Traffic Accident Severity Prediction

Team Name: The Collective

Team Members and Email id's:

1. Susmitha Srirangam - 00764501 - ssrir5@unh.newhaven.edu
2. Kaivalya Reddy Maddireddy - 00760616 - kmadd3@unh.newhaven.edu
3. Thatikonda Akanksha - 00764338 - athat2@unh.newhaven.edu

Abstract:

Traffic accidents are one of the leading causes of death. The frequency of occurrence of these accidents can be reduced, by predicting severity of accidents based on speed limit weather conditions, work zones and surface conditions using scientific knowledge. Here we are going to anticipate the probability of occurrence of a fatal or nonfatal accidents. This type of predictions is required for issuing safety guidelines for travelers along the route through the entire weather conditions. To determine the level of severity, we are using several algorithms like Decision Tree, Logistic Regression, K Nearest Neighbor, Naive Bayes algorithm on the data.

Keywords – Naïve Bayes, K Nearest Neighbor, Logistic Regression, Random Forest, Decision Tree.

Introduction:

Predicting severity of accidents allows us to take safety precautions necessary, to make the roads safer, accident prediction is crucial for public transportation optimization, permitting better routes, and cost-effectively enhancing the transportation infrastructure. Accident analysis and prediction has received a lot of attention due to its importance.

By examining how environmental factors, such as weather, traffic, and characteristics of the road network, affect the frequency of traffic accidents during the past years, we can predict the future incidents that might happen under the same conditions.

A decision regarding the safety of a driver may be made by using the data mining technique to create a prediction model in these traffic accident data. Considering this, we can all agree that road safety can boosted by creating a model for accident severity prediction. Data mining, according to, is a method for extracting information from a collection of data that aids in decision-making. The four steps of data mining are classification, clustering, estimation, and association.

To build a prediction model in our research we are using different prediction models like Decision tree, Naive Bayes, K Nearest Neighbor and Logistic Regression here we even tried to figure out which prediction model gives the best accuracy when compared with the rest of the algorithm models.

Related Research Work and Literature Review:

Research Paper – 1:

(PDF) Traffic Accident Severity Prediction Using Naive Bayes Algorithm - A Case Study of Semarang Toll Road (researchgate.net)

Title: Traffic Accident Severity Prediction Using Naïve Bayes Algorithm – A case Study of Semarang Toll Road

Authors: W Budiawan, S Saptadi, Sriyanto, C Tijioe and T Phommachak

Affiliations: W Budiawan, S Saptadi, Sriyanto, C Tijioe and T Phommachak
Diponegoro University, JL.Prof.H Soedarto, SH. Semarang 50275, Indonesia
Student at Department of Architectural and Civil engineering, Toyohashi University of Technology, Japan

Publishers Name: IOP Conference Series Materials Science and Engineering

Publish Date: September 2019 **Literature**

Review:

In this research paper they have focused on traffic accidents happening at toll road Indonesia. To analyze the probability of accidents happening they have considered the data from 2007-2017. Their aim is to propose a prediction model which can determine the probability and severity of the accident based on different factors like days, type of road, weather condition of road, time of accident, sex of driver and type of vehicle. Naïve Bayes algorithm was used to make the model which can predict accident severity like weather it is material damage, Minor injuries, Major Injuries and fatal. The result of this applied model is with an accuracy of 39.49% to predict the severity and probability of accident.

Data Set Selected: Traffic accident in Semarang Toll Road from 2007 to 2017
Research Paper -2

Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights
(arxiv.org)

Title: Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights

Authors: Sobhan Moosavi, Mohamad Hossein, Srinivasan Parthasarathy, Radu Teodorescu, Rajiv Ramnath.

Affiliations: The Ohio State University

Publishers Name: 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '19)

Publish Date: November5-8, 2019

Literature Review

The main goal of this research paper is to reduce traffic accidents by predicting the impact of accidents under various conditions like environment, Driver's status, vehicles status and many other attributes. Their solution relies on a deep-neural network model while comparing traditional machine learning algorithms like line regression. The data set they have used is US- Accident data set and an extensive set of experiments across several large cities. Their study results showed significant improvements to predict rare accident events on traffic information, time, and points-of-interest data for real time accident prediction.

Data Set Selected: US- Accidents took place within the contiguous United States Between February 2016 and March 2019.

Research Paper-3

https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/stamp/stamp.jsp?tp=&arnumber=9795842

Title: Analysis of Data in Solving the problem of Reducing the Accident Rate Through the use of Special Means on Public Roads.

Publishers: Vladislav Kukartsev, Anton Mikhalev, Alexander Stashkevich, Kristina Moiseeva, Igor Kauts

Publishers Name: 2022 IEEE International IOT, Electronics and Mechatronics Conference

Publish Date:2022

Literature Review: This study's primary goal is to compile statistics on traffic collisions in the Russian Federation. Next, think of measures to cut down on accidents. the evaluation of statistical data on the number of collisions, the population, the length of the roads, the number of vehicles besides motorcycles, and the general condition of the roads today and a few years ago, as well as the main causes of collisions and prevention strategies. This investigation produced accurate outcomes and ideas that will help the situation on the roads.

Data set selected: Russian federation statistical data given in the period from 1991 to 2019.

Research Paper-4

https://www.ijser.org/researchpaper/Analysis-of-Road-Accidents-using-AprioriNaive-Bayes-and-K-Means.pdf

Title: Analysis of Road Accidents using Apriori, Navie-Bayes and K-Means

Publishers: Neha Patil, prof. Deepesh Jagadala

Affiliations: International Journal of Scientific & Engineering Research Volume

Publishers Name: IJSER@2021

Publish Date: March 3,2021

Literature Review

The overall goal is to use predictive modeling to examine the impact of road-related factors on accident severity and to determine risk factors for fatal traffic accidents. To estimate accident severity using various data mining strategies we used Apriori, Naive Bayes classifier, and other association and classification data mining algorithms are used to anticipate the pattern of future road accidents. K means, which have excellent scalability. even if we are at work on a database containing millions of records with specific properties, this classifier can produce the best outcomes. There are types of model's issue instances, which are represented as vectors of feature values, are given class names, and the class labels are selected from a limited set.

Data set Selected: The data is collected from police stations restricted to an area

Research Paper -5

Review On Road Accident Analytics Using Data Mining Technique | International Journal on Future Revolution in Computer Science & Communication Engineering (ijfrcsce.org)

Title: Review on Road Accident Analytics Using Data Mining Technique

Publishers: Prof. Reena Thakur, Sonal Paryani, Department of C.S.E, GNIT, India

Publishers Name: International Journal on Future Revolution in Computer Science & Communication Engineering

Publish date: January 2018 **Literature**

Review:

The main goal of this research is to make use of association and classification rules to discover the patterns between road accidents and as well as predict road accidents for new roads. Here they predicted the probability of common accidents that might occur on new roads with the help of Naïve Bayes algorithm and to find the association possibilities among the road accidents using aprioiri algorithm. By this research they want to change the current government system which is manual to an automated one that is necessary to take the precautionary measures and in turn reduce the number of accidents.

Data Set Selected: Ledger data from government of India including details of weather, road conditions and many more

Performance Metrics of the research papers:

Combinedly all the research paper's goal is analyzed and predict the possible ways to reduce the number of accidents occurred due to factors like surface conditions, weather, and many other constraints.

Here to predict the probability of future accidents occurrences and avoid them with better measures we feel that Naïve Bayes Algorithm is the best procedure to give us expected results. While in research papers they have used many other algorithm approaches like decision trees and other methods. Apriori can give best results when we are trying to find the association between accidents and attributes effecting the accidents.

Research Question: Our aim is to predict the severity of accident based on the data collected by applying various data mining techniques.

Dataset information:

This dataset contains 11 attributes with values recorded based on time while considering hours of travel like rush hour, workdays, weekends, surface conditions, weather conditions, speed limits and predicting the severity of accidents weather they are fatal or non-fatal.

