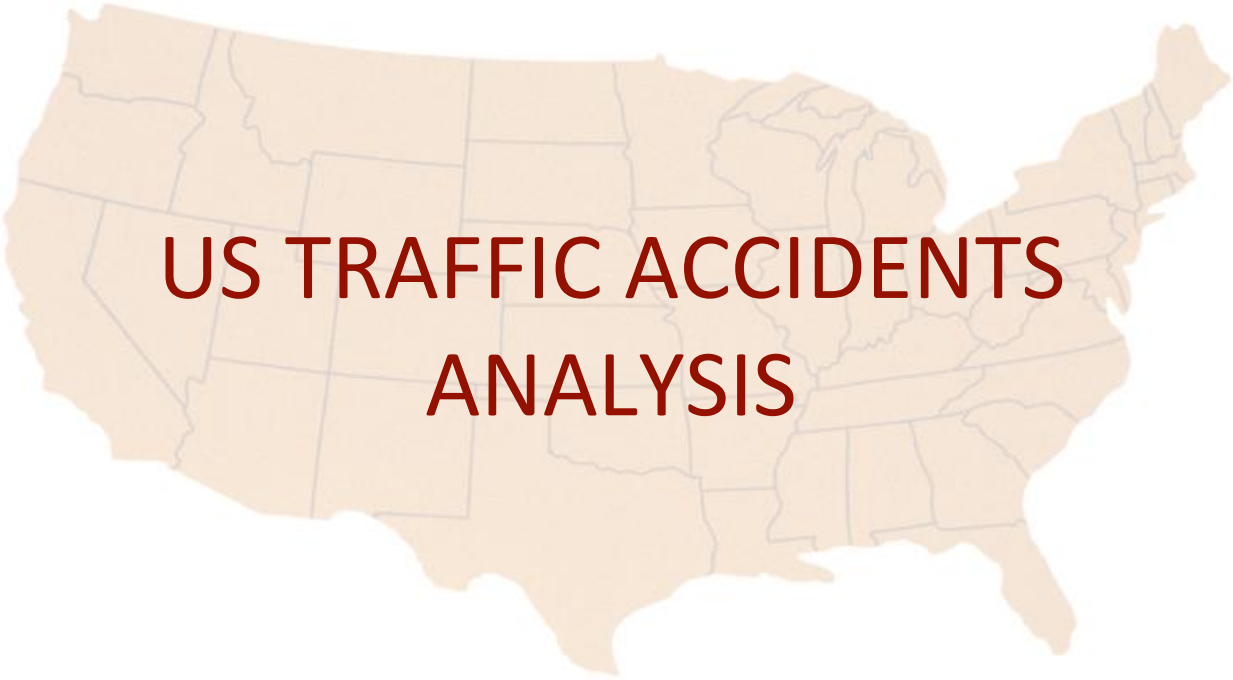

US TRAFFIC ACCIDENTS ANALYSIS

CSP 571 – Fall 2021

Kemen Goïcoechea
Luis Marés de la Cruz
Caridad Arroyo Arévalo

ABSTRACT

Traffic accidents remain one of the leading causes of death in the United States, only behind disease/injury or firearms (CDC. National Center for Health Statistics 2021). The analysis conducted in this project seeks to, first, assess how weather conditions affect US car accidents and then, to identify which factors have the greatest impact.

This project contributes to the development of a classification model for the severity of US accidents. The severity of an accident is a number lying between 1 and 4, where 1 indicates the least impact on traffic —a short delay— and 4 indicates a significant impact on traffic —a long delay—.

The analysis makes use of data collected by Kaggle, and it is carried out in RStudio, the R language-based development environment.

KEY WORDS

Severity, *weather*, *R*, *imbalanced*, *PCA*, *Decision Tree*, *Random Forest*.

INDEX

FIGURES

R OUTPUT SCREEN CAPTURES

TABLES

1. OVERVIEW

As previously introduced, car accidents are still a major cause of mortality in the United States. In light of this fact, data mining techniques appear as a very useful tool, focused on discovering unknown properties and patterns of large data sets.

The dataset covers 49 U.S. states and collects accidents between February 2016 and December 2020. The accidents have been collected from a variety of sources, including U.S. and state departments of transportation, law enforcement, traffic cameras, and traffic sensors within the road-networks.

There are two main papers to be used as references:

- A Countrywide Traffic Accident Dataset. (Sobhan Moosavi 2019)
- Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights. (M. H. Sobhan Moosavi 2019)

The project is structured as follows.

Section 2 describes the first stage of the data workflow: Data Preprocessing. It involves searching for missing data, removing noisy and redundant data to obtain high-quality data which can be meaningful for the subsequent analysis —Data Cleaning and Data Reduction steps—. In addition, Data Preprocessing includes the Data Transformation step, where data is prepared to be fed to the data mining algorithms. The format, structure or values of data are here changed.

Section 3 outlines the exploratory data analysis. By means of statistics and visualizations, initial investigations on the accidents data can be performed. Relationships between *Severity* and some important features are displayed, along with the distribution of the accidents according to different locations: time zones, Stated and cities.

Section 4 presents the core stage of the project: Data Mining stage. PCA is first studied to reduce the large dimensionality of the data set. Then, the test/train split is executed. Feature selection is achieved by evaluating linearity and multicollinearity among predictors, and their significance after fitting a linear model. Sampling techniques are also used to deal with class imbalance issued. Next, model selection is tackled through two supervised learning algorithms: a pruned Decision Tree and a Random Forest. Their processes and results are described and compared. Lastly, model validation takes place by means of confusion matrices and ROC curves.

The last section summarizes and highlights the conclusions derived from the analysis, in addition to listing the documentation, academic literature and bibliography that enable the development of this project.

2. DATA PREPROCESSING

Data preprocessing accounts for a key stage of the data workflow. It is the process of transforming raw data into an understandable format so as to improve reliability and effectiveness. The data set contains 1,516,064 observations and 47 attributes. Please find the description of such features on section 6.1.

The three main steps of data preprocessing are described hereunder.

2.1. Data Cleaning

2.1.1. Missing Values

There are 8 numerical attributes containing NA fields. These are presented below.

Nearly 70% of the observations lack *Number*. This is a high number of observations to which no accurate replacement method could be applied. As a result, this column is simply removed.

	missing_values	percentage
Number	1046095	69.001
Precipitation.in.	510549	33.676
Wind_Chill.F.	449316	29.637
Wind_Speed.mph.	128862	8.500
Humidity...	45509	3.002
Visibility.mi.	44211	2.916
Temperature.F.	43033	2.838
Pressure.in.	36274	2.393

R Output 1: Data Cleaning – NA Values

With respect to the remaining seven variables, they are all related to weather conditions. A first approach involves replacing these missing values by a central tendency. Although it is a valid technique, a second approach is found to yield more precise results. Since weather conditions are similar in close areas on close dates —and since the data set was sorted by date (first) and location (next)— these features' missing values are filled in by means of linear interpolation.

