

COMPUTATIONAL DATA ANALYSIS

PROJECT REPORT

CRASH REPORTING ANALYSIS: A MACHINE LEARNING APPROACH TO AUTOMATED CRASH REPORT TYPE PREDICTION

INTRODUCTION:

This project's main objective is to analyze a large dataset focusing on driver crash reports by utilizing machine learning approaches. This dataset captures a broad range of characteristics that are closely related to traffic accidents. The main goals of this dataset's careful evaluation are three categories: first, to extract valuable information from the data; second, to predict and classify the automated crash report type that happen; and third, to identify the critical elements that have a significant impact on the rate and type of crash occurrences. Through the use of advanced machine learning algorithms, the goal of this study is to derive useful information from the combination of features in the dataset. The ultimate goal is to identify patterns, correlations.

DATASET DESCRIPTION:

The dataset chosen for this project, titled "Crash Reporting -Drivers Dataset," focuses on motor vehicle collisions at Traffic. This data set is obtained from the website named Data.gov [1]. This dataset originates from the Automated Crash Reporting System (ACRS) managed by the Maryland State Police, collating reports provided by the Montgomery County Police, Gaithersburg Police, Rockville Police, or the Maryland-National Capital Park Police.

The dataset comprises 168,515 records and 43 columns has shown in Figure-1, featuring details of reported traffic incidents. Each entry includes specific information such as the report number,

agency name, ACRS (Automated Crash Reporting System) report type, crash date and time, location details including route type, road name, and geographical coordinates (latitude and longitude). Additional data encompasses factors like weather conditions, surface state, lighting, traffic control mechanisms, and substances involved.

For individuals involved, it provides a person ID, indication of fault, severity of injuries sustained, circumstances leading to the incident, distractions, and state of their driver's license. Vehicle-related data includes IDs, damage extent, impact locations, body type, movement direction, speed limit, year, make, and model. Information regarding equipment problems, driverless or parked vehicle status, and substance abuse among non-motorists is also available. It provides a valuable resource for detailed analysis and potential insights into the circumstances surrounding traffic collisions.

DATA PRE-PROCESSING:

The data is set to pre-process for analysis. Columns with information that might not contribute significantly to the analysis were removed, like report numbers, off-road descriptions, municipality details, and identifiers like person IDs and vehicle IDs in Figure-2.

Duplicates in the dataset were identified and removed to ensure unique records. The column names were standardized to simplify analysis, ensuring consistency and clarity across the dataset . For example, Column name is ACRS Report Type is converted as ACRS_Report_Type has shown in Figure 3. Categorical columns were converted to numeric format for statistical analysis, using a process known as one-hot encoding to transform categorical data into a numerical format that the computer can understand seen in Figure 4.

To understand relationships between variables, correlation analysis was performed. This involved calculating the correlation matrix to see how each variable relates to others and particularly to the target variable, ACRS Report Type shown in Figure 5. The correlation matrix helps identify which variables might be more strongly related to the type of accident reports. Also selected top five features which have high correlation with target variable shown in Figure 6.

To simplify further analysis, a subset of the data was taken (10% of the original) for computational efficiency while maintaining representativeness. This subset was split into training and testing sets (70% and 30%, respectively) to build and validate predictive models.

MODELS USED:

The decision to use the decision tree algorithm is an intuitive choice for classification tasks and offers insights into feature importance. The code utilizes the 'rpart' package to implement the decision tree algorithm. The 'rpart. Control' function sets parameters for tree complexity, such as complexity parameter ($cp=0.00001$) and minimum split($min_split=1$), ensuring the tree's adequacy for prediction. The 'rpart' function trains the model on the provided training dataset ('train_data').

The logistic regression algorithm is selected for effectiveness in binary classification tasks. The 'glm' function is utilized to implement logistic regression. It fits the model to the training dataset ('train_data'), specifying the family argument as "binomial" for binary classification. The model is trained on the provided dataset.

The trained Naive Bayes model is evaluated on the test set ('test_data') using evaluation metrics such as accuracy and precision. The confusion matrix is generated to assess the model's

classification performance. It demonstrates the counts of correct and incorrect predictions for each class.

