PREDICTING THE SEVERITY OF ACCIDENTS

Abstract

Every year, road accidents result in deaths and related economic losses globally. Therefore, it is the foremost area of societal concern when considering loss prevention. Modeling accident severity prediction and upgrading the model are crucial to the successful operation of road traffic systems for increased safety. In accident severity modeling, the input vectors consist of many data related to the accident, such as driver traits, roadway conditions, and environmental factors. The output vector represents the specific class of accident severity. In this chapter, we have created two classifiers, a decision tree classifier and a KNN classifier, for the purpose of predicting the severity of accidents. Both classifiers exhibit high accuracy rates, with KNN achieving a superior accuracy of 89.3% compared to the 85.5% accuracy of the decision tree classifier. The identification of the primary elements that impact the severity of accidents may provide valuable insights for Government Departments/Authorities such as the Police, Roads and Buildings, and Transport, from a public policy perspective. The Departments may use the findings of research and modeling to implement effective actions aimed at mitigating the effects of accidents and thus enhancing traffic safety.

Keywords: Road Accidents, Prediction Technques, KNN , Decision Tree, artificial intelligence algorithms .

Authors

Ponnaboyina Ranganath
Research Scholar
Acharya Nagarjuna University
Guntur, India.
ranganathponnaboyina@gmail.com

G N R Prasad
Associate Professor
Department of MCA
Chaitanya Bharathi Institute of Technology
Hyderabad, India.
gnrp@cbit.ac.in

Venkata Pavan Kumar Savala
Associate Professor & Head
Department of CSE-DS
Siddhartha Institute of Engineering & Technology, Ibrahimpatnam
Hyderabad, India.
venkatapavankumarsavala@gmail.com

P. V. Ravi Kumar
Associate Professor
Department of CSE
Krishna Chaitanya Institute of Technology &Sciences
Markapur, India.
putta.msc@gmail.com

P M Yohan
Professor & Principal
CSI Wesley Institute of Technology and Sciences
Secunderabad, India.
pmyohan@rediff.com

SK Althaf Hussain Basha
Professor and R&D Coordinator
Krishna Chaitanya Institute of Technology & Sciences
Markapur, India.
althafbashacse@gmail.com

I. INTRODUCTION

The primary goal of machine learning is to understand the underlying patterns in data and translate that data into a comprehensible model. This technology is a distinct field within computer science that differs from conventional processing techniques. This kind of processing involves a well-defined set of instructions that are used to solve issues. The machine learning (ML) algorithms enable training on data inputs and use statistical analysis to derive principles that are inherent within a certain dataset. ML aids in the creation of models using data to enable automated decision-making based on factual knowledge.

Machine learning has provided users with valuable information in several parts of facial recognition systems, enabling them to enhance their ticketing experience and add to their friends' picture collections. The textual content of photos is transformed into mobile technology with excellent characterisation. In addition to these elements, other platforms use machine learning applications to address diverse problems. Machine learning applications in the present day are very valuable for data analysis, comprehension of intricate data, and the generation of computer-interpretable instructions.

The primary objective of this chapter is to forecast cardiovascular disease in patients utilizing artificial intelligence algorithms. This notion is beneficial for mobile applications that aim to detect cardiac problems by analyzing a person's heartbeat. It operates dynamically by using categorization methods. It is more cost-effective to create techniques for detecting cardiac problems of this kind.

II. IMPORTANCE OF MACHINE LEARNING ALGORITHMS

Every day, many cars traverse various sorts of roadways. Incidents may occur in any location and at any moment. Over the last decade, there has been a significant exponential rise in the occurrence of vehicle accidents. Over the previous decade in India, the annual number of traffic accidents has been 1.7 lakh. The frequency of road accidents in several Indian towns has been steadily rising due to inadequate safety measures. Road safety is a multifaceted problem in 2019. In 2018, the number of road accidents in India reduced by 3.27%, with 534,810 collisions compared to 520,552 in 2017. This represents a reduction in the accident rate of 2.1%. In 2019, the number of fatalities due to traffic accidents in India rose to about 189,000. These data illustrate the issue that the nation is grappling with in terms of traffic accidents. In 2018, vehicle accidents resulted in a mortality rate of 7 fatalities per 10-minute interval. Every individual desires to evade road accidents and ensure personal safety, as nobody wants to experience death or endure suffering as a result of such incidents. By using classification algorithms on accident data, we can accurately forecast the extent of injuries and contribute to the effective management of road accidents. Several factors contribute to accidents, including violations of traffic regulations. However, road conditions and traffic congestion are significant contributors to fatalities and injuries worldwide. These incidents arise as a result of the dynamic design and expansion of the vehicle industry. A traffic collision occurs as a result of several factors, such as the collision between two cars on a road, a pedestrian, an animal, or any other natural obstruction.

A Decision Tree is a kind of supervised machine learning method that recursively splits data based on certain conditions. Decision Trees consist of two components: decision nodes and leaves. Leaves represent the determinations or ultimate results.

III. ASSUMPTIONS, USING A DECISION TREE

- At first, regard the whole set as the primary training.
- Preferred concepts for understanding characteristics.
- Attribute values records are distributed forcefully.
- Arithmetical procedures determine the hierarchy of characteristics.