In addition, there are 11 categorical attributes with empty fields. These are shown below.

A further analysis of these empty fields shows that most of them —in particular, fields from features with less than 0.3% empty characters— belong to the same observations. In other words, these observations turn to be low-quality data which should be disregarded. Thus, those observations are dropped.

This decision leaves three categorical features with empty fields: *Weather_Condition*, *Wind_Direction* and *Weather_Timestamp*. Their empty fields are treated as special values, i.e., a new category "*Unknown*" is created.

	empty_values	percentage
Weather_Condition	44007	2.903
Wind_Direction	41858	2.761
Weather_Timestamp	30264	1.996
Airport_Code	4248	0.280
Timezone	2302	0.152
Zipcode	935	0.062
City	83	0.005
Sunrise_Sunset	83	0.005
Civil_Twilight	83	0.005
Nautical_Twilight	83	0.005
Astronomical_Twilight	83	0.005

R Output 2: Data Cleaning – Empty Characters

2.1.2. Noisy Data and Anomalies

The second and last step of Data Cleaning implies identifying and dealing with noisy data/anomalies. Boxplots along with histograms are used for outlier detection in numerical features. Figure 1 shows the boxplots of the 7 numerical features having outliers.

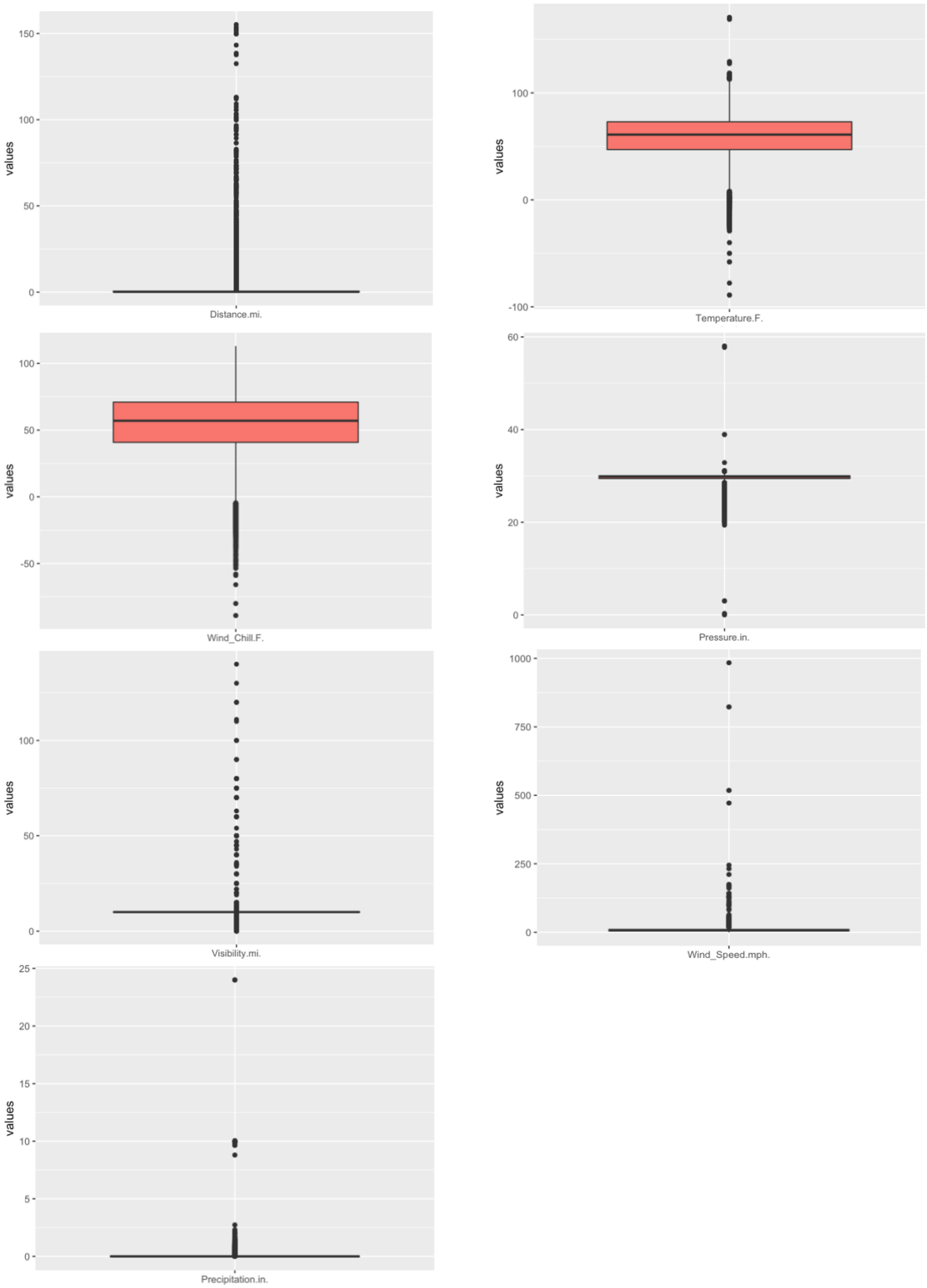

Figure 1: Data Cleaning - Outlier Detection

There are two ways in which to handle these anomalies. On one hand, outliers in *Distance.mi.*, *Temperature.F*, *Wind_Chill.F.* and *Wind_Speed.mph.* are considered to be those observations lying outside the minimum $(Q1 - 1.5IQR)$ and maximum $(Q3 + 1.5IQR)$ boxplot values. Such observations are removed. On the other hand, outliers in *Pressure.in.*, *Visibility.mi.* and *Precipitation.in.* are considered to be those observations outside their corresponding standard ranges. The standard pressure range is 28-21 inches; the standard visibility range is 0-10 miles; and the standard precipitation range is 0-2 inches. Such observations are accordingly removed.

This concludes Data Cleaning. The next step of Data Preprocessing is Data Reduction.

2.2. Data Reduction

The previous step manages to reduce the data set to 1,244,342 observations and 46 features. Yet, it is still a considerable amount of data to use when training the model.

In a first stage, this Data Reduction is determined as unnecessary. However, when the Data Mining process is to be done, computational issues arise. Let us recall that the scope of the project is to analyze the impact of weather conditions on traffic accidents so as to create a real-time accident prediction tool. Therefore, only features related to weather, geographical location and time are kept. Results are optimal, as the number of features decreases significantly, from 46 to 28.

The subsequent steps proceed from this status of the data set. However, computational issues appear again when reaching the Data Mining stage, which leads to the need of further dimensionality reduction. It is first assessed through Principal Component Analysis. PCA's inadequate performance leads to a reconsideration of Data Reduction: drop unnecessary features. These strategies are described in section 4.1.1.

2.3. Data Transformation

The third and last stage of data preprocessing deals with converting the data into the form which best suits data modeling. Among data transformation techniques, the following three stand out:

2.3.1. Appropriate Data Format

Three variables are transformed into a more appropriate, useful and understandable format.