Attribute Information:

- **Rush Hour:** Shows weather the hour of travel is comes under rush hour or non-rush hour so, we can get to know the traffic conditions of the area.
- **Work Zone:** Shows if the area of travel is coming under work zone or not from this, we can decide how the timings impact traffic in this zone
- **Workday:** Shows if the day is workday or not because of which we can determine the which zones can see traffic during the working days.
- **INT_HWY**
- **LGTCON_day**

- **Level:** The level of road from the surface which can also tend to accidents when having the impact of weather
- **Speed Limit:** Travel speed of vehicles during the traffic and while having the impact of other conditions mostly lead to accidents
- **Surface Conditions:** Shows weather the surface condition of the road is good or wearied off including the climate conditions like snow, rainy or greasy roads
- **Traffic Two Way:** This shows if the traffic is one way or two ways to rule out and analyze the conditions opposite side
- **Maximum Severity:** With analysis of all these conditions the severity of the accident is determined weather it is fatal or non-fatal.

List of data mining techniques used:

Until today the data mining techniques used are:

- Decision Trees: It is predictive machine learning model that is used for classification of data based on considered attributes
- Logistic regression: This is a statistical model used for classification and predictive analytics, it estimates the probability of event occurring based on a given dataset of independent variables.
- Naïve Bayes Algorithm: This provides us with probability of a prediction from the underlying evidence, as observed in the data.
- K Nearest Neighbor Algorithm: This is a classifier algorithm where the learning is based on how similar a data from other available attributes is.

Description of parameters and hyperparameters:

Decision Trees: Hyper parameter are responsible for controlling the model. In decision tree the hyper parameters are most basic ones are:

- **Criterion:** This function is used to calculate the uncertainty on the rule selected
- **Max Features:** The number of features to consider when searching for the best split rule.
- **Max Depth:** To determine the maximum depth of the tree to get best outcome possible.

- **Cpp_alpha:** Here the node with high complexity is pruned and less than cpp value will be pruned.

Logistic Regression:

Logistic regression is a classification algorithm that is used to predict a outcome where there are only two possible scenarios that is either the event happens or it does not happen. To calculate the outcome we can equally split the data into intervals. We can continue to split the data until we find appropriate outcomes. These intervals are termed as iterations. In our model it took 4 iterations to get to an optimized solution.

Naive Bayes:

Naive Bayes is a probabilistic algorithm which can provide us prediction really quick, In general Naive Bayes doesn’t have any hyper parameters to tune, however we can control the data that is used by the algorithm by setting parameters like Maximum_Input_Attributes, Maximum_Output_Attributes and Minimum_States.

K Nearest Neighbor :

The KNN technique, which is used to address classification model problems, establishes an illogical border to classify the data. The program will try to forecast the closest border line as additional data points are received.

The adjustable parameters for KNN are:

- N_neighbor: This regulates number of neighbors that are checked when an item is being classified. The default value is 5. This is also known as ‘K’ Value.
- Weights: This determines how weights are distributed among the neighbor values. The default value is uniform.

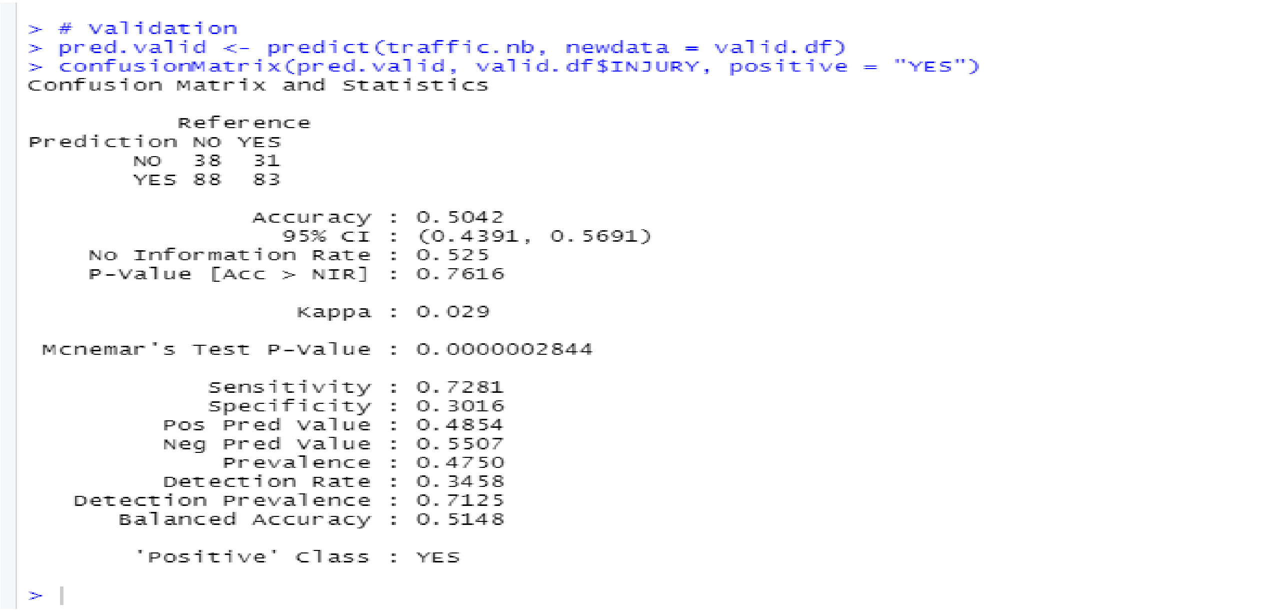
Brief description of hardware used:

11th Gen Intel® Core™ i7, 16 GB RAM, Windows 11, 64-bit operating system.

Outcomes of data mining techniques and performance metrics:

Naïve Bayes Technique:

Here we have implemented Naïve Bayes Algorithm in R studio by taking 60 percent of training data and 40 percent of validation data. The Performance metrics are as follows

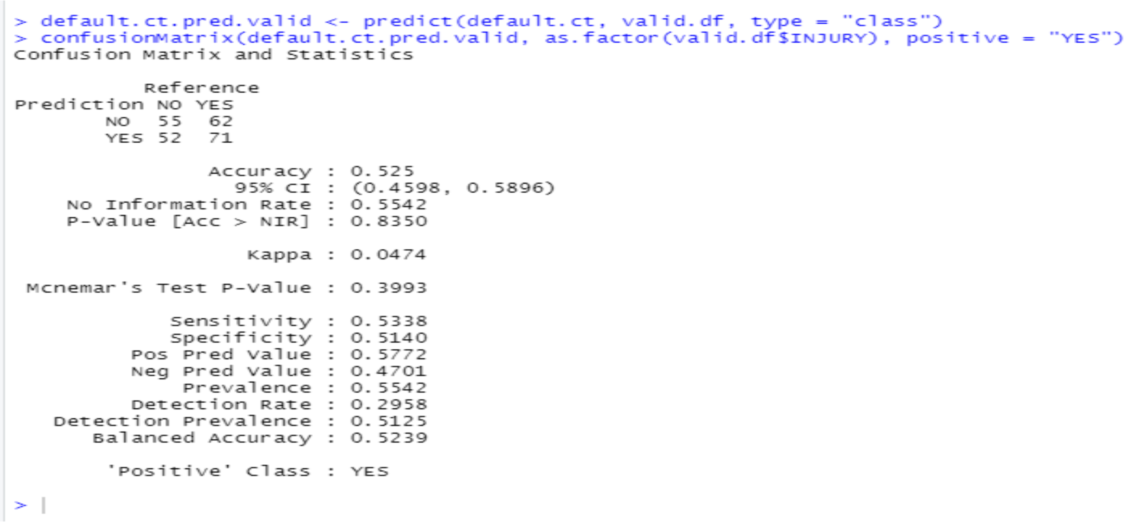


```
> # validation
> pred.valid <- predict(traffic.nb, newdata = valid.df)
> confusionMatrix(pred.valid, valid.df$INJURY, positive = "YES")
Confusion Matrix and Statistics

          Reference
Prediction NO YES
       NO  38  31
       YES 88  83

               Accuracy : 0.5042
                 95% CI : (0.4391, 0.5691)
    No Information Rate : 0.525
    P-Value [Acc > NIR] : 0.7616

                  Kappa : 0.029

 Mcnemar's Test P-Value : 0.0000002844

            Sensitivity : 0.7281
            Specificity : 0.3016
         Pos Pred Value : 0.4854
         Neg Pred Value : 0.5507
             Prevalence : 0.4750
         Detection Rate : 0.3458
   Detection Prevalence : 0.7125
      Balanced Accuracy : 0.5148

       'Positive' Class : YES

> 
```

The accuracy is not up to the mark and coming to precision and recall that are Sensitivity and Specificity. The values are not perfectly balanced out. But the probability of injury in most of the cases is 'YES'.

