ROAD ACCIDENT ANALYSIS AND CLASSIFICATION USING MACHINE LEARNING

1. **Introduction**

Road accidents are a major cause of injury and death worldwide, according to the World Health Organisation (WHO), 50 million more people suffer injuries or disabilities as a result of road traffic accidents each year, killing over 1.35 million people. Given the gravity of this issue, applying machine learning (ML) to enhance our capacity to anticipate and prevent traffic accidents is gaining popularity.

By analysing and classifying road accidents into different categories, we can gain We will learn more about the causes and contributing elements of accidents and develop prevention strategies by analysing and categorising road accidents into distinct groups (Najaf P. et al. 2017).

Several studies have focused on using ML techniques to classify road accidents and predict their severity for example, a study conducted by Singh and colleagues (2021) used a dataset of road accidents in India to develop a classification model that could predict accident severity with high accuracy. This study builds upon these previous studies by developing a prediction framework. It aims to use Machine learning techniques to analyse road accident data and classify the road traffic accidents into different severity classes. There are three objectives for this project as follows:

* To gather and pre-process real life accident data datasets
* Use machine learning techniques to build multiple classification models
* Evaluation and analysis of models based on its performance

1. **Literature review**

The use of traffic accident data has changed recently as a result of the increase in traffic accidents. Until recently, basic statistical analysis was the main use for this type of data, which offered only a few insights into patterns and statistics (Jamal A et al., 2019). However, it is now widely acknowledged that traffic accidents are an important developmental and public health issue that demands the attention of researchers, civic organisations, automakers, governments, and business groups everywhere. The main emphasis is on figuring out the key causes of these incidents and creating measures to deal with them (Kumeda B. et al, 2019).

Understanding the various aspects, such as driver behaviour, road conditions, and weather conditions that contribute to varying levels of injury severity, can be facilitated by using data analysis and machine learning approaches to analyse traffic accident data sets. Decision-makers can use this information to create more effective traffic safety control policies.

Machine Learning (ML), a subset of artificial intelligence, has gained popularity for its capacity to create clever predictive algorithms for a variety of applications. In high-dimensional and multivariate settings, such as industrial environments, ML approaches can manage complicated and dynamic data and reveal hidden links within the data (Wuest et al., 2016). As a result, ML offers a strong strategy for a variety of applications. The performance of these applications depends on picking the right ML technique (Xiao wen. et al., 2021).

An overview of various machine learning methods that have been applied to the analysis and prevention of traffic accidents is given in (Ijaz, M. et al. 2021). They go over the benefits and drawbacks of various methodologies, such as decision trees and random forests. They also talk about the difficulties in gathering and pre-processing accident data.

In order to investigate the viability of using support vector machine (SVM) models to assess accident injury severity, (Li et al., 2012) conducted a study. The authors used the same dataset to collect crash data from 326 motorway diverging locations and create SVM and ordered probit (OP) models to forecast the severity of injuries related to specific crashes. When the SVM and OP models' abilities to predict collision injury severity were evaluated, the researchers found that the SVM model performed better than the OP model.

(Kumeda et al. 2019) did a study with the aim of classifying data on road traffic accidents using machine learning techniques. The authors used a variety of machine learning approaches to categorise the accidents based on the severity of the injuries experienced, using data from year 2016 traffic accident records in the United Kingdom. The Fuzzy-FARCHD, Random Forest, Hierarchal LVQ, RBF Network, Multilayer Perceptron, and Naive Bayes algorithms, according to their findings, showed a high level of classification accuracy.

Road Accident Severity Prediction — A Comparative Analysis of Machine Learning Algorithms (Malik, S. et al., 2021) examines the usefulness of various machine learning algorithms for predicting road traffic accidents. The authors compared the effectiveness of six models (Bagging, Decision Tree, AdaBoost, Naive Bayes, Random Forest, and Logistic Regression) from data gotten from the Department of Transportation They discovered that, across all performance parameters, random forest, decision tree, bagging, and AdaBoost beat the other techniques.

