Fatality Prediction on Car Crash Data

Amit Dighe | Bhushan Sontakke | Neel Patel | Smit Vasani | Dr.James Foulds

Department of Information Systems

IS 733 Data Mining

***Abstract*-** Safety of people is the priority of today's engineers and researchers while designing a car. Accidents is an unavoidable circumstance even after full loaded safety. Patterns involved in car accidents can be used to detect what would be the fatality rate of various car accidents. This could be helpful to enhance the traffic safety and accident control policies. This project, presents some models to predict fatality that occurred during car accidents. The accuracy of fatality prediction on car crash data is important implication for initial rule making for drivers and passengers safety. Model prediction was conducted with various parameters such as speed of car, seat belt, driver or passenger position, accident from front or not, driver age. The goal of this project was to examine fatality rate by considering various factors given in the car crash dataset and find out actionable knowledge so that based on this knowledge, further steps can be conducted with regards to passenger and driver safety. [3]

I. Introduction and Motivation

Fatalities and injuries caused due to car accidents qualify among the top 5 causes of deaths in most countries. Nearly 1.3 million people die in road crashes each year. On average, they are around 3287 deaths per day. [7] To reduce the number of car crashes in the United States, the government makes the data available for analysis. This data can be used by independent individuals or agencies to come up with various trends to showcase the impact of safety measures in an accident. This data consists primarily of two entities, namely: the occupants and the vehicle. The occupant data comprises of the number of occupants in the car, the status of each occupant, condition of the driver etc. Whereas, the vehicle data consists of details about the make of the vehicle, the condition of the vehicle after accident, safety features in the vehicle and the deployment status of the safety features.

II. Background and related work

Due to the rapid popularity in masses about the importance of data analytics, there have been many research conducted in this field by numerous researchers. Some of the related research are as listed below:

Huelke et al. takes into consideration the crash investigation files available at the University of Michigan Transportation institution (UMTRI). This research focuses on reviewing the deployed airbags in the car accidents where the principal direction of force was at 11-1 o'clock. They found out that of the total drivers involved in frontal crashes 83% were having seat belts on them.

Also, the drivers that have seat belts restrained on them at the time of the accident have higher chances of upper extremity injuries. [2]

Chong et al. used traffic accident data to predict the severity of injuries. The researchers used three primary methods for prediction, they are: neural networks trained hybrid learning approaches, decision trees and a concurrent hybrid model involving decision trees and neural networks. They found out that the actual speed of the vehicle during an accident would have proven to be a principal factor for accurate predictions, as 67% of the rows had unknown speeds recorded. [3]

Our work primarily tries to predict if the driver or the passenger in the car is injured fatally or not. This prediction is mainly done based on factors such as speed of the vehicle, availability of safety measures such as seatbelt and airbags and whether these safety measures were deployed or not. Our prediction methods primarily consisted of Decision trees, K nearest neighbor method and Support Vector Machines. Our focus is on identifying various patterns which could help in reducing the car occupants suffering from fatal injuries.

III. Data Source

This dataset is a subset of dataset obtained from national inflation factor by NASS (National Automotive Sampling System).

A. Description of data set

The main dataset consists of 417,670 cases. This dataset is a subset of the main dataset. Dataset contains 15 variables and 26,217 instances for accidents and factors influencing the accident fatalities in United States for year 1997-2002. [3] This dataset has following factors:

Dvcat: Speed of the vehicle, Weight: Weight of the person involved in accident, Airbag: Whether the vehicle had an airbag or not, Seatbelt: whether the person was belted or not belted, Frontal: whether the impact was from front or not, Sex: Sex of the person, ageOFocc: Age, yearacc: Year of accident, yearVeh: Year of Vehicle, abcat: whether the airbag was available or not, occRole: whether the person was a passenger or driver, deploy: airbag deployed or not, injSeverity: divided into 6 levels i.e. 0: none, 1: possible injury, 2: no incapacity, 3: incapacity, 4: killed; 5: unknown, 6: prior death; caseid and the outcome variable as Dead: The person is dead or alive.