MODEL COMPARISION:

The Table-1 shows a comparison of the accuracy, Confusion matrix, ROC curve and precision of three machine learning models: decision tree, Naive Bayes, and logistic regression. The Naive Bayes model has the highest accuracy (0.9986), followed by decision tree model has the accuracy (0.8046), and the logistic regression model (0.2070). The predicted accuracy of the Naive Bayes model is significantly higher than the actual accuracy. The Naive Bayes model has the highest precision (1.0000) for all three classes, followed by the logistic regression model and the decision tree model.

In Figure 7, If the Injury_Severity variable value is less than 2.5, then the prediction is that the ACRS_Report_Type3 is equal to 3. If the value of this variable is greater than or equal to 2.5, then the next decision point is based on the value of the Vehicle_First_Impact variable. If the value of this variable is equal to 13, then the prediction is that the ACRS_Report_Type3 is equal to 3. If the value of this variable is not equal to 13, then the next decision point is based on the value of the Vehicle_Second_Impact variable. If the value of this variable is equal to 14, then the prediction is that the ACRS_Report_Type3 is equal to 3. If the value of this variable is not equal to 14, then the prediction is that the ACRS_Report_Type3 is not equal to 3.

The Figure 8 shows a graph of the relative error of a tree model as a function of the x-axis. The x-axis is labeled "X-val", which stands for "Cost complexity". The y-axis is labeled "Relative Error". The relative error of the tree model increases as the cost complexity value increases. This means

that the tree model is less accurate for larger values of the x-variable. The Naïve Bayes has the highest AUC (Area Under the ROC Curve), followed by the Decision Tree and the logistic regression model.

This suggests that the decision tree model is better at predicting the overall accuracy of a model, while the Naive Bayes model is better at predicting the accuracy of individual classes. Additionally, the analysis shows that the logistic regression model has the lowest accuracy and precision of the three models. This suggests that the logistic regression model may not be as well-suited for this particular task. This means that the Naïve Bayes is the best performing model out of the three.

DISCUSSION:

A deeper tree will be more accurate, but it may also be more likely to over fit the training data. A minimum leaf size will help to prevent overfitting, but it may also make the tree less accurate. A Naïve Bayes used to smooth the probabilities of the features. This can help to improve the accuracy of the model, but it can also make the model more likely to over fit the training data. A Logistic regression technique used to prevent overfitting.

There are a few possible reasons why the relative error of the tree model increases as the x-validation value increases. One possibility is that the tree model is overfitting to the training data. This means that the tree model is learning the noise in the training data, rather than the true patterns in the data. Another possibility is that the tree model is not complex enough to capture the true patterns in the data.

A Decision Tree is Easy to interpret and Can also handle both numerical and categorical features. But, Can be prone to overfitting and also Can be slow to train on large datasets. Naive Bayes is Fast to train and Can also handle both numerical and categorical features. But, Can be biased if the features are not independent. Logistic regression is to be interpreted using coefficients and can handle both numerical and categorical features. But, Can be sensitive to outliers.

CONCLUSION:

The conclusion is that the Naïve Bayes is the best performing classification model out of the three. It has the highest AUC (Area under the ROC Curve), which means that it is the best at distinguishing between positive and negative instances. The logistic regression model and the Decision Tree model also perform well, but they are not as good as the Naïve Bayes.

This also shows that the Naïve Bayes is more robust to noise in the data than the other two models. The Target variable as ACRS_Report_Type classified in to 3 class as property damage, Injury, fatal crash, is better classified in Naive Bayes then the other models.

Overall, it suggests that the Naive Bayes classifier is a more accurate and reliable method for classification than the decision tree and logistic model. However, it is important to note that the accuracy of a classifier can vary depending on the features dataset used. As, in this we use a best correlated features.

