

TRAFFIC SEVERITY ANALYSIS



INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Group Number -4
Chaitanya Garg(2021248)
Rudra Jyotirmay(2021280)
Deepanshu Dabas(2021249)
Arpan Kumar(2021020)



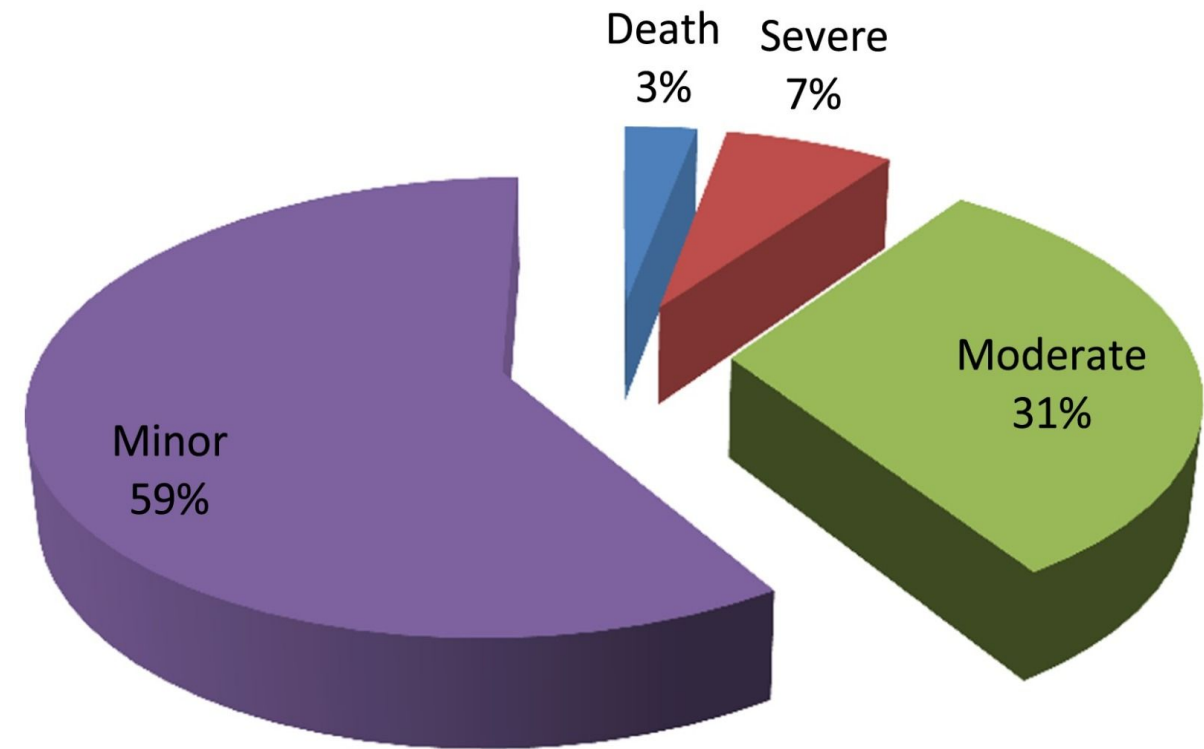
Motivation



Traffic Severity significantly impacts society through economic expenses, physical and mental health disorders, a loss of productivity and most importantly, the loss of many valuable lives.

Take the example of the USA, where the National Highway Traffic Safety Administration released its latest report estimating that there had been about 42,795 fatalities in 2022 due to motor vehicle crashes.

