

**Road Traffic Accident Prediction**

A project report submitted in partial fulfilment of the requirements for the Degree of Bachelor of Engineering & Technology

in

**COMPUTER ENGINEERING (CP) SEM – VI**

by

**Name: Heet Aghara, Darshan Bhuva Enrollment No.: 12202040501005, 12202040501013**

Under supervision of

**Prof. Priyang Bhatt**

# Academic Year 2024-25 (EVEN)

Department of Computer Engineering

G H Patel College of Engineering & Technology Bakrol Road, Vallabh Vidyanagar

# CERTIFICATE

This is to certify the project work embodied in this report entitled, “ **Road Accident Using Linear Regression**” was carried out by **Heet Aghara** and **Darshan Bhuva**. Enrollment No. **12202040501005** and **12202040501013** at G H Patel College of Engineering & Technology for partial fulfilment of B.Tech. degree to be awarded by Charutar Vidya Mandal University. This seminar work has been carried out under my supervision and is to the satisfaction of department.

## Guide Head of Department

Prof. Priyang Bhatt Dr. Maulika Patel

# ACKNOWLEDGEMENT

The completion of any project work depends upon cooperation, co-ordination, and combined efforts of several sources of knowledge. I would like to express my deepest thanks to **Prof. Priyang Bhatt** for her valuable inputs, guidance, encouragement, wholehearted cooperation, and constructive criticism throughout the duration of our project.

I hope that this seminar work report will provide all necessary information required to readers to fulfil their aspiration. Theory and practices are essential and complimentary to each other.

**Introduction**

The objective of this mini project is to predict the number of deaths caused by traffic accidents using historical accident data and machine learning techniques. Traffic accidents remain a significant public safety issue, and predictive analytics can help in policy-making and resource allocation. In this project, we apply a Linear Regression model to estimate the number of deaths based on various features such as road type, weather conditions, alcohol involvement, speed limits, and more.

**Key Concepts Covered in the Project:**

1. Descriptive Statistics:
   * Descriptive analysis was conducted to understand the distribution of features and the target variable (Number\_of\_Deaths). Measures like mean, standard deviation, and frequency counts were used.
2. Data Preparation:
   * The dataset was cleaned by removing duplicates, encoding categorical variables, and handling missing values to prepare it for modeling.
3. Algorithm Used:
   * A Linear Regression model was used for prediction. Linear Regression attempts to establish a linear relationship between input features and the target variable.
4. Performance Measures:
   * Model performance was assessed using Mean Squared Error (MSE), Root Mean Squared Error **(RMSE), and the R² Score.**

**Data Preparation**

Dataset:

The dataset used in this project contains historical traffic accident records. Key columns in the dataset include:

* State: The state where the accident occurred
* Speed\_Limit: The speed limit in the area
* Alcohol\_Involved: Indicates if alcohol was involved
* Weather\_Condition, Road\_Type, Light\_Condition: Environmental and infrastructural conditions
* Number\_of\_Deaths: The target variable representing fatalities in the accident

**Data Cleaning:**

1. Duplicate Removal:
   * Duplicate records were removed to maintain data integrity.
2. Handling Categorical Variables:
   * Label encoding and one-hot encoding techniques were used to convert categorical variables into numeric format.
3. Missing Value Handling:
   * Missing values were handled using forward-fill methods or dropped based on relevance.

**Exploratory Data Analysis (EDA):**

* Distribution Plots:
  + Visualizations such as count plots and bar charts were used to observe the distribution of deaths and other variables.
* Correlation Heatmap:
  + A heatmap was generated to analyze correlations between features and the target variable, helping to identify impactful predictors**.**

**Algorithm Implementation**

Model Choice:

The Linear Regression model was chosen due to its simplicity and interpretability. It provides a baseline model to understand how well linear relationships can predict traffic fatalities.

**Training and Testing the Model:**

The dataset was split into training and testing sets (80% training, 20% testing) using train\_test\_split from sklearn.

python

CopyEdit

X = encoded\_features

y = df['Number\_of\_Deaths']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

model = LinearRegression()

model.fit(X\_train, y\_train)

**Model Evaluation:**

After training, predictions were made using the test data, and performance was measured using evaluation metrics.

python

CopyEdit

predictions = model.predict(X\_test)

mse = mean\_squared\_error(y\_test, predictions)

rmse = np.sqrt(mse)

r2 = r2\_score(y\_test, predictions)

**Performance Metrics:**

* MSE: 1.18
* RMSE: ~1.09
* R² Score: 0.06  
  A low R² score indicates that the linear model does not capture much **Challenges Faced**

1. Data Cleaning and Preprocessing:
   * Handling categorical and missing data effectively.
   * Encoding strategies that preserved useful information.
2. Feature Selection:
   * Selecting relevant features that actually influence accident outcomes.
   * Ensuring the model does not overfit due to noisy or irrelevant features.
3. Model Training and Optimization:
   * The linear model showed limitations due to complex, non-linear relationships in the data.
4. Evaluation and Interpretation:
   * Understanding the limitations of linear regression when applied to real-world datasets.
   * Recognizing the importance of low R² in regression interpretation.
5. Visualization:
   * Creating clear, insightful visuals to compare predictions and actual values.