Determination of major causes of traffic accidents in the US and use of Machine Learning to predict the severity of such

Leonardo Gama García – A01366891

B.S. in Mechatronics Engineering

Tecnológico de Monterrey

Abstract

As one of the leading causes of death both in the United States of America and worldwide, it is important to understand the factors involved in traffic accidents, as well as to determine the severity of those accidents in order to avoid dangerous situations on the road and have more caution when those factors are present.

I. Introduction

According to data from the World Health Organization from 2021, it is estimated that around 1.3 million people die each year as a result of road traffic accidents worldwide, while between 20 and 50 million people suffer non-lethal injuries, many of which may result in a disability to the victim. Road traffic accidents are estimated to be the world's 8th leading cause of death, while they are the leading cause of death to people aged from 1 to 54 in the United States of America.

Not only is this a health hazard, but an economic one as well. In 2010, the economic cost caused by motor vehicle crashes totaled to $247 billion (USD), which includes costs such as productivity losses, medical costs, legal and courts, emergency services, insurance, property damage and so on. As the number of vehicles in operation in the US increases, reaching over 283.8 million vehicles in the third quarter of 2021, the probability of road traffic accidents occurring increases as well.

The objective of this implementation is to visualize and understand the factors involved in road traffic accidents in the US using data collected in real time from February 2016 to December 2021 using multiple Traffic APIs, in order to prevent them using information analysis, as well as classification machine learning algorithms, in order to predict the severity of such accidents to take the corresponding measures, and to compare which method is the best to implement for the application.

II. Theoretical framework

Information analysis

Information analysis is defined as the systematic process of discovering and interpreting information. It consists of tasks such as finding the relationships between the raw data, grouping in categories and adding context, which is interpreting what the values are telling us, using tools such as ratios or indicators. Once information has been analyzed, knowledge can be generated, which adds interpretation to the analysis and allows insight gain in order to take decisions.

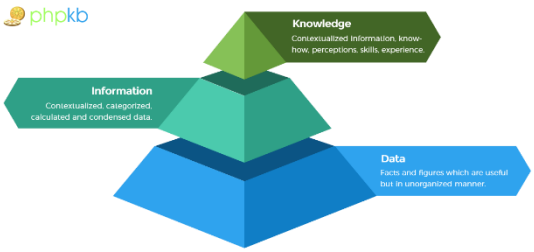

Fig. 1: Data, information and knowledge definitions (PHPKB, 2021)

K nearest neighbor

The K nearest neighbor algorithm is a classification method used to determine what group a data point belongs to depending on the data points nearest to it.

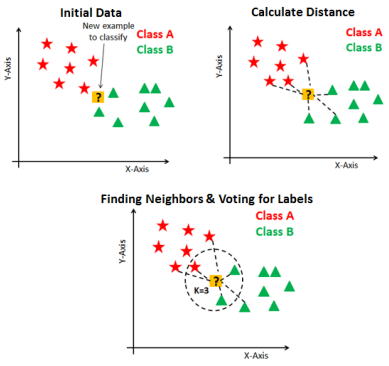

Fig. 2: K nearest neighbor algorithm explanation (DataCamp, 2018)

Gaussian Naïve Bayes

Gaussian Naïve Bayes is a simple probabilistic classification algorithm that follows the Bayes Theorem to categorize the probability of events, using the following formula:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

In it:

- A and B are two events
- P(A|B) is the probability of event A provided event B has already happened.
- P(B|A) is the probability of event B provided event A has already happened.
- P(A) is the independent probability of A
- P(B) is the independent probability of B

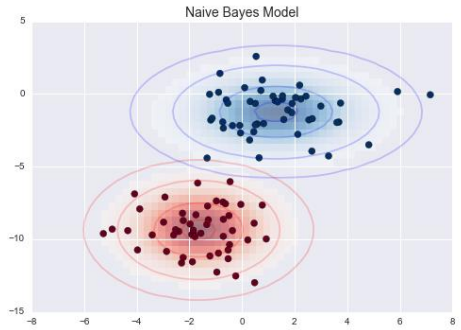

Fig. 3: Naïve Bayes Model (Jake VanderPlas)

This algorithm assumes that each features contributes equally and independently from each other to the predicted output, which most of the time can make accurate predictions,

and can be beneficial to use in large datasets.