Studies that only predict accidents based on one or two characteristics, which may not adequately reflect real-world circumstances, are among the gaps that have been discovered through an analysis of the existing literature. Additionally, a sizable proportion of research ignore the problem of class inequality. The current study aims to close these research gaps in order to address them.

1. **Methodology**

The goal of this study is to use machine learning techniques to analyse data on traffic incidents and classify them according to their severity. The stages of the methodology are as follows:

1. Acquiring traffic accident data of the police departments of Addis Ababa Sub-city.
2. Cleaning and pre-processing the data to create models.
3. Feature Selection to determine the model's most crucial features and data balance
4. Algorithms for machine learning.
   1. **Dataset details**

In this study, the Road Traffic Accident Dataset of Addis Ababa City it used. Based on the documentation, The data set was prepared from manual records of road traffic accident of the year 2017-20 and it’s a public based dataset downloadable from the National Academic Research and Collaborations Information System (NARCIS) databank (https://www.narcis.nl/dataset/RecordID/oai:easy.dans.knaw.nl:easy-dataset:191591/Language/en). The dataset contains 12316 instances of the accident and 32 characteristics. The following table displays the dataset's characteristics.

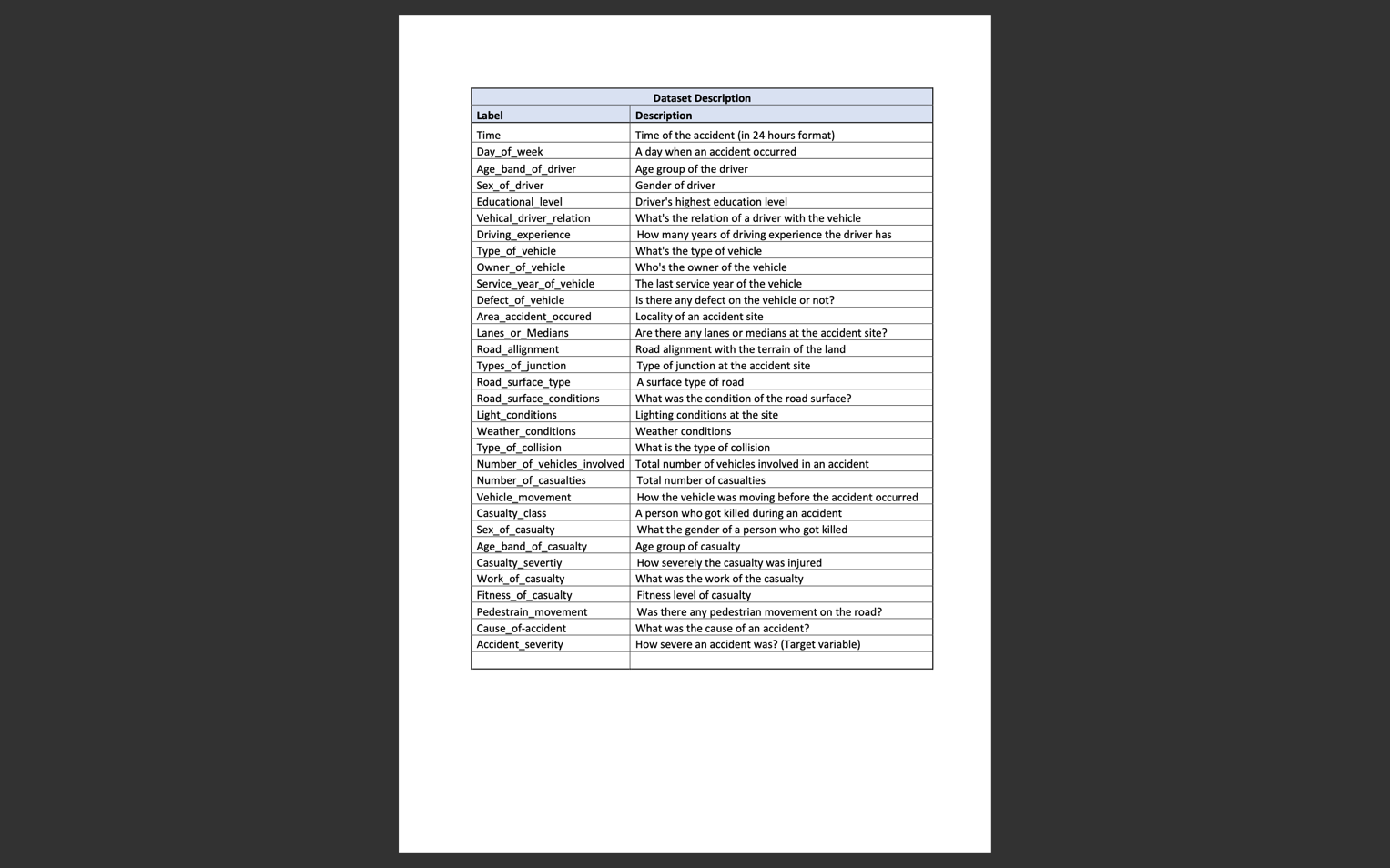


Table 1: Brief description of Dataset

* 1. **Pre-processing**

Data pre-processing and cleaning are essential processes in transforming the raw data into a useable and efficient format for modelling, and they play a significant role in machine learning modelling. Raw data cannot be converted into a format appropriate for machine learning modelling without proper pre-processing.

The dataset on traffic accidents is available in CSV format. We dealt with the problem of missing values after downloading the dataset by imputed the missing values with the word "unknown." This strategy was chosen since the missing value features lack numerical values, and we don't want to throw away any data. Furthermore, the feature value is then converted from string to numeric form. In data analysis and machine learning, this is a typical task. Categorical characteristics are converted via one-hot encoding, and then target encoding using label encoding

* 1. **Feature selection & Data balancing**

