**Ramaiah Institute of Technology**

(An Autonomous Institute, Affiliated to VTU)

MSR Nagar, MSRIT post, Bangalore-54

A Dissertation Report on **Traffic Accident Analysis by**

**Decision Tree Induction**

Submitted by

**Ankush Kumar 1MS15CS020**

**Manish Raj Mohanty 1MS15CS064**

**Manollas M 1MS15CS066**

**Kumar Gaurav 1MS15CS145**

*Bachelor of Engineering in Computer Science & Engineering*

Under the guidance of :  **Sowmya B.J**

**Assistant Professor**

**Computer Science**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**M.S.RAMAIAH INSTITUTE OF TECHNOLOGY**

**(Autonomous Institute, Affiliated to VTU)**

**BANGALORE-560054**

**www.msrit.edu, 2017**

**Introduction:**

Road Traffic Accidents killed more than 1.2 million people, and injured between 20 and 50 million others in 2004, thereby becoming the ninth most common cause of death in that year. Road traffic accidents remains among the most central public health problems in the world. A tragic fact is that among the young people aged between 15 and 29 years, road traffic accident is one of the most common causes of death worldwide.

For identification of major cause of accidents, the large amount of data is collected from NHs which is very complex and inefficient to analyze manually. For the analysis, data mining technique can be used to take full advantage of this data set. The results of data mining technique can help highway authority in safety improvements. The traffic accident data analysis can investigate different reasons of traffic accidents. The identification and understanding of these different contributory factors can help public and individual drivers in prevention of major accidents.

Road traffic accidents, the inadvertent crash involving at least one motor vehicle, occurring on a road open to public circulation, in which at least one person is injured or killed; intentional acts (murder, suicide) and natural disasters excluded, is indisputably one of the most frequent and most damaging calamities bedeviling human societies, in particular, US, today. It is therefore, of paramount importance to seek to identify the root causes of road traffic accidents in order to proffer mitigating solutions to address the menace. This research, aimed at predicting the likely causes of road accidents, its prone locations and. In this study data mining decision tree algorithm was used to predict the causes of the accidents, its prone locations and time for effective decision making.So we have used R software ctree for construction of decision tree.

Capacity partykit::ctree is a reimplementation of (a large portion of) party::ctree utilizing the new party framework of the partykit foundation. In spite of the fact that the new code was at that point widely tried, it isn't yet as develop as the old code. On the off chance that you see contrasts in the structure/forecasts of the subsequent trees, please contact the bundle maintainers. See likewise vignette ("ctree", bundle = "partykit") for a few comments about the internals of the distinctive usage. Restrictive derivation trees appraise a relapse relationship by twofold recursive dividing in a contingent induction structure. Generally, the calculation acts as takes after: 1) Test the worldwide invalid theory of autonomy between any of the information factors and the reaction (which might be multivariate also). Stop if this theory can't be rejected. Generally select the info variable with most grounded relationship to the reaction. This affiliation is measured by a p-esteem relating to a test for the fractional invalid theory of a solitary info variable and the reaction. 2) Implement a twofold split in the chose input variable. 3) Recursively repeate stages 1) and 2).

**Data Set Description:**

We have taken Traffic accident data from US traffic database.

**Source of Dataset:** <https://www.kaggle.com/milenski/accident-data/data>

**Attributes Description:**

Accident Severity: It describes the Severity of accident on the person like

Value 1 is Fatal ,Value 2 is Serious ,Value 3 is Slight

No\_of\_Vehicles: It defines the number of vehicles involved in accidents. It has two values

Low ( number of vehicles involved less than 3),High( number of vehicles involved is greater than 3)

Number\_of\_casualities: It defines number of person injured in an accident.

It has 2 values Low (number of persons less than 3) High( number of persons injured greater than 3)

Day\_of\_week: it defines on which day the accident occurs. It has 2 values Weekdays,Weekend

Speed\_limit: It defines the speed of the vehicles travelling on road. It has 2 values High(<=30), Low(<30)

Road\_Type: It defines the type of road the vehicle is travelling, It has two values single\_carriage\_way and all others roads are considered as others.

Pedestrian\_crossing.human\_control: It defines whether the traffic is controlled by human or traffic lights. It has 2 values, Yes and No.

Light\_condition: It defines the intensity of light. It gives two choices it was daylight or artificial light(darkness).