In this dataset 95% of the instances are for alive and the rest 5% is for dead.

Our task was to build models that could classify the person's status after accident as dead/alive accurately. This will also help to learn the relations between other influences like airbag, seatbelt, age etc. on accident fatalities

B. Data Preprocessing

Data cleaning is the most important part of any analysis. This dataset had no conflicts between any columns as the data was in categorical form. The dataset columns had vague nomenclature which was processed and were given easy understanding names. For e.g. dvcat: Vehicle.Speed, dead: Condition, ageOfocc: Occupant.Age, Yearacc: Accident.Year, and occRole:

Occupant.Role (fig 2). Importing dataset caused few rows to change their format which were replaced by correct values. E.g.: In speed column 1-9km/hr was replaced with 1-9, 24-Oct was replaced with 10-24. The dataset consisted of 154 NA values i.e. 0.5% of the total dataset which were removed (fig 1). Next step before performing feature selection, we saw the overall distribution of the data. Total we had 26216 observations in our dataset, out of which, 25036 were in Alive class and the remaining 1180 were in Dead class. Because of this imbalance in the number of instances, we applied SpreadSubSample filter on our data to equalize the Alive and Dead classes instances.

IV. Feature Selection

After carefully selecting 8 predictors from 15, we then performed Collinearity check to ensure that these predictors are not related

to each other. Fig 6 shows the results for the collinearity check.

In Injury Severity column, there are 6 levels as described in description. 6 which means prior death had only 2 instances and 5 which described unknown factor of injury had 135 instances. Number of instances were very less to predict according to the Injury Severity levels and also Injury Severity and Condition column are highly correlated to each other. So we decided to drop Injury Severity column and go with Condition column.

```
#Data preprocessing for Sex column
levels(DA$Sex) <- c(levels(DA$Sex), "M")
DA$Sex[DA$Sex=="m"]   <- "M"

levels(DA$Sex) <- c(levels(DA$Sex), "F")
DA$Sex[DA$Sex=="f"]   <- "F"

DA$Sex <- factor(DA$Sex)

#Data preprocessing for Occupant.Role column
levels(DA$Occupant.Role) <- c(levels(DA$Occupant.Role), "Driver")
DA$Occupant.Role[DA$Occupant.Role=="driver"]   <- "Driver"

levels(DA$Occupant.Role) <- c(levels(DA$Occupant.Role), "Passanger")
DA$Occupant.Role[DA$Occupant.Role=="pass"]   <- "Passanger"

DA$Occupant.Role <- factor(DA$Occupant.Role)

#Command to remove NA's from the D
DA <- na.omit(DA)


str(DA)
attach(D)
```

Fig 1: Removed NA values

```r
colnames(DA)[1] <- "Vehicle.Speed"
colnames(DA)[2] <- "Condition"
colnames(DA)[3] <- "Airbag"
colnames(DA)[4] <- "Seatbelt"
colnames(DA)[5] <- "Frontal"
colnames(DA)[6] <- "Sex"
colnames(DA)[7] <- "Occupant.Age"
colnames(DA)[8] <- "Occupant.Role"
colnames(DA)[9] <- "Airbag.Status"

#Changing column datatypes
DA$Frontal <- as.factor(DA$Frontal)
DA$Airbag.Status <- as.factor(DA$Airbag.Status)

#Data preprocessing for Vehicle.Speed column
levels(DA$Vehicle.Speed) <- c(levels(DA$Vehicle.Speed), "1-9")
DA$Vehicle.Speed[DA$Vehicle.Speed=="1-9km/h"]  <- "1-9"

levels(DA$Vehicle.Speed) <- c(levels(DA$Vehicle.Speed), "10-24")
DA$Vehicle.Speed[DA$Vehicle.Speed=="24-Oct"]  <- "10-24"

DA$Vehicle.Speed <- factor(DA$Vehicle.Speed)

DA$Vehicle.Speed <- factor(DA$Vehicle.Speed, levels=c("1-9", "10-24", "25-39", "40-54", "55+"))

#Data preprocessing for Airbag column
levels(DA$Airbag) <- c(levels(DA$Airbag), "Airbag")
DA$Airbag[DA$Airbag=="airbag"]  <- "Airbag"

levels(DA$Airbag) <- c(levels(DA$Airbag), "No Airbag")
DA$Airbag[DA$Airbag=="none"]  <- "No Airbag"

DA$Airbag <- factor(DA$Airbag)


#Data preprocessing for Seatbelt column
levels(DA$Seatbelt) <- c(levels(DA$Seatbelt), "Belted")
DA$Seatbelt[DA$Seatbelt=="belted"]  <- "Belted"

levels(DA$Seatbelt) <- c(levels(DA$Seatbelt), "Not Belted")
DA$Seatbelt[DA$Seatbelt=="none"]  <- "Not Belted"

DA$Seatbelt <- factor(DA$Seatbelt)
```