The primary challenge with this notion is to accurately identify the property for the root node at each level. This process is referred to as attribute selection. The processes for selecting attributes are Entropy Reduction and Gini Index is a statistical measure used to assess the level of inequality in a distribution.

1. **Entropy Reduction:** The entropy fluctuations within subsets dictate the division of training examples into smaller pieces using a node in a decision tree. Let S be a set of instances and A be a feature. Sv is the subset of S where A has a value of v.

   The gain of (S, A) is calculated by subtracting the entropy of S from a certain value. Entropy is a concept in thermodynamics that measures the level of disorder or randomness in a system. The user's text is empty.

   Entropy is a concept in thermodynamics that measures the degree of disorder or randomness in a system.

   Gain (S, A) = Entropy(S) $-\sum_{v\epsilon Values\,(A)} \frac{S_v}{S}$. Entropy ($S_v$)……. (1)

   The term refers to the quantification of a random variable, specifically focusing on the spread of an irrational collection. The higher the entropy, the more additional information there is.

   Information gain is computed while constructing a decision tree

   At the start of the procedure, each training instance is linked to the root node.

   - To assign a tag to each attribute node based on the data obtained for decision making.
   - The root should not obstruct the inclusion of distinct attributes on two separate occasions.
   - Every subset of the subtree for training examples is categorized coercively.

2. **The Gini Index:** The Gini Index is a metric used to quantify the extent to which a randomly selected element might be incorrectly identified.

   - The characteristic with a lower Gini index should be prioritized.
   - Skelton provides assistance for calculating the Gini Index.

   The Gini Index formula is shown below.

   Gini Index = 1- $\sum_j p_j^2$ ……………… (2)

The Gini Index is calculated as 1 minus the sum of squared proportions of each category in a distribution. The number is 2.

3. **K-Nearest Neighbor (KNN):** This classifier is exhibiting delayed performance while doing trained classification on two distinct beginning stages for training and testing in a unique manner. The KNN classifier assigns weights to each data point based on the principle that they are treated as neighbors. The technique incorporates the idea of boundaries (ranges) to calculate neighboring elements. In this classification measure, the distance of each element is calculated using both Euclidean distance and Manhattan distance. The following formula is used to determine the distance:

$EucledianDistance = D_{\mathrm{a,\ b}}$
$= (\mathrm{a}_i - \mathrm{b}_i)\ 2\mathrm{l}_i = 1$-------------------------- (3)

Where,
L= number of cluster
a, b = sample spaces of co-ordinate
$Manhattandistance = (x_{i-}) = 1$ -------------- (4)

A & B are, distances of Minkowski usually called Euclidian distance
$Min = (-\mathrm{b}_{ip})\ 1p$------------------------ (5)

The KNN algorithm utilizes a cluster model based on a super class to enhance training and provide accurate results. This model generates less noise during the classification process. Below are the stages of the KNN algorithm.

- The variable D represents the set of training data points, whereas k indicates the number of closest neighbors.
- Every instance of a subclass is derived from a superclass.
- Compute the Euclidean distance for each training data.
- The majority of neighbors belong to the same class and are classified based on their class traits.

According to this algorithm, the process of training and testing allows for the classification of data in many ways. The training phase involves computing the distance between data points and storing the data for further testing, ensuring its availability. In order to determine the unknown data point in a classification problem, one might compute its distance from the neighboring data points. Various metrics are used to compute the distance of a data point. The accident dataset has been collected from the Kaggle website for our study. The qualities are as follows:

- **Collision_Ref_No:** This number serves as a unique identifier for each accident.
- **Policing Area:** denotes the jurisdictional area of the police station where the accident took place.
- Collision Severity refers to the extent of harm resulting from an accident.
- Weekday_of_Collision refers to the specific day of the week on which the event happened.

- **Day_of_Collision:** denotes the specific day within a month when the event took place.
- Month_of_Collision refers to the specific month in which the accident took place.
- **Hour_of_Collision:** denotes the specific hour at which the incident took place.
- **Carriageway_Type:** relates to the kind of roadway.
- **Speed Limit:** The maximum speed allowed on the road where the accident took place.
- **Junction_Detail:** details on the junction
- **Junction_Control:** details on the management of traffic at the intersection
- **Ped_Crossing_HC:** data indicating if a pedestrian is currently crossing the road
- **Ped_Crossing_PC:** denotes the presence of a pedestrian crossing sign
- **Light_Conditions:** The illumination level present at the time when the event occurred.
- **Weather Conditions:** The state of the weather at the time of the accident.
- **Road_Surface_Conditions:** pertains to the state or quality of the road surface.
- **Special Conditions at Site:** Additional circumstances present at the time of the accident.

The objective is to gather a data collection including many characteristics related to accidents. The dataset was obtained via a Kaggle competition. After obtaining the dataset, proceed to do attribute selection on it. Attribute selection is the process of choosing the relevant qualities from a dataset that are valuable for predicting the result. Filter the dataset to include just the characteristics that are relevant to the severity attribute. Preprocess the data after doing feature selection.

