Accident Probability Analysis from Road Conditions

Abdullah Salih Öner
Comp. Eng. Dept., Eng. Faculty
Gazi University
Ankara, Turkiye
riyadlioner00@gmail.com

Buğra Burak Başer
Comp. Eng. Dept., Eng. Faculty
Gazi University
Ankara, Turkiye
bugraburakbaser@gmail.com

Melike Beria Ayas
Comp. Eng. Dept., Eng. Faculty
Gazi University
Ankara, Turkiye
melikeberiaayas@gmail.com

***Abstract*—Traffic accidents are an important issue affecting human life and predicting the trend of these accidents is critical for accident prevention. In this study, we will try to predict the accident trend with machine learning algorithms using traffic accident data collected in the USA between 2016 and 2023. Decision Trees, Random Forest, Logistic Regression, Support Vector Machines, Nearest Neighbor and Naive Bayes techniques will be used and the results will be evaluated with Precision, Recall and F1-Score metrics. This study emphasizes the importance of machine learning algorithms in accident propensity prediction.**

***Index Terms*—data mining, random forests, accidents, svm, correlation**

I. Introduction

Traffic accidents are seen as a major public safety problem that both threatens human life and causes economic losses. Millions of accidents occur worldwide every year, resulting in serious injuries and loss of life. Understanding the causes of traffic accidents is critical to reducing these incidents and improving road safety. Therefore, analyzing and predicting the conditions under which accidents occur can be of great benefit for both traffic management and urban planning.

This study aims to develop a model to predict the probability of traffic accidents using data on accident frequency, weather conditions and time of day. This model will help to identify particularly risky periods and regions and will allow measures to be taken to prevent accidents. It is planned to predict the probability of accidents by using data mining algorithms.

II. Problem Definition

In addition to careless drivers and speed demons, traffic accidents can be influenced by many other factors such as road conditions, weather and time of day. In order to prevent accidents, knowing the probability of these accidents in advance can save lives by increasing the frequency of the measures to be taken. The main objective of this study is to analyze the relationship between road conditions (weather, time of day, accident density, etc.) and traffic accident occurrence trends and to draw conclusions.

Understanding how these conditions affect crash frequency will allow for predicting high-risk situations and making recommendations for safety improvements. In this study, models will be developed to identify when and where traffic accidents are likely to occur, focusing on environmental and temporal factors, using historical accident data.

III. Methods

In this study, various machine learning models will be used to analyze and predict traffic accident trends. Within the scope of the project, we aim to determine the best approach by analyzing the performance between these models.

A. Dataset

The US Accidents (2016 - 2023) dataset contains approximately 7.7 million traffic accident records from 49 US states. The dataset is collected through various APIs that provide real-time traffic incident information from sources such as the US Department of Transportation, law enforcement agencies, traffic cameras and road sensors. The dataset contains detailed information such as accident locations, times, weather conditions, and road infrastructure, and is well suited for analyzing and predicting traffic accidents based on road conditions, weather, and time of day.

B. Valuable Variables in the Project

The variables in our project include environmental factors such as the start and end of the accident, the locations where the accident occurred, and weather conditions. Road infrastructure variables such as traffic lights, intersections, and crossings will also be analyzed as factors affecting the probability of accidents occurring. Also, the severity of the accident is one of the main variables that the model tries to predict.

C. Data Exploration and Preprocessing

Key variables will include environmental factors (e.g., weather, temperature, visibility), road conditions (e.g., traffic signals, intersections), and accident details (e.g., severity, time, location). Missing data will be managed through list-based extraction or substitution, and outliers will be handled using Z-score and IQR. Feature engineering will derive new variables like ”time of day” and ”season” [1].

D. Models

- **Accident Severity Prediction:** Models such as logistic regression, decision trees, and random forests will be applied. Decision trees will provide clear insights into factor relationships, while random forests will improve accuracy by reducing overfitting. SVM will be used to model non-linear data relationships, and Naive Bayes will efficiently handle large datasets for predicting accident probabilities.
- **Density Zones Detection:** Kernel Density Estimation (KDE) and clustering algorithms (K-Means and DBSCAN) will be employed to identify high-density accident zones. Moran's I test will assess the spatial concentration of accidents [2].
- **Accident Trends by Time:** Time-series analysis with ARIMA [3] and seasonal decomposition will evaluate accident risks across different times of the day and year, helping to reveal when accidents are most likely to occur.

