we used was by extracting the values from the trees generated by J48. We started by running this method for all 6 Age Groups with Severity as the class and storing the trees as text files. From there we wrote a Java Program to parse the tree and determine the important values. The code for this program can be found in Appendix C. This code works by keeping track of the variables in each branch as it iterates through the text.

Our first measure of 'importance' for variables were ones that appeared earlier in the tree. We decided to do 10 to see how well it did. In the end, our results varied a fair deal. The Method did work well for the groups containing drivers younger than 15 and of unknown ages, but the

others all had a decrease in how well it predicted Fatalities. We reasoned that the aforementioned groups had improved results because their trees were both heights less than 10.

For a better option, we then rewrote the program to determine whether each branch of a tree ended up being used to predict a Fatality or not. This was due to fatalities not being properly classified. In focusing on fatalities, we hoped to give them more weight and, thus, classify them better. If so, all the attributes contained in the branch would be considered important attributes. This method was better as it allowed us to reduce the number of total attributes while still predicting fatalities rather well. It did, however result in different sizes for the attributes that were considered important. For instance, the Age Groups with drivers 15 and younger and the unknown age drivers had fewer than 10 splits while the group containing drivers aged 21-64 had a bit over 200. The other 3 age groups, however, typically were between 40 and 70 splits.

3.7 Cost Matrices

While using J48 to measure variables, we considered whether having a weighted cost matrix would help us boost the classification accuracy for fatalities. Cost Matrices are a technique by which different classification errors are weighted more heavily than others. These are often done in situations where the classifier needs to take into account that one type of error is less desired than another, such as in our current situation. We decided that FMeasure would be the best measurement of accuracy to count all the classes, using both the individual FMeasures and the weighted FMeasure. We compared the following 3 Cost Matrices

Figure 3.4: Original Cost Matrix

0	1	1
1	0	1
1	1	0

Figure 3.5: '10' Cost Matrix

-10	10	20
10	-1	5
20	5	0

Figure 3.6: '789' Cost Matrix

-78.9	78.9	175.2
78.9	-1	2.22
175.2	2.22	0

The 10 Cost Matrix was the result of weighting the correct predictions of Crashes with Injuries as more important than correct predictions of Crashes without Injuries by making them negative and giving correct Fatality predictions a stronger weighting than correct non-fatal injury Predictions. In addition, to try and maximize predictivity of fatalities while hopefully keeping a high overall FMeasure, we weighted a fatality predicted as no injury as worse than a fatality predicted as a non-fatal injury, which in turn was weighted as worse than a non-fatal injury

predicted as a injury-free crash. This matrix was originally conceived as an example for how to weight a cost matrix, which is why most of its values are divisible by 5. There is unfortunately no clear method to predict good cost matrices, which is why we picked numbers such as these. The 789 Cost Matrix was made using a similar pattern to the 10 Matrix, but by choosing the numbers by ratio of sizes of each group, making sure it has either 3 significant figures or at least one decimal value. We hope this will result in better predictions as a result.

Table 3.3: Ages 21-64, Cost Matrix Comparison

	1	10	789
Fatal	0.339	0.379	0.361
Injury	0.647	0.644	0.641
None	0.877	0.888	0.891
Weighted	0.809	0.816	0.818
Unweighted	0.621	0.637	0.631

Table 3.4: Ages 65-74, Cost Matrix Comparison

	1	10	789
Fatal	0.222	0.333	0.3
Injury	0.638	0.634	0.631
None	0.872	0.881	0.884
Weighted	0.8	0.806	0.807
Unweighted	0.577	0.616	0.605

We ran J48 with each of these cost matrices on the full tree, comparing the FMeasures of each. We thought 10 was the best Cost Matrix from a quick glance and went with it for some earlier results, but we decided to average all the results after this to be certain.

Table 3.5: Cost Matrix Comparison, Weighted Averages

	1	10	789
A (<=15)	0.834	0.82	0.82
B (16-20)	0.791	0.795	0.796
C (21-64)	0.809	0.816	0.818
D (65-74)	0.8	0.806	0.807
E (>=75)	0.802	0.802	0.804
U (?)	0.99	0.988	0.991
Average	0.838	0.838	0.839

Table 3.6: Cost Matrix Comparison, Averages

	1	10	789
A (<=15)	0.630	0.609	0.682
B (16-20)	0.598	0.585	0.589
C (21-64)	0.621	0.637	0.631
D (65-74)	0.577	0.616	0.605
E (>=75)	0.577	0.625	0.615
U (?)	0.653	0.649	0.654
Average	0.609	0.620	0.629

This could be altered by the deceptively high value for 789's predictions of fatalities for ages 15 and under, but it looks like, overall, 789 outperforms 10 by a bit in both categories after all. So, to that end, future results will be using the 789 Cost Matrix in future attempts.