Traffic Accident Severity Proportions



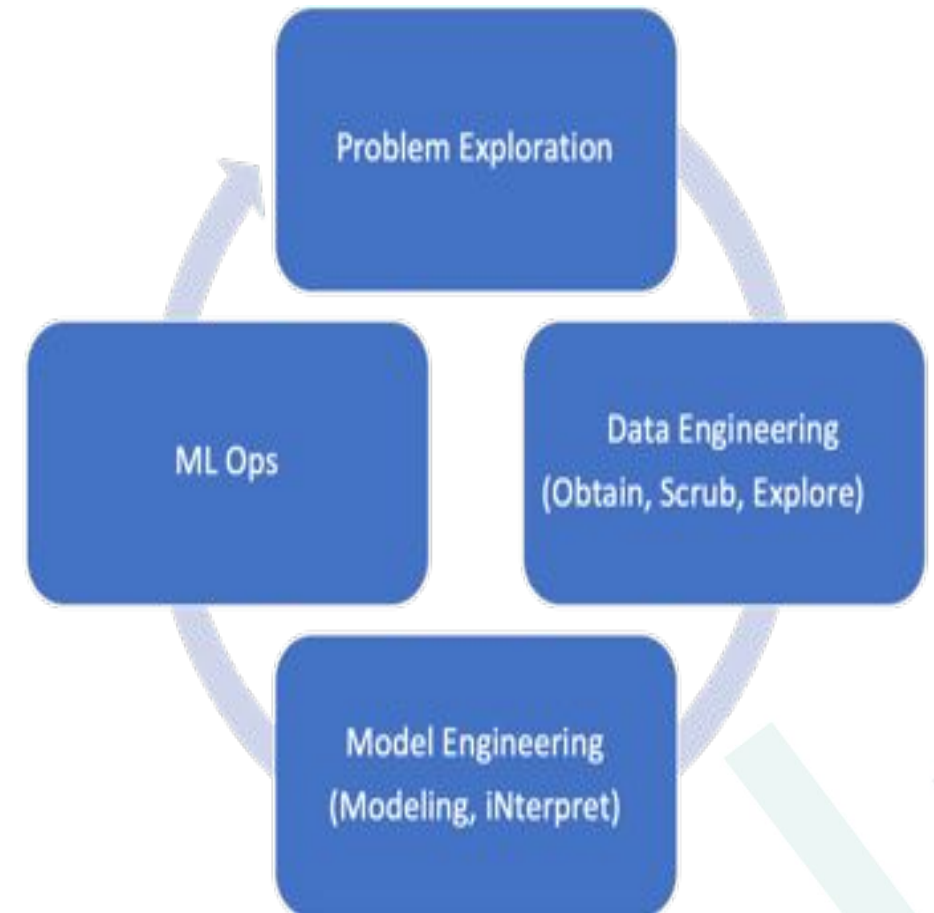
(Source : ResearchGate)

Motivation



In order to come up with effective approaches to prevent such accidents, it is critical that we perform Traffic Severity Analysis to identify frequent patterns and trends.

This involves carefully examining the dataset and identifying the key factors/variables that influence traffic severity, and then using these factors to create a model which can accurately predict the severity of any future traffic situations.



(Source : DataScience.org)

Literature Survey



1. Improved naive Bayes classification algorithm for traffic risk management by

Hong Chen , Songhua Hu , Rui Hua and Xiuju Zhao

- Paper highlights Naive Bayes' advantages of minimal parameter estimation from limited training data.
- However, simple Naive Bayes faces some obvious shortcomings. So, it introduces an "Improved Naive Bayes Classifier" with feature weighting for enhanced feature importance.
- Further addresses accuracy issues in small sample, large attribute scenarios using Laplace calibration.
- Offers promising potential for more accurate traffic risk prediction.

Feature-weighted naive Bayes classification algorithm

w_j represents the proportion of the number of samples in the total number of samples when attribute A_j is x_j .

$$P(C_i|X) = \alpha \prod_{j=1}^k w_j \frac{N(C = C_i, A_j = x_j)}{N(C_i)} \cdot \frac{N(C_i)}{N(D)}$$
$$= \alpha \prod_{j=1}^k \frac{N(A_j = x_j)}{N(D)} \cdot \frac{N(C = C_i, A_j = x_j)}{N(C_i)} \cdot \frac{N(C_i)}{N(D)}$$

Laplace calibration

$$P(C_i|X) = \alpha \frac{N(C_i)}{N(D)} \prod_{j=1}^k \frac{N(A_j = x_j) + 1}{N(D) + q_j} \cdot \frac{N(C = C_i, A_j = x_j) + 1}{N(C_i) + q_j} i$$

q_j represents the number of possible values of attribute A_j .

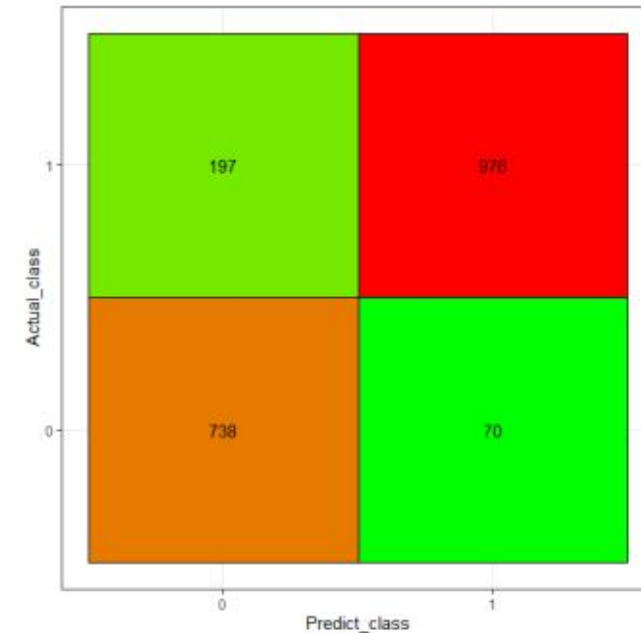
Literature Survey



2. Traffic Accidents Severity Prediction using Support Vector Machine Models *by*

Zeinab Farhat, Ali Karouni, Bassam Daya, Pierre Chauvet, Nizar Hmadeh

- Paper explores SVM models for predicting accident fatality rates, comparing radial basis function (RBF) and linear kernels.
- Dataset from Lebanon in 2016-2017 underwent preprocessing with normalization and outlier removal.
- SVM seeks to maximize hyperplane margin for optimal class separation. This model employs a binary SVM.
- Paper employs SVM kernels (linear, RBF) for data prediction model.
- Model achieved 91% accuracy on the testing set with RBF kernel and 84.6% with the linear kernel.



