

PROJECT ON ROAD ACCIDENTS DUE TO WEATHER CONDITIONS

SUBMITTED BY:

GOUTAM MOHANTY

GULNAZ AHMAD

ROSHAN KUMAR

5th Semester CSE (BATCH: 2022-2026)

*Under the guidance of:*

DR.JAYSHREE PIRI

****

**STUDENT DECLARATION**

I/ We hereby declare that the project report entitled "Handwritten Digit Recognition" submitted in partial fulfillment of the requirements for the degree of Bachelor of Technology in Computer Science and Engineering to SILICON UNIVERSITY is our original work and not submitted to any other university or lnstitute for the award of any degree or diploma.

BY:

GOUTAM MOHANTY

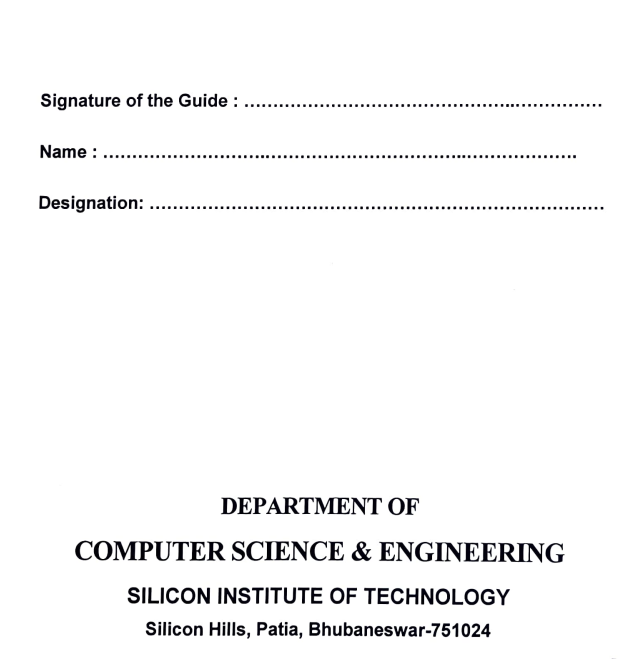
GULNAZ AHMAD

ROSHAN KUMAR

****

*CERTIFICATE*

This is to certify that Goutam Mohanty, Gulnaz Ahmad, Roshan Kumar have undertaken and successfully completed the project entitled "Handwritten Digit Recognition" under my supervision. This work is original and is being submitted as a part of 4 Semester project for the undergraduate curriculum.

**

*ACKNOWLEDGEMENT*

We are really grateful because we managed to complete our ROAD ACCIDENTS DUE TO WEATHER CONDITIONS project within the time given by our guide DR. JAYSHREE PIRI. This project cannot be completed without the effort and co-operation from our group members. We also sincerely thank our guide DR. JAYSHREE PIRI for the guidance and encouragement in finishing this project. Last but not the least, we would like to express our gratitude to our friends and respondents for the support and willingness to spend some times with us to fill in the questionnaires.

NAME:- GULNAZ AHMAD NAME:- GOUTAM MOHANTY

SIC:- 22BCSH62 SIC:- 22BCSF16

SEC:- A(2),CSE SEC:- A(2),CSE

****

**ABSTRACT**

Road accidents and their associated societal and economic impacts have been persistent global challenges over the past two to three decades. These issues have affected both developing and developed countries, with densely populated urban centers bearing the brunt. In the United States, all major cities, including New York City, Dallas, Texas, Miami, and others, face significant challenges related to road safety. With the United States as our study area, this research aims to address road safety concerns by analyzing and determining the degree of severity of road accidents.

Our study employs machine learning techniques to assess the impact of weather on road accident risks and identify high-risk conditions. Preliminary results reveal a strong association between adverse weather—such as heavy rainfall, snow, fog, and reduced visibility—and increased accident frequency. These findings emphasize the critical need to integrate weather-related considerations into urban planning and road safety strategies to mitigate accident risks effectively. Additionally, the role of extreme temperature conditions and their influence on accident patterns has also been examined.