Fig 2: Process for change column names

	X	dvcat	weight	dead	airbag	seatbelt	frontal	sex	ageOFocc	yearacc	yearVeh	abcat	occRole	deploy	injSeverity	caseid
1	1	25-39	25.069	alive	none	belted	1	f	26	1997	1990	unavail	driver	0	3	2:3:1
2	2	10-24	25.069	alive	airbag	belted	1	f	72	1997	1995	deploy	driver	1	1	2:3:2
3	3	10-24	32.379	alive	none	none	1	f	69	1997	1988	unavail	driver	0	4	2:5:1

Fig 3: Before processed data

Vehicle.Speed	Condition	Airbag	Seatbelt	Frontal	Sex	Occupant.Age	Occupant.Role	Airbag.Status
10-24	alive	Airbag	Belted	1	M	18	Driver	1
25-39	alive	No Airbag	Not Belted	1	M	20	Driver	0
10-24	alive	Airbag	Belted	0	M	38	Driver	0

Fig 4: After processed data

```
> str(DA)
'data.frame':   26216 obs. of  9 variables:
 $ Vehicle.Speed: Factor w/ 5 levels "1-9","10-24",..: 2 3 2 2 2 2 2 4 2 4 ...
 $ Condition    : Factor w/ 2 levels "alive","dead": 1 1 1 1 1 1 1 2 1 1 ...
 $ Airbag       : Factor w/ 2 levels "Airbag","No Airbag": 1 2 1 1 1 2 1 2 2 2 ...
 $ Seatbelt     : Factor w/ 2 levels "Belted","Not Belted": 1 2 1 1 1 1 1 1 1 2 ...
 $ Frontal      : Factor w/ 2 levels "0","1": 2 2 1 2 2 2 2 2 1 2 ...
 $ Sex          : Factor w/ 2 levels "M","F": 1 1 1 1 1 2 2 1 1 2 ...
 $ Occupant.Age : int  18 20 38 35 40 26 77 49 51 46 ...
 $ Occupant.Role: Factor w/ 2 levels "Driver","Passanger": 1 1 1 1 1 1 2 1 1 1 ...
 $ Airbag.Status: Factor w/ 2 levels "0","1": 2 1 1 2 2 1 1 1 1 1 ...
>
```