Testing confusion matrix

Literature Survey



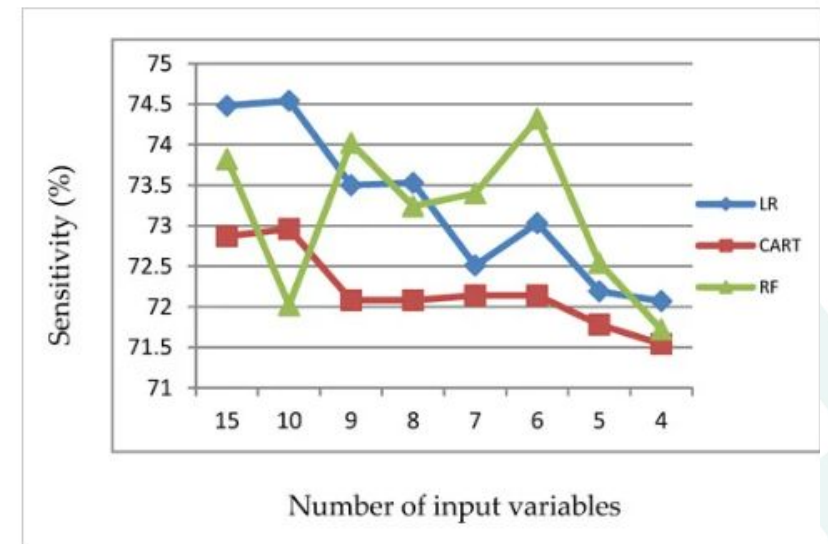
3. Modeling Road Accident Severity with Logistic Regression by Mu-Ming Chen and Mu-Chen Chen

- The paper analyzes accident severity with Logistic Regression (LR), making use of the logit function to get the probability of a serious accident.
- Divides data into training and validation sets, highlighting significant variables with probability values < 0.05 .
- Categorized accidents as "serious" or "minor" for balance.
- Uses the Spearman's rank correlation coefficient method to identify strongly correlated features.

The probability p of a serious accident is given by

$$Y = \text{logit} = \ln\left(\frac{p}{1-p}\right) = \beta X$$

where Y is the dependent variable (accident severity; $Y = 1$, if severity is serious; $Y = 0$, if severity is minor)



DATASET DESCRIPTION



The dataset includes traffic severity information covering 49 states of the US from the years 2016 - 2018. There are a total of 314285 rows and 18 features in the dataset to begin with. On the basis of these attributes, an accident is classified as having a severity between 1 - 4 (inclusive).

The features have been divided into 4 types of attributes depending upon the kind of information that they provide.