Chi-squared test is a popular statistical method used to determine whether there is a significant association between two categorical variables. It is used for feature selection in this machine learning problem to produce top 50 features from the encoded dataset.chisquare is mathematically represented as:

chi2 = sum((O\_i - E\_i)^2 / E\_i)

where:

* O\_i is the observed frequency of the i-th category in a contingency table
* E\_i is the expected frequency of the i-th category under the assumption of independence between the two categorical variables
* The sum is taken over all categories

The imbalanced dataset is upsampled using the Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTENC). SMOTENC generates synthetic samples of the minority class by interpolating between neighboring samples there by balancing the datset

* 1. **Machine Learning algorithm**

After data pre-processing, feature selection and data balancing, the data was split into 80% training and 20% testing sets. In this research work, we have used four different classification algorithms to train and compare results.

* MLPClassifier: multilayer perceptron (MLP) based on the feedforward artificial neural network model to classify data into multiple classes. We passed in two hidden layers each with 50neurons, maximum iterations of 1000 with a random state set to 42.
* Random Forest: based on ensemble learning method that combines multiple decision trees to improve the accuracy and robustness of the model. Parameters passed includes 800decision trees, max depth of 20 with a random state set to 42.
* Support Vector Classifier: based on the concept of finding the optimal hyperplane that separates data into different classes. the kernel function is set to 'linear', with a random state set to 42.
* KNeighbors: finding the k closest data points neighbours to a new input data point, and then assigning a label

1. **Results and Discussion**

The table below shows a summary of the results

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy** | **F-1 Score** |
| Random Forest Classifier | 0.88446151 | 0.88 |
| Support Vector Classifier | 0.68346936 | 0.68 |
| Multi-layer Perceptron classifier | 0.8668587 | 0.87 |
| KNeighborsClassifier | 0.80588894 | 0.79 |

Table 2: Summary of the model performance

The models are evaluated using the F-1score and as seen in the table above its accuracy and scores, the Random Forest Classifier performed the best at 88%.