Decision Tree:

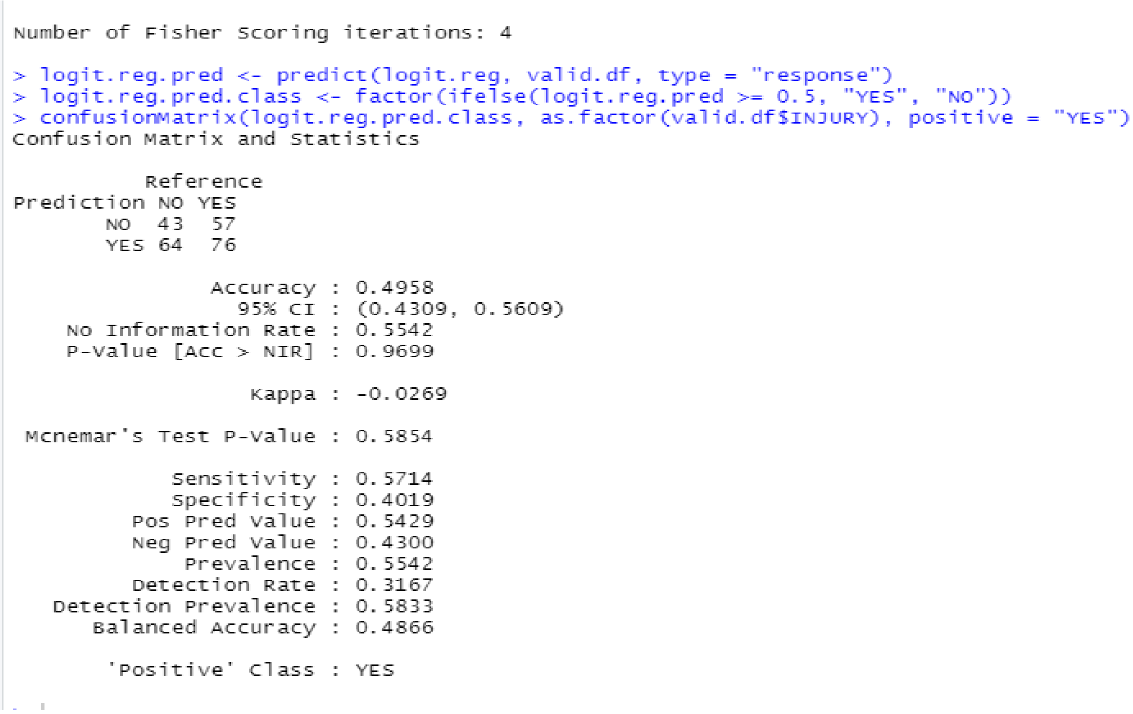
The Decision tree algorithm also gave us the positive class out come that is yes. Here the Accuracy, Precession and Recall values are at the same value. Which makes it difficult to understand the performance.



```
> default.ct.pred.valid <- predict(default.ct, valid.df, type = "class")
> confusionMatrix(default.ct.pred.valid, as.factor(valid.df$INJURY), positive = "YES")
Confusion Matrix and Statistics

          Reference
Prediction NO  YES
       NO   55  62
       YES  52  71
                                          
               Accuracy : 0.525           
                 95% CI : (0.4598, 0.5896)
    No Information Rate : 0.5542          
    P-Value [Acc > NIR] : 0.8350          
                                          
                  Kappa : 0.0474          
                                          
 Mcnemar's Test P-Value : 0.3993          
                                          
            Sensitivity : 0.5338          
            Specificity : 0.5140          
         Pos Pred Value : 0.5772          
         Neg Pred Value : 0.4701          
             Prevalence : 0.5542          
         Detection Rate : 0.2958          
   Detection Prevalence : 0.5125          
      Balanced Accuracy : 0.5239          
                                          
       'Positive' Class : YES             
                                          
> |
```

Logistic Regression:

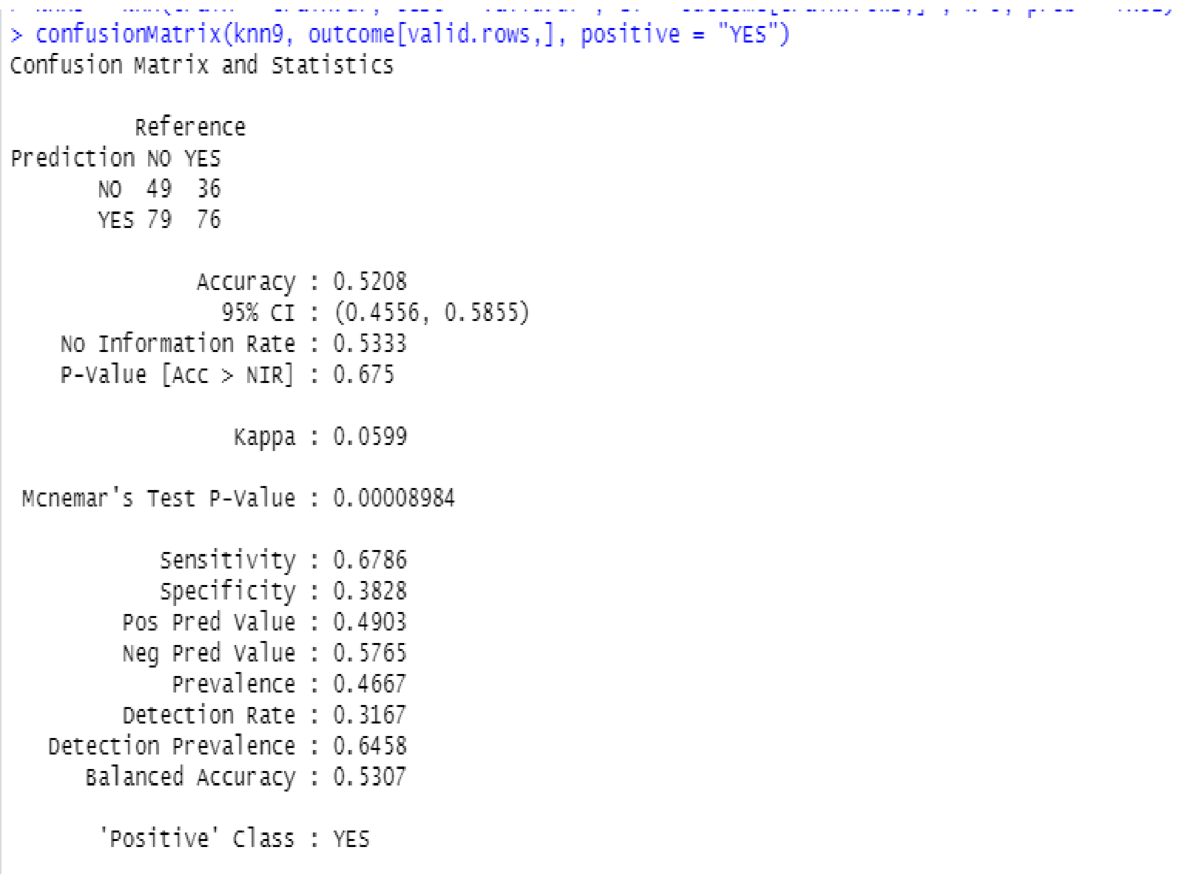
Here the number of iterations is four and the performance outcome is 'YES'. But accuracy here is low compared to above applied algorithms. Similarly, the precision and Recall values are also not perfectly balanced.



```
Number of Fisher Scoring iterations: 4

> logit.reg.pred <- predict(logit.reg, valid.df, type = "response")
> logit.reg.pred.class <- factor(ifelse(logit.reg.pred >= 0.5, "YES", "NO"))
> confusionMatrix(logit.reg.pred.class, as.factor(valid.df$INJURY), positive = "YES")
Confusion Matrix and Statistics

          Reference
Prediction NO  YES
       NO   43  57
       YES  64  76
                                          
               Accuracy : 0.4958          
                 95% CI : (0.4309, 0.5609)
    No Information Rate : 0.5542          
    P-Value [Acc > NIR] : 0.9699          
                                          
                  Kappa : -0.0269         
                                          
 Mcnemar's Test P-Value : 0.5854          
                                          
            Sensitivity : 0.5714          
            Specificity : 0.4019          
         Pos Pred Value : 0.5429          
         Neg Pred Value : 0.4300          
             Prevalence : 0.5542          
         Detection Rate : 0.3167          
   Detection Prevalence : 0.5833          
      Balanced Accuracy : 0.4866          
                                          
       'Positive' Class : YES             
```