- *Start_Time*: from this variable, 4 new features can be extracted: *Year*, *Month*, *Date* and *Time*.
- *Zipcode* encompasses two distinct types of zip codes: standard zip codes and zip+4 codes. Zip+4 codes identify a geographic segment within a standard zip code and are often subject to change. In other words, they give no additional information. Their five first digits correspond to the standard zip code they belong to, and the following digits make the (+4)-suffix. Therefore, this feature is edited so that zip codes are composed only by the first five digits.
- *Wind_Direction* has duplicated values: same values but in different formats, e.g., "W" and "West". These values are corrected.

2.3.2. Aggregation

Aggregation refers to the process of combining existing attributes in order to come up with a new attribute that may yield better results.

Firstly, the start and end time of the accident (*Start_Time* and *End_Time*) are aggregated to create a new feature, which measures the duration in minutes of the accident. It is called *Duration*. The other two aggregated features are *Lat* and *Lng*, which measure the exact location of the accident byc computing the mean value between *Start_Lat* and *End_Lat;* and *Start_Lng* and *End_Lng*, respectively.

Besides, two additional features are removed. *Weather_Timestamp* shows the timestamp of the weather observation record. However, this information is already covered by the date and time of the accident (*Year*, *Month*, *Date* and *Time* features). *Country* turns to be irrelevant as well, since all observations, all accidents, take place in the US. It is a feature with a single value.

2.3.3. Discretization and Normalization

The final technique employed in Data Transformation is related to discretizing/normalizing attributes. This technique is very simple: all categorical features are converted from characters to factors. This way, they can be used in the upcoming model.

The following table provides a brief summary of Data Preprocessing:

Data Set Size Before	Stage	Data Set Size After
1,516,064 x 47	Data Cleaning (Beginning of Data Preprocessing)	1,244,342 x 46
1,244,342 x 46	Data Reduction	1,244,342 x 28
1,244,342 x 28	Data Transformation (End of Data Preprocessing)	1,244,342 x 27

Table 1: Data Preprocessing - Summary

3. EXPLORATORY DATA ANALYSIS

Exploratory data analysis is a part that helps to better understand the dataset and provide some insights. Firstly, statistics are extracted from the observations. This is key to see if the results obtained later in the project are plausible. Then, some visualizations are made to have insight about the domain studied.

3.1. Statistics

3.1.1. Repartition statistics of features

In the first place, one could ask some basic information about the repartition of the observations for the given features. Below is a table regrouping the mean, maximum and minimum values for important features of the dataset.

Feature	Minimum value	Mean	Maximum value
Severity	1	2.217	4
Month	1	7.731	12
Day	1	16.28	31
Duration (minutes)	5.2	236.5	1421955
Temperature (°F)	8.0	60.6	112.0
Humidity (%)	1	65.29	100
Visibility (mi)	0	9.058	10

Table 2: Exploratory Data Analysis – Basic statistics of important features

As seen in Table 2, there are some shifts in the mean of the data for certain features. For instance, one could expect the mean of the *Month* feature to be around 6.0, but the reality is that more accidents happen in the end of the year than in the beginning. This is also true for the feature *Humidity*.

In this project, the variable of interest is *Severity*. The following table indicates the number of observations for each class of this feature:

Severity	1	2	3	4
Number of Observations	21,111	1,011,432	132,817	78,982

Table 3: Exploratory Data Analysis – Number of observations for each class of Severity

The class 2, is far more present that the other three. Thus, having disparities in the number of predictions for each class is expected. There is an important problem of class imbalance.

3.1.2. Advanced statistics

To have more insights about the domain studied, it is possible to calculate statistics about the happening of events. One example is the average number of accidents per day in the US. This value is calculated by dividing the total number of observations by the number of unique dates. According to this dataset, this value is 721.78.

3.2. Visualizations

Visualizations are the best way of understanding the distribution of the data under certain features. All of the following figures are done with the *ggplot2* library. This part is focused on extracting information about the causes of severe accident.

3.2.1. Data distribution

The distribution of some subjectively important features according to the *Severity* is presented below:

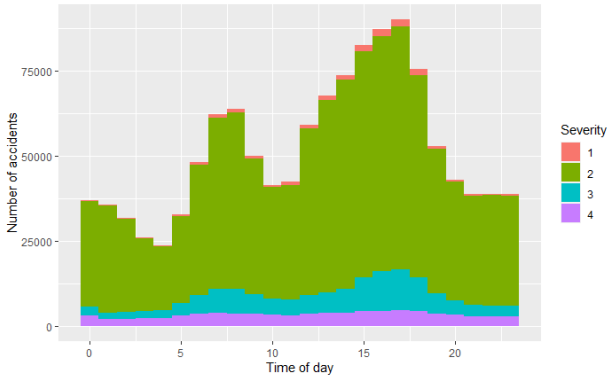

Figure 2: Exploratory Data Analysis - Time of Day vs Accident Severity

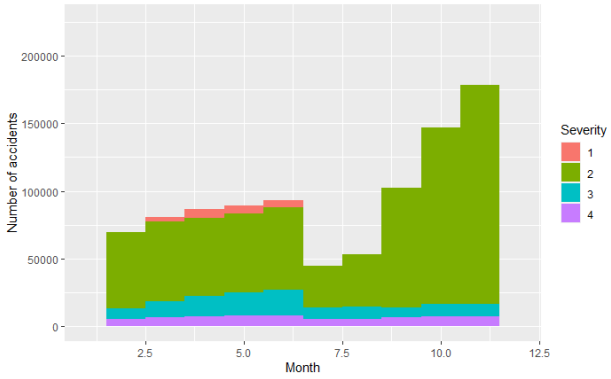

Figure 3: Exploratory Data Analysis - Month vs Accident Severity

A direct observation of the right histogram is that the Summer sees a lot less accidents than the Winter. Moreover, as seen on the left histogram, the distribution according to the feature *time of day* is not uniform as well. The peak hours at approximately 8am and 5pm are the most impacted in term of accidents.