The study does, however, suggest that more data and hyperparameter adjustment could be used to further enhance model performance.

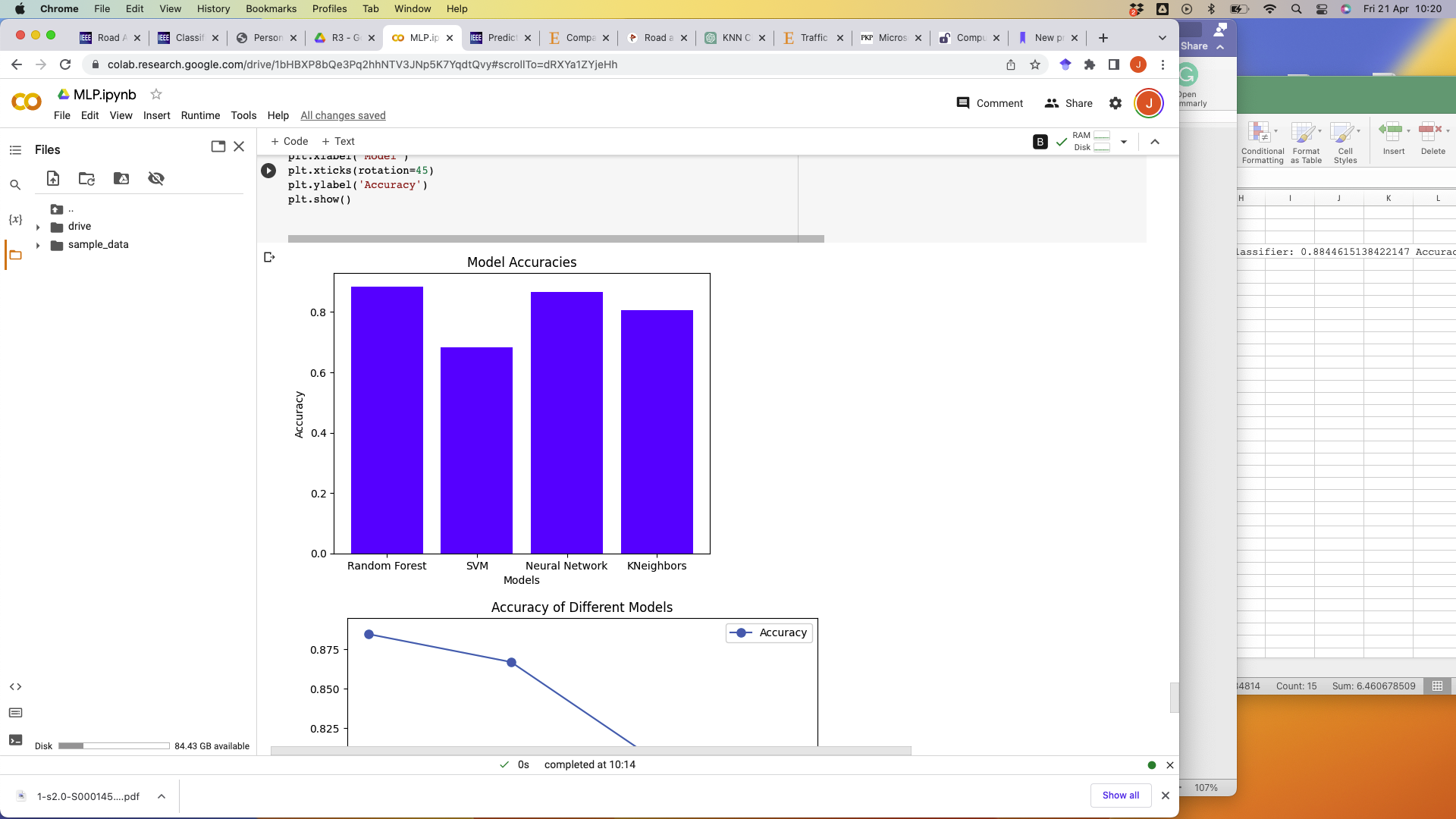


Fig-1: Plot of model’s accuracy- bar plot

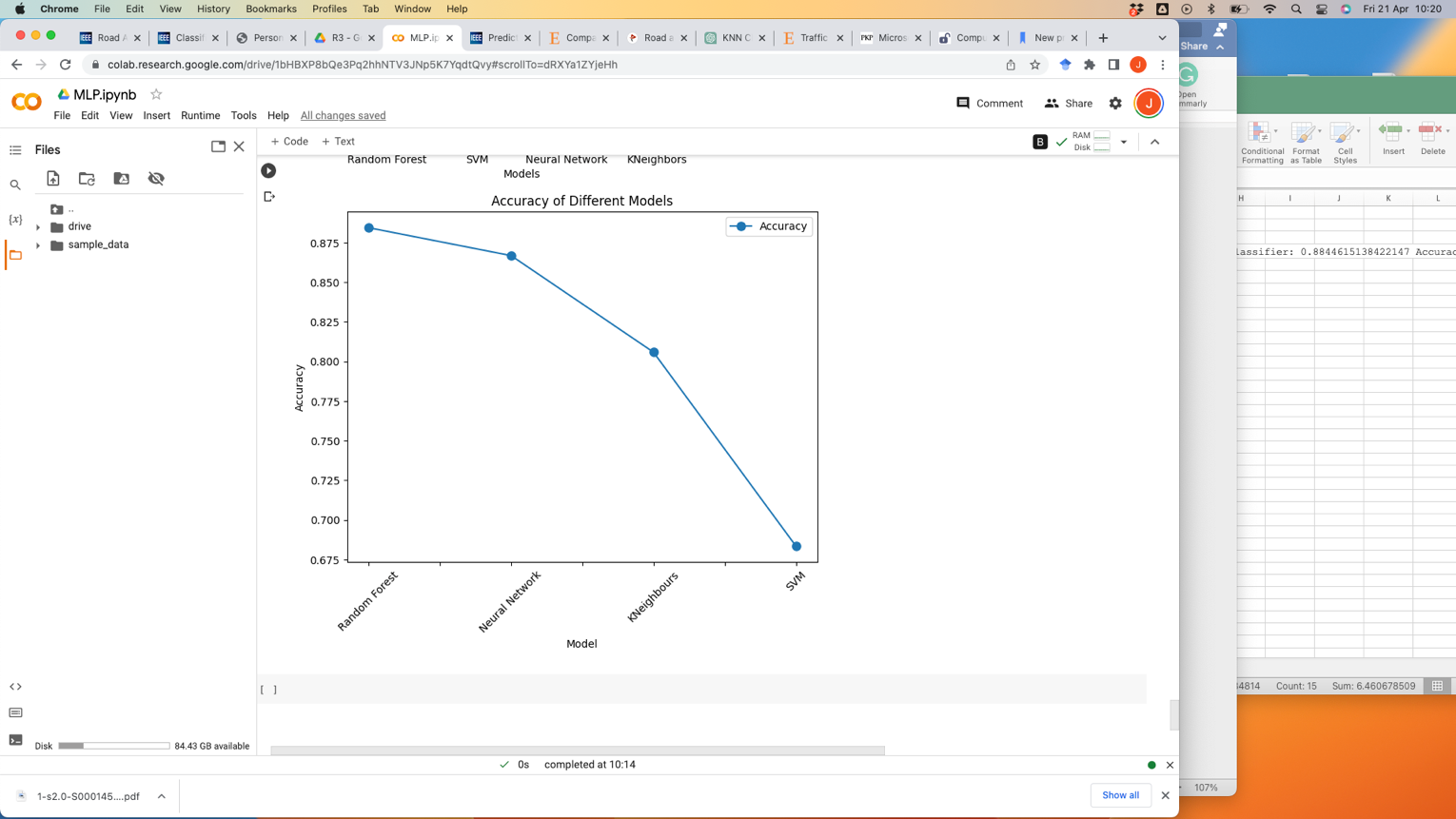


Fig-1: Plot of model’s accuracy- line plot

1. **Conclusion**

Using the Addis Ababa City Road Traffic Accident Dataset, a prediction framework was developed in this study to evaluate the efficacy of four machine learning (ML) algorithms for forecasting accident severity. The SMOTENC approach was used to remedy the dataset's class imbalance issue. The results showed that the Random Forest algorithm performed better than the other ML algorithms examined, with an F1 score of 88% and the best accuracy.

1. **Reference**

Anon. 2023. NARCIS > About NARCIS. [online] www.narcis.nl. Available at: <https://www.narcis.nl/about/Language/en> [Accessed 21 April 2023].

Chong, M.M., Abraham, A. and Marcin Paprzycki, 2005. Traffic Accident Analysis Using Machine Learning Paradigms. Informatica (lithuanian Academy of Sciences), 29(1), pp.89–98.