K Nearest Neighbor:

We have considered K value as 9. Here the accuracy, precession and recall values are comparatively balanced out.

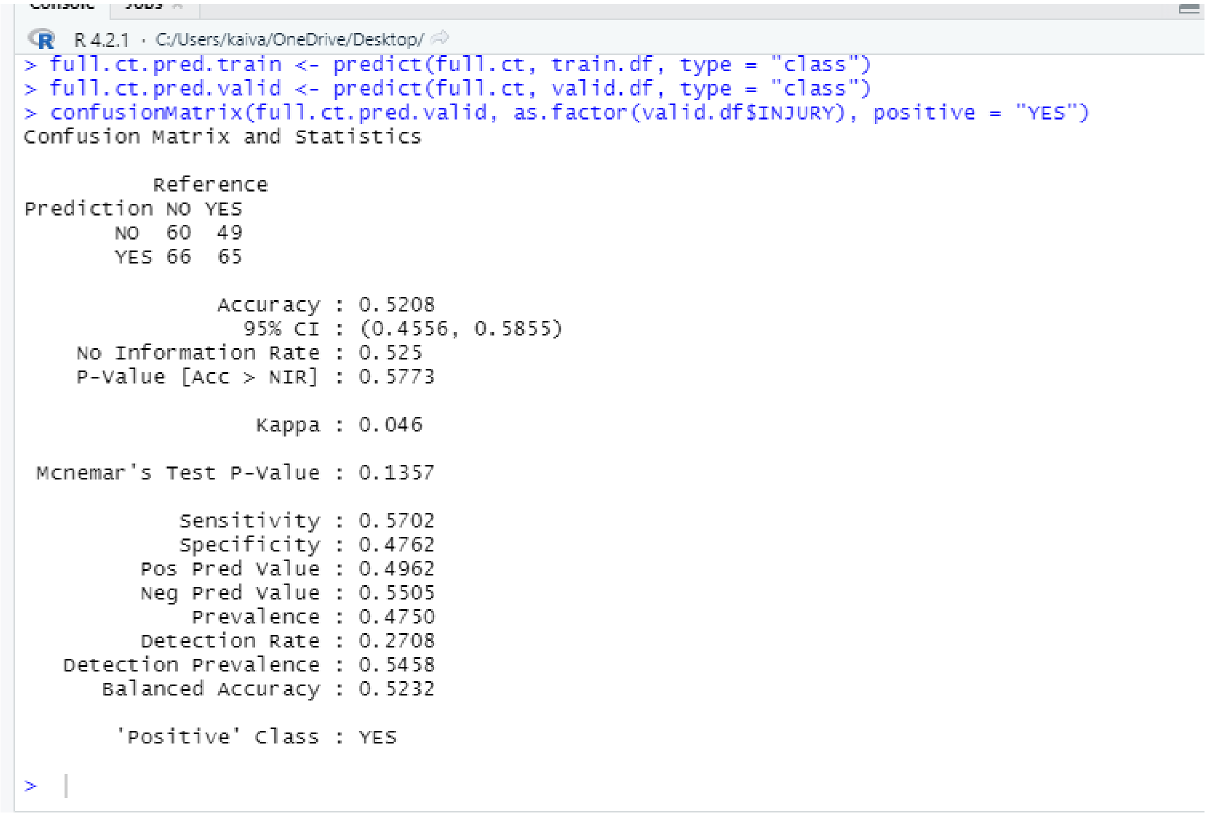


```
> confusionMatrix(knn9, outcome[valid.rows,], positive = "YES")
Confusion Matrix and Statistics

          Reference
Prediction NO YES
       NO  49  36
       YES 79  76

               Accuracy : 0.5208
                 95% CI : (0.4556, 0.5855)
    No Information Rate : 0.5333
    P-Value [Acc > NIR] : 0.675

                  Kappa : 0.0599

 Mcnemar's Test P-Value : 0.00008984

            Sensitivity : 0.6786
            Specificity : 0.3828
         Pos Pred Value : 0.4903
         Neg Pred Value : 0.5765
             Prevalence : 0.4667
         Detection Rate : 0.3167
   Detection Prevalence : 0.6458
      Balanced Accuracy : 0.5307

       'Positive' Class : YES
```

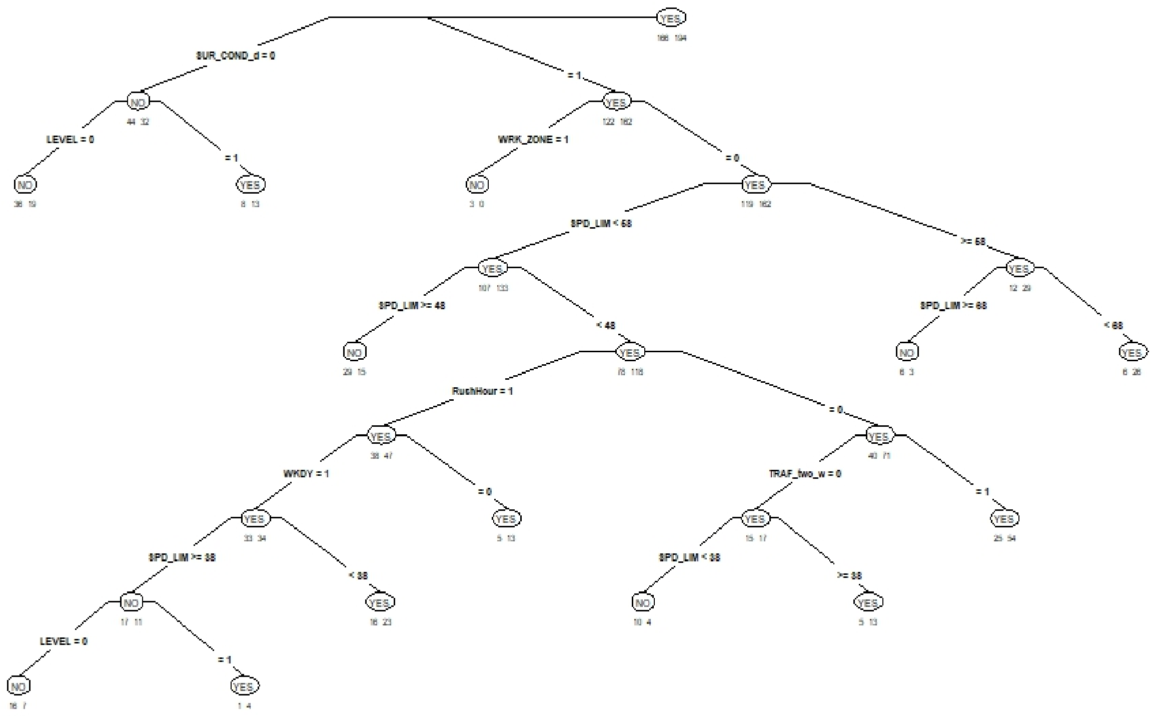
List of optimization techniques:

By comparing the values of accuracy, sensitivity, and specificity of all the techniques used. By reimplementing and changing the hyperparameter values we tried optimizing the values.