Fig 5: Structure of dataset after preprocessing

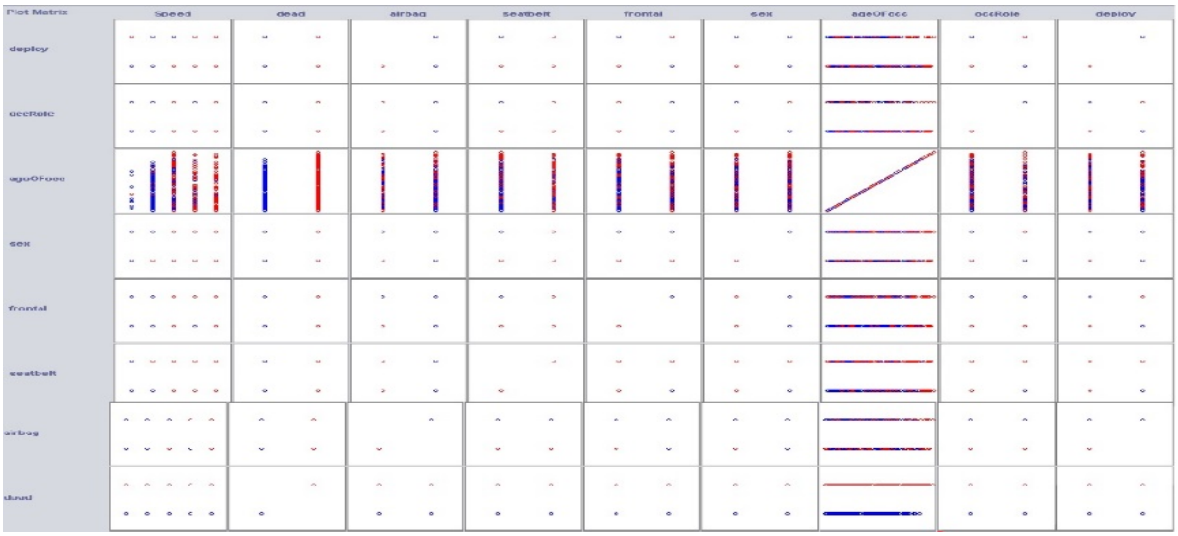


Fig 6: Co-relation between predictors

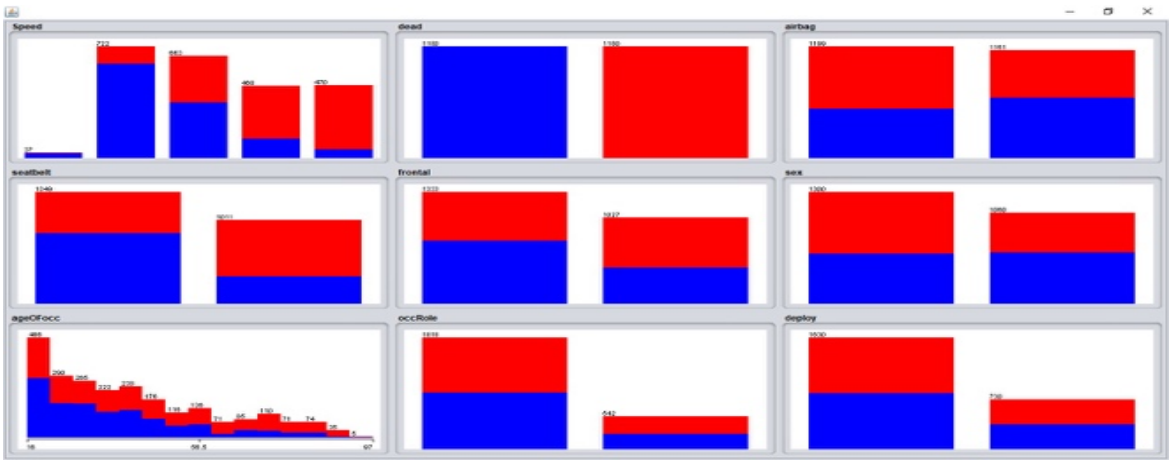


Fig 7: Data Distribution

V. Methodology

After the careful features selection and data preprocessing work, we focused on developing predictive models. Since our dataset already has output variable, we identified that it will be supervised learning. Our output variable is categorical, so we thought of performing classification algorithms on it. Overall, we performed 6 algorithms out of which we finalized 3 like, Decision Tree, Support Vector Machines and K Nearest Neighbors. All the above methodologies were performed using WEKA.