The research is underpinned by a decade's worth of accident-related data, analyzed through the spatial tools of Geographical Information Systems (GIS). Using GIS-based mapping and machine learning models such as **logistic regression**, **Support Vector Machine (SVM)**, and **Random Forest**, we have identified the spatio-temporal distribution of accident-prone locations alongside their degrees of severity. Each of these models has contributed uniquely to the analysis:

* **Logistic Regression**: Used for predicting the likelihood of accidents under varying conditions.
* **KNN**: Helped in classifying accident severity based on high-dimensional data features.
* **Random Forest**: Played a key role in identifying critical contributing factors and determining accident hot spots.

Our experimental results highlight accident hot spots, correlating them with various contributing factors such as timing (e.g., peak traffic hours), accident duration, wind speed, visibility, vehicle speed, weather conditions, and weather timestamps.

This comprehensive analysis not only identifies high-risk conditions and locations but also provides actionable insights for policymakers and urban planners. By leveraging machine learning, spatial analysis, and the collective strength of the models employed, the study aims to contribute significantly to reducing road accident occurrences and improving safety measures for both urban and suburban areas.

|  |  |  |
| --- | --- | --- |
| S.no. | TOPIC | PAGE |
| 1 | ABSTRACT | 1-1 |
| 2 | INTRODUCTION | 3-3 |
| 3 | PROBLEM STATEMENT | 3-3 |
| 4 | REVIEW OF RELATED WORKS | 3-3 |
| 5 | DATASET ANALYSIS |  |
| 6 | ANALYSIS OF MODEL USED | 4-4 |
| 7 | ALGORITHM | 4-4 |
| 8 | RESULT | 4-7 |
| 9 | SCOPE FOR FUTURE WORKS |  |
| 10 | REFERENCES |  |

**TABLE OF CONTENTS**

**INTRODUCTION**

* **Global Road Safety Challenge:** Road accidents have caused significant societal and economic impacts globally over the past decades, with urban centers being most affected.
* **Study Focus:** This research focuses on the United States, analyzing the role of weather in influencing road accident severity.
* **Data Overview:** A decade’s worth of accident-related data is analyzed using machine learning techniques and GIS-based spatial tools.
* **Weather-Accident Correlation:** Adverse weather conditions, including heavy rainfall, snow, fog, and extreme temperatures, are strongly associated with increased accident frequency and severity.
* **Key Insights:** Insights aim to assist policymakers and urban planners in integrating weather-related considerations into road safety strategies.

**Machine Learning Models Used:**

* **Logistic Regression:** Predicts the probability of accidents under varying weather and traffic conditions.
* **K-Nearest Neighbors (KNN):** Classifies accident severity based on patterns in high-dimensional data.
* **Random Forest:** Identifies critical factors contributing to accidents and locates high-risk hotspots.
* **GIS-Based Mapping:** Supports spatio-temporal analysis by identifying accident-prone locations and their degrees of severity.
* **Objective:** Provide actionable insights for reducing accidents, improving road safety, and informing urban planning and traffic management strategies.

**PROBLEM STATEMENT**

Road accidents are significantly influenced by adverse weather conditions, such as rain, fog, and wind, which impair visibility and compromise road safety. Understanding the relationship between weather conditions and accident severity is crucial for improving road safety measures. This project aims to develop a machine learning model to predict the severity of road accidents using historical weather data, geographical information, and accident-related features. The model will assist in identifying high-risk conditions, enabling proactive safety measures to mitigate accidents and reduce their impact.

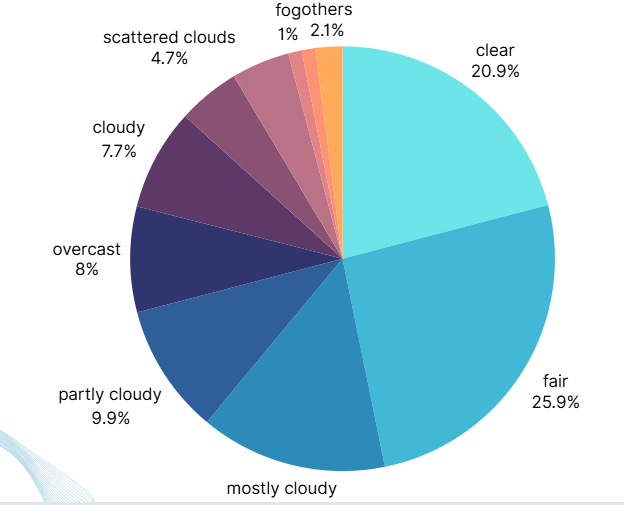
Accidents caused by weather-related factors continue to be a major concern, leading to loss of life, injuries, and substantial economic costs. With increasingly frequent extreme weather events due to climate change, understanding how weather conditions affect accident severity is more important than ever. By predicting accident severity with machine learning, we can provide valuable insights for city planners, traffic management authorities, and insurance companies. Proactively addressing high-risk conditions can lead to safer road environments, reduced accident-related costs, and improved public safety.

