***Analysis of Accidents Based on Weather, Light and Road Conditions***

***Data Warehousing and Data Mining Lab (ICT 3262)***

***Dept. of Information & Communication Tech.***

|  |  |
| --- | --- |
| ***Name*** | ***Registration Number*** |
| ***Pasupuleti Gomathi Sowmya*** | ***200911124*** |
| ***Sesha Mohan Srivatsava Maddali*** | ***200911176*** |
| ***Pooja Bhogadi*** | ***200911178*** |

***Submission date: 05-April-2023***

***Introduction:***

There are several features that favour the occurrence of road accidents, which are a top reason for deaths and serious harms/injuries globally. One of these elements is the weather, which can have an impact on, among other things, driving behaviour, road conditions, and visibility. The intention of this study is to scrutinise traffic accidents depending on atmospheric conditions like weather, light, etc. to comprehend how they are related to accident severity along with considering some other factors as well like location, number of vehicles, speed limit, etc.

The weather can have a significant impact on driving behaviour, road conditions, and accessibility, all of which can contribute to traffic accidents. In this study, we will look at the relationship between atmospheric variables and traffic accidents, taking other factors like visibility, road conditions, and other drivers into account.

We will first browse the dataset to look for patterns and trends before we start our analysis. To comprehend the connection between atmospheric conditions and traffic accidents, we will use statistical analytic techniques including correlation and regression analysis. In order to examine the data and spot any patterns or outliers, we will also employ data visualisation tools like scatter plots, bar graphs, correlation matrices and heat maps.We will also take into account additional elements that could affect the frequency and seriousness of incidents, such as location, the number of vehicles, speed limits, etc.

In this study, we will also look at weather-related factors such as fog, rainfall, snow, and severe winds. We'll examine how these factors impact the frequency, seriousness, and location of accidents as well as their number. For instance, we might discover incidents are more serious and frequent when it rains, and they happen more frequently on particular roads or highways when it snows.

The outcomes from this investigation can direct decisions made by the government regarding road safety. For example, policymakers could adjust speed limits for certain weather conditions, budget money for road maintenance in areas with a high accident rate for certain weather conditions, or design targeted safety and health campaigns to increase awareness among the populace about the harms associated with driving in specific weather conditions. This project's overarching goal is to improve public safety and decrease the frequentness and seriousness of such accidents.

Apart from atmospheric conditions, there are several other factors that could possibly contribute to the frequentness and severity of accidents. For instance, the location of the accident, the number of vehicles involved, the speed limit, and road conditions are all important factors to consider. Additionally, the speed limit could affect the likelihood and severity of accidents, with higher speed limits potentially leading to more serious accidents. By analysing these factors in conjunction with atmospheric conditions, we can gain a more comprehensive understanding of the causes and effects of such accidents, and try to mitigate their impacts.

***About The Data Set:***

Data concerning the weather and road conditions when an accident occurred, such as rainfall, fog, snow, and wind, as well as data about the road conditions, such as slippery or wet roads, often are included in accident reports collected by the officials. Other variables that may affect accidents, like time or, the accident's location, and types of cars involved, may also be included in the data. These can also be considered in the analysis to achieve higher accuracy. To find trends and patterns in accidents related to weather and road conditions, this data is often gathered by government organisations and put into databases which can be examined by academics and analysts.

**Dataset 1:**

We used one of such datasets called “UK\_Accidents.csv” which is available from “UK Road Safety” online. It contains a total of 32 columns. Below are the columns along with their initial data types:

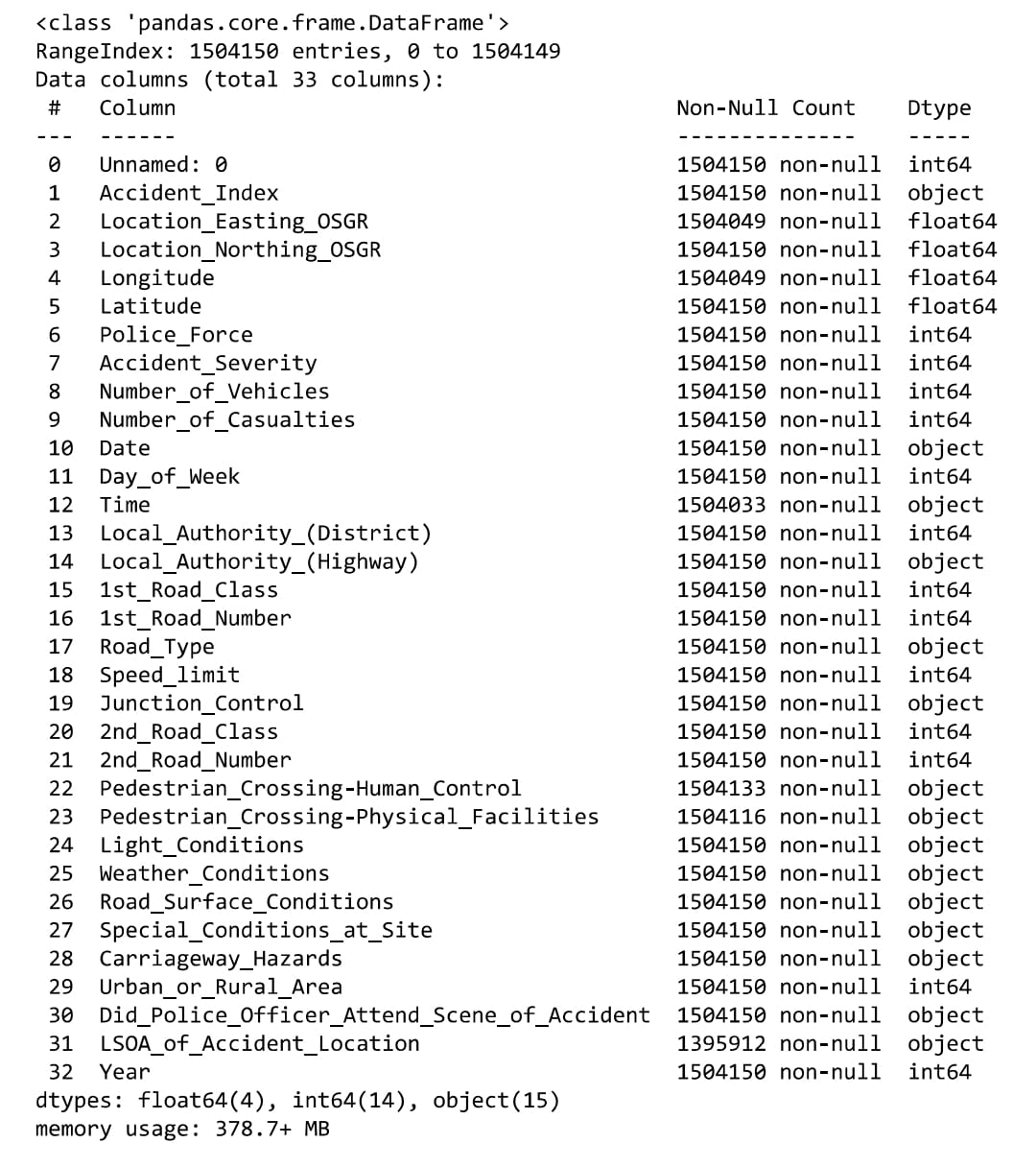
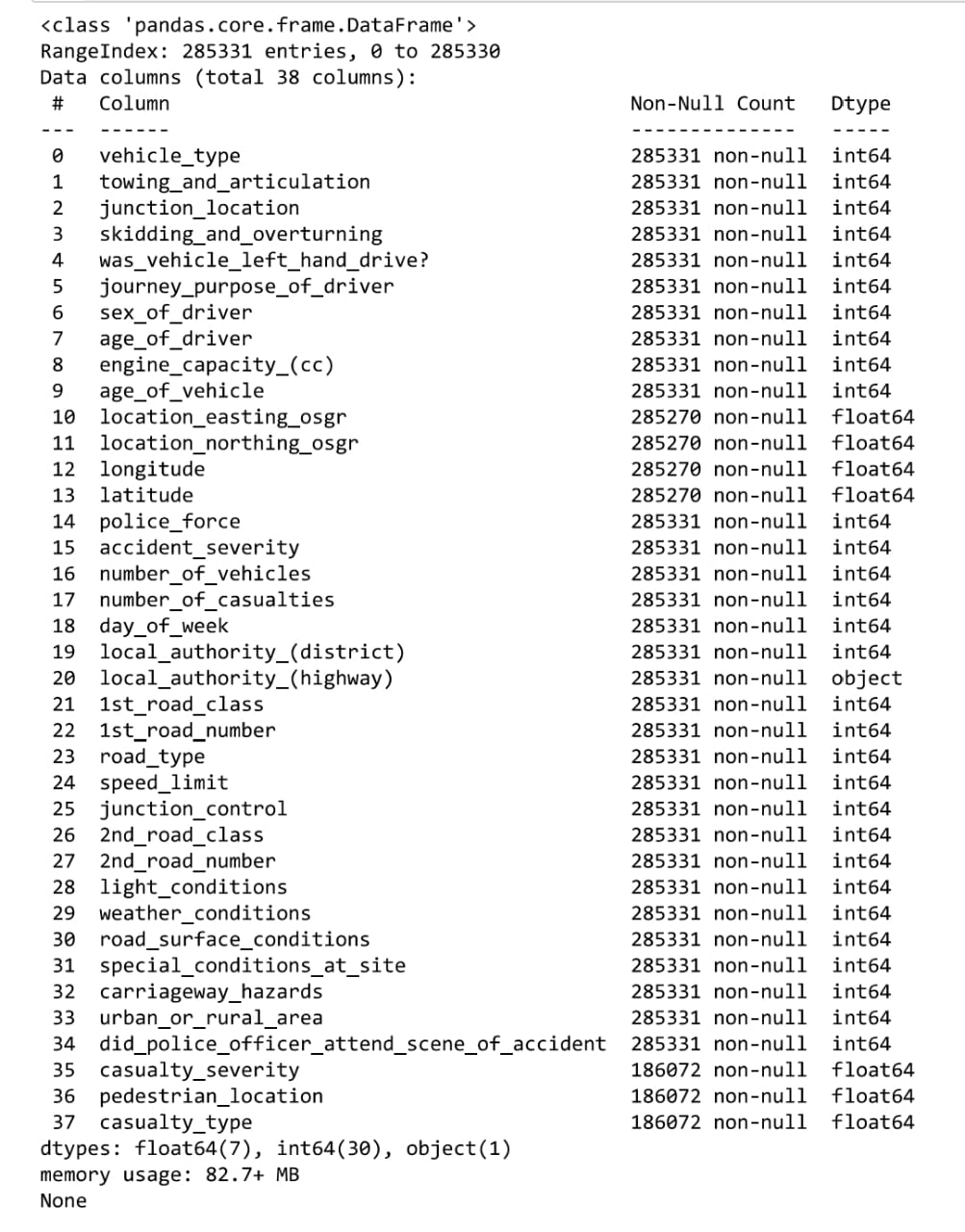