Decision tree

In data science, a decision tree is a type of classification algorithm used to categorize or predict based on how a previous set of questions were answered by the dataset.

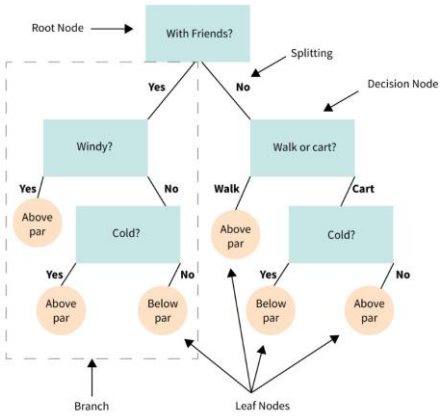

Fig. 4: Decision tree diagram (Master's in Data Science, 2022)

The visual presentation of the results imitates the human way of thinking, which makes it easier for data scientist to analyze, understand and interpret the information.

III. Implementation

Information analysis

It is important first to analyze the information given in the dataset in order to grasp the most frequent characteristics that make up for the road traffic accidents, since knowing this information helps both drivers and pedestrians to avoid or to be more aware in situations in which accidents are more likely to occur.

First, having a visual representation of the number of accidents in the US is an important tool in order to see which are the places of the country in which most accidents occur. Therefore, a map is made using the start longitude and altitude of each accident by using a scatterplot.

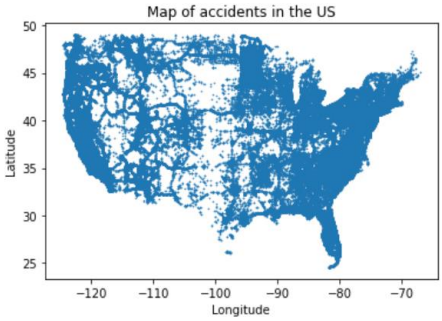

Fig. 5: Map of accidents in the US

The map is an indicator of where the accidents mostly occur thanks to the density of points in different states, such as in the state of California and the East Coast of the country, although it is not accurate enough. Therefore, bar plots are used in order to obtain both the states as well as the cities with the highest amount of road traffic accidents with actual numbers.

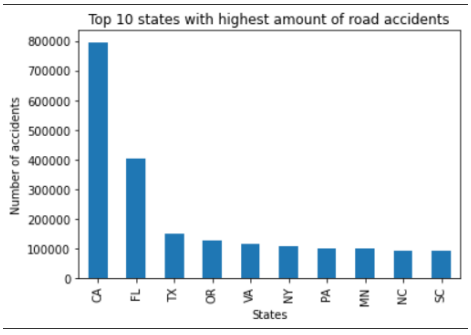

Fig. 6: Bar plot of 10 top states with most road traffic accidents

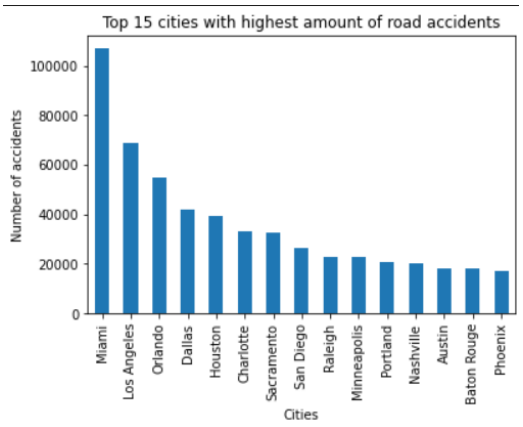

Fig. 7: Bar plot of 15 top cities with most road traffic accidents

As it can be observed from figure 5, the state with the highest amount of motor vehicle accidents is California, followed by Florida, Texas, Oregon and Virginia. Additionally, the city with most road traffic accidents is Miami, with Los Angeles, Orlando, Dallas and Houston being the second, third, fourth and fifth cities with most accidents respectively, as it can be seen in figure 6.