FIGURES AND TABLES

```

$ Report Number      : chr [1:168515] "MCP3040003N" "EJ/8850038" "MCP2009002G" "MCP3201004C"
$ Local Case Number  : num [1:168515] 1.9e+08 2.3e+08 2.3e+08 2.3e+08 2.3e+08
$ Agency Name        : chr [1:168515] "Montgomery County Police" "Gaithersburg Police Depart" "Montgomery County Police" $ ACRS
Report Type          : chr [1:168515] "Property Damage Crash" "Property Damage Crash" "Property Damage Crash" "Property Damage Crash"
...
$ Crash Date/Time    : chr [1:168515] "05/31/2019 03:00:00 PM" "07/21/2023 05:59:00 PM" "07/20/2023 03:10:00 PM" "07/23/2023
12:10:00 PM"
$ Route Type         : chr [1:168515] NA "Maryland (State)" "Maryland (State)" "County"
$ Road Name          : chr [1:168515] NA "FREDERICK RD" "GEORGIA AVE" "CRYSTAL ROCK DR"
$ Cross-Street Type  : chr [1:168515] NA "Unknown" "Maryland (State)" "County"
$ Cross-Street Name  : chr [1:168515] NA "WATKINS MILL RD" "NORBECK RD" "WATERS LANDING DR"
$ Off-Road Description : chr [1:168515] "PARKING LOT OF 3215 SPARTAN RD" NA NA NA
$ Municipality       : chr [1:168515] NA "N/A" "N/A" "N/A"
$ Related Non-Motorist : chr [1:168515] NA NA NA NA
$ Collision Type      : chr [1:168515] "OTHER" "STRAIGHT MOVEMENT ANGLE" "STRAIGHT MOVEMENT ANGLE"
"STRAIGHT MOVEMENT ANGLE"
$ Weather            : chr [1:168515] "CLEAR" "CLEAR" "CLEAR" "CLEAR" ...
$ Surface Condition  : chr [1:168515] NA "DRY" "DRY" "DRY"
$ Light              : chr [1:168515] "DAYLIGHT" "DAYLIGHT" "DAYLIGHT" "DAYLIGHT"
$ Traffic Control    : chr [1:168515] "N/A" "TRAFFIC SIGNAL" "TRAFFIC SIGNAL" "NO CONTROLS"
$ Driver Substance Abuse : chr [1:168515] "UNKNOWN" "NONE DETECTED" "NONE DETECTED" "NONE DETECTED"
$ Non-Motorist Substance Abuse : chr [1:168515] NA NA NA NA ...
$ Person ID          : chr [1:168515] "DE2A24CD-7919-4F8D-BABF-5B75CE12D21E" "E7058A8E-4F18-4D2A-954E-04A099CFED12"
"2B404D6D-8DB5-4CB6-9E71-9F1B8D0A8925" "637D8107-0381-4B8D-848A-B4A93B4D53CE"
$ Driver At Fault    : chr [1:168515] "Yes" "No" "Yes" "Yes"
$ Injury Severity    : chr [1:168515] "NO APPARENT INJURY" "NO APPARENT INJURY" "NO APPARENT INJURY" "NO"
$ Circumstance       : chr [1:168515] "N/A" "N/A" "N/A" "N/A"
$ Driver Distracted By : chr [1:168515] "UNKNOWN" "NOT DISTRACTED" "NOT DISTRACTED" "LOOKED BUT DID NOT SEE"
$ Drivers License State : chr [1:168515] NA "MD" "MD" "MD"
$ Vehicle ID         : chr [1:168515] "165AD539-A8C8-4004-AF73-B7DCAAA8B3CC" "1C3C3E2F-9A23-4ED0-9BB3-
B6C370D99C37" "0483CE47-E0FC-4BCA-BAB0-B7541820FEE6" "4406AA84-07F8-45F4-88A2-09AD89AC9AAF"
$ Vehicle Damage Extent : chr [1:168515] "SUPERFICIAL" "DISABLING" "FUNCTIONAL" "FUNCTIONAL"
$ Vehicle First Impact Location : chr [1:168515] "ONE OCLOCK" "THREE OCLOCK" "TWELVE OCLOCK" "TWELVE OCLOCK" ...
$ Vehicle Second Impact Location : chr [1:168515] "ONE OCLOCK" "TWO OCLOCK" "TWELVE OCLOCK" "TWELVE OCLOCK"
$ Vehicle Body Type   : chr [1:168515] "PASSENGER CAR" "PASSENGER CAR" "PICKUP TRUCK" "PASSENGER CAR"
$ Vehicle Movement    : chr [1:168515] "PARKING" "MAKING LEFT TURN" "ACCELERATING" "STARTING FROM LANE"
$ Vehicle Continuing Dir : chr [1:168515] "North" "East" "North" "East"
$ Vehicle Going Dir   : chr [1:168515] "North" "South" "North" "East"
$ Speed Limit         : num [1:168515] 15 40 35 40 35 30 25 35 35 30
$ Driverless Vehicle  : chr [1:168515] "No" "No" "No" "No"
$ Parked Vehicle      : chr [1:168515] "No" "No" "No" "No"
$ Vehicle Year        : num [1:168515] 2004 2011 2019 2016 2016
$ Vehicle Make        : chr [1:168515] "HONDA" "GMC" "FORD" "KIA"
$ Vehicle Model       : chr [1:168515] "TK" "TK" "F150" "SW"
$ Equipment Problems  : chr [1:168515] "UNKNOWN" "NO MISUSE" "NO MISUSE" "NO MISUSE"