Traffic Attributes

Severity, Distance, Traffic
Signal, Time Difference

Location Attributes

Start Latitude, Start
Longitude, Street, City, County,
State, Airport Code

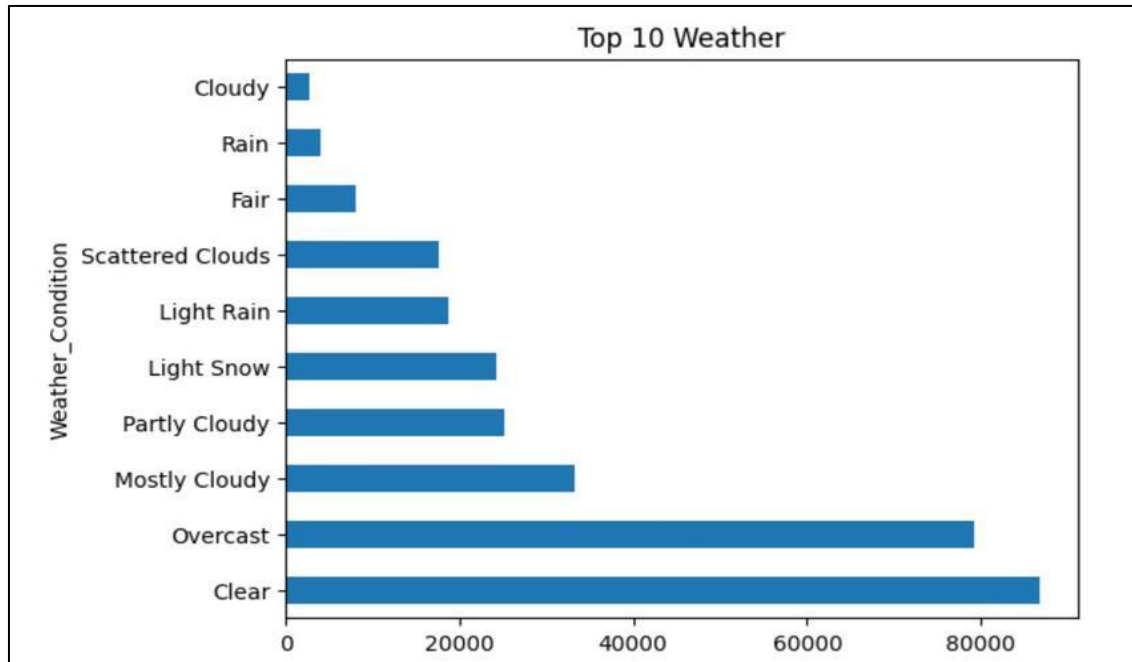
Weather Attributes

Temperature, Wind Chill,
Visibility, Wind Direction,
Weather Condition

Time Attributes

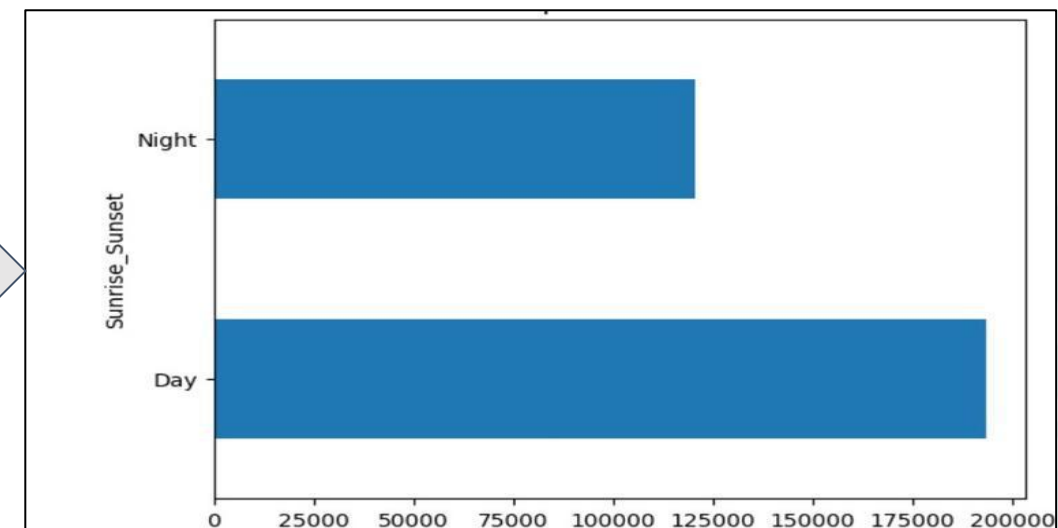
Year, Sunrise_Sunset

DATA INFERENCES

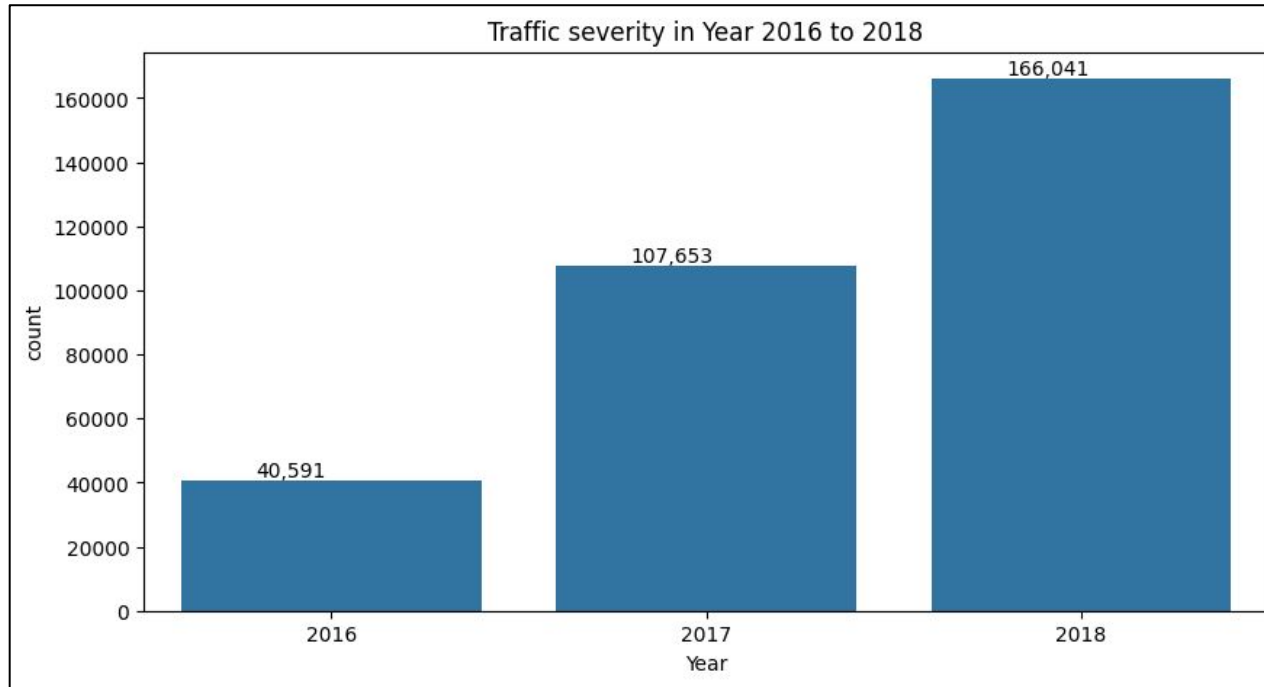


The bar graph representing weather conditions reveals that the majority of accidents occur during clear or overcast weather conditions.

The bar graph depicting sunrise and sunset times indicates that the majority of accidents occur during daylight hours.