Not only is the location of the accidents important, but also the time in order to grasp at what time and what days it is more likely to be involved in an accident.

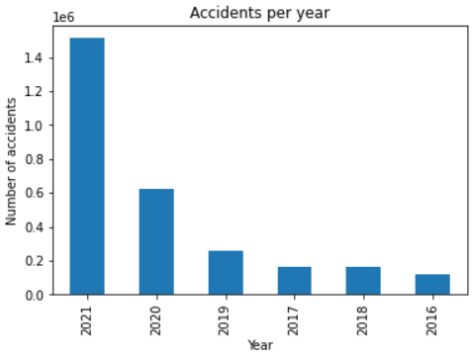

Fig. 8: Bar plot of accidents per year

First, a bar plot of the accidents per year shows an increase in road traffic accidents each year since 2019. This tendency is likely to continue due to the ever-increasing number of cars in circulation in the US.

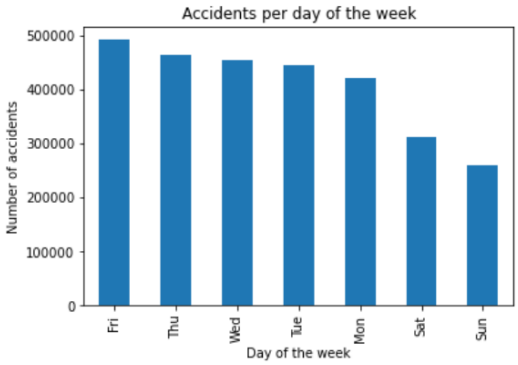

Fig. 9: Bar plot of accidents per weekday

As seen in figure 7, most accidents occur on Friday, while the least number happen on Sunday and Saturday.

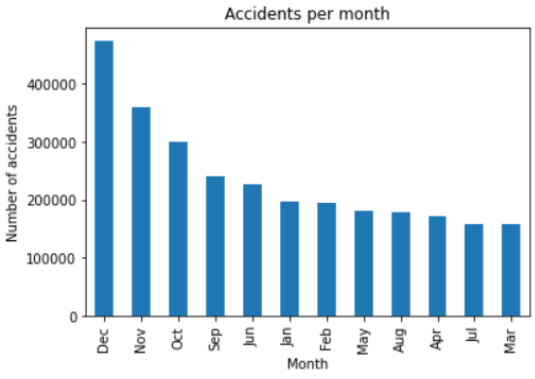

Fig. 10: Bar plot of accidents per month

As for the accidents per month, most accidents happen on the month of December, with November being the second highest and October the third.

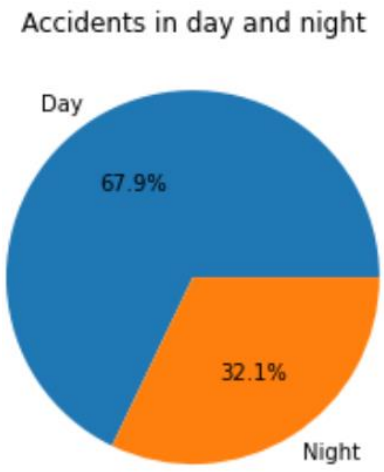

Fig. 11: Pie plot of accidents in day and night

Figure 10 shows that 67.9% of accidents happen during daytime, while 32.1% happen during nighttime. For this plot, the civil twilight parameter was used, since it indicates the period when enough natural light remains that artificial light is not needed.