IV. DATA FILTERING IS THE PROCESS OF SELECTIVELY EXTRACTING OR REMOVING CERTAIN DATA FROM A LARGER DATASET BASED ON CERTAIN CRITERIA OR CONDITIONS

Data filtering involves applying practical transformations to our data prior to inputting it into the algorithm. Oftentimes, the data we acquire may consist of raw data, meaning it has not been processed or analyzed. The dataset may include both noisy and incomplete data. The accuracy of our developed model will be compromised if the data includes noise or missing values. The model's precision will be modest. To mitigate this issue, it is imperative to address the null values that exist inside our collection. There are several approaches to address the issue of missing values, which include:

- Remove all instances with missing values.
- Replace the incomplete values with statistical measures.
- Utilize a machine learning technique to make predictions for the missing data.

The dataset has a very low proportion of missing values, therefore we may remove the entries that have these values. If the proportion of misplaced data is substantial, it is not feasible to discard all the missing values. In order to address the missing data, we may use statistical measures such as mean or mode, or alternatively, employ a machine learning method to forecast the missing values.

Following pre-processing, the dataset should be partitioned into two distinct components: one for model generation and the other for model testing. The dataset used for constructing the model is referred to as the training dataset, while the dataset employed for evaluating the model is known as the testing dataset. Once the dataset has been divided, construct the classification model using the training data and then evaluate the model's performance using the test dataset. To verify the model, we make predictions on the test set and compare these predictions with the actual outcomes included in the test dataset.

V. ANALYZING THE OUTCOMES OF TWO MODELS

This study used a Decision Tree and KNN model to accurately estimate the severity of accidents using the information that was acquired from the input. The Decision Tree model has an accuracy rate of 85.68%, but the KNN model has an accuracy rate of 90.33%. Both models exhibited high accuracy rates, however, the KNN classifier outperformed the Decision tree classifier.

Calculation of the Correlation Matrix

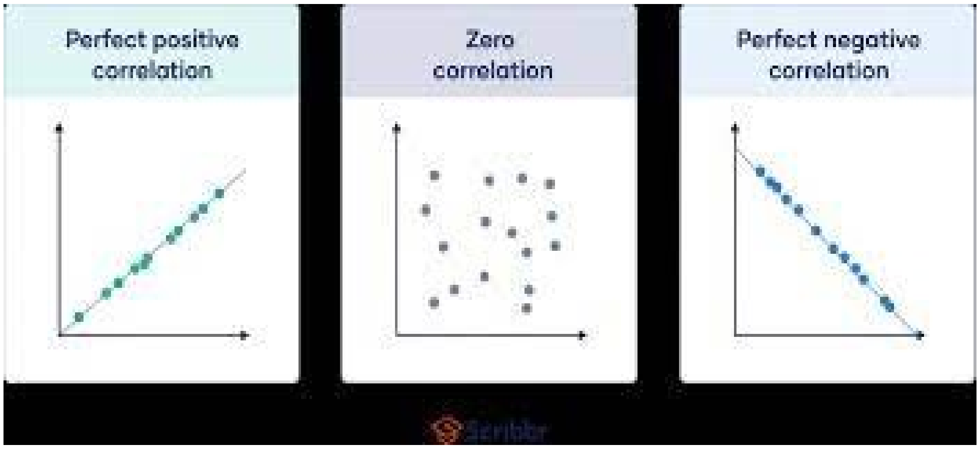

Figure 1: The correlation coefficient between a variable and itself is always equal to one

The autocorrelation of a variable is always one. Autocorrelation refers to the degree in which values of the same variable across different observations correlated with one another.

Autocorrelation frequently comes up when doing time series analysis. Very often time-based data such as temperature measured at different times of the year display contain autocorrelation. For example, the temperature at the 1st day of the year is most likely similar to the temperature on the 2nd day of the year.

In multiple linear regression analysis, autocorrelations are problematic. One of the key major assumptions behind linear regression is that each obervation is independent. Autocorrelation invalidates that assumption, causing any predictions and insights extracted from the model to be biased.

Autocorrelation can be addressed by adding lagged parameters/features of the independent and/or dependent variable to the linear regression model.

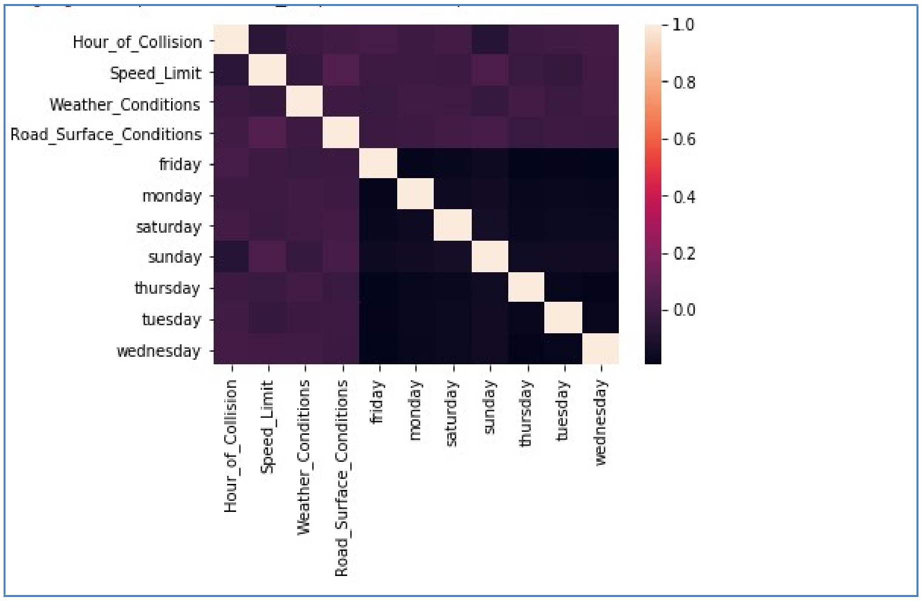

Figure 2: Schematic representation of the association between variables

Table 2: Matrix Displaying the Classification Results of a Decision Tree Classifier