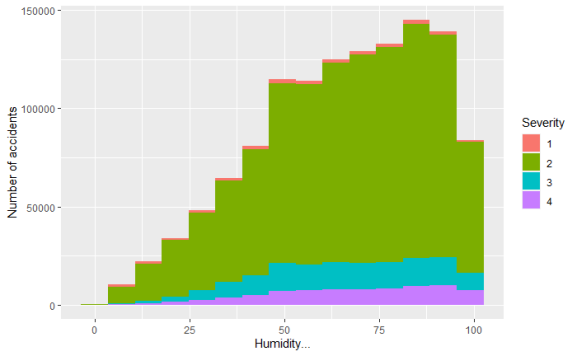

Figure 4: Exploratory Data Analysis - Humidity vs Accident Severity

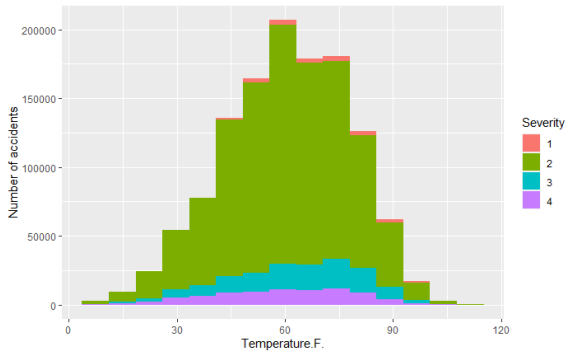

Figure 5: Exploratory Data Analysis – Temperature (°F) vs Accident Severity

These two histograms regroup information about the weather at the moment of the accident. The right histogram displays the number of accidents and their *Severity* according to the *Temperature* in Fahrenheit. One observation is that the most occurred temperature is shifted between accidents with a *Severity* of 2 and a *Severity* of 3 and 4. An outside-temperature of approximately 75°F seems to cause the most severe accidents, whereas this temperature is around 60°F for accidents of severity 2. Concerning the left histogram representing the *Humidity* as a factor of the severity of accidents, it is clear

that a high percentage of humidity causes the most accidents. It is possible that when there is 100% of humidity, which means that it rains, the drivers are more cautious on the roads. This would explain the abrupt decline between 90% and 100%.

3.2.2. Location of the accidents

Firstly, the distribution of accidents across different regions of the country is hereunder presented.

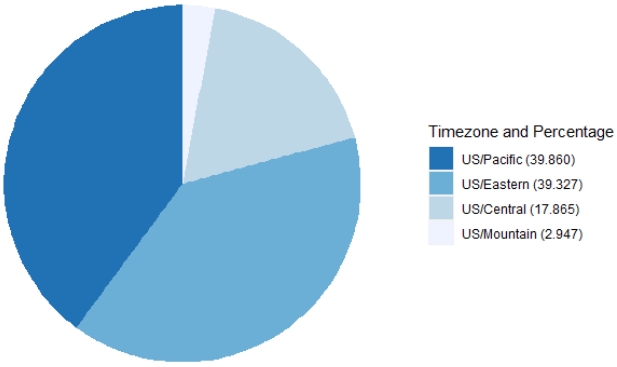

Figure 6: Exploratory Data Analysis – Distribution of accidents according to the time zone

As seen in the pie plot, the distribution is not homogeneous. The Pacific and Eastern regions regroup almost 80% of the car accidents. Now, let's precise where the accidents occurred:

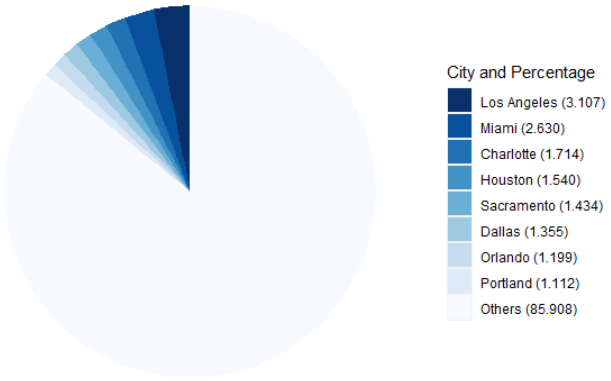

Figure 7: Exploratory Data Analysis – Distribution of accidents according to the city

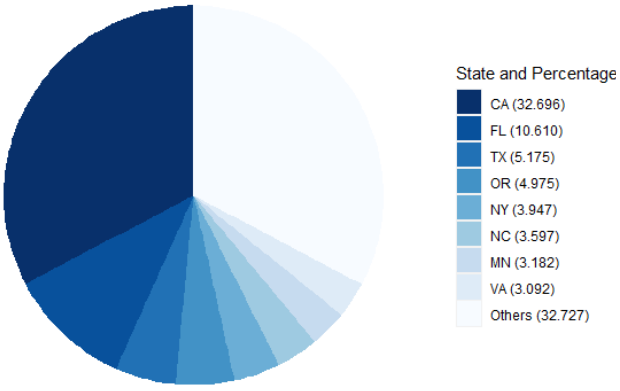

Figure 8: Exploratory Data Analysis – Distribution of accidents according to the State

It is clear that the state of California has, by far, the most accidents in the United-States. These accidents happened in heavily populated cities such as Los Angeles or Sacramento. However, the population of a city is not the only factor because the most populated city in the country, New-York, is not present in the 8 cities with the most accidents.

Because the longitude and latitude of accidents are given, it is possible to plot them in order to form a map of the United States. The map on the bottom left side compares accidents with a *Severity* of 1 to 3 to accidents with a *Severity* of 4.

The map is consistent with the previous pie plots, as the repartition of accidents between the states is not homogeneous. Even though, the East coast is not the most impacted region in the country, it is clear that the density of *Severity* 4 accidents is higher than in the West coast.

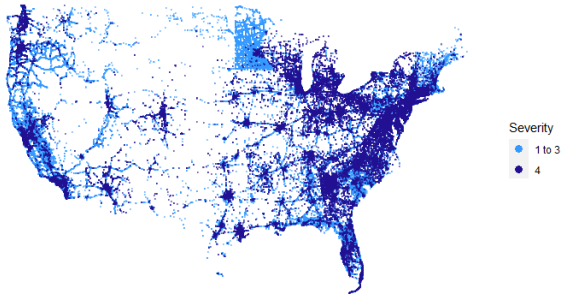

Figure 9: Exploratory Data Analysis – Location and severity of accidents

4. DATA MINING

This third stage entails the process of applying techniques that enable the extraction of patterns from the data in order to answer the questions that were initially posed. It is comprised by two different phases: model training and model validation.

4.1. Model Training

To address the previously established target inquiries, a model must be carefully selected from a collection of suitable candidates for a training data set. A previous step, feature selection, is required to determine which attributes of the data set provide useful information and which, conversely, do not.

4.1.1. Feature selection

At the beginning of this stage of the project the data set comprises 1,244,342 observations and 27 features. An attempt is made to continue with the development of the project with this large data set, but this turns out not to be possible as it demands a high computational cost. For this reason, it is decided to backtrack a few steps and reduce the dimensionality of the data set before proceeding with the feature selection per se.