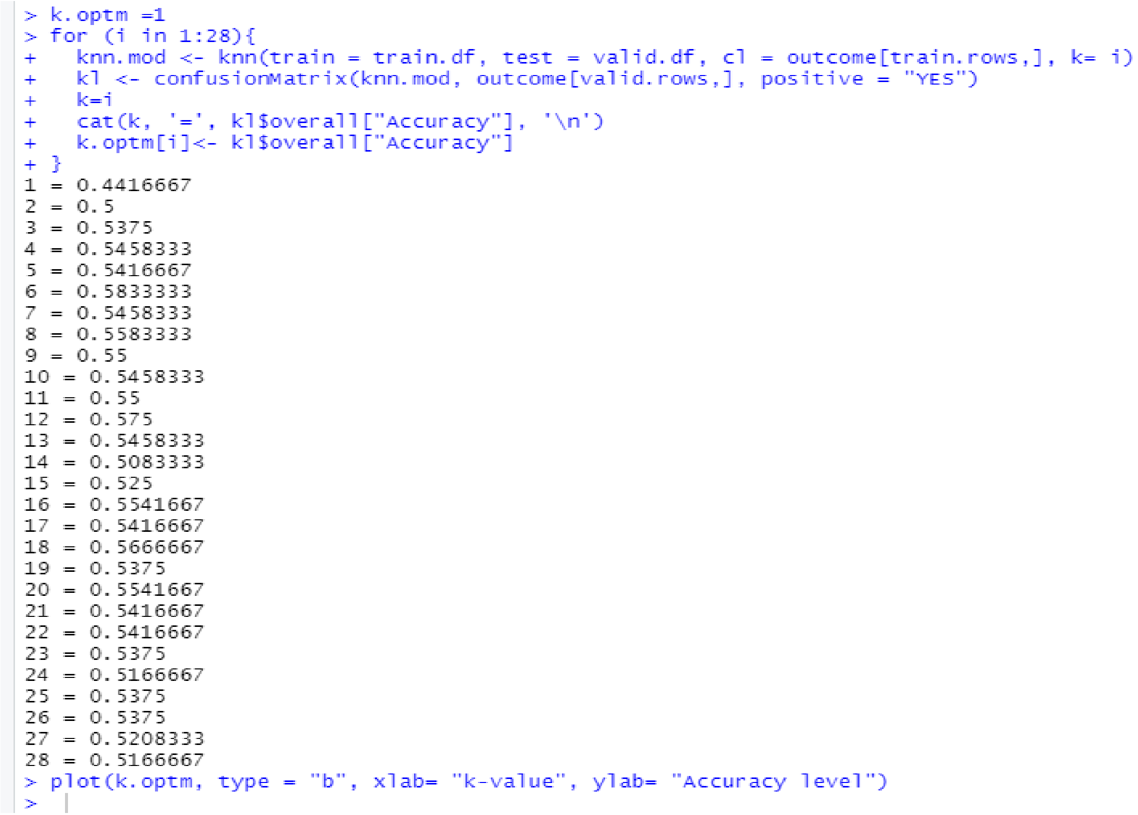
Firstly, for decision tree, we can try to optimize it by changing the depth values so we have added the cp value which gives impact on accuracy and specificity values, this when compared to the previous execution without mentioning the cp value.



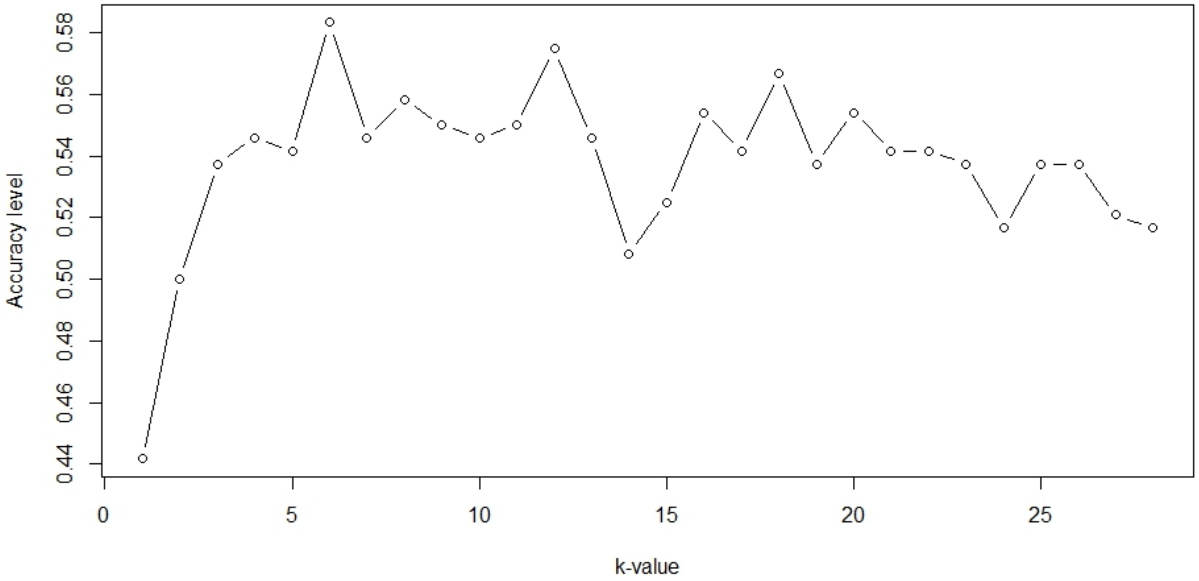
```
> full.ct.pred.train <- predict(full.ct, train.df, type = "class")
> full.ct.pred.valid <- predict(full.ct, valid.df, type = "class")
> confusionMatrix(full.ct.pred.valid, as.factor(valid.df$INJURY), positive = "YES")
Confusion Matrix and Statistics

          Reference
Prediction NO YES
       NO  60  49
       YES 66  65

               Accuracy : 0.5208
                 95% CI : (0.4556, 0.5855)
    No Information Rate : 0.525
    P-Value [Acc > NIR] : 0.5773

                  Kappa : 0.046

 Mcnemar's Test P-Value : 0.1357

            Sensitivity : 0.5702
            Specificity : 0.4762
         Pos Pred Value : 0.4962
         Neg Pred Value : 0.5505
             Prevalence : 0.4750
         Detection Rate : 0.2708
   Detection Prevalence : 0.5458
      Balanced Accuracy : 0.5232

       'Positive' Class : YES

> 
```



