**DMPA LAB PROJECT REPORT**

**Advanced Programming Lab (ICT 3167)**

**Project Title:** Accident Severity And Driver Reliability

Aniket Bhattacharya - 210953156

Pranav Chavare - 210953138

Karan Chawla - 210953096

## **Introduction:**

Road traffic accidents are a big concern in cities, especially in metropolitan areas. It is important to understand the factors that influence the severity of accidents and subsequently improve road safety. We have used a dataset that has been sourced from Addis-Ababa police departments, made available on Kaggle. The dataset comprises comprehensive features such as driver information like education and other factors such as weather and the different collision types.

The purpose of this project is to use different machine learning algorithms to predict driver reliability and severity of accidents. Furthermore, we have a comparison between the algorithms to conclude which algorithm is best for the domain of road accidents. We have used the Random Forest, XGBoost and Extra Trees algorithms. We have derived insights after running these algorithms and contribute to improve safety on the roads.

We have used SMOTE and SMOTETomek to deal with the imbalance in the dataset. We have made an effort to promote data based decision making in traffic management and safety efforts.

## **Literature Survey:**

[1] This research focuses on the main national highways in the Krishna area and tackles the worldwide problem of traffic accidents. Finding trends in accident data while taking into account variables like weather, road conditions, and driver emotions is the aim. By utilising machine learning methods such as association rule mining and clustering, the research finds hidden links in the data. Although there are several drawbacks, such as missing attribute data, the study emphasises how useful techniques like K-medoids clustering are. The results highlight how a mix of driver mistake, vehicle malfunctions, and infrastructural problems frequently causes accidents. In order to improve pattern identification, the paper recommends future research into features and clusters as well as the possible application of deep learning techniques. All things considered, this study offers insightful information to guide preventative measures.

[2] In order to address the increasing number of accidents and fatalities, especially those caused by driver drowsiness, this research focuses on the concerning problem of road accidents in India. Using machine learning and deep learning classifiers, such as Support Vector Machine, k-nearest neighbours, Decision Trees, and the Autoencoder model, the method divides 60-minute EEG data into distinct time zones. The Autoencoder classifies driver weariness with an astounding accuracy rate of 99.7%. Although the study effectively tackles the crucial issue of road safety, it recognises several limits, including the restricted interpretability of deep learning results and their lack of real-time application. The results highlight the potential for early driver tiredness identification to dramatically improve road safety and avoid accidents, pointing to possible areas for further study to increase practicality and reliability.

[3] This study tackles the pressing issue of detecting driver drowsiness early on to improve road safety. It focuses on analyzing physiological signals, behavioral traits, and driving performance to classify drivers as either alert or slightly drowsy, particularly within short 10-second intervals. Using machine learning algorithms, including Random Forest, the study achieves an impressive accuracy of 81.4% in distinguishing between these states. While emphasizing the importance of early detection for preventing accidents, the study acknowledges limitations such as a focus on laboratory-based simulations and a lack of real-time application discussion. Despite these, the findings underscore the potential of early drowsiness detection, suggesting avenues for future research to enhance accuracy and real-world applicability.

[4] This study addresses the escalating risks of road accidents due to the growing global population and increasing vehicle numbers, emphasizing the need for early accident information to prevent incidents and save lives. The proposed solution introduces a comprehensive system incorporating data collection on user behavior, GPS tracking for identifying danger zones, and the application of the Random Forest algorithm to swiftly classify events as normal or accidents, providing early warnings to users in potential "black spot" danger areas. Despite achieving a commendable 93.45% classification accuracy with the BF-GD algorithm, the study is limited by a lack of real-time implementation discussion and technical system requirements. While focusing on algorithm performance, the study highlights the proactive nature of the proposed system in alerting users ahead of accidents, potentially averting them and enhancing road safety. Future work should prioritize real-world implementation, effectiveness evaluation, and addressing technical requirements to ensure practical applicability. This proactive approach holds promise for significantly improving road safety outcomes and saving lives.

[5] This research paper explores the effectiveness of road tests and neuropsychological assessments in predicting older drivers' responsibility for collisions, emphasizing specific accident types for improved prediction accuracy. Cognitive issues, notably attention and perception, are identified as significant contributors to elderly driver accidents, with distinct neuropsychological measurements correlating with different accident categories. The study underscores the importance of comprehensive accident data for validating on-road driving proficiency tests and acknowledges the significance of self-reporting in accident reporting. The evaluation process involves participants contacting the CARA department, completing questionnaires, and undergoing morning neuropsychological tests followed by an afternoon road test. The use of Chronbach's alpha reveals high inter-item consistency, while discriminant analyses show that neuropsychological tests outperform TRIP factors in predicting accidents. Despite the challenges in predicting car accidents, the study suggests that specifying accident types moderately enhances predictability. The findings advocate for a comprehensive assessment of driving ability and strategic compensation training for older adults, acknowledging the exploratory nature of the study and calling for prospective research due to sample size limitations and the complexity of accident prediction.