Weather\_condition: It has three values 1-sunny,2-cloudy, 3-other .

Road\_surface\_condition: It has two values 1-concrete(well-maintained),2-other

Urban\_rural\_area: Self-explanatory . It has 2 values 1 and 2 where 1 defines urban and 2 defines rural.

Sex\_of\_casuality: It defines the gender of the person who has had casualities due to accident.

It has two values 1 and 2 1 represent male and 2 represents female.

Age\_of\_casuality: it defines the age of the injured person.

Casuality\_Severity: It defines the level of the casuality.1-fatal,2-serious,3-slight

Vehicle\_type: It defines the type of vehicle involved in accident. It has 2 values car and motorcycle.

Age\_of\_driver: It defines the age of the driver who is driving the vehicle involved in accident.

Sex\_of driver: It defines the gender of driver. It ha s2 values 1 defines male and 2 defines female.

Casualty\_class:It define the casualty belongs to which category.It has three values 1-driver or rider,2-passenger,3-pedestraian

Age\_of vehicle: It defines the age of vehicle.

Data Set size in terms of Bytes and Number of Tuples:

Data set size: 840,45,459 bytes (80.1mb)

Number of tuples: 10,48,575

**Inferences:**

**i>** **accident\_severity**

From the given decision tree we can clearly infer the accident severity (i.e 1.fatal, 2.serious & 3.slight) under dIfferent scenario.

1. If Number\_of\_Vehicles is high and Light\_Condition is daylight and Speed\_Limit is low then probablity is (fatal~nil, serious~0.2, slight>0.8)

2. If Number\_of\_Vehicles is high and Light\_Condition is daylight and Speed\_Limit is high then probablity is (fatal<0.00001, serious<0.2, slight~0.8)

3. If Number\_of\_Vehicles is high and Light\_Condition is darkness then probablity is (fatal<0.005, serious~0.2, slight~0.8)

4. If Number\_of\_Vehicles is low and Urban\_or\_Rural\_Area is urban and Light\_Condition is daylight and Weather\_Condition is 1 and Road\_Type is other then probablity is (fatal<0.0001, serious<0.05, slight<0.9)

5. If Number\_of\_Vehicles is low and Urban\_or\_Rural\_Area is urban and Light\_Condition is daylight and Weather\_Condition is 1 and Road\_Type is single\_carriage\_way then probablity is (fatal<0.0001, serious<0.05, slight<0.8)

6. If Number\_of\_Vehicles is low and Urban\_or\_Rural\_Area is urban and Light\_Condition is daylight and Weather\_Condition is 2,3 then probablity is(fatal<0.0001, serious<0.05, slight<0.9)

7. If Number\_of\_Vehicles is low and Urban\_or\_Rural\_Area is urban and Light\_Condition is darkness then probablity is (fatal<0.0001, serious<0.05, slight<0.9)

8. If Number\_of\_Vehicles is low and Urban\_or\_Rural\_Area is rural then probablity is (fatal<0.0005, serious<0.05, slight~0.8)

**ii>.Urban\_Rural\_Area:**

1. If speed limit is low and if Road\_Surface\_Conditions is 1 and nmber\_of \_casualities is

high then the probability of accident occurring in urban area is .8 and in rural area is .2.

2. If speed limit is low and if Road\_Surface\_Conditions is 1 and nmber\_of \_casualities is

low and light condition is daylight and if day\_of\_week is weekend then probability of

accident occurring in urban area is .2 and in rural area is .8.

3. If speed limit is low and if Road\_Surface\_Conditions is 1 and nmber\_of \_casualities is

low and light condition is daylight and if day\_of\_week is weekday weekend then

probability of accident occurring in urban area is .8 and rural area is .2.

4. If speed limit is low and if Road\_Surface\_Conditions is 1 and nmber\_of \_casualities is

low and light condition is darkness then probability of accident occurring in urban area is

.8 and rural area is .2.

5. If speed limit is low and if Road\_Surface\_Conditions is 2 then probability of accident

occurring in urban area is .8 and rural area is .2.

6. If speed limit is high and road\_type is others then probability of accident occurring in

urban area is .4 and rural is .6.

7. If speed limit is high and road\_type is single\_carriage\_road then probability of accident

occurring in urban area is .2 and rural is .8.

**iii> Vehicle\_type**

1. If sex\_of\_driver is female and age\_of\_vehicle is & lt;= 4 then probability that the vehicle

is motorcycle is .2 and probability of it being a car is .8.

