Car Accident Analysis

Noah Shoap and Jonathan Graham

***Abstract*— This project uses car accident analysis data stored on Kaggle to predict how severe an accident will be based on environmental conditions such as weather and time of day [1]. The data was trimmed and processed using decision trees to find how accurately the given features could predict the severity of an accident. The work done here also generated ideas for how to do more work on the data in the future.**

I. Introduction

From the year 2016 to 2021 there were roughly 34 thousand motor vehicle accidents per year, with 29.5 thousand in 2021. [2]. These accidents caused between 37 and 42 thousand deaths each year. With such an astounding number of deaths it only makes sense to find what underlying causes impact the severity of an accident to reduce the death toll.

This project aims to find a machine learning approach to find a way to predict how sever an accident will be on any day given certain environmental conditions based on previously collected data.

II. Data

To find data for this project the team searched Kaggle for a suitable data set. After some searching a data set was found containing 7.7 million vehicular accidents, with a ranking for severity, over the years 2016 to 2023.

A. *Trimming the data*

Once the data was found it had to be trimmed to a more usable size and format to allow for machine learning methods to be applied. The first step was to reduce the number of entries to a workable number. The number 20,000 was chosen as it provides a large amount of data to work with, but it is also still small enough to rework easily.

The second step was to remove any entries that contained NA entries, followed by dropping the first column which was just used to number the entries in the table for easy referencing.

After those issues were resolved, the data was further trimmed to remove any column that stored data as a string, such as the source of the data or the state it occurred in. Some of the data that was removed, such as the state or the column labelled "Weather Condition", was removed because the data there is covered by other columns, such as the latitude/longitude and wind conditions/precipitation.

B. *Finalized Data*

After the trimming, the data was cut down to roughly 11 thousand usable entries with 31 different features. The chosen features describe the time of day, the weather conditions, and any potential distractions around the accident. These distractions range from speed bumps to stop signs to any amenities in the vicinity that may have signs to read etc.

Once the data was fully timed and prepped the team did an 80 20 split on the data for training and testing the chosen machine learning algorithm.

The features used to describe the data and a description of that feature are listed in Table 1 below.

Table 1 Chosen Features

Feature	Description
Start_Lat	Shows latitude in GPS coordinate of the start point.
Start_Lng	Shows longitude in GPS coordinate of the start point.
Temperature(F)	Shows the temperature (in Fahrenheit).
Wind_Chill(F)	Shows the wind chill (in Fahrenheit).
Humidity(%)	Shows the humidity (in percentage).
Pressure(in)	Shows the air pressure (in inches).
Visibility(mi)	Shows visibility (in miles).
Wind Speed	Shows wind speed (in miles per hour).
Precipitation(in)	Shows precipitation amount in inches, if there is any.
Amenity	A POI annotation which indicates presence of amenity in a nearby location.
Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.
Crossing	A POI annotation which indicates presence of crossing in a nearby location.
Give Way	A POI annotation which indicates presence of give_way in a nearby location.
Junction	A POI annotation which indicates presence of junction in a nearby location.
No Exit	A POI annotation which indicates presence of no_exit in a nearby location.
Railway	A POI annotation which indicates presence of railway in a nearby location.
Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.
Station	A POI annotation which indicates presence of station in a nearby location.
Stop	A POI annotation which indicates presence of stop in a nearby location
Traffic Calming	A POI annotation which indicates presence of traffic_calming in a nearby location.
Traffic Signal	A POI annotation which indicates presence of traffic_signal in a nearby location.
Turning Loop	A POI annotation which indicates presence of turning_loop in a nearby location.
Sunrise Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.
Civil Twilight	Shows the period of day (i.e. day or night) based on civil twilight.
Nautical Twilight	Shows the period of day (i.e. day or night) based on nautical twilight.
Astronomical Twilight	Shows the period of day (i.e. day or night) based on astronomical twilight.
Month	Month of the accident.
Start Day	The day the accident started on.
End Day	The day the accident is over on.
Start Seconds	The second of the day the accident starts.
End Seconds	The second of the day the accident is over.

III. ANALYSIS

A. INITIAL IDEAS

In order to analyse the data, the team wanted to run some sort of machine learning algorithm on the data. Several ideas were brought up, from decision trees to a multilevel perceptron.

Through examination of the data and discussion on how complicated the approach needed to be the team settled on running decision trees on the data.

B. HYPERPARAMETERS

Once it was decided to use decision trees to predict the severity of future accidents, the hyperparameters had to be fine-tuned to find the best tree to use. The main hyperparameter that was changed was the maximum depth the tree was allowed to grow to. The team generated trees with maximum depths of 1, 2, 4, 8, 16, 32, 64, 128, 256, and no maximum.

