Cell Phone Distraction: Data Mining Application on Fatality Analysis Reporting System (FARS)

Anas Alrejjal

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Civil Engineering

Eastern Mediterranean University
June 2018
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Assoc. Prof. Dr. Ali Hakan Ulusoy
Acting Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science in Civil Engineering.

Assoc. Prof. Dr. Serhan Şensoy
Chair, Department of Civil Engineering

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Master of Science in Civil Engineering.

Asst. Prof. Dr. Mehmet Metin Kunt
Supervisor

Examining Committee

1. Asst. Prof. Dr. Şevket Can Bostancı

2. Asst. Prof. Dr. Mehmet Metin Kunt

3. Asst. Prof. Dr. Eriş Uygar

ABSTRACT

Distracted driving has become one of the ubiquitous concerns in terms of traffic safety as the presence of portable technology and its emergence while driving considerably increase. The objective of this thesis was to investigate the relationship between this prevalent distraction and motor vehicle accidents and how cell distraction influences driver performance across the United States from years 2011 to 2015 by one of the most reliable databases the Fatality Analysis Reporting System (FARS). Applying data mining techniques was used to discover and explore the role of distracted drivers in fatal crashes. Classification algorithms utilized which are C5.0, C4.5 and Ctree decision trees in determining the most related variables for the manner of collision and predicting the vehicle collision patterns occurred in fatal accidents. In addition, study the most contributed attributes related to the most harmful event that happens due to cell phone distraction, classifying and predicting the most harmful consequences in fatal crashes for those distracted drivers.

The results show that the driver-related factors' contribution continuously increased in the five-year period to determine the manner of collision and the most harmful event regardless of the gender. The dangerous use of cell phone while driving was demonstrated at intersections and the contribution of the intersection to increase the likelihood to get involved in angle crash was illustrated. Additionally, most of the crashes for the distracted driver are not with a motor vehicle in motion due to the inability of maintaining lanes or improper lane change during driving task because the driver's eyes and mind during the use of cell phone are off the road for extended periods of time. The results demonstrate that cell phone distraction does not just have

dangerous impacts on drivers' lives or vehicles but on pedestrians as well. This thesis sheds light on a new cause of overcorrection and subsequent rollover.

For a better understanding of the issue of cell phone use while driving, in order to create more precise data with respect to drivers' awareness to the cell phone use risk, it is essential to set up the extent of drivers' cell phone use more precisely. In order to truly assess the share of cell phone distraction crashes in the total number of crashes, cell phone use ought to be recorded accurately in accident reports.

Keywords: distracted driving, data mining techniques, driver-related factors, decision trees, overcorrection.

ÖZ

Gunumuzde dikkatsiz sürüş taşınabilir teknolojinin varlığı ve sürüş sırasında ortaya çıkması nedeniyle yaygın olmustur. Bu tezin amacı, bu yaygın distraksiyon ve motorlu taşıt kazaları arasındaki ilişkiyi incelemek ve hücre distraksiyonunun Amerika Birleşik Devletleri genelinde 2011'den 2015'e kadarki en güvenilir veritabanlarından biri olan Fatality Analysis Reporting System (FARS) ile sürücü performansını nasıl etkilediğini araştırmaktır. Veri madenciliği teknikleri ölümcül kazalardaki dikkatsiz suruculerin rolunu anlamak icin kullanildi. sınıflandırma algoritmaları çarpışma şekli için en ilgili değişkenlerin belirlenmesi ve ölümcül kazalarda meydana gelen araç çarpışma paternlerinin tahmin edilmesi icin kullanildi. Ayrıca, en zararlı olaylarla ilgili en önemli özelliklerin incelenmesi, cep telefonu rahatsızlığının ölümcül çökmelerdeki en zararlı olaylara nasıl katkıda bulunduğunu tahmin etmek ve sınıflandırmaktır. Sonuç olarak, sürücüye ilişkin faktörlerin, beş yıllık dönemde, cinsiyete bakılmaksızın, çarpışma şeklini ve en zararlı olayı belirlemek için sürekli olarak arttığını göstermektedir. Sürüş sırasında cep telefonunun kullanımı, tehlikeli kesişme noktaların'da gösterilmiş ve kavşağın açısının, kazanın karışma olasılığını arttırmaya olan katkısı gösterilmiştir.

Ek olarak, dikkati dağılmış sürücünün çöküşlerinin çoğu hareket halinde olan bir motorlu araçla değil, çünkü , cep telefonu kullanımı sırasında sürücünün gözleri ve aklı yolun dışındadır. Sonuç olarak, cep telefonu rahatsızlığının sadece sürücülerin hayatları veya araçları üzerinde değil, yayalarda da tehlikeli olduğunu göstermektedir. Bu tez, yeni bir aşırı düzeltme ve müteakip devrilmeye ışık tutmaktadır.

Sürüş sırasında cep telefonu kullanımının zararlarının daha iyi anlaşılması için, sürücülerin cep telefonu riskine karşı farkındalıklara daha kesin veriler oluşturmak ve sürücülerin cep telefonu kullanımının kapsamını daha kesin bir şekilde belirlemek önemlidir. Cep telefonu kullanımının, kazalardaki oranını ve cep telefonu risklerini gerçekten değerlendirmek için kaza raporlarının doğru bir şekilde kaydedilmesi gerekir.

Anahtar Kelimeler: dikkat dağıtıcı sürüş, Veri madenciliği teknikleri, sürücü ile ilgili faktörler, Karar ağaçları, aşırı düzeltme.

DEDICATION

To my parents

ACKNOWLEDGMENT

I would like to express my gratitude to my supervisor Asst. Prof. Dr. Mehmet M. Kunt, whose sympathetic, proficiency and patience, added prominently to my knowledge. I appreciate all his skills and efforts in various areas and his support and assistance in my methodology. All my efforts could have been short-sighted without his valuable recommendation and supervision.

I would like to declare my sincere appreciation to my family for supporting me through my whole life and in particular. Their inspiration and encouragement are limitless to state.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BAC	Blood Alcohol Content
CDS	Crashworthiness Data System
DM	Data Mining
DOT	Department of transportation
FARS	Fatality Analysis Reporting System
GES	General Estimating System
NASS	National Automobile Sampling System
NHTSA	National Highway Traffic Safety Administration
OPIGP	Occupant Protection Incentive Grant Program
PAR	Police Accident Report
SHSOs	State Highway Safety Offices

Chapter 1

INTRODUCTION

In spite of safety advancements in roads and design of vehicles, the total number of fatal crashes still rises. The expanding number of fatalities illustrates that the safety of driving embodies a persistent and vital issue in the United States. Decreasing crash involvement would advantage millions of individuals across the world. Despite the fact that most motor-vehicle crashes are credited to multiple causes [1]. Based on thinking of drivers in a situation in the traffic way will go and what other drivers will do, they must persistently make decisions. In a matter of seconds a traffic situation can end up life-threatening, therefore drivers must subsequently focus their attention on traffic persistently.

The same goes when the individuals on the road frequently tend to do things that unrelated to driving, like using a cellphone for talking, checking apps or taking photos, watching navigation screen while driving and manipulating with go-pro cameras that casting while driving. So much attention may be needed in performing an additional task that the driving performance decreases and dangerous circumstances happen. In such case, the term of distraction is utilized.

Distraction is an impairment that has gotten progressively more relative especially with the technology inside the vehicle presentation like systems of navigation, display

screens, video blogging and radio and audio controls and expanding attention has been drawn by policymakers and human figure researchers in the transportation safety field.

Distraction of driver diverts the attention of driver from the activities basic for driving in safe [2] which contributes about 13-50% of all crashes, resulting $40 billion in damages and 10,000 fatalities each year [1]. Distracted driving can be is categorized into four main forms. Firstly visual-distractions that means the eyes of the driver are taken off the trafficway, secondly auditor-distractions that the aural perception of the driver from the relative tasks of driving is taken, thirdly cognitive distractions which the mind of the driver is taken off the driving task (operating a mobile keypad to enter number or text) and lastly manual distractions that the hands of the driver are taken off the wheel [3].

The first two forms of distraction delay the driver from getting fundamental information for the driving task, the third influences the preparation of this approaching information and delays the driver to take corrective action required by the driving situation [4]. When the drivers cannot give adequate concentration to maintain the performance of driving they can be classified as distracted drivers by a non-driving action. There are two situations when they occur the distraction happens, the first one when the non-driving action is so complicated that the driving task needs as much attention to engage in activities that are non-driving related or the driver cannot give adequate concentration to the driving tasks [5]. Regardless of the distraction cause, the impairment that resulting is hazardous since the capacity of the driver to focus and keep up a safe driving performance and travel way are limited. The National Highway Traffic Safety Administration (NHTSA) has decided that the driver distraction as a fundamental causation for a majority of accidents [6].

1.1 Risks of Distracted Driving

Distracted driving is a widespread concern resulting in 3,477 fatalities and an estimated 391,000 injuries in crashes that involve drivers with distraction in 2015 in the United States alone. Of these fatalities, 476 were in crashes attributed to cell phone use or other cell-phone-affected activities as distractions while driving. There were 442 fatal crashes detailed to have involved cell phone usage distraction which represent 14% of all fatal distraction-related crashes. On behalf of these crashes that the distraction is the main cause, the crash report of police expressed that the driver was talking conversations on a cell phone, listening to a call, or otherwise cell phone manipulation at the time of the crash. At any given moment in 2014 during daylight hours, the number of drivers who is distracted by using a cell phone while driving is more than 587,000 vehicles which were reported by The Department of Transportation (DOT) in the United States that. Driver inattention was recognized to be a causal reason in 78% of motor vehicle crashes and 65% of near crashes by a previous investigations [7, 8].

Cell phone use has been a growing interest in the past few years, because the technological advancements have come with risk to traffic safety. People have more access to cell phones than they can conceivably handle. One place where this is apparent is distracted driving, since 72 percent of grown-up cell phone users 18 years or more in the United States admit to use a cell phone while driving. Of these grown-ups, 25% admitted sending or receiving text messages while driving [9].

Smartphone use is very widespread. There are 261.9 million smartphones in use in the US nowadays. These smartphones are using 102 times more data than a current basic

mobile gadget [10]. On average last year, each month 3.87 GB of data were used by a smartphone. Since 2010 this translates to an increment of 1,400 %, because of the escalation of networks, more advanced phones, and modern services and applications. In the mid-1980s, mobile phones were presented to the market in the United States. By 1986, there were 681,825 wireless-subscribers in U.S. The number of American-subscribers since that time has increased significantly to 395,881,497 and the total annual minutes, messages and megabits of wireless activity has come to 13,719 Billion by the end of December of 2016 compared to 388 Billion in 2010. Figure 1 shows the percentage of U.S. adults who own cellphones and smartphones which illustrates a sharp increase in the usage of both [11]. A growing number of studies of simulated driving tasks and investigative counting naturalistic studies provide proof that the behavior of driver is influenced by utilization of cell phone, for instance, slowing response time of drivers or taking away the eyes of drivers from the roadway more often [12, 13]. Figure 1 shows the increment in owning cell phones and smart phones for U.S. adults which indicates the widespread using smart phones that most of Americans people have cell phones that increases the possibility of cell phone use while driving.

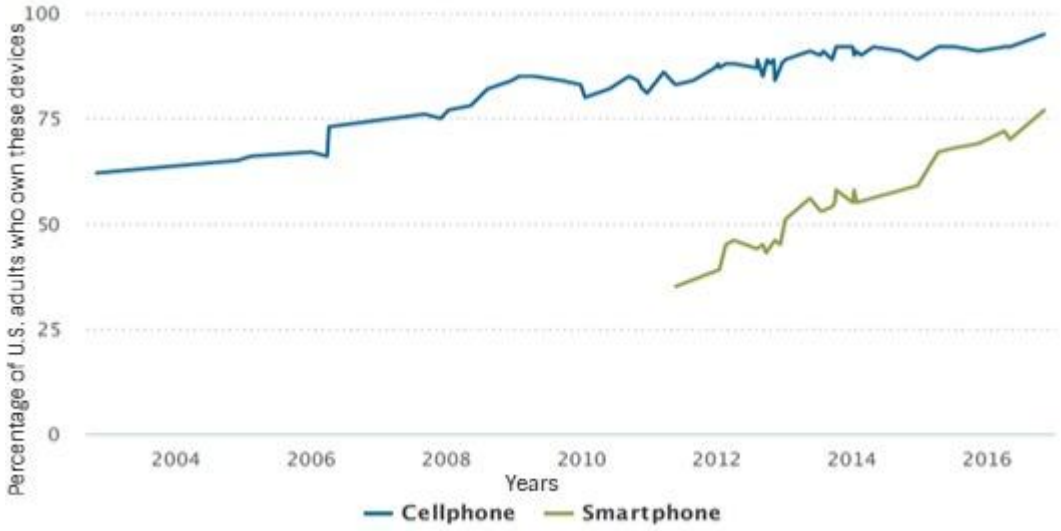


Figure 1: Adults users of cellphone and smartphone[11]

Data are not so precise for the number of drivers who using their cell phone while driving, even though data about the precise number of subscribers who using cell phone do exist. For estimating these numbers, three major sources were proposed [14]: Self-reports about the cell phone use while driving, police accident records and observational studies.

Distracted driving research by NHTSA presents fatal crash data for distraction-related crashes by the age of driver for the year 2015. There were 290 out of 3,183 which represent 9% of all drivers 15 to 19 years old involved in deadly crashes were distracted at the time of the crash. The largest percentage of distracted drivers among all age groups was this age group and the group of drivers under 30 was an overrepresentation in distraction-related fatal crashes. Also, 14% of all the cell phone-distracted drivers were 15 to 19 years old in other words they were 64 of the 456 distracted drivers by using cell-phone in fatal crashes. Correspondingly, drivers from 20-29 years old make up 24% of drivers in all fatal crashes but represent 27% of distracted drivers and 33% of cell phone-distracted drivers in fatal crashes [14].

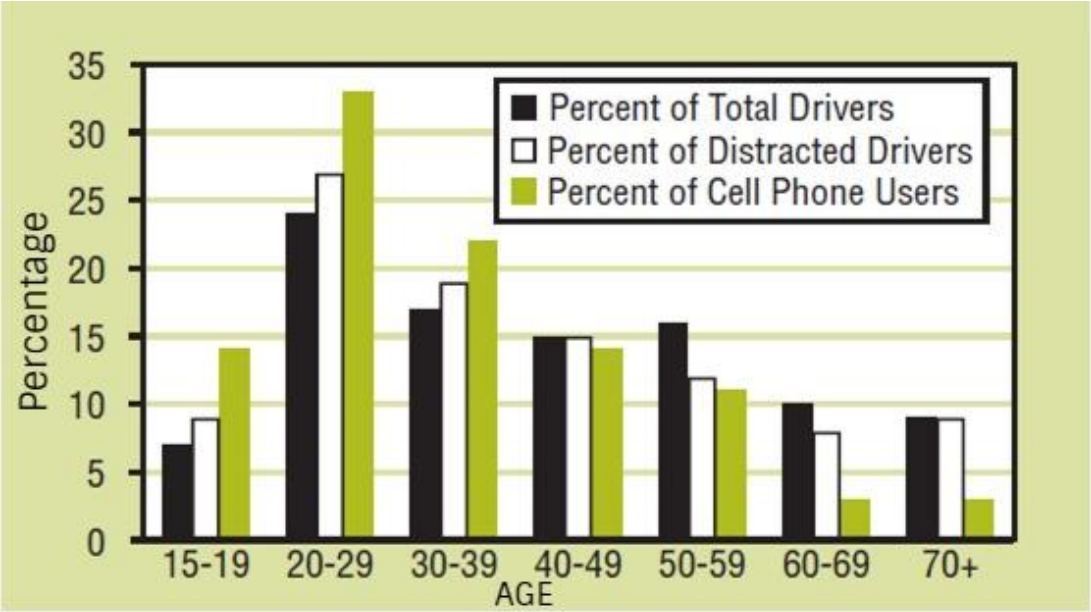


Figure 2: Percent distribution by age, distraction and cell phone use of driviers involved in fatal crashes[14]

1.2 Distraction-Affected Crash Data Collection

In the context of crash data in United States, there are three main sources: NHTSA's Fatality Analysis Reporting System (FARS), National Automobile Sampling System (NASS) and General Estimating System (GES) [14]. The Fatality Analysis Reporting System (FARS) provides data on U.S. road fatalities. FARS database records crashes that result fatalities within 30 days of the crash and the crashes involved motor vehicles and take place on an open trafficway. FARS gives data on imperturbability testing and components contributing to each crash as chosen by crash examiners since the start of use of mobile phones in motor vehicles in 1991 [14].

In relation to distraction, NHTSA proceeds to refine the collection of data about distracted driving in crashes that with police report, this incorporates enhancements to the distraction coding in FARS database. Prior to 2010, fatal vehicle crashes which are covered by FARS and data about a sample of all severities of crashes with police report which are coved by the NASS and GES, coded distraction data in distinctive formats.

About inattentive behavior, FARS was more commonly use and more comprehensive, while particular distracted-driving behaviors were distinguished by GES.

In 2010, when FARS adopted the GES framework, the two strategies of systems for coding distraction were combined. Starting in 2010 for both frameworks, when looking at distraction-related crashes, in both FARS and GES the driver is distinguished as "Yes-Distracted," "No-Not Distracted," or "Unknown in case Distracted". In the event that the driver is recognized as distracted, in order to recognize the particular action that was distracting the driver while driving, advance coding is performed. This was changed in FARS but was not for the GES data coding. On the Police Accident Report (PAR), the information collected was not modified. GES can be compared over the years, since the information was not alter in this frame work.

Of extra note is the phrasing with respect to distraction. For FARS and GES information, starting with the framework of 2010, the crash is referred to as a distraction-related crash if a crash in which a driver was recognized as distracted at the time of crash. Cell phones use is also more particularly noted beginning with the 2010 framework. With the current coding, the usage of a cell phone is more defined and in case the particular association cannot be decided. There are restrictions on collecting and detailing FARS and GES data that were recognized by NHTSA with respect to distracted driver. The FARS and GES data are based on PARs and after the crashes have happened, the data are made [14].

Non-pedestrian and non-cyclist included crashes are in general categorized into singular, angular, sideswipe, rear-end, rear-rear, and head-on crashes. FARS is the data

source for this thesis, also implements this classification. More details on this dataset are delivered in Chapter 3.

1.3 State Laws on Distracted Driving

Numerous States confine cell phone utilized by drivers. The enforcement categorizes in two forms: primary and secondary enforcement As of March 2016, no State totally prohibited all shapes of cell phone utilized by drivers:

Hand-held Cell Phone Use: 15 states, D.C., Guam, the U.S. Virgin Islands and Puerto Rico, forbid using hand-held cell phones for all drivers while driving. All bans are primary enforcement laws this means without any other traffic offense proceeding, a driver is cited by an officer for using a hand-held cell phone.

All Cell Phone Use: No state forbids all use of cell phone for all drivers, but for novice drivers, 38 states and D.C. ban all cell phone use while driving for them, and for school bus drivers, 20 states and D.C. forbid it.

Text Messaging: In 2007, the first state was Washington to pass a texting ban. Presently, 47 states, D.C., Guam, Puerto Rico and the U.S. Virgin Islands forbid text messaging for all drivers while driving. Primary enforcement is implemented by all states but 4 of them do not. Novice drivers are forbidden of text messaging by two of the three states.

Collection of Crash Data: Except 2 states, crash report forms of police contain at least one category for distraction on all others states, though the definite data collected diverges. The best practices are offered on collecting information of distraction by The Model Minimum Uniform Crash Criteria (MMUCC) guide.

Preemption Laws: There are preemption laws in some states that the local jurisdictions are forbidden from passing their own bans of driving by distracted drivers. States with such laws comprise Oregon, Mississippi, Florida, Pennsylvania, Iowa, South Carolina, Kentucky, Nevada, Louisiana and Oklahoma.

1.4 Data Mining Approach

Data mining is drawing meaningful hidden patterns from a huge database. Using roadway, driver and vehicle characteristics is valuable in the investigation of fatalities in traffic safety. Data mining is a valuable tool to specify the need for filtering important data like patterns that are hidden from the enormous databases currently present [15].

Data mining of traffic crash data is vital for understanding why traffic crashes occurred frequently in certain vehicle conditions, driving and environment. There are numerous reasons that lead to be involved in a crash, and the relationships between them are complex, it is exceptionally difficult to construct a model with correct assessment. In order to overcome this issue, several statistical models have been broadly utilized such as decision tree, random forest, neural network and fuzzy logic on such crash data to explore the patterns of road crashes.

Chapter 2

LITERATURE REVIEW AND OBJECTIVES

2.1 Distracted Driving

In 2015, 3,196 deadly crashes happened on U.S. roadways that distraction was involved and that represent 10% of every deadly crash. 3,263 distracted drivers was included in these accidents, as a few accidents included more than one driver who was distracted. Distraction was accounted for 7% of the drivers engaged in fatal accidents which is 3,263 out of 48,613. In these crashes that affected by distraction, 10% of general fatalities which is 3,477 fatalities happened by distraction [16].

By drawing on the concept of distracted driving, Young et al. [5] and Regan et al. [17] show driver distraction happens when the attention of a driver is diverted away from the operation of driving by an object or an occurrence to the degree that the driver is not any more capable to achieve the driving operation sufficiently or carefully and is a particular type of driver inattention.

There is expanding proof that driver inattention and driver distraction are main contributing components in car and truck accidents and occurrences [18] and the issue will increase as more technologies discover their way into vehicles. In response, several studies investigating driver distraction have been achieved on an explosion in research on these topics, culminating lately in the publication [19].

Several studies [20, 21] have revealed that there are three main types of distraction were distinguished in driver distraction, cognitive distraction, visual distraction and manual distraction: Visual diversion happens when the distracted driver focuses for an extended period of time on another visual object and ignores to see at the road. Cognitive distraction incorporates all considerations that deviate the attention of the driver. This can avoid the driver from being able to explore safely through the road arrange and response time may be reduced.

Manual distraction happens when a distracted driver physically manipulate an object by removing one or both hands from the steering wheel which decreases focusing on the essential task which is driving safely. Impacts of manual distraction incorporate directing in the off-base heading or not changing gears. One of the major sources of manual distraction is sending a text massage [5].

Cognitive distraction influences more visual filtering conduct whilst visual distraction has more impact on measures of lateral control. Some other interesting considerations come from a previous study by authors [22], who stresses that, among the three types of distraction, only visual and manual distraction can be reduced partially but cannot remove totally.

In general, in terms of driver attention, driver crash risk and driver behavior driver distraction measures the effect on them. Notably, the particular measures utilized essentially vary [23].

According to Ranney [24], 70% of rear-end or a single vehicle crashes implicated inattention and 30% of drivers due to a distracted driving, detailed having to take avoidance action in a crash [25].

Inattention Blindness

Inattention blindness is a curious phenomenon and was examined to determine the impact of this phenomenon on driving. Strayer and Drews [26] investigated this phenomenon by explaining that the distracted driver fails to see and recognize objects which are obvious their environment of driving. In order to consider the influence of this distraction on the driving performance, understanding how the capacity of cognitive of the brain handles the information that has taken while driving is needed for the researchers. They tasked many subjects and recognized object in their involvement of driving and after that trying to assess and review them. They put these subjects through some distracting scenarios such as conversations in cell phone and discussions inside the vehicle. The outcomes of this study reported that conversations by cell phone weaken the work of cognitive. On the other hand since discussions inside the vehicle had less loading on the capacities of cognitive of the brain, it did not disable driving.

Strayer and Drews [26] determined that inside vehicle discussion could be handled with driving requests rather than cell phone conversations. Moreover, single and dual task actions were distinguished by them. As the name infers, first one which is single task actions refer to drivers physically perform only one task at one time whilst when they perform two actions simultaneously while driving, the dual task happens. A single task would involve the driving task while making a call or a conversation through a cell phone while driving that would be a dual task. In dual tasks, Subjects who are

locked in this task are twice as likely to not making a recognition to the roadway signs, particularly when the drivers engage with their phones. “Inattention blindness” happens when they fail to recall what they saw which presents eventually a perspective of distraction of the driving involvement. More cognitive assets are required in dual tasks of a person which means more distraction for a driver will be when requiring more cognitive assets.

Luis Garcia- Larrea [27] conducted a study that count the response times of people attempting to observe particular targets on a screen while driving. In terms of distractions or no distractions and through different scenarios, ten subjects were set. The results show that a driver’s responsiveness is delayed by phone use. Garcia-Larrea concluded this study expressing that the attention of the driver would be diminished by cell phone conversations.

Several studies investigating cell phone distraction have been carried out, table 1 shows some of previous studies with their findings.