3.8 minNumObj

For further refining, we looked at the minNumObj Attribute of the J48 Tree. This counts the minimum number of instances that need to be in a potential leaf of a branch for it to split. The default value for this is 2. Given that we ended up deciding a severity set with 3 different possibilities, this would not guarantee that a branch has a majority of one class as opposed to a 2 or 3-way split in a branch. We decided to increase the minNumObj to 4 to avoid 3-way ties at the leaf nodes.

CHAPTER 4 EVALUTION

In this chapter, we will evaluate the various methods that we outlined previously.

4.1 SevGroup Evaluation

Due to the large amount of data and general similarity between them, we have only presented a few of the tables here. The additional tables can be found in Appendix D. One important thing to state is how well crashes without injuries are predicted in comparison to the other categories. This is because the majority of the crashes, around 70% for most Age Groups and 80% for the Unknown Age Group, did not involve an injury. As a result, these results are predicted much better than the others. One thing to note about these results is that the values 1 through 5 generally mean different things across each set, so the best comparison of values come in the Average and Weighted Average.

Table 4.1: Severity Group Evaluation for Ages 21-64, J48, no Cost Matrix

	All Severities	Fatal, Injury, None	Major, Minor, None
1	0.317	0.339	0.393
2	0.222	0.647	0.561
3	0.293	0.877	0.875
4	0.330		
5	0.883		
Weighted	0.718	0.809	0.779
Unweighted	0.409	0.621	0.610

Table 4.2: Severity Group Evaluation for Ages >= 75, J48, no Cost Matrix

	All Severities	Fatal, Injury, None	Major, Minor, None
1	0.149	0.192	0.283
2	0.183	0.667	0.565
3	0.303	0.871	0.866
4	0.344		
5	0.874		
Weighted	0.265	0.802	0.760
Unweighted	0.371	0.577	0.572

Looking at these, there is a definite improvement in reducing the number of variables from 5 to 3. However, it is less clear which compression of 3 variables yields the better result. To solve this, we will examine the weighted and Averages below.

Table 4.3: Severity Group Evaluation, Weighted Average

	All Severities	Fatal, Injury, None	Major, Minor, None
A (<=15)	0.35	0.834	0.78
B (16-20)	0.694	0.791	0.761
C (21-64)	0.718	0.809	0.779
D (65-74)	0.708	0.8	0.774
E (>=75)	0.265	0.802	0.76
U (?)	0.938	0.99	0.705
Average	0.612	0.838	0.76

As shown using the above chart, reducing the number of options for the class variable improved FMeasure. As well, while SevGroup2 usually did increase Group 1's FMeasure, this was often at the cost of overall FMeasure. As a result, we decided to use SevGroup1 as our class variable going forward. Around this time we also decided to merge some Crash Level variables that had initially been split, such as the MajorCause X variables, back into a single value

4.2 Comparison of Values

From here, we will attempt evaluate each set of values on each of the 6 sets along with all the values to see whether the variable set generated by a given age group best predicts its own age group.

Table 4.4: minNumObj 5, Cost Matrix 789, Diff Values, A (<=15)

	All	A (<=15)	B (16-20)	C (21-64)	D (65-74)	E (>=75)	U (?)
Fatal	0.5	0.571	0.333	0.571	0.444	0.400	0.000
Injury	0.76	0.716	0.716	0.747	0.734	0.760	0.444
None	0.896	0.893	0.888	0.892	0.892	0.893	0.870
Weighted	0.846	0.828	0.824	0.839	0.834	0.843	0.742
Unweighted	0.719	0.727	0.646	0.737	0.690	0.685	0.438

As we can see, we get a much better picture by comparing the FMeasures with all the values as we can easily tell which of the variable sets performed the best. Similar to the original set, the only values that improved FMeasure were those of its own and the drivers aged 21-64, with the latter performing the best overall.