Fig 1 (Attributes of dataset-1)

**Dataset 2:**

This dataset has around 70 columns and 250,000 rows, and primarily captures road accidents in the UK between 1979 and 2015. This dataset was already merged and cleaned and was available online. Below are the 37 columns we considered along with their initial data types:

 Fig 2 (Attributes of dataset-2)

***Literature review:***

**[1]** discusses the effects of conditions like weather and light on road accidents. The study examines the relationship between weather conditions and crashes and uses regression techniques to analyse the data. The paper presents the relevant literature, the methodology used, data processing, results, and conclusions. The article concludes that conditions like weather and light have a considerable impact on road accidents and recommends further research to determine the most appropriate statistical models for analysis.

**[2]**analyses the correlation of various road weather conditions to their impact on the risk of accidents. A novel approach is utilised, whereby the “palm distribution” of various weather conditions is computed. Then it is compared against the distribution of accidents under similar weather conditions. The study carried out on 43 primary roads in Finland during the period of 2014-2016, analysing all single and multi-vehicle accidents reported by the police by collecting hourly road and weather conditions from nearby road weather stations, indicated that the greatest accident risk is associated with slippery road conditions and icy rain. Furthermore, the risk of accidents is higher on motorways than on multi-lane and two-lane roads. Additionally, it is found that single-vehicle accidents are more dangerous than multi-vehicle accidents.

**[3]** from 21 years of daily counts, this study examines how environmental conditions affect highway safety in Athens, Greece. The impacts of weather on accident categories—vehicle accident, vehicle deaths, pedestrian collisions, and pedestrian fatalities—were estimated using an integer autoregressive model (INAR). According to the study, the variables that consistently showed the most significance and influence were the mean daily precipitation height and its lag value. Increases in rainfall, in contrast to earlier research, were found to decrease both the overall amount of accidents and the number of pedestrians’ accidents and fatalities. This was linked to the safety offsetting hypothesis brought on by slower and more cautious driving. The study also discovered that as temperature rose, accidents rose as well.

**[4]**is based on the analysing of the aggregate effect of atmospheric circumstances on the amount of accidents. One drawback of this method is that it can’t talk about the effect of these conditions separately. This method should be used only when the relevant atmospheric conditions are known for the data and should be dichotomous. An analysis was done by trying to find a connection between the daily accumulated travel and road fatality data.

**[5]** analysed the effects of weather conditions on accidents in Maryland between 2007-2010 by the use of three models, examining relevant variables and found that the Multilayer Perceptron Model is performing the best as the accident severity model. The research has arrived at a conclusion that all weather factors except visibility and heavy precipitation significantly impacted accident severity, and the MNL regression model was more better at finding this than other models.

**[6]**explores the circumstances that cause accidents and their severity, using data from Setúbal, Portugal, from 2016 to 2019 by utilising machine learning algorithms, including supervised methods such as “decision trees, random forests, logistic regression, and naive Bayes, as well as unsupervised techniques like DBSCAN and hierarchical clustering”, to develop models that can predict the severity of an accident depending on influential factors. Results show that the “C5.0 algorithm” can correctly detect the most influential factors for finding accident severity, and the “RF model” is a good tool for predicting accident occuring hotspots.

**[7]** discusses the utilisation of machine learning techniques to try and estimate road accidents and their severity. The paper provides a review of the research in prediction of accident using clustering techniques and machine learning algorithms. The authors review related work and propose a system to analyse the data using Python, pandas, numpy and regression analysis. The authors used regression analysis, specifically logistic regression analysis.

**[8]** evaluates risk factors that have a substantial impact on crash severity levels. It also looks at the reasons, conditions, and distribution of locations with a high frequency of accidents. Two supervised ML systems, random forest and AdaBoost, are used and contrasted to forecast the severity level and upcoming crashes. An unsupervised algorithm called Association rules is also applied.

**[9]** uses generalised additive models and combined overall effect of volume of traffic and weather related parameters on periodic probabilities of around 78 different types of crashes. The authors have found that snow has had the biggest effect on single truck crashes,while on the other hand raining has had a much-higher effect on “single crashes” and that “sun glare” increased the occurrence of multiple car crashes especially at increased speed limits.

**[10]**focuses on the issue of rapid detection of accidents to decrease congestion, secondary crashes, and the wastage of resources. The paper proposed a deep learning approach which relies on a “long short-term memory neural network” to identify potential accident occurrences by forecasting speed values on freeway links and extracting features such as weather and traffic flow.

**[11]** In order to forecast the severity of weather-related collisions in Connecticut, this paper compares the effectiveness of some ML-based models: “random forest (RF)” and “bayesian additive regression tree (BART)”. The findings indicate that in terms of predicting probabilities and skill score, the RF model performs better than BART. The results indicate that people can utilise the “RF model” to get more accurate crash severity estimations.

**[12]** suggests a smart system that makes use of a visibility range estimating system. The study reviews the body of exisitng research on evaluating visibility distance in cloudy weather, then proposes a neural network method for estimating visibility range with a camera that may be fixed to an RSU or positioned on a vehicle. It is possible to make 4-wheelers and other vehicles intelligent by integrating the suggested technology.

**[13]**investigates important features, such as crash severity, the amount of fatalities, and the vehicles involved, in order to analyse data patterns related to road automobile accidents and suggest a predictive model. As a result, a model is created to transform raw data utilising interquartile outlier removal, generalising data attribute, and removing missing and nonsensical features. Four classification techniques—”decision trees”,” random forests”, “multinomial logistic regression”, and “naive Bayes”—are employed and analysed. With the exception of the naïve Bayes classifier, the results indicate adequate accuracy levels for car accident prediction.

**[14]** “Convolutional Long Short-Term Memory neural network” model is used in this paper to conduct an extensive investigation on the traffic accident prediction problem. Several specific details, including the climate, surroundings, and road-surface condition as well as traffic are collected from large datasets covering Iowa over an eight-year period. They present a Hetero-ConvLSTM framework, which adds a few additional concepts to the core ConvLSTM model to solve the problem of spatial heterogeneity in the data. This study suggests that this proposed model greatly outperforms baseline approaches in terms of prediction accuracy and provides reasonably accurate forecasts.