Table 1: Previous studies with their findings

Authors	Findings
Dingus et al. [28]	Drivers are engaging in distracting activities more than 50% of the time while they are driving
J. Stutts et al [29]	Drivers were engaged in one or more potentially distracting activities 34.5% of the total time that their vehicles were moving.
Isa et al. [30]	Majority of drivers using their mobile phones while driving (68.6%), do not use ‘‘hands-free” device.
Nelson et al.[31]	Apart from being aware that a mobile phone conversation while driving is posing a danger, drivers continued to make conversations on their mobile phones while driving, if they deemed these calls important.
Hallet et al.[32]	there is a significant difference between the average number of text messages (sms), sent by male and female drivers while driving (male drivers send more text messages (sms), on an average)
Grøndahl and Sagberg [33]	Male drivers more likely use their mobile phones than female drivers.
Bener et al. [34]	Drivers who used their mobile phones while driving have a lower seat belt wearing rate (49.9%), than drivers who did not use their mobile phones while driving (57.7%).

2.2 Crash Scenarios of Distracted Driving

Due to the later expansion relatively of distraction-affected factors in the databases of crash and in portion to the deficient consequences of these factors, these new expanded databases are attempting to better decide the relative recurrence of distraction-related crashes by a crash situation.

Single Vehicle Crash in the Road Side Scenario: In the United States it was shown that for almost 23% of crashes are off roadway crashes [35]. Najm et al. [36] examine Crashworthiness Data System (CDS) data and GES data in a research that related to crashes in U.S. They found that the major factor in 12% of CDS to 14% of GES was inattention in the case of single vehicle crashes. In this way, one of the most three components of involving in a crash in this study was distraction. Many studies [37, 38] that have examined "pre-crash situations" based upon the movements of vehicle and basic events that happen before the single vehicle crash for highways and non-highways independently. Wang et al. [6] have compared crashes that related to distraction to no distracted crashes by crash type. Distraction crashes account for approximately 13% of crashes in the U.S was reported by this study and in terms of crash type, distraction-related single vehicle crashes are 16% of all distraction crashes was also investigated by this study. Therefore, the study showed a common distraction-related crash situation which is single vehicle run off the road crash situation.

Front to End Scenario

The most common crash scenario is front to end crashes scenario, about 30 percent of crashes in the U.S was accounted [35]. In this context, 21% of front to end crashes in which the vehicle was moving have a distraction contribution and in 24 percent of

crashes in which the vehicle was stopped have the same contribution, all of these results was found by a previous study [6]. Furthermore, front to end crashes were the second most common distraction crash situation and the first most common was single vehicle crashes. Consequently, a large percentage of all front to end crashes were by distracted drivers, no matter if the lead vehicle is moving or not.

Intersection Scenario

Based on the data from GES, the third most common type of crash in the U.S. is intersection crashes in other words crashes where vehicles crossways [35]. According to a previous study on analyses of intersection crash by contributing figure appeared around 7% of these crashes have distraction involvement [6].

Lane-Change Merge Scenario

Previous study [35] reported in U.S. that crashes including a vehicle merging or improper lane usage around 9 percent of crashes. In a previous research [6] by utilizing GES data, inattention distraction was 29 % in lane changing crash scenarios.

2.3 Cell Phone Distraction-Related Crashes

The rise in smartphone utilization over the past few years has taken over all other internet accesses. Today, 2 hours and 32 minutes per day typically were spent by American people, accessing the net on their smartphones or using the widespread apps, a figure that has folded two times compared to the previous year alone. Additionally, in March 2016, it was found that 71% of people's time spend online is from a smartphone, for that month mobile minutes exceeded 1 billion [39].

In 2015, due to cell phone distraction it was reported that there were 442 fatal crashes which are 14% of all fatal crashes that are distraction-related. The police crash report,

for these distraction-related crashes, expressed three main activities or the distracted driver at the time of the crash: talking on a cellphone, listening to a cellphone, or something else manipulating a cell phone. A total of 476 individuals died in deadly crashes that included cell phone use or any other activity that related to a cell phone as a distraction. 30,000 individuals were estimated as injuries in 2015 in cell phone distraction crashes which are 8% of all individuals injured in distraction-related crashes [14].

All cell phone related activities performed by drivers most often associated with three main activities that lead to crashes that are proposed as cell phone distraction crashes are: receiving a call, dialing and talking which have decreasing frequency. Receiving a call while driving is a high risky task, because of the difficulty for drivers to get their cell phones when they often left them in a place that is difficult to reach like on the passenger seat or in a jacket pocket and when the drivers think that like being in home or an office when their cell phone rings while driving, they have tendency to leave whatever they are doing to answer it and subsequently lead to the impairment of safe driving. Although this sort of mechanical effect can be reduced by hands-free cell phones adoption, the cognitive distraction involved cannot be reduced while involving in a conversation [40].

When using a cell phone while driving the risk of is up to 3.6 times higher than not using a cell phone and with the frequency of calls, the risk of involving in a crash increases [28]. The reaction times of distracted drivers by cell phone were more than 40% longer compared to drivers who are not distracted [41].

Previous researches [24, 42] have documented that at any given time, 10% of drivers have been stated as using a cell phone while driving and 33% of crashes and 27% of near-crashes has been recorded as involvement in the cell phone distraction as a secondary task [7]. One of the most interesting approaches to this issue has been proposed by a study [43] showed that the most distraction-related activities experienced while driving were rated by respondents. Three types of activates were categorized as the most distracting activates: using a hand-held portable phone, reading and writing text messages and using a hand-held cell phone. The percentage were 53%, 62%, and 41% of respondents detailing undertaking these behaviors respectively.

Mobile phone use while driving could hence negatively influence driving performance. Nevertheless, a major part in this process was played by the demands of the driving task and the demands and content of the conversation of cell phone. The complexity degree of making a conversation by cell phone while driving which is a cognitive demand task is the vital figure that also determines the impact degree of this distracted activity on driving performance [44].

Regan [17] reported that in responding to the message content, the cognitive demand is significant figure. There are numerous cognitive distractions manual and visual may happen while driving. Moreover, for smart phones, text messaging activity only presents one task of many tasks activates concurrently.

In a previous study [5] the authors investigated that making a conversation on a mobile phone is less distracting than sending an SMS. The crash hazard among distracted drivers due to sending text messages is anticipated to be high although it is obscure as well. While sending a message both mind and the eyes of driver are involved for a

significant length of time with other things than they ought to be involved with the task of driving. Many of activities of a smartphone are proposed a case of this, like accessing social media, web and e-mail. Of late, the impacts of being locked in these tasks of the road safety while driving have been picking up pace[45].

Some preliminary work that was carried out several years ago shows that drivers recognize that dialing a hand-held phone is more distracted as distracted-related activity than dialing a hands-free model [46]and consequences have been revealed in lane keeping and hazard recognition [47]. As mentioned by Tison, et al., [48] sending text messages while driving represents over 60% of distracted drivers as repeatedly, and answering calls while driving represents 66% of distracted drivers. Several studies [49, 50] among car drivers signify that sending text messages strongly amplifies the contribution of being in an accident by a factor of more than 23.

Although text messages have reduced by 6.3% when compared to 2010 for subscribed devices [51]. Further many users are relying on ample of apps in the cell phone for example Messenger, Snapchat, WhatsApp, and Facebook to send messages and that is the main reason for the reduction in texting messages through cell phone which indicate that more potential cell phone distraction will occur. In these days social media is one of the significant distracted factors for drivers as the drivers check their accounts in the social media, take photos, blog videos and many other distracted activates while driving which are very effective to the performance of driving and lead to hazardous consequences.

In 2015, the percentage of drivers who manipulate handheld devices or text a messaging is 2.2%. Whilst, in 2014 for drivers using hand-held cell phone the

percentage decreased from 4.3% to 3.8% in 2015, this was not a statistically significant decrease. Figure 3 represents the percentage of drivers using handheld cell phone vs. drivers visibly manipulating electronic device while driving [16].

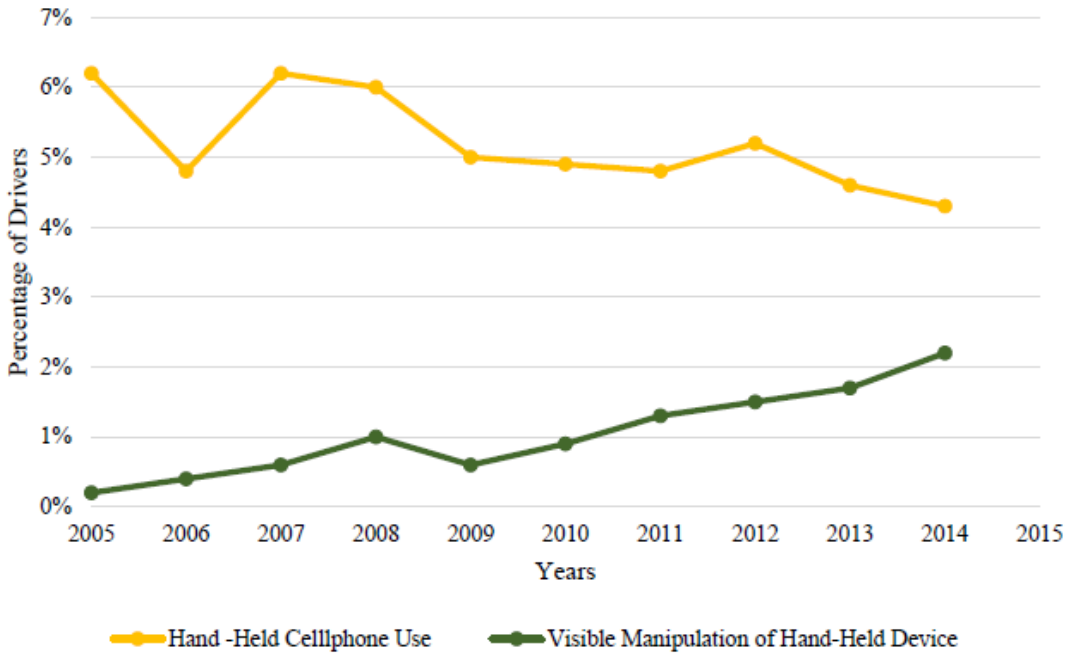


Figure 3: Drivers observed using electronic device while driving (2005-2014) [16]

A naturalistic study was conducted by NHTSA [52] proposed to check the impact of several potential distractions on crashes and pre-crashes. Dialing a phone while driving lead drivers to incorporate a crash by 2.8 times, and talking on the phone while driving lead drivers to incorporate a crash by 1.3 times, compared to the ordinary driving without distraction. In comparison, Strayer et al. [53] shows 5.4 times and in Redelmeier and Tibshirani it was 4.3 times [54] for the same incorporation category.

In a study [55] carried out in 2010 among 405 drivers, 22% of the drivers making a hand-held phone call in their cars at least once a week and 40% making hand-free calls while driving. In a previous study [56] a simulator used with 20 subjects revealed that lane deviations were significantly more likely to happen when subjects were talking

on a cell phone and the same results was for both hand-held and hands-free calls while driving. In another driving simulator study [57], 48 subjects were inquired to distinguish signals blazing at the edge of the simulator screen when they were talking on a cell phone, the likelihood of missing the signal doubled. Additionally, a field test [58] with 12 subjects showed a diminished recurrence of mirror looks and an increased heart rate during a cell phone discussion task.

Lesch and Hancock [59] inspected the effects of cell phone conversations in a simulator experiment with 36 subjects. By 0.18 s, brake responses were slowed and by 0.34 s stopping times were diminished which demonstrating that the brake pedal was pressed harder by the distracted drivers. Although these members braked harder, they still finished up approximately 50% closer to intersections, and the compliance of stop light fell by 14%.

These outcomes demonstrate that the compensation was basically not sufficient, despite the fact that the attempting of distracted drivers to compensate for the delay in their beginning reactions by braking harder. Essentially, a previous study [60] found that talking with a passenger and talking on a cell phone both possibly expanded reaction time to a pedestrian attack event by the same degree.

A limited considerable amount of literature was published on distraction for drivers in the past years, despite the fact that several studies [61-64] examined the impact of different components like crash types, speed, seatbelt use, vehicle types, and drivers' characteristics. Differential impacts of distractions on the behavior of distinctive age groups are anticipated to impact crash results in an unexpected way.

Researchers [65-67] show that young road drivers engage in all distracting activities type as much as older road users, but that the activities are distinct. Young drivers, for example, utilize advanced devices like smartphones or music players more habitually. With beginner drivers, many activities such as dialing a call or getting a cell phone, texting and distracting by looking at a roadside object were all connected with a risk increase severely of a crash or near-crash. Only cell phone dialing for experienced drivers was related to hazardous consequences[68].

There 92% of 18-29 year-old having a smartphone headed the smartphone adoption, the percentage of this adoption for 30-49 year-old was 88% and for 50-64 year-old was 74% [69]. Callaway et al. draw our attention to impressive results that for sending messages 18-24 years old users send twice as many messages as sent by 25-34 years old users and more messages by 10 times than the users of ages 55 and over [70].

For distracted drivers at the time of the crashes, the percentage all drivers 15-19 years old that included in fatal crashes were detailed 9 % as well as the biggest extent of drivers was this group who at the time of the fatal crashes were distracted [14].

Figure 4 shows the percentage of three age groups which 16-24, 25-69 since 2005 who have been observed manipulating electronic devices. The figure illustrates how the first age group (young drivers) has the highest rates compared to the older drivers [16].

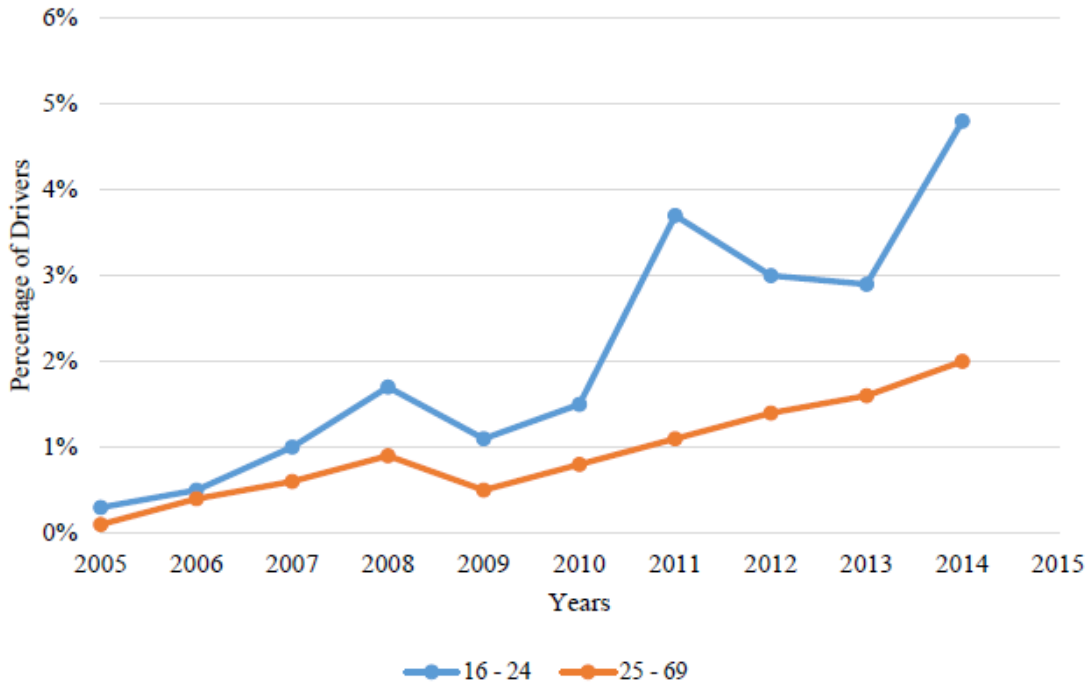


Figure 4: Drivers visibly manipulating handheld device by age group [16]

2.4 Alcohol Impairment Role in Distracted Driving

A study performed by [71] looked for comparing the impact of using cell phone use while driving and the consumption of alcohol and its impact on the activity of driving. The study decided that when the blood alcohol content (BAC) rates increased, the ability of drivers was hampered to drive. It was found that in response time there are many delays in several factors such as braking response time which was 811 milliseconds for drunk drivers compared to 777 milliseconds for ordinary drivers and other driving tasks [71]. By considering the same factors the impacts of mobile phone utilization were also compared with alcohol consumption. The results showed that mobile phone use diminished driving performance in any way with alcohol consumption. With phone use the driving performance decreased similar to drunk-driving [71]. Moreover, the results of previous study [53] which examining the effects of using a cell phone while driving and drinking decided that the driving ability can be disabled when using a cell phone for drivers as a similar way to drunk-driving.

2.5 Cell Phone Distraction Laws

By all increasing risks of cell phone distraction many states presented different sorts of the legislation pointed at confining using a cell phone while driving. The ban on handheld mobile phones in vehicles is the most common legislative degree. Other measures incorporate forbidding using a cell phone for drivers in distinctive driver classes, like novice drivers who have a permission for learners or school bus drivers or other drivers who have particular duties. It has been perceived that the adequacy of legislation could be expanded in case upheld by publicity campaigns in reaction to concerns that cell phone distraction for drivers has become hazardous to traffic safety, therefore plenty of states have endorsed bans and laws that restrict drivers from all sort of cell phone distraction such as talking on hand-held cell phone or texting messages.

In 2014 the American Car Association Foundation [72] showed that distraction because of using a cell phone is much more predominant than is reflected in insights of the official government. It is hard to decide when the cause for a crash is distracted driving. Most individuals do not declare that before they get in a crash those drivers were distracted. U.S. DOT and other safety partners work hard to end distracted driving by cell phone.

Bans were started to pass by many states for using a cell phone include bans for hand-held and texting that forbid drivers from using cell phones if they are behind the steering control of the car. For texting bans, drivers were restricted to send or read messages on cell phones, for handheld bans, all drivers were restricted to engage in phone conversations, whichever by tuning in or talking, on hand-held cell phones. All of these bans has become vital punishments for this violation and checking fines which

expands from $20 to $500 in the adopting states with suspension for license, and without a doubt jail suspension. Cheng discovered that the percentage was diminished by 60% for texting on cellphone and 50 % for talking on hand-held cell phone [73].

In an investigation [74] into state laws restricting mobile use while driving Jennifer et al. show thirty-nine states and the District of Columbia have at least one form of constraint on using cell phone and other connections of devices. These laws alter in the types of connections activities and categories of a driver in time, furthermore authorizing components and punishments. No state totally bans the use of mobile connections devices by all drivers.

A study [75] by US Gallup showed that for a ban on hand-held cell phone use, 70% of the public upheld this ban by drivers. Other similar results were obtained which is 69% by a national survey of ABC News. Two American studies [76] and [77] are some of the exceptional endeavors to not just look at the cell phone bans effects but also look at the viability of these sort of legislations in the long-term. In the term of hand-held cell phone use ban New York in November 2001 was the first state which passed this ban in the US [76]. The law went with by noteworthy presentation and it included two levels which the first one is warning in one month warning and the second one is period of three months in which the driver can delay these fines on a situation that confirmation of purchasing a headset or speakerphone was provided by the driver. The outcomes of this study show that the anticipated result was obtained from New York's ban for the first months after it was enforced. Figures 5 and 6 represent the cell phone laws in each state compulsory on drivers[78].

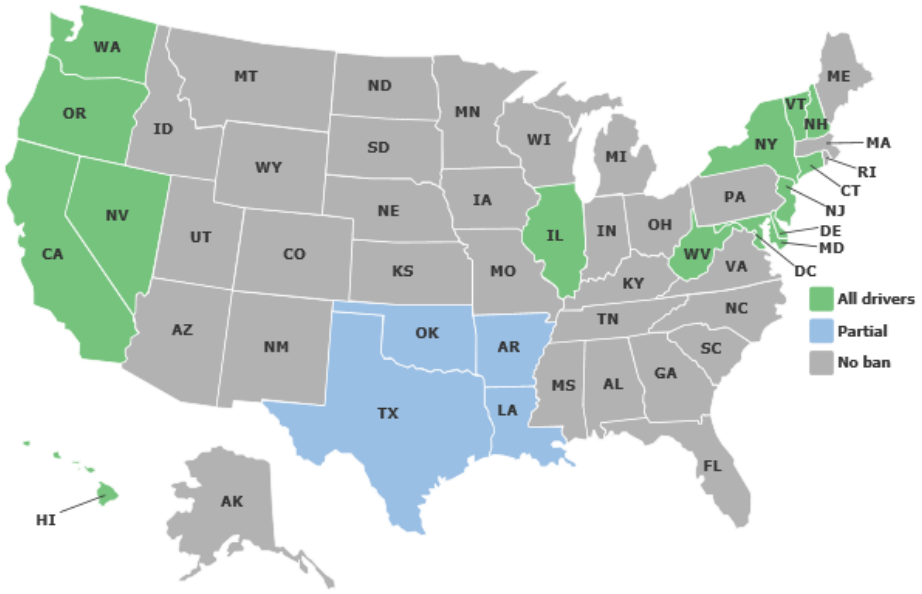


Figure 5: Cellphone handheld ban types imposed by states in US.[78]

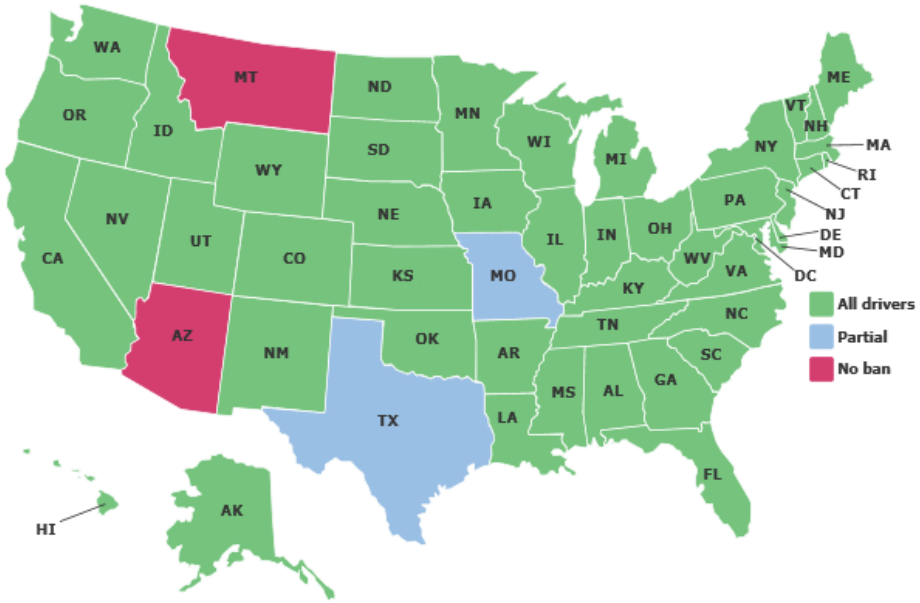


Figure 6: Cellphone texting message ban types imposed by states in US.[78]

Occupant Protection Incentive Grant Program (OPIGP) was proposed by The U.S. government in order to empower all states to execute productive and fruitful programs pointed at reducing number of injuries and fatalities on all roadways in the U.S. [79].

The program were adjusted by the NHTSA at the level of government, and for the level of state the State it was been abroad by Highway Safety Offices (SHSOs). For the financing qualification, governments of the state must accomplish the criteria of requirement. One of those is that particular laws and directions to deny using electronic devices while driving and texting for all novice drivers who are under 18 years age must be arranged by the state [80]. In case that criteria is met by states, at that point, of the funding of distracted driving plans, the states allocates 8.5 % of the financial requirement. For national media campaigns, 5M$ have been chosen on distracted driving issue. Since 2011, month of April was distinguished as "National Distracted Driving Awareness Month." by NHTSA [81].

Burger et al. [82] found no verification that the hand-held cell phone use ban diminished accidents. Moreover, no significant effect of fines for the laws violation which can be low as 20$ per violation. Furthermore, driver's record points commonly is not incorporated by tickets. A powerless necessity may as well restrain the laws' viability. A low probability of endorsement attached with small fines may not allow more forcing constrain for the compliance of driver. At long last, bans of distracted driving may be inadequate on the situation that drivers do not tend to be compliant with the law and in the same way those distracted drivers are completely amenable towards driving of crash-prone.

Indeed in case drivers have completely the required compliance with the law, it is conceivable that the rate of accidents would not diminish. To begin with, in the event that both hands-free and hand-held of using cell phone the effect of distraction is the same [83], the rate of accidents by changing from one strategy to the other may be unchanged. As Hahn et al. [84] and Prieger et al. [85] propose, a hazard may shift

over distracted drivers and those cell phone distracted drivers may essentially be more careless and inclined to be involved in an accident. In this situation, the minimal impact of cell phone usage for hazardous drivers is little, at that point, the prohibition would have a small impact on results. Once more, researchers need more information to appraise the impacts of this distraction and the validation of these bans.