Index	Prediction 1	Prediction 2	Prediction 3
Actual : 1	3	10	13
Actual : 2	7	63	128
Actual : 3	30	156	1785

The matrix used to assess the accuracy of a classification model, displaying the number of true positives, true negatives, false positives, and false negatives.

The confusion matrix is a tabular representation that assesses the performance of a classification model using test data sets. It helps identify the actual values and presents the performance visually in a table style. It also supports the identification of discrepancies across classes, which may sometimes result in mislabeling of data points owing to confusion. This matrix use categorization to forecast the result. Divide each class to get the number of accurate and inaccurate predictions in the categorization. The confusion matrix is a crucial component of a classification model since it enables accurate prediction evaluation. During this procedure, two distinct sorts of mistakes occur, which are particularly useful for identifying further faults in categorization.

The matrix above provides a clear depiction of the mistakes committed by the decision tree classifier. The classifier produced 9 errors by incorrectly identifying the severity as 2 instead of 1, and 12 errors by classifying it as 3 instead of 1.

When the true severity level is 2, the classifier correctly identified 63 instances as 2. However, it incorrectly forecasted 6 instances as 1 and 156 instances as 3.

With an actual severity level of 3, the classifier accurately predicted 1785 values, misclassifying 30 values as 1 and 156 values as 2.

Table 3: Matrix Displaying the Classification Results of a K-Nearest Neighbors (KNN) Classifier.

Index	Prediction 1	Prediction 2	Prediction 3
Actual : 1	1	9	21
Actual : 2	2	43	185
Actual : 3	2	56	1952

The matrix above displays the faults committed by the KNN classifier. Given an actual value of 1, the KNN classifier made predictions of 2 for 6 values and 3 for 17 values.

The KNN classifier correctly forecasted 23 instances as 2 and misclassified 172 instances as 3, when the true value was 2. Given an actual value of 3, the KNN classifier correctly predicted 1 value as 1, 47 values as 2, and 1818 values as 3.

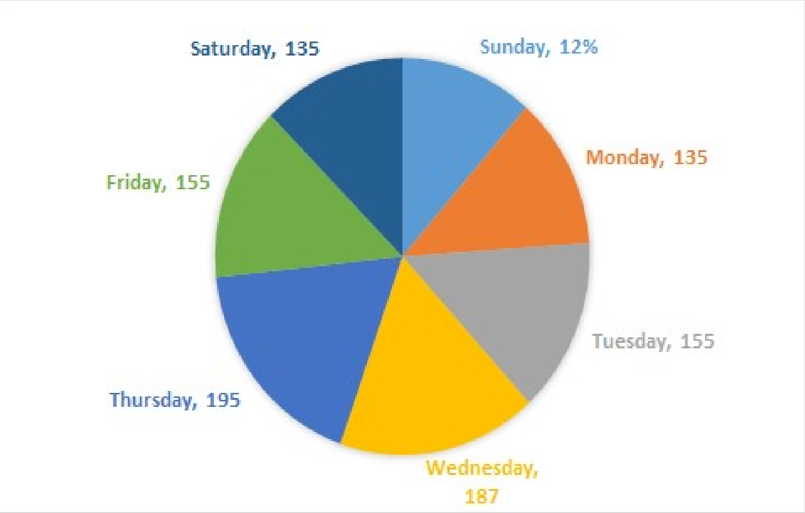

Figure 3: Each Day of the Week Corresponds to the Accidents that Happened

Displaying a pie chart illustrating the frequency of accidents that have happened based on the day of the week. Friday has the greatest frequency of accidents, while Sunday has the lowest frequency.