**[15]** suggested a strategy that uses an “ensemble learning classifier” built from “sequential minimal optimisation” and “decision trees”. The “Lebanese Road Accidents Platform (LRAP)” database of 8482 road crashes occurrences served as the basis for the model's construction, training, testing, and validation. The outcome variable used in the analysis was the frequency of fatalities. To investigate the effects of various conditions on the occurrence of deaths, a sensitivity study was carried out. Seven of the nine independently chosen variables—”crash type”, “injury severity”, “spatial cluster-ID”, and “crash time”—were significantly linked to the occurrence of fatalities (hour).

**[16]**studies aboutroad accidents, a significant public concern in Bedouin Arabian Gulf nations, are highlighted in this study's abstract, especially in the United Arab Emirates. (UAE). Oil has been found in the area, which has altered lifestyles by increasing wealth and the quantity of vehicles on the highways. But because typical patterns of behavior have not changed as quickly, incidents involving aggressive driving and traffic violations have occurred. The most significant contributing factor to RTAs, accounting for more than 35% of all events, according to this investigation's current information on RTAs and roadway traffic behavior in the United Arab Emirates, was excessive speed.

**[17]**emphasizes how climate change may affect the incidence of RTAs in Saudi Arabia, a nation with a high rate of RTA-related fatalities and frequent meteorological extremes. Despite the potential societal consequences of RTAs, no prior research has looked at how climate change may affect this problem in Saudi Arabia. The impact of rainfall, temperature, sandstorms, and vehicle density on RTAs is examined using panel regression models and yearly data across 13 locations in Saudi Arabia between 2003 and 2013. These results demonstrate that each of these factors had a significant impact on RTAs in the nation throughout the period of study. It also demonstrates that, within and outside of cities, both, RTAs greatly increase the risk of fatalities.

**[18]** emphasises India's worrisome road accident condition, which varies by age, gender, time of day, and region. The study concludes that men are more likely to die and sustain injuries than women in the 30- to 59-year-old age range. In addition, the results show that there is a considerable variance in fatality risk between Indian states and cities, with 16 of the country's thirty-five states and union territories showing a greater fatality risk than the average. The document emphasises the urgent need for expanded efforts and innovative initiatives to address the deteriorating condition of road crashes and fatalities in India because, in the absence of proper action, the overall number of road traffic deaths is anticipated to reach 250,000 by 2025.

**[19]**emphasises the value of using machine learning and open data to solve complicated issues in a variety of fields, including traffic. The research specifically suggests a hybrid model, for forecasting the danger of road accidents in traffic based on various atmospheric parameters using historic and current data. A two-agent system is used to implement the model, with one of the agents which learns from historic data while the other one gathers from real-time data. This model was evaluated by the authors using a case-study for the Serbian city - Ni, and it was then put into use as a web citizens application. Overall, the study offers a strategy that uses real-time data and machine learning to enhance traffic safety.

**[20]** examines the association between climate factors, social development indicators, and the frequency of accidents using a negative binomial framework and a “log-change model”. The findings demonstrate that non-climate variables like beer consumption, the ratio of rural to urban vehicle kilometres driven, and performance of vehicles have a considerable effect on fatal traffic accidents, as do climate variables like the average temperature and average precipitation.

**[21]**Using epidemiological data released between 1989 and 2013, this systematic review attempted to compare the crash risk for “young drivers riding with passengers” versus “solo driving”. The findings showed that teenage drivers who had a minimum of a single passenger had a risk of fatal collisions that ranged from 1.24 to 1.89 times higher than those who drove alone. For groups of two or more passengers, the risk increased, with values ranging from 1.70 to 2.92. The study also discovered that male passengers and drivers who were younger were likely to be at risk of collisions. Based on this, the authors recommended licensing regulations that limit the amount of young passengers that “young drivers” may have in order to increase road safety.

**[22]**suggests a method for determining the average accident probability of highway sections using “fault tree” and “energy approaches”. The approach formulates the average crash probability of segments using fault tree theory while accounting for variables including lateral instability, rollover, and rear-end crashes. Additionally, it suggests an energy-based instability factor to quantify the degree of steering instabilities and tyre sideslip. The average crash risk for each portion was determined using a multi-body dynamics model created with CarSim of the driver, vehicle, and road. The suggested approach was found to increase evaluation results' accuracy while offering a theoretical framework for pinpointing problem areas on both in-service and design-phase roadways.

**[23]**adds to the body of prior knowledge on Spanish crosstown highways by analysing a hitherto unresearched viewpoint: that of the driver. The investigation's primary goal is to pinpoint the elements that increase the risk of a tragic outcome in single-vehicle collisions that happen on Spanish crosstown roadways.

**[24]** combines “Long Short-Term Memory (LSTM)” and “Gated Recurrent Units”, to identify accidents occurring in Chicago. (GRUs). The study makes use of a dataset from the Chicago motorways that includes spatiotemporal information of traffic, weather-related data, and traffic status data and includes 241 incident and 6,038 non-accident incidents. “Synthetic Minority Over-sampling Method” is used to address the unbalanced dataset. (SMOTE). The study indicated that both models perform significantly well, however the GRU model clearly is better than the LSTM model in terms of detection rate. The study demonstrates the potential benefits of LSTM and GRU models for road safety applications and contributes to the body of knowledge on identifying accidents using deep learning techniques.

In **[25]**this study, accident information from the Setbal district of Portugal were examined using a variety of machine learning approaches from 2016 to 2019. The main aim of the study was to create predictive models to predict future accidents based on historical data and to choose relevant parameters for categorising accident severity.

***Methodology:***

***FOR DATASET 1:***

**Initial steps:**

* For executing our codes for this project, we used Anaconda’s Jupyter notebook.
* The first step here, is to install all the required Python libraries like Pandas, Numpy, Matplotlib, etc
* After setting the path to our csv file and taking it into a frame, we then proceed to the data cleaning process where we fill all the missing values with appropriate values like zero, None, Unknown, etc.
* Make relevant plots that would assist in our analysis.
* Plotting the days of the week against total casualties, we get a bar graph indicating that *most accidents occur on Saturdays*.

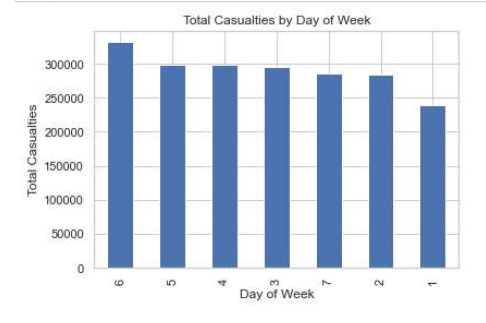


Fig 3 (Bar Graph representing the total casualties by Day of Week)

**Correlation Analysis:**

* To further study the relation between light and weather conditions on accident severity, we did a correlation analysis by constructing a correlation matrix and a heatmap on it. Here are the results:

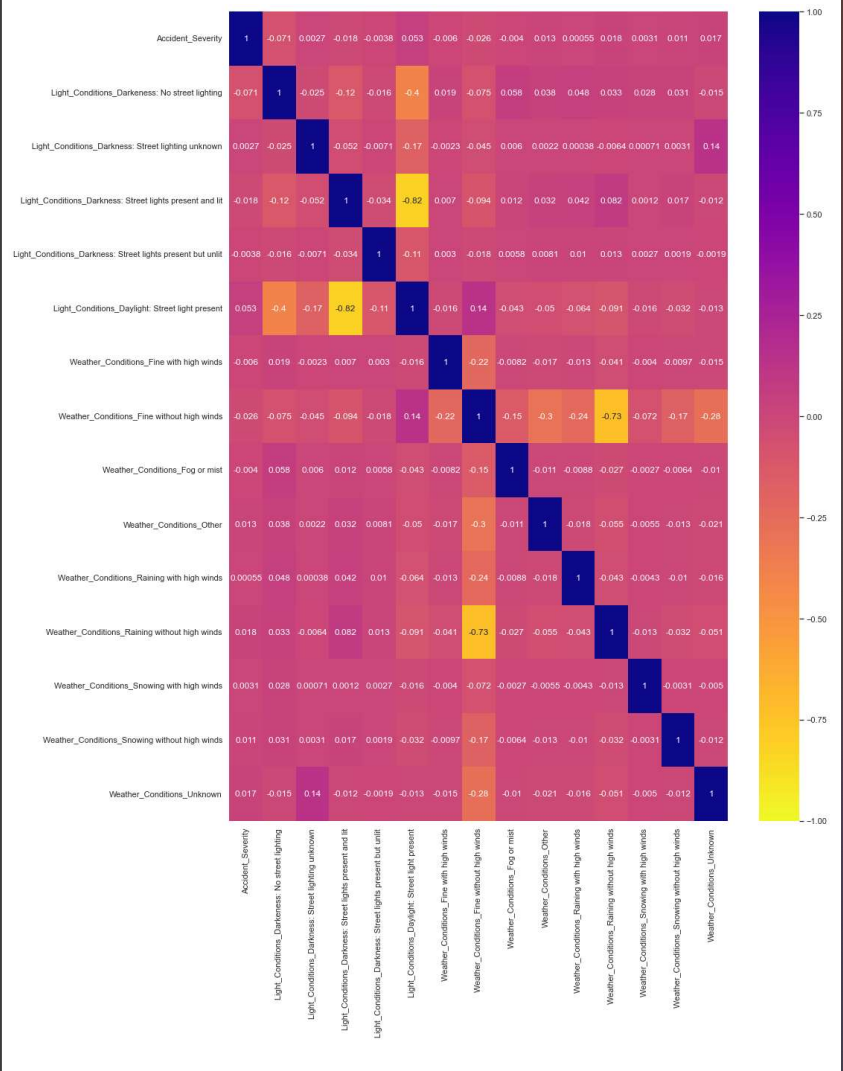


Fig 4 (Heatmap showing the correlation)

* The above results show a little to no correlation indicating that weather and light conditions alone do not play a part in accident severity. This made us look into other features of the dataset and consider them as well for our analysis.
* It is a well known perception that road conditions might affect accident severity. So we made a bar graph on road conditions and accident severity. Here are the results:

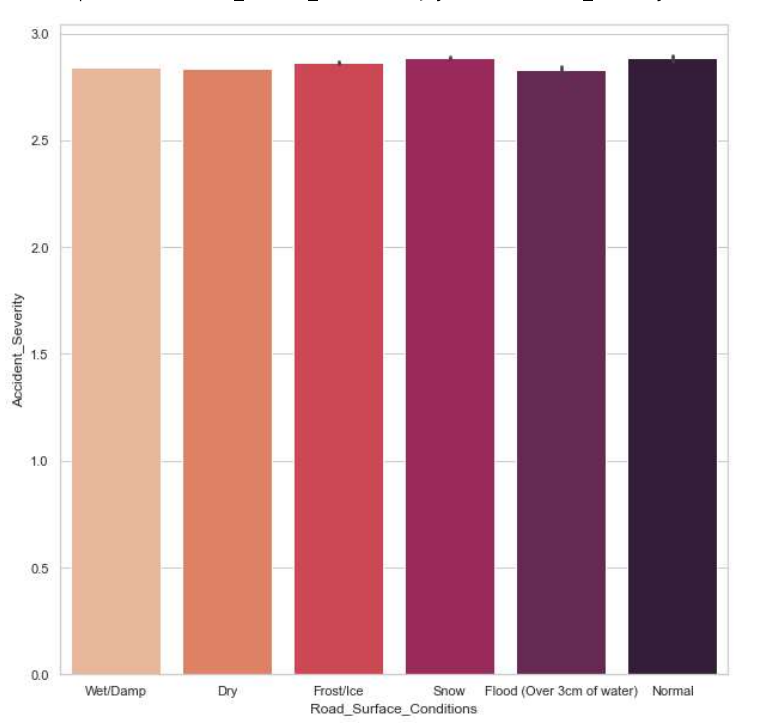


Fig 5 (Bar Graph on Road Conditions and Accident Severity)

* The above results are quite counter intuitive as they show that accidents occur as frequently on wet, dry, frost, snow road conditions as they do on normal roads. All these indicate that no single condition has a giant effect on the accident severity. Rather it is an aggregate effect of all the relevant features that would help us predict accident severity.
* Here is a correlation matrix having more numeric features into consideration:

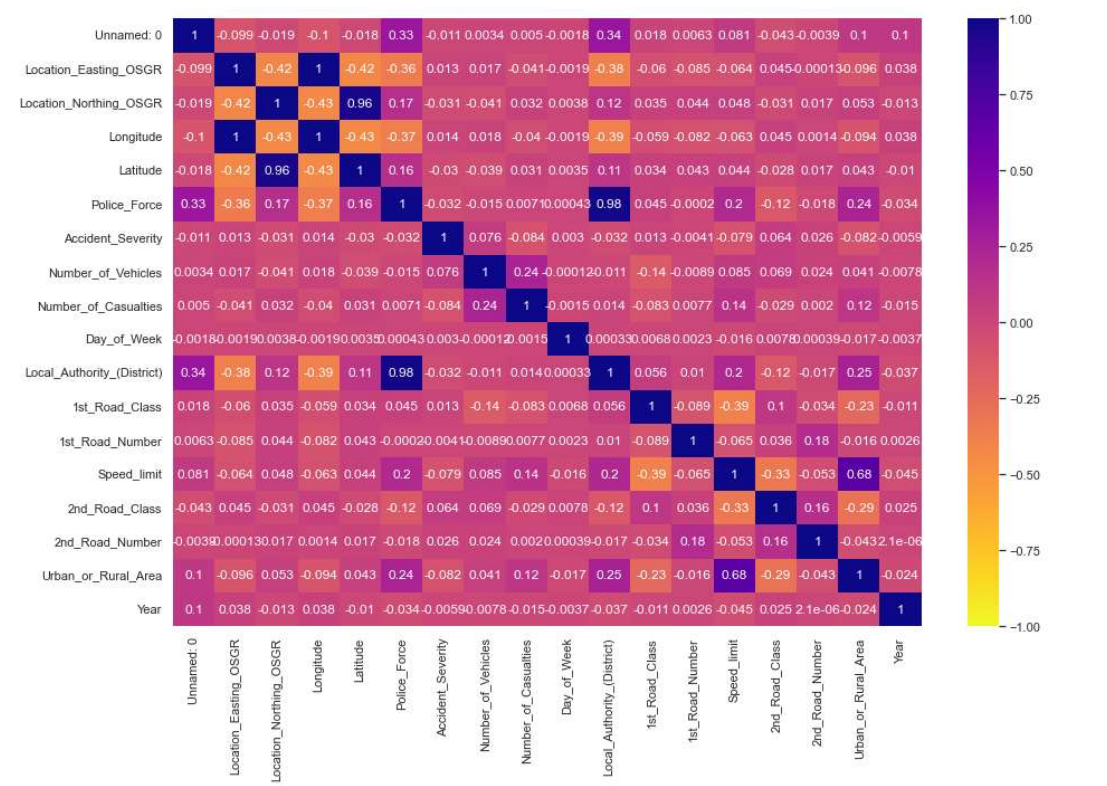


Fig 6 (Heatmap Correlation Matrix )

* As we can see from the above heatmap, it clearly indicates an improved correlation between various features and accident severity. This further strengthens the hypothesis that all the relevant features need to be taken into consideration for improved accuracy.

**Data Cleaning and Encoding:**

* To prepare the data for further data analysis, we need to clean the data for missing values, Nan values and we need to encode the non numeric data into numeric form (float64 or int64).
* First we are dropping the columns that are not related to our analysis or those that would affect our analysis in a negative way. These are:

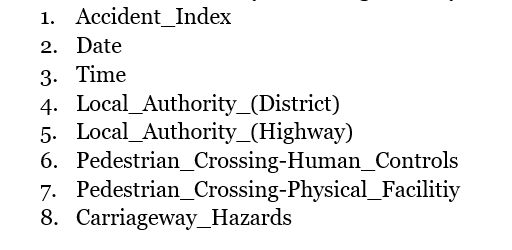


Fig 7 (Attributes to be dropped)

* Then for the remaining columns, we are required to encode the non numeric data types columns so that they can be useful for data analysis. For this purpose we used “***label encoding***” where we encoded each of the string values to a nominal value. This is done for all non numeric columns.
* Here are the string values and their nominal labels for the column “*Road\_Type*”:

*'Single carriageway': 1*

*'Dual carriageway': 2*

*'One way street': 3*

*'Roundabout': 4*

*'Slip Road': 5*

*'Unknown': 6*

* Here are the string values and their nominal labels for the column *“Junction\_Control”*:

*'None': 1*

*'Automatic traffic signal': 2*

*'Giveaway or uncontrolled': 3*

*'Stop Sign': 4*

*'Authorised person': 5*

* Here are the string values and their nominal labels for the column *“Light\_Conditions”:*