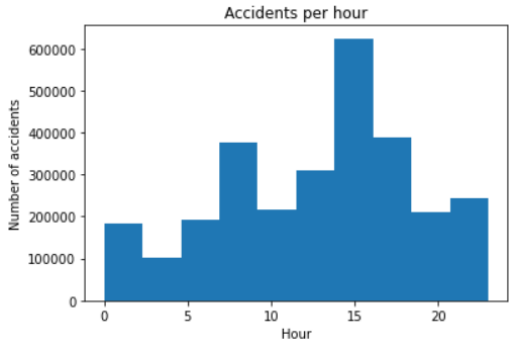

Fig. 12: Histogram of accidents per hour

To be more precise of the moment when accidents are more likely to occur, a histogram shows that most accidents happen between 12pm and 6pm.

Moving on to the characteristics present in accidents, the first one is the side of the road in which the accident occurred. For this, a pie chart is plotted.

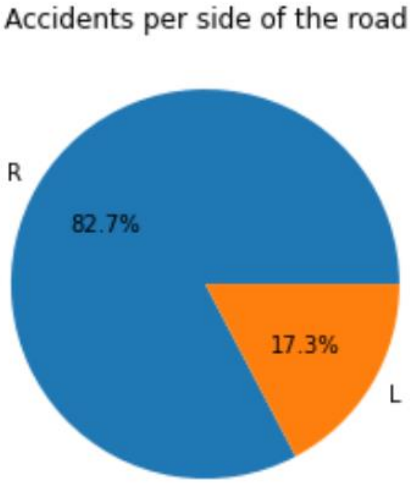

Fig. 13: Pie chart of percentage of accidents on each side of the road

The pie chart shows that 82.7% of the road traffic accidents recorded happened on the right side of the road, while only 17.3% of them where on the left side, which means that it is more dangerous to drive on the right lanes than on the left lanes of the road.

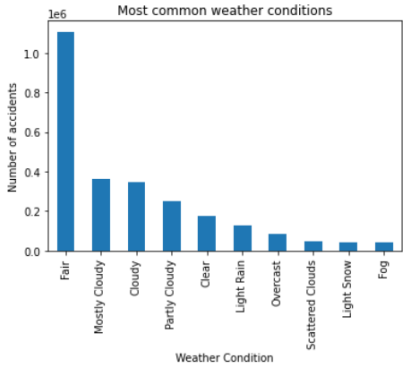

Fig. 14: Bar plot of most common weather conditions

The next is the weather conditions, which shows that most accidents happen when the weather is fair,

followed by mostly cloudy, cloudy and partly cloudy weathers.

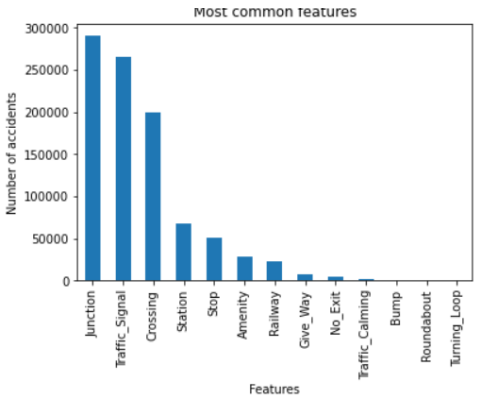

Fig. 15: Bar plot of most common road features present in motor vehicle accidents

Figure 14 shows the most common road features present during road accidents, showing how junctions, traffic signals and crossing are the most present features.

Lastly, since the target is to predict the severity of road accidents in the US, it is important to analyze this data.

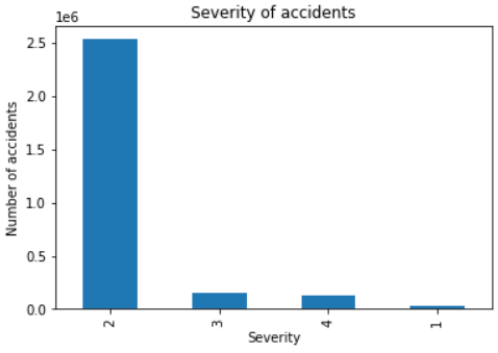

Fig. 16: Bar plot of severity of accidents

Figure 5 shows a bar plot of the number of accidents per severity. This shows that most accidents are severity type 2, which is a medium severity, while there is a big gap with the other classification of severities.