**REVIEW OF RELATED WORKS**

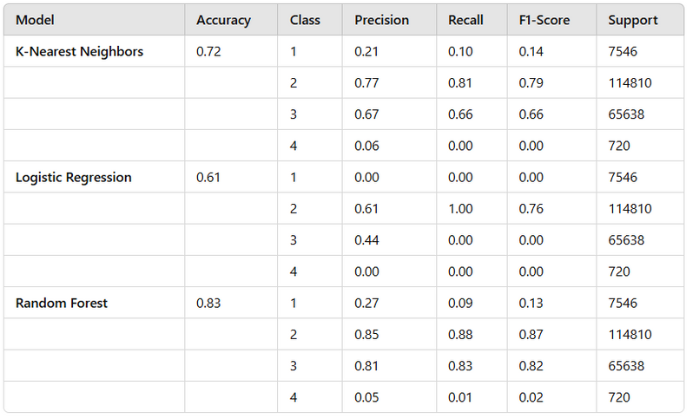
1. **P. B. Parmar et al. [7]**  
   Identified accident-prone locations (blackspots) on S.P. Ring Road, Ahmedabad using the heatmap plugin in QGIS. The study suggested remedial measures, such as determining locations where enforcement steps are needed and identifying areas where signboards for speed restriction and traffic signs should be installed.
2. **J. Choudhary et al. [8]**  
   Geocoded accident locations for the years 2009 to 2013 over a digitized map of Varanasi. They used Kernel Density Estimation (KDE) and the heatmap plugin to analyze spatial densities and clustering of accidents. The study successfully isolated accident hotspots along road stretches using these methods.
3. **Jerome Ballarta et al. [9]**  
   Identified accident hotspots in Katipunan Avenue, Quezon City. The authors utilized standard GIS geoprocessing techniques to determine the accident-prone areas effectively.
4. **V. Prasannakumar et al. [10]**  
   Conducted a comparative study on temporal and spatial aspects of road accidents in Thiruvananthapuram. By employing cluster analysis and spatial data statistics, they highlighted the differences between accident-vulnerable and non-vulnerable locations over time and space.
5. **Sanjay Kumar Singh and Ashish Misra [11]**  
   Analyzed road accidents in Patna city using annual data from 1996 to 2000. The study provided a comprehensive overview of the road accident scenario in India and evaluated the existing transport system in Patna.
6. **Deelesh Mandloi and Rajiv Gupta [12]**  
   Developed a GIS-based model to predict accident-vulnerable locations using road accident-related parameters. The study also proposed remedial steps to improve road safety and reduce accidents.
7. **Yen Chen et al. [13]**  
   Discussed the application of geocoding technology for preparing spatial information related to traffic accidents. The authors presented a method to identify blackspots by associating road network features with accident-prone areas using GIS data storage techniques.
8. **Ela Ertung et al. [14]**  
   Conducted a GIS-based analysis of intersection road accidents in Antalya, Turkey, using fatal and injury traffic accident data from 2009 to 2010. The study identified hotspots for intersection accidents and carried out statistical evaluations of the accident data.
9. **Anik Vega and Dwi Cahyono [15]**  
   Used Multiple-Attribute Utility Theory (MAUT) to map dense and accident-prone traffic roads. The approach helped analyze spatial and attribute data to identify alternative routes to minimize traffic density and reduce accident risks.
10. **Michal Bil et al. [16]**  
    Evaluated and organized clusters of traffic accidents based on their significance. Using a KDE-based strategy, they identified and ranked the most hazardous spots, emphasizing the importance of verifying and ordering accident clusters.
11. **Hao Yu et al. [17]**  
    Conducted a comparative study of spatial analysis techniques for identifying hotspots using data from a 622.2-km section of the A1 highway in the UK. Between 2001 and 2010, the authors analyzed 7930 crashes within the selected highway zones to identify accident-prone areas.
12. **Gholam A. Shafabakhsh et al. [18]**  
    Performed spatial analysis of traffic accidents in Mashhad, Iran, using GIS methods. This study, the first of its kind for the city, provided a detailed analysis of various accident types and their spatial patterns, offering insights into the road safety challenges in the region.