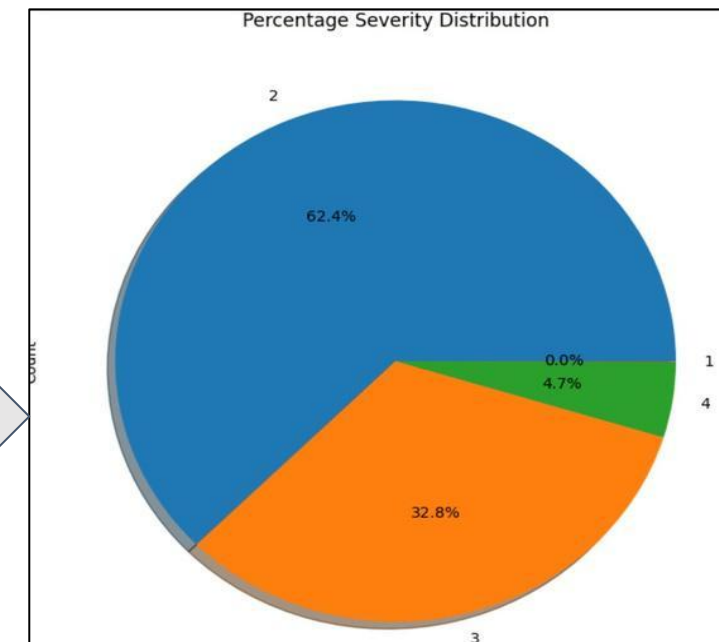


DATA INFERENCES

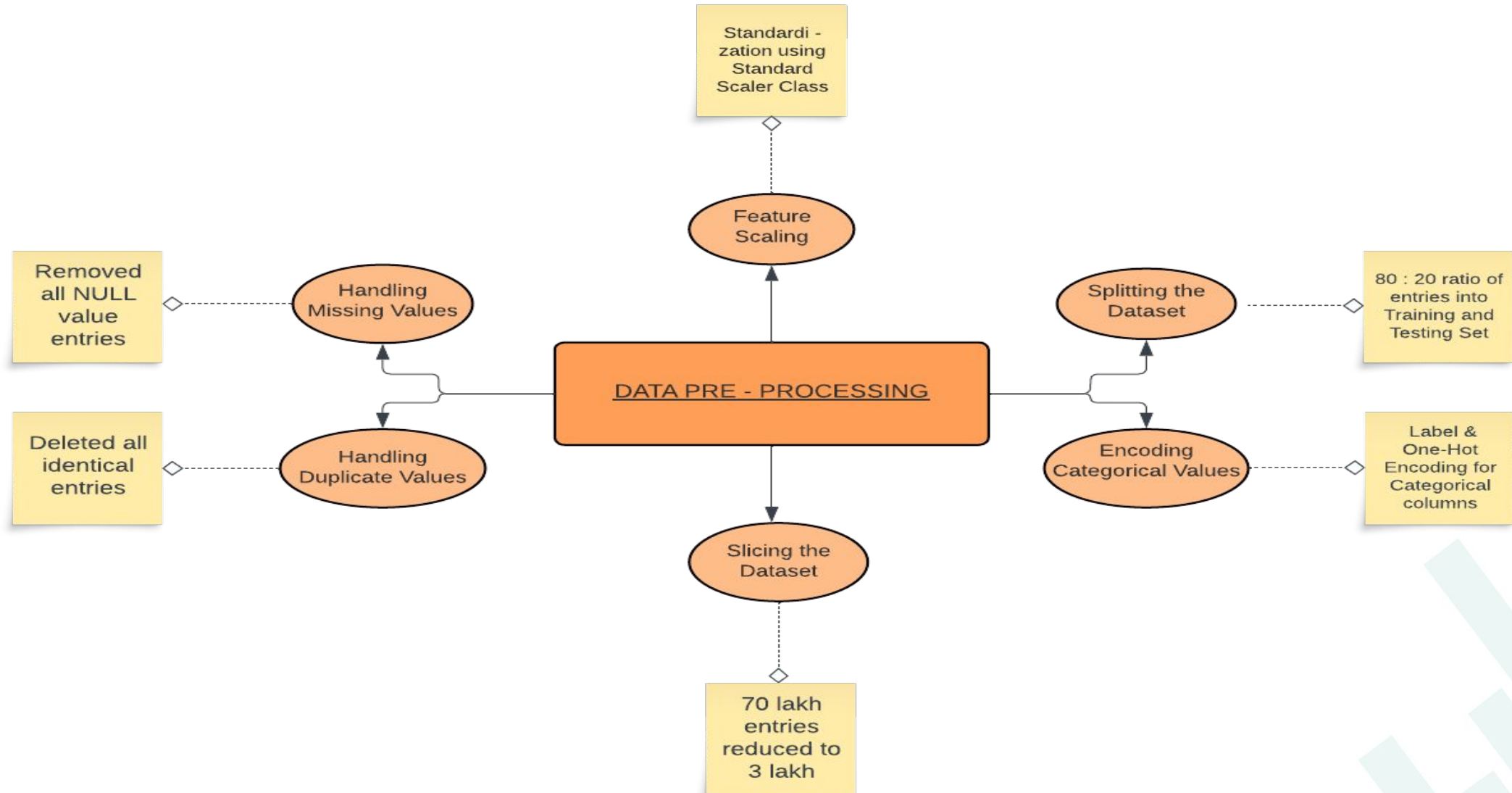


The bar graph of the year wise traffic distribution shows that the traffic severity has kept on increasing over the years, with the 2018 count being about 4 times the 2016 count.

The pie-chart of the percentage severity distribution tells us