**Spring 2023**
**IST 718 Group Project Proposal**
**Group 1: Harim Shin, Kishan Rathor, and Rich Granato**

# Using Machine Learning to Reduce the Number of Daily Car Accidents

## 1. Objective

In 2021, about 43,000 people died in the United States due to car accidents. (Forbes, 2023) The goal of this project is to use supervised machine learning to provide both short and long-term insights to reduce the number of car accidents each year and ultimately reduce the number of resulting fatalities. Using five years of car accident data, we will predict the severity rating of a given car crash, which is generally a 1-4 rating assigned by-hand after a car accident scenario has already concluded. In the short term, we can notify various companies of highly severe car accidents (severity = 3 or 4) so proper roadblocks and civilian communications can be carried out. In the long term, we can try to educate young drivers about which combinations of weather, seasonal, and locational variations result in the greatest risk of getting into a car accident on any given day. There are very likely variables that are not included in our dataset that contribute to the likelihood of getting into a car accident, such as drunk driving and distracted driving, but as long as there are other factors that contribute, we can try to help the average driver who follows the rules not feel as though they are risking their lives each time they get into their car.

## 2. Data Set Description

### a. Overview / Description

The dataset we will use for this project provides many different types of variables that we can utilize as features for our prediction model.

First, the data has information on the timeframe of the car accidents. It tells when each accident started and ended.

Second, the data provides rich information on the geographical location where the accidents happened. We can find the latitude/longitude of both the start and the end points of the accident, and the distance measured by those two points as well. In addition, geolocation information of the accidents are provided, such as street number, name of the street, the relative side of the street, the city, the county, the time zone, and the airport code.

Third, we can check the different aspects of the weather when each accident occurred, such as the wind, such as wind chill, wind direction, and wind speed. We can also obtain other information like temperature, humidity, air pressure, visibility, precipitation, and weather conditions. Most of the weather-related variables are continuous, so we could verify how those variables are related to each other by deriving a correlation matrix and a pair plot for each one.

Fourth, we can learn the locational characteristics related to the traffic for each car accident. The dataset tells if there is any amenity, bump/hump, crossing, give way sign, junction, no exit sign, railway, roundabout, station, stop sign, traffic calming, traffic signal, or turning loop. All the variables related to the locational characteristics are boolean-type (True/False) variables.

Finally, the variables including the information related to the period of day are included in the dataset. The variables tell if a car accident occurred during the day or night based on the sunrise/sunset, civil twilight, nautical twilight, or astronomical twilight.

### b. The shape of the dataset

The dataset has 47 variables and 2,845,342 data points in total. We had to take a sample from this dataset for summary statistics due to its size which is greater than 1GB.

### c. Sample predictors (does not need to be an exhaustive list)

- Distance(mi): The length of the road extent affected by the accident.
- Precipitation(in): Shows precipitation amount in inches, if there is any.
- Roundabout: A POI annotation that indicates the presence of a roundabout in a nearby location.
- Sunrise_Sunset: Shows the period of day (i.e. day or night) based on sunrise/sunset.

### d. A link to the dataset

https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents?select=US_Accidents_Dec21_updated.csv

### e. Anything interesting or surprising about the data

We have noted that the data only contains characteristics about the environment and the crash itself, and does not contain information about the driver. While there is a lot of valuable information missing about the driver, such as their past driving history and whether or not they were under the influence, the nature of this project focuses on improving situational safety as opposed to trying to change human behavior.

## 3. Preliminary Data Exploration

We explored the dataset before starting to build a prediction model.

First, we showed the probability of a car accident to be severe by state. A car accident is defined to be ‘severe’ when the severity is 3 or 4, which is greater than 2, the most common degree of severity based on the data. As we can verify from Figure 1, in Colorado, Wisconsin, and Illinois, when a car accident occurs, it is likely to be a severe one at approximately 50% of probability. Environmental factors such as weather conditions could cause this pattern, but it is also possible that the criteria with which the police officers determine if an accident is severe can be looser in those states.

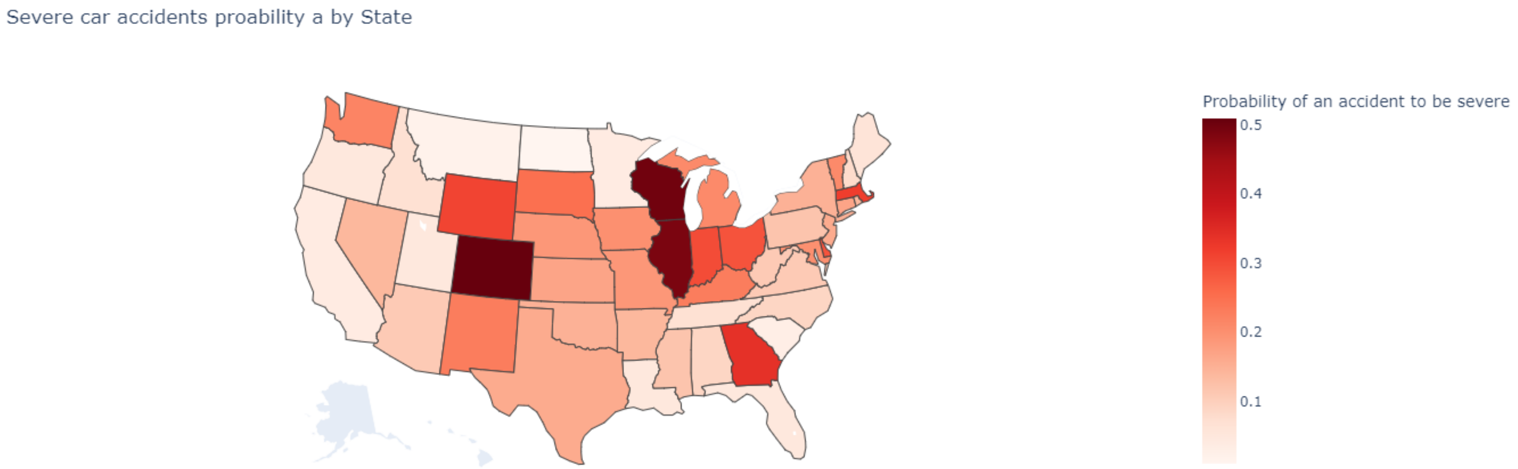
Severe car accidents proability a by State
Probability of an accident to be severe
0.5
0.4
0.3
0.2
0.1