```

Figure-1 Dataset with Rows: 168515 Columns: 43

```

library(dplyr)
crash_data <- select(crash_data, -c(`Report Number`, `Off-Road Description`,
`Municipality`, `Related Non-Motorist`, `Non-Motorist Substance Abuse`, `Person ID`,
`Circumstance`, `Vehicle ID`, `Location`))
crash_data

```

A tibble: 168,515 × 34

Figure-2 Removal of unnecessary columns

[1] "Local_Case_Number"	"Agency Name"	"ACRS_Report_Type"
[4] "Crash Date/Time"	"Route Type"	"Road Name"
[7] "Cross-Street Type"	"Cross-Street Name"	"Collision_type"
[10] "Weather_rep"	"Surface Condition"	"Light_rep"
[13] "Traffic_Control"	"Driver Substance Abuse"	"Driver At Fault"
[16] "Injury_Severity"	"Driver_distracted_by"	"Drivers License State"
[19] "Vehicle_damage_extent"	"Vehicle_First_Impact"	"Vehicle_Second_Impact"
[22] "Vehicle Body Type"	"Vehicle_movement"	"Vehicle Continuing Dir"
[25] "Vehicle Going Dir"	"Speed_limit"	"Driverless Vehicle"
[28] "Parked Vehicle"	"Vehicle Year"	"Vehicle Make"
[31] "Vehicle Model"	"Equipment_problems"	"Latitude"
[34] "Longitude"		