*'Day\_light: Street light presentt': 1*

*'Darkeness: Streeet lights presentt and lit': 2*

*'Darkeness: Street lighting not\_known': 3*

*'Darkeness: Street lights present but not\_lit': 4*

*'Darkeness: No street lightning': 5*

* Here are the string values and their nominal labels for the column *“Weather\_Conditions”:*

*'Raining without high\_wind': 1*

*'Fine without high\_wind': 2*

*'Unknown': 3*

*'Snowing without high\_wind': 4*

*'Other': 5*

*'Fine with high\_wind':6*

*'Raining with high\_wind':7*

*'Fog or mist':8*

*'Snowing with high winds':9*

* Here are the string values and their nominal labels for the column *“Road\_Surface\_Conditions”:*

*'Wet/Damp': 1*

*'Dry': 2*

*'Frost/Ice': 3*

*'Snow': 4*

*'Flood (Over 3 cm of water)': 5*

*'Normal':6*

* Here are the string values and their nominal labels for the column *“Special\_Conditions\_at\_Site”. A label of 0 was used for no special condition and 1 for any special condition that exists:*

*'None': 0*

*'Ol or diesel': 1*

*'Auto traffic singal partly defective': 1*

*'Road\_surface defactive': 1*

*'Auto\_traffic sygnal out': 1*

*'Permanent sign , marking defective or obscure': 1*

*'Mud': 1*

*'Roadworks':1*

* Here are the string values and their nominal labels for the column *“Did\_Police\_Officer\_Attend\_Scene\_of\_Accident”:*

*'No': 0*

*‘Yes': 1*

* After encoding all the non numeric columns,here are the data types of all the columns:

*Location\_Easting float*

*Location\_Northing float*

*Longitude float*

*Latitude float*

*Polece\_Force int*

*Accident\_Sevearity int*

*No\_of\_Vehicles int*

*No\_of\_Casualties int*

*Day\_of\_The\_Week int*

*1s\_Road\_class int*

*1st\_Road\_number int*

*Road\_type int*

*Speed\_Limit int*

*Junction\_Control int*

*2\_Road\_Class int*

*2n\_Road\_number int*

*Light\_condition int*

*Weather\_condition int*

*Road\_Surface\_Condition int*

*Special\_Conditions\_at\_The\_Site int*

*Urban\_r\_Rural\_Area int*

*Did\_a\_Police\_Officer\_Attend\_Scene int*

*Year int*

***FOR DATASET 2:***

**Initial steps:**

* Similar to what we did with dataset 1, we plotted the days of the week against total casualties. It showed that most accidents occur on *Saturdays*.

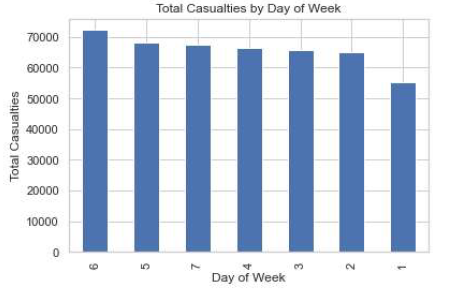


Fig 8 (Bar Graph on total casualties of the week)

* Since the dataset 2 is already cleaned we didn't need to perform any additional data cleaning and encoding.

**ALGORITHMS AND EVALUATION METRICS:**

* **Algorithms used:**

**1. Multiclass Logistic regression:** It is a statistical method used for predicting the categorical outcome of a dependent variable with more than 2 categories. The model uses a softmax function to model the probability of each category. We split the columns into two groups, one group containing *'Accident\_Severity'* and the other group containing all the remaining columns. We used get\_dummies to convert categorical variables to binary variables. Then we used cross validation to fit the regression model and evaluated the model. We used 1000 iterations to calculate cross-validation scores. The logistic regression model is a generalised linear model that estimates the probability of an event occurring based on one or more predictor variables. The logistic regression equation:

p(x) = 1 / (1 + e^(-xβ))

**2. Decision Tree:** First we imported the decision tree classifier. Then we divided the columns into two groups X and Y similar to the groups made in multiclass logistic regression. Then we split the data, for testing and training. Then we used the “decision tree classifier” and fit the classifier on the training data and computed the accuracy score on y\_test, y\_predict. The basic flow for a Decision Tree is as follows:

First, select the best attribute in the dataset as the root node of the tree. This attribute should have the highest information gain (IG) or the lowest Gini index, depending on the criterion chosen. Create a branch for each possible value of the root node attribute. Recursively apply steps 1 and 2 to each branch, using only the instances that are associated with that branch (i.e., prune the dataset). Stop the recursion when all attributes have been used as nodes, or when adding another node does not improve the classification accuracy. Here are the formulas for Gini Index, a commonly used criteria for selecting the root node attribute:

Gini Index:

*G(T) = 1 - ∑ p\_i^2*, where p\_i is the proportion of instances in T that belong to class i

*G(T|X) = ∑ p\_j \* G(T\_j)*, where T\_j is the subset of T where attribute X has value j and p\_j is the proportion of instances in T that have value j for attribute X

**3.Random Forest:** First we imported the random forest classifier. Then we divided the columns into two groups X and Y similar to the groups made in multiclass logistic regression. Then we split the data for testing and training. Then we used the “random forest classifier” and fit the classifier on the training data and computed the accuracy score on y\_test, y\_predict. Note that Random Forest is an “ensemble learning method” that combines “multiple decision trees”, so it doesn't have a single formula like a decision tree. Instead, it follows the above algorithm to build multiple decision trees.

**4.K-nearest neighbours:** First we imported the k-neighbours classifier. Then we divided the columns into two groups X and Y similar to the groups made in multiclass logistic regression. Then we split the data. Then we created a KNN classifier with 5 nearest neighbours and fit the classifier on the training data and computed the accuracy score on y\_test, y\_predict. KNN uses distance metrics to check for similar neighbours.

Euclidean Distance:

d(x,y) = sqrt((x\_1 - y\_1)^2 + (x\_2 - y\_2)^2 + ... + (x\_n - y\_n)^2)

here, x and y are the two instances being compared, and n is the no. of attributes/features. x\_i and y\_i are the values of the i-th attribute for x and y, respectively.

Note that other distance metrics, like Manhattan distance or cosine similarity, can also be used in KNN.

* **Evaluation metrics used:**

**Accuracy:**

Accuracy is a measure of how often the model correctly predicts the outcome.

*Accuracy = (TP + TN) / (TP + TN + FP + FN)*

**Precision:**

Precision is a measure of how often the model correctly predicts positive cases.

*Precision = TP / (TP + FP)*

**Recall:**

Recall is a measure of how often the model correctly identifies positive cases.

*Recall = TP / (TP + FN)*

**F1-score:**

The F1-score is a harmonic mean of precision and recall.

*F1-score = 2 \* (Precision \* Recall) / (Precision + Recall)*

**Confusion Matrix:**

A confusion matrix is a table that summarises the model's predictions and their actual outcomes and contains all four: TP, FP, TN, FN

Here, TP is the number of true positives, FP is the number of false positives, FN is the number of false negatives, and TN is the number of true negatives.

**RESULTS:**

***FOR DATASET 1:***

**Multiclass Logistic regression**

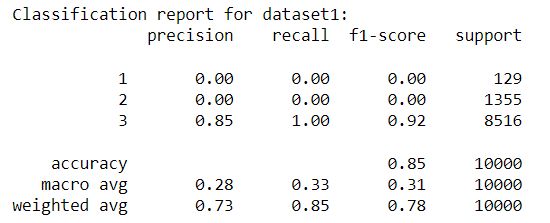


Fig 9 (Classification report for multiclass regression)

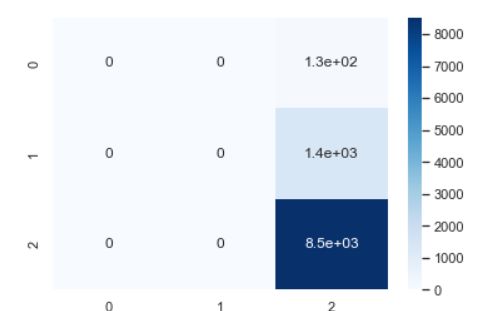


Fig 10 (Confusion matrix for multiclass logistic regression)

**Decision tree**

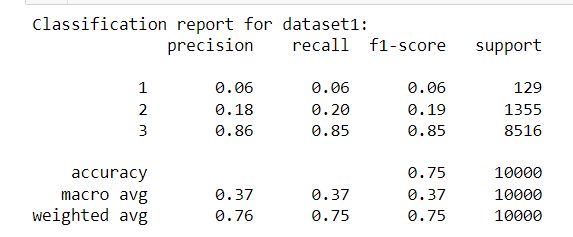
****

Fig 11 (Classification matrix for Decision Tree)

