Abstract

1.32 lakh individuals lost their lives in traffic accidents in 2020, the fewest in 11 years. The lowest number was 1.26 lakh in 2009. The number of road accidents decreased to 3.66 lakh last year, the smallest amount in the previous 20 years. The strong restrictions concealed in these widely used item sets usually expose the connections between the factors that influence accidents, which may be used to break them and reduce the frequency of accidents. The recommendations may also be utilised to look into regular accident spots and execute the necessary security upgrades to reduce accidents and, as a consequence, improve traffic safety in the city. Generally speaking, association rule mining can provide a lot of weak rules; As a result, the study first developed a method for determining the minimal Support value of training parameters, and then it suggested a strategy for automatically generating strong rules. The outcomes of the trial demonstrated the efficacy of the research's suggested techniques. In order to facilitate the effective use of association rule mining in intelligent transportation systems, an automated modelling approach based on association rules was created.

Keywords
Accident prediction, machine learning, regression, classification, smart vehicle, automobile

Introduction

One of the main causes of fatalities and economic losses worldwide, both in developed and developing nations, has been identified as traffic accidents. Malaysia had 363,319 incidents in 2007, with an average of 18 people dying in road accidents every single day. In addition, in 2007 has resulted in 6,282 fatalities, with a societal cost of almost $8 billion. According to the Royal Malaysia Police, there were 23.6 fatalities in Malaysia for every 100,000 people in 2006, which is among the highest rates in the world when compared to the Netherlands and Japan, which had 4.5 and 5.7 deaths per 100,000 people, respectively. In Malaysia, motorbikes account for 49% of all registered vehicles, while cars are listed in second place with 45% of the total. The majority of traffic collision fatalities—67%—are caused by cars, followed by motorbikes, which account for more than 16% of all fatalities (Royal Malaysia Police, 2009) In order to recommend safety measures, the authorities, including the State Department of Transportation, may be interested in identifying problem locations.

Transportation engineer would be interested in identifying those factors (traffic flow, speed, road geometric and etc) that influence accident frequency to improve the roadway design and provide safer driving environment. Previous studies have proved that improvement of highway design should effect

significant reduction in the number of crashes. The paper concerned with investigating major factors contributing to highway accident, this study concentrating the relationship between road condition, traffic flow, accident rates and their predicting, using multiple non-linear regression (MLR). The ability to predict accident rates is very important to transportation planner and engineers, because it can help in identifying hazardous location , sites which require treatment and as well as ranking the black spot locations.
Traffic flow, speed, road geometry, and other variables that affect accident frequency should be identified by a transportation engineer in order to enhance road design and provide a safer driving environment. Previous research has demonstrated that a considerable reduction in accidents should result from improved roadway design. This study focused on the link between road quality, traffic flow, accident rates, and their prediction using multiple non-linear regression to investigate the primary causes causing highway accidents (MLR). For transportation planners and engineers, the ability to estimate accident rates is crucial since it may be used to identify dangerous places, sites that need to be treated, as well as rank the black spot locations.

Data description:
The data has been collected from the website called Bitbucket. Bitbucket is a hosting platform that offers developers access to

both public and private repositories. The 3 CSV files are taken from the Bitbucket data of the user Abdul Wahed.
The Data consists of the following 3 csv files:
Accident.csv:
Accident location, date, time, police visited, no of vehicles involved, weather condition, surface condition, light condition
Vehicle.csv:
Information about vehicle, vehicle age, driver's information, if vehicle was left handed,enginer capacity, propulsion code, age band of driver
Casualties.csv:
Casualty age, type , reference, if pedestrian was present, location of pedestrian,casualty home area
The following columns are present in the merged data:
['accident index', 'location easting osgr', 'location northing osgr', 'logitude', 'latitude', 'police force', 'accident severity', 'number of vehicles', 'number of causalities', 'date', 'day of week time', 'local authority (district)', 'local authority (highway)', '1st road class', '1st road number', 'road type', 'speed limit', 'junction detail', 'junction control', '2nd road class', '2nd road number', 'pedestrian crossing human control', 'pedestrian crossing physical facilities ', 'light conditions', 'weather conditions', 'road surface conditions', 'special conditions at site', 'carriageway hazards', 'urban or rural area', ' did police attend the scene', 'lsoa of accident location', 'vehicle reference x', 'vehicle type', 'towing and articulation', 'vehicle maneuver', 'vehicle location