2. If sex\_of\_driver is female and age\_of\_vehicle is & gt; 4 then probability that the vehicle

is motorcycle is .1 and probability of it being a car is .9.

3. If sex\_of\_driver is male and age\_of\_vehicle is & lt;= 4 and age of driver is &lt;=16 then

probability that the vehicle is motorcycle is .95 and probability of it being a car is .05.

4. If sex\_of\_driver is male and age\_of\_vehicle is & lt;= 4 and age of driver is &gt;16 and then

probability that the vehicle is motorcycle is .2 and probability of it being a car is .8

5. If sex\_of\_driver is male and age\_of\_vehicle is &gt;4 and & lt;=7 and age then probability

that the vehicle is motorcycle is .3 and probability of it being a car is .7

6. If sex\_of\_driver is male and age\_of\_vehicle is &gt;7 and age then probability that the

vehicle is motorcycle is .2 and probability of it being a car is .8

**iv>light\_conditions**

1. If Road\_Surface\_Conditions is 1 and Day\_of\_Week is weekend and Number\_of\_Vehicles is low then probability is(daylight<0.8 and darkness<0.2)

2.If Road\_Surface\_Conditions is 1 and Day\_of\_Week is weekend and Number\_of\_Vehicles is high then probability is(daylight<0.7 and darkness<0.3)

3. If Road\_Surface\_Conditions is 1 and Day\_of\_Week is weekday and Accident\_Severity is 1,2 then probability is(daylight<0.8 and darkness<0.2)

4. If Road\_Surface\_Conditions is 1 and Day\_of\_Week is weekday and Accident\_Severity is 3 and Weather\_condition is 2 then probability is(daylight<0.8 and darkness<0.2)

5. If Road\_Surface\_Conditions is 1 and Day\_of\_Week is weekday and Accident\_Severity is 3 and Weather\_condition is 1,3 and Number\_of\_Vehicles is high then probability is(daylight>0.8 and darkness<0.2)

6. If Road\_Surface\_Conditions is 1 and Day\_of\_Week is weekday and Accident\_Severity is 3 and Weather\_condition is 1,3 and Number\_of\_Vehicles is low then probability is(daylight>0.8 and darkness<0.2)

7. If Road\_Surface\_Conditions is 2 and Number\_of\_Vehicles is high then probability is(daylight<0.6 and darkness<0.4)

8. If Road\_Surface\_Conditions is 2 and Number\_of\_Vehicles is low then probability is(daylight<0.6 and darkness<0.4)

**Algorithm Description:**

A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.

Generating a decision tree form training tuples of data partition D

Algorithm : Generate\_decision\_tree

Input:

Data partition, D, which is a set of training tuples

and their associated class labels.

attribute\_list, the set of candidate attributes.

Attribute selection method, a procedure to determine the

splitting criterion that best partitions that the data

tuples into individual classes. This criterion includes a

splitting\_attribute and either a splitting point or splitting subset.

Output:

A Decision Tree

Method

create a node N;

if tuples in D are all of the same class, C then

return N as leaf node labeled with class C;

if attribute\_list is empty then

return N as leaf node with labeled

with majority class in D;|| majority voting

apply attribute\_selection\_method(D, attribute\_list)

to find the best splitting\_criterion;

label node N with splitting\_criterion;

if splitting\_attribute is discrete-valued and

multiway splits allowed then // no restricted to binary trees

attribute\_list = splitting attribute; // remove splitting attribute

for each outcome j of splitting criterion

// partition the tuples and grow subtrees for each partition

let Dj be the set of data tuples in D satisfying outcome j; // a partition

if Dj is empty then

attach a leaf labeled with the majority

class in D to node N;