Data preprocessing

Before continuing with any machine learning algorithm, the first step is to preprocess the data in order to simplify the dataset as much as possible, as well as to scale the numbers to have a more proportionate relation between the data.

First, a correlation matrix is plotted in order to see if there is data directly proportional to one another. In it, temperature and wind chill are directly proportional, which means that one of them can be omitted. Also, the start latitude and longitude are completely related to end latitude and longitude of the accident, which means that the end longitude is not relevant. Furthermore, the matrix shows that the turning loop has no values other than false, which makes it irrelevant for the analysis.

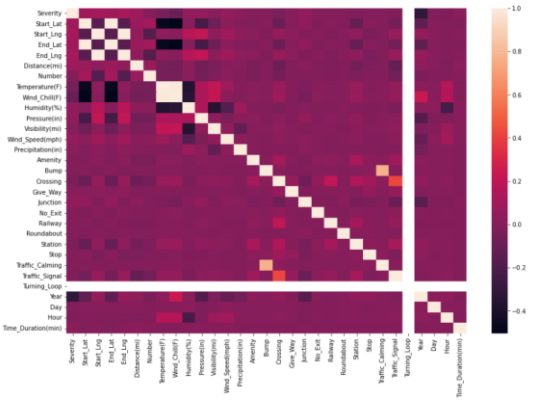

Fig. 17: Correlation matrix of data

Along with wind chill, end latitude and longitude, and turning loop, other features were omitted, since they were not important for the analysis, such as the description of the accidents since the information is not very useful, start and end time since that information is

already featured in separate features, irrelevant features related to the location or other information such as the country, number, state, zipcode, timezone or weather stamp, and every feature related to the definition of day and night except civil twilight, since they all have the same purpose.

Due to the high number of variables in the dataset, plenty of irrelevant features were omitted, such as the zipcode, timezone, country, and so on. Furthermore, string-based features such as the weather conditions or the wind direction were omitted as well, since the machine learning algorithms can't process string variables, and transforming each would complicate the process, as well as risking overfitting. Also, many binary features were not considered, since some had a very low percentage of appearance. Therefore, to maintain a low number of variables, the most significant ones were selected, including:

- Severity (target)
- Distance (mi)
- Temperature (°F)
- Humidity (%)
- Pressure (in)
- Visibility (mi)
- Wind speed (mph)
- Precipitation (in)
- Crossing
- Junction
- Traffic Signal
- Hour

After dropping the features that are not used, the next step is to remove lines of data with missing information, since replacing the data with another value will alter the results of the predictions, as well as to remove duplicates in the dataset.

Then, as seen in the information analysis when analyzing the severity of accidents, it is evident that the dataset is imbalanced, since most of the accidents are severity 2, while the gap with the other severities is very high. Therefore, the solution is to undersample the majority classes so that every classification of severity has the same number of instances as the lowest type, which is severity 1.

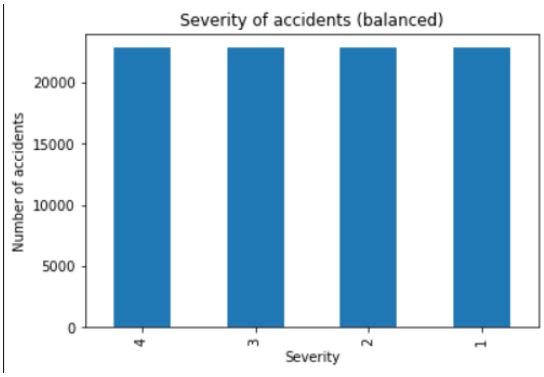

Fig. 18: Balanced target data

Once normalized, the data is divided into the training set and the test set in order to start with the algorithms.