Figure-3 Rename column names

```

$ Local_Case_Number      : num 77494 77459 77527 77544 77257 ...
$ Agency Name           : num 2 6 6 6 6 6 2 6 6 6 ...
$ ACRS_Report_Type      : num 3 3 3 3 3 3 2 3 3 3 ...
$ Crash Date/Time       : num 42038 41803 42474 42689 39818 ...
$ Route Type            : num 4 4 1 1 1 4 4 4 1 4 ...
$ Road Name             : num 1075 1120 697 1910 3082 ...
$ Cross-Street Type     : num 9 4 1 1 1 10 9 1 1 4 ...
$ Cross-Street Name     : num 5792 3671 5783 857 70 ...
$ Collision_type        : num 18 18 18 5 11 11 17 11 11 18 ...
$ Weather_rep           : num 3 3 3 4 3 3 3 3 3 3 ...
$ Surface Condition     : num 1 1 1 1 1 1 1 1 1 1 ...
$ Light_rep             : num 5 5 5 5 5 6 1 5 5 5 ...
$ Traffic_Control       : num 9 9 3 9 3 3 3 4 9 9 ...
$ Driver Substance Abuse : num 10 10 10 10 10 10 10 2 10 10 ...
$ Driver At Fault       : num 1 3 3 3 1 3 3 3 1 1 ...
$ Injury_Severity       : num 2 2 2 2 2 2 4 2 2 2 ...
$ Driver_distracted_by  : num 10 10 8 10 10 10 17 7 10 10 ...
$ Drivers License State : num 28 28 28 8 28 28 28 28 28 28 ...
$ Vehicle_damage_extent : num 2 3 3 3 3 2 1 7 3 2 ...
$ Vehicle_First_Impact  : num 12 13 13 13 10 13 13 13 10 4 ...
$ Vehicle_Second_Impact : num 14 13 13 13 10 13 4 13 10 3 ...
$ Vehicle Body Type     : num 20 21 20 1 28 1 20 1 31 20 ...
$ Vehicle_movement      : num 7 1 19 7 21 18 10 18 21 1 ...
$ Vehicle Continuing Dir : num 1 2 1 5 1 3 2 2 5 1 ...
$ Vehicle Going Dir     : num 3 2 1 2 1 3 2 2 2 1 ...
$ Speed_limit           : num 9 8 9 8 7 6 8 7 8 8 ...
$ Driverless Vehicle    : num 1 1 1 1 1 1 1 1 1 1 ...
$ Parked Vehicle        : num 1 1 1 1 1 1 1 1 1 1 ...
$ Vehicle Year          : num 80 88 85 85 85 83 82 91 82 86 ...
$ Vehicle Make          : num 522 371 801 1519 481 ...
$ Vehicle Model         : num 5223 2312 5043 5223 1094 ...
$ Equipment_problems    : num 6 6 6 6 6 6 11 6 6 5 ...
$ Latitude              : num 58901 45357 65078 61518 6915 ...
$ Longitude             : num 10998 44582 3923 15187 56706 ...
$ ACRS_Report_Type1     : logi FALSE FALSE FALSE FALSE FALSE ...
$ ACRS_Report_Type2     : logi FALSE FALSE FALSE FALSE FALSE ...
$ ACRS_Report_Type3     : logi TRUE TRUE TRUE TRUE TRUE ...
$ Vehicle_damage_extent : num 2 3 3 3 3 2 1 7 3 2 ...
$ Injury_Severity       : num 2 2 2 2 2 2 4 2 2 2 ...
$ Vehicle_First_Impact  : num 12 13 13 13 10 13 13 13 10 4 ...
$ Vehicle_Second_Impact : num 14 13 13 13 10 13 4 13 10 3 ...
$ encoded_data$ACRS_Report_Type : num 3 3 3 3 3 3 2 3 3 3 ...
$ Vehicle_damage_extent : num 2 3 3 3 3 2 1 7 3 2 ...
$ Injury_Severity       : num 2 2 2 2 2 2 4 2 2 2 ...
$ Vehicle_First_Impact  : num 12 13 13 13 10 13 13 13 10 4 ...
$ Vehicle_Second_Impact : num 14 13 13 13 10 13 4 13 10 3 ...
$ encoded_data$ACRS_Report_Type : num 3 3 3 3 3 3 2 3 3 3 ...