Decision Tree:

This is a supervised learning algorithm mainly used for regression as well as classification problems. Main idea behind using Decision Tree is to train a model which can predict output variable or class by identifying decision rules from the training data. (Fig 7)

J.48:

Preprocessing:

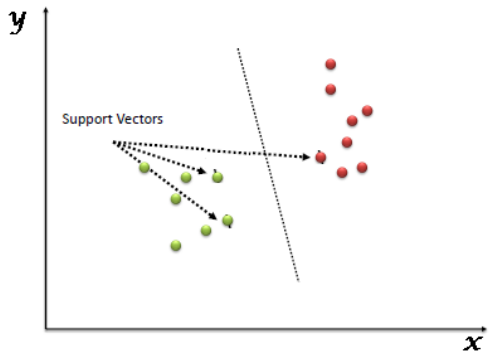
We applied InfoGainAttributeEval attribute evaluator to identify the root node for the tree. Search Method was selected as Ranker as per the Weka suggestions. Below is the ranking of the attributes. [5]

average merit	average rank	attribute
0.049 +- 0.001	1 +- 0	1 Speed
0.012 +- 0	2 +- 0	4 seatbelt
0.006 +- 0	3 +- 0	7 ageOFocc
0.003 +- 0	4 +- 0	5 frontal
0.002 +- 0	5 +- 0	3 airbag
0.001 +- 0	6 +- 0	6 sex
0 +- 0	7.1 +- 0.3	8 occRole
0 +- 0	7.9 +- 0.3	9 deploy

After ranking the attributes, we chose J.48 classifier and used 10-fold cross-validation for testing the predictions.

Support Vector Machines:

Support Vector Machines (SVM) is also a supervised machine learning algorithm mainly used in classification problems. In this, we plot each data item as a point in n-dimensional space. Then we perform classification by finding the hyperplane which clearly differentiate two classes. [7]



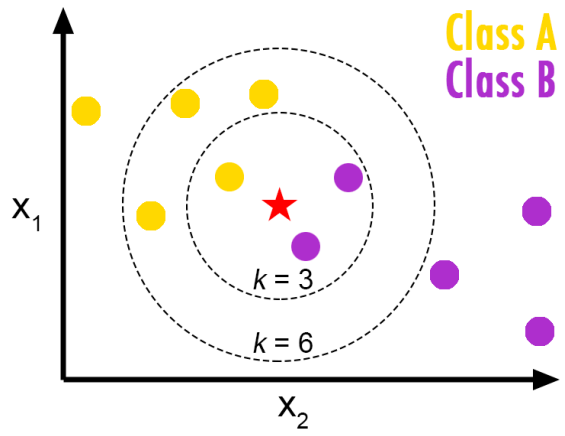
SMO:

As our output variable is binary categorical variable, SMO is the appropriate method to use for prediction. We used 10-fold cross validation for test. Accuracy rate is mentioned in the Analysis part. [6]

K-Nearest Neighbors:

K-Nearest Neighbors is a simple algorithm which we can use for classification problem. In this algorithm, it classifies new data from the previous stored data by calculating the distance from a similar data point. The value

of K determines how many points to consider. [4]



IBK:

We tested our dataset by putting different K values starting from 1, 5, 10, but we got maximum accuracy by keeping K value as 48 which is square root of the total number of instances 2360. Accuracy rate is mentioned in the Analysis part.