else

attach the node returned by Generate

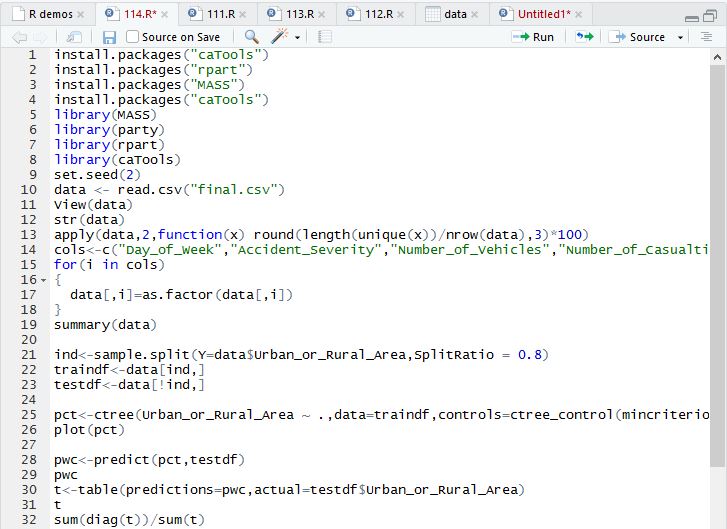
decision tree (Dj, attribute list) to node N;

end for

return N;

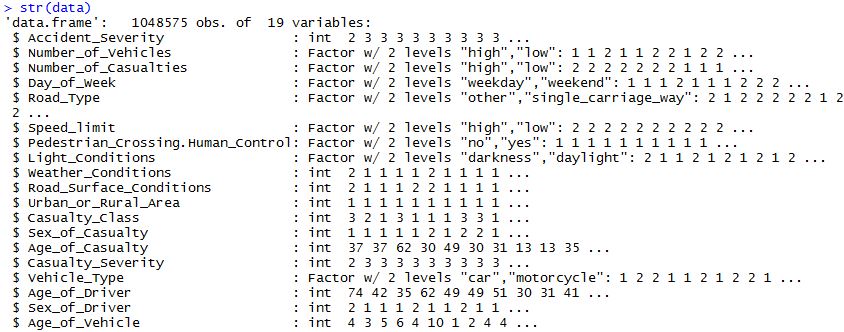
Snapshots:

Algorithm code:

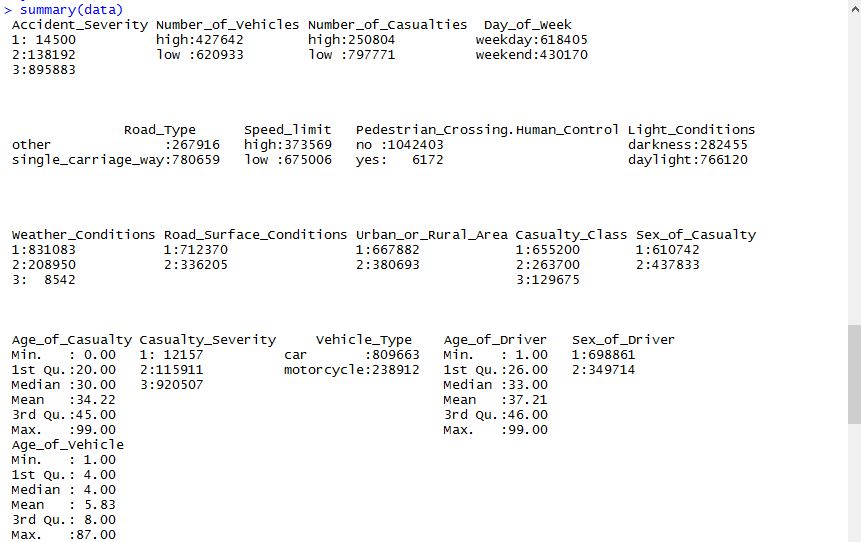


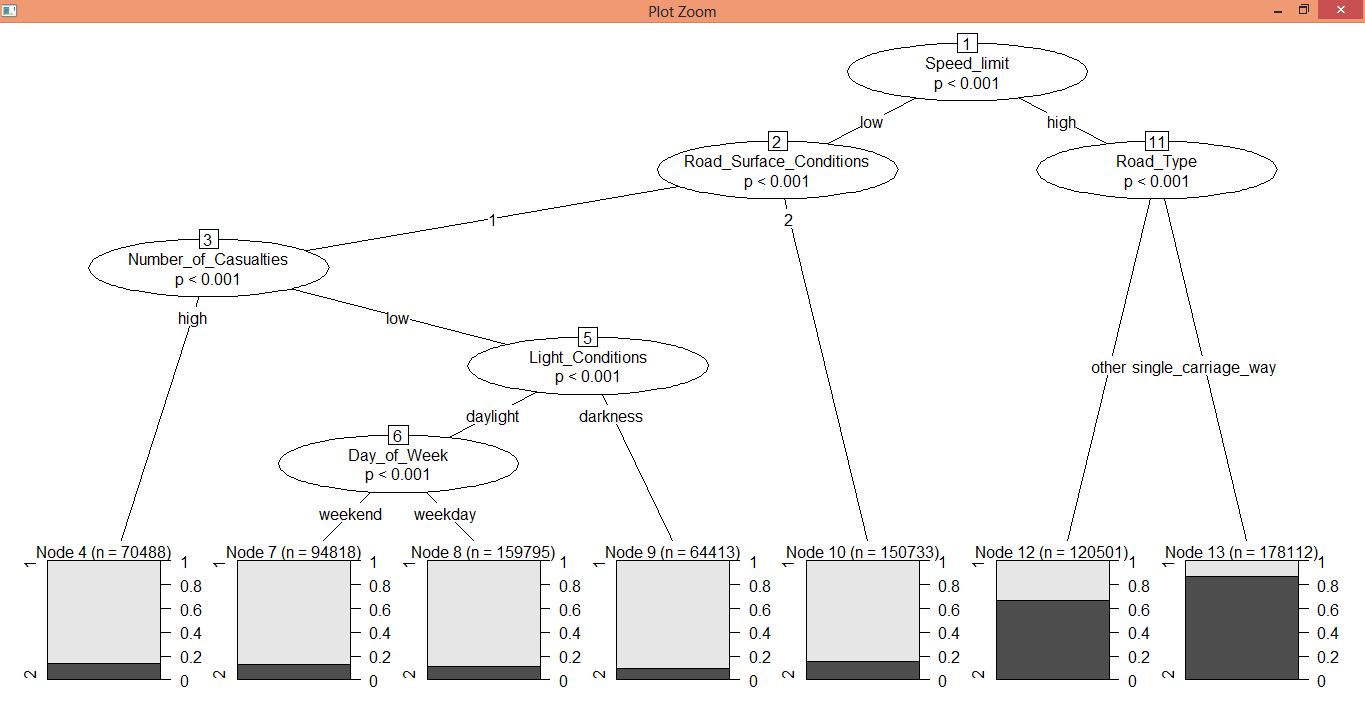
RESULT SNAPSHOT: ATTRIBUTE(URBAN\_RURAL\_AREA)

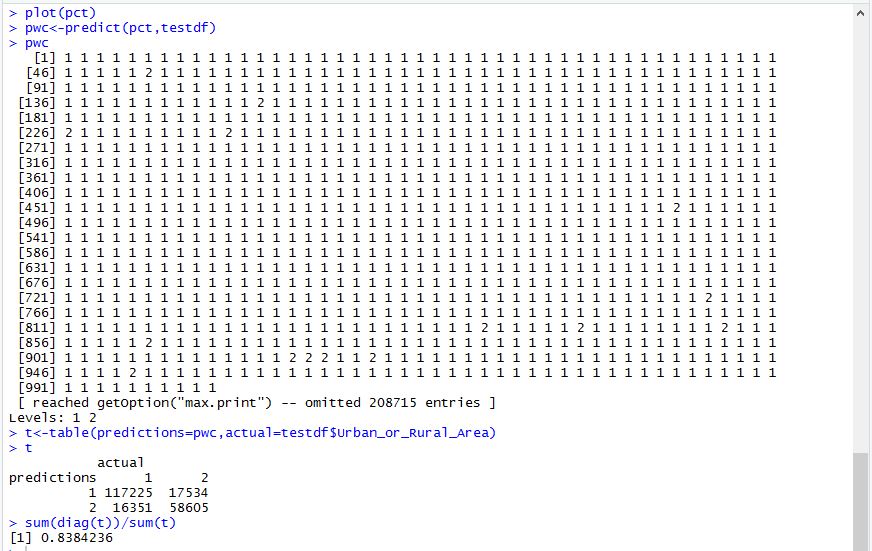
Dataset structure:



Dataset summary:







1. If speed limit is low and if Road\_Surface\_Conditions is 1 and nmber\_of \_casualities is

high then the probability of accident occurring in urban area is .8 and in rural area is .2.

2. If speed limit is low and if Road\_Surface\_Conditions is 1 and nmber\_of \_casualities is

low and light condition is daylight and if day\_of\_week is weekend then probability of

accident occurring in urban area is .2 and in rural area is .8.