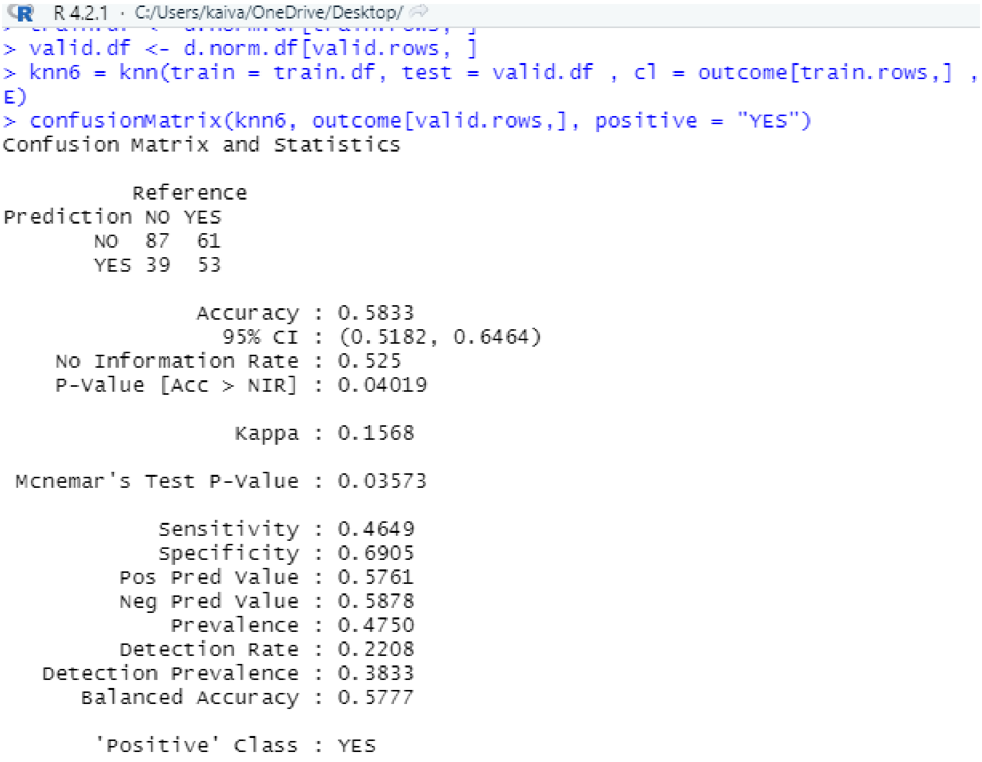
Similarly, we can calculate accuracy values in the KNN Algorithm for different values of K. When implemented we have noticed the accuracy value between K value 5 and K value 10 is high so when implemented By changing the K values, observing the changes in accuracy values. We observed that the accuracy value at K=6 is highest.



```
> k.optm =1
> for (i in 1:28){
+   knn.mod <- knn(train = train.df, test = valid.df, cl = outcome[train.rows,], k= i)
+   kl <- confusionMatrix(knn.mod, outcome[valid.rows,], positive = "YES")
+   k=i
+   cat(k, '=', kl$overall["Accuracy"], '\n')
+   k.optm[i]<- kl$overall["Accuracy"]
+ }
1 = 0.4416667
2 = 0.5
3 = 0.5375
4 = 0.5458333
5 = 0.5416667
6 = 0.5833333
7 = 0.5458333
8 = 0.5583333
9 = 0.55
10 = 0.5458333
11 = 0.55
12 = 0.575
13 = 0.5458333
14 = 0.5083333
15 = 0.525
16 = 0.5541667
17 = 0.5416667
18 = 0.5666667
19 = 0.5375
20 = 0.5541667
21 = 0.5416667
22 = 0.5416667
23 = 0.5375
24 = 0.5166667
25 = 0.5375
26 = 0.5375
27 = 0.5208333
28 = 0.5166667
> plot(k.optm, type = "b", xlab= "k-value", ylab= "Accuracy level")
> 
```



So the optimization values after the K value is changed to 6 is:



```
> train.df <- d.norm.df[train.rows, ]
> valid.df <- d.norm.df[valid.rows, ]
> knn6 = knn(train = train.df, test = valid.df , cl = outcome[train.rows,] ,
E)
> confusionMatrix(knn6, outcome[valid.rows,], positive = "YES")
Confusion Matrix and Statistics

          Reference
Prediction NO YES
       NO  87  61
       YES 39  53

               Accuracy : 0.5833
                 95% CI : (0.5182, 0.6464)
    No Information Rate : 0.525
    P-Value [Acc > NIR] : 0.04019

                  Kappa : 0.1568

 Mcnemar's Test P-Value : 0.03573

            Sensitivity : 0.4649
            Specificity : 0.6905
         Pos Pred Value : 0.5761
         Neg Pred Value : 0.5878
             Prevalence : 0.4750
         Detection Rate : 0.2208
   Detection Prevalence : 0.3833
      Balanced Accuracy : 0.5777

       'Positive' Class : YES
```

Experimental Values:

Non Optimized	Accuracy	Sensitivity	Specificity
Decision Tree	52.5	53.38	51.4
Logistic Regression	49.58	57.14	40.19
KNN	52.08	67.86	38.28
Naïve Bayes	50.42	72.81	30.16

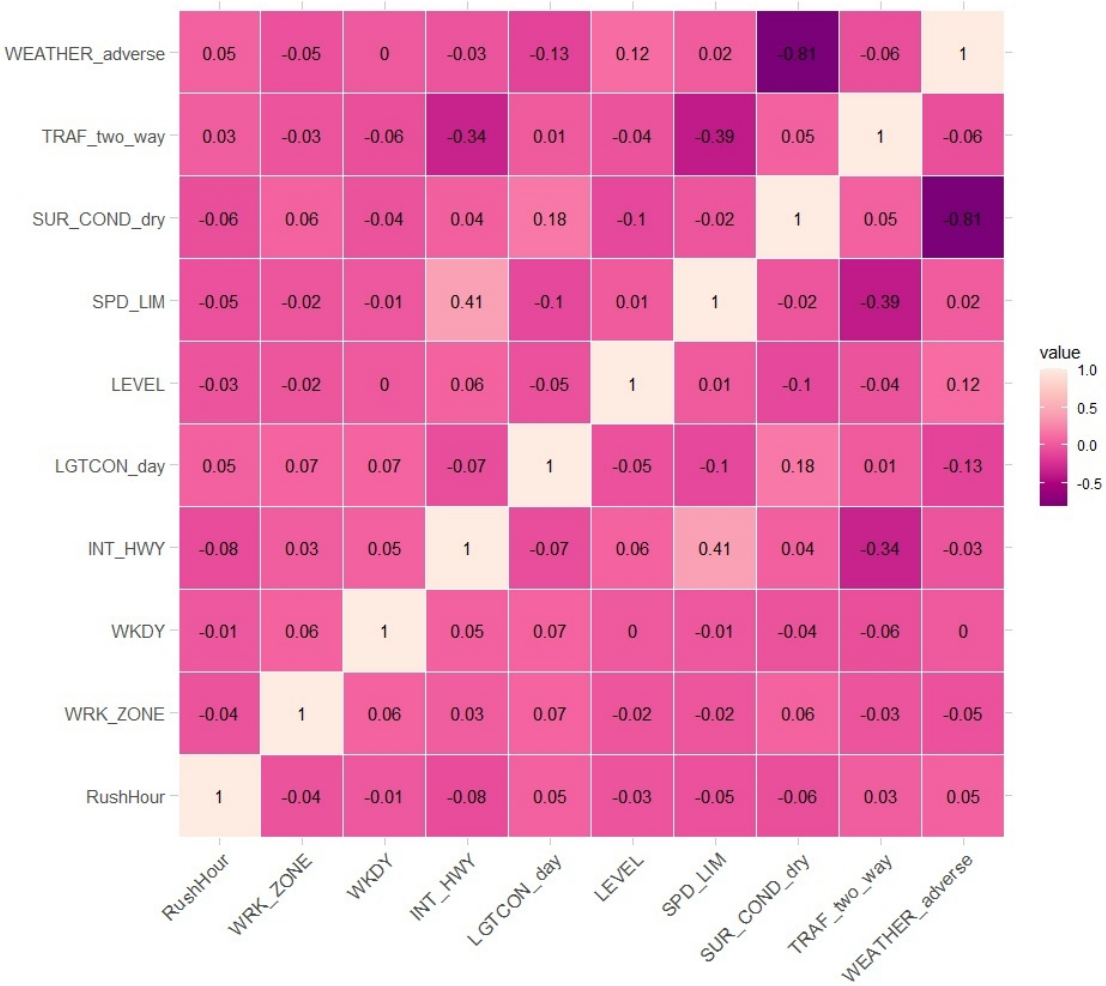
Optimized	Accuracy	Sensitivity	Specificity
Decision Tree (Cp=10)	52.08	57.02	47.62
KNN (K=6)	58.33	46.49	69.05