restricted time', 'junction location', 'skidding and overturning', 'hit object off carriageway', '1st point of impact', 'was vehicle left hand drive', 'journey purpose of driver', 'sex of driver', 'age of driver', 'age band of driver', 'engine capacity cc', 'propulsion code', 'age of vehicle', 'driver imd decile ', 'driver home area type', 'vehicle reference_y', 'casualty reference', 'casualty class', 'sex of casualty', 'age of casualty', 'age band of casualty', 'casualty severity', 'pedestrian location', 'pedestrian movement', 'car passenger', 'bus or coach passenger', 'pedestrian road maintenance worker', 'casualty type', 'casualty home area type']

Data cleaning-

Data cleaning is the procedure of removing inaccurate or irrelevant information from the data before analysis. This type of information typically reinforces a false belief, which might have a detrimental effect on the model or algorithm it is given into. In addition to eliminating large portions of irrelevant data, data cleaning also frequently refers to correcting inaccurate data in the train-validation-test dataset and minimising duplicates.

Before conducting any type of analysis on data, data cleansing is a crucial step.

In pipelines, datasets are frequently gathered in small groups, combined, and then fed into a model. When several datasets are combined, duplicates and redundancies are created in the data that must be deleted.

Data preprocessing steps

1. The “Accident.csv.” data contains information about accident location, road conditions, weather conditions, light conditions, etc that contributes to the accident of the vehicle. The “Vehicle.csv” data includes information about the vehicle that has been involved in the accident along with its specification as well as information about the driver. The “Casualty.csv” data contains information about the Accident severity, Casualties included into accident, Casualty type, Age of the Victim in the accident along with the information about the location and locality. All these data contributes to the accident prediction model but are present individually so there is need to using all these data together, so all the 3 datasets are merged together based on the accident index.

The accident index shows the instance of the same accident so “Inner join” is applied to all the tables and merged together to form a dataset containing 64 columns.

The next step in preprocessing is removal of the irrelevant columns in the data. Irrelevant column would be the column that cause very less effect in the prediction of the accident. The following columns were removed from the dataset resulting into final 55 columns:

[Police_Force, Did_Police_Attend_the_Scene, Age_of_Driver, Propulsion_Code, Driver_imd_Decile, Driver_home_area_type, Age_of_Casualty, Casulty_home_area_type]

Machine learning techniques used:

Ml definition:
Computer science's branch of machine learning tries to teach computers how to learn and behave without having to explicitly programme them. Machine learning is a method of data analysis that entails creating and modifying models so that computers may "learn" from their experiences. Building algorithms for machine learning entails customising their models to enhance prediction accuracy.
A computer programme is said to learn from experience E with regard to some task T and some performance measure P if its performance on T, as measured by P, increases with experience E, according to Tom Mitchell, professor of Computer Science and Machine Learning at Carnegie Mellon. A software is said to apply machine learning if it becomes better at solving problems as it gains experience.
Linear Regression
The linear regression procedure, often known as linear regression, demonstrates a linear relationship between a dependent (y) and one or more independent (y) variables. Given that linear regression demonstrates a linear connection, it may be used to determine how the dependent variable's value changes as a function of the independent variable's value.
Lasso Regression
A shrinkage and variable selection technique for linear regression models is called lasso regression analysis. Finding

the subset of predictors that minimises prediction error for a quantitative response variable is the aim of lasso regression. The lasso does this by placing a restriction on the model's parameters, which leads some variables' regression coefficients to converge to zero. After the shrinking procedure, variables having a regression coefficient of zero are not included in the model. The variables most closely related to the response variable are those with non-zero regression coefficients.

Stochastic Gradient Descent

The best parameter setting for a machine learning algorithm may be found via stochastic gradient descent. To reduce the inaccuracy of the network, progressively modest tweaks are made to the setup of a machine learning network.

RANSAC Regression

By removing outliers from the training dataset, the RANSAC (RANdom SAmple Consensus) method elevates the linear regression technique. The coefficients and parameters that were learnt during training are affected by the presence of outliers in the training dataset. Therefore, it is advised to spot and eliminate outliers during the exploratory data analysis phase.