In a first attempt, Principal Component Analysis (PCA) technique is used. An 80/20 test-train is performed on the data set and a subset comprising only the numerical attributes is created. The attributes are scaled since the mean and the variance values for the features are highly different. After applying PCA, the principal component loadings are obtained:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Year	-0.044915952	0.07742333	-0.60754480	0.28140590	0.0750301649	-0.012181155	0.0641454171
Month	0.022745266	-0.05482667	-0.31545873	-0.18037590	-0.3507099202	-0.007306598	-0.0005684392
Day	0.021476864	-0.03735640	-0.11919243	0.02031394	0.1017179924	-0.150551640	-0.9711755764
Time	-0.196133739	-0.07121624	0.13425012	0.33756585	0.1826340813	0.008138913	-0.0088983290
Duration	0.006674744	-0.01339575	0.04442320	-0.02385417	-0.0382907546	-0.986997499	0.1423102880
Lat	0.265261421	0.25713407	0.22748855	0.42784512	0.0701997010	-0.017043209	0.0041118726
Lng	0.279157669	-0.61099652	-0.08363431	0.11850025	-0.0004040836	0.009737004	0.0516029495
Zipcode	-0.284420544	0.59970152	0.08670688	-0.13770688	0.0074904647	-0.009450080	-0.0563382182
Temperature.F.	-0.502575546	-0.25403108	0.01068953	-0.19267060	0.1428231216	-0.006463065	0.0087985039
Wind_Chill.F.	-0.463436087	-0.16970864	-0.23786574	-0.13815267	0.1817970979	-0.012700514	0.0468132749
Humidity...	0.417450406	0.03591605	-0.21614895	-0.30789681	0.1192651073	-0.001339559	-0.0126245092
Pressure.in.	0.071183848	-0.11742181	0.39300083	-0.43443155	-0.3526693981	0.046194593	-0.1080105168
Visibility.mi.	-0.272280393	-0.15445331	0.13062914	0.25095303	-0.4996215225	0.007704975	-0.0785165207
Wind_Speed.mph.	-0.056221852	-0.22615021	0.37065342	0.22050699	0.2811519862	-0.005278366	-0.0553535830
Precipitation.in.	0.072928794	-0.04319691	0.13636570	-0.32549321	0.5481826467	0.001229267	0.0565196209

PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15
0.094284961	0.064192614	0.045126731	0.49289220	-0.346194219	-0.24777890	-0.306637607	-0.0231246870
-0.711186676	-0.220116231	-0.379344122	-0.14764964	-0.142786218	-0.01324030	-0.029481078	0.0018670359
-0.010826644	-0.018331977	0.050240902	-0.04641999	0.015228719	-0.04248888	0.012604996	0.0010299086
-0.503619147	0.707224609	0.012068133	0.09627817	0.090966235	0.13524115	0.034712369	0.0011748714
0.002369667	0.026060932	0.002727002	0.02501208	-0.002180705	-0.00519412	-0.001181004	-0.0003732258
-0.096387858	-0.113606988	0.166806108	-0.44628535	-0.566600959	0.10366993	-0.199774312	-0.0509376385
0.022176569	0.006588493	0.086502222	-0.08550295	0.043903892	-0.05361797	0.020411115	-0.7102195111
-0.021234266	-0.019171144	-0.128407245	0.11639302	0.029495176	0.08540046	-0.006499559	-0.7004303914
0.069277815	-0.009720773	-0.007333580	-0.29674639	0.022000704	0.10683712	-0.721264018	-0.0024661554
0.131685567	0.049560710	0.037231009	-0.21976160	-0.488447070	0.14669511	0.561664437	-0.0176389002
0.118158560	0.226621159	-0.135550468	0.10582690	-0.102573536	0.73912461	-0.128579913	0.0169667819
0.047984922	0.352328560	0.045903487	0.22283991	-0.486923209	-0.26981506	-0.104955347	-0.0259983372
-0.046793011	-0.283453030	0.353313666	0.34354090	-0.032805317	0.48948985	0.004938795	0.0057699291
0.068435250	-0.261014445	-0.679321157	0.32188255	-0.184468698	0.07066255	0.009789082	0.0231153982
-0.419592655	-0.326207450	0.437447468	0.29041174	-0.075039770	-0.02756437	-0.003608587	-0.0012003578

R Output 3: Model Training - PCA loadings

Moreover, the screen plots of the Proportion and Cumulative Proportion of Variance Explained by each of the principal components are visualized:

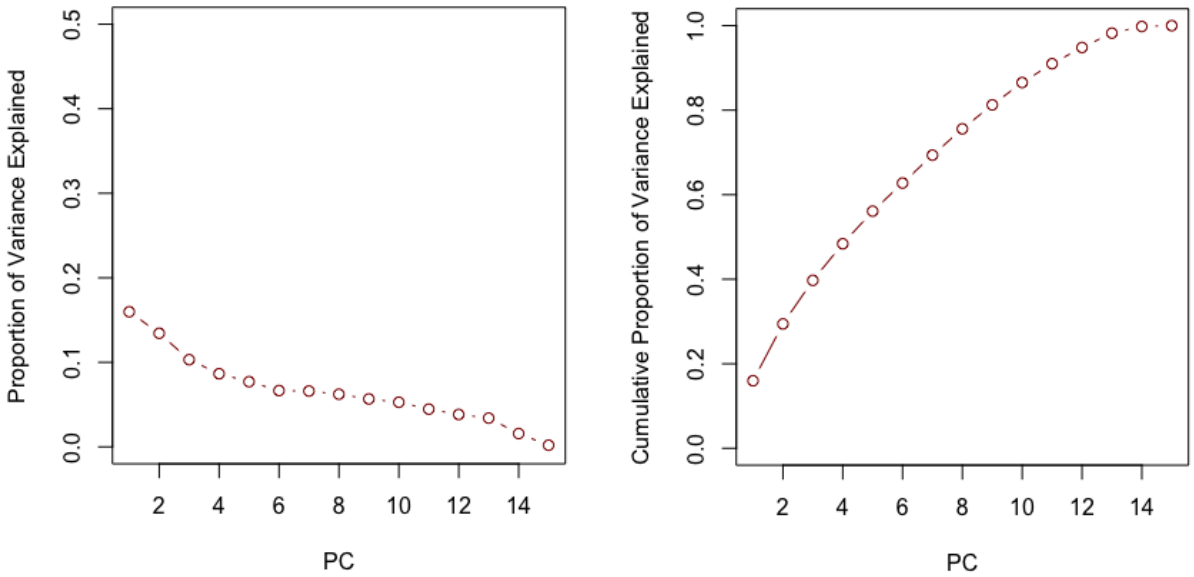

Figure 10: Model Training - PCA screen plots

Commonly, most of the variance —around 80%— is explained by the first two or three principal components. For this scenario, unfortunately, it is not. On the contrary, a proportion of variance high enough to serve for dimensionality reduction purposes is achieved with the first nine principal components, as can be seen on the screen plots. The reduction that would be reached by considering the main components up to PC9 is not sufficiently high. Therefore, PCA technique is considered of no use.

Recalling the results obtained in the exploratory stage, it can be accurately concluded that meteorological and environmental factors have high impact on traffic accidents. Thus, in order to reduce the dimensionality of the complete data set—before the train-test split—, it is decided to consider only the features related to weather.

The objective of the project is then focused on creating a tool capable of predicting traffic accidents in real time by considering weather conditions. For this purpose, a pre-selection of attributes is carried out, discarding those that are not related to these variables. The following features were maintained: *Timezone, Temperature.F., Wind_Chill.F., Humidity…, Pressure.in., Visibility.mi., Wind_Direction, Wind_Speed.mph., Precipitation.in., Weather_Conditions, Sunrise_Sunset, Civil_Twilight, Nautical_Twilight* and *Astronomical_Twilight*. The number of features is lowered to 14.