K Nearest Neighbor

The first classification method used is the K Nearest neighbor using manual definitions. When the algorithm is ran, it presents the following results:

```
KNN Accuracy = 0.2488
```

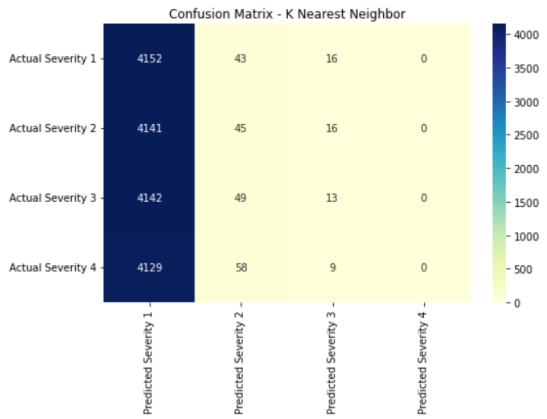

Fig. 19: K Nearest Neighbor confusion matrix

When running the KNN algorithm, it presents an accuracy of around 25%, where the results are highly biased towards only one type of severity. This is because the data doesn't have good distinctions between its groups, which means that it interprets all the classes as only one. This is evident when representing the data in a 2D space using PCA.

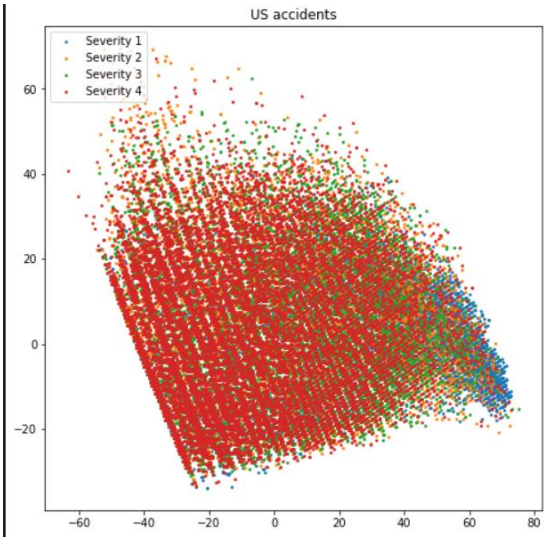

Fig. 20: US accidents dataset PCA visualization

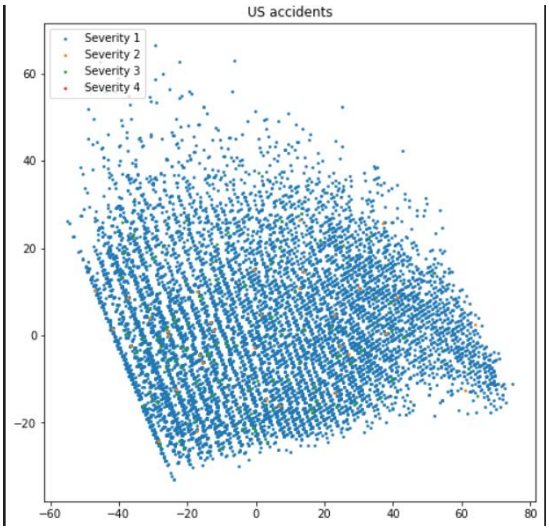

Fig. 21: KNN prediction PCA visualization

With an accuracy so low and very biased predictions, it is evident that K Nearest Neighbor is not the best algorithm to solve this problem.