In the case of the lack of information in the driver compliance, progress investigation into this issue may be important to unravel these effects and offer an assistance guiding public approach to cell phone use. For illustration, in case of the issue is in the compliance, in order to reduce rates of crashes extend the fines could be satisfactory [85].

2.6 Data Mining (DM) Approach for Cellphone Distraction

Ordinarily, two categories can be distinguished from data mining portrayal and expectation. In other words directed and undirected data mining, the first one regularly called predictive modeling which is a top-down process and when the extreme outcome is predicted this process is utilized. Undirected data mining is the second phase, utilizing when it is requested to portray the information due to the merit of bottom-up process. The most broadly utilized tasks are portrayal, forecast, classification, clustering, segmentation, association. As it is known that due to rich data and poor information the best solution to support knowledge workers is data mining process which is a multidisciplinary field and provide a method to extract information. By the wide range of data analysis techniques, these tools provide a method to discover the relevant information. Data mining technique consider any method that used to extract patterns and relationships from a source that has a given data.

Functionalities: As mentioned before, although many data mining tasks are used in data analysis, it is classified into two categories descriptive and predictive. Descriptive data mining tasks illustrate in the database the general characteristics of the given data. Predictive data mining tasks in order to make prediction which obvious from its name accomplish inference on the given data.

Main data mining functionalities are as follows: Data characterization which it is one of the first techniques that summarizes the classes of a given data from the study. Data discrimination which is the second functionality that makes comparison between the classes with one or more sets of the given data. Association analysis is the third one which shows the conditions of attributes value in a given data that frequently happen together by discovering the association rules. Classification is the widespread process that commonly used to find bunch of models that provide the description of the class of a given data and distinguish concepts of these set of data in order to predict the class of attributes when the class label is unknown by using the current model which has gotten by this process.

Prediction technique is one of the most significant functionalities that used in data mining process. It aims to predict the missing information or any unavailable data it is nothing but in many applications, users may wish to predict some missing or unavailable data by the given data. Evolution analysis which asses the models that have gotten before and discover the trends of the attributes that have changeable behavior over time [86].

Classification technique that is commonly used to classify categorical data and discover the relationships among different data whether it is categorical data or

continuous numeric values. Several classification techniques have gained large acceptance in many fields of data mining performs like decision trees and random forest. Decision trees provide superior advantage among these classification techniques. Creating rules that very interpretable and producing logic statements for a given data by the classification tree [87].

2.7 Data Mining Approach in Traffic Crashes

In road safety, the major challenges in research are: how to recognize the most frequent patterns of the accident, individuation of the most substantial elements of traffic accidents, and in order to address the most related issues, how to allocate the resources necessary. Typically, traffic accidents have been regarded as random events and statistical models have been extensively employed to investigate the determinants of fatal and injury accidents.

There have been several attempts in the late of the last decade and the beginning of this century to use the techniques of data mining in the area of traffic safety. In particular, frequent patterns in accident data have been searched by implementing spatial data mining [88], decision trees such as Clarke et al., in 1998 and Bayam et al., in 2005 [89, 90].

There are numerous research papers [91, 92] about the technologies of data mining of traffic data such as, traffic safety optimization, traffic jam visualization, plan of street, GPS assisted navigation and road design. An alternative approach is constituted by data mining techniques that in recent years, increasing attention has been received by researchers. Smith et al. [93] show that high potential was hold by advanced techniques

of data mining in order to deliver automated tools that were helpful in signal control system operations and design for traffic engineers.

Four machine learning models were evaluated by Chong et al. [94] applied to modeling the injury severity that happened during traffic crashes: decision trees, support vector machines, neural networks, and a hybrid model involving decision trees and neural networks.

S.Shanthi, et al [95] highlight the importance of the classification algorithms of data mining to predict the patterns of vehicle collision happened for a given accident data. Several classification algorithms have been applied like random forest, C4.5, CTree, CS-MC4 and Naïve Bayes in order to predict the patterns of vehicle collision and the given data were obtained from FARS. A previous research [96] highlights the significance of classification algorithms of data mining to predict the attributes which affect the accidents of road traffic and specifically to the severity of injury. The performance of classification algorithms were precisely compared which were C4.5, CTree, CS-CRT, Naïve Bayes, MC4, and random forest that applied to modeling the severity of injury that happened during accidents of road traffic. The results show that random forest in the context of feature selection outperformed the other approaches.

2.8 Data mining approach for distracted driving

An important research [97] associated with distraction while driving shows essential results for distraction crashes. Data mining strategies were applied in order to find the relationships between the distracted driver inattention and crashes by vehicles from FARS dataset from 2000 to 2003. The research focused on Maryland and Washington, DC zone.

Clustering technique was first done by utilizing the Kohonen networks. After that, decision tree and neural network models investigate the rules and designs of the given data. Result proposes that when the driver has at the same time physical or mental conditions beside inattention, the type of the crash that included a distracted driver would be with fixed objects. Moreover, with respects to the type of the crash, for the first harmful event the percentage of the relative importance of involving in a crash with moving vehicle and involving in a crash with a fixed objects is 2:1. The significance of this research that discover the probable relationships and rules for drivers that having inattention and vehicle crashes [97].

One of the most significant research related to distraction while driving is the research conducted by Ghazizadeh and Boyle [98] utilizing crash data for five-year period from 2001 to 2006 and the data was particularly in the state of Missouri. The study uncovered that using cell phone while driving or other electronic devices or had a passenger in the car were the most noteworthy reasons to be involved in a distraction-related crash. The methodology of the study was multinomial logit model in order to predict the type of crash that the distracted driver would be involved which were front to rear, angular, single crash and made a comparison with those crash types. The outcomes of the study demonstrated that crash type could be changeable by influence of the distraction. More particularly, angular crashes are the main type if the driver distracts by a passenger and a cell phone. Single crashes would be the particular type of crashes by the presence of other electronic devices as the distraction type [98].

In a previous research[99], data mining was applied to a given data in Saudi Arabia in order to investigate the crash severity and the factors that leaded to it. J48, CHAID, and Naive Bayes were the three classification techniques. After that all provided

models by the previous techniques were assessed and then compared. The outcomes of the study highlights how distraction could be dangerous to driver's life. As a result by the obtained models, if the distracted driver hits an electric tower the consequences would be disastrous and lead to death. Moreover, if the distracted driver hits fence or motor vehicle the consequences would be serious injuries. One of the significant factors was the age of the car. The results would be injure or death if the crash was with older cars are more likely. At last to evaluate the models, the study has gotten high accuracy for all models.

2.9 Using Random Forest for Variable Ranking

As mentioned before there are many data mining techniques that utilized in traffic crash analysis and by searching one of the most effective technique is random forests (RF) which has several functionalities such as forecast, classification, considering variable importance and variable selection. In terms of variable importance, plenty of studies have been conducted by random forest.

Variable importance rankings particularly when the number of trees in RF is small, depend on how many factors utilized in splitting a node when users are designing RF. When utilizing variable importance assessments this truth should be noticeable for data understanding and investigating. In any case, the number of the factors in splitting a node may vary essentially from the value which is the default. Hence, when using the optimal number of factors in splitting a node a distinctive ranking of factors may be gotten [100].

A previous simulation study [101] utilized some levels of association between the true predictor attributes and the dichotomous response as well as many levels of correlation

for the predictor attributes. In term of classification and variable importance random forest technique was shown as very important technique especially if the purpose of the study to get an accurate models and afford vision by the ability of the predictor attributes.

A previous study [102] that discussed several factors like characteristics of the environments associated with avoidance maneuvers of the crash and other factors such as drivers and vehicles. Different types of crash were examined, front to rear crash, front to front crash and angular crash by utilizing decision trees and variable importance on two variables which are evasive and no evasive actions. Furthermore, by setting crash avoidance maneuver as the dependent variable the random forest was utilized as variable importance in order to rank the factors driver, vehicle and environments' characteristics.

The outcomes of the study showed by analyzing the results that for drivers, three phases were connected to crash avoidance maneuvers which are distraction, physical impairment and visibility obstruction and consequently connected to the three types of crash. Additionally, front to rear crash connected to speed limit for avoidance maneuvers and front to front crash and angular crash connected to the vehicle type for avoidance maneuvers. One of the limitations of this study that it did not determine the type of distraction and their effects on avoidance and that because of the sample size was limited [102].

2.10 Limitation observed in the Literature

A newly elevated alarm is considerably rising which is cell phone distraction within the context of traffic safety issues and in the future it is a must to do substantial

enhancements. In terms of distracted driving, there are two major factors correlate to each other which are cell phone use and performance of drivers. Although many previous studies examined impact of diverse types of distraction while driving, a noticeable drawback of many previous studies lacks taking cell phone distraction as a type of distraction to be more accurate in determining the distracted driving behaviors.

Although, there are numerous researches that have been conducted on distraction while driving there is an absence of information in defining specifically the relationships between the distracted driver by cell phone and the driver related factor and how the manner of collision will be affected by the performance of drivers distracted by cell phone. Moreover, discover the connection between different types of crashes and the most harmful event in this cell phone distraction crash through data mining approach.

For distracted drivers there is an important phase to interpret the increase in the risk with the different crash type's frequency. Therefore, the most frequent crash types are: angular crash, front to rear crash and single crash which is associated with fixed objects[98]. However, there are many studies that examine the role of drivers in traffic safety issue but omit the role of distraction in determining the response of drivers in crashes like [102]. As mentioned in the previous section, many studies did not discriminate the distraction types and their impacts on the performance of the driving task.

Neyens et al. [63] and Ghazizadeh et al. [98] represent two of few studies which considered the associations between driver distraction and crash type. Neither of these studies investigated how particularly cell phone use is proposed as a major type affecting drivers. In addition, many of the significant factors that affect the driver

behavior were omitted in these studies. Although [97] examined the distraction role in drivers performance and how driver-related factor has an important impact in determining the most harmful event but this study talked about distraction in general which makes determining the effect of the cellphone as a widespread distraction type in particular more difficult. Selecting the most important attributes in this study was weak by just choosing four attributes to investigate the manner of collision for distraction crashes.

Another research [103] takes drowsiness as the dependent factor of distraction crashes. The methodology was multinomial logit model in order to focus on vehicles which were in a single crash, vehicles that hit a lead vehicle from the rear, and vehicles that hit another vehicle or were hit in an angular crash. The results showed drowsiness/fatigue as a distraction type given that it results in progressive withdrawal of attention from the roadway and without taking cell phone distraction as a factor of distraction.

One of the most vital limitations in the literature was that the data was taken from particular states like [97] which focused on Maryland and Washington, DC and [98] in which the crash data was in the state of Missouri. Thus to our knowledge, there is no research that takes the whole united states to investigate cell phone distraction associated with the crash type.

To fill the gaps in the literature, research in this dissertation seeks to examine the distracted drivers by cell phone in specific to detect the relationships by data mining among many attributes selected carefully by one of the most effective methods in feature selection for the whole United States.

2.11 Objectives and Scope of the Study

Rationale of the Study

The problem of distracted drivers that they may not realize the hazard of using cell phone while driving as they think that multitasking is easily possible without any dangerous consequence which it is a wrong recognition. That is right in some cases when the distracted drivers by cell phone reach to their destination without involving in a crash. An acceptance is leaded to those drivers that they can handle multitasking successfully by this circumstance. However, the fact of traffic safety that crashes and accidents do not happen frequently in the roadway and that fact is not realized by the distracted drivers. As a result of these untrue recognitions and with the development of cell phone industry, the number of distracted drivers that utilizing cell phones while driving has expanded. Therefore, distraction-related crashes have gotten to be a critical concern for administering offices and NHTSA. Several laws and solid regulations were made by numerous states over the U.S. which forbid using cell phone while driving. The main problem in this topic that lack of self-reporting and many regulation were made, there is a requirement to know the portion of using cell phone whether by sending messages, making calls or any other activity related to cell phone distraction which is more important to evaluate the performance of the distracted driving. Due to safety concerns with field studies, Data mining application on one of the most reliable datasets (FARS) study was conducted in order to investigate the effect of using a cell phone on the performance of driving (i.e., the most harmful event, the manner of collision, driver-related factor). The results will give vital information in understanding the relationships among the most important attributes caused by cell phone distraction, and the overall effect on roadway safety.

Study Objectives

Cell phone distraction is examined in order to investigate the relationship between this prevalent distraction and motor vehicle accidents and how cell distraction influences driver performance. This research sheds new light on determining the essential factors among seventeen attributes.

Discovering and exploring the role of distracted drivers on fatal crashes was done by applying data mining techniques. Classification algorithms utilized in determining the most related variables for the manner of collision and predicting the vehicle collision patterns occurred in fatal accidents across the United States. In addition, study the most vital attributes related to the most harmful event, classifying and predicting how cell phone distraction contributes with the most harmful event in fatal crashes.

The classification algorithms C4.5, Ctree, C5 tree have been applied in forecasting the patterns of the crash type and the most harmful event. Random forest and cforest have been utilized to determine the importance of attributes.

Originality

The significance of this research that utilizing the techniques of data mining to discover and explore the potential relationships between driver that distracted by cell phone and motor vehicle crashes as one of the first research in this field.

This up-to-date research investigates cell phone distraction by one of the most reliable databases (FARS) and study the relationships among many attributes affect the driver who is distracted by the cellphone.

Chapter 3

METHODOLOGY

3.1 Dataset and Tools Used

The data in this thesis was obtained by FARS where the data is available in FARS website and categorized by several attributes and for each year separately. The attributes was chosen for five years and associated with specific type of distraction which is cell phone distraction. FARS database records crashes that result fatalities within 30 days of the crash and the crashes involved motor vehicles and take place on an open trafficway. Distracted drivers that involved in any crash have been recognized by FARS since 2010 which refer to distraction-related crash.

Therefore 2063 records that occurred during 2011 to 2015 were identified of distracted drivers by cell phone that have been involved in a fatal crash and reporting driver while doing any of three categories: talking or listening to cell phone ,manipulating cell phone while driving and other activity related to cell phone. According to FARS manual the explanation of all types of cell phone distraction using in the thesis were explained. The following paragraph explains each type of cell phone distraction according to FARS that is used in this thesis.

Talking or listening to cell phone while driving: Talking or listening on cell phone which includes talking or listening on whether a hand-held, hands-free phone or Bluetooth-related device.

Manipulating cell phone while driving: By dialing or text messaging on cell phone or any wireless device which includes using manual button control on phone being.

Other activity related to cell phone: Used when none of the identified codes in the police report are related like calling or texting messages on cell phone but the police report recognizes a distraction from the driver due to cell phone engagement.

FARS dataset was selected because it is the most extensible database and it is available for anyone to use it with the merit of the most updated information for fatal crashes. Additionally, because if the analysis is limited to just fatal crashes that would simplify data mining techniques task in identifying crash patterns and the contributed factors related to the crash without the "noise" which is recommended by several researches [97, 104]. In other words, when fatal crash happened all the 2063 drivers that were observed from FARS had been distracted by cell phone.

Seventeen variables were identified for data mining process. These are seventeen attributes that provide the detailed description of the crashes, drivers, vehicles and pre-crash. Pre-crash variables describe what the vehicle was doing just prior to a crash, what made the vehicle's situation critical.

Numerous variables were considered among those seventeen attributes such as driver characteristics, road characteristics and traffic conditions. Table 2 shows the list of the attributes used in this thesis and their descriptions with the values for each category.

Table 2: Attributes and their description

Attributes	Description	Sub categories	Values
AGE	Driver's age in a 10-year intervals	1 (10 to 19) 2 (20 to 29) 3 (30 to 39) 4 (40 to 49) 5 (50 to 59) 6 (60 to 69) 7 (70 and more)	293 728 436 277 206 84 38
event	The most harmful event applies to the vehicle	Motor Vehicle Fixed Object Pedestrian Rollover other	1046 363 266 367 20
MAN_COLL	Manner of collision	Not collision with vehicle Front and Rear Front to Front Angle Sideswipe	1049 293 297 362 61
SEX	Gender of the driver	Male Female	1244 818
DR_CF	Driver-related factors in the crash	None Careless Driving Failure in Traffic Laws Improper Lane Usage Over Correction Failure To Yield Right-of-Way Object Other	784 350 105 298 159 169 35 162
WEATHER	At the time of crash how weather condition was	Clear Rain Fog Cloudy Snow	1593 135 20 300 14
CLASS	At the time of crash what was the class of road	Locale Interstate highway	746 292 1024
Light_Cond	Light condition at the time of crash	Day light Dark	1218 844
INTERSECTION	Intersection existence	NO Intersection Intersection	1642 20
AIRBAG	Airbag deployment	Not Applicable Deployed Not Deployed	220 1026 816
ALCOHOL	Driver alcohol involvement	Not Involved Involved	1580 482
RESTRAINT	Restraint system usage	Not Used Used	132 1930
DAMAGE	Extent of damage in the vehicle	No Damage Damage is Minor Damage is Functional Damage is Disabling	15 101 228 1718
ALIGNMENT	Roadway alignment	Straight Curved	1502 560
TRAFFIC_DIS	Description of trafficway	Two-Way, Not Divided Two-Way, Divided One-Way Trafficway Entrance/Exit Ramp	1401 619 24 18
RELATED_TRAFFIC	Relation to trafficway	On the Roadway On the Shoulder On the Median On the Roadside	1242 60 100 660
SPD_REL	Relation to driver speed	NO Yes	1474 588

3.2 Data Mining Techniques Used in Traffic Accident

Hamid and Barko [105] cautioned that in order to get a successful data mining project, many component have to be accomplished. The missing or incorrect values or wrong data should not be in the process of data mining. Thus, data mining application in this thesis was conducted with fundamental steps for the dataset to get the most use of the techniques.

For our research, it was used RStudio which is an R language implemented software and one of the most widely used DM softwares to apply data mining techniques of the given data which resulting from the FARS database in two models[106].

Firstly, it was applied random forest to investigate the importance of variables and study the effective change of the variables and the contribution of each during the five-year period (2011-2015) according to the dependent variable which is the manner of collision in the first model and the most harmful event in the second model.

According to Mitchell [107] to reduce the bias of random forest as an importance variable ranking, *cforest* technique was applied for subsampling without replacement for 2015 year according to the dependent variables.

After that, decision trees was applied to categorize relationships among the selected variables which have been recognized and identified previously. For classification process, manner of collision (MAN_COLL) is the dependent variable in the first model and the most harmful event (event) is the dependent variable in the second model and the if-then rules of other independent variables determine the dependent variables.

Lastly, it was applied the confusion matrix according to the applied decision trees to assess our models.

Data preparation or preprocessing is extremely important both in data mining and in the pattern recognition process. However, there are numerous types of preprocessing tasks like treating missing values, diminishing noises, dimensionality reductions, variable aggregations, feature establishment, sampling and feature selection, attribute transformation.

Data Preprocessing

For applying data mining the original dataset in the beginning is not ready, a single fixed value set could be noticed in some cases for all the records. These variables require being modified by preprocessing to get the dataset ready. Data preprocessing is critical and in some cases a challenging task in data mining.

Data Preparation

The variables are categorized according to the manner in which the collision happened and has four categories, the most harmful has five categories as the dependent variables, it was shown the elements of preparation in the following sections.

Data Cleaning: the missing values should be filled in, any noise data should be smoothed and inconsistencies in the data should be adjusted [108]. In order to get rid of any deviation in the results a few cases with missing values were excluded.

Data Transformation process: It changes over the given data into suitable forms for mining the data. The dataset utilized contained two types of attributes value: integer

values and categorical values. In FARS database the representation for some attributes in numerical way with which each number representing a categorical value. So it was distinguished categorical factors and coded them by categorizing them as factors and for other attributes like AGE it was derived the input values into intervals as factor values as well. Comparative transformations have been done to the categorical factors in order to deal with them as factors in the software, not integers.

Relevance Analysis**:** Feature selection is a very important method of supervised classification, by reducing the space of attribute in a feature set. The purpose of variable selection has three phases: the performance of prediction is developed for the predictors, the predictors would become more effective and quicker and the basic process which made the data would be more understandable [108].

Splitting of data: Part of the given data should be assigned to training and the remaining part of testing. The given data is divided into two sets the first one is a training set (70% of total records) and the second one is a testing set (30% of total records) and this 70/30 split is the most accurate split for decision trees and provide the most accurate predictions for decision trees that recommended by many studies [97, 109] . Building the model is done by the training set and assessment of the model for correctness is done by the testing set.

3.3 Variable Importance with Random Forest Method

Ensemble learning methods have received a lot of interest because they generate several classifiers and aggregate their result. Breiman [110] investigated a well-known method called bagging of classification trees (CTs). CTs were demonstrated by many researchers [111, 112] as unstable learners which by acquiring samples of bootstrap

Lb from the sample of the original learning L that can be stabilized, the predictive model is improved on Lb then the class prediction is being averaged over the b= 1,….,B of predictors. This procedure is called bagging or bootstrap aggregation.

An extra layer of randomness to bagging was proposed by Breiman [111] that which is the random forest. The construction of the classification trees are changed by random forests. In a random forest, each node is split by using the best split among a subset of predicto rs which is selected randomly at that node instead of using the best split among all variables which is the process in standard trees.

To be clear, RF not only captures the most significant relationships but also those more delicate to develop predictions. As in Bagging, the number of trees to figure is determined by the user. Compared to various classifiers such as support vector machines, neural networks and discriminant analysis, this technique turns out to perform superbly and overfitting is not a problem because the random forest is robust against it [111].

Out-of-Bag Estimate of Performance:

Preferably, utilizing a large independent testing data set which was not utilized in the training is done in order to evaluation of performance for a prediction algorithm. In practice, some sort of cross-validation is ordinarily utilized when the data is constrained. A sort of cross-validation is performed by random forest in parallel with the training step by utilizing the samples that are called Out-Of-Bag (OOB) samples.

Particularly, each tree is developed utilizing a specific bootstrap test in the preparation of training step. As the bootstrapping process is sampling with substitution from the

data of training set, from the test a few of the particles is going to be left-out, while a few others in the sample will be repeated.

In order to simplify the process, the construction of random forest starts with taking samples from the training subset randomly with replacement so the chosen cases from sampling will be called in the bag due to randomness at each node and use them in constructing tree. The left out cases that were not chosen from sampling which are out of bag that are not used in the tree, they will be passed to the make the assessment as the true values are existed and OOB error will be the difference between the true values and OOB samples.Practically, by utilizing around 2/3 of particles that in the training set, each tree is developed, and clearing out the rest of them which are 1/3 as OOB.

Algorithm of random forest

Figure 1 shows an architecture of random forest in general, where B presents the number of trees in RF and k1, k2, kB and k are class labels. As the number of trees in random forest escalates, the set error rates converge to a limit, therefore that means in large random forest there is no over-fitting. For accuracy, two necessary conditions low bias and low correlation. In terms of catching low bias, maximum depth of trees should be done. Moreover, for reaching low correlation, applying randomization is requested by: in the training set each tree of random forest is grown on a sample of bootstrap and after that when growing a tree, certain number of variables that is called mtry are randomly selected out of the P variables available at each node. Commonly, mtry<<P so in order to start with a specific mtry it is suggested that mtry= (log2 (P) +1) or mtry=√P and then increasing and decreasing mtry until achieving the minimum error for the data set of OOB.