**DATASET DESCRIPTION**

* Coverage: 49 states of the USA from February 2016 to March 2023.
* Data Sources:
  + Departments of Transportation
  + Law enforcement agencies
  + Traffic cameras
  + Road sensors
* Size: Approximately 7.7 million accident records.
* Features: 45 total, including:
  + Numeric: Distances, wind speed, temperature, etc.
  + Categorical: Weather conditions, city, state, etc.
* Insights: Provides detailed information on weather conditions, traffic patterns, and accident severity.
* Types of weather responsible for accidents:



**ANALYSIS OF MODELS USED**



Why choose Random Forest over the other models?

1. **Handling Non-linearity**:

* **Random Forest** is a non-linear model, meaning it can capture complex relationships in the data that might not be captured by linear models like **Logistic Regression**.
* **KNN**, while also capable of handling non-linearity, can become computationally expensive with large datasets and may be sensitive to the choice of distance metric.

1. **Model Flexibility**:

* Random Forest builds an ensemble of decision trees and aggregates their predictions, which allows it to handle a variety of feature types and distributions better.
* **Logistic Regression** is limited to linear relationships and might not perform well if the true relationship between features and the target is non-linear.
* **KNN** requires careful consideration of the distance metric and number of neighbors, and its performance can degrade when dealing with high-dimensional data (curse of dimensionality).

1. **Robustness**:

* **Random Forest** is less likely to overfit compared to a single decision tree. It can also deal with noisy data and missing values better.
* **KNN** might suffer from overfitting if the number of neighbors is too small or underfitting if it's too large.
* **Logistic Regression** could be affected by outliers or collinearity between features, leading to poor model performance.

1. **Interpretability**:

* Although **Random Forest** is often considered a "black-box" model due to the ensemble of trees, it offers interpretability through feature importance scores, which help understand how features contribute to the model’s predictions.
* **Logistic Regression** offers coefficients that are easily interpretable, but only works well when features have linear relationships with the target.
* **KNN** provides very little interpretability, as it relies on the proximity of data points in feature space without providing an explicit model structure.

1. **Accuracy**:

* Random Forest generally provides higher predictive accuracy for most datasets compared to **Logistic Regression** (which is limited to linear decision boundaries) or **KNN** (which can struggle in high-dimensional spaces or with noisy data).

**ALGORITHM**

**Step 1: Load the Data**

1. **Load the dataset**: Import the CSV file containing accident and weather data using pd.read\_csv('US\_Accidents\_March23.csv').

**Step 2: Preprocessing**

1. **Filter Data Based on Weather Condition**:
   * Use value\_counts() to get the count of each weather condition.
   * Filter out rows where the weather condition occurs more than 100 times.
2. **Handle Missing Values**:
   * Drop rows where any of the relevant columns ('Start\_Lat', 'Start\_Lng', 'Weather\_Condition', 'Wind\_Speed(mph)', 'Severity') have missing values.
3. **Encode Categorical Data**:
   * Use LabelEncoder to encode the Weather\_Condition column to numeric values. This is necessary for machine learning models, which require numerical data.
4. **Separate Features and Target Variable**:
   * Define X as the independent variables (Start\_Lat, Start\_Lng, Weather\_Condition, Wind\_Speed(mph)).
   * Define y as the target variable (Severity).

**Step 3: Exploratory Data Analysis (EDA)**

1. **Correlation Matrix**:
   * Calculate and display the correlation matrix of numeric features.
   * Identify the correlation between each feature and the target variable (Severity).
2. **Chi-Square Test**:
   * Apply the Chi-Square test for independence between categorical features and the target variable. This helps identify relationships between categorical features and Severity.