Discussion:

After exploring and implementing various optimization techniques like by changing the hyper parameter values. We can get the highest accuracy by implementing the KNN algorithm with K value 6 but here the sensitivity the specificity values don’t seem to be good. But at the decision tree with cp specified, all the values seem to be regulated. So, we think the decision tree with specified cp value as 10 will give the us the better true positive results.

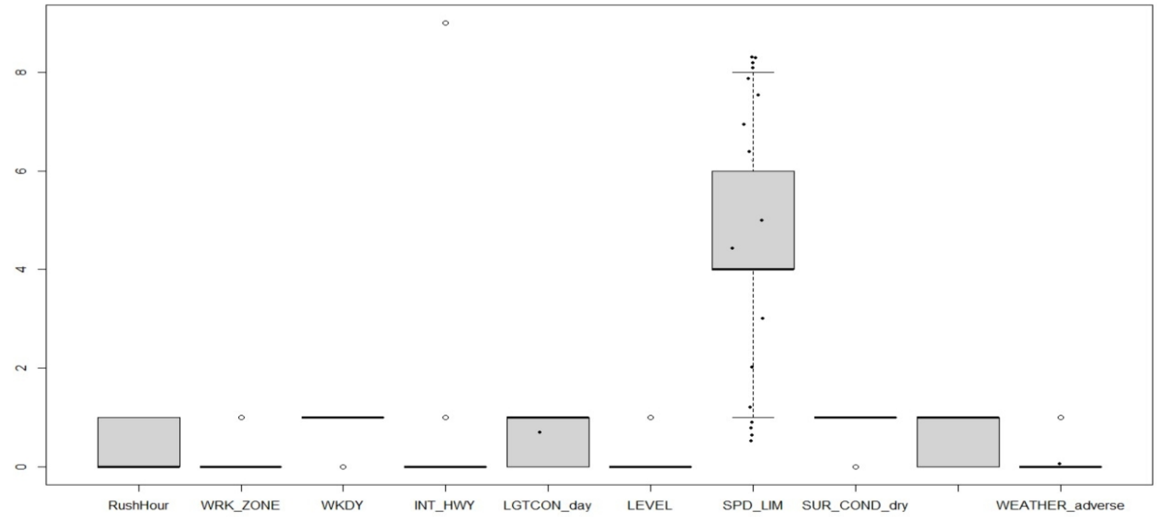
Visualization techniques used:

Here we have applied correlation to find out similarity between the unique features so, we can eliminate exploring any similar attributes and in turn narrowing down the search.

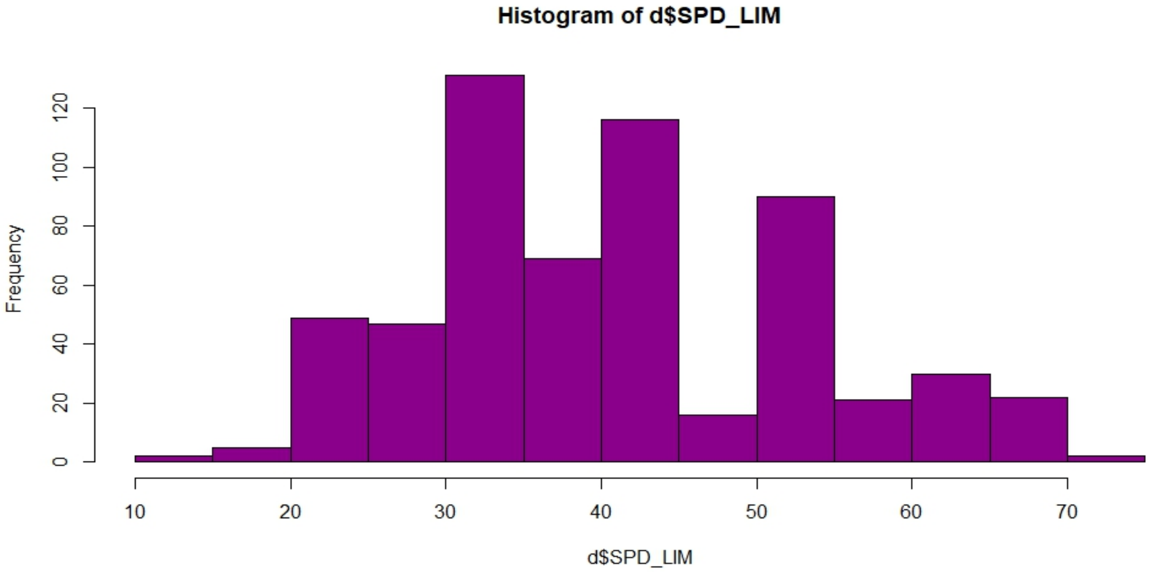


According above graph the similarities between the attributes is exceptionally low so, we can use all the attributes in predicting the severity of the accidents occurred. By considering all these features our goal is to predict the future possibility of accidents that might occur.

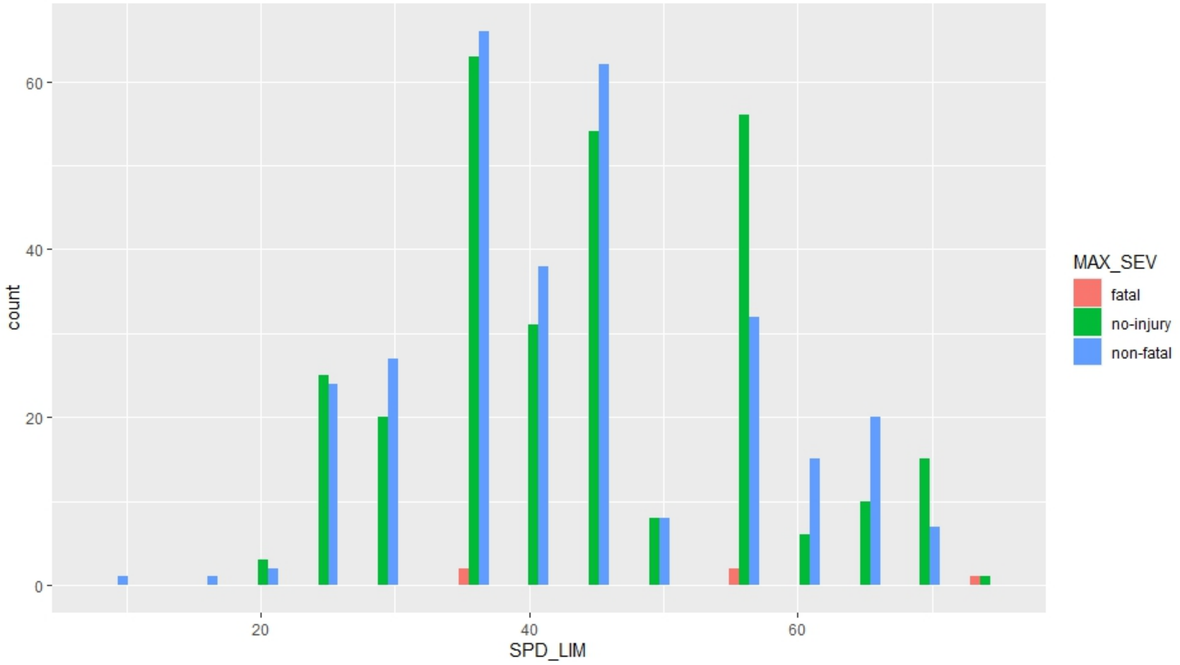
Outliers Detection: Our next process is to find if there are any outlining data present in the dataset. Outlining data might occur due to some conditions like road works or any other blockings that might occur due to accidents happened on that day on the road.