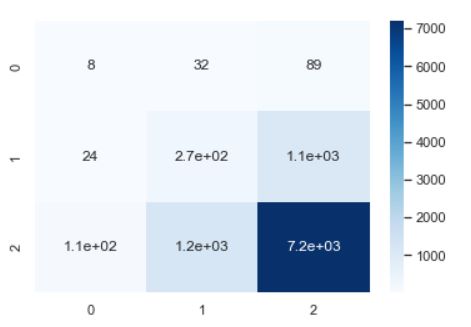
****

Fig 12(Confusion matrix for decision matrix)

**Random Forest**

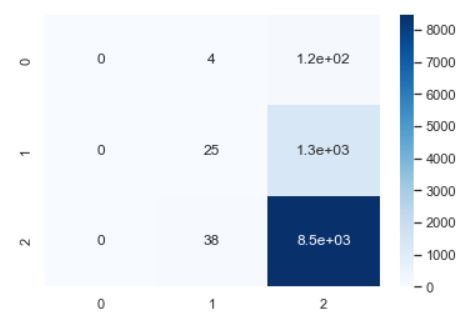
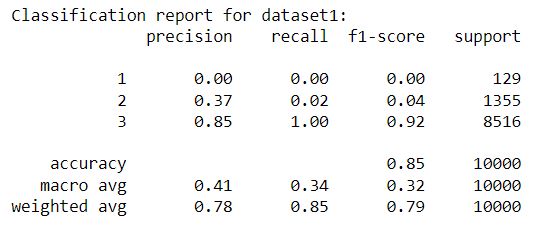
Fig 13 ( Classification report for random forest)

Fig 14 ( Confusion matrix for random forest)

**K-nearest neighbours**

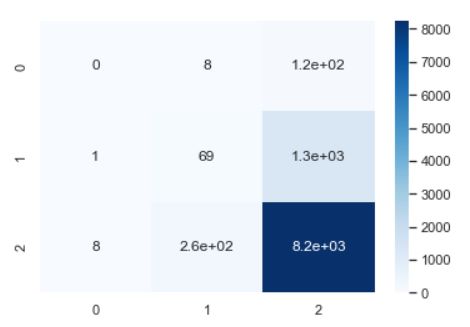
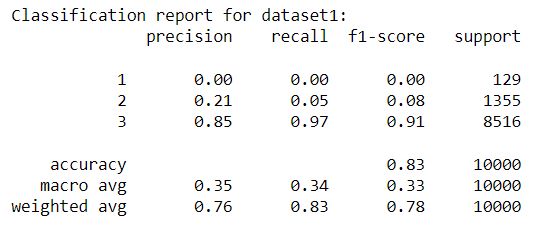
Fig 15(Classification Report for K-Nearest Neighbours)

Fig 16 (Confusion matrix for K-Nearest Neighbours)

**Voting classifier:**

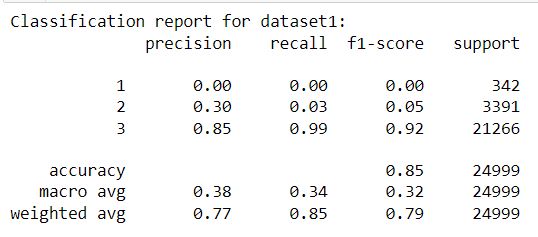
****

Fig 17(Classification Report for dataset 1)

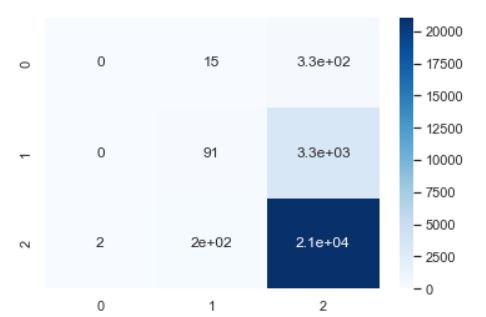
****

Fig 18(Confusion matrix for dataset 1)

***FOR DATASET 2:***

**Multiclass Logistic regression**

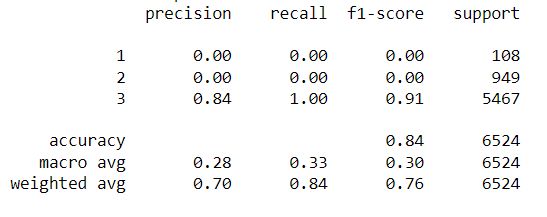
****

Fig 19(Classification Report for multiclass logistic regression)

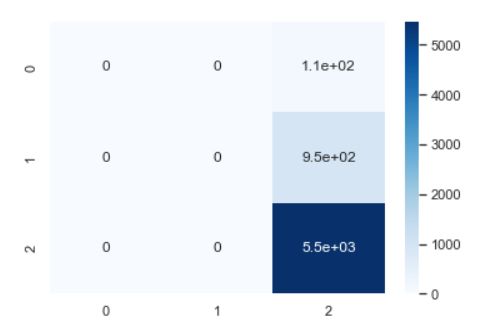
****

Fig 20(confusion matrix for multiclass logistic regression)

**Decision Tree**

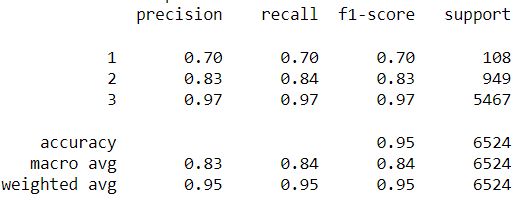
****

Fig 21(classification report for decision tree)

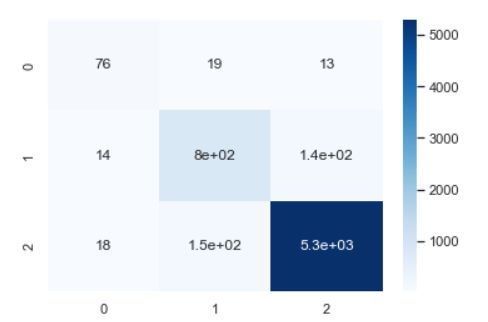
****

Fig 22(confusion matrix for decision tree)

**Random Forest**

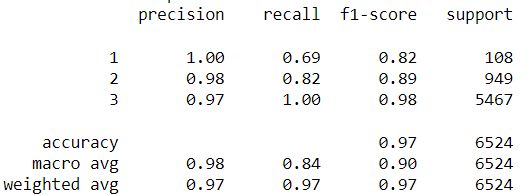
****

Fig 23(classification report for random forest)

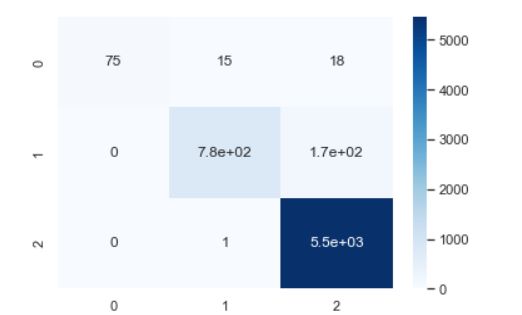
****

Fig 24(confusion matrix for random forest)

**K-nearest neighbours**

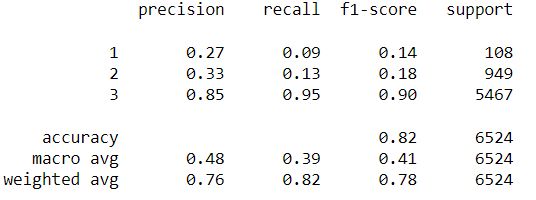
****

Fig 25(classification report for k-nearest neighbours)

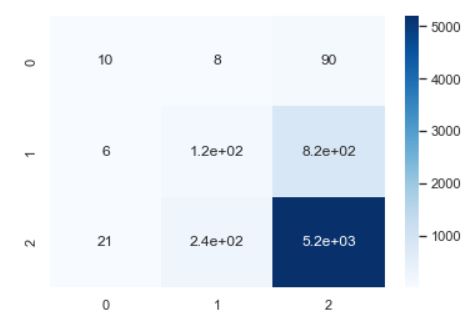
****

Fig 26(confusion matrix for k-nearest neighbours)

**Voting classifier:**

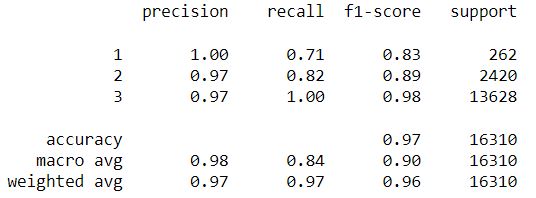
****

Fig 27(classification report for data set 2)