Gaussian Naïve Bayes

Using the Sklearn library, the first framework used is the Gaussian Naïve Bayes algorithm. When ran using the same features, it gives the following results.

```
Gaussian Naive Bayes Accuracy = 0.3429
```

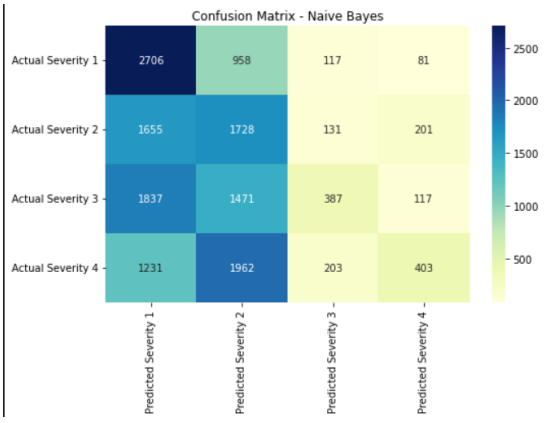

Fig. 22: Gaussian Naïve Bayes confusion matrix

As it can be seen, the accuracy of this algorithm is close to 34%, which although is better than the results from K Nearest Neighbor, it is not good enough to have a viable prediction. This has to do with the same issue encountered with KNN, as the data is not distributed enough for the GNB algorithm to find significant distinctions among the data.

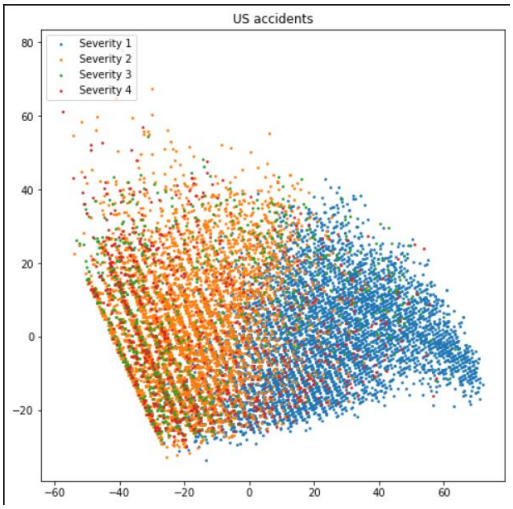

Fig. 24: Gaussian Naïve Bayes confusion matrix

Decision tree

The next classification algorithm developed with Sklearn was the decision tree algorithm. In this algorithm, the feature “Time Duration (min)” was added, since it increased the accuracy of the significantly. Furthermore, the decision tree algorithm works well to automatically identify the most significant features when limiting the depth of the tree.

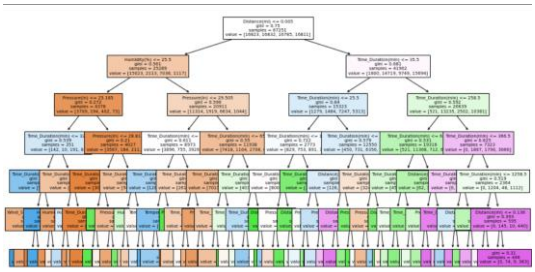

Fig. 25: Decision tree classification algorithm results

The results of the prediction using the decision tree classifier are the following:

```
Decision Tree accuracy = 0.6306
```

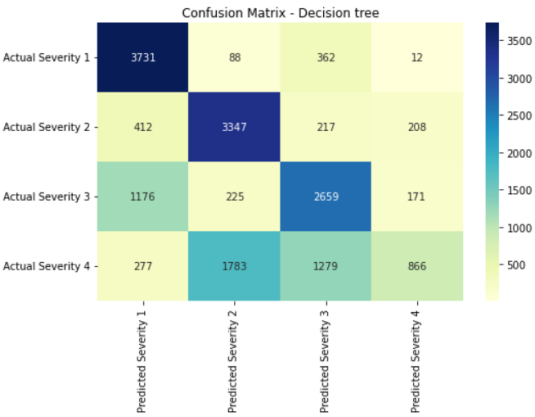

Fig. 26: Gaussian Naïve Bayes confusion matrix

The results show a higher accuracy of up to 63% when limiting the max depth to 6. This accuracy can be increased to nearly 65% when increasing the max depth to 8, and further increasing the depth did not increase the accuracy any further. However, the tradeoff of 2% increase in accuracy was not worth making the algorithm more complex.