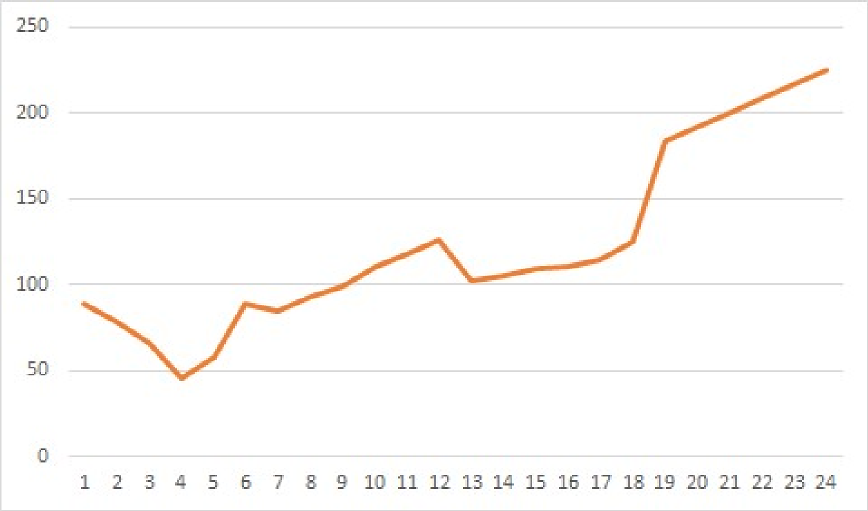

Figure 4: Graph Depicting the Frequency of Accidents

Graph depicting the frequency of accidents that happened throughout various time intervals during the day. The graph illustrates a surge in accidents between hours 6 and 8, followed by a decline from hour 13 to 18, and thereafter another spike. The highest frequency of accidents occurs during hour 00. Accidents are less frequent between hours 4-7 and 13-19.

Comparison of the overall accident count with the designated speed restriction

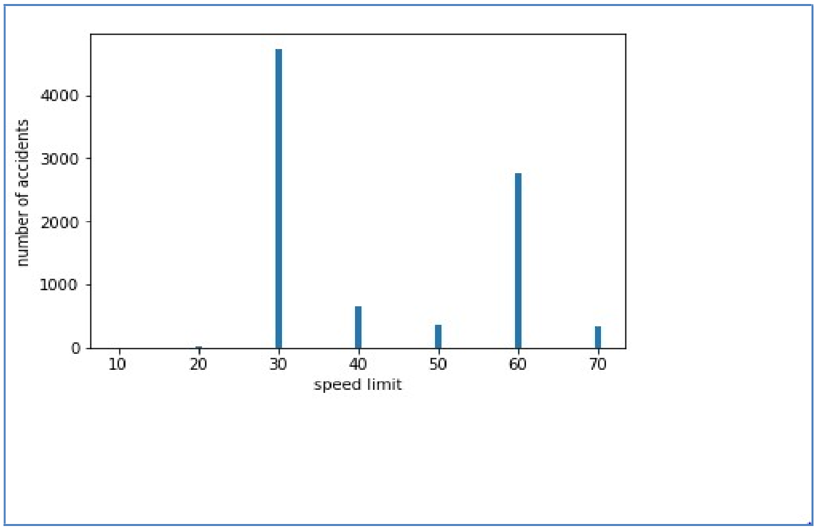

Figure 5: Comparison of the overall accident count with respect to the designated speed limit.

The bar graph below illustrates the frequency of accidents in relation to the speed limit. The majority of accidents occur when the speed limit is raised. The incidence of accidents decreases when the speed limit is set between 30 and 50 km/hr.

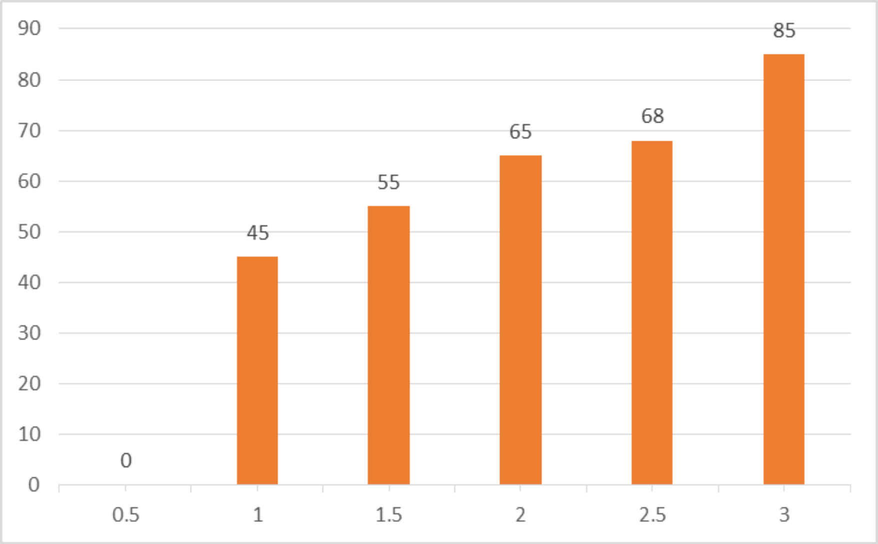

Figure 6: Frequency of incidents with varying degrees of severity occurring specifically on Fridays

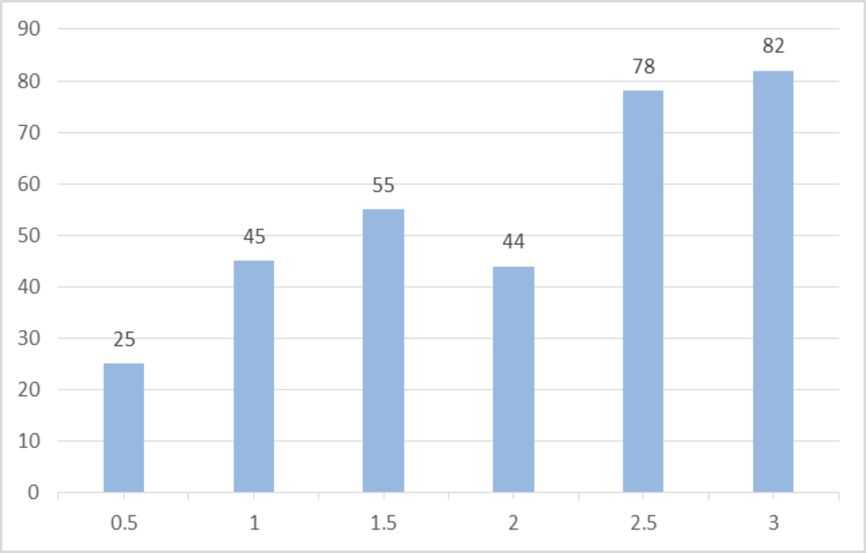

Figure 7: Frequency of incidents with varying degrees of severity occurring on Saturdays