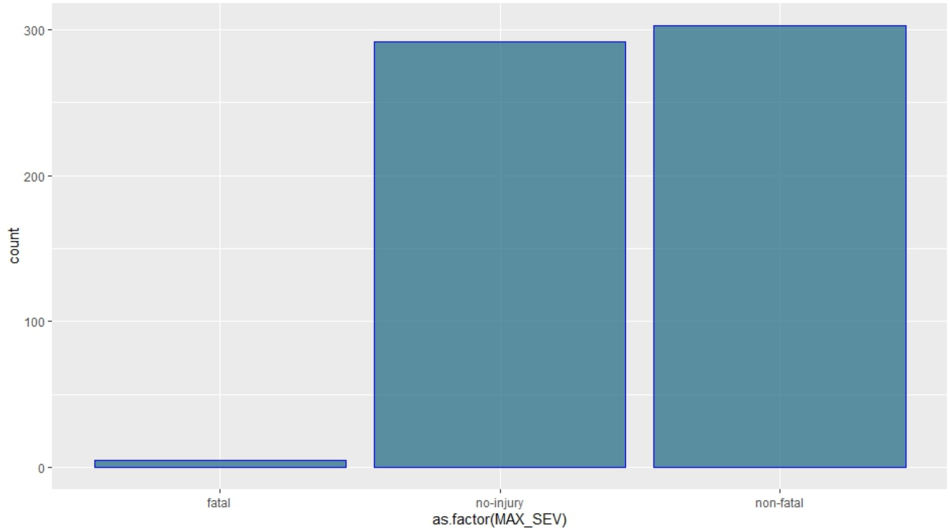
Histogram Analysis: Here we have compared the happening of injuries because of speed limit from below diagram the conclusion is that the frequency of accidents occurring is from the speed above 40



This histogram gives count of frequency of the accident severity when compared to speed limit



Bar Chart Analysis:



Conclusion and Future Works:

The severity of the accidents like fatal or non-fatal is identified from data set that we have taken from Kargil.com and we have used the decision tree and KNN algorithm. Attributes that affect the traffic accidents are environmental factors such as weather, traffic, characteristics of road network, rush hour etc. Prediction model can predict the severity based on traffic attributes which are time, day, weather and road surface etc, with an accuracy of 58.33%. In future we plan to incorporate other publicly available sources of data(e.g. demographic information and annual traffic reports) for the task of real-time accident prediction.

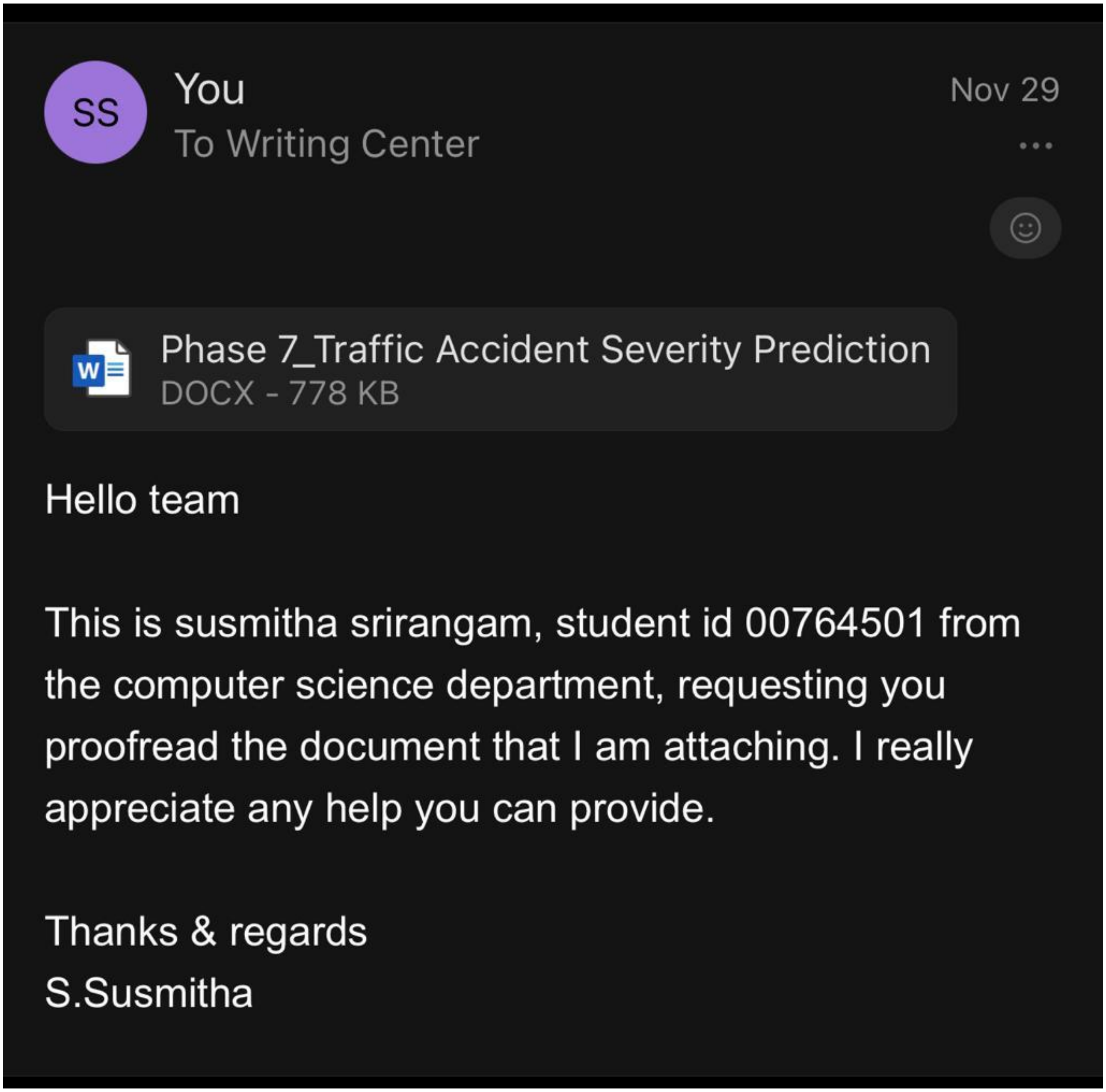
References:

1. https://www.researchgate.net/publication/335671728_Traffic_Accident_Severity_Prediction_Using_Naive_Bayes_Algorithm_-_A_Case_Study_of_Semarang_Toll_Road
2. https://arxiv.org/pdf/1909.09638.pdf
3. https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/stamp/stamp.jsp?tp=&arnumber=9795842
4. https://www.ijser.org/researchpaper/Analysis-of-Road-Accidents-using-Apriori-Naive-Bayes-and-K-Means.pdf
5. http://www.ijfrcsce.org/index.php/ijfrcsce/article/view/1027

Git Repository:

https://github.com/susmitha7599/TheCollective

Writing center review screen shot.



You

To Writing Center

Nov 29

Phase 7_Traffic Accident Severity Prediction

DOCX - 778 KB

Hello team

This is susmitha srirangam, student id 00764501 from the computer science department, requesting you proofread the document that I am attaching. I really appreciate any help you can provide.

Thanks & regards
S.Susmitha