The test-train split is then performed, allocating 80% of the observations to the first subset and the remaining 20% to the second one. From this point, the proper feature selection is carried out.

Firstly, the correlation between predictors is assessed so as to avoid multicollinearity issues. Multicollinearity should be prevented since non-independent predictors can increase the variance of the estimated coefficients and make them very sensitive to minor changes in the model. The correlation matrix obtained is presented hereunder:

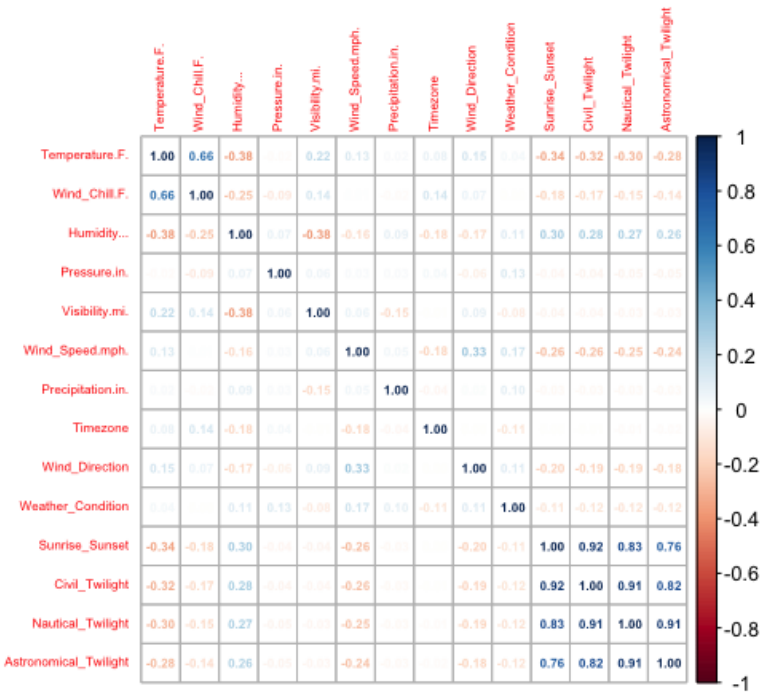

Figure 11: Model Training - Correlation matrix

The correlation matrix presents high correlation between several predictors. As mentioned some lines above, multicollinearity can have detrimental effects on the model. Two features are considered correlated if their correlation value is below -0.75 —high negative correlation— or above 0.75 —high positive correlation—. Consequently, the features *Civil_Twilight, Nautical_Twilight* and *Astronomical_Twilight* are disregarded.

Moreover, only statistically significant attributes should be employed. The statistical significance can be defined as the assertion that an outcome of data generated by testing or experimentation is not likely to occur by chance or random chance but is expected to be traceable to a specific cause. Therefore, building a model that considers only statistically significant features contributes to support the assumption that the outcomes obtained are realistic and not caused by chance or randomness. Significance can be determined by observing p-values, where a p-values lower than around 0.05 indicates that the attribute is statistically significant.

To assess which attributes are statistically significant for this project, a linear regression is performed considering *Severity* as the response value and all the features as predictors. The p-values obtained for each feature can be found in the right-hand side image, where it can be noted that all 11 predictors are statistically significant.

After this last step, the feature selection stage is finalized, and the model selection and building can be carried out.

```
Coefficients:
                    Estimate Std. Error  t value Pr(>|t|)
(Intercept)       -5.840e-01  3.974e-02  -14.698  < 2e-16 ***
Temperature.F.     1.542e-03  4.738e-05   32.534  < 2e-16 ***
Wind_Chill.F.     -1.857e-03  3.797e-05  -48.917  < 2e-16 ***
Humidity...       -8.318e-05  3.042e-05   -2.734  0.00626 **
Pressure.in.       1.012e-01  1.338e-03   75.648  < 2e-16 ***
Visibility.mi.    -2.525e-03  2.776e-04   -9.094  < 2e-16 ***
Wind_Speed.mph.    3.414e-03  1.373e-04   24.873  < 2e-16 ***
Precipitation.in.  1.439e-01  1.138e-02   12.644  < 2e-16 ***
Timezone          -8.579e-02  5.070e-04 -169.205  < 2e-16 ***
Wind_Direction    -1.424e-03  1.962e-04   -7.257 3.95e-13 ***
Weather_Condition  8.671e-04  2.861e-05   30.306  < 2e-16 ***
Sunrise_Sunset     9.033e-03  1.299e-03    6.954 3.56e-12 ***
```

R Output 4: Model Training - p-values

4.1.2. Classification

Once the most relevant features have been identified, the model can be built. Two approaches are carried out. For the first one, a decision tree is obtained and pruned, while for the second one Random Forest learning method is employed.

During this stage, several problems are encountered. Due to the large amount of observations —1,244,342— the computational cost involved in building the models is too high, which makes it impossible to use the complete data set. As a result, the data set is reduced to 20,000 observations.

Prior to the model building, an analysis of the class-imbalance of the data set is performed. Class-imbalance is an issue that arises when the number of observations for the training data set for each class differs highly. This circumstance can lead to a model that provides an excellent performance when predicting the predominant classes, but a poor one when the class to be predicted is the one whose number of observations is low.

Analyzing the training data set it is noticed that it is extremely unbalanced. To overcome this issue, both over and under sampling techniques are performed, for which the function *ovun.sample*[1] is employed. This function requires that the data set is comprised for only two classes, while the training dataset of the project has four —*Severity* has four classes: 1, 2 3 and 4—. Consequently, two subsets are created, one containing the observations belonging to classes 1 and 2, and a second one with the observations of classes 3 and 4. The first subset is balanced with no sample size constraints, as class 2 is the predominant one, while for the second subset the sample size is set to the length of the balanced first subset. Finally, both subsets are combined into the data set which will be used as the training one. The graphs below depict the percentage of observations belonging to each class before and after the imbalance tackling.