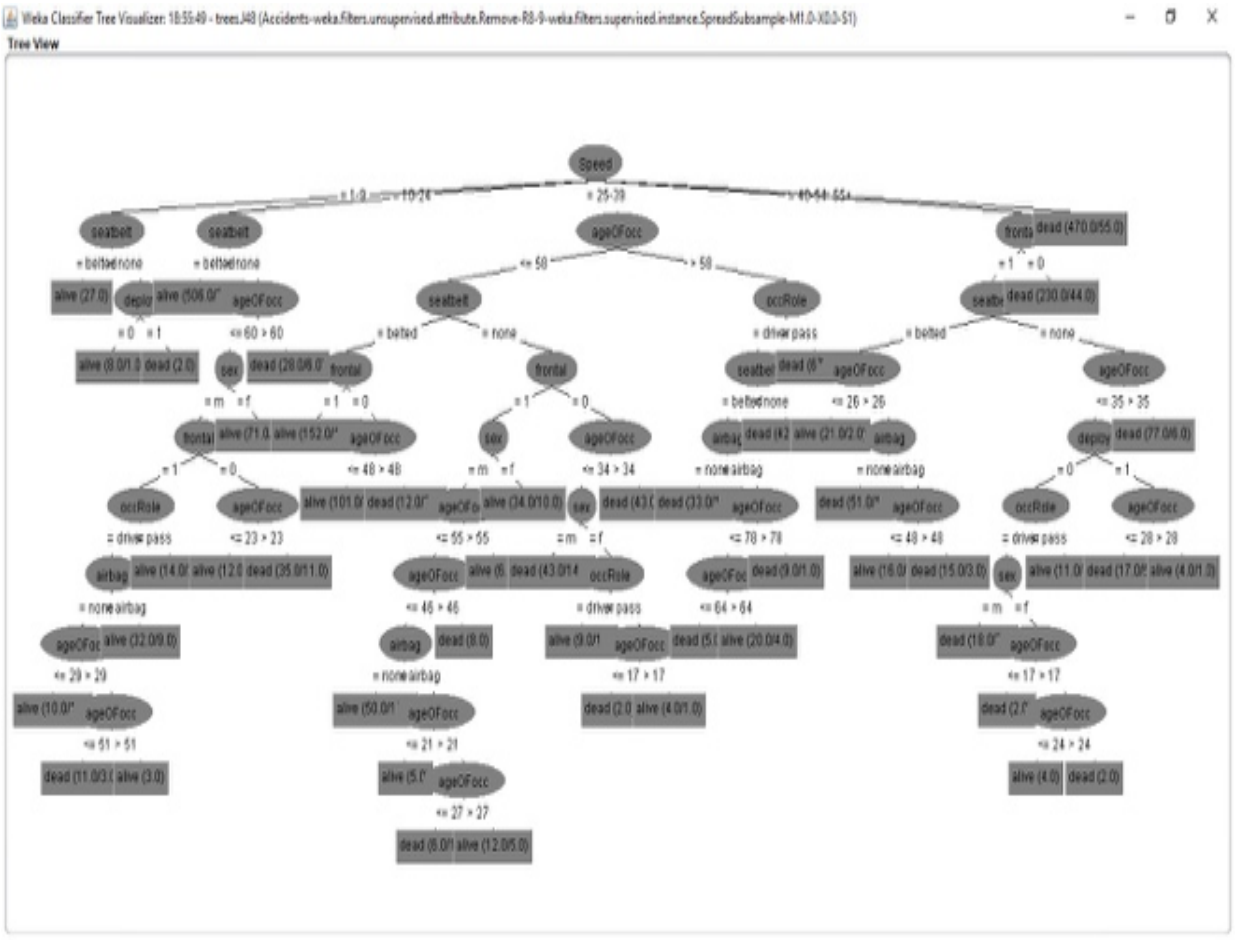


Fig 7: Decision- Tree

VI. Experimental Analysis

Models Performances:

J.48 Decision Tree:

```
Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1877               79.5339 %
Incorrectly Classified Instances       483               20.4661 %
Kappa statistic                          0.5907
Mean absolute error                      0.2818
Root mean squared error                  0.3927
Relative absolute error                 56.3644 %
Root relative squared error             78.5473 %
Total Number of Instances             2360

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.773    0.182    0.809      0.773    0.791      0.591    0.847     0.836     alive
                 0.818    0.227    0.783      0.818    0.800      0.591    0.847     0.797     dead
Weighted Avg.    0.795    0.205    0.796      0.795    0.795      0.591    0.847     0.817

=== Confusion Matrix ===

   a   b   <-- classified as
 912 268 |   a = alive
 215 965 |   b = dead
```

SVM:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1910               80.9322 %
Incorrectly Classified Instances       450               19.0678 %
Kappa statistic                          0.6186
Mean absolute error                      0.1907
Root mean squared error                  0.4367
Relative absolute error                 38.1356 %
Root relative squared error             87.3334 %
Total Number of Instances             2360

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.786    0.168    0.824      0.786    0.805      0.619    0.809     0.755     alive
                 0.832    0.214    0.796      0.832    0.814      0.619    0.809     0.746     dead
Weighted Avg.    0.809    0.191    0.810      0.809    0.809      0.619    0.809     0.751