[6] This paper introduces a hybrid strategy that combines Random Forest classification with K-means clustering to solve the global concern of road accident severity, especially in developing nations. The process entails adding a new feature to the training set and first using K-means clustering to uncover hidden patterns in data on traffic accidents. Maximum distance is used to calculate the ideal number of clusters, and the resulting Random Forest model outperforms previous classification methods in predicting the severity of accidents with an astounding accuracy of 99.86%. Key factors determining severity are identified through the interpretation of the model findings, including driving experience, day of the week, lighting conditions, driver age, and year of vehicle service. The study emphasises how well this method works to forecast the severity of traffic accidents, providing useful information for insurance companies and road transport companies to use when creating safety plans. The advantages of the hybrid model are its scalability, simplicity, and resilience to overfitting; when combined, these factors allow the combined technique to perform better than typical machine learning algorithms. The study's conclusion emphasises the possibility of more advancement with more

datasets to improve model performance.

[7] This research looks at traffic accidents and emphasises a thorough methodology that takes into account both external (environmental) and internal (biological) factors inside the driving system, such as user interactions, road infrastructure, and vehicle components. While most accident categorization schemes concentrate only on immediate mistakes, this study combines qualitative and quantitative analysis to offer a comprehensive picture of human variables that affect accident rates. The technique enables customised classification, identifying inattention as the most frequent contributing factor. Differences have been noted according to age, gender, road familiarity, and risk behaviour. The study highlights the need of taking into account both short- and long-term repercussions in traffic education programmes and road safety management by identifying notable differences in causal elements for various demographic groups. In the end, the goal is to remove underlying issues and improve road safety. These findings have consequences for legislative revisions, namely for resolving health eligibility criteria among senior drivers with health problems.

[8] This research paper presents a system designed to quickly detect when drivers are not paying enough attention while driving, using advanced computer techniques. The process involves gathering information about the vehicle's location through GPS, considering factors like speed and distance. To make sure the system works well, a method called Smote is used to handle imbalanced data. The system then looks at various features like how fast the car is going and any changes in altitude. It uses a smart method, the Random Forest algorithm, to spot events related to inattention. For the critical task of recognizing when drivers are not paying attention, a special computer model called Inception v3 is used, along with a technique called mix-up, which makes the system better at recognizing patterns. The system also takes into account additional information from outside sources to improve its predictions. Although the approach is effective in handling imbalanced data and accurately spotting inattention, one thing to consider is that it might require a lot of computing power. Despite this, overall, the system seems to hold great potential for quickly detecting when drivers are not paying enough attention.

[9] This paper explores if drivers who tend to get distracted are less skilled at predicting traffic hazards. Using a self-reported measure called the Attention-Related Driving Errors Scale (ARDES-Spain), the research compares distraction levels between new and experienced drivers. While the study offers insights into factors influencing road accidents, relying on self-reported data and the simplicity of the Hazard Prediction test are limitations. Findings indicate that new drivers report more attentional errors and struggle more with hazard prediction than experienced drivers. The study highlights the importance of driving experience, emphasizing the need for stricter regulations on roadside ads to maintain drivers' awareness and enhance road safety.

[10] Road accidents worldwide are largely caused by distracted driving, thus it's important to understand how this affects hazard prediction, particularly when taking experience into account. In this research paper, 120 participants—half of them were seasoned drivers, the other half were beginners—completed a hazard prediction test while working on a distracting job. The exam assessed their capacity to identify risks in the midst of distractions. The study discovered that while experienced drivers were less affected by distractions and performed better overall, inexperienced drivers' hazard prediction was more negatively impacted by distractibility. The study's findings highlight the significance of driving experience in reducing the effects of distraction, despite the fact that cognitive differences were not taken into account. These results can direct interventions aimed at improving danger prediction abilities and decreasing distracted driving, especially among novice drivers. Future research could explore individual differences and test various training programs for improving hazard prediction skills.