Table 4.5: minNumObj 5, Cost Matrix 789, Diff Values, B (16-20)

	All	A (<=15)	B (16-20)	C (21-64)	D (65-74)	E (>=75)	U (?)
Fatal	0.289	0.226	0.242	0.286	0.314	0.314	0.000
Injury	0.63	0.491	0.592	0.633	0.632	0.632	0.081
None	0.878	0.866	0.877	0.878	0.878	0.878	0.821
Weighted	0.799	0.748	0.787	0.800	0.800	0.800	0.589
Unweighted	0.599	0.528	0.571	0.599	0.608	0.608	0.301

Here we seem to see a different pattern, however. While the values from drivers aged 2164 performs about the same as the original and its own values perform worse, we get a better result for the values from the older 2 age groups.

Table 4.6: minNumObj 5, Cost Matrix 789, Diff Values, C (21-64)

	All	A (<=15)	B (16-20)	C (21-64)	D (65-74)	E (>=75)	U (?)
Fatal	0.358	0.160	0.408	0.382	0.357	0.371	0.000
Injury	0.637	0.449	0.604	0.641	0.615	0.634	0.113
None	0.896	0.876	0.894	0.895	0.895	0.895	0.842
Weighted	0.82	0.751	0.809	0.821	0.813	0.818	0.632
Unweighted	0.63	0.495	0.635	0.639	0.623	0.633	0.318

The Values perform on their own about as well as expected, with Group C's Values predicting itself the best. The Drivers from 16to 20 and 75 and older also improved our results.

Table 4.7: minNumObj 5, Cost Matrix 789, Diff Values, D (65-74)

	All	A (<=15)	B (16-20)	C (21-64)	D (65-74)	E (>=75)	U (?)
Fatal	0.327	0.125	0.340	0.275	0.370	0.292	0.000
Injury	0.614	0.444	0.608	0.631	0.620	0.619	0.090
None	0.888	0.870	0.888	0.890	0.888	0.887	0.834
Weighted	0.805	0.742	0.804	0.812	0.808	0.806	0.614
Unweighted	0.61	0.480	0.612	0.599	0.626	0.599	0.308

Unlike the others, this group isn't improved by the values from the group with drivers aged 21-64. It seems to have been predicted the best by its own values, with the values from drivers aged 16-20 also performing well.

Table 4.8: minNumObj 5, Cost Matrix 789, Diff Values, E (>=75)

	All	A (<=15)	B (16-20)	C (21-64)	D (65-74)	E (>=75)	U (?)
Fatal	0.298	0.136	0.204	0.259	0.091	0.213	0.000
Injury	0.654	0.464	0.627	0.656	0.638	0.669	0.082
None	0.887	0.865	0.884	0.889	0.885	0.891	0.821
Weighted	0.81	0.736	0.799	0.812	0.802	0.816	0.589
Unweighted	0.613	0.488	0.572	0.601	0.538	0.591	0.301

We didn't seem to have much luck with this group on reducing values. Every group that we attempted this for performed worse than the original. However, Age Group 21-64's values performed the best of each of the values.

Table 4.9: minNumObj 5, Cost Matrix 789, Diff Values, U (?)

	All	A (<=15)	B (16-20)	C (21-64)	D (65-74)	E (>=75)	U (?)
Fatal	0	0.000	0.000	0.000	0.000	0.000	0.000
Injury	0.972	0.642	0.972	0.975	0.928	0.914	0.879
None	0.998	0.967	0.997	0.997	0.992	0.991	0.988
Weighted	0.993	0.928	0.992	0.993	0.982	0.980	0.973
Unweighted	0.657	0.536	0.656	0.657	0.640	0.635	0.622

For this one, none of the values improved on the FMeasure, but the Values from Age Group 21-64 performed as well as the full values.

Altogether, we can determine that the largest age group's values performed the best on most of the sets. It does this for itself, drivers aged 15 and younger, and drivers of unknown ages. Interestingly, though, the values from drivers aged 65-74 also performed well on some groups, its own and drivers aged 16-20, indicating that the best group of values isn't necessarily the largest encompassing one. The drivers aged 75 and older were a bit strange compared to all this, however, only losing precision from the decrease in values, even with their own set.