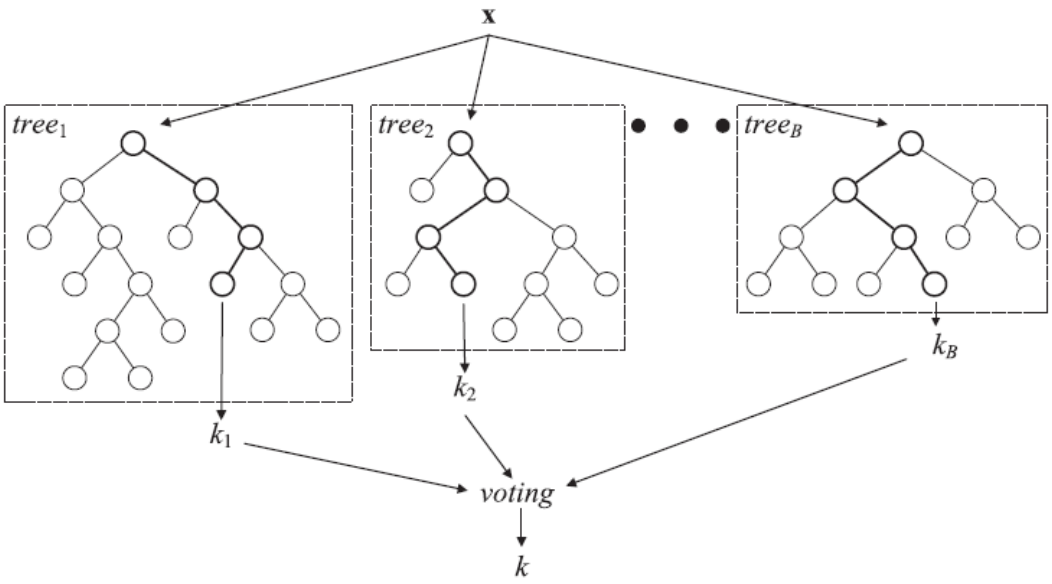


Figure 7: An architecture of a random forest in general [113]

The algorithm of random forests begins with the original data which n tree bootstrap samples is established then grow an unpruned classification tree for each of the bootstrap samples and after that randomly sample mtry from the predictors and choose the best split among these variables where it is proposed that when obtaining mtry = P, the number of predictor, bagging presents as a special case of random forest. After that, by aggregating the predictions of the n tree trees, make the required prediction of the new data which has the majority votes for classification.

Based on the training data, obtaining an estimation of the rate of error by the following:

- The OOB data is predicted at each bootstrap iteration.
- Aggregating the OOB predictions from the previous step which is around 1/3 of the sample for each data point and after that computing the error rate which is called the OOB estimate of error rate[114].

Variable importance

Several variable importance processes were implemented in random forest. Based on the node impurity and the accuracy of classification of OOB data, the Gini index was chosen to determine the importance of the variables which is called Gini importance. The improvement in the splitting criterion of Gini index is described by the Gini importance [115]. Gini importance is a measure of how often a randomly chosen variable from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. The Gini importance can be computed by summing the probability Pi of an item with label i being chosen times the probability of a mistake in categorizing that item [107].

A node t is given and class probabilities is estimated as p(k/t) where k from 1 to Q then the Gini index is defined as

$$G(t) = 1 - \sum_{k=1}^{Q} p^2 \left(\frac{k}{t}\right) \quad (1)$$

Where Q is the number of classes.

In order to compute the Gini index, the decrease in the Gini index is calculated at each node for variable Xj which used to split a node. The Gini index-based variable importance measure $\overline{\Delta_j}$ is given by the average Gini index decrease in the forest, where the split of a node was made by using the variable Xj [113].

Nicodemus and Shugart [116] show how the Gini importance has the ability to distinguish the most effective predictor attributes in the related variables and exceedingly mtry which is the number of chosen splitting attributes is significant in determining the Gini importance.

The user indicates the number of attributes that are chosen in random at each split which is mtry. Thus, to get unbiased variable importance process it was optimized mtry to be more accurate by R software because the mtry parameter has the most significant effect on the ability of actual predictive process according to numerous past considers usage [107]. The number of attributes inspected mtry has the biggest effect on the true prediction error. An unbiased assessment of OOB for the true prediction error is claimed. Therefore, the mtry was optimized based on the minimum rate of OOB error. Subsampling without replacement which is mtry is the solution for this bias to be reduced and from each group the same number of observations is chosen.

Conditional inference forests (cforest)

Based on the Gini index principle which forms Gini importance variable in this algorithm is engaged in the algorithms of classification tree. The objective of Gini importance is to reduce the essential Gini index splitting principle bias when the measurement scale or the number of categories of the predictor variables are varied [115, 117, 118].

Therefore, instead of bootstrap sampling, subsampling without replacement is more reliable measure of variable importance for uncorrelated predictors and therefore unbiased trees are used in constructing the forest. Thus, it was proposed by Strobl [115] to use conditional inference forest (cforest) to reduce the bias according to variable importance in random forest as much as possible. From the party library in R language, the reduction was achieved with the *cforest* function [119].

In contrary to random forest, the cforest function forms random forests not by using the classification trees based on Gini criterion which is the formation principle in random forest [120]. Thus, the permutation importance was chosen for cforest.

The importance measure $\overline{D}_j$ for variable Xj by having bootstrap samples where b is from 1 to B, is calculated by setting b=1 and the OOB data points L_b^{oob} is found then categorize L_b^{oob} by using Tb tree and count the correct classification number R_b^{oob}. For variable Xj, j=1,….,P, permut Xj values in L_b^{oob} and the results of the permutation into L_{bj}^{oob} then using Tb to categorize L_{bj}^{oob} and count the correct classification number R_{bj}^{oob} then repeat the same process for b from 2 to B. Features which produce large values for this score are ranked as more important than features which produce small values. The permutation importance $\overline{D}_j$ is given for Xj variable by

$$\overline{D}_J = \frac{1}{B}\sum_{b=1}^{B}(R_b^{oob} - R_{bj}^{oob}) \quad (2)$$

3.4 Classification Techniques

Decision trees (DT) Make a tree of conditional (If . . . then) explanations that partition the training data by ground truth labels. Each partition is chosen using data gain, which is a degree of the decrease of uncertainty in the data after making a part. After a parcel is chosen, the information characterized by the partition is assessed for extra splits.

The process stops when the training data are completely divided by their labels or when the greatest depth of the tree has reached three types of decision tree algorithms[121]. Therefore, the following three different types of the decision tree were utilized in this study.

C4.5 tree algorithm

By changing numbers of branches at each node the C4.5 algorithm produces a tree. For categorical factors, for each value of the factor, C4.5 adopts one branch and comes with a program of companion in order to turn trees into rules. By using the algorithm of divide-and-conquer, C4.5 creates an initial tree by letting S be the set of cases associated at the node, with the most frequent class in S the tree is a leaf labeled if all the cases in S is related to the same class or S is small.

After that, based on a single factor with two or more outputs, choose a test. The root of the tree will be this test with one branch for each test outcome, corresponding subsets S1, S2… is partitioned in S according to the test outcome for each case, the same procedure is applied to each subset concurrently.

Generally, in the last step there are numerous tests that can be selected. Two heuristic measures is utilized by C4.5 to make possible tests ranking: attribute selection measure is used by information gain, which the total entropy of the subset Si is minimized and the default gain ratio that by the information provided by the outcomes of test divides information gain. The function Gain (A) describes the algorithm of information gain, as which is shown below [122]:

- By the highest information gain, the attribute is chosen and Si tuples of class Ci is contained by S where i from 1 to m
- In order to classify any arbitrary tuple, expected information or information measure is required:

  $$I(s_1.s_2 \dots s_m) = -\sum_{i=1}^{m} \frac{si}{s} log_2 \frac{si}{s} \quad (3)$$

- With values a1,a2,…,av, entropy of A attribute:

$$E(A) = \sum_{j=1}^{v} \frac{S1_j + \cdots + Sm_j}{S} I\left(S1_j \ldots . Sm_j\right) \quad (4)$$

- how much can be gained is the concept of information gain by branching on A attribute**:**

  Gain (A) =I(s1, s 2,..., sm) - E(A) (5)

Algorithm of C5.0 tree

Algorithm of C5.0 is an extension of the algorithm of C4.5 which applying this algorithm to big set of data as a classification algorithm. The efficiency, speed and memory all those properties are better in C5.0 algorithm than C4.5 algorithm. Splitting the given sample based on the field is how C5.0 model works which the maximum information gain is provided. On basis of the biggest information gain field, the samples in C5.0 model can be splitted. From the former split that was gotten the sample subset and this subset sample afterward will be split. This process will stop when the sample subset cannot be split. Finally, the lowest level of the split is examined and is rejected any subset of the sample that has not remarkable contribution to the model. Missing attribute from dataset and multivalued factor is easily handled by C5.0. Therefore the formula of the algorithm is the same for C4.5 and C5.0 trees [123].

Conditional Inference Tree (Ctree) Algorithm

An algorithm is provided by Conditional inference trees (CTREEs) which split a large group of observations into such groups by using statistical tests. For this purpose, CTREE is one of the newer classification algorithms [118].

The recursive binary partitioning is the concept of CTREE which means splitting into two recursively will happen to the full group of observations until reaching a stop

criterion. A binary decision tree is formed and information is presented by this tree which descriptive variables that is grouped together.

Three steps describe the algorithm of CTREE [118], firstly the independence global null hypothesis is tested between any of the descriptive variables with the response which if this hypothesis cannot be rejected where $p>0.05$ then it should be stopped. Otherwise with the strongest association to the response, the input variable is chosen. A p-value measures this association corresponding to a test for the partial null hypothesis of the response and a single input variable. The p-value and the stopping criterion can be modified.

Secondly, a binary split is implemented in the input variable that is selected before. A permutation test is used by the algorithm in order to find the optimal binary split in the input variable that is selected before. Lastly, the first two steps should be repeated recursively until identifying a stop criterion.

In order to assess manner of collision and the most harmful event, these classification models were made using decision trees technique. The data were assessed and then were added to a clean dataset included 476 records for the year of 2015 which included the largest number of fatal crashes with drivers distracted by cell phone in the whole five-year period.

The dependent variable MAN_COLL (Manner of collision) has five categories: “Not A Collision With Motor Vehicle In-Transport”, “Front-to-Rear”, “Front-to-Front”, “Angle” and “Sideswipe”. Front-to-Rear crash is defined as happen when two vehicles in the same direction in the roadway and the first one coming into a contact with the

lead vehicle. Front-to-Front crash is defined as the two vehicles in the opposite direction in the roadway contact. Angular crashes include crashes by vehicles that are not in the same direction when traveling. Not a collision with motor vehicle in-transport crashes which is called singular crash occur between a vehicle and a fixed object or any other thing but not a vehicle.

For the second process to examine "event" (the most harmful event) the dependent variable has five categories: “Motor vehicle in transport”, “Fixed object”, “Pedestrian”, “Rollover” and “others”.

In order to obtain more accurate results for decision trees in our models and to avoid noise in our analysis, it was picked up the most important attributes. Thus, based on the contribution of the input attributes that make the formation of the decision tree in random forest, feature selection is done.

The evaluation of the three types of decision tree algorithms was done by confusion matrix for the manner of collision model and the most harmful event model. The accuracy (ACC) is calculated for the confusion matrix by the following relation:

$$\text{Accuracy} = (TP + TN)/ (TP + TN + FP + FN) \quad (6)$$

Where:

Positive (P): Reference is positive

Negative (N): Reference is not positive

True Positive (TP): Positive reference, and is predicted to be positive.

False Negative (FN): Positive reference, but is predicted negative.

True Negative (TN): Positive reference, and is predicted to be negative.

False Positive (FP): Negative reference, but is predicted positive.

Consequently, after the evaluation process, they were presented with the most accurate decision tree and the results obtained from it. *IF-THEN* rules are the manner to illustrate our results of decision tree due to easily interpret advantage and convert to understandable.

Figure 8 illustrates the whole process of data mining and provides a clear description step by step to the methodology of this thesis.

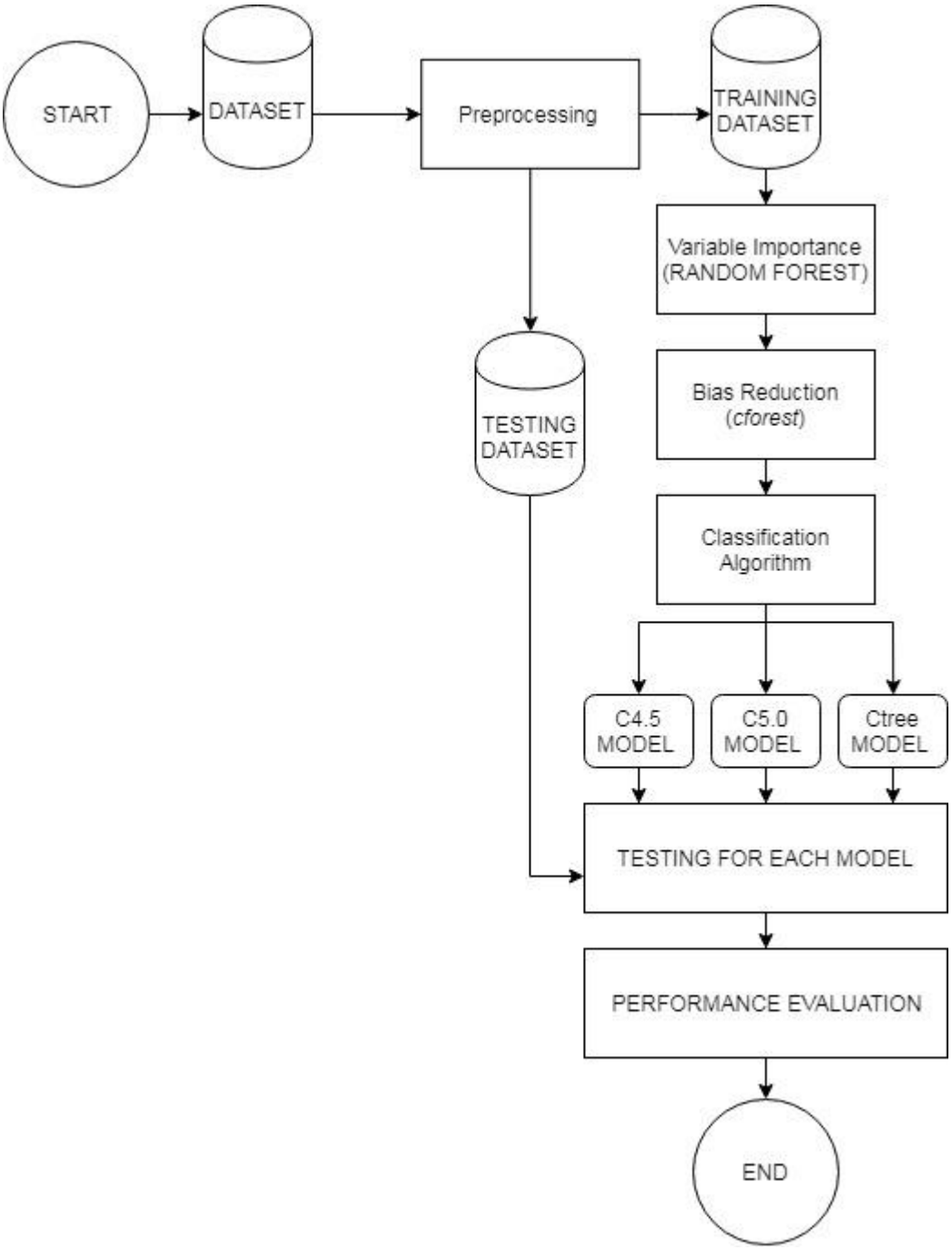


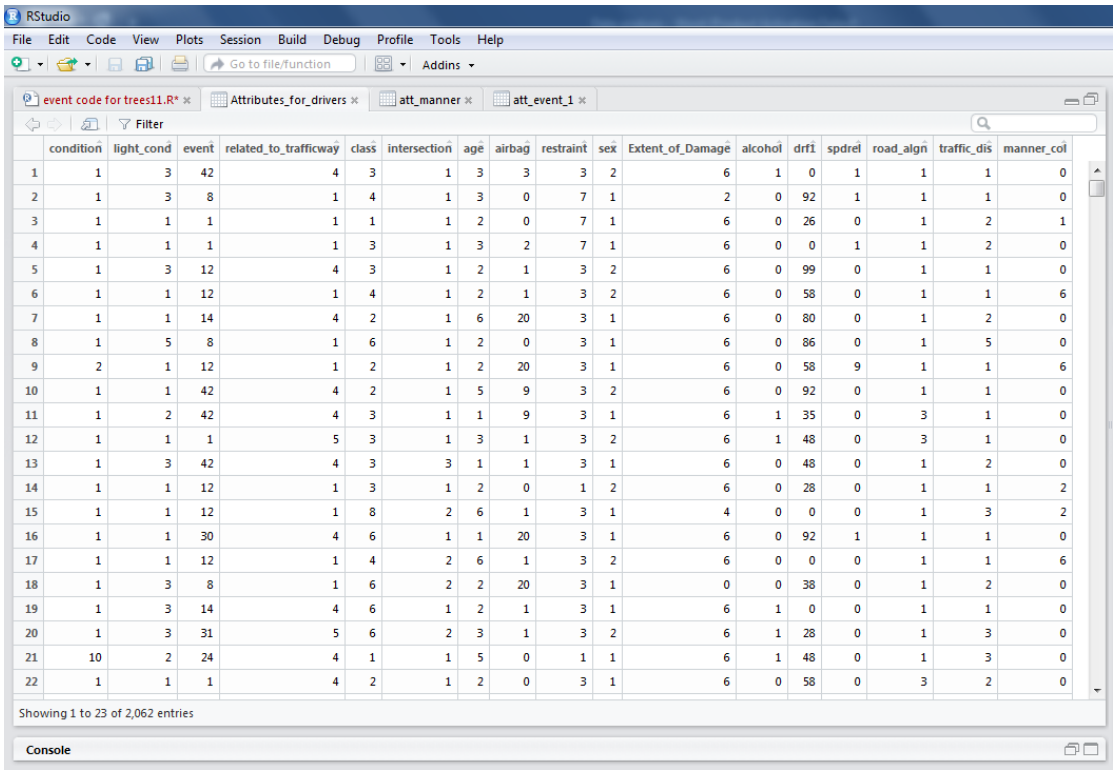
Figure 8: Data mining process proposed in this thesis

3.5 Data Analysis

First of all, it was inserted the seventeen attributes with 2064 cases for a five-year period into Rstudio from Excel file with (.csv) extension after importing the original file from FARS database and then began with the preprocessing step and preparing the data to be analyzed. Since all attributes are presented as numbers and each number

represents a particular category inside the attribute, these categorical attributes should be treated as factors to deal with in the model.

Figure 9 shows a sample of the data that was inserted in Rstudio at the beginning of the analysis process as columns representinmg the attributes and rows representing the cases as integer variables



	condition	light_cond	event	related_to_trafficway	class	intersection	age	airbag	restraint	sex	Extent_of_Damage	alcohol	drf1	spdrel	road_algn	traffic_dis	manner_col
1	1	3	42	4	3	1	3	3	3	2	6	1	0	1	1	1	0
2	1	3	8	1	4	1	3	0	7	1	2	0	92	1	1	1	0
3	1	1	1	1	1	1	2	0	7	1	6	0	26	0	1	2	1
4	1	1	1	1	3	1	3	2	7	1	6	0	0	1	1	2	0
5	1	3	12	4	3	1	2	1	3	2	6	0	99	0	1	1	0
6	1	1	12	1	4	1	2	1	3	2	6	0	58	0	1	1	6
7	1	1	14	4	2	1	6	20	3	1	6	0	80	0	1	2	0
8	1	5	8	1	6	1	2	0	3	1	6	0	86	0	1	5	0
9	2	1	12	1	2	1	2	20	3	1	6	0	58	9	1	1	6
10	1	1	42	4	2	1	5	9	3	2	6	0	92	0	1	1	0
11	1	2	42	4	3	1	1	9	3	1	6	1	35	0	3	1	0
12	1	1	1	5	3	1	3	1	3	2	6	1	48	0	3	1	0
13	1	3	42	4	3	3	1	1	3	1	6	0	48	0	1	2	0
14	1	1	12	1	3	1	2	0	1	2	6	0	28	0	1	1	2
15	1	1	12	1	8	2	6	1	3	1	4	0	0	0	1	3	2
16	1	1	30	4	6	1	1	20	3	1	6	0	92	1	1	1	0
17	1	1	12	1	4	2	6	1	3	2	6	0	0	0	1	1	6
18	1	3	8	1	6	2	2	20	3	1	0	0	38	0	1	2	0
19	1	3	14	4	6	1	2	1	3	1	6	1	0	0	1	1	0
20	1	3	31	5	6	2	3	1	3	2	6	1	28	0	1	3	0
21	10	2	24	4	1	1	5	0	1	1	6	1	48	0	1	3	0
22	1	1	1	4	2	1	2	0	3	1	6	0	58	0	3	2	0

Figure 9: Descriptive states data table

In order to convert the integer variables to categorical variables, the *as.factor* function of R language in RStudio was applied for each category in the attribute and the process was repeated to all other attributes. The following script is a sample of converting a manner of collision attribute to categorical variables.

```
> dat$manner_col<- as.character(dat$manner_col)
> dat$manner_col[dat$manner_col==0] <- 'Not'
```

```
> dat$manner_col[dat$manner_col==1] <- 'front_to_rear'
> dat$manner_col[dat$manner_col==2] <- 'front_to_front'
> dat$manner_col[dat$manner_col==6] <- 'Angle'
> dat$manner_col[dat$manner_col==7] <- 'sideswipe'
> dat$manner_col[dat$manner_col==8] <- 'sideswipe'
> dat$manner_col[dat$manner_col==9] <- 'sideswipe'
> dat$manner_col[dat$manner_col==10] <- 'sideswipe'
> dat$manner_col[dat$manner_col==11] <- 'fron_to_rear'
> dat$manner_col[dat$manner_col==99] <- 'sideswipe'
> dat$manner_col<- as.factor(dat$manner_col)
```

The summary was provided after preparing the dataset in the appropriate way to build the models. Figure 10 shows the attributes with the related categories and the number of cases involved in each category. The figure shows how to get the summary of the data and the results will be illustrated specifically in the next chapter.

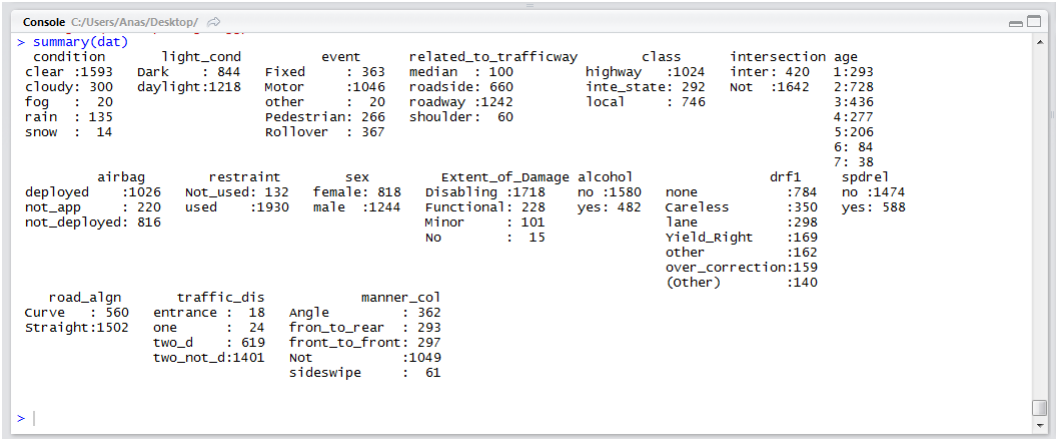


Figure 10: Data summary

For all 2063 records that was obtained from FARS there were 1704 distracted drivers that got involved in a crash in a state that does not have a ban for hand-held cell phone

use which means there was no ban for using cell phone for 82% of those fatalities. Therefore, those states in U.S. that do not have this important ban have to impose it in order to decrease the huge number of fatalities that was caused by cell phone distraction while driving.

The dataset should be split into two sets, one for training and the other for testing, therefore, the splitting was done as 70% of the dataset for training and 30% of the dataset for testing, using random forest as variable importance for the manner of collision and the most harmful event as dependent attributes.

In order to use random forest for ranking the variables "randomForest" package should be applied to use randomForest function but first mtry should be specified to get the most accurate results as mentioned before, mtry was optimized according to the least OOB error by tuneRF function in "randomForest" package for the training data based on manner of collision and the most harmful event attributes. The second input is the number of trees and it was shown for these cases that this input has no effects on the results, the same results was noted as many other researches so it was proposed to use the default number which is 20 trees.

Subsequently, the best mtry that was received from the ranking process can take place with randomForest function for the training data. In order to figure out the change in the contribution of each independent attribute according to the dependent attributes during the five-year period, the variable importance was done for each year separately.

The randomForest function provides the Gini importance method by showing the mean decrease of Gini index $\overline{\Delta_j}$ for each attribute. Figure 11 illustrates the variable

importance process and the mean decrease of Gini index for a five-year period and the entire results of the changing in contribution will be illustrated in the next chapter.

```
> modelrandom <- randomForest(manner_col ~ .,data = TrainData ,mtry=4,ntree=20)
There were 50 or more warnings (use warnings() to see the first 50)
>    importance(modelrandom)
                        MeanDecreaseGini
condition                       21.63972
light_cond                      17.44003
event                          274.43294
related_to_trafficway          118.13373
class                           34.48594
intersection                    62.45993
age                             56.18304
airbag                          31.59603
restraint                        6.85465
sex                             18.60265
Extent_of_Damage                25.77209
alcohol                         12.88283
drf1                            95.52587
spdrel                          23.35920
road_algn                       20.26306
traffic_dis                     31.76674
> |
```

Figure 11: The mean decrease of gini index

The purpose of using conditional inference forest cforest is to minimize the bias according to variable importance. Thus, cforest function through party package was utilized for the training dataset based on the dependent attribute for the year of 2015 to get the most important attributes. Since the cforest function does not employ the Gini criterion, the permutation importance $\overline{D}_j$ was selected in cforest function. Figure 4 shows $\overline{D}_j$ values for all independent attributes according to the manner of collision as dependent attributes.