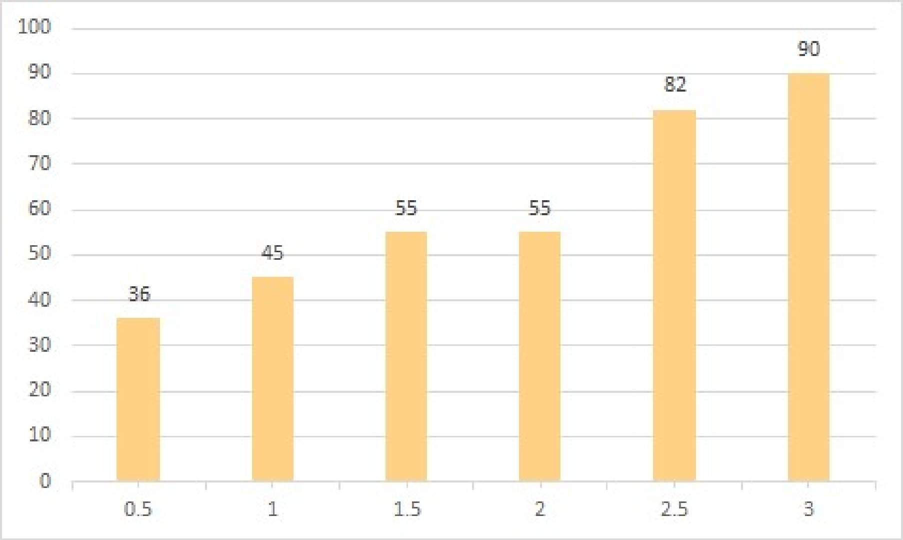

Figure 8: Frequency of incidents with varying degrees of severity occurring on Sundays

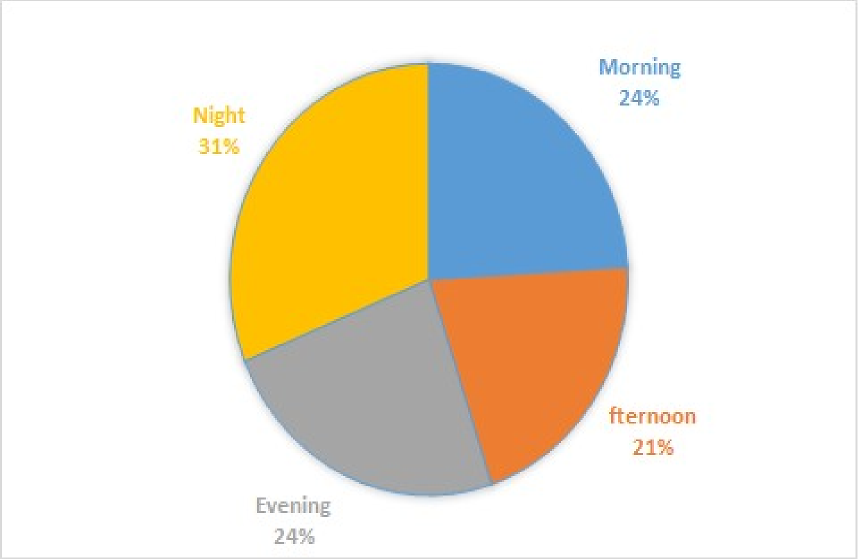

Figure 9: Frequency of accidents throughout different time periods: morning, afternoon, evening, and night.

The number of accidents is classified according to the time of day: morning, afternoon, evening, and night. The pie-chart clearly illustrates a significant decrease in the frequency of accidents at nighttime in comparison to other periods of the day. The afternoon period has the greatest incidence of accidents in India.