Model performance will be evaluated using accuracy, precision, recall, and ROC-AUC metrics, with a focus on minimizing false negatives in accident predictions.

E. Evaluation Methods

In this study, various evaluation methods will be used to analyze traffic accident probabilities. Accuracy measures the overall success of the model by the proportion of correct predictions. Precision evaluates the ratio of correct positive predictions to total positive predictions. Sensitivity shows how many true positives are correctly predicted and aims to minimize false negatives. The F1-Score offers a balanced performance of precision and sensitivity. The ROC Curve and AUC Score show the model's ability to discriminate between classes. The Confusion Matrix provides detailed information on correct and incorrect predictions. These methods are important because accidents are rare and critical events.

IV. Time Plan and Roles

TABLE I
Time plan of the Project

Date	Phase
01.11.2024 – 07.11.2024	Data Cleaning
08.11.2024 – 14.11.2024	Data Analysis and Preparation
15.11.2024 – 05.12.2024	Modeling
06.12.2024 – 12.12.2024	Comparison of Models
13.12.2024 – 19.12.2024	Evaluation of Results
20.12.2024 – 25.12.2024	Demo and Presentation Preparation
20.12.2024 – 25.12.2024	Final Report Preparation
25.12.2024	Code Submission

Buğra Burak Başer, Melike Beria Ayas and Abdullah Salih Öner will take an active role in every stage of the project. Each team member will collaborate by taking responsibility for tasks such as data cleaning, analysis, modeling, model comparisons, evaluation of results and presentation preparations. The phases of the project are shown in Table 1.

V. Backup Plan

- **Data Loss or Access Issues:** Data will be backed up with cloud storage (Google Drive, AWS). If access to the data is completely lost, USDOT and national traffic databases will be used as an alternative [4].
- **Model Failure or Performance Issues:** Models may underperform or overfit. More powerful algorithms such as XGBoost and LightGBM will be tried [5], and hyperparameter optimization (Grid Search, Random Search) will be applied [6]. Cloud computing platforms (AWS, Google Cloud) will be used for large data sets [7].
- **Problems in Time Series Analysis:** If seasonality and trend components cannot be decomposed correctly, alternative models such as SARIMA or Exponential Smoothing will be used [8]. In addition, decomposition techniques such as LOESS and STL will be preferred [9].
- **Problems in Spatial Analysis:** Dense datasets may give erroneous results. If K-Means or DBSCAN are insufficient, alternative algorithms such as Hierarchical Clustering and Mean-Shift will be applied [10]. Datasets will be divided into smaller areas by analyzing on the basis of geographical regions of the USA.
- **Inadequacy of Weather Data:** For missing or inaccurate weather data, data will be completed and verified from sources such as NOAA and WeatherAPI [11], [12].
- **Visualization Issues:** Large data sets may cause performance issues. To improve speed, data sampling techniques will be applied, and visualization will be done using tools like Plotly, Tableau, or PowerBI.

VI. Github Repository

The source code and related materials are available at: https://github.com/burakbasher/team12-accident-probability

References

[1] R. J. Hyndman and G. Athanasopoulos, "Forecasting: Principles and Practice," OTexts, 2018.

[2] P. A. P. Moran, "Notes on continuous stochastic phenomena," Biometrika, vol. 37, no. 1/2, pp. 17-23, 1950.

[3] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, "Time Series Analysis: Forecasting and Control," 5th ed., Wiley, 2015.

[4] "USDOT Traffic Data, U.S. Department of Transportation," [USDOT], Available: https://www.transportation.gov/data.

[5] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 785-794, 2016.

[6] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," Journal of Machine Learning Research, vol. 13, pp. 281-305, 2012.

[7] "AWS Cloud Computing Services," [Amazon Web Services], Available: https://aws.amazon.com.

[8] R. J. Hyndman and G. Athanasopoulos, "Forecasting: Principles and Practice," OTexts, 2018, Chapter on Seasonal ARIMA Models.

[9] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning, "STL: A Seasonal-Trend Decomposition Procedure Based on Loess," Journal of Official Statistics, vol. 6, no. 1, pp. 3-73, 1990.

[10] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," 2nd ed., Springer, 2009.

[11] "NOAA National Weather Service Data," [NOAA], Available: https://www.noaa.gov.

[12] "WeatherAPI: Accurate Weather Data for Developers," [WeatherAPI], Available: https://www.weatherapi.com.