Final Project: Proposal

Rajan Grewal, Ibrahim Ali, Arshdeep Chhokar, Alex Sichitiu

1. Research questions

List 3 questions that you intend to answer (1 point)

- What times of the day do the most car crashes happen?
- Are car crashes more common on weekdays or weekends?
- Will a driver get in a car crash in the next 5 years?
- Which weather patterns correlate the most with car crashes?

2. Dataset utilization

List all the datasets you intend to use (1 point)

Dataset: https://www.kaggle.com/datasets/saurabhshahane/road-traffic-accidents
Description: The data is collected from Addis Ababa, Ethiopia. It is from the years 2017-2020. It has 32 features and contains a plethora of information about the driver, weather, and vehicle.

3. Methodology

Give us a rough idea on how you plan to use the datasets to answer these questions. (2 points)

The dataset that we are using was obtained from Kaggle.

We have to conduct EDA in order to understand key characteristics of the data. We will also need to uncover any missing values, underlying distributions of the categorical and numerical data, as well as trends in traffic incidents over time.

Data cleaning will be done to handle missing values by either removing rows that have many missing fields or filling values with a statistic such as the mean or median. The dataset separates the day of the week and time of the accident, so we will need to consolidate this into a single column with the proper format. We will also need to normalize categorical variables such as the vehicle-driver relation and educational level into numerical values using techniques such as one-hot-encoding. We will also need to remove duplicate records from the dataset if they exist.

Data integration can be performed from additional data sources such as a weather dataset or a traffic density dataset by joining them on the date-time and location attributes. This will enhance our insights by providing us with additional information to analyze. We will have to do some work to ensure that the data is consistent and accurate.

The type of analysis that we wish to do is a car accident analysis that uses various factors to determine whether or not a car accident will occur. We will do this by developing a machine learning model which uses a clustering algorithm such as K-means to determine which attributes determine a car accident. In order to evaluate this, we will use metrics such as precision, recall, and F1-score for our predictive machine learning model. We will also perform statistical analyses in order to produce visualizations such as histograms, pie charts, and scatter plots.

The final data product that we will be producing is an interactive dashboard which visualizes the different trends in traffic incidents, hotspot regions, and numerical values representing the total number of accidents, casualties, and injuries reported.

4. Expected impact

Think about that once your project is complete, what impacts it can make. Pick up the greatest one and write it down. (1 point)

The biggest impact of this project is the potential to improve road safety. By analyzing patterns in car crashes, like when and where they happen most often, and what factors contribute to them, we can help create safer roads. For example, if we find that crashes are more common during certain weather conditions or at specific times of the day, authorities can take steps like adjusting traffic signals, raising public awareness, or improving road design. In the long run, these insights could help reduce accidents and save lives.

5. Potential challenges

Identify any anticipated obstacles and how you plan to address them. (1 point)

One challenge we might face during this project is dealing with missing or incomplete data, which could affect the accuracy of our analysis and predictions. To handle this, we plan to fill in missing numerical values with the mean or median, and for categorical fields, we'll use the most common value. If some records are missing too much important information, we'll remove them to keep the dataset reliable. Another potential issue is integrating external datasets, like weather data, since date formats or location details might not match perfectly. To address this, we'll standardize the date and time formats and make sure the location markers align correctly. Lastly, when training the machine learning model, there's a risk of overfitting, but we'll use techniques like cross-validation and parameter tuning to keep the model accurate and reliable. By planning ahead for these challenges, we hope to avoid major roadblocks and get the best possible insights from our data.