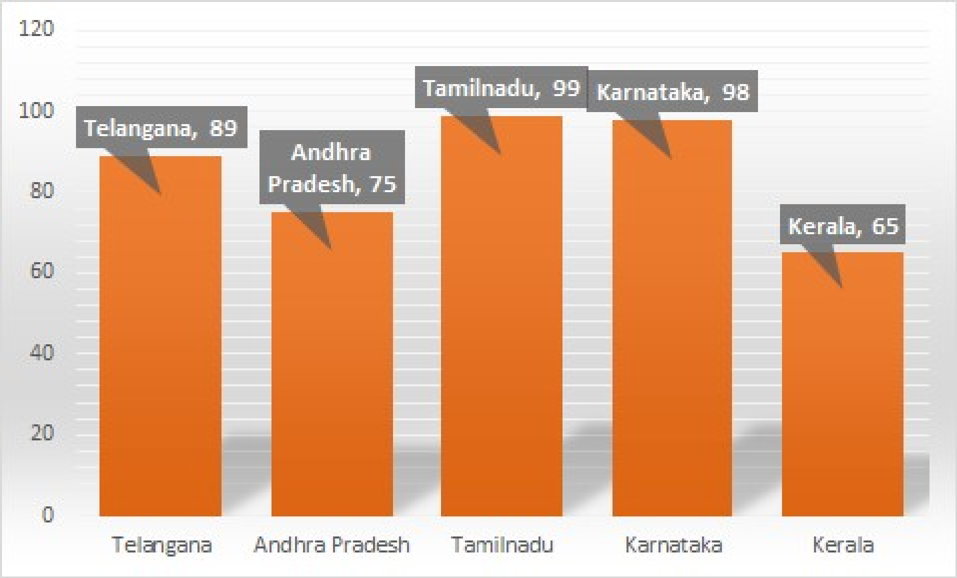

Figure 10: Peak summer accident rates

The following states/union territories have the greatest number of accidents during the summer season: Bihar has the greatest incidence of accidents during the summer, followed by Mizoram, Jharkhand, Madhya Pradesh, and Sikkim.

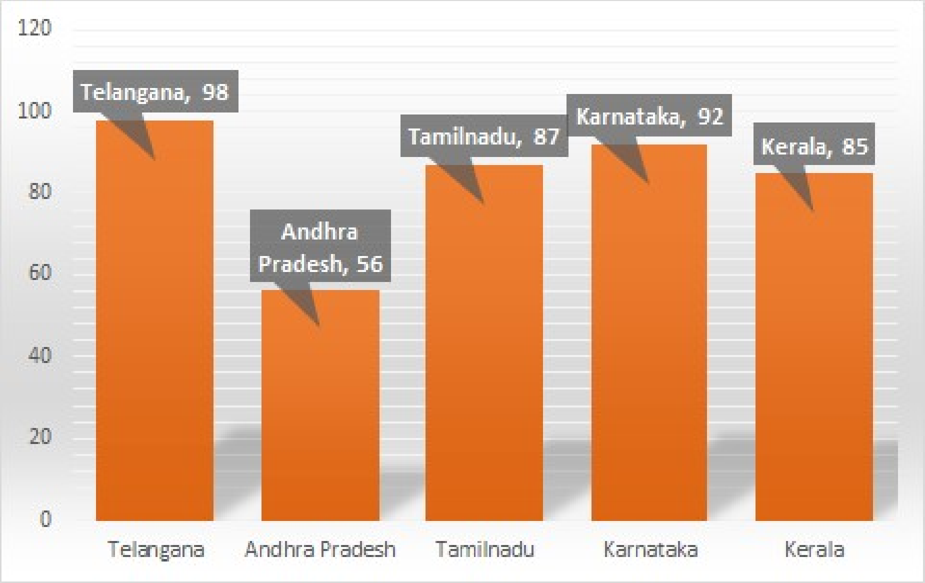

Figure 11 : Record-breaking winter of accidents

States and union territories having the greatest rate of accidents during the winter season. Lakshadweep ranks top, followed by Mizoram, Nagaland, D&N Haveli, and Daman & Diu.

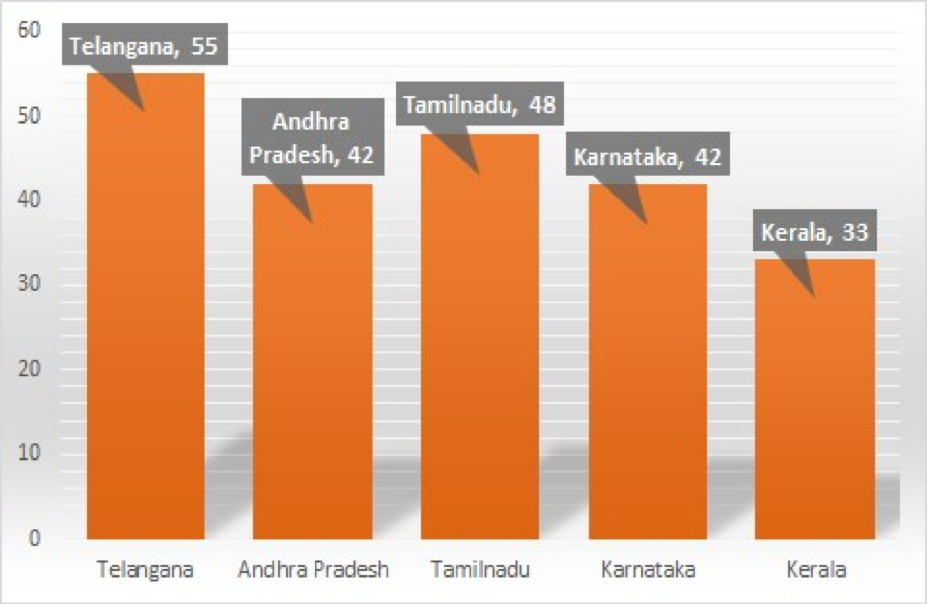

Figure 12: Record-breaking surge in autumn accidents

States/Union Territories in India having the greatest incidence of accidents during the fall season. Lakshadweep has the highest ranking, followed by Punjab, Himachal Pradesh, Haryana, and Jammu & Kashmir.