that most of the traffic observed on the roads is of severity level 2 (62.4%) and severity level 3 (32.8%). Traffic severity levels of 1 and 4 are rarely observed.



DATA PRE-PROCESSING



Methodology-Support Vector Machine

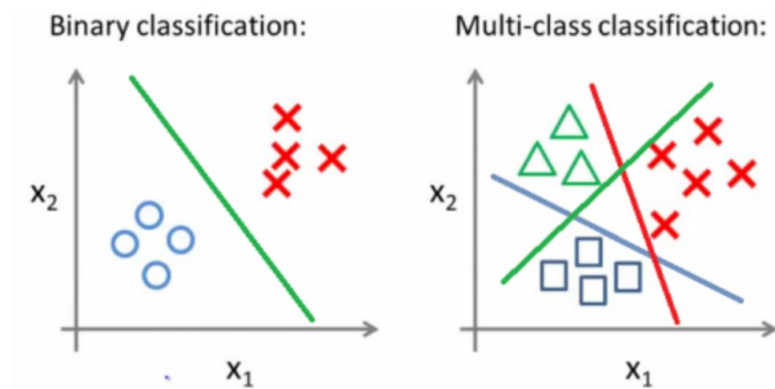
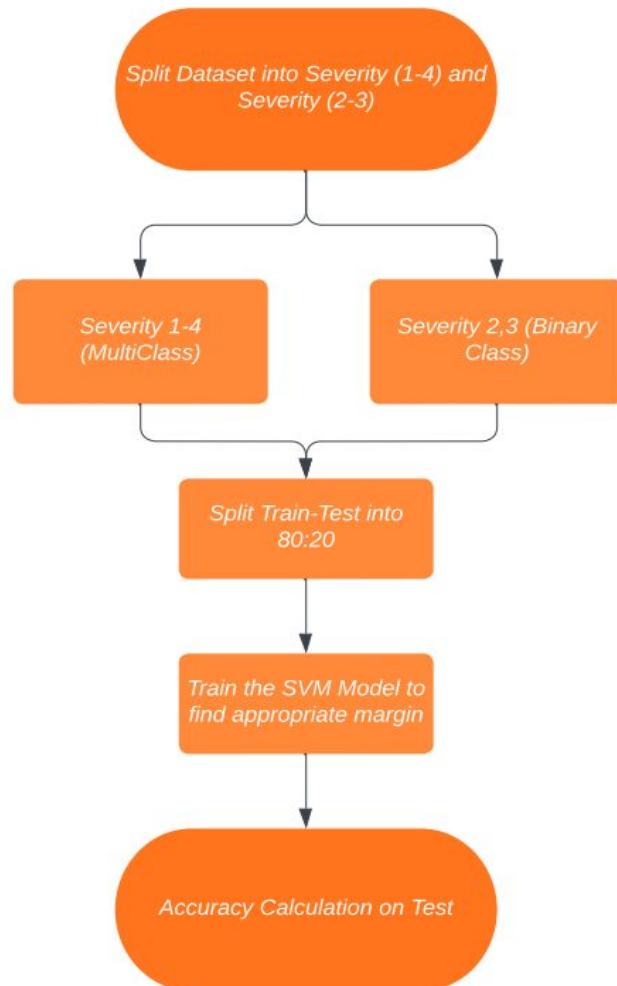
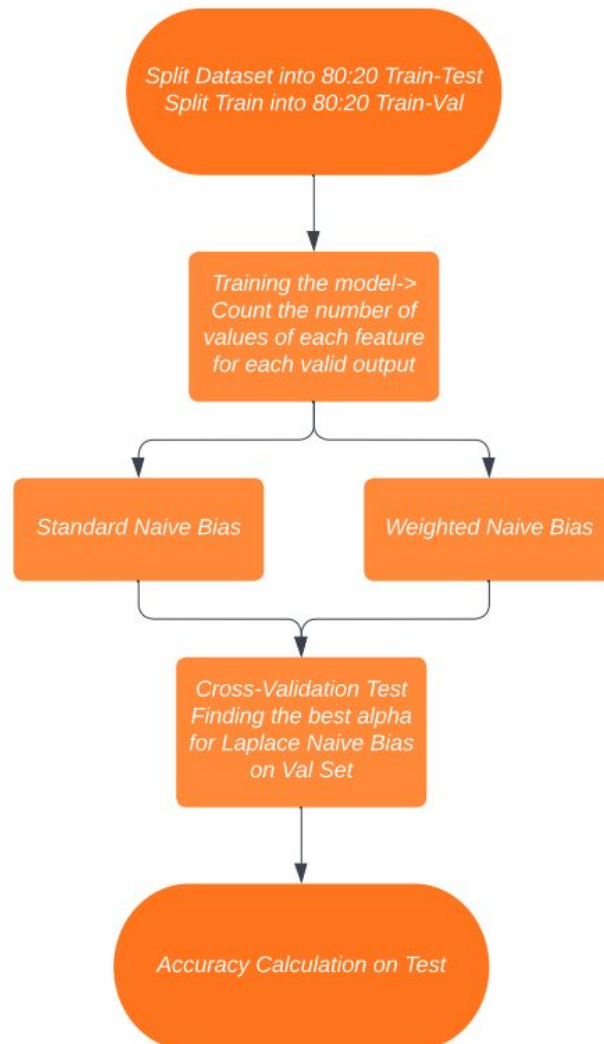