```

Figure-4 Categorical to Numeric conversion of features

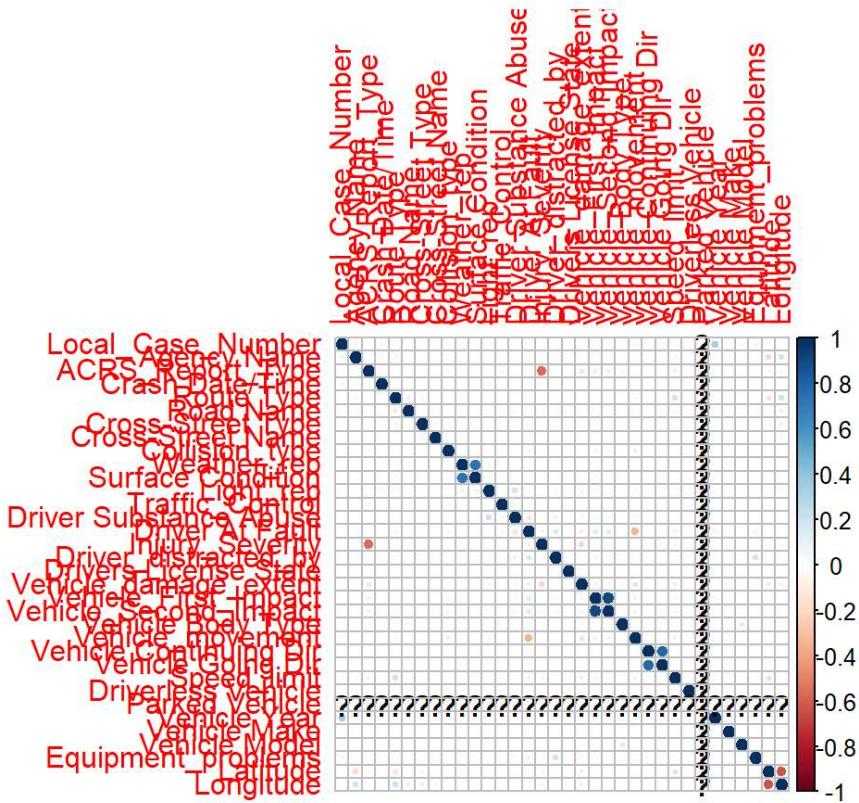


Figure-5 Correlation Matrix

```
# Simple named list:
list(mean = mean, median = median)

# Auto named with `tibble::lst()`:
tibble::lst(mean, median)

# Using lambdas
list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
Warning: the standard deviation is zero
Warning: the standard deviation is zero
[1] "ACRS_Report_Type"      "Injury_Severity"
"Vehicle_damage_extent"
[4] "Vehicle_Second_Impact" "Vehicle_First_Impact"
```

Figure-6 Feature Selection

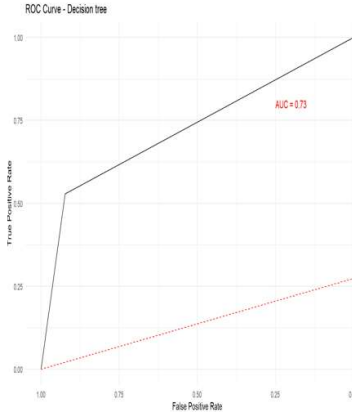
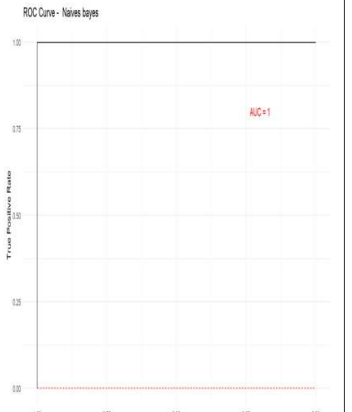
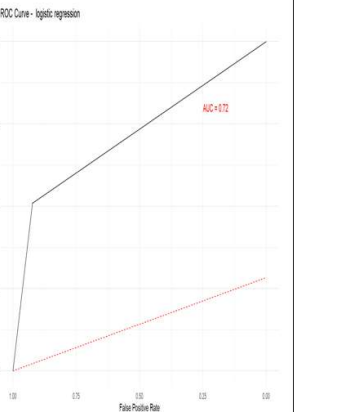
Parameters	Decision Tree	Naïve Bayes	Logistic Regression
Accuracy	0.8046004	0.9986059	0.2070167
Confusion matrix	<pre> actual predictions 1 2 3 1 0 0 0 2 1 913 16 3 12 812 2550 </pre>	<pre> actual predictions 1 2 3 1 13 2 4 2 0 1723 0 3 0 0 2562 </pre>	<pre> actual predictions 1 2 3 FALSE 12 846 2566 TRUE 1 879 0 </pre>
ROC Curve			
Precision	<pre> 1 2 3 NaN 0.9817204 0.7557795 </pre>	<pre> 1 2 3 0.6842105 1.0000000 1.0000000 </pre>	<pre> FALSE TRUE 0.003504673 0.998863636 </pre>