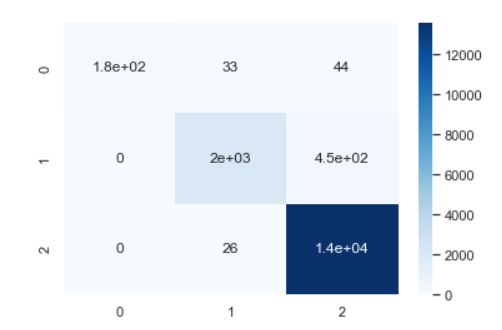
****

Fig 28(confusion matrix for dataset 2)

***EVALUATING THE RESULTS:***

***Dataset-1:***

Table 1 (Results for dataset -1 )

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***MODEL*** | ***ACCURACY*** | ***PRECISION*** | ***RECALL*** | ***F1-SCORE*** |
| Multinomial Logistic Regression | 0.851 | 0.73 | 0.85 | 0.78 |
| Decision Tree | 0.7453 | 0.76 | 0.75 | 0.75 |
| Random Forest | 0.8476 | 0.78 | 0.85 | 0.79 |
| KNN | 0.8303 | 0.76 | 0.83 | 0.78 |
| Voting classifier | 0.8462 | 0.77 | 0.85 | 0.79 |

***Dataset-2:***

Table 2 (Results for dataset - 2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***MODEL*** | ***ACCURACY*** | ***PRECISION*** | ***RECALL*** | ***F1-SCORE*** |
| Multinomial Logistic Regression | 0.833 | 0.70 | 0.84 | 0.76 |
| Decision Tree | 0.9432 | 0.95 | 0.95 | 0.95 |
| Random Forest | 0.9694 | 0.97 | 0.97 | 0.97 |
| KNN | 0.8153 | 0.76 | 0.82 | 0.78 |
| Voting classifier | 0.9663 | 0.97 | 0.97 | 0.96 |

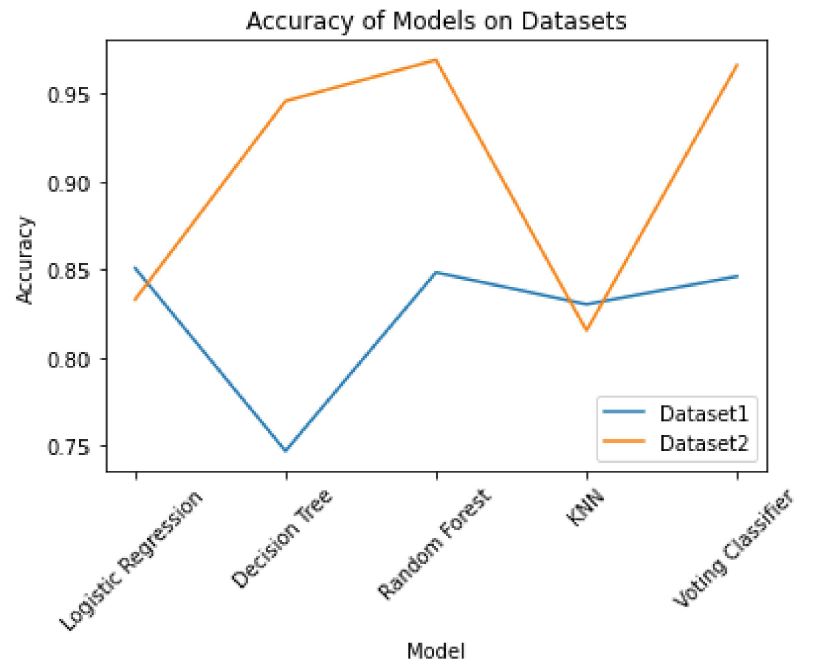
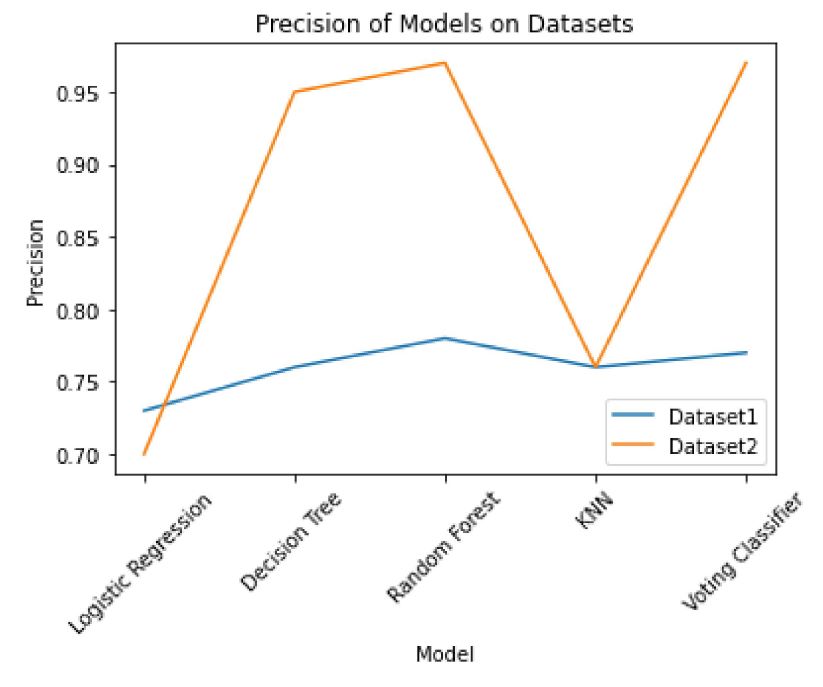
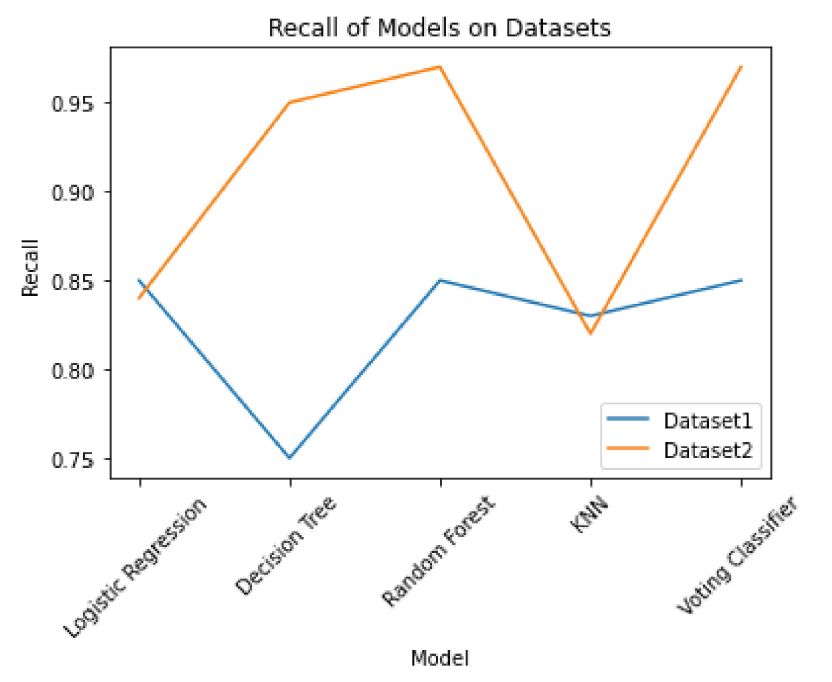
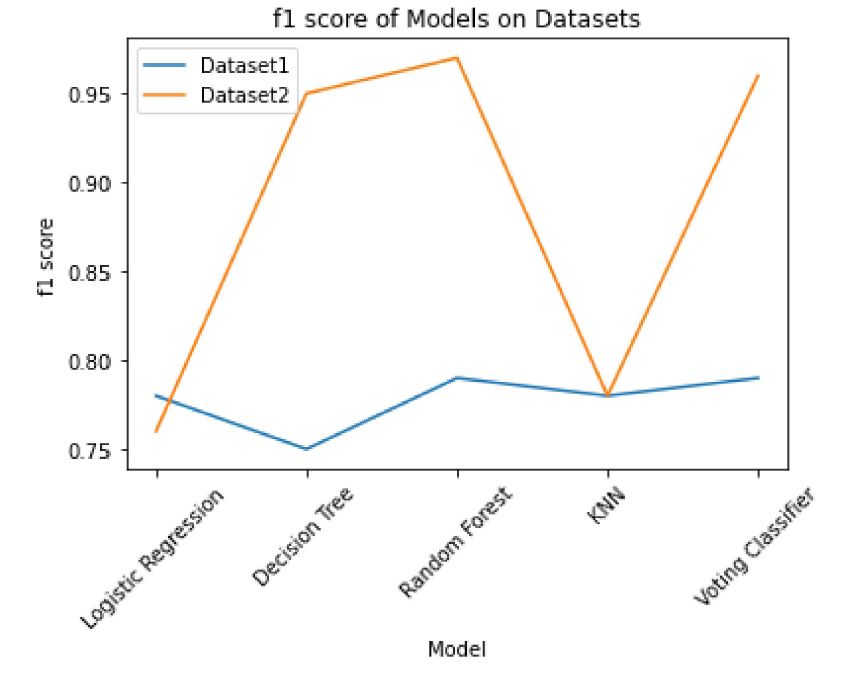