Fig : Multiclass breaks into many binary classes problems (Source: Medium)

	TrainSet	TestSet	SVM type
Severity(1-4)Outliers	0.6775	0.681	linear
Severity(1-4) w/o Outliers	0.689	0.657	linear
Severity (2,3)	71.43	0.669	linear

Methodology-Mixed Naive Bayes



Naive Bayesian classification

$$P(C_i|X) = \alpha \prod_{j=1}^k \frac{N(C = C_i, A_j = x_j)}{N(C_i)} \cdot \frac{N(C_i)}{N(D)}$$

**Feature-weighted
Naive Bayesian classification**

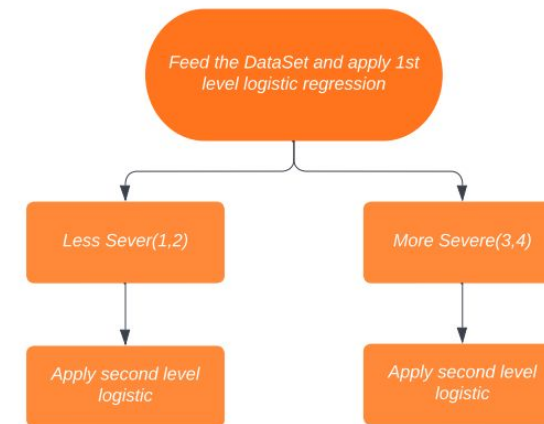
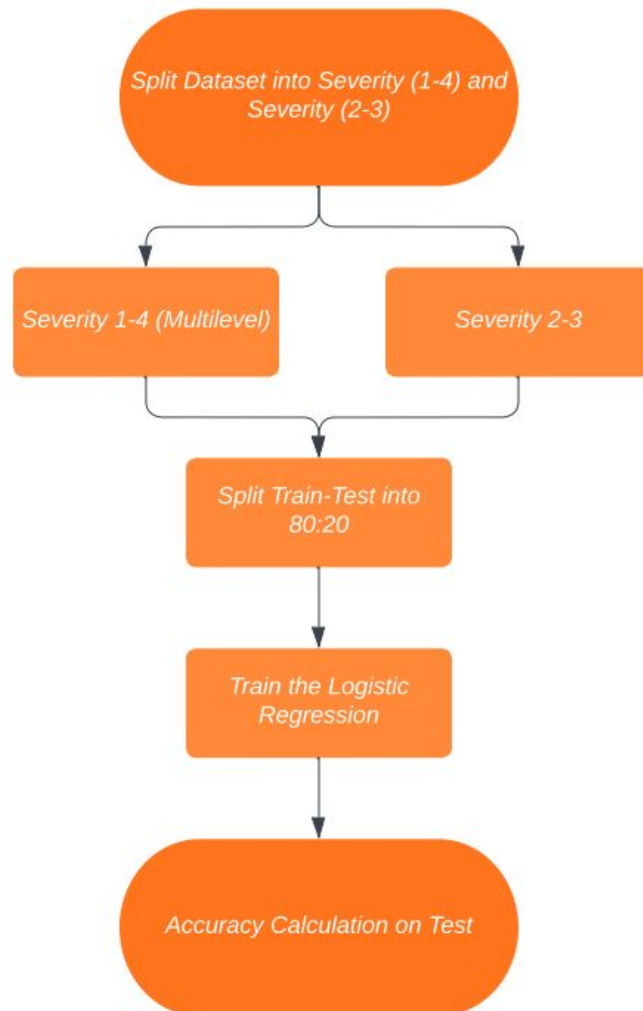
$$P(C_i|X) = \alpha \prod_{j=1}^k w_j \frac{N(C = C_i, A_j = x_j)}{N(C_i)} \cdot \frac{N(C_i)}{N(D)}$$