**Step 4: Train-Test Split**

1. **Split the Data**:
   * Split the dataset into training and testing sets using train\_test\_split (80% training, 20% testing).

**Step 5: Scaling Features**

1. **Scale Numeric Features**:
   * Use StandardScaler to normalize the numeric features (Start\_Lat, Start\_Lng, Wind\_Speed(mph)) to ensure all features are on the same scale for better model performance.

**Step 6: Model Training**

1. **Train Random Forest Model**:
   * Initialize and train a RandomForestClassifier.
   * Make predictions on the test set (y\_pred\_rf).
   * Evaluate model performance using accuracy score and classification report.
2. **Train Logistic Regression Model**:
   * Initialize and train a LogisticRegression model.
   * Make predictions on the test set (y\_pred\_log\_reg).
   * Evaluate model performance using accuracy score and classification report.
3. **Train K-Nearest Neighbors Model**:
   * Initialize and train a KNeighborsClassifier model.
   * Make predictions on the test set (y\_pred\_knn).
   * Evaluate model performance using accuracy score and classification report.

**Step 7: Model Evaluation**

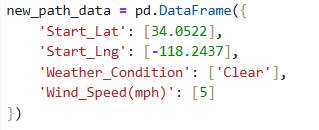
1. **Evaluate All Models**:
   * Print the accuracy scores and classification reports for each model (Random Forest, Logistic Regression, KNN).

**Step 8: New Data Prediction**

1. **Predict for New Path Data**:
   * Prepare new path data with Start\_Lat, Start\_Lng, Weather\_Condition, and Wind\_Speed(mph) (example coordinates and weather condition).
   * Encode the weather condition for new data using the same LabelEncoder as the training data.
   * Scale the numeric features for new data using the same StandardScaler.
   * Use all three models (Random Forest, Logistic Regression, KNN) to predict the severity for the new path data.

**RESULT**

Now upon giving an input such as:



The Output displayed will :

Severity: [2]

**SCOPE FOR FUTURE WORKS**

1. Incorporate Infrastructure Data :

* Integrate road condition data
* Include traffic signals, street lighting and other urban planning features to improve risk prediction.

1. Real Time Weather Forecasting API:

* Implement a real-time weather API to improve up-to-date data on temperature, rainfall, wind speed and visibility.

1. Real time re-routing:

* Re-routing of alternative paths with safer chances.

1. Local Implementation:

* Implementation in cities with high population density such as Bhubaneswar, Cuttuck, etc.

**REFERENCES**

* Dataset:- https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents
* Shaaban, A. Siam, and A. Badran, "Analysis of Traffic Crashes and Violations in a Developing Country," Transportation Research Procedia, vol. 55, pp. 1689-1695, 2021.
* P. Thomas, A. Morris, R. Talbot, and H. Fagerlind, "Identifying the causes of road crashes in Europe," Annals of advances in automotive medicine, vol. 57, p. 13, 2013.
* K. Shaaban and M. Ibrahim, "Analysis and Identification of Contributing Factors of Traffic Crashes in New York City," Transportation Research Procedia, vol. 55, pp. 1696-1703, 2021.
* Yu H, Liu P, Chen J, Wang H. Comparative analysis of the spatial analysis methods for hotspot identification. Accid Anal Prev. 2014 May;66:80-8. doi: 10.1016/j.aap.2014.01.017. Epub 2014 Jan 29. PMID: 24530515.
* Organization, W.H., ”Global Status Report on Road Safety 2018,” WHO: Geneva, Switzerland. 2019
* Cheng Z, Zu Z, Lu J (2018) Traffic crash evolution characteristic analysis and spatiotemporal hotspot identification of urban road intersections. Sustainability 2019(11) :160. https://doi.org/10.3390/su11010160
* Choudhary, Jayvant & Ohri, Anurag & Kumar, Brind. (2015). Identification of Road Accidents Hot Spots in Varanasi using QGIS. Proceedings of National Conference on Open-Source GIS: Opportunities and Challenges; Department of Civil Engineering IIT(BHU), Varanasi; October 9-10, 2015. ISBN:978-81-931-2500-7
* Parida M, Jain SS, Kumar CN (2012) Road traffic crash prediction on national highways. Indian Highway Road Congress 40:93–103