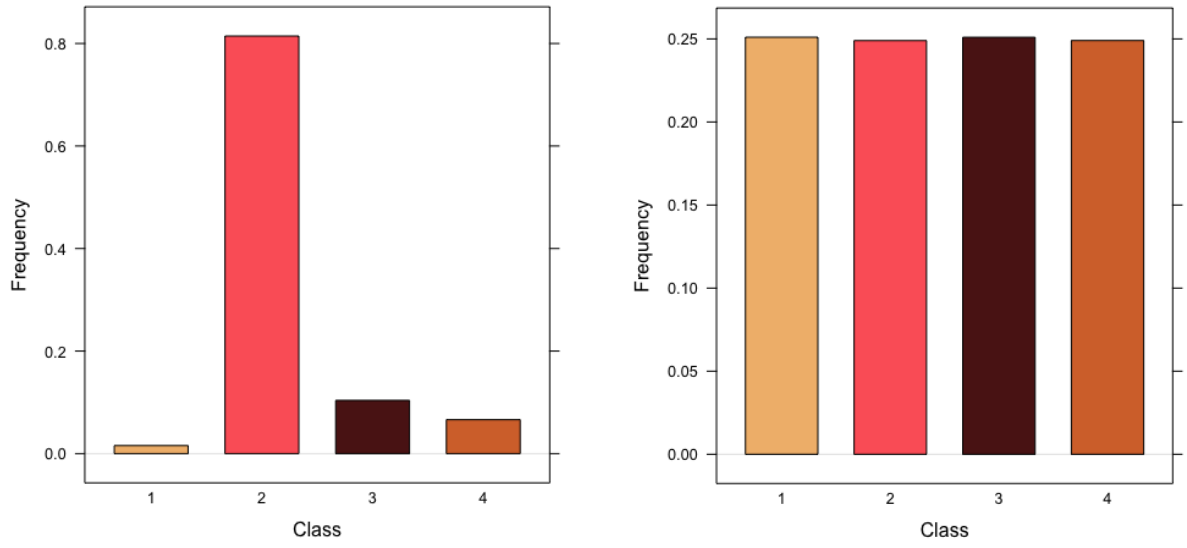

Figure 12: Classification – Imbalanced and balanced classes

[1] *ovun.sample* belongs to the *ROSE* packet. Its functionality description can be found in https://www.rdocumentation.org/packages/ROSE/versions/0.0-4/topics/ovun.sample

Model building

For the first approach, a decision tree is built and then pruned afterwards. For the pruning, the optimal complexity parameter is determined as the one that minimizes the cross-validated relative error. Analyzing the variation of the error as a function of the complexity parameter, the optimal value for the latter is found to be 0.01. The pruned decision tree, along with a graph depicting the cross-validated relative error as a function of the complexity parameter, can be found hereunder.

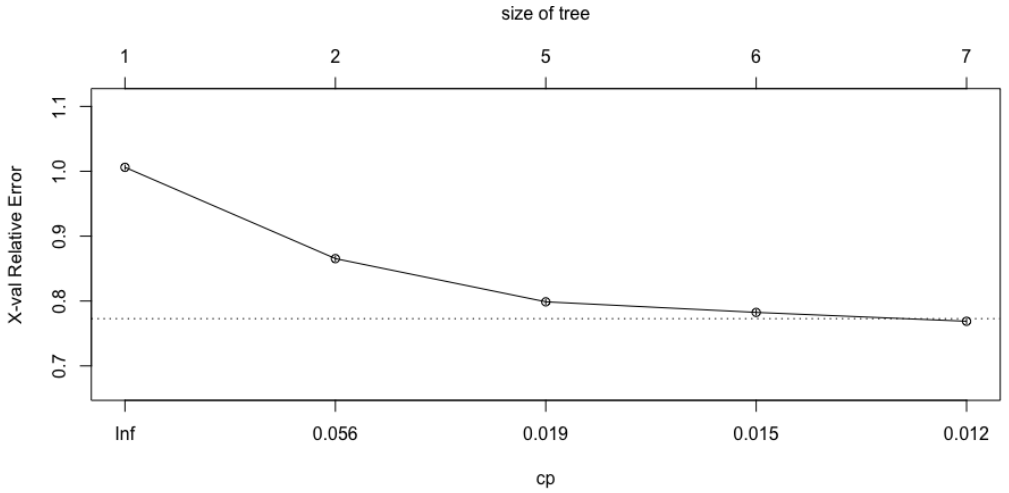

Figure 13: Classification - Complexity parameter

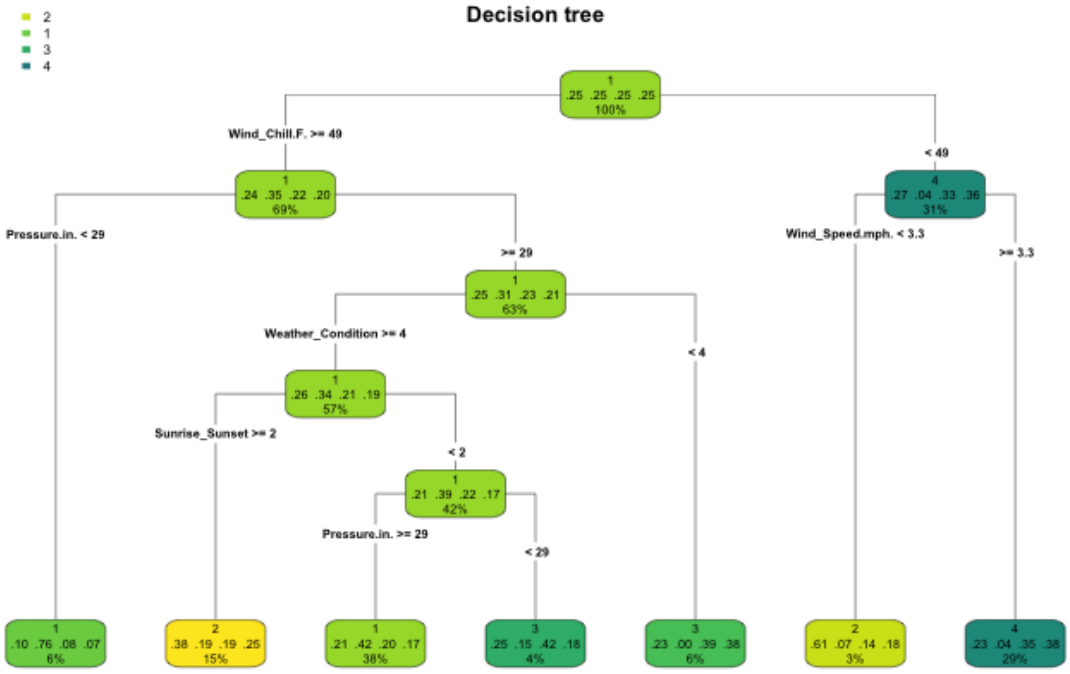

Figure 14: Classification - Decision tree

The decision trees obtained before and after pruning are the same, as the complexity parameter set by the *rpart()* function is the same as the one that minimizes the cross-validated relative error. Therefore, no improvement is achieved with the pruning stage.

The second approach is based on building a random forest. This method relies on the *wisdom of crowds* idea: a large number of uncorrelated decision trees are built, where each one yields a prediction, and the class predicted by the majority is proposed as the final prediction. The number of variables randomly sampled as candidates at each split —referred to as *mtry* parameter— is one of the most relevant parameters in random forest building. Its value is chosen as the optimal value with respect to Out-of-Bag error estimate by means of the function *tuneRF()*, that proposes a value of *mtry* equal to 3 for an Out-of-Bag error 3.63%. The random forest output is not as visual as the decision tree, but its performance is evaluated when fitting the model in the next step.