Multinomial Naïve Bayes

A specific kind of naive bayes called multimodal naive bayes, sometimes known as multinomial naive bayes, is created to handle text data utilising word counts as its fundamental

technique of computing probability. Popular in Natural Language Processing is the Bayesian learning method known as the Multinomial Naive Bayes algorithm (NLP). Using the Bayes principle, the computer makes an educated prediction about the tag of a text, such as an email or news article. It determines the likelihood of each tag for a certain sample and produces the tag with the highest likelihood.

PERFORMANCE METRICS:

Metrics for machine learning are used to evaluate how successfully a model learned from the input data that was given to it. By adjusting the hyperparameters or changing the input dataset's properties, the performance of the model may be enhanced in this way. A learning model's primary objective is to perform well on data that has never been seen before. Metrics of performance aid in assessing how effectively the model generalises to new data.

Precision:

The accuracy is computed as the ratio of Positive samples that were correctly categorised to all samples that were classified as Positive (either correctly or incorrectly). The precision gauges how well the model categorises a sample as positive.

Recall

The recall is determined as the proportion of Positive samples that were properly identified as Positive to all Positive samples. The recall gauges how well the model can identify Positive samples. The more positive samples that are identified, the

larger the recall. Only the classification of the positive samples is important to the recall. The classification of the negative samples has no bearing on this.

Accuracy

An indicator of the model's performance across all classes is accuracy. When all classes are equally important, it is helpful. The number of accurate forecasts divided by the total number of predictions is used to compute it.

Cohen Kappa score:

When two raters are assessing the same amount, Cohen's Kappa, a statistical metric, is used to assess the reliability of the two raters and determine how frequently the raters agree. This function calculates Cohen's kappa, a measure of how well two annotators agree on a categorization task. It's described as

where is the predicted agreement when both annotators give labels randomly, and p0 is the empirical probability of agreement on the label assigned to each sample (the observed agreement ratio). A per-annotator empirical prior over the class labels is used to estimate p0.

Modern humans arrived on the Indian subcontinent from Africa no later than 55,000 years ago.[26][27][28] Their long occupation, initially in varying forms of isolation as hunter-gatherers, has made the region highly diverse, second only to Africa in human genetic diversity.[29] Settled life emerged on the subcontinent in the western margins of the

Indus river basin 9,000 years ago, evolving gradually into the Indus Valley Civilisation of the third millennium BCE.[30] By 1200 BCE, an archaic form of Sanskrit, an Indo-European language, had diffused into India from the northwest.[31][32] Its evidence today is found in the hymns of the Rigveda. Preserved by a resolutely vigilant oral tradition, the Rigveda records the dawning of Hinduism in India.[33] The Dravidian languages of India were supplanted in the northern and western regions.[34] By 400 BCE, stratification and exclusion by caste had emerged within Hinduism,[35] and Buddhism and Jainism had arisen, proclaiming social orders unlinked to heredity.[36] Early political consolidations gave rise to the loose-knit Maurya and Gupta Empires based in the Ganges Basin.[37] Their collective era was suffused with wide-ranging creativity,[38] but also marked by the declining status of women,[39] and the incorporation of untouchability into an organised system of belief.[i][40] In South India, the Middle kingdoms exported Dravidian-languages scripts and religious cultures to the kingdoms of Southeast Asia.[41]

In the early medieval era, Christianity, Islam, Judaism, and Zoroastrianism became established on India's southern and western coasts.[42] Muslim armies from Central Asia intermittently overran India's northern plains,[43] eventually founding the Delhi Sultanate, and drawing northern India into the cosmopolitan networks of medieval Islam.[44] In the 15th

century, the Vijayanagara Empire created a long-lasting composite Hindu culture in south India.[45] In the Punjab, Sikhism emerged, rejecting institutionalised religion.[46] The Mughal Empire, in 1526, ushered in two centuries of relative peace,[47] leaving a legacy of luminous architecture.[j][48] Gradually expanding rule of the British East India Company followed, turning India into a colonial economy, but also consolidating its sovereignty.[49] British Crown rule began in 1858. The rights promised to Indians were granted slowly,[50][51] but technological changes were introduced, and modern ideas of education and the public life took root.[52] A pioneering and influential nationalist movement emerged, which was noted for nonviolent resistance and became the major factor in ending British rule.[53][54] In 1947 the British Indian Empire was partitioned into two independent dominions,[55][56][57][58] a Hindu-majority Dominion of India and a Muslim-majority Dominion of Pakistan, amid large-scale loss of life and an unprecedented migration.[59]