C. Testing the Models

Once the team had selected the trees to test, the depths needed to be tested for the greatest accuracy score. For each depth the team generated a decision tree fit to the training data and then predicted the labels for the testing data. The predicted labels were then compared to the actual labels to give an accuracy score for each max depth.

Below in Table 2 the different depths and their accuracy scores are compared from a chosen running of the decision trees.

Table 2 Decision Trees

Max Depth	Actual Depth	Accuracy Score
1	1	0.5890
2	2	0.5890
4	4	0.7123
8	8	0.8356
16	16	0.7808
32	15	0.7808
64	16	0.7945
128	17	0.7945
256	15	0.8082
None	15	0.7808

An interesting note: despite the max depth of each tree being set, the deepest tree that was generated was of depth 17 on the tree with a max depth set at 128.

D. Chosen Tree

The different depths of decision trees were generated and fit to the data multiple times with their accuracy scores being calculated each time. After doing this the team noticed that a decision tree with a max depth of 8 consistently did better than any of the other trees at predicting the severity of an accident based on the given 31 features.

Figure 1 on the next page is one of the trees generated with a max depth of 8 and an accuracy score matching Table 2.

Looking at the tree, the first 4 decisions made are based off the longitude or latitude. This means that the most important decider of the severity of an accident, according to this decision tree, is where it is in the country. To reinforce this idea, later decisions in this tree are also based on positional data. Other factors that are used to predict the severity include the time of day, the weather, and a few road features, such as if the accident is at a junction or not.

IV. Results

Looking at all the data collected and the results of the decision tree modeling there are a few concepts to take away. Firstly, using a decision tree to predict the severity of an accident seems like a reasonable approach to take. Getting an accuracy of roughly 83% means that using that model for predicting how severe an accident is going to be will be right about four out of five times. The prediction of accident severity could be used by law enforcement to determine where to set police cars to slow down traffic as well as to be closer to potentially deadly accident sites. Hospitals could also use this to know when to staff more EMS and emergency room personnel.

Another take away is how the location in the country can determine how severe an accident is likely to be. People who are looking to move to safer places could use this data as an additional data point to find those safer places to live. Realtors could use this data to market the safety of their locations to potential buyers as well.

A last take away from this data analysis is that whereas weather does impact the severity of an accident, it does not have as much of an impact as where you wreck. Any smart driver with access to this data would use it to determine where to be more aware of their surroundings and practice better defensive driving practices.

V. Major issues

The main issue in this project was trimming the data to meaningful and useable data. The initial data file was too large to analyze as a whole, and thus it had to be cut into smaller pieces to determine how much data was actually needed. After that the data still had to be trimmed and adjusted, turning True/False into 1/0 etc., in order to apply any machine learning algorithms to it.

VI. Generated Ideas

After having worked with the data and the results from the machine learning analysis of it, a few ideas emerged for future work.

One idea is to look at accident data based on the population density of a given area. The location having such a large impact on the severity of an accident lead the team to wonder if it was because of the way drivers drive in those areas, or if it had to do with how many drivers were in the area to be impacted. Perhaps the location is more indicative of the population density than the driving habits of that population.

In order to research this further there would need to be more data collection done on population density, number of accidents, and severity of those accidents. That would allow for the team to determine if population density affects the severity as well as rate of accidents. The results of that analysis could then be compared to the results found here to try and determine whether the population density of an area or the driving habits of an area are more telling in the severity of accidents.