**Figure 1. Severe car accidents probability by State**

Second, we also tested whether there is any serious multicollinearity between the variables related to weather. Figure 2 is the heatmap drawn with the correlation between those variables. We can see that temperature and wind chill are almost perfectly linear, but other variables do not show very serious multicollinearity. Based on the results, we will consider this high correlation between temperature and wind chill when building a prediction model.

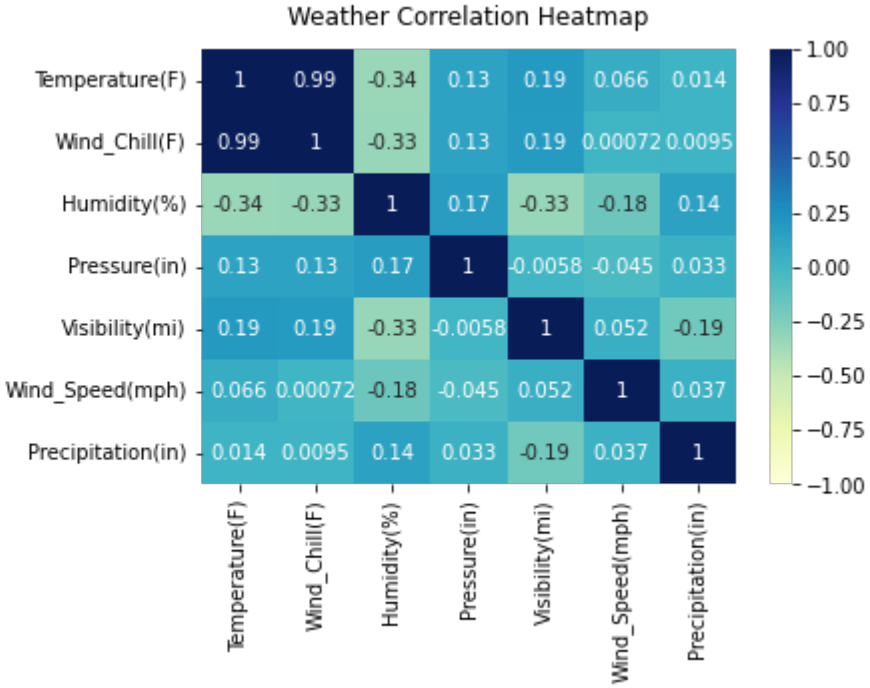

**Figure 2. Heatmap with the correlation between variables related to the weather**

## 4. Predictions

The project has the goal of developing predictive models that can accurately anticipate the severity of car accidents. The purpose of this is to provide valuable insights to inform stakeholders and prevent future accidents. The project has a particular focus on accurately identifying the most serious and problematic accidents. To achieve this, the project will identify patterns or correlations between variables and accident severity. Variables such as weather conditions, location, time of day, and other factors will be taken into consideration. The project will also predict the likelihood of accidents based on various factors. This information could be used to inform drivers on when it may be dangerous to drive. In addition, the project will provide real-time warnings or alerts to drivers, emergency services, or other stakeholders based on predicted accident severity and location. The effectiveness of certain interventions or measures to prevent or reduce accidents will also be assessed. Examples of these interventions include road improvements, traffic signal changes, or public awareness campaigns. Overall, the project aims to use data and predictive models to make our roads safer for everyone.

## 5. Inference

There are two main inferences that we want to draw from our finalized learning model. The first inference is the ability to accurately identify highly severe car accidents right away, and send the information off to both the police and news media outlets. Police would be notified to block off the area of a highly severe car accident and set up a detour route as soon as possible. News media outlets would be informed so they can broadcast immediate communications to the public through television and radio to avoid the location for an estimated period of time. The hope would be that this deters additional traffic accidents at or near the sight of the initial reported accident and protects drivers the moment a car accident has been detected.

The second inference is the ability to identify commonalities between certain types of car accidents and communicate our findings to both educational and automobile industries. Educational industries would be provided tools that can be used to educate young drivers on driving conditions to avoid and actions to perform when driving under these conditions is necessary. Automobile industries that work on self-driving cars can utilize our results to program their vehicles to better identify potentially dangerous driving conditions. Overall, the main goal here would be to use information on prior car accidents to promote the prevention of car accidents in the longer-term, more distant future.

## 6. Non Spark Packages

- chart_studio: Interactive charts and maps for Python, R, Julia, Javascript, ggplot2, F#, MATLAB®, and Dash.

## 7. Citations

Simon, S. (2023, January 27). *How many people die from car accidents each year?* Forbes. Retrieved March 4, 2023, from https://www.forbes.com/advisor/legal/auto-accident/car-accident-deaths/