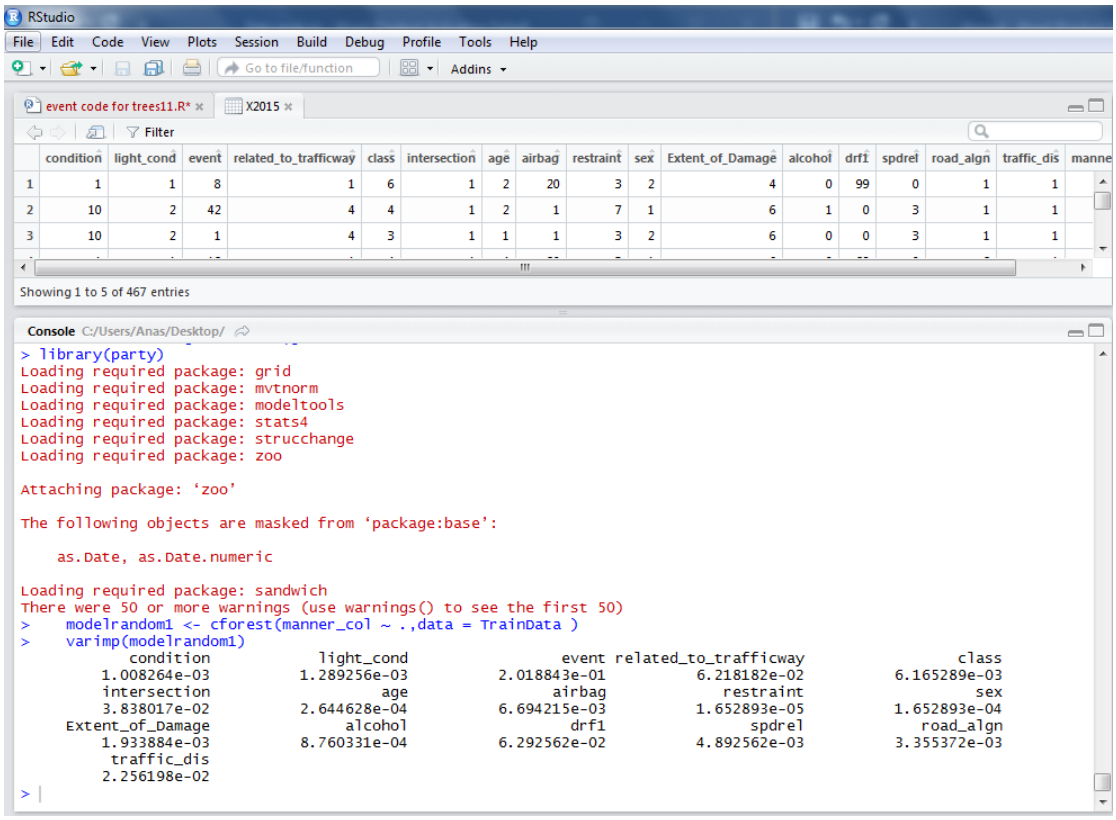


Figure 12: variable importance by cforest

Decision tree models

After selecting the most important attributes for the 2015 year it was implemented three types of C4.5, C5.0 and Ctree decision in order to get the most accurate results from those algorithms for the manner of collision and the most harmful event as dependent attributes. The following section is of coding in RStudio for the three algorithms and their results according to the manner of collision as well as the confusion matrix to check the accuracy of the decision tree.

C4.5 tree algorithm (manner of collision is the dependent variable)

The following figures show the code and the results of applying C4.5 algorithm as well as the accuracy of the tree by the confusion matrix for the testing data.

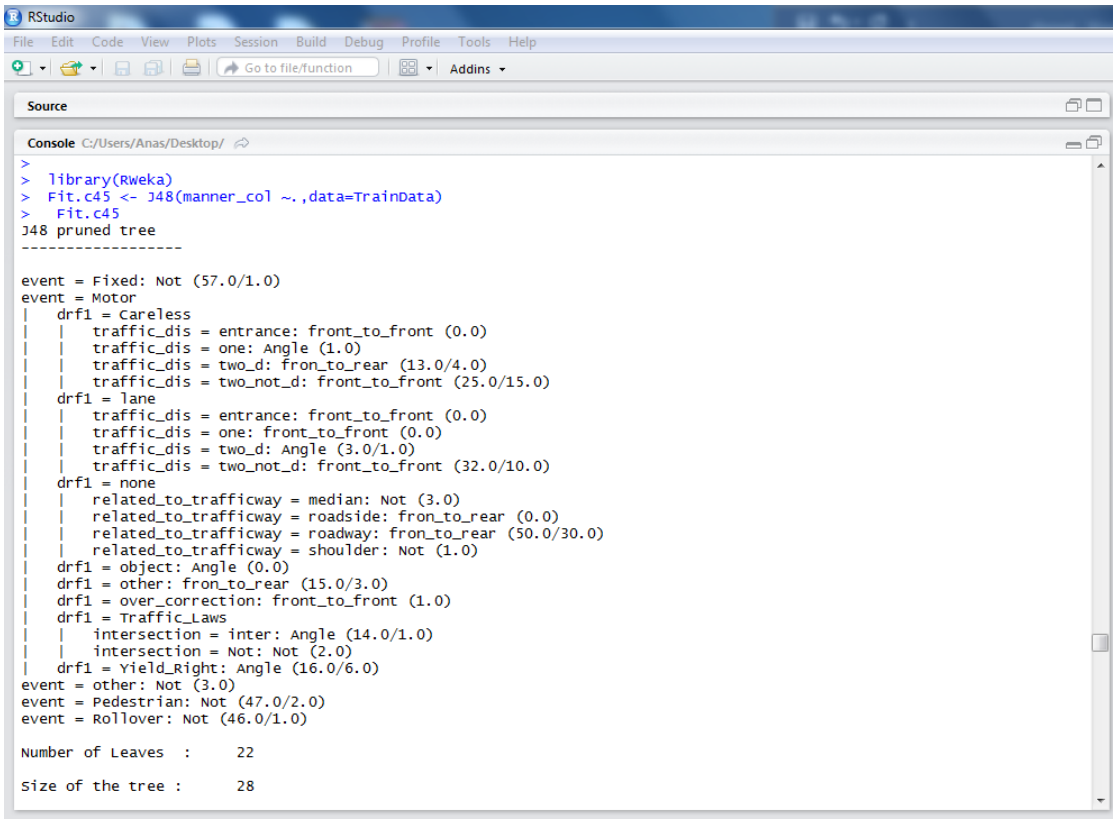


Figure 13: C4.5 tree algorithm (manner of collision is the dependent variable)

Next figure shows the confusion matrix and the accuracy of the decision tree that was applied by Rstudio software.

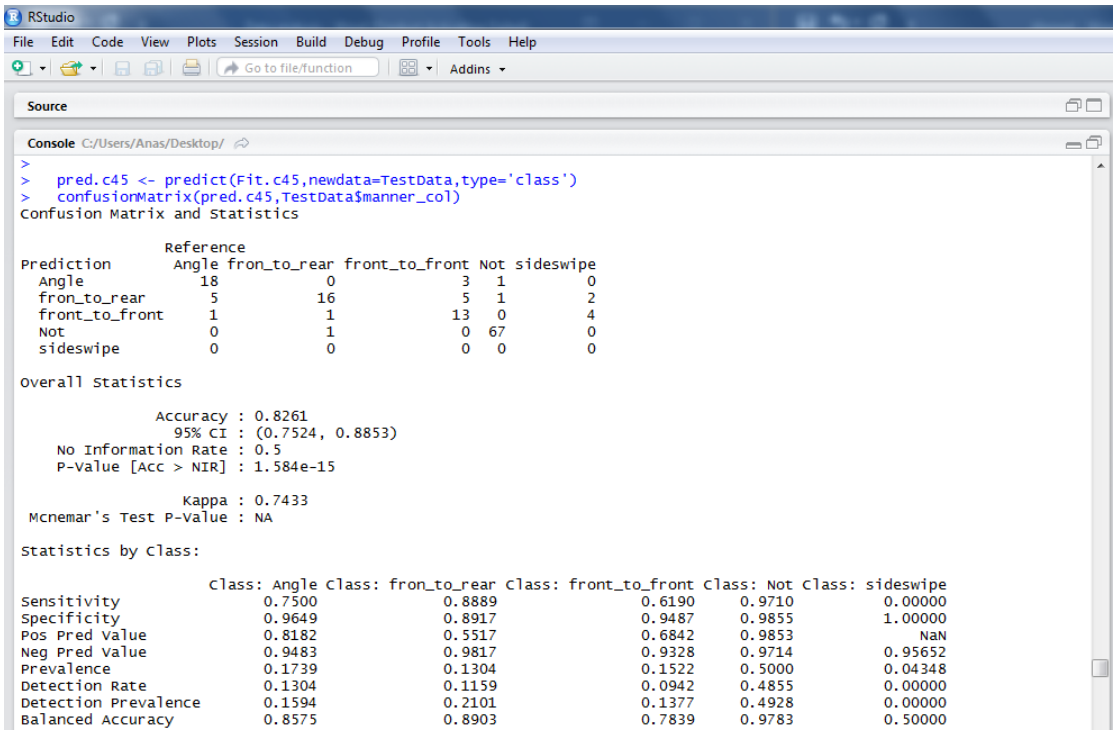


Figure 14: Applying C4.5 tree to the testing data of the manner of collision

Ctree algorithm (manner of collision is the dependent variable)

The following figures show the code and the results of applying Ctree algorithm as well as the accuracy of the tree by the confusion matrix for the testing data.

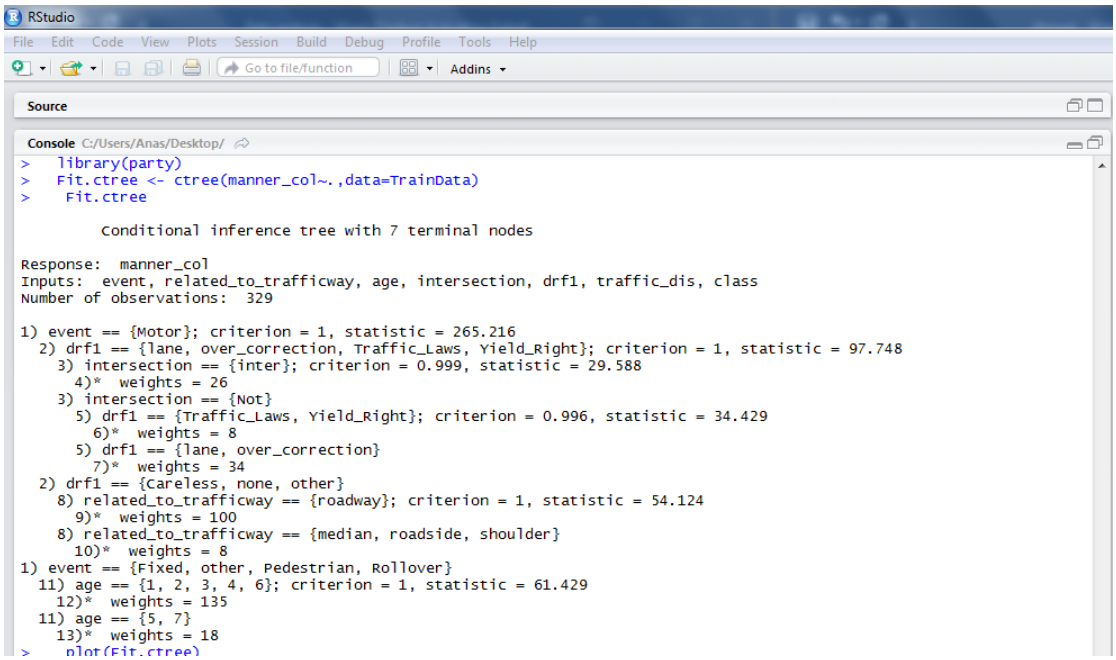


Figure 15: Ctree algorithm (manner of collision is the dependent variable)

Figure 16 shows the results of Ctree as the first leaf provided angular crashes and the last two nodes provide single crashes. In this section it was shown how the process works and in the next chapter the results will be demonstrated particularly.

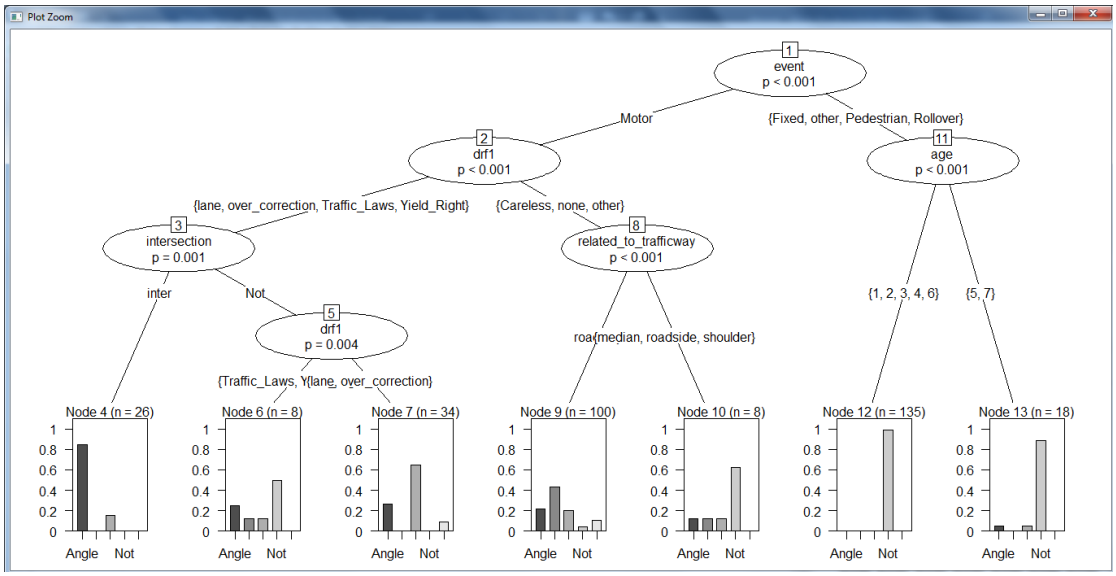


Figure 16: Ctree appearance (manner of collision is the dependent variable)

Next figure shows the confusion matrix and the accuracy of the decision tree that was applied in Rstadio software.

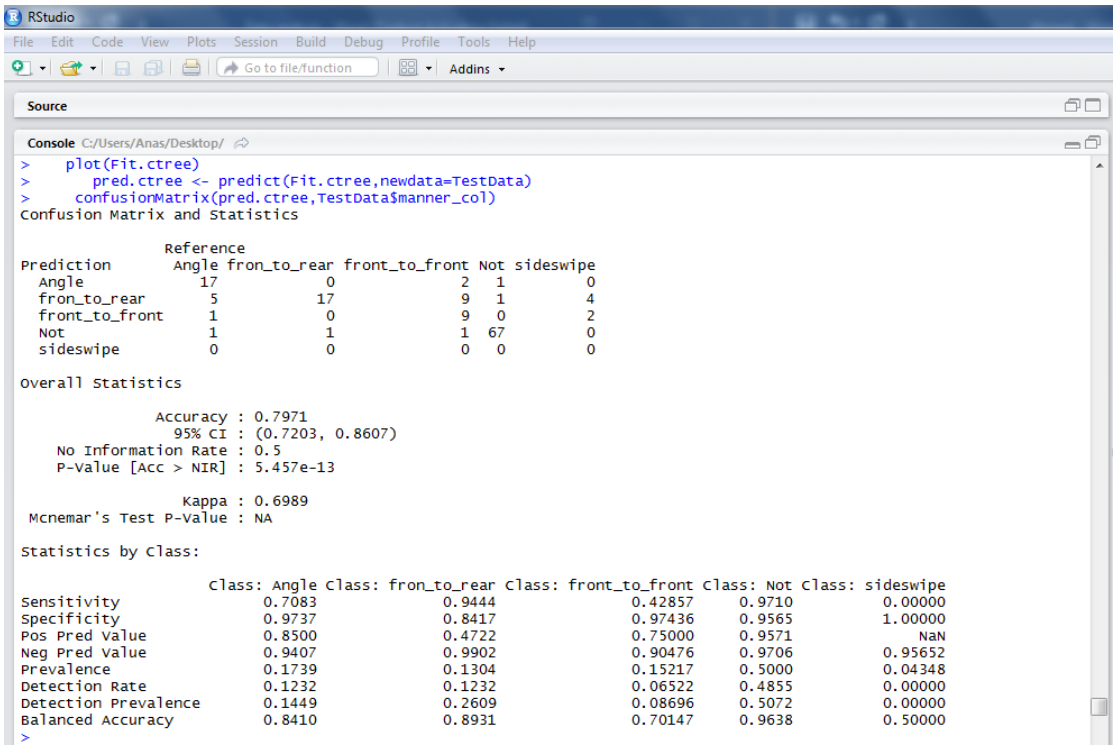


Figure 17: Applying Ctree to the testing data of the manner of collision

C5.0 algorithm (manner of collision is the dependent variable)

The following figures show the code and the results of applying C5.0 algorithm as well as the accuracy of the tree by the confusion matrix for the testing data.

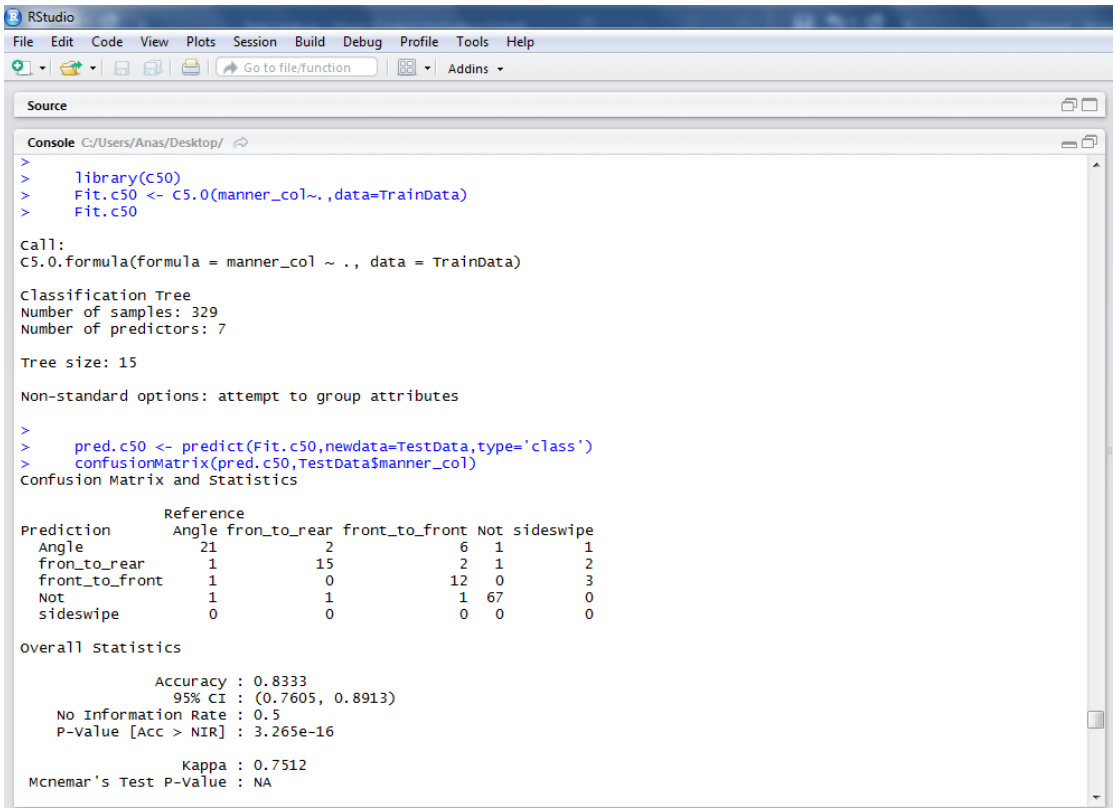


Figure 18: Applying C5.0 tree to the testing data of the manner of collision

Figure 19 shows the application of C5.0 tree algorithm in Rstudio and the structure of this tree with the results.

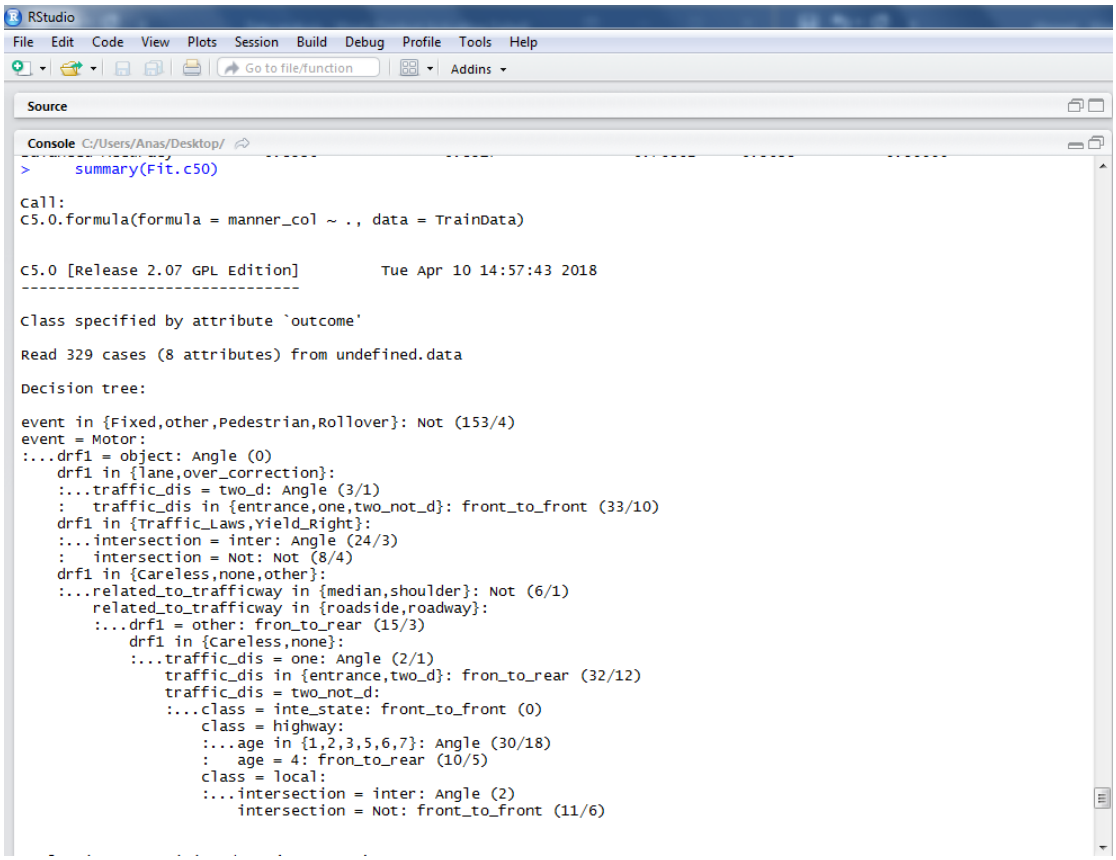


Figure 19: C5.0 algorithm (manner of collision is the dependent variable)

C4.5 tree algorithm (the most harmful event is the dependent variable)

The following figures show the code and the results of applying C4.5 algorithm as well as the accuracy of the tree by the confusion matrix for the testing data.