## 

## 

## 

## **Methodology:**

The project methodology for predicting road accidents involves several steps, including data exploration, preprocessing, feature engineering, model selection, training, and evaluation. Here's a breakdown of the methodology:

* **Data Collection:**
  + The dataset was obtained from the Addis-Ababa police departments, available on Kaggle.
  + Features include driver information, weather conditions, collision types, and other relevant details.
* **Data Exploration:**
  + Used Python libraries such as Pandas, Matplotlib, Seaborn to explore the dataset.
  + Checked for missing values, examined summary statistics, and visualized the distribution of different features.
  + Outlier Detection and Handling:
    - Identified outliers in numerical features using techniques like IQR.
    - Mitigated the impact of outliers on analysis and modeling through appropriate handling.
  + Correlation Analysis:
    - Explored feature correlations using a correlation matrix.
    - Examined relationships between features and the target variable ('Accident\_severity').
    - Visualized correlations using heatmaps to detect multicollinearity and significant associations.
* **Data Preprocessing:**
  + Handled missing values and replaced them with appropriate values (e.g., mode for categorical features).
  + Converted categorical variables into numerical format using label encoding.
  + Applied ordinal encoding for ordinal categorical variables.
  + Addressed class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE) and SMOTE Tomek Links.
  + Feature Scaling:
    - Applied feature scaling to normalize numerical features.
    - Utilized Min-Max scaling or Standardization techniques.
    - Ensured separate scaling for training and testing sets to prevent data leakage.
  + Handling Categorical Variables:
    - Employed one-hot encoding for categorical variables without ordinal relationships.
    - Ensured proper encoding to avoid introducing ordinal relationships where not applicable.
    - Checked and addressed the cardinality of categorical features using techniques like frequency encoding.
* **Feature Engineering:**
  + Created new features such as time-of-day categories based on the 'Time' feature.
  + Derived insights from the data exploration phase to identify relevant features for prediction
  + Interaction Terms:
    - Investigated the creation of interaction terms between features to capture combined effects.
    - Introduced new features representing interactions between relevant variables.
    - Analyzed the impact of specific feature combinations on predicting accident severity.
  + Temporal Features (Time Series):
    - Utilized time-series data for features.
    - Organized time data, converting it into meaningful categories like morning, afternoon, evening.
    - Enhanced the interpretability and relevance of time-related features for accident severity prediction
  + .
* **Data Visualization:**
  + Utilized Matplotlib and Seaborn for visualizing accident frequency and fatality ratios across different categories.
  + Plotted bar charts to show the distribution of accidents by various factors.
* **Model Selection:**
  + Choose three machine learning algorithms: Random Forest, XGBoost, and Extra Trees for accident severity prediction.
  + Utilized the scikit-learn library for implementation.
* **Model Training:**
  + Split the dataset into training and testing sets.
  + Trained each selected model on the training data.
* **Upsampling Techniques:**
  + Implemented SMOTE and SMOTE Tomek Links to address class imbalance in the dataset.
  + Evaluated the impact of upsampling on model performance.
* **Model Evaluation:**
  + Used F1-score as the evaluation metric to account for the imbalanced nature of the dataset.
  + Assessed the performance of each model on the test set**.**
* **Insights and Recommendations:**
  + Derived insights from the trained models to identify factors contributing to accident severity.
  + Provided recommendations for improving road safety based on the analysi**s.**
* **Comparison of Models:**
  + Compared the performance of Random Forest, XGBoost, and Extra Trees models.
  + Selected the best-performing model for the given dataset and problem.
* **Project Conclusion:**
  + Summarized the findings, highlighted key insights, and discussed the implications for road safety efforts.
  + Concluded with the potential impact of the project on data-driven decision-making in traffic management.

By following this methodology, the project aims to contribute to the understanding and prediction of road accidents, ultimately enhancing safety measures and reducing the severity of accidents on the roads.

## 

## 

## 

## **Results and Discussion:**

The first step was to divide the dataset into two different sets for training and testing in an 80:20 ratio. It resulted in 9852 data samples for training and 2464 samples for testing. The preprocessing was important to handle categorical variables properly. There were two datasets for the purpose of experimentation: one using mean encoding and another using a combination of mean and label encoding. The combination dataset captured a wider range of information in categorical variables.

In mean encoding strategy, Synthetic Minority Over-Sampling Technique (SMOTE) exhibited promising results. The F1 scores for the algorithms were as follows: 0.85 for Random Forest, 0.85 again for XGBoost and the Extra Trees algorithm had 0.84. These scores are reflective of an impressive performance across different classifiers. It indicates that mean encoding has the ability to successfully convert categorical information to a numerical state. This enables the models to generalize on unseen data. The consistency strengthens the reliability of the models.