4.3 Variable Set Comparison

Now we will compare the variables generated by the trees calculated by integrating all of our methods. These variables will likewise have more clear names than before. One thing of note, however, is that this variable set will not include the variables generated from the largest data set, C. This is because it covers almost all the variables and there are many values that apply only to it, making it difficult to say anything about the other classes compared to it. The full table, with the values generated by C as well, can be found in Appendix E with the other additional tables.

Figure 4.1: Variable Set without C (21-64), minNumObj 5, Cost Matrix 789

	A (<=15)	B (16-20)	D (65-74)	E (>=75)	U (?)
TOccupants			X		
AvgPassCount				X	
Drug		X	X	X	
Alcohol		X		X	
DAOther		X			
DANone		X			
NonColl			X		
Rural			X		
PavedRoad		X			
UnpavedRoad					X
WeatherClear				X	
WeatherRain				X	
WeatherBlowPrtcls			X		
PassAge18To70		X			
BSeat	X	X	X	X	
SeatDriver		X		X	
SeatFrontMid		X			
SeatFrontPass	X	X	X	X	
NumPedestrians	X	X	X	X	X
NumPedacyclists		X	X	X	X
OccProcNone		X	X	X	
OccProcShoulderLapBelt				X	
OccProcShoulderBelt		X			
OccProcHelmet	X	X	X	X	
EjectNot		X	X		
EjectPart		X			
EjectFull	X	X	X	X	X
EjectOth			X	X	
EjectPathNone		X			
EjectPathSide		X	X	X	

Figure 4.1 cont: Variable Set without C (21-64), minNumObj 5, Cost Matrix 789

	A (<=15)	B (16-20)	D (65-74)	E (>=75)	U (?)
AirbagDeployFront			X	X	
AirbagDeployFrontSide			X		
TrapNonMech		X	X	X	
TrapMech	X	X	X	X	
PassengerCar				X	
FourTireLightTruck		X		X	
Van/MiniVan				X	
SUV				X	
Tractor/SemiTrailer				X	
Motorcycle		X	X	X	
Moped/ATV	X		X	X	
FarmVehicle/Equip			X		
EmerVehNA			X		
EmerVehOth		X			
CargoNA		X		X	
TrailerCargoOth				X	
NoDefects				X	
DefectOther		X			
SpeedLess25		X			
Speed30To40			X	X	
NoTrafficControls			X		
TrafContNoPassZone		X			
InitImpactPassMid		X		X	
InitImpactDriverRear			X		
InitImpactDriverMid	X			X	
MostDamageFront				X	
MostDamagePassMid				X	
MostDamageDriverRear			X		
MostDamageDriverFront			X		
DamageMinor				X	

Figure 4.1 cont: Variable Set without C (21-64), minNumObj 5, Cost Matrix 789

	A (<=15)	B (16-20)	D (65-74)	E (>=75)	U (?)
DamageDisabling		X	X	X	
DamageTotaled		X	X	X	
MinAge		X			
MaxAge	X	X	X		
DAge16To20		X			
DriverMale		X		X	
DCCSpeeding		X			
DCCCenterline		X			
DCCFTYROWStopSign			X		
DCCOth		X			
VisionNotObscured		X			
DriverApparentlyNormal				X	
DriverDUI		X			
OtherCond		X	X	X	

Without the largest age group, we can get a better picture of the others. It seems like there are a decent mix of variables that seem important to all of the groups as well as some that only seem significant to certain groups. For ease of reading, we will outline the overlap or lack thereof of variables below.

Figure 4.2: Variables in 3 or More Groups

All 5	All But U (?)	All But A (<=15)	A, B, D (<=20, 65-74)	A, D, E (<=15, >=65)	B, D, E (16-20,>=65)
NumPedestrians	BSeat	NumPedacyclists	MaxAge	Moped/ATV	Drug
EjectFull	SeatFrontPass				OccProcNone
	OccProcHelmet				EjectPathSide
	TrapMech				TrapNonMech
					Motorcycle
					DamageDisabling
					DamageTotaled
					OtherCond

From these variables, we can tell that the number of pedestrians involved in a crash as well as whether someone is fully ejected from a car seem to be important in determining the severity of a crash across the board. These definitely make sense as pedestrians don't have the protectiveness of a car to help them. Drugs, Seating, Occupant Protection, the degree of damage, and presence of vehicles like mopeds and Motorcycles also seem to be of general importance in several categories.