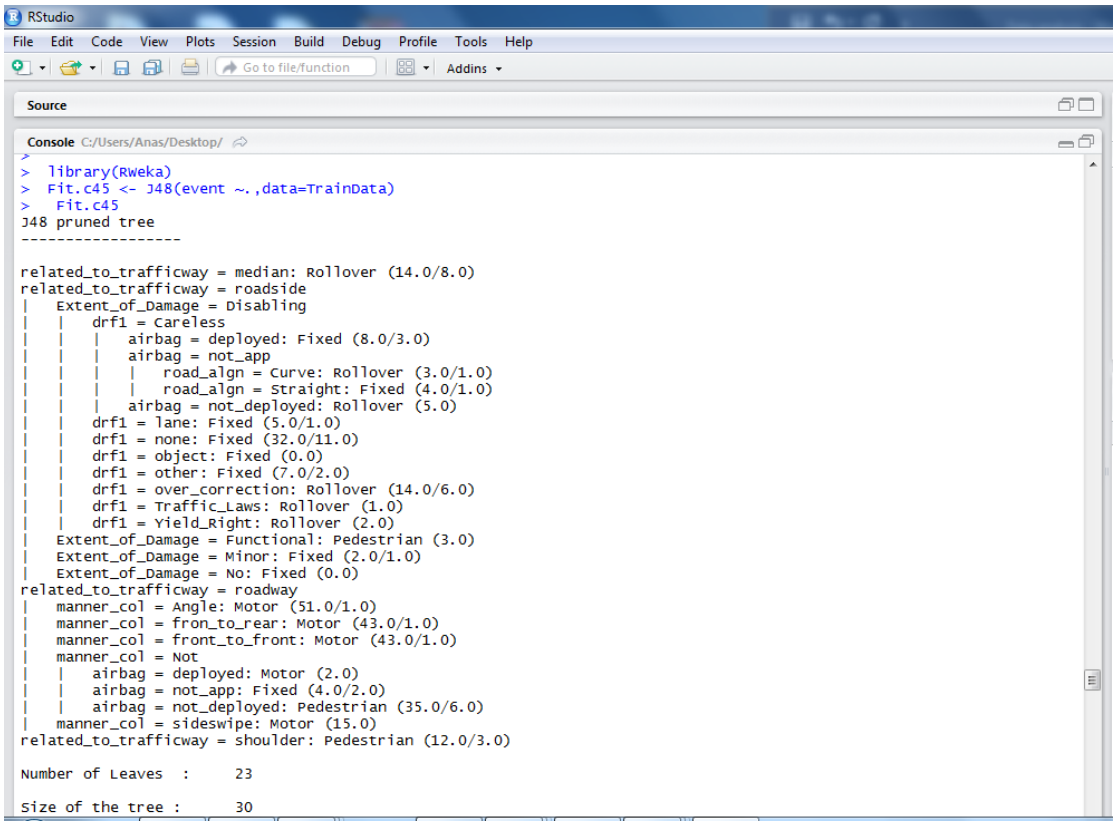


Figure 20: C4.5 tree algorithm (the most harmful event is the dependent variable)

Next figure shows the confusion matrix and the accuracy of the decision tree that was applied by Rstudio software.

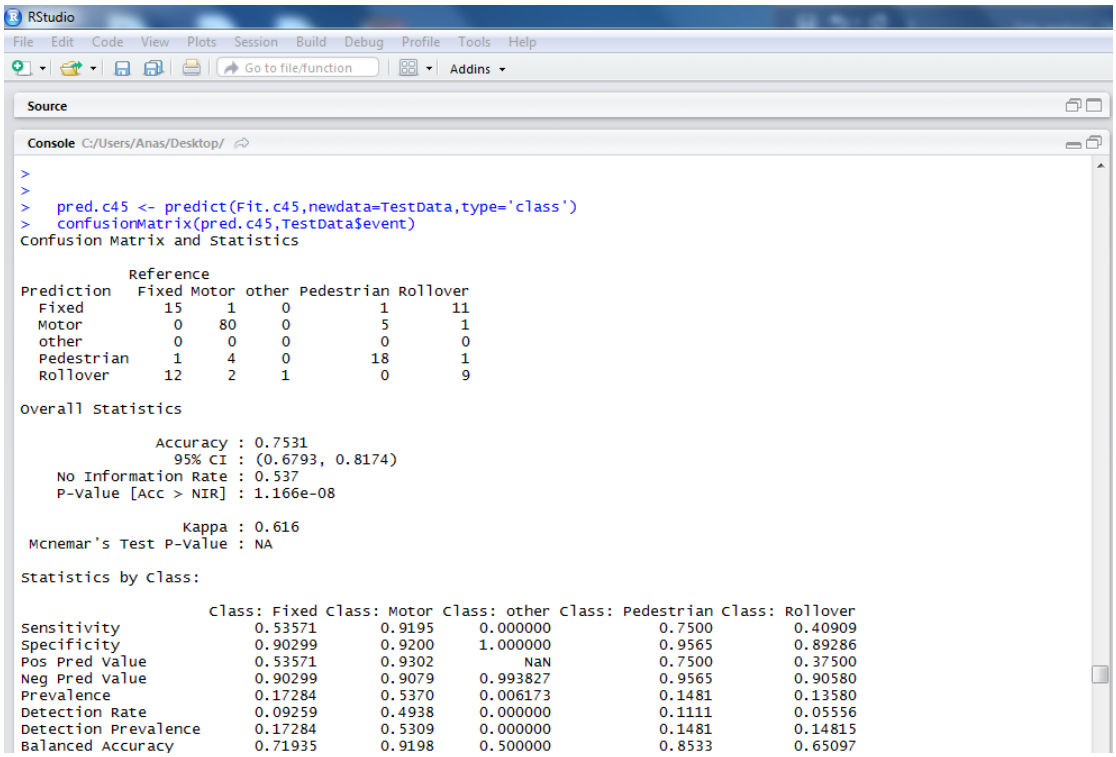


Figure 21: Applying C4.5 tree to the testing data of the most harmful event

Ctree algorithm (the most harmful event is the dependent variable): The following figures show the code and the results of applying Ctree algorithm as well as the accuracy of the tree by the confusion matrix for the testing data.

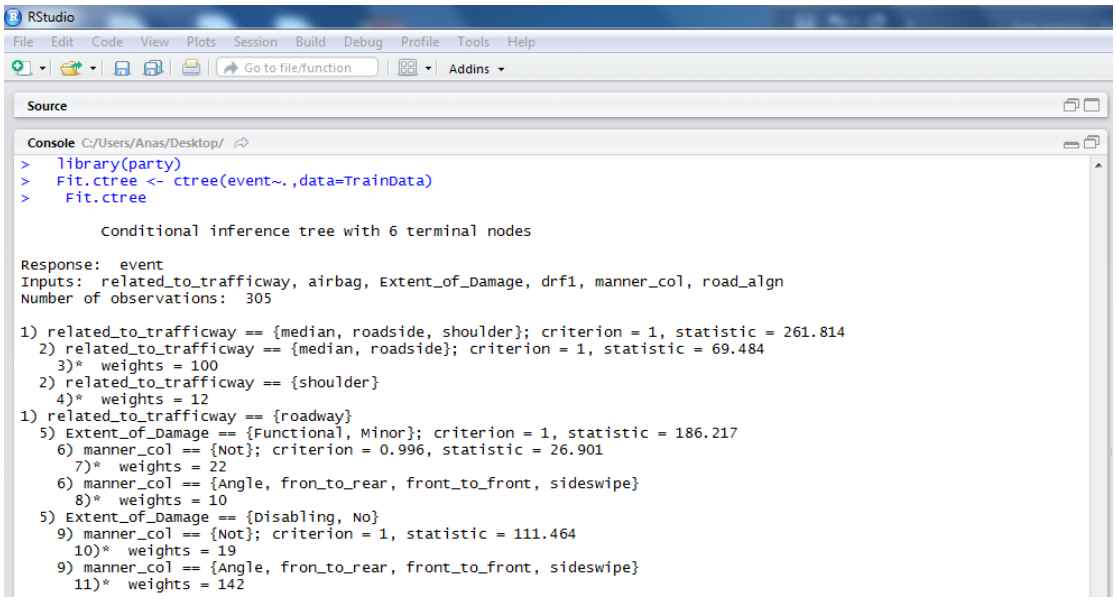


Figure 22:Ctree algorithm (the most harmful event is the dependent variable)

Figure 16 shows the results of Ctree as the first leaf provide angular crashes with 26 cases and the last two nodes provide single crashes. In this section it was shown how the process works and in the next chapter the results will be demonstrated particularly.

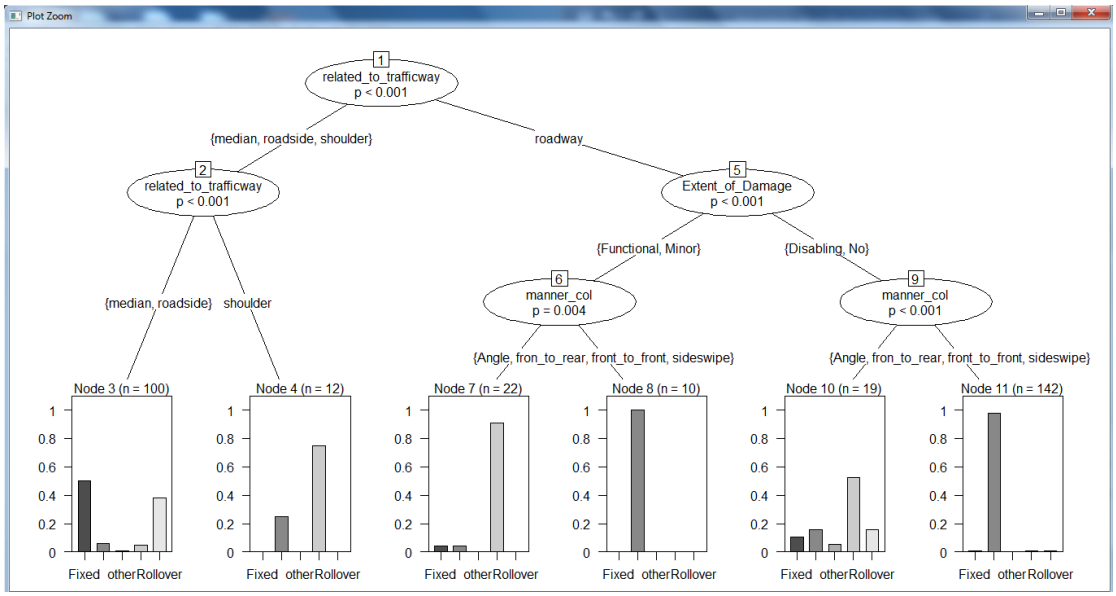


Figure 23: Ctree appearance (the most harmful event is the dependent variable)

Next figure shows the confusion matrix and the accuracy of the decision tree that was applied by Rstudio software.

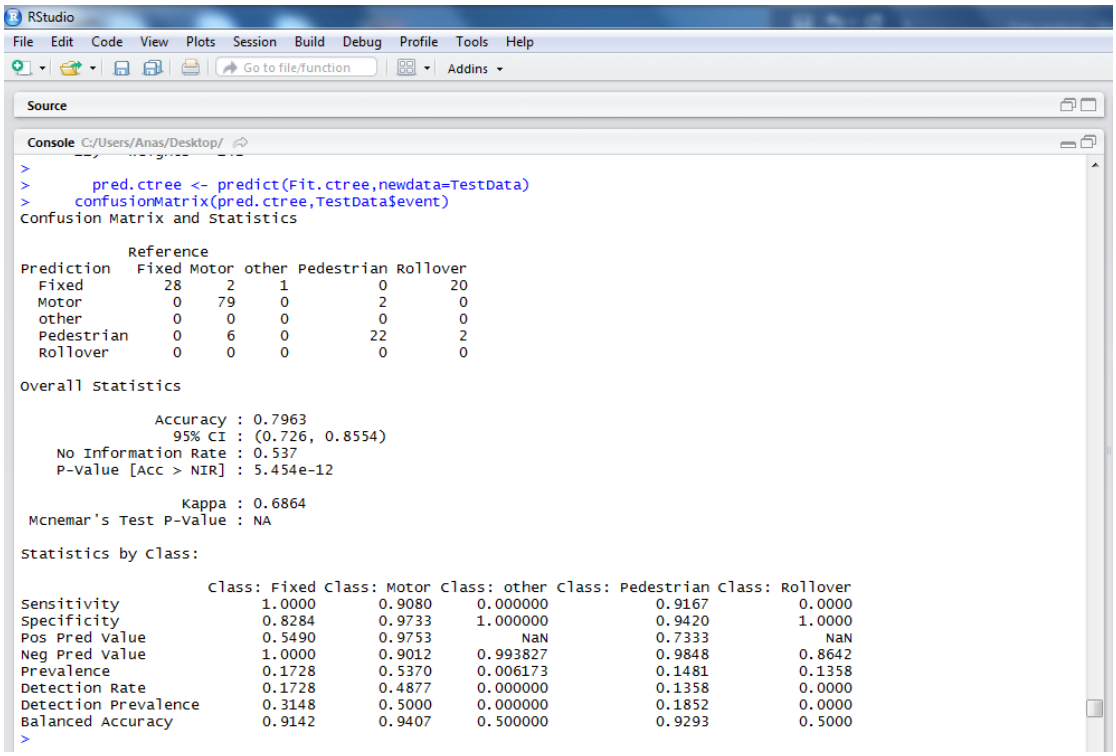


Figure 24: Applying Ctree to the testing data of the most harmful event

C5.0 tree algorithm (the most harmful event is the dependent variable): The following figures show the code and the results of applying C5.0 algorithm as well as the accuracy of the tree by the confusion matrix for the testing data.

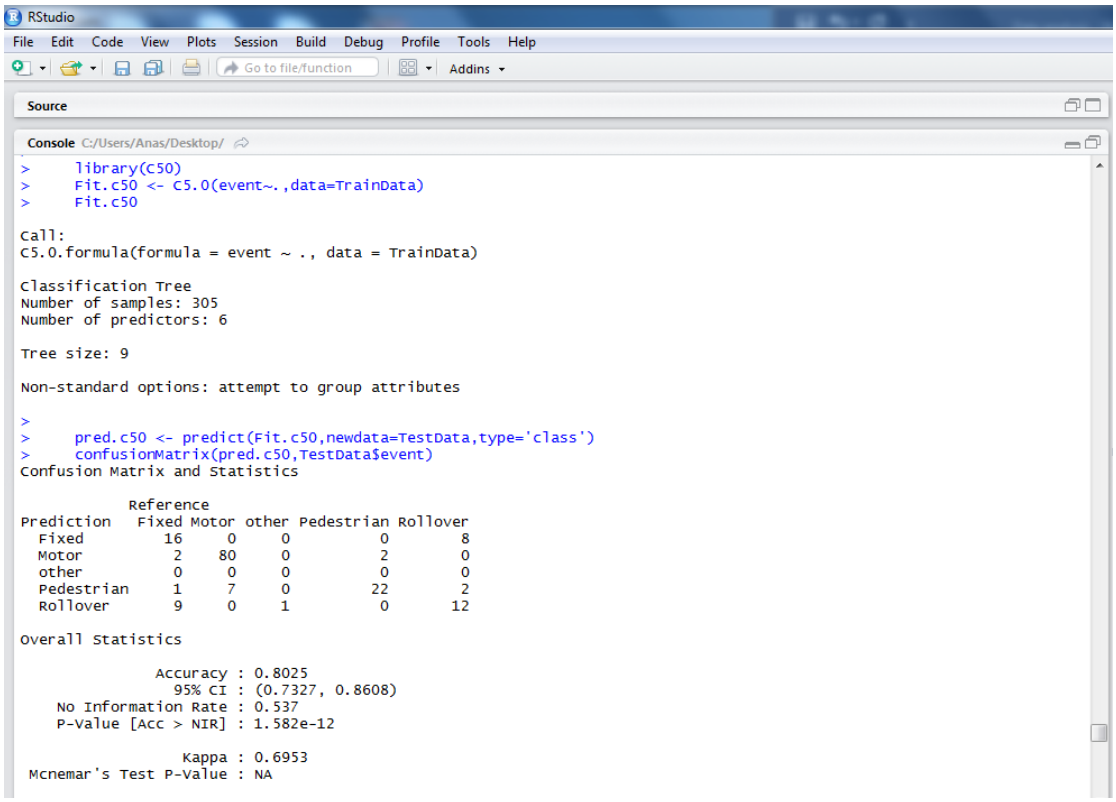


Figure 25: Applying C5.0 tree to the testing data of the most harmful event

Figure 26 illustrates the structure of C5.0 tree when the dependent variable is the most harmful event with the outcomes for each leaf.

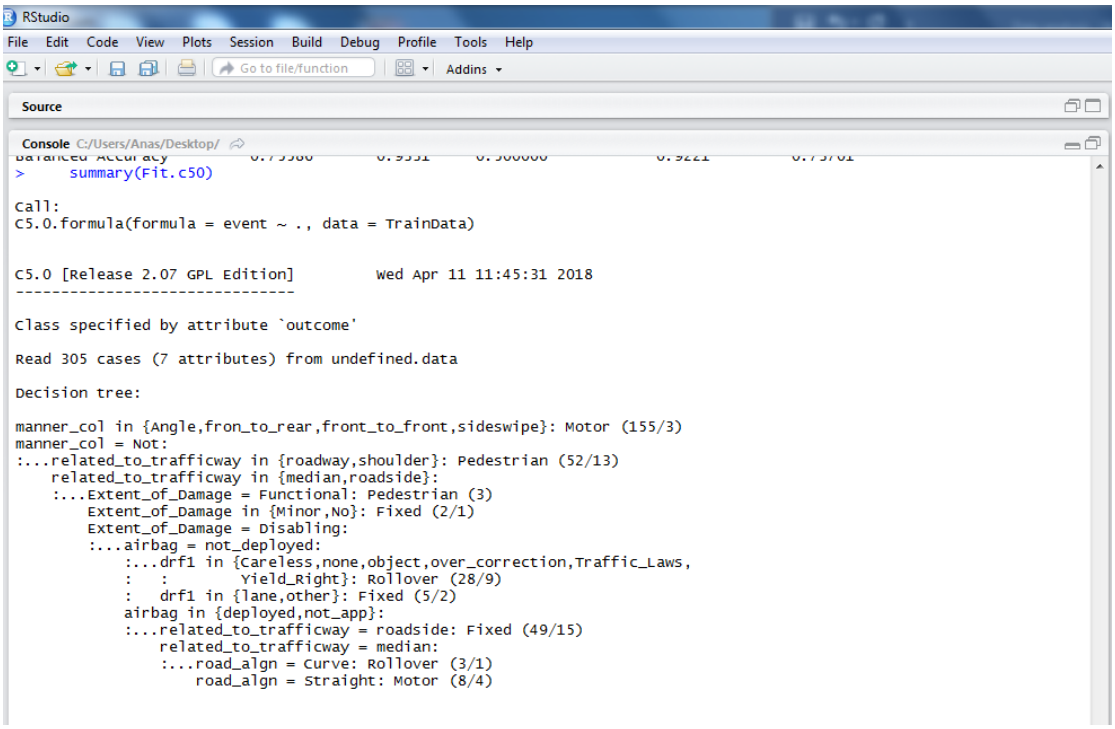


Figure 26: C5.0 tree algorithm (the most harmful event is the dependent variable)

Chapter 4

RESULTS AND DISCUSSION

4.1 Manner of Collision

4.1.1 Using Random Forest for Variable Importance

Random forest as variable importance is conducted in order to explore and investigate the contribution of the seventeen attributes on a distracted driver by cell phone. Additionally, to study the change in this contribution during a five-year period (2011-2015). Furthermore, random forest was applied for the training dataset where the manner of collision is the dependent attribute.

Firstly, due to the significant effect of the number of the chosen attributes in each split (mtry), mtry was optimized by RandomForest function in RStudio instead of increasing and decreasing mtry to get the least OOB error rate. Thus, in order to obtain the most accurate result, least error of OOB was used therefore less OOB error means more accuracy for results. It appeared that when the least OOB error was characterized which equals 0.69%, mtry = 4 and that gave better results with the least error conducted by the optimization as well as more accuracy in choosing the most important attributes.

Table 3 shows the importance of attributes by mean Gini importance which is one of the most accurate methods in random forest as variable importance method. It was

demonstrated by percentage for each year according to the dependent variable which is the manner of collision.

Table 3: Gini importance percentage for attributes

Attribute	2011	2012	2013	2014	2015
DR_CF	13.61	14.72	10.51	12.42	15.08
AGE	7.32	7.07	7.54	6.12	8.62
RELATED_TRAFFIC	15.58	11.58	12.91	10.09	8.39
INTERSECTION	5.78	4.31	8.65	9.48	4.87
CLASS	5.04	3.70	5.22	4.63	4.65
AIRBAG	5.30	3.77	4.80	6.14	4.62
WEATHER	3.47	2.96	2.84	2.91	3.83
TRAFFIC_DIS	2.94	3.93	3.91	3.90	3.61
SEX	2.04	2.91	2.08	1.90	3.41
Light_Cond	2.30	2.61	2.33	1.72	3.23
SPD_REL	2.68	2.63	1.76	3.79	2.83
ALIGNMENT	3.37	3.49	2.26	2.06	2.75
DAMAGE	1.60	1.99	3.09	2.68	2.10
ALCOHOL	2.03	1.85	1.64	1.64	1.46
RESTRAINT	0.61	0.34	0.73	0.92	0.77

From table 2 it can be figured out the significant increase in driver-related factor contribution through five years. This finding indicates how the driver factor rises noticeably year by year especially the significant increment in 2015 by 21% compared with the average of the other years. Thus, driver-related factor has a huge contribution to determine the manner of collision particularly in the year 2015. Previous studies

showed that driver factors were the major causes in 65-75% of road accidents [124, 125].

As it is shown in table 2, DR_CF, AGE and RELATED_TRFFIC have the highest ranks of attributes. Driver alcohol-impairment has one of the least contribution in specifying the manner of collision due to the percentage of drivers' alcohol-impairment being just 30% of all distracted drivers. Gender has an insignificant contribution that means cell phone distraction happens whether the driver is male or female.

One of the noticeable findings is that the speed related factor is insignificant to determine the manner of collision. This finding articulates the results of many researches which illustrated that on distracted driving, adapting slower speed to get more available reaction time is the most common pattern [126]. Compensation for the reaction time and providing control over their driving performance are the main purpose of this strategy for the distracted drivers. Moreover, drivers show greater variability of speed and adjust their control of the vehicle when they are talking on the cell phone that was showed by many studies [127-129].

Therefore, compensatory behavior is the correct expression that explains the behavior of the distracted driver when the driving behavior is adjusted by them when they are using the cell phone in order to be able to perform this additional task and increase the control while doing this extra task. When drivers have a tendency to drive at greater distances of headway and with slower speed and sometimes more speed variation are the most understandable cases of a compensatory behavior [45].

4.1.2 Decision Tree Model

The year of 2015 was chosen, due to having the highest number of fatalities in motor vehicle crashes in which the driver was distracted by cell phone among the five-year period to investigate the relationships between all chosen attributes.

In order to avoid misclassification rate, feature selection by random forest has been conducted with all the algorithms to determine the relevant attributes for the classification. According to variable importance by random forest, the most important attributes were picked to be RELATED_TRAFFIC, DR_CF, AGE, INTERSECTION, TRAFFIC_DIS and CLASS as independent variables and MAN_COLL as the dependent variable for the decision model.

In order to reduce bias as much as possible, the importance of chosen attributes was checked again by cforest in party package. Figure 27 shows the importance percentage for each attribute in 2015 which means the attribute with the highest percentage is the most significant attribute according to the dependent attribute.

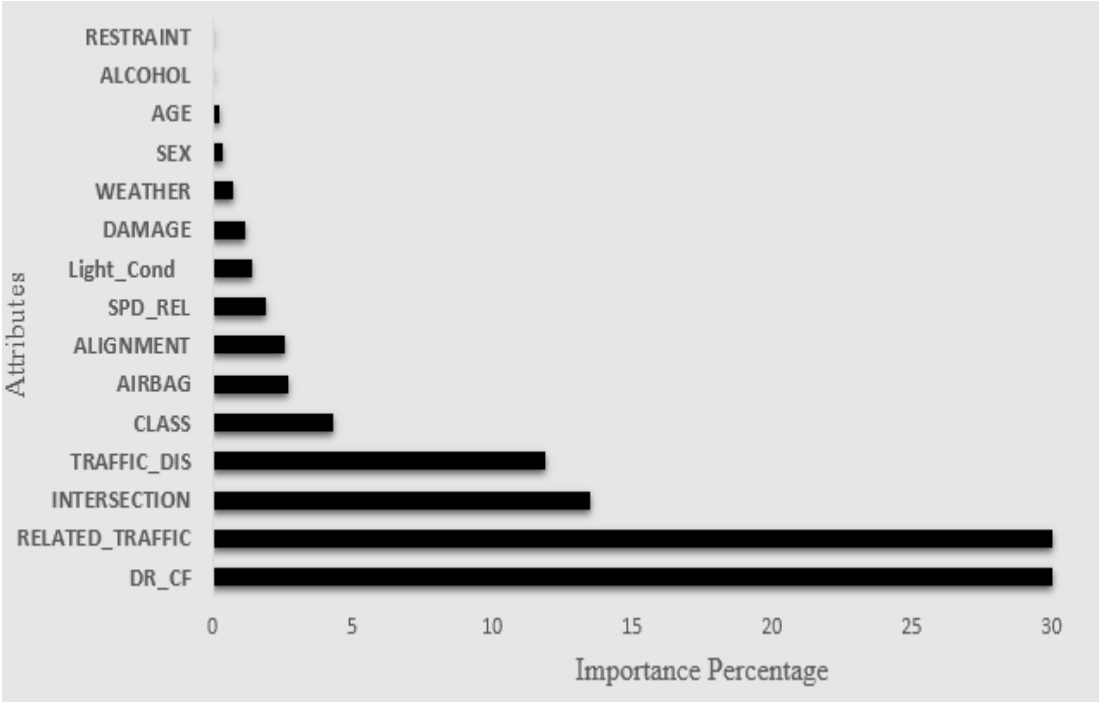


Figure 27: Importance percentage for the attributes (the manner of collision is the dependent attribute)

As figure 27 shows the same results for the most important attributes except for AGE. Therefore, it was excluded and this explanation illustrates the validity of [115] to reduce the bias because AGE includes seven categories thereby it got more rank in importance variable in random forest.

