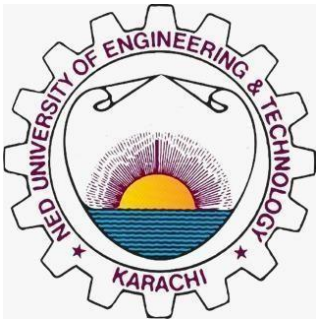
PROJECT REPORT



ACCIDENT ANALYSIS

COURSE CODE / TITLE	MACHINE LEARNING (CS-324)
SUBMITTED BY	AYMEN FATIMA HASSAN (CS-053) HAFSA HABIB (CS-081) ABBAS RAZA(CS-097)
YEAR / BATCH	THIRD YEAR / 2021
GROUP	G3
SUBMITTED TO	Mahnoor Malik
WEEK NO.	15TH

INTRODUCTION:

Our innovative machine learning model predicts vehicle accident rates by accurately identifying individual vehicles and analyzing a comprehensive set of data, including historical accident records, vehicle specifications, and real-time traffic information. By leveraging advanced algorithms and sophisticated data processing techniques, our model can recognize complex patterns and correlations within these diverse datasets. This enables the provision of precise and actionable insights that significantly enhance road safety and mitigate potential risks.

The application can help insurance companies, fleet managers, and individual drivers. For insurance companies, the model facilitates the development of more accurate risk assessments and personalized insurance policies, thereby improving customer satisfaction and operational efficiency. Fleet managers gain the ability to optimize their vehicle management strategies, reducing accident-related costs and enhancing overall fleet safety. Individual drivers receive tailored recommendations that promote safer driving practices, reducing their likelihood of being involved in accidents.

DATASET:

The Accident Analysis dataset was obtained using Kaggle. Features like Vehicle Type, Speed Limit, Road Type, Accident Severity, Accident Injuries, Number of Casualties, Death Toll, Hour of Accident, and so forth were included in the dataset. A comprehensive foundation for analysis and model training was offered by the dataset. We were able to investigate past trends, spot patterns, and build predictive models with a strong data base thanks to the databases.

FEATURE ENGINEERING:

1. Handling Missing Values:
    - Dropped unbalanced column like `Carriageway_Hazards`.
    - Filled missing values in `Road_Surface_Conditions` and `Road_Type` with their most frequent values.
    - Filled missing values in `Weather_Conditions` using a logical mapping from `Road_Surface_Conditions` and defaulted to 'Other' for remaining gaps.
2. Data Cleaning:
    - Corrected typo in `Accident_Severity`.
    - Dropped irrelevant columns: `Latitude`, `Longitude`, `Junction_Control`, `Local_Authority_(District)`, `Police_Force`.
3. Feature Transformation: Extracted hour from `Time` and converted to float for numerical analysis, filling missing values with the mean hour.
4. Renaming Columns: Renamed columns for clarity:
    - `Accident_Severity` to `Accident_Injuries`

- `Number_of_Casualties` to `death_toll`
- `Accident_Index` to `Accident_ID`
- `Time` to `Hour_of_Accident`

5. Categorical Mapping: Consolidated `Vehicle_Type` categories for simplicity and consistency.

EXPLORATORY DATA ANALYSIS (EDA):

Data visualization is shown through bar charts and graphs of Road_Surface_Conditions Percentage of Accidents by Day of the Week Number of Accidents per Each Hour in the Day (Sorted) Accident by Vehicle Type, Accident by Light Condition percentage of Accidents by Date Accident count by Injuries Accident count by Speed Limit and Count of accident by junction details for each day in the week.

MODEL SELECTION:

1.Decision Tree Regressor: A model that splits the data into subsets based on feature values, creating a tree-like structure.

2.Logistic Regression: It models the probability of a binary outcome based on one or more predictor variables, The model uses a logistic function to map predicted values to probabilities

MODEL TRAINING:

The dataset used to predict traffic accidents was used to train each model. Cross-validation was used to optimize the hyperparameters in order to guarantee optimal performance. We split the data into training and testing sets, fine-tuned the models, and made sure they performed well on data that had not yet been seen. In total, we trained two models

EVALUATION:

Models were evaluated using accuracy, confusion matrix, and classification report. Accuracy measures the proportion of correctly predicted instances. The confusion matrix provides insight into the types of errors made by the model, showing the counts of true positive, true negative, false positive, and false negative predictions. The classification report includes precision, recall, F1-score, and support for each class, providing a comprehensive view of model performance.

LIMITATION:

- Generalizability: Region-Specific Models: Models trained on data from specific regions may not generalize well to other regions with different driving conditions, road types, and weather patterns.
- Handling of Time: Treating Hour_of_Accident as a continuous variable might not capture the cyclic nature of time effectively, potentially impacting model accuracy.
- Category Simplification: Grouping various vehicle types and other features into broader categories might result in a loss of valuable, granular information.

WEB APPLICATION:

We developed a web application to showcase our predictions, visualizations, and analysis. The website features interactive charts, allowing users to explore the historical trends. User can view performance metrics for each model, providing a comprehensive understanding of our findings. The application serves as a valuable tool for both educational and practical purposes.

CONCLUSION:

This experiment effectively illustrated how different machine learning approaches may be applied to forecast traffic accidents. Accuracy, confusion matrix, and classification report were used to evaluate the models, giving a thorough picture of their performance. An approachable means of exploring and comprehending the analysis and forecasts is provided via the web application. To increase forecast accuracy, future study may incorporate more advanced models and broaden the feature set.