NAMES OF PARTICIPANTS: KALULU GARVEN (PROJECT MANAGER), BOGDAN VAKHRIN/LIKOLA MISHECK(BACKEND), I ANTON (ML), VADIM RYZHOV (FRONTEND), 5130203/20101.

A COMPREHENSIVE REPORT OF TRAFFIC ACCIDENT PREDICTION.

Introduction

Traffic accidents pose a critical challenge to public safety worldwide, causing significant loss of life, injuries, and economic impact. Predicting where and when accidents are likely to occur can play a crucial role in reducing these risks. By enabling proactive measures, such as targeted road safety improvements or real-time alerts, lives can be saved, and accidents prevented.

This report delves into the use of machine learning models to forecast the likelihood of traffic accidents. These predictions are informed by a range of factors, including environmental conditions, driver behavior, and vehicle attributes. This report is organized into four key sections:

Understanding the Problem: Why traffic accident prediction matters and the specific challenges it seeks to address.

Methodology and Approach: How machine learning models are developed and the data they rely on.

Findings and Insights: The results of the predictive models and their implications for accident prevention.

Conclusion: Key takeaways and future opportunities to enhance road safety.

Through this exploration, we aim to highlight how predictive analytics can transform traffic management and make our roads safer for everyone.

PROBLEM STATEMENT

This project aims to create a reliable system that predicts the likelihood and severity of traffic accidents, helping to reduce their occurrence and impact. Here's a breakdown of the problem we seek to address:

Context

Traffic accidents are increasing globally due to the growing number of vehicles on the road and varying environmental conditions. Understanding the factors that contribute to these incidents—such as road type, weather, and driver behavior—is essential for implementing effective preventative measures.

Dataset Overview

The dataset used for this project includes a range of features that influence traffic accidents, such as:

Environmental Factors: Weather conditions and road type.

Temporal Variables: Time of day and traffic density.

Vehicle and Driver Factors: Speed limits, number of vehicles, and driver alcohol consumption.

Outcome: Accident severity.

The dataset was divided into two parts: a training set for building the model and a testing set for validation.

The Challenge

Developing an accurate predictive model involves overcoming key obstacles:

Diverse Data Types: The dataset includes a mix of categorical and numerical features that require careful preprocessing.

Imbalanced Outcomes: Accident severity and occurrence may be unevenly represented in the data, demanding strategies like resampling or weighted models to address bias.

The Goal

To tackle these challenges, we aim to build a machine learning pipeline that:

Preprocesses the Data: Cleans and prepares the dataset for analysis.

Selects Significant Features: Identifies the most influential factors contributing to accidents.

Applies Predictive Models: Uses advanced machine learning algorithms to forecast the likelihood and severity of traffic accidents.

By achieving these objectives, the project seeks to provide actionable insights that can enhance road safety and save lives.

METHODOLOGY OF SOLUTION

To develop an accurate and effective traffic accident prediction system, we followed a systematic machine learning workflow. Below is a breakdown of the key steps and techniques used:

1. Data Preprocessing

Handling and preparing the data for analysis was a critical initial step:

Dealing with Missing Values:

Missing data was visualized and analyzed using tools implemented in preprocess.py.

Appropriate techniques were applied to filter and manage missing values for consistency.

Feature Engineering:

Categorical Data: Categorical columns were carefully analyzed and encoded to make them suitable for machine learning models.

Correlation Analysis: Relationships between numerical features were studied to understand dependencies and inform feature selection.

Distribution Analysis:

Histograms provided insights into the distribution of numerical features.

Pie charts and frequency plots highlighted the distribution of categorical variables.

Data Visualization:

Tools like Matplotlib and Seaborn were used to create intuitive visualizations, including histograms and correlation matrices, to aid data exploration.

2. Model Development

We explored various machine learning models, tuning them to achieve optimal performance:

Models Implemented:

K-Nearest Neighbors (KNN):

Fine-tuned for parameters such as the number of neighbors (n_neighbors), weighting strategy (weights), and distance metric (metric).

Support Vector Machines (SVM):

Examined multiple kernels (linear, radial basis function, polynomial) and adjusted regularization parameters to improve performance.

Random Forest:

Optimized for the number of trees (n_estimators), tree depth, and feature splits.

Gradient Boosting Models:

XGBoost and LightGBM:

Parameters like learning rate, tree depth, and leaf nodes were fine-tuned for enhanced accuracy and efficiency.

Optimization Techniques:

Hyperparameter Tuning:

Conducted using GridSearchCV to systematically evaluate combinations of parameters.

Cross-Validation:

Employed to ensure models generalized well to unseen data, avoiding overfitting.

3. Evaluation

The effectiveness of each model was measured using a range of metrics:

Performance Metrics:

Accuracy, Precision, Recall, Specificity, and F1-Score were calculated to assess model quality.

Confusion Matrices:

Provided a detailed view of true positives, true negatives, false positives, and false negatives for each model.

4. Technologies and Tools

A variety of technologies were leveraged to implement and evaluate the solution:

Languages: Python served as the primary language for development.

Libraries:

Data Processing: Pandas, NumPy

Visualization: Matplotlib, Seaborn.

Machine Learning: Scikit-learn, XGBoost, LightGBM.

Frameworks:

Flask was used for backend integration, enabling potential deployment of the model as a web application.

Database:

PostgreSQL provided robust options for data storage and retrieval, supporting scalability.

This structured methodology ensured a comprehensive approach to solving the problem, from data preparation to model deployment, offering a reliable framework for traffic accident prediction.

RESULTS

The performance of the machine learning models was evaluated using key metrics such as accuracy and F1-score. The summarized results are as follows:

K-Nearest Neighbors (KNN)

Accuracy: 70%

Precision:0.510101

Recall: 0.504791

Specificity: 0.504791

F1-Score:0.485763

Balanced performance between precision and recall.

Support Vector Machine (SVM)

Accuracy: 74%

Precision:0.37156

Recall: 0.493902

Specificity: 0.493902

F1-Score: 0.848168

Particularly robust for smaller datasets due to its ability to find optimal decision boundaries.

Random Forest

Accuracy: 75%

Precision:0.745455

Recall: 0.5

Specificity: 0.5

F1-Score: 0.854167

Best-Performing Model

Random Forest emerged as the standout model, achieving the highest accuracy and F1-score among all candidates. Its ability to efficiently handle a wide range of features and deliver quick, accurate predictions made it the most suitable choice for this use case.

CONCLUSION

The "Traffic Accident Prediction" project successfully developed a machine learning-based system capable of forecasting the likelihood and severity of traffic accidents. Through meticulous data preprocessing, model optimization, and evaluation, the project delivered meaningful insights and practical predictions.

Key Takeaways

Influential Factors: Environmental and driver-related factors, such as weather conditions, road type, and alcohol consumption, were identified as significant contributors to accident severity.

Model Performance: Random Forest emerged as the most effective model, demonstrating the value of advanced boosting algorithms in predictive analytics.

Impact

This predictive system has the potential to make a tangible difference in road safety:

For Transportation Authorities: The model can help identify high-risk conditions, enabling targeted interventions to prevent accidents.

For Policymakers: Insights derived from the predictions can guide data-driven policy decisions aimed at improving road safety.

Future Directions

Real-Time Predictions: Incorporating real-time data, such as live traffic updates and weather conditions, could make predictions more dynamic and actionable.

Exploring Deep Learning: Leveraging deep learning models might further enhance predictive accuracy and adaptability.

This project demonstrates how machine learning can address critical real-world challenges like traffic accident prevention. By enabling smarter and safer transportation systems, it underscores the transformative potential of technology in saving lives and improving public safety.