In order to include both mean and label encoding, the dataset returned better model performance. The F1 scores were 0.85 for XGBoost, 0.85 for Extra Trees and 0.84 for Random Forest on incorporating SMOTE to cover oversampling. This small improvement suggests the models ability to comprehend intricate relationships within the data. This encoding offers a better representation of the categorical features.

Oversampling through SMOTE displayed the models’ resilience to handle the data imbalances. The addition of label encoding boosts the performance and it plays a critical role.

## **Conclusion:**

## In conclusion, our project focused on predicting car accident severity using real-world traffic data. We meticulously prepared the data, addressing missing values and considering factors like road conditions. We explored two modeling approaches, one involving advanced machine learning models like XGBoost and Random Forest, and another incorporating techniques like SMOTE and SMOTETomek to address class imbalance.

The models showed promising results, with specific improvements noted when using SMOTE and SMOTETomek. Notably, these techniques enhanced the models' ability to predict accident severity accurately. While precise accuracy values varied among models, the detailed examination of driver-related factors revealed intriguing patterns. For instance, younger and less experienced drivers were identified as more prone to severe accidents.

The study highlights how oversampling methods effectively balance dataset disparities. It also emphasizes the role of data-driven approaches in shedding light on accident severity and driver reliability. Through oversampling, the project not only boosts overall traffic safety but also offers a detailed look into the factors shaping accidents and driver dependability. This nuanced insight enriches our understanding of traffic dynamics, contributing to a more comprehensive approach to road safety improvement.

**References:**

1. Vasavi, S. "Extracting hidden patterns within road accident data using machine learning techniques." *Information and Communication Technology: Proceedings of ICICT 2016*. Springer Singapore, 2018.

**Link:** <https://link.springer.com/chapter/10.1007/978-981-10-5508-9_2>

1. R. Bhardwaj, S. Parameswaran and V. Balasubramanian, "Performance Comparison of Machine Learning and Deep Learning While Classifying Driver’s Cognitive State," 2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS), Rupnagar, India, 2018, pp. 89-93, doi: 10.1109/ICIINFS.2018.8721374.

**Link:** <https://ieeexplore.ieee.org/abstract/document/8721374>

1. J. Gwak, M. Shino and A. Hirao, "Early Detection of Driver Drowsiness Utilizing Machine Learning based on Physiological Signals, Behavioral Measures, and Driving Performance," 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 2018, pp. 1794-1800, doi: 10.1109/ITSC.2018.8569493.

**Link:** <https://ieeexplore.ieee.org/abstract/document/8569493>

1. S. Alagarsamy, M. Malathi, M. Manonmani, T. Sanathani and A. S. Kumar, "Prediction of Road Accidents Using Machine Learning Technique," 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2021, pp. 1695-1701, doi: 10.1109/ICECA52323.2021.9675852.

**Link:** <https://ieeexplore.ieee.org/abstract/document/9675852>

1. De Raedt, Rudi, and Ingrid Ponjaert-Kristoffersen. "Predicting at-fault car accidents of older drivers." Accident Analysis & Prevention 33.6 (2001): 809-819.

Link: <https://www.sciencedirect.com/science/article/pii/S0001457500000956>

1. Yassin, Salahadin Seid, and Pooja. "Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach." SN Applied Sciences 2 (2020): 1-13.

**Link:** <https://link.springer.com/article/10.1007/s42452-020-3125-1#Sec16>

1. Bucsuházy, Kateřina, et al. "Human factors contributing to the road traffic accident occurrence." Transportation research procedia 45 (2020): 555-561

Link: <https://www.sciencedirect.com/science/article/pii/S2352146520302192>

1. Jiang, Landu, et al. "Smart diagnosis: Deep learning boosted driver inattention detection and abnormal driving prediction." *IEEE Internet of Things Journal* 9.6 (2021): 4076-4089.

**Link:** <https://ieeexplore.ieee.org/abstract/document/9509762>

1. Chand, Arun, S. Jayesh, and A. B. Bhasi. "Road traffic accidents: An overview of data sources, analysis techniques and contributing factors." Materials Today: Proceedings 47 (2021): 5135-5141.

**Link:** <https://www.sciencedirect.com/science/article/pii/S2214785321040153?via%3Dihub>

1. Castro, Candida, et al. "How are distractibility and hazard prediction in driving related? Role of driving experience as moderating factor." Applied ergonomics 81 (2019): 102886.

**Link:** <https://www.sciencedirect.com/science/article/pii/S0003687019301188>