Table-1 Comparisons of models

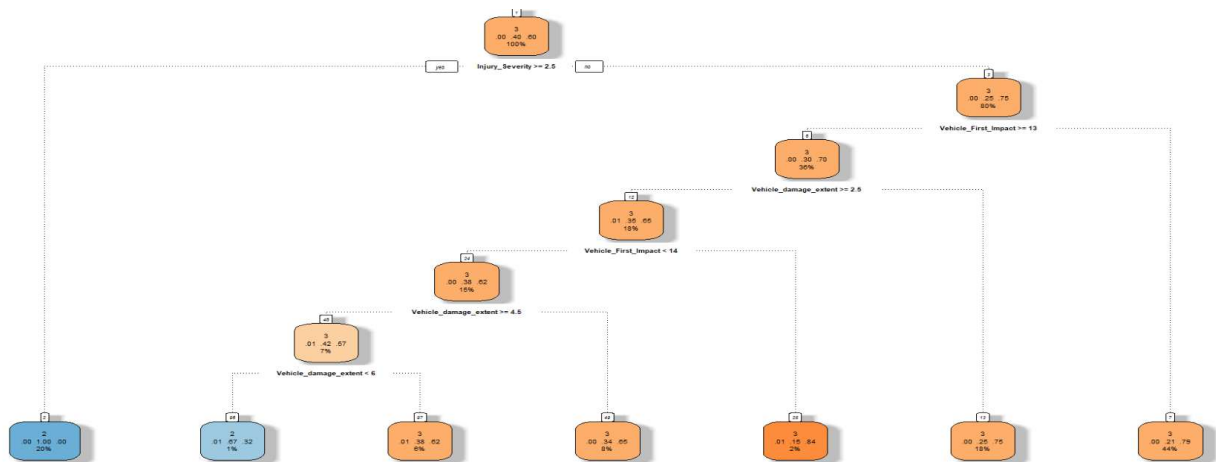


Figure-7 Tree structure

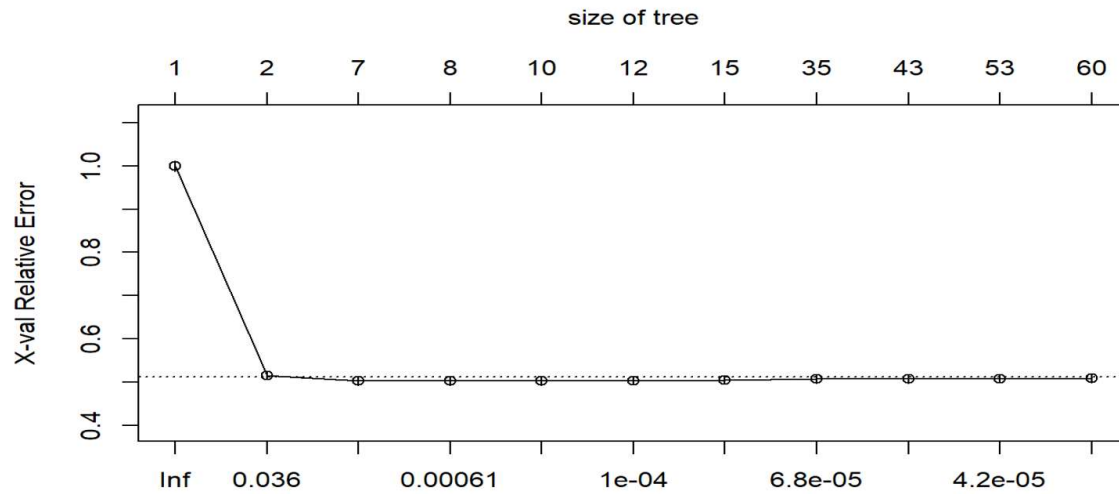


Figure-8 Plot cost complexity vs Error

Citation of Dataset:

[1] <https://data.montgomerycountymd.gov/Public-Safety/Crash-Reporting-Drivers-Data/mmzv-x632>