Ijaz, M., lan, L., Zahid, M. and Jamal, A., 2021. A comparative study of machine learning classifiers for injury severity prediction of crashes involving three-wheeled motorized rickshaw. Accident Analysis & Prevention, 154, p.106094. https://doi.org/10.1016/j.aap.2021.106094.

Jamal, A., Rahman, M.T., Al-Ahmadi, H.M. and Mansoor, U., 2019. The Dilemma of Road Safety in the Eastern Province of Saudi Arabia: Consequences and Prevention Strategies. International Journal of Environmental Research and Public Health, 17(1), p.157. https://doi.org/10.3390/ijerph17010157.

Kumeda, B., Zhang, F., Zhou, F., Hussain, S., Almasri, A. and Assefa, M., 2019. Classification of Road Traffic Accident Data Using Machine Learning Algorithms. [online] IEEE Xplore. https://doi.org/10.1109/ICCSN.2019.8905362.

Li, Z., Liu, P., Wang, W. and Xu, C., 2012. Using support vector machine models for crash injury severity analysis. Accident; Analysis and Prevention, [online] 45, pp.478–486. <https://doi.org/10.1016/j.aap.2011.08.016>.

openai, 2023. ChatGPT. [online] chat.openai.com. Available at: <https://chat.openai.com/>.

Malik, S., El Sayed, H., Khan, M.A. and Khan, M.J., 2021. Road Accident Severity Prediction — A Comparative Analysis of Machine Learning Algorithms. 2021 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT). https://doi.org/10.1109/gcaiot53516.2021.9693055.