=== Confusion Matrix ===

   a   b   <-- classified as
 928 252 |   a = alive
 198 982 |   b = dead
```

KNN with K=1:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1744               73.8983 %
Incorrectly Classified Instances       616               26.1017 %
Kappa statistic                          0.478
Mean absolute error                      0.2626
Root mean squared error                  0.4917
Relative absolute error                 52.5245 %
Root relative squared error             98.3414 %
Total Number of Instances             2360

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.778    0.300    0.722      0.778    0.749      0.479    0.776     0.728     alive
                 0.700    0.222    0.759      0.700    0.728      0.479    0.776     0.722     dead
Weighted Avg.    0.739    0.261    0.740      0.739    0.739      0.479    0.776     0.725

=== Confusion Matrix ===

   a   b   <-- classified as
 918 262 |   a = alive
 354 826 |   b = dead
```

KNN with K = 48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1837               77.839  %
Incorrectly Classified Instances       523               22.161  %
Kappa statistic                          0.5568
Mean absolute error                      0.3326
Root mean squared error                  0.3917
Relative absolute error                 66.5258 %
Root relative squared error             78.332  %
Total Number of Instances             2360

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Cl
                 0.780    0.223    0.778      0.780    0.779      0.557    0.863     0.870     al
                 0.777    0.220    0.779      0.777    0.778      0.557    0.863     0.841     de
Weighted Avg.    0.778    0.222    0.778      0.778    0.778      0.557    0.863     0.855