**Laplace calibration
Naive Bayesian classification**

$$P(C_i|X) = \alpha \frac{N(C_i)}{N(D)} \prod_{j=1}^k \frac{N(A_j = x_j) + 1}{N(D) + q_j} \cdot \frac{N(C = C_i, A_j = x_j) + 1}{N(C_i) + q_j} i$$

	TestTrain	ValTrain	TestSet	Alpha
With Outliers	0.67	0.645	0.623	1.5
Without Outlier	0.687	0.676	0.679	1
Improved NBC	0.728	0.683	0.672	1

Methodology-Logistic Regression



Multilevel Logistic Regression procedure

	TrainSet	TestSet
Basic With Outliers	0.65	0.603
Basic Without Outliers	0.67	0.62
Multilevel	0.57	0.53

Results & Analysis



- Data Preprocessing is crucial since a significant improvement in accuracy is observed before and after removal of outliers.
- We here deal with a classification problem with discrete outputs so linear regression can't be used and hence is avoided.
- We have employed the models of SVM, Naive Bayes and Logistic Regression till now.
- It is observed that upon the use of weighted Naive Bayes and Laplace calibration, bias decreases by about 5%.
- For all SVM models, linear kernel is the best fit out of linear, polynomial and rbf kernels.
- Multilevel logistic regression is not an idealistic method for multiclass classification

Results & Analysis



- Our best model is Naive Bayes Classifier trained upon Dataset without outliers, its is preferred upon SVM on Severity(1-4) as the latter has a poor f1-score of 0.28. (f1-score is the H.M. of precision and recall)

	TrainSet(in %)	TestSet(in %)
Naive Bayes Without Outliers	68.7	67.9
SVM on Severity (1-4)	67.75	68.1
Improved Naive Bayes	72.8	67.2

$$\begin{bmatrix} 0 & 3 & 0 & 0 \\ 0 & 2456 & 15 & 0 \\ 2 & 1250 & 109 & 0 \\ 0 & 2 & 0 & 0 \end{bmatrix}$$

Naive Bayes Fit Confusion Matrix

$$\begin{bmatrix} 0 & 10 & 0 & 0 \\ 0 & 7830 & 72 & 0 \\ 0 & 3735 & 350 & 0 \\ 0 & 2 & 1 & 0 \end{bmatrix}$$

SVM Best Fit Confusion Matrix

Timeline



TILL NOW		
Week Number	Topic Covered	Status
1-2	Data Collection & Cleaning	Completed
3-4	Pre-processing & Visualization	Completed
5	Feature Extraction, Analysis & Correlation	Completed
6-7	Logistic Regression, Naive Bayes & Support Vector Machines	Completed
FUTURE TIMELINE		
8	Random Forest & Decision Trees	In Progress
9-10	Neural Network & kNN	To Be Completed
11	Overfitting, Underfitting & Analysis	To Be Completed
12	Final Report	To Be Completed

Individual Contributions



- Data Collection - Chaitanya, Deepanshu, Rudra, Arpan
- Data Preprocessing & Cleaning - Arpan, Rudra
- Data Visualization - Chaitanya, Deepanshu
- Naive Bayes - Chaitanya
- Support Vector Machine - Deepanshu
- Logistic Regression - Arpan, Rudra
- Multilevel Logistic Regression - Chaitanya, Rudra
- Report - Chaitanya, Deepanshu, Rudra, Arpan
- Presentation - Chaitanya, Deepanshu, Rudra, Arpan