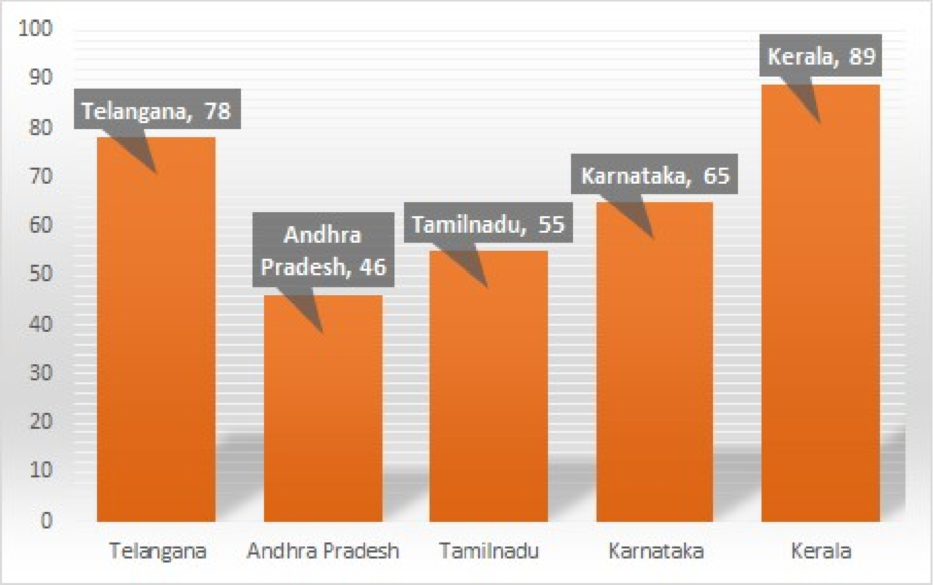

Figure 13: Highest incidence of accidents due to rain

The following data displays the States/Union Territories with the greatest number of accidents during the rainy season. Jammu & Kashmir has the greatest incidence of accidents during the rainy season, followed by Himachal Pradesh, Pondicherry, Delhi UT, and Tamil Nadu.

VI. CONCLUSION

Every year, road accidents result in deaths and related economic losses globally. Therefore, it is the foremost area of societal concern when considering loss prevention. Modeling accident severity prediction and upgrading the model are crucial to the successful operation of road traffic systems for increased safety. In accident severity modeling, the input vectors consist of many data related to the accident, such as driver traits, roadway conditions, and environmental factors. The output vector represents the specific class of accident severity. We have created two classifiers, a decision tree classifier and a KNN classifier, for the purpose of predicting the severity of accidents. Both classifiers exhibit high accuracy rates, with KNN achieving a superior accuracy of 88.3% compared to the 84.5% accuracy of the decision tree classifier. The identification of the primary elements that impact the severity of accidents may provide valuable insights for Government Departments/Authorities such as the Police, R&B, and Transport, from a public policy perspective. The Departments may use the findings of research and modeling to implement effective actions aimed at mitigating the effects of accidents and thus enhancing traffic safety. Insurers benefit from it via less claims, improved underwriting, and more accurate rate setting. This approach may be used to forecast the severity of accidents caused by different circumstances, hence aiding in the improvement of accident management.

REFERENCES

[1] B Pang and L Lee, “Opinion mining and sentiment analysis,” Foundations and Trends in Information Retrieval 2(1-2), 2008, pp 1–135.

[2] M Hu and B Liu, "Mining and summarizing customer reviews," Proceedings of the tenth ACM international conference on Knowledge discovery and data mining, Seattle, 2004, pp 168-177.

[3] B Pang, L Lee, and S Vaithyanathan, “Thumbs up? sentiment classification using machine learning techniques,” Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol10, 2002, pp 79-86.

[4] Jie Yang University of Wollongong, Australia “Mining Chinese social media UGC- a big-data framework for analyzing Douban movie reviews”, Journal of Big Data Springer, 2016.

[5] Kia Dashtipour Scotland, United Kingdom “Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques”, Springer, 2016.

[6] Kigon Lyu Korea University, Korea “Sentiment Analysis Using Word Polarity of Social Media”, Springer, 2016.

[7] Monu Kumar Thapar University, Patiala “Analyzing Twitter sentiments through big data”, IEEE, 2016.

[8] Minhoe Hur Seoul National University “Box-office forecasting based on sentiments of movie reviews and Independent subspace method”, Information Sciences, 2016

[9] Jorge A Balazs University of Chile “Opinion Mining and Information Fusion- A survey”, 2015.

[10] Donglin Cao Xiamen University, China “A cross-media public sentiment analysis system for microblog”, Springer, 2014.

[11] Ashraf I, Hur S, Shafiq M & Park Y. Catastrophic factors involved in road accidents: Underlying causes and descriptive analysis. PLoS one 2019 14(10), e0223473.

[12] Singh H, Kushwaha V, Agarwal AD & Sandhu SS. Fatal road traffic accidents: Causes and factors responsible. Journal of Indian Academy of Forensic Medicine 2016)., 38(1), 52-54.