Thus, the final selected attributes were RELATED_TRAFFIC, DR_CF, INTERSECTION, TRAFFIC_DIS and CLASS as independent variables and MAN_COLL as the dependent variable.

In order to investigate the significant relationships between the chosen attributes and the manner of collision, the data set is analyzed using C4.5, C5.0, and Ctree Decision tree algorithms by having MAN_COLL as the dependent variable and all others were set as independent variables. Accuracy is measured using confusion matrix for the three models.

4.1.3 Model Interpretation

It was discussed the results obtained from the classification models. The following if-then rules show some of the interesting rules derived from the C4.5, C5.0 and Ctree models**:**

Rule 1: IF the driver related factor is carelessness or improper lane usage and the traffic description is two lanes not divided

THEN the manner of collision is Front to Front.

Rule 2: IF the driver related factor is carelessness or impaired lane usage and the traffic description is two lanes divided

THEN the manner of collision is Front to Rear.

Rule 3: IF the crash occurs in the roadway and the road class is highway

THEN the manner of collision is Angle.

Rule 4: IF the crash occurs in the roadway and the road class is local

THEN the manner of collision is Front to Front.

Rule 5: IF at an intersection, the distracted driver faild to yield the right-of-way or break the traffic law

THEN the manner of collision is Angle.

Rule 6: IF the distracted driver broke the right of traffic or broke the traffic law and there is no intersection

THEN the manner of collision is Not with a motor vehicle.

Rule 7: IF the distracted driver is careless or has improper lane usage and the crash is not on the roadway

THEN the manner of collision is Not with a motor vehicle.

The models show significant and detailed relationships between drivers distracted by cell phone and manner of collision to portray the manner of collision according to many attributes with mentioning the driver related factors and how these factors are capable to affect the manner of collision.

One of the most significant results that was derived from these models is that if a driver is careless or has improper lane usage, the traffic description determines the manner of collision. Thus, if the roadway is not divided two-lane, they are more likely to be involved in Front to Front crash or if the roadway is divided two-lane, they are more likely to be involved in Front to Rear crash.

Thus drivers who use cell phones frequently will have a challenging time attending to the trafficway since they are busy with additional task and more concentrated on using their cell phone and hence may have a problem with lane keeping to maintain their position in the lane if they are in a not divided roadway. This corresponds with the findings that teenage drivers using cell phones are more likely to be associated with Front to Rear crashes as well [130].

In addition, if the roadway is not divided then the careless drivers who use cell phones are more likely to make improper overtaking through the opposite roadway and consequently face vehicles in that roadway to make Front to Front crashes. This differs from the findings of Ghazizadeh and Boyle [98] who showed that improper ability of lane-keeping and decision making are the main causes to be involved in an angular crashes.

Neither of previous two studies [98, 130] mentions the role of intersections in investigating the manner of collision for cell phone distracted drivers. In this study, it was found that if those drivers fail to yield the right-of-way or break the traffic law at intersections then they are more likely to be associated with angle crashes.

Drivers educated of the intersection safety that further steps can be taken to prevent collisions and their further caution can lead to less vehicle to vehicle collisions and fewer fatalities. Indeed, the safety in intersections is undervalued by many drivers that they drive every day and they are not aware of the characteristics of intersections that more precaution should be taken by drivers due to more possibility of involving in a collision with deadly consequence.

In general, drivers are more likely to be in crashes not with moving vehicles like colliding with fixed objects if there is no intersection and they are careless and fail to yield right-of-way and break traffic laws while driving.

These rules demonstrate the capability of data mining to show the complex relationships among crash, traffic and road variables.

4.1.4 Performance Evaluation

The performance of the obtained models was evaluated using confusion matrix by investigating accuracy for each model. As mentioned previously, the accuracy computed by

$$\text{Accuracy} = (TP + TN)/ (TP + TN + FP + FN) \quad (6)$$

Overall accuracy results show that all the models have similar performance and the most accurate model was C5.0 tree with an estimated accuracy of the model being 83.3

%. The confusion matrix for this model for testing data for model assessment of actual and predicted outcomes of manner of collision is illustrated in table 4. 97 % of the not with motor vehicle crashes, 83.3% of the Front to Rear crashes, 83% of Front to Front, 87.5% of the angle crashes and without any prediction of the Sideswipe were correctly predicted. The lower accuracy rate for Sideswipe crashes may be due to the limited number of observed cases.

Table 4: Confusion matrix for manner of collision

	Reference				
Prediction	Angle	front_to_rear	front_to_front	Not	sideswipe
Angle	21	2	6	1	1
front_to_rear	1	15	2	1	2
front_to_front	1	0	12	0	3
Not	1	1	1	67	0
sideswipe	0	0	0	0	0

Table 6 demonstrates the accuracy applying the testing subset to test the three classification models. The accuracy of the three models shows privileged accuracy than models in a previous study where the drivers were disracted. Tseng et al. illustrate that the accurate prediction percentage for front to rear collision is 78.31% and 77.35 % for angle collision[97].

Table 5: Accuracy comparison for the three models with the manner of collision as the dependent attribute

model	Accuracy
C4.5	82.61%
C5.0	83.33%
Ctree	79.71%

4.2 The Most Harmful Event

4.2.1 Using Random Forest for Variable Ranking

Random forest was used with the purpose of examining the involvement of the seventeen attributes on cell phone distracted drivers. In addition, study the change in this involvement during a five-year period (2011-2015). Therefore, random forest was applied for the training dataset and the most harmful event being the dependent variable.

Firstly, due to the significant effect of the number of the chosen attributes in each split (mtry), mtry was optimized by RandomForest function in RStudio instead of increasing and decreasing mtry to get the least OOB error rate. Thus, in order to obtain the most accurate result, least error of OOB was used therefore less OOB error means more accuracy for results.

It appeared that when the least OOB error was characterized which equals 1.9%, mtry = 4 and that gave better results with the least error conducted by the optimization as well as more accuracy in choosing the most important attributes.

Table 7 shows the importance of attributes by mean Gini importance which is one of the most accurate methods in random forest as variable importance method. It was demonstrated by percentage for each year according to the dependent variable which is the most harmful event.

Table 6: Gini importance percentage for attributes

	2011	2012	2013	2014	2015
RELATED_TRAFFIC	15.84	21.01	18.67	19.39	23.92
AIRBAG	8.13	7.82	5.90	8.29	9.76
DR_CF	12.39	10.31	10.76	9.31	7.36
DAMAGE	4.18	7.66	6.38	7.24	6.54
AGE	8.99	8.76	7.35	7.06	5.90
CLASS	5.27	3.06	4.88	3.26	3.33
Light_Cond	2.83	2.76	3.43	2.24	3.01
TRAFFIC_DIS	2.75	2.02	1.89	2.39	2.55
SPD_REL	2.03	2.16	2.25	3.20	2.47
ALIGNMENT	4.15	1.63	2.51	3.27	2.47
WEATHER	2.63	2.40	3.31	2.87	2.31
ALCOHOL	2.88	3.14	1.70	3.08	2.29
SEX	2.74	1.73	2.89	2.85	2.02
RESTRAINT	1.41	1.11	1.08	0.85	1.53
INTERSECTION	2.20	1.13	2.49	1.54	1.04

From table 6 it is noticed that the significant increase in the involvement of the location of the crash either on the roadway, roadside, shoulder or median in determining the most harmful event through five years. This result indicates how the relation to trafficway involvement gets higher clearly year by year particularly the major increment in 2015 by 27% compared with the average of the other years. Thus, the relation to trafficway where the crash happens has a vast involvement to determine the most harmful event.

As it is shown in Table 2 related to traffic way and airbag deployment have the highest ranks among attributes. The presence of an intersection has the least contribution in identifying the most harmful event. The same in determining the manner of collision the gender, speed-related factor and driver alcohol-impairment have unimportant involvement in determining the most harmful event.

4.2.2 Decision Tree Model

Same for the manner of collision, decision tree models used the 2015 dataset. The reason for this selection was that the 2015 dataset has the highest number of fatality crashes among the drivers distracted by cell phone among the five-year period. Therefore, to achieve the main purpose which is investigating the relationships between all chosen attributes.

Once more, in order to avoid a misclassification rate, feature selection by random forest has been involved into all the algorithms to select the features that are relevant for the process of classification. According to variable importance by random forest the most important attributes were picked which are, RELATED_TRAFFIC, AIRBAG, DR_CF, AGE and DAMAGE as independent variables and the most harmful event as the dependent variable for the decision model.

So, to reduce bias as much as possible the importance of chosen attributes was verified again by cforest in *party* package. Figure 28 shows the importance percentage for each attribute in 2015 which means the attribute with the highest percentage is the most significant attribute according to the dependent attribute.

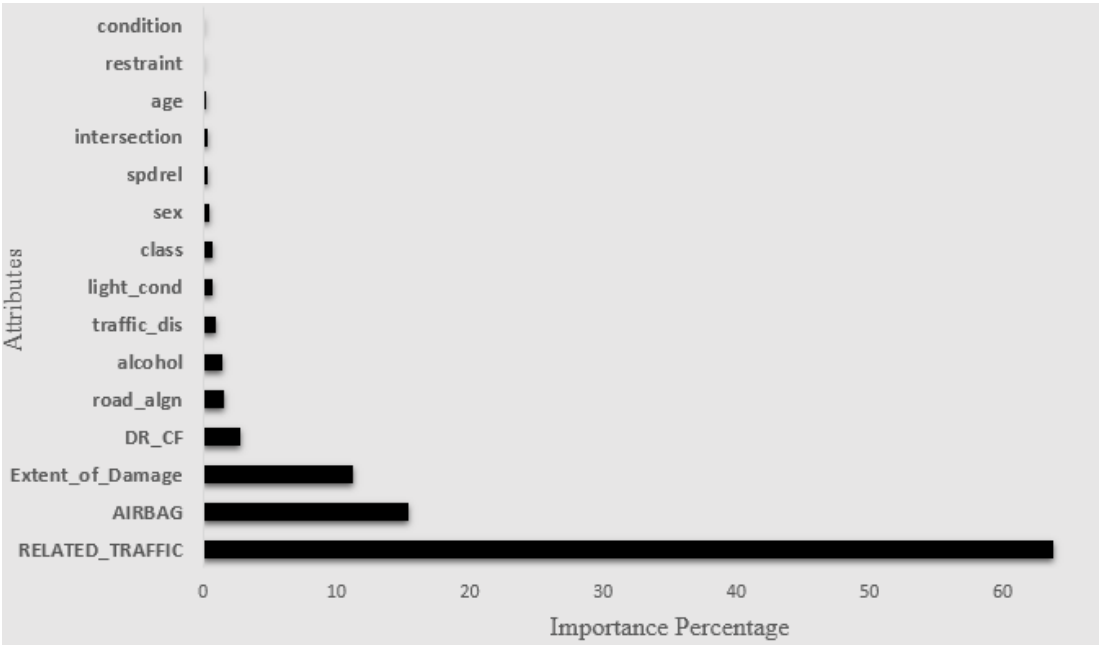


Figure 28: Importance percentage for the attributes (the most harmful event is the dependent attribute)

As figure 28 shows the same results to pick the most important attributes but instead of AGE the attribute ALIGNMENT took place so the final selected attributes were RELATED_TRAFFIC, AIRBAG, DR_CF, DAMAGE and ALIGNMNET as independent variables and the most harmful event as the dependent variable.

The data set is analyzed using C4.5, C5.0, and Ctree decision tree algorithms by having EVENT as the dependent variable and all others were set as independent variables. Accuracy is measured using confusion matrix.

4.2.3 Model Interpretation

The results obtained from the classification models was discussed. The following if-then rules show some of the interesting rules derived from the C4.5, C5.0 and Ctree models:

Rule 1: IF the driver gets involved in a single crash (not with a vehicle in transport) and the distracted driver deviate from the road to the shoulder

THEN the most harmful event is hitting a pedestrian.

Rule 2: IF the driver gets involved in a single crash (not with a motor vehicle in transport) and the distracted driver deviate to the roadside and the vehicle has disabling damage from the crash and the airbag system is deployed

THEN the most harmful event is hitting a fixed object.

Rule 3: IF the road alignment is curved and the driver is careless and deviate to the median and the vehicle has a disabling damage

THEN the most harmful event is Rollover.

Rule 4: IF the road alignment is straight and the driver is careless and has improper lane usage the vehicle has a disabling damage with a deployed airbag system

THEN the most harmful event is hitting motor vehicle in the transport.

Rule 5: IF the distracted driver deviate to the median or roadside and the driver related factor is an overcorrection

THEN the most harmful event is Rollover.

Rule 6: IF the distracted driver deviate to the roadside and the driver is careless and there is minor or functional damage to the vehicle without airbag deployment

THEN the most harmful event is hitting a pedestrian.

The models show major and comprehensive relationships between cell phone distracted drivers and the most harmful event by using the classification algorithms to explore the relationships between the different categories of the most harmful events and when the drivers distracted by cell phone according to many attributes.

One of the most significant results that were derived from these models is if a driver is careless or has improper lane use, the road alignment determines the most harmful event. Thus, if the road alignment is curved, they are more likely to be involved in

rollover when the distracted driver deviates to the median or if the road alignment is straight, he or she is more likely to be involved in hitting motor vehicle with airbag deployment and disabling damage of the vehicle.

So, if there is a curved road, the distracted drivers who utilize cell phones repeatedly will not have an adequate time attending to the trafficway. As the distracted drivers are busy and more concentrated on using their cell phone and hence when they face curved road deviate to the median and rollover with disabling damage and airbag deployment in the vehicle which indicates more severe crashes and more fatalities.

Additionally, one of the significant findings that was obtained is regarding the pedestrian issue which is one of the most significant concerns for the safety of traffic. On behalf of pedestrians, the roadside and shoulder are very dangerous to their lives if there is a driver distracted by cell phone because of the carelessness of the distracted driver and failure to avoid deviation from the road, thereby losing attention to the presence of a pedestrian on the roadside.

A fundamental finding in our research is that hitting fixed objects is more hazardous than hitting motor vehicle and getting involved in a fatal crash and is associated with disabling damage in the involved vehicle. When the distracted driver deviates from the road to the roadside and hits objects like trees, barriers, concrete, utility poles, signs, and guardrails the crash will be dangerous to the driver and the occupants.

Our findings correspond to a previous study which illustrates that collisions with roadside objects have a higher risk to be involved in a fatal collision than with another motor vehicle or fixed objects. In an investigation into the most harmful event,

Danielloa [131] found that reported in the crash, being involved with collisions that hit with guardrail were 7 times more likely to be involved with a fatal collision than without a fixed object present. Furthermore, 15 times more likely to be involved in a fatal collisions if the collision was with trees than without a fixed object present.

As a result, the most harmful event is hitting a fixed object either in the roadside or not as previously reported by a study which indicates that get involved in a crash with a fixed object is more harmful than getting involved with other types of crashes [131].

In this thesis, light was shed on a new approach in the causing of overcorrection and linked over overcorrection with cell phone distraction as the main cause, especially with prevalent cell phone use while driving these days. Therefore, if the distracted driver is careless and deviates from the road and figures out this deviation belatedly and tries to come back to the lane by overcorrection (oversteering) that leads to rollovers.

In one of the most recent studies [132] discussing the main causes of overcorrection, the authors illustrated that when drivers are ill they are more likely to be in overcorrecting or oversteering event by 2.22 times and when they under fatigue they are more likely to be in overcorrecting or oversteering event by 3.44 times. Moreover, when drivers fallen asleep they are more likely to be in overcorrecting or oversteering event when they fallen asleep by 1.61 times comparing with normal conditions. Consequently, there is no mention of cell phone distraction while driving which proposed a new insight to link overcorrection with this newer cause.

In accordance with the present results, Brookhuis [58] has demonstrated high distraction detected when the driver dialed numbers on the cell phone. In such cases, the steering wheel handling stability decreases and that leads to deviation and the crash being more likely to be in the roadside or the median and subsequently may hit fixed objects or pedestrians on the roadside.

Several previous results [59, 133, 134] showed that dialing a cell phone impairs the ability of driver to keep lateral position on the road properly and maintain consistent speed. The present findings seem to be consistent with those studies and this provides us an understanding of how improper lane usage by distracted drivers has significant outcomes to the manner of collision and the most harmful event. Moreover, inability to sustain consistent speed for cell phone distracted drivers has a disastrous influence on traffic safety and leads to several types of collisions.

As a result, these rules reveal the capability of data mining to show the complex relationships between crash, traffic and road variables and shows distinguishable results between these classification models.

4.2.4 Performance Evaluation

The performance of the obtained models was assessed using confusion matrix by investigating accuracy for each model.

Overall accuracy results show that all the models have significant performance and the most accurate model was C5.0 tree with estimated accuracy of the model being 80.25%. The confusion matrix for this model for the testing data for model assessment of actual and predicted outcomes of the most harmful event is illustrated in table 8. 92% of crashes with motor vehicle in motion, 91% of hitting pedestrian, 55% of

rollover, 54% of hitting fixed objects and without any prediction of the others were correctly predicted. The lower accuracy rates for the *others* category may be due to few observed cases being considered *others*.

Table 7: The confusion matrix for the most harmful event

	Reference				
Prediction	Fixed	Motor	others	Pedestrian	Rollover
Fixed	16	0	0	0	8
Motor	2	80	0	2	0
others	0	0	0	0	0
Pedestrian	1	7	0	22	2
Rollover	9	0	1	0	12

Table 9 demonstrates the accuracy of applying the testing subset which is the second subset to test the classification models. Table 10 demonstrates the accuracy applying the second subset of the data to test the three classification models.

Table 8: Accuracy comparison for the three models with the most harmful event as the dependent attribute

model	Accuracy
C4.5	75.31%
C5.0	80.25%
Ctree	79.63%

Chapter 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusion

This thesis was focused on the detection of relationships between motor vehicle crashes and drivers distracted by cell phone in the United States. This most up-to-date study will make contributions to the growing literature on distracted driving and specifically, cell phone distraction. While latest research on driving with a distraction effect has focused on driver health and wellness, this thesis investigates their occupational safety for a pervasive risk factor that has been understudied, which is cell phone use while driving and its consequences.

The data-mining tool RStudio was used to explore the data derived from FARS, which is one of the most reliable databases that illustrates real data from the field without needing to implement simulations. Major data mining techniques for crash analysis were conducted in this thesis. The random forest model is used to expose the importance of the input attributes as well as study the change of contribution for the input attributes in the five-year period from 2011 to 2015 for two output attributes which are the manner of collision and the most harmful event.

The decision tree models were then applied by C4.5 tree, C5.0 tree and Ctree algorithms to classify the output attributes of the manner of collision and the most harmful event through the most important input variables determined previously by

random forest and developed by *cforest* in order to reduce the bias as much as possible to get more accuracy in decision tree classification results by the chosen attributes for 2015, the most fatal year for distracted drivers by cell phone shown by FARS till now.

The results show that the driver-related factors' contribution continuously increased in the five-year period to determine the manner of collision and the most harmful event which implies drivers' role when they are distracted by cell phones while driving which affects their lives and leads to tragic consequences. Therefore, the driver related factors' are among the main factors that affect traffic safety and their effect is rising year by year which is emphasized by several findings that show at least sometimes more than two thirds of drivers use a cell phone while driving regardless to their gender, which is shown to have a low contribution on the our outcome [135, 136].

The safety in intersections is undervalued by many drivers that they drive every day and they are not aware of the characteristics of intersections that more precaution should be taken by drivers due to more possibility of involving in a collision with fatal consequence. The results show the dangerous of cell phone use while driving at intersections and illustrate the contribution of the intersection to increase the likelihood to get involved in angle crash. The cause of this role is that the distracted driver certainly fails to yield the right-of-way or break the traffic law.

Additionally, most of the crashes for the distracted driver are not with a motor vehicle in motion due to the inability of maintaining lanes or improper lane change during driving task because during cell phone use, the eyes and mind of driver are off the roadway for extended periods of time. The consequences in our results were very deadly for drivers' lives and disability damage to their vehicles which means more cost

lost when hitting fixed objects, all of these miserable outcomes are due to the carelessness of the distracted drivers.

The results demonstrate that cell phone distraction does not just have dangerous impacts on drivers' lives or vehicles but on pedestrians as well. The results illustrate that the pedestrians on the roadside are more likely to get hit by distracted drivers who deviate from the road, meaning that careless drivers who are using their cell phone while driving threaten pedestrian lives more often. Therefore, this implies the huge impact of this distraction on one of the most significant concerns in traffic safety.

This thesis sheds light on a new cause of overcorrction and subsequent rollover that was not mentioned specifically in previous researches, which is cell phone distraction. So, if the careless distracted driver deviates from the road to the median or is unable to maintain the lane when driving on a curved road and overcorrects that leads to rollovers which have fatal outcomes.

Although there is no big difference in the accuracy for C5.0, C4.5 and Ctree, the highest accuracy percentage was obtained by C5.0 decision tree for the manner of collision and the most harmful event which are 83 and 80 percent, respectively, which both are significant percentages of accuracy in classification for the models.

Algorithm of C5.0 is an extension of the algorithm of C4.5 which applying this algorithm to big set of data as a classification algorithm which indicates that C5.0 tree is more capable to deal with big dataset than C4.5 tree which proposed to be a limitation in C4.5 tree . The efficiency, speed and memory all those properties are better in C5.0 algorithm than C4.5 algorithm [123]. For Ctree algorithm, the

conditional inference trees seem to be more appropriate for specific purposes than others which indicates that it is uncomprehensive [118].

In the end, it is wanted to highlight that the results are geared to diminish the number of collisions of the type of collision that outcomes in the highest number of fatalities because it is very challenging to drop the total number of collisions.

5.2 Limitations

Using the Fatality Analysis Reporting System (FARS) database produces both challenges and great opportunities in the analysis. A first limitation of this database is that it only captures crashes involving fatalities. Crashes involving no injuries or non-fatal injuries are much more common than fatal crashes [16]. In a report for NHTSA [137], Ascone and colleagues find that both FARS and distraction data could be limited in their reliance on police reports for distraction data in crashes. Furthermore, Ascone raises concerns about the consistency of police reporting of distractions across jurisdictions.

5.3 Recommendations for future studies

For a superior understanding of the matter of using cell phone while driving, in order to create more precise data with respect to drivers' awareness to using cell phone risk, it is essential to set up the extent of the cell phone use more accurately for drivers. In order to truly assess of the cell phone crashes share in the total number of crashes, using a cell phone ought to be noted and recorded accurately in reports of accident by police. Moreover, increase the awareness of drivers for the threats of using cell phone and of other activities that proposed as distracted activates; drivers could be uninformed of their performance decrements while driving. Additionally, the driving educational programs should have this distraction as an official portion of their

program and they must educate drivers about the possible distraction impacts and their relative capacity of cognitive to compensate for it.

According to the results, many states have to impose bans for hand-held cell phone use as there is no ban for 82% of distracted drivers in this thesis's dataset for using their cellphone while driving. Therefore, base legislation with respect to mobile phone use on scientific proof is needed. Drawing consideration to the threats of using hands-free phones in case hands-free phones are as unsafe as handheld phones.