=== Confusion Matrix ===

   a   b   <-- classified as
 920 260 |   a = alive
 263 917 |   b = dead
```

So the overall accuracy is:

Model	Accuracy %
J.48 Decision Tree	79.53
SVM	80.93
KNN	77.83

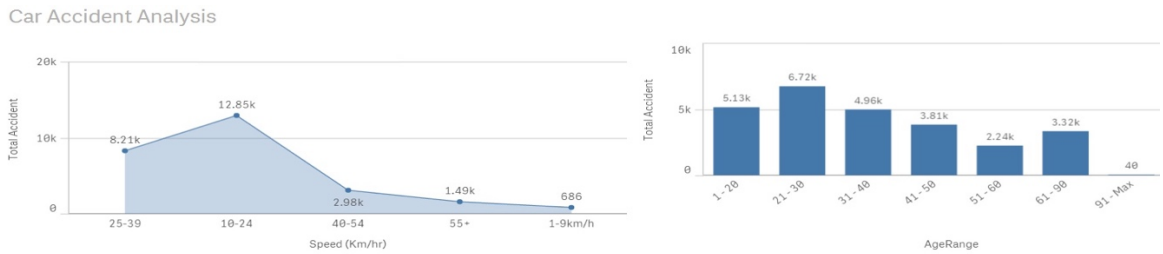


Fig: 8 Car accident Analysis graph

VII. Analysis

1. For further analysis, we found that from fig 8 that out of overall accidents, 4% resulted in fatality due to unidentified causes.
2. Maximum accidents happened at speed of 10-24 km/hr but only 0.8% resulted in fatalities.
3. Minimum accidents happened at speed of 55+ km/hr in contrast, the death rate was 35%.
4. The main cause of death was “no seatbelt” which caused overall 60% of deaths. Majority of the dead was from the age range of 16-30, which shows young adults drive cars without seatbelt. (Fig. 9)
5. 40% of overall deaths occurred because of malfunctioning of airbags deployment and no seatbelt at a speed of 55+ km/hr. (Fig 10)
6. Even with successful deployment of airbags and seatbelts, majority of the death were at the age of 61-90 and over 67% of them were actually driving the car. This proves that Age plays a crucial role in fatality along with technical malfunctioning. [1]

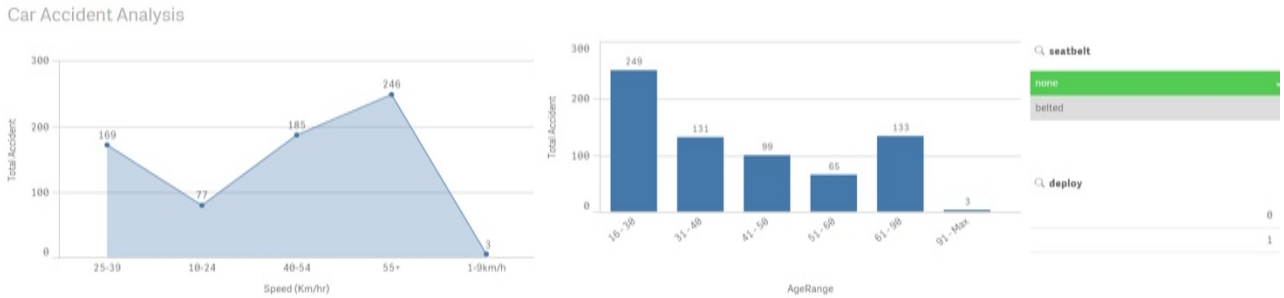


Fig: 9 Fatality rate Vs Age

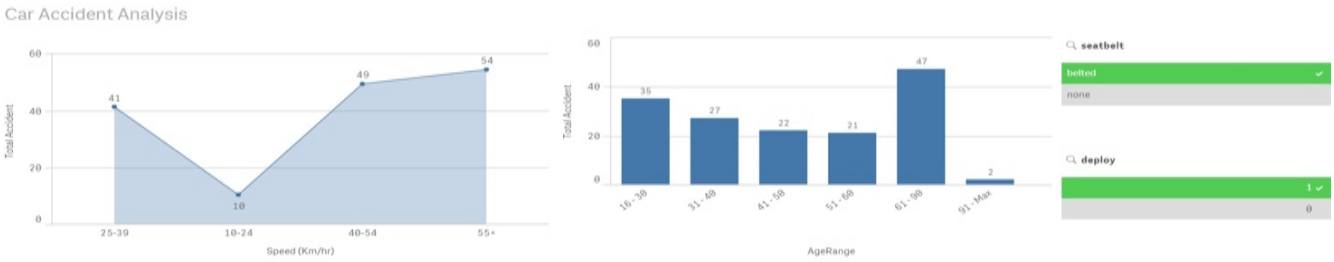


Fig: 10 Fatality rate Vs Speed

VIII. Conclusion

We collected and cleaned National Automotive Sampling System data of car accidents which happened from year 1997 to 2002 and tested number of predictive models. From our implemented methodology, we found that support vector machine is given us good accuracy result to predict pattern of fatality than other methodology i.e. decision tree, KNN through WEKA. We focused on many factors/ attributes related to car and driver. Result or analysis of this project would really help to automobile industry or any stakeholder who are deal with road safety. Based on final analysis, they can come with new rules and safe guide to prevent accidents. For future scope of this project would be combination of road related factors with driver and car information to predict fatality and to find relation between attributes.

References

[1] Federico Vaca-Craig Anderson-Harold Herrera-Chirag Patel-Eric Silman-Rhian DeGuzman-Shadi Lahham-Vanessa Kohl - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2691511/

[2] Huelke, D. F. (2001). The Effects of Belt Use and Driver Characteristics on Injury Risk in Frontal Airbag Crashes.

[3] Traffic Accident Data Mining Using Machine Learning Paradigms. (n.d.). Retrieved December 07, 2017, from http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.59.9074

[4] A Short Introduction to K-Nearest Neighbors Algorithm. (n.d.). Retrieved December 07, 2017, from https://helloacm.com/a-short-introduction-to-k-nearest-neighbors-algorithm/

[5] (2016, December 19). Retrieved December 07, 2017, from http://weka.sourceforge.net/doc.dev/weka/classifiers/functions/SMO.html

[6] Joachims, T. (2002). Text Classification. Learning to Classify Text Using Support Vector Machines, 7-33. doi:10.1007/978-1-4615-0907-3_2

[7] Road Crash Statistics. (n.d.). Retrieved December 07, 2017, from http://asirt.org/initiatives/informing-road-users/road-safety-facts/road-crash-statistics