3. If speed limit is low and if Road\_Surface\_Conditions is 1 and nmber\_of \_casualities is

low and light condition is daylight and if day\_of\_week is weekday weekend then

probability of accident occurring in urban area is .8 and rural area is .2.

4. If speed limit is low and if Road\_Surface\_Conditions is 1 and nmber\_of \_casualities is

low and light condition is darkness then probability of accident occurring in urban area is

.8 and rural area is .2.

5. If speed limit is low and if Road\_Surface\_Conditions is 2 then probability of accident

occurring in urban area is .8 and rural area is .2.

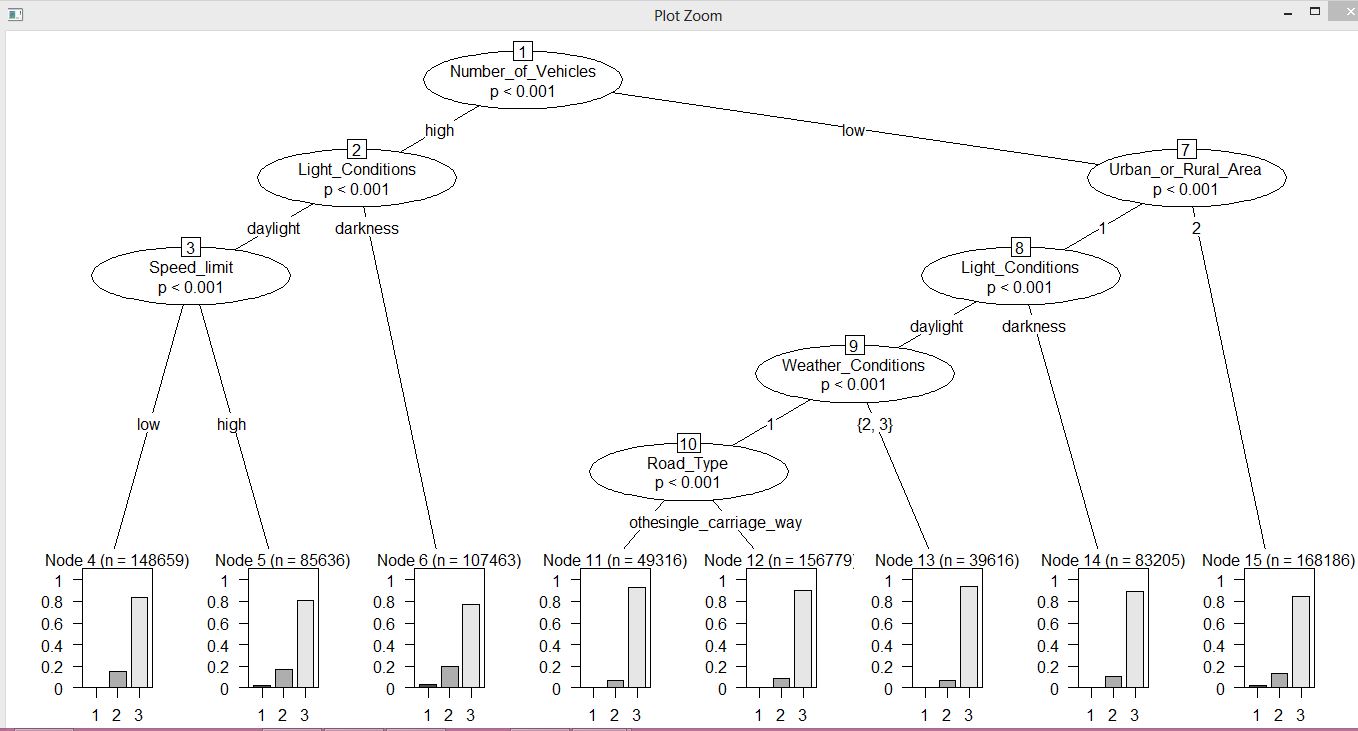
6. If speed limit is high and road\_type is others then probability of accident occurring in