Figure 4.3: Variables in 2 Groups

A & E (<=15,>=75)	B & D (16-20, 65-74)	B & E (16-20, >=75)	D & E (>=65)
InitImpactDriverMid	EjectNot	Alcohol	EjectOth
		SeatDriver	AirbagDeployFront
		FourTireLightTruck	Speed30To40
		CargoNA	
		InitImpactPassMid	
		DriverMale	

There are fewer variables common among only two of the groups, but they do tell us important info. From this it seems that the oldest and youngest drivers seem to both be impacted by alcohol, hits to the side of the vehicle, and the driver being male. The older drivers also seem to be impacted by whether airbags deploy and speed limits between 30 and 40 MPH.

Figure 4.4: Variables in 1 Group

B (16-20) Only D (65-74) Only E (>=75) Only U (?) Only

DAOther	TOccupants	AvgPassCount	UnpavedRoad
DANone	NonColl	WeatherClear	
PavedRoad	Rural	WeatherRain	
PassAge18To70	WeatherBlowPrtcls	OccProcShoulderLapBelt	
SeatFrontMid	AirbagDeployFrontSide	PassengerCar	
OccProcShoulderBelt	FarmVehicle/Equip	Van/MiniVan	
EjectPart	EmerVehNA	SUV	
EjectPathNone	NoTrafficControls	Tractor/SemiTrailer	
EmerVehOth	InitImpactDriverRear	TrailerCargoOth	
DefectOther	MostDamageDriverRear	NoDefects	
SpeedLess25	MostDamageDriverFront	MostDamageFront	
TrafContNoPassZone	DCCFTYROWStopSign	MostDamagePassMid	
MinAge		DamageMinor	
DAge16To20		DriverApparentlyNormal	
DCCSpeeding			
DCCCenterline			
DCCOth			
VisionNotObscured			
DriverDUI			

Lastly, we have the variables only found in one of the groups, which seem to be somewhat differentiated. One thing that seems to affect all groups, however, is the number of passengers, as PassAge18To70, TOccupants, and AvgPassCount all seem to indicate this. There might be some factor differentiating these that might lead them to have been selected differently, however.

For the drivers aged 16-20, their factors seem to be focused on urban driving, drugs and alcohol, and traffic violations, which seem to particularly involve speeding, crossing the centerline, and no passing zones. Wearing the shoulder belt seems to be another important factor, as does driving in areas with a speed limit less than 25.

Drivers aged 65-74 seem to have a more rural focus, with Rural appearing for them, as well as Farm Equipment being involved in crashes, a lack of traffic controls, as well as failing to yield right of way at a stop sign.

Drivers aged 75 and up seem to be more affected by the weather, the type of vehicles involved in the crash, with Passenger Car, Van, SUV, and Tractor/Semi Trailer being listed. Shoulder and lap belt seems to be important as well, along with where the car is most damaged and whether the driver is in normal condition.

Some limitations of this method are due to the variables that were selected. As some of the variables were more directly tied to severity rather than factors occurring beforehand, such as the MostDamageX set of variables. For future work, these variables should be removed for future analyses.

The only factor unique to drivers of unknown ages seems to be the Unpaved Road. This could be due to something like accidents on unpaved roads being more likely to not check the ages of drivers or something like this.

CHAPTER 5 CONCLUSION

Looking at our results overall, it seems that our choice to analyze the groups by age was overall a positive one. We were able to predict most of the groups better with our variable selection and the classes we were unable to do this for can still gain benefits from being analyzed separately from the others.

Group A, featuring drivers aged below 16, was the smallest group, and predictably so, as these drivers are legally only able to drive with a permit in the state of Iowa. It was interesting how we were able to predict this class fairly well with only around 10 variables.

Group U was an interesting class as well, featuring only incidents where none of the drivers ages are known. We were able to reduce it to only 4 variables, though we are unsure how well this succeeds when it is unable to predict any of the fatalities among the class and results in an overall low percentage. That said, none of the classifiers we attempted were able to predict the fatalities among Group U, so it is fairly functional.

Group E, containing drivers older than 75 was another interesting class. It had around as many variables as B and D, being most similar in composition to D, though, like U, any attempts to classify it based on fewer variables just hindered the overall results. Perhaps there is an element of unpredictability to this class we could not properly analyze with fewer variables, thought we are unable to determine this effectively in the current study.