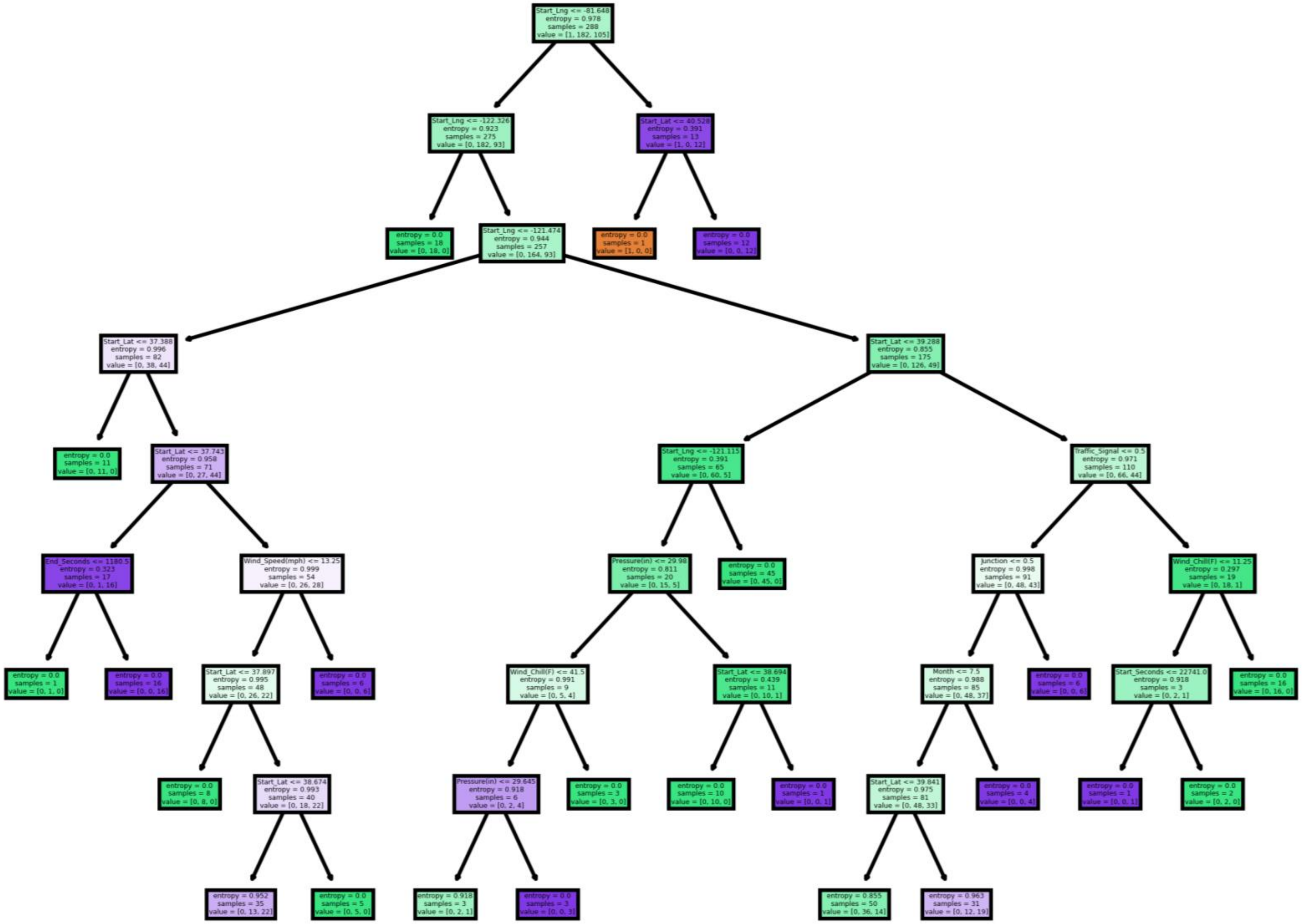

Figure 1 Decision Tree Max Depth 8

Another idea to research in the future is how location and weather affects the rate of accidents in any given area. Those results could then be compared to the results found here to determine how much of an impact weather has compared to the location of an accident. It is generally accepted that poor weather conditions lead to more accidents, but this data suggests that it doesn't impact the severity of those accidents as much as other factors. Getting more data and analyzing it as suggested above could potentially answer those questions better and more empirically.

A final idea that was generated from this data analysis is to trim this data into two major categories: weather and location. Then those two major categories could be sued to see if one of them separate from the other is better at predicting the severity of an accident or if it is the combination of the two that gives the best result.

VII. ORGANIZATION OF WORK

Noah and Jonathan split the work for this assignment into five major parts as shown below in Table 3. The first part was the acquisition of data, which Noah finished by October 1st. The second part was the trimming and filtering of the data, which Jonathan finished by November 15th. The third part was analyzing the data using machine learning algorithms, which Noah lead and was completed by November 20th. The final portion was writing this paper, which Jonathan lead and was jointly finished by November 27th. The final presentation of the project was given by both Jonathan and Noah in class on November 30th.

Table 3 has what portion of the project is being delivered, who was responsible for its completion, and when it was completed.

Table 3 Org Chart

Deliverable	Owner	Completion Date
Data Acquisition	Noah	October 1st
Data Filtering	Jonathan	Nov. 15th
Data Analysis	Noah	Nov. 20th
Paper	Jonathan	Nov. 27th
Presentation	Both	Nov. 30th

REFERENCES

[1] "US Accidents (2016 - 2021)," www.kaggle.com. https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents?resource=download

[2] IIHS, "Fatality Facts 2017: Yearly snapshot," IIHS-HLDI crash testing and highway safety, 2017. https://www.iihs.org/topics/fatality-statistics/detail/yearly-snapshot