Fig 29(Accuracy of models on datasets) Fig 30(Precision of models on datasets)

Fig 31(Recall of models on datasets) Fig 32(f1 score of models on datasets)

***CONCLUSION:***

Road accidents are increasing day by day even with the increasing regulations and precautionary measures. So, trying to understand the causal relationship between various conditions and accident severity will greatly help in reducing both the frequency and severity of road accidents. In our work, we have used four machine learning models and used one voting classifier to predict the accident severity using various conditions. We have considered weather conditions, light conditions and many more. Other conditions like pedestrians and carriageways might also affect the accident severity. The results of our analysis, as displayed above, show that for dataset 1, multinomial logistic regression and the voting classifier perform the best. But for dataset 2, the voting classifier clearly outperforms the other models, although it is observed that random forest and decision tree also perform relatively well. The results of our analysis might greatly help in shaping public road safety laws and also help reduce the severity of accidents.

***References:***

[1]Brijs, Stiers & Hermans, Elke & Brijs, Tom & Stiers, Tim & Offermans, Col. (2007). The Impact of Weather Conditions on Road Safety Investigated on an Hourly Basis.

[2]Fanny Malin, Ilkka Norros, Satu Innamaa, Accident risk of road and weather conditions on different road types, Accident Analysis & Prevention, Volume 122, 2019, Pages 181-188, ISSN 0001-4575,Available at:https://doi.org/10.1016/j.aap.2018.10.014.

[3]Karlaftis, M.G., & Yannis, G. (2010). Weather Effects on Daily Traffic Accidents and Fatalities: Time Series Count Data Approach.

[4]Frits Bijleveld, Tony Churchill, The influence of weather conditions on road safety, R-2009-9

[5]R. Pokhrel, "Impact of Weather on Severity of Crashes: A Statistical Approach." Order No. 28316991, Southern Illinois University at Edwardsville, United States -- Illinois, 2021.

[6] D. Santos, J. Saias, P. Quaresma, and V. B. Nogueira, “Machine Learning Approaches to Traffic

Accident Analysis and Hotspot Prediction,” Computers, vol. 10, no. 12, p. 157, Nov. 2021, doi: 10.3390/computers10120157.

[7] M. F. Labib, A. S. Rifat, M. M. Hossain, A. K. Das and F. Nawrine, "Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh," 2019 7th International Conference on Smart Computing & Communications (ICSCC), Sarawak, Malaysia, 2019, pp. 1-5, doi: 10.1109/ICSCC.2019.8843640.0

[8] Al-Mistarehi BW, Alomari AH, Imam R and Mashaqba M (2022) Using Machine Learning Models to Forecast Severity Level of Traffic Crashes by R Studio and ArcGIS. Front. Built Environ. 8:860805. doi: 10.3389/fbuil.2022.860805

[9] Becker, N., Rust, H.W. & Ulbrich, U. Weather impacts on various types of road crashes: a quantitative analysis using generalized additive models. Eur. Transp. Res. Rev. 14, 37 (2022). https://doi.org/10.1186/s12544-022-00561-2

[10]Khaghani, Farnaz. A Deep Learning Approach to Predict Accident Occurrence Based on Traffic Dynamics. Diss. Virginia Tech, 2020.

[11] Mondal, A.R., Bhuiyan, M.A.E. & Yang, F. Advancement of weather-related crash prediction model using nonparametric machine learning algorithms. SN Appl. Sci. 2, 1372 (2020). https://doi.org/10.1007/s42452-020-03196-x

[12] Dr. A. Pranayanath Reddy, Dr. Shekhar, Dr. Suresh Babu(2019). Machine Learning Approach to Predict the Accident Risk during Foggy Weather Conditions. Alliance International Conference on Artificial Intelligence and Machine Learning(AICAAM), April. 2019

[13]S. Pourroostaei Ardakani et al., “Road Car Accident Prediction Using a Machine-Learning-EnabledDataAnalysis,”Sustainabilieedings/Papers/Machine-Learning-Approach-to-Predict.pdfty, vol. 15, no. 7, p. 5939, Mar. 2023, doi: 10.3390/su15075939.

[14 ]Zhuoning Yuan, Xun Zhou, and Tianbao Yang. 2018. Hetero-ConvLSTM: A Deep Learning the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18). Association for Computing Machinery, New York, NY, USA, 984–992. https://doi.org/10.1145/3219819.3219922

[15] A. J. Ghandour, H. Hammoud, and S. Al-Hajj, “Analyzing Factors Associated with Fatal Road Crashes: A Machine Learning Approach,” International Journal of Environmental Research and Public Health, vol. 17, no. 11, p. 4111, Jun. 2020, doi: 10.3390/ijerph17114111.

[16]Bener, Abdulbari & Crundall, David. (2005). Road traffic accidents in the United Arab Emirates compared to Western countries. Advances in Transportation Studies an international Journal Section A. 6.

[17]Md. M. Islam, M. Alharthi, and Md. M. Alam, “The Impacts of Climate Change on Road Traffic Accidents in Saudi Arabia,” Climate, vol. 7, no. 9, p. 103, Aug. 2019, doi: 10.3390/cli7090103.

[18]Singh, Sanjay. (2017). Road Traffic Accidents in India: Issues and Challenges. Transportation Research Procedia. 25. 4712-4723. 10.1016/j.trpro.2017.05.484.

[19]A. Aleksić, M. Ranđelović, and D. Ranđelović, “Using Machine Learning in Predicting the Impact of Meteorological Parameters on Traffic Incidents,” Mathematics, vol. 11, no. 2, p. 479, Jan. 2023, doi: 10.3390/math11020479. [Online]. Available: http://dx.doi.org/10.3390/math11020479

[20]Y. Zou, Y. Zhang, and K. Cheng, “Exploring the Impact of Climate and Extreme Weather on Fatal Traffic Accidents,” Sustainability, vol. 13, no. 1, p. 390, Jan. 2021, doi: 10.3390/su13010390. [Online]. Available: http://dx.doi.org/10.3390/su13010390

[21]Ouimet MC, Pradhan AK, Brooks-Russell A, Ehsani JP, Berbiche D, Simons-Morton BG. Young Drivers and Their Passengers: A Systematic Review of Epidemiological Studies on Crash Risk. J Adolesc Health. 2015 Jul;57(1 Suppl):S24-35.e6. doi: 10.1016/j.jadohealth.2015.03.010. PMID: 26112735; PMCID: PMC4483197.

[22]Chen, Zhijian & Ge, Yinen & Wang, Kai & Song, Jiahua. (2019). Evaluating safety performance of highway alignment utilising fault tree analysis and energy method. Advances in Mechanical Engineering. 11. 168781401985426. 10.1177/1687814019854268.

[23]N. Casado-Sanz, B. Guirao, and M. Attard, “Analysis of the Risk Factors Affecting the Severity of Traffic Accidents on Spanish Crosstown Roads: The Driver’s Perspective,” Sustainability, vol. 12, no. 6, p. 2237, Mar. 2020, doi: 10.3390/su12062237. [Online]. Available: http://dx.doi.org/10.3390/su12062237

[24]Parsa, A.B., Chauhan, R.S., Taghipour, H., Derrible, S., & Mohammadian, A.(. (2019). Applying Deep Learning to Detect Traffic Accidents in Real Time Using Spatiotemporal Sequential Data. ArXiv, abs/1912.06991.

[25]D. Santos, J. Saias, P. Quaresma, and V. B. Nogueira, “Machine Learning Approaches to Traffic Accident Analysis and Hotspot Prediction,” Computers, vol. 10, no. 12, p. 157, Nov. 2021, doi: 10.3390/computers10120157. [Online]. Available: http://dx.doi.org/10.3390/computers10120157

30