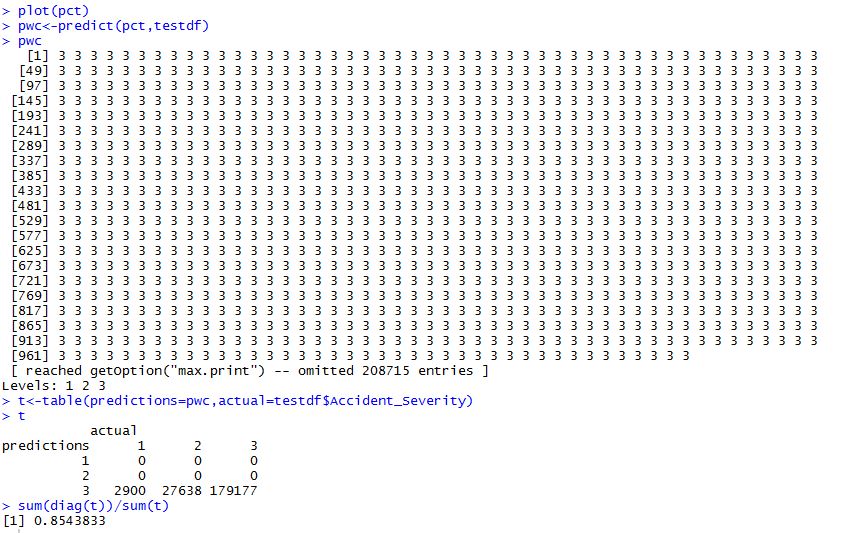
urban area is .4 and rural is .6.

7. If speed limit is high and road\_type is single\_carriage\_road then probability of accident

occurring in urban area is .2 and rural is .8.

ATTRIBUTE(ACCIDENT\_SEVERITY):





1. If Number\_of\_Vehicles is high and Light\_Condition is daylight and Speed\_Limit is low then probablity is (fatal~nil, serious~0.2, slight>0.8)

2. If Number\_of\_Vehicles is high and Light\_Condition is daylight and Speed\_Limit is high then probablity is (fatal<0.00001, serious<0.2, slight~0.8)

3. If Number\_of\_Vehicles is high and Light\_Condition is darkness then probablity is (fatal<0.005, serious~0.2, slight~0.8)

4. If Number\_of\_Vehicles is low and Urban\_or\_Rural\_Area is urban and Light\_Condition is daylight and Weather\_Condition is 1 and Road\_Type is other then probablity is (fatal<0.0001, serious<0.05, slight<0.9)

5. If Number\_of\_Vehicles is low and Urban\_or\_Rural\_Area is urban and Light\_Condition is daylight and Weather\_Condition is 1 and Road\_Type is single\_carriage\_way then probablity is (fatal<0.0001, serious<0.05, slight<0.8)

6. If Number\_of\_Vehicles is low and Urban\_or\_Rural\_Area is urban and Light\_Condition is daylight and Weather\_Condition is 2,3 then probablity is(fatal<0.0001, serious<0.05, slight<0.9)

7. If Number\_of\_Vehicles is low and Urban\_or\_Rural\_Area is urban and Light\_Condition is darkness then probablity is (fatal<0.0001, serious<0.05, slight<0.9)

**Implementataion:**

i>After selecting the data we need to preprocess them.We have used python programming language for preprocessing .we have replaced all missing values with mean or mode depending upon the attribute type.

ii>We have 10 lakh tuples,out of ehich we have used 80% training and 20% for testing.

ind<-sample.split(Y=data$Urban\_or\_Rural\_Area,SplitRatio = 0.8)

traindf<-data[ind,]

testdf<-data[!ind,]

**iii>**We have used ctree for construction of tree .The **ctree()** function in party package can be used to model binary, nominal, ordinal and numeric variables. The nature of the tree depends on the type of response variable. Pruning the tree is not required with this approach.

pct<-ctree(Urban\_or\_Rural\_Area ~ .,data=traindf,controls=ctree\_control(mincriterion=0.99,minsplit=200000))

plot(pct)

iv>Prediction on new data or test data.

pwc<-predict(pct,testdf)

t<-table(predictions=pwc,actual=testdf$Urban\_or\_Rural\_Area)

sum(diag(t))/sum(t)

**Social impact:**

Investigation on the reasons that would cause traffic accidents, together with their probability distribution, is of great help to understand and prevent such accidents for both public departments and individual drivers. However, the amount of accidental data collected on the roads or highways is extremely huge and the data are complex involving various factors, it is very difficult if not impossible for a data analysis to be carried out manually. To take effective and full advantage of these resources, a data mining technology, by obtaining the reasons that cause traffic accidents and their probability distribution, can be utilized to provide decision references for relevant industrial and governmental users and drivers.