Najaf, P., Duddu, V.R. and Pulugurtha, S.S., 2017. Predictability and interpretability of hybrid link-level crash frequency models for urban arterials compared to cluster-based and general negative binomial regression models. International Journal of Injury Control and Safety Promotion, 25(1), pp.3–13. https://doi.org/10.1080/17457300.2017.1285789.

Roh, J. W., Bessler, D. A., & Gilbert, R. F. (1999). Traffic fatalities, Peltzman’s model, and directed graphs. Accident Analysis & Prevention, 31(1-2), 55-61.

Santos, D., Saias, J., Quaresma, P. and Nogueira, V.B., 2021. Machine Learning Approaches to Traffic Accident Analysis and Hotspot Prediction. Computers, 10(12), p.157. https://doi.org/10.3390/computers10120157.

Wen, X., Xie, Y., Jiang, L., Pu, Z. and Ge, T., 2021. Applications of machine learning methods in traffic crash severity modelling: current status and future directions. Transport Reviews, pp.1–25. https://doi.org/10.1080/01441647.2021.1954108.

Wuest, T., Weimer, D., Irgens, C. and Thoben, K.-D., 2016. Machine learning in manufacturing: advantages, challenges, and applications. Production & Manufacturing Research, [online] 4(1), pp.23–45. https://doi.org/10.1080/21693277.2016.1192517.