It is recommend using the 'technology against technology' attitude. In order to solving the distraction problem while driving or at least partially, it is not hard to visualize that technology could also offer the answer of that with new technologies becoming accessible every day. Regarding mobile phones use, without being distracted by continuous cell phone ringing tones, solutions can be gotten by permitting drivers more time to answer incoming calls to integrating some applications of cell phones with car system to prevent receiving calls while changing the lane and notice the distance with the front car or any complicated task.

Many applications in the smart phones were created in order to warn the driver to stop manipulating in the mobile phone while driving like Focus-Screen free driving, it works in an effective way. It launches when the driver starts driving, and if the mobile phone is touched, a voice sternly says, "hang up and drive."

Accident mitigation technology with autonomous braking monitors the area in front of the vehicle to notify drivers of danger ahead (including vehicles and pedestrians) and

can automatically brake to help drivers avoid crashes. This crash avoidance system will intervene when the drivers are most likely to be distracted.

Future studies should be conducted in Cyprus and it depends on getting a reliable datasets from the reports of police and that needs to include cell phone distraction as a cause for accidents in the police reports.

Also, the relatively new field of self-driving cars could be a major solution to driver distraction by taking charge of the entire driving process, but this field is still new and in need of more development and study.

More studies should be conducted to examine the overcorrection caused by cell phone distraction and how to avoid rollovers by new technologies which is one of the most harmful events to cell phone distracted drivers.

REFERENCES

[1] Lee, J. D. (2006). Human factors and ergonomics in automation design. *Handbook of Human Factors and Ergonomics, Third Edition*, 1570-1596.

[2] Regan, M. A., Lee, J. D., & Young, K. (2008). *Driver distraction: Theory, effects, and mitigation*: CRC Press.

[3] Association, T. G. H. S. (2011). *distracted driving what research shows and what states can do*. Retrieved from Washington, DC:

[4] Pettitt, M., Burnett, G. E., & Stevens, A. (2005). *Defining driver distraction.* Paper presented at the 12th World Congress on Intelligent Transport SystemsITS AmericaITS JapanERTICO.

[5] Young, K., Regan, M., & Hammer, M. (2007). Driver distraction: A review of the literature. *Distracted driving*, 379-405.

[6] Wang, J.-S., Knipling, R. R., & Goodman, M. J. (1996). *The role of driver inattention in crashes: New statistics from the 1995 Crashworthiness Data System.* Paper presented at the 40th annual proceedings of the Association for the Advancement of Automotive Medicine.

[7] Klauer, S. G., Dingus, T. A., Neale, V. L., Sudweeks, J. D., & Ramsey, D. J. (2006). The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data.

[8] Analysis, N. C. f. S. a. (2017). *2016 fatal motor vehicle crashes: Overview* (DOT HS 812 456). Retrieved from Washington,DC:

[9] Council, N. S. (2012). Majority of U.S. adults engage in distracting activities behind the wheel.

[10] Cisco. (2017). VNI Mobile Forecast Highlights, 2016-2012.

[11] CTIA. (2016). CTIA's Wireless Industry Survey.

[12] Hosking, S. G., Young, K. L., & Regan, M. A. (2009). The effects of text messaging on young drivers. *Human factors*, *51*(4), 582-592.

[13] Just, M. A., Keller, T. A., & Cynkar, J. (2008). A decrease in brain activation associated with driving when listening to someone speak. *Brain research*, *1205*, 70-80.

[14] Analysis, N. C. f. S. a. (2017). *Distracted driving 2015*. Retrieved from Washington, DC:

[15] Gartner. (1995). Evolution of data mining. *Gartner Group Advanced Technologies and Applications Research Note*.

[16] Pickrell, T. M., Li, R., & KC, S. (2016). *Driver electronic device use in 2015* (DOT HS 812 326). Retrieved from Washington, DC: National Highway Traffic Safety Administration:

[17] Regan, M. A., Hallett, C., & Gordon, C. P. (2011). Driver distraction and driver inattention: Definition, relationship and taxonomy. *Accident Analysis & Prevention, 43*(5), 1771-1781.

[18] OPERATIONS, C. V. (2009). DRIVER DISTRACTION IN COMMERCIAL VEHICLE OPERATIONS.

[19] Rupp, G. L. (2010). Performance metrics for assessing driver distraction: The quest for improved road safety. *Training, 2017*, 09-29.

[20] Young, K. L., & Salmon, P. M. (2012). Examining the relationship between driver distraction and driving errors: A discussion of theory, studies and methods. *Safety Science, 50*(2), 165-174.

[21] Tchankue, P., Wesson, J., & Vogts, D. (2011). *The impact of an adaptive user interface on reducing driver distraction.* Paper presented at the Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications.

[22] Llaneras, R., Lerner, N., Dingus, T., & Moyer, J. (2000). *Attention demand of IVIS auditory displays: An on-road study under freeway environments.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

[23] Papantoniou, P., Papadimitriou, E., & Yannis, G. (2017). Review of driving performance parameters critical for distracted driving research. *Transportation Research Procedia, 25*, 1801-1810.

[24] Ranney, T. A. (2008). *Driver distraction: A review of the current state-of-knowledge*. Retrieved from

[25] Robertson, R. D., Marcoux, K. D., Vanlaar, W. G., & Pontone, A. M. (2011). Road Safety Monitor 2010: Distracted Driving. *TIRF road safety monitor, 2011*(11F), 1-42.

[26] Strayer, D. L., & Drews, F. A. (2007). Cell-phone–induced driver distraction. *Current Directions in Psychological Science, 16*(3), 128-131.

[27] García-Larrea, L., Perchet, C., Perrin, F., & Amenedo, E. (2001). Interference of cellular phone conversations with visuomotor tasks: An ERP study. *Journal of Psychophysiology, 15*(1), 14.

[28] Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences, 113*(10), 2636-2641.

[29] Stutts, J., Feaganes, J., Reinfurt, D., Rodgman, E., Hamlett, C., Gish, K., & Staplin, L. (2005). Driver's exposure to distractions in their natural driving environment. *Accident Analysis & Prevention, 37*(6), 1093-1101.

[30] Isa, K. A. M., Masuri, M. G., Aziz, N. A. A., Isa, N. N. M., Hazali, N., Tahir, M. P. M., . . . Fansuri, H. (2012). Mobile phone usage behaviour while driving among educated young adults in the urban university. *Procedia-Social and Behavioral Sciences, 36*, 414-420.

[31] Nelson, E., Atchley, P., & Little, T. D. (2009). The effects of perception of risk and importance of answering and initiating a cellular phone call while driving. *Accident Analysis & Prevention, 41*(3), 438-444.

[32] Hallett, C., Lambert, A., & Regan, M. A. (2012). Text messaging amongst New Zealand drivers: Prevalence and risk perception. *Transportation research part F: traffic psychology and behaviour, 15*(3), 261-271.

[33] Backer-Grøndahl, A., & Sagberg, F. (2011). Driving and telephoning: Relative accident risk when using hand-held and hands-free mobile phones. *Safety Science, 49*(2), 324-330.

[34] Bener, A., Crundall, D., Özkan, T., & Lajunen, T. (2010). Mobile phone use while driving: a major public health problem in an Arabian society, State of Qatar—mobile phone use and the risk of motor vehicle crashes. *Journal of Public Health, 18*(2), 123-129.

[35] Smith, D., Najm, W., & Glassco, R. (2002). Feasibility of driver judgment as basis for a crash avoidance database. *Transportation Research Record: Journal of the Transportation Research Board*(1784), 9-16.

[36] daSilva, M., Campbell, B., Smith, J., & Najm, W. (2002). *Analysis of Pedalcyclist Crashes*. Retrieved from

[37] Koopmann, J. A., & Najm, W. G. (2002). *Analysis of off-roadway crash countermeasures for intelligent vehicle applications* (0148-7191). Retrieved from

[38] Najm, W., Koopmann, J., Boyle, L., & Smith, D. (2002). *Development of test scenarios for off-roadway crash countermeasures based on crash statistics*. Retrieved from

[39] comScore. (2017). *Mobile's Hierarchy of Needs*. Retrieved from https://www.comscore.com/Insights/Presentations-and-Whitepapers/2017/Mobiles-Hierarchy-of-Needs:

[40] Tijerina, L. (2000). Issues in the evaluation of driver distraction associated with in-vehicle information and telecommunications systems. *Transportation Research Center Inc*.

[41] Haque, M. M., & Washington, S. (2014). A parametric duration model of the reaction times of drivers distracted by mobile phone conversations. *Accident Analysis & Prevention, 62*, 42-53.

[42] Goodman, M. J., Tijerina, L., Bents, F. D., & Wierwille, W. W. (1999). Using cellular telephones in vehicles: Safe or unsafe? *Transportation Human Factors, 1*(1), 3-42.

[43] Lansdown, T. C. (2012). Individual differences and propensity to engage with in-vehicle distractions–A self-report survey. *Transportation research part F: traffic psychology and behaviour, 15*(1), 1-8.

[44] Dragutinovic, N., & Twisk, D. (2005). Use of mobile phones while driving–effects on road safety. *SWOV Institute, Leidschendam*.

[45] research, S. i. f. r. s. (2017). *Use of the mobile phone while driving*. Retrieved from The Netherlands:

[46] Ranney, T., Watson, G. S., Mazzae, E. N., Papelis, Y. E., Ahmad, O., & Wightman, J. R. (2004). *Examination of the distraction effects of wireless phone interfaces using the national advanced driving simulator-preliminary report on freeway pilot study*. Retrieved from

[47] Greenberg, J., Tijerina, L., Curry, R., Artz, B., Cathey, L., Kochhar, D., . . . Grant, P. (2003). Driver distraction: Evaluation with event detection paradigm. *Transportation Research Record: Journal of the Transportation Research Board*(1843), 1-9.

[48] Tison, J., Chaudhary, N., & Cosgrove, L. (2011). *National phone survey on distracted driving attitudes and behaviors*. Retrieved from

[49] Hickman, J. S., Hanowski, R. J., & Bocanegra, J. (2010). Distraction in commercial trucks and buses: Assessing prevalence and risk in conjunction with crashes and near-crashes.

[50] Olson, R. L., Hanowski, R. J., Hickman, J. S., & Bocanegra, J. L. (2009). *Driver distraction in commercial vehicle operations*. Retrieved from

[51] Urs, H. R., & Urs, A. (2016). *Effect of Cellphone Conversation and Text Messaging on Driver Behaviour: Distracted Driving.* Ohio University.

[52] Dingus, T. A., Klauer, S. G., Neale, V. L., Petersen, A., Lee, S. E., Sudweeks, J., . . . Gupta, S. (2006). *The 100-car naturalistic driving study, Phase II-results of the 100-car field experiment*. Retrieved from

[53] Strayer, D. L., Drews, F. A., & Crouch, D. J. (2006). A comparison of the cell phone driver and the drunk driver. *Human factors, 48*(2), 381-391.

[54] Redelmeier, D. A., & Tibshirani, R. J. (1997). Association between cellular-telephone calls and motor vehicle collisions. *New England journal of medicine, 336*(7), 453-458.

[55] Biervliet, N., Zandvliet, R., Schalkwijk, M., & Gier, M. d. (2010). Periodiek Regionaal Onderzoek Verkeersveiligheid PROV 2009: hoofd-en bijlagenrapport. *Directoraat-Generaal Rijkswaterstaat, Dienst Verkeer en Scheepvaart DVS, Delft*.

[56] Abdel-Aty, M. (2003). Analysis of driver injury severity levels at multiple locations using ordered probit models. *Journal of safety research, 34*(5), 597-603.

[57] Strayer, D. L., & Johnston, W. A. (2001). Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular telephone. *Psychological science, 12*(6), 462-466.

[58] Brookhuis, K. A., de Vries, G., & De Waard, D. (1991). The effects of mobile telephoning on driving performance. *Accident Analysis & Prevention, 23*(4), 309-316.

[59] Lesch, M. F., & Hancock, P. A. (2004). Driving performance during concurrent cell-phone use: are drivers aware of their performance decrements? *Accident Analysis & Prevention, 36*(3), 471-480.

[60] Laberge, J., Scialfa, C., White, C., & Caird, J. (2004). Effects of passenger and cellular phone conversations on driver distraction. *Transportation Research Record: Journal of the Transportation Research Board*(1899), 109-116.

[61] Hanley, P. F., & Sikka, N. (2012). Bias caused by self-reporting distraction and its impact on crash estimates. *Accident Analysis & Prevention, 49*, 360-365.

[62] Liu, Z., & Donmez, B. (2011). *Effects of distractions on injury severity in police-involved crashes.* Paper presented at the Proceedings of the Transportation Research Board 90th Annual Meeting, Washington, DC.

[63] Neyens, D. M., & Boyle, L. N. (2008). The influence of driver distraction on the severity of injuries sustained by teenage drivers and their passengers. *Accident Analysis & Prevention, 40*(1), 254-259.

[64] Zhu, X., & Srinivasan, S. (2011). A comprehensive analysis of factors influencing the injury severity of large-truck crashes. *Accident Analysis & Prevention, 43*(1), 49-57.

[65] Goldenbeld, C., Houtenbos, M., Ehlers, E., & De Waard, D. (2012). The use and risk of portable electronic devices while cycling among different age groups. *Journal of safety research, 43*(1), 1-8.

[66] McEvoy, S. P., Stevenson, M. R., & Woodward, M. (2006). The impact of driver distraction on road safety: results from a representative survey in two Australian states. *Injury prevention, 12*(4), 242-247.

[67] Young, K. L., & Lenné, M. G. (2010). Driver engagement in distracting activities and the strategies used to minimise risk. *Safety Science, 48*(3), 326-332.

[68] Klauer, S. G., Guo, F., Simons-Morton, B. G., Ouimet, M. C., Lee, S. E., & Dingus, T. A. (2014). Distracted driving and risk of road crashes among novice and experienced drivers. *New England journal of medicine, 370*(1), 54-59.

[69] Center, P. R. (2017). *Mobile Fact Sheet*. Retrieved from http://www.pewinternet.org/fact-sheet/mobile/:

[70] Callaway, J., Rushing, S., & Stallmann, A. (2014). Fatal Distraction.

[71] Leung, S., Croft, R. J., Jackson, M. L., Howard, M. E., & McKenzie, R. J. (2012). A comparison of the effect of mobile phone use and alcohol consumption on driving simulation performance. *Traffic injury prevention, 13*(6), 566-574.

[72] David L. Strayer, J. M. C., Jonna Turrill, James Coleman, Nate Medeiros Ward, and Francesco Biondi. (2014). *Measuring Cognitive Distraction in the Automobile*. Retrieved from AAA Foundation for Traffic Safety Washington, DC:

[73] Cheng, C. (2015). Do cell phone bans change driver behavior? *Economic Inquiry, 53*(3), 1420-1436.

[74] Ibrahim, J. K., Anderson, E. D., Burris, S. C., & Wagenaar, A. C. (2011). State laws restricting driver use of mobile communications devices: distracted-driving provisions, 1992–2010. *American journal of preventive medicine, 40*(6), 659-665.

[75] Gillespie, W., & Kim, S.-E. (2001). Cellular phone use while driving: should it be banned or restricted in Georgia?

[76] McCartt, A. T., Braver, E. R., & Geary, L. L. (2003). Drivers' use of handheld cell phones before and after New York State's cell phone law. *Preventive Medicine, 36*(5), 629-635.

[77] McCartt, A. T., & Geary, L. L. (2004). Longer term effects of New York State's law on drivers' handheld cell phone use. *Injury prevention, 10*(1), 11-15.

[78] Insurance Institute of Highway Safety, H. L. D. i. (2016). Cellphones and Texting Laws

[79] Association, G. H. S. (2013). Occupant Protection Incentive Grants.

[80] Association, G. H. S. (2013). National Priority Safety Program.

[81] Council, N. S. (2014). National Distracted Driving Awareness Month.

[82] Burger, N. E., Kaffine, D. T., & Yu, B. (2014). Did California's hand-held cell phone ban reduce accidents? *Transportation research part A: policy and practice, 66*, 162-172.

[83] Strayer, D. L., Drews, F. A., & Johnston, W. A. (2003). Cell phone-induced failures of visual attention during simulated driving. *Journal of experimental psychology: Applied, 9*(1), 23.

[84] Hahn, R. W., & Prieger, J. E. (2006). The impact of driver cell phone use on accidents. *The BE Journal of Economic Analysis & Policy, 6*(1).

[85] Prieger, J. E., & Hahn, R. W. (2007). Are drivers who use cell phones inherently less safe?

[86] Jayasudha, K., & Chandrasekar, C. (2009). An Overview of data mining in road traffic and accident analysis. *Journal of Computer Applications, 2*(4), 32-37.

[87] Khan, M. N. A., Qureshi, S. A., & Riaz, N. (2013). Gender classification with decision trees. *Int. J. Signal Process. Image Process. Patt. Recog, 6*, 165-176.

[88] Zeitouni, K., & Chelghoum, N. (2001). *Spatial decision tree-application to traffic risk analysis.* Paper presented at the Computer Systems and Applications, ACS/IEEE International Conference on. 2001.

[89] Zhang, J., Fraser, S., Lindsay, J., Clarke, K., & Mao, Y. (1998). Age-specific patterns of factors related to fatal motor vehicle traffic crashes: focus on young and elderly drivers. *Public health, 112*(5), 289-295.

[90] Bayam, E., Liebowitz, J., & Agresti, W. (2005). Older drivers and accidents: A meta analysis and data mining application on traffic accident data. *Expert Systems with Applications, 29*(3), 598-629.

[91] Abdel-Aty, M., & Keller, J. (2005). Exploring the overall and specific crash severity levels at signalized intersections. *Accident Analysis & Prevention, 37*(3), 417-425.

[92] Chang, L.-Y. (2005). Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. *Safety Science*, *43*(8), 541-557.

[93] Smith, K. A., Woo, F., Ciesielski, V., & Ibrahim, R. (2001). *Modelling the relationship between problem characteristics and data mining algorithm performance using neural networks, in: C. Dagli, et al.* Paper presented at the Eds.), Smart Engineering System Design: Neural Networks, Fuzzy Logic, Evolutionary Programming, Data Mining, and Complex Systems.

[94] Chong, M., Abraham, A., & Paprzycki, M. (2005). Traffic accident analysis using machine learning paradigms. *Informatica*, *29*(1).

[95] Shanthi, S., & Ramani, R. G. (2011). Classification of vehicle collision patterns in road accidents using data mining algorithms. *International Journal of Computer Applications, 35*(12), 30-37.

[96] Shanthi, S., & Ramani, R. G. (2012). *Feature relevance analysis and classification of road traffic accident data through data mining techniques.* Paper presented at the Proceedings of the World Congress on Engineering and Computer Science.

[97] Tseng, W.-S., Nguyen, H., Liebowitz, J., & Agresti, W. (2005). Distractions and motor vehicle accidents: Data mining application on fatality analysis reporting system (FARS) data files. *Industrial Management & Data Systems, 105*(9), 1188-1205.

[98] Ghazizadeh, M., & Boyle, L. (2009). Influence of driver distractions on the likelihood of rear-end, angular, and single-vehicle crashes in Missouri.

Transportation Research Record: Journal of the Transportation Research Board(2138), 1-5.

[99] Al-Turaiki, I., Aloumi, M., Aloumi, N., & Alghamdi, K. (2016). *Modeling traffic accidents in Saudi Arabia using classification techniques.* Paper presented at the Information Technology (Big Data Analysis)(KACSTIT), Saudi International Conference on.

[100] Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences, 43*(6), 1947-1958.

[101] Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis, 52*(4), 2249-2260.

[102] Harb, R., Yan, X., Radwan, E., & Su, X. (2009). Exploring precrash maneuvers using classification trees and random forests. *Accident Analysis & Prevention, 41*(1), 98-107.

[103] Olson, R. L., Morgan, J. F., Hanowski, R. J., Daily, B., Zimmermann, R., Blanco, M., . . . Flintsch, A. M. (2008). Assessment of a Drowsy Driver Warning System for Heavy Vehicle Drivers.

[104] Prato, C. G., Bekhor, S., Galtzur, A., Mahalel, D., & Prashker, J. N. (2010). *Exploring the potential of data mining techniques for the analysis of accident patterns.* Paper presented at the Proceedings of the 12th World Conference on Transport Research, Lisbon, Portugal.

[105] Nemati, H. R., & Barko, C. D. (2003). Key factors for achieving organizational data-mining success. *Industrial Management & Data Systems, 103*(4), 282-292.

[106] r-project. (2017). R language.

[107] Mitchell, M. W. (2011). Bias of the Random Forest out-of-bag (OOB) error for certain input parameters.

[108] Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*: Elsevier.

[109] Wong, R. C.-W., Fu, A. W.-C., & Wang, K. (2005). Data mining for inventory item selection with cross-selling considerations. *Data mining and knowledge discovery, 11*(1), 81-112.

[110] Breiman, L. (1996). Bagging predictors. *Machine learning, 24*(2), 123-140.

[111] Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

[112] Hothorn, T., & Lausen, B. (2003). Bagging tree classifiers for laser scanning images: a data-and simulation-based strategy. *Artificial intelligence in medicine, 27*(1), 65-79.

[113] Verikas, A., Gelzinis, A., & Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern Recognition, 44*(2), 330-349.

[114] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news, 2*(3), 18-22.

[115] Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics, 8*(1), 25.

[116] Nicodemus, K. K. (2007). *Evaluation of statistical methods to detect epistasis and application to the DISC1 pathway and risk for schizophrenia.* The Johns Hopkins University.

[117] Kim, H., & Loh, W.-Y. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association, 96*(454), 589-604.

[118] Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics, 15*(3), 651-674.

[119] Hothorn, T., & Bühlmann, P. (2006). Model-based boosting in high dimensions. *Bioinformatics*.

[120] Breiman, L. (2017). *Classification and regression trees*: Routledge.

[121] Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in medicine, 29*(3), 337-346.

[122] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., . . . Philip, S. Y. (2008). Top 10 algorithms in data mining. *Knowledge and information systems, 14*(1), 1-37.

[123] Brijain, M., Patel, R., Kushik, M., & Rana, K. (2014). A survey on decision tree algorithm for classification.

[124] Zimasa, T., Jamson, S., & Henson, B. (2017). Are happy drivers safer drivers? Evidence from hazard response times and eye tracking data. *Transportation research part F: traffic psychology and behaviour, 46*, 14-23.

[125] Walker, G. H., Stanton, N. A., & Salmon, P. M. (2011). Cognitive compatibility of motorcyclists and car drivers. *Accident Analysis & Prevention, 43*(3), 878-888.

[126] Hogema, J., & van der Horst, A. (1994). *Driver behaviour under adverse visibility conditions.* Paper presented at the World Congress on Applications of

Transport Telematics and Intelligent Vehicle-Highway Systems. Towards an intelligent transport system Vol. 4.

[127] Haigney, D., Taylor, R., & Westerman, S. (2000). Concurrent mobile (cellular) phone use and driving performance: task demand characteristics and compensatory processes. *Transportation research part F: traffic psychology and behaviour, 3*(3), 113-121.

[128] Rakauskas, M. E., Gugerty, L. J., & Ward, N. J. (2004). Effects of naturalistic cell phone conversations on driving performance. *Journal of safety research, 35*(4), 453-464.

[129] Beede, K. E., & Kass, S. J. (2006). Engrossed in conversation: The impact of cell phones on simulated driving performance. *Accident Analysis & Prevention, 38*(2), 415-421.

[130] Neyens, D. M., & Boyle, L. N. (2007). The effect of distractions on the crash types of teenage drivers. *Accident Analysis & Prevention, 39*(1), 206-212.

[131] Daniello, A., & Gabler, H. C. (2011). Fatality risk in motorcycle collisions with roadside objects in the United States. *Accident Analysis & Prevention, 43*(3), 1167-1170.

[132] Penmetsa, P., Pulugurtha, S. S., & Duddu, V. R. (2017). Factors associated with crashes due to overcorrection or oversteering of vehicles. *IATSS Research*.

[133] Green, P., Hoekstra, E., & Williams, M. (1993). *Further on-the-road tests of driver interfaces: examination of a route guidance system and a car phone*. Retrieved from

[134] Reed, M. P., & Green, P. A. (1999). Comparison of driving performance on-road and in a low-cost simulator using a concurrent telephone dialling task. *Ergonomics, 42*(8), 1015-1037.

[135] McKnight, A. J., & McKnight, A. S. (1993). The effect of cellular phone use upon driver attention. *Accident Analysis & Prevention, 25*(3), 259-265.

[136] Pöysti, L., Rajalin, S., & Summala, H. (2005). Factors influencing the use of cellular (mobile) phone during driving and hazards while using it. *Accident Analysis & Prevention, 37*(1), 47-51.

[137] Ascone, D., Lindsey, T., & Varghese, C. (2009). *An examination of driver distraction as recorded in NHTSA databases*. Retrieved from