As mentioned before, the decision tree is good in finding the most relevant characteristics of a group, which is why

it can make good predictions for severity 1 accidents, since the most significant features such as distance and time duration are what clearly define this type of severity. However, the algorithm still has issues identifying key elements of other severities, such as severity 4.

IV. Results

Once each algorithm has been run, the accuracy of each one can be compared.

Algorithm	**Accuracy**
K Nearest Neighbor	25%
Gaussian Naïve Bayes	34%
Decision tree	63%

Tab. 1: Algorithm accuracy comparison

As it can be seen, the decision tree is the best algorithm for this specific dataset, since it can identify the key features that make up for each severity and make predictions accordingly.

V. Conclusions

This implementation shows that there is no specific machine learning algorithm that is good at solving every problem, but instead there are algorithms that perform better than others depending on the dataset.

The results also show that there are some flaws in this specific road traffic accidents dataset. The first one involves how imbalanced the data is with respect to the severity by classifying most accidents as level 2 severity. This causes the dataset to be impractical and not very informative. A better way to classify each type of accident, such as widening the range of the severity would improve the results and the usability of the dataset, or changing the way severity is classified altogether.

Furthermore, having algorithms with an accuracy so low means that the parameters given by the dataset are not distinct enough to return good predictions. In other words, the data that is being collected with respect to road traffic accidents does not have much influence over the target, which is to know the severity. Therefore, different parameters could be collected in order to have better results.

VI. Sources

Road traffic injuries. (2021, June 21). World Health Organization. Retrieved April 12, 2022, from https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

Road Traffic Injuries and Deaths—A Global Problem. (2020, December 14). Centers for Disease Control and Prevention. Retrieved April 12, 2022, from https://www.cdc.gov/injury/features/global-road-safety/index.html#:%7E:text=Road%

20Traffic%20Injuries%20and%20Deaths%E2%80%94A%20Global%20Problem&text=Road%20traffic%20crashes%20are%20a,citizens%20residing%20or%20traveling%20abroad.

Blincoe, L. J., Miller, T. R., Zaloshnja, E., & Lawrence, B. A. (2015, May). The economic and societal impact of motor vehicle crashes, 2010. (Revised) (Report No. DOT HS 812 013). Washington, DC: National Highway Traffic Safety Administration.

Statista. (2022, April 25). *U.S.: vehicles in operation Q4 2017-Q4 2021*. Retrieved April 27, 2022, from https://www.statista.com/statistics/859950/vehicles-in-operation-by-quarter-united-states/

Spacey, J. (n.d.). *10 Types of Information Analysis*. Simplicable. Retrieved April 15, 2022, from https://simplicable.com/new/information-analysis#:%7E:text=Information%20analysis%20is%20the%20systematic%20process%20of%20discovering%20and%20interpreting%20information.

Techopedia. (2017, March 14). *K-Nearest Neighbor (K-NN)*. Techopedia.Com. Retrieved April 24, 2022, from https://www.techopedia.com/definition/32066/k-nearest-neighbor-k-nn

KNN Classification Tutorial using Sklearn Python. (n.d.). Datacamp. Retrieved April 25, 2022, from https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn

Sharma, P. (2021, November 29). Implementing Gaussian Naive Bayes in Python. Analytics Vidhya. Retrieved April 27, 2022, from https://www.analyticsvidhya.com/blog/2021/11/implementation-of-gaussian-naive-bayes-in-python-sklearn/

In Depth: Naive Bayes Classification | Python Data Science Handbook. (n.d.). Python Data Science Handbook. Retrieved April 25, 2022, from https://jakevdp.github.io/PythonDataS

cienceHandbook/05.05-naive-bayes.html

What Is a Decision Tree? (2022, April 19). Master's in Data Science. Retrieved April 30, 2022, from https://www.mastersindatascience.org/learning/introduction-to-machine-learning-algorithms/decision-tree/