Group C was the largest Age Group overall and also had the greatest number of variables associated with it. This large number of variables unfortunately made us unable to properly contrast the different classes as almost every variable was in one class was also in C. Whether this is due to the size of C or there are potential subdivisions in C that could be classified more similarly to B and D would likewise be an interesting direction of further study. Regardless,

however, removing it from consideration in the final section did greatly help yield clearer trends in the data.

Future Work

As we were documenting our results, we discovered that the Unknown Age Group predicted crashes without injuries rather well, even if it did not predict any fatalities correctly. We reasoned from there it might be possible to combine it with another tree that was generated in order to improve classification for all three classes. It seems like the drivers aged 15 and younger performed the best for prediction of the injury classes, so this seems like a good candidate for combination.

As well, while our data leads us in a promising direction and shows that this might be a good area for future study, more information will have to be examined from it in order to have a breakthrough. Integration with other data sources such as traffic volume, population characteristics, or more specific weather conditions would be good future directions to take this research in. Analyses involving other segregating attributes such as rural and urban crashes would also be a good direction, as it is unclear whether the propensity for younger drivers to be involved in more urban crashes and older drivers to be involved in more rural is solely a matter of demographics or not.

BIBLIOGRAPHY

1. U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine 17(3): 37-54 (1996)
2. M. A. Hall (1998). Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand.
3. M. A. Hall (n.d.). GainRatioAttributeEval. Retrieved June 02, 2016, from http://weka.sourceforge.net/doc.dev/weka/attributeSelection/GainRatioAttributeEval.html
4. Frank, E. (n.d.). ChiSquaredAttributeEval. Retrieved June 02, 2016, from http://weka.sourceforge.net/doc.dev/weka/attributeSelection/ChiSquaredAttributeEval.ht ml
5. Ray, A. (2006, August 22). Understanding Chi Square. Retrieved June 02, 2016, from http://practicalsurveys.com/reporting/chisquare.php
6. http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/J48.html
7. http://weka.sourceforge.net/doc.dev/weka/classifiers/bayes/NaiveBayes.html
8. By Jason Brownlee on March 5, 2014 in Machine Learning Process. "A Simple Intuition for Overfitting, or Why Testing on Training Data Is a Bad Idea - Machine Learning Mastery." *Machine Learning Mastery*. N.p., 2016. Web. 08 Nov. 2016.
9. Baratloo, Alireza, Mostafa Hosseini, Ahmed Negida, and Gehad El Ashal. "Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity." *Emergency*. Shahid Beheshti University of Medical Sciences, 2015. Web. 12 Nov. 2016.
10. "Supervised and Unsupervised Machine Learning Algorithms - Machine Learning Mastery." *Machine Learning Mastery*. N.p., 2016. Web. 12 Nov. 2016.
11. "J48 Decision Tree - Mining at UOC." *Mining at UOC*. N.p., n.d. Web. 17 Nov. 2016.
12. "Weka 3: Data Mining Software in Java." *Weka 3 - Data Mining with Open Source Machine Learning Software in Java*. N.p., n.d. Web. 21 Nov. 2016.
13. Curry, Allison E., Jessica Hafetz, Michael J. Kallan, Flaura K. Winston, and Dennis R. Durbin. "Prevalence of Teen Driver Errors Leading to Serious Motor Vehicle Crashes." *Accident Analysis & Prevention* 43.4 (2011): 1285-290. Web.
14. Lam, Lawrence T. "Factors Associated with Young Drivers' Car Crash Injury: Comparisons among Learner, Provisional, and Full Licensees." *National Center for Biotechnology Information*. U.S. National Library of Medicine, n.d. Web. 26 Nov. 2016.
15. K. Haleem and A. Gan, "Identifying traditional and nontraditional predictors of crash injury severity on major urban roadways," *Traffic Inj Prev.*, vol. 12, no. 3, pp. 223-234, 2011.
16. D. Bose et al., "Increased risk of driver fatality due to unrestrained rear-seat passengers in severe frontal crashes," *Accident Analysis and Prevention*, vol. 53, pp. 100-104, 2013.
17. J. Brady and G. Li, "Trends in Alcohol and Other Drugs Detected in Fatally Injured Drivers in the United States, 1999–2010," *American Journal of Epidemiology,* vol. 179, no. 6, pp. 692-699, 2014.