Model fitting

Both models — pruned decision tree and random forest— are fitted on the test data set, and their accuracies obtained:

Model	Accuracy
Pruned decision tree	0.2805
Random forest	0.7245

Table 4: Classification - Model accuracies

Although none of the proposed models has a high accuracy, random forest will be the approach of choice. It is also important to consider that the data set had to be reduced to 20,000 observations, and that a higher accuracy could be expected if the training data set had been higher. However, as already mentioned, the computers employed to deploy this project do not have the required computational capacity to build models with the large amount of data the data set was initially comprised by.

After selecting random forest as the approach to follow, the question arises as to whether cross-validation needs to be applied or not. As already mentioned, random forest is an ensemble method that combines multiple *weak* learners — characterized either by a high bias or a high variance— into a more powerful learning method. Its performance is based on bagging, where several trees are independently grown on random samples of observations, and each split on each tree is caried out using a random subset of the features. It could be then stated that random forests perform their own cross-validation, and therefore there is no logic in doing it again.

4.2. Model Validation

This part is dedicated to evaluating the model previously created. Since the Random Forest model has by far the best accuracy, it is the one who is evaluated.

4.2.1. Confusion Matrix

Here are the results of the model for predicting the class *Severity*:

Prediction \ Reference	1	2	3	4
1	11	37	5	1
2	61	2741	311	175
3	5	308	82	35
4	2	148	40	38

Table 5: Model Validation - Confusion Matrix for the Random Forest model

Table 4 indicates that the model is more inclined to predict a *Severity* of 2.

This classification has an overall accuracy of 72.58%. More precisely, these are some key statistics for each class:

Class	1	2	3	4
Sensitivity	15.19%	85.34%	21.69%	14.46%
Specificity	98.93%	30.03%	90.40%	95.28%

Table 6: Model Validation: Classification Statistics for the Random Forest model

The sensitivity and specificity observed confirmed that the classification is biased over the second class. This is maybe due to the disparity of the number of observations for each class observed in the Data Analysis part.

4.2.2. ROC Curves

This model is predicting 4 classes of the feature *Severity*. Hence, to be able to represent the ROC curve in 2 dimensions, 4 models are deduced and are predicting one class of *Severity* against the other three.

The 4 ROC curves calculated from the prediction on the testing set are represented below:

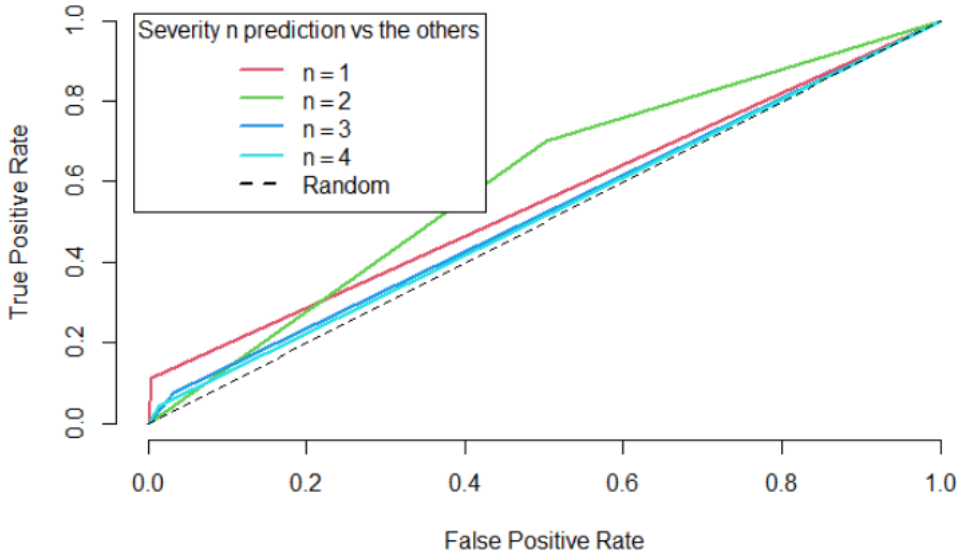

Figure 15: Model Validation – ROC curves for the Random Forest model

The resulting curves are disappointing. As observed in table 6, the model is better at predicting the class *Severity=2*, but for the other 3 it is almost random. This can be explained by the large cut of observations in the dataset to be able to run the algorithm on our machines. This implies a loss of explained variance in the data that the model cannot catch. Also, the fact that the distribution of observation over the class *Severity* cause a bias in the classification.

5. CONCLUSIONS

Throughout the project, the group has fought against the *curse of dimensionality*. Numerous drawbacks have been encountered due to the high computational cost required by the large dataset chosen. Not only the large number of observations, but also the amount of features, have been translated in issues that have impeded the development of the tasks that were initially stablished. Nonetheless, this has served as a lesson and has allowed to put into practice numerous ideas learned during the course.

During the first stage, a thorough cleaning and modification of the data set is carried out. It allowed a better understanding of the data set, as well as a reduction of the dimensionality, both in observations and features. The Exploratory Data Analysis stage aided significantly in the posterior decision of deciding which features where the most relevant, in order to perform a severe reduction, needed due to the mentioned computational issues.

Regarding the results obtained after the completement of the project, it could be said that are good but not excellent. The accuracy yielded by the random forest —0.7245— is appropriate, but after the stage of model validation it has been noticed that the imbalance of the data set, although theoretically addressed, is translated into a good behavior when predicting the predominant class but a worse one when it comes to predicting the remaining classes. At first, it was believed that this low accuracy value could be improved if using a higher number of observations and features, but the imbalance would probably yield to the same results.

As a conclusion, a final thought is extracted: several tools and concepts have been learned, and will continue to be learnt, as the time goes on. The application of these will probably not yield perfect results, since real life does not have perfect data sets, but this should not serve as a setback, but rather as an engine to try to find combinations of these tools that will allow us to obtain the results we desire. It could be said that this course has given us a toolbox that we must learn to use, which will be achieved with practice.

6. DATA SOURCES

The dataset employed in this project can be found in Kaggle under the title *US Accidents: A Countrywide Traffic Accident Dataset.* A more thorough description is presented hereunder.

6.1. Description of the data set

The data set[2] to be analyzed comprises information related to traffic accidents occurred in 49 states of the US from February 2016 to December 2020. In order to gather this data, use has been made of APIs that disseminate traffic events captured by various entities. The data set is composed by 1,516,064 observations and 47 attributes. The description of the attributes can be found in the table hereunder:

Nº	Attribute	Description	Nullable
1	ID	This is a unique identifier of the accident record.	No
2	Severity	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).	No
3	Start_Time	Shows start time of the accident in local time zone.	No
4	End_Time	Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow was dismissed.	No
5	Start_Lat	Shows latitude in GPS coordinate of the start point.	No
6	Start_Lng	Shows longitude in GPS coordinate of the start point.	No
7	End_Lat	Shows latitude in GPS coordinate of the end point.	Yes
8	End_Lng	Shows longitude in GPS coordinate of the end point.	Yes
9	Distance(mi)	The length of the road extent affected by the accident.	No
10	Description	Shows natural language description of the accident.	No
11	Number	Shows the street number in address field.	Yes
12	Street	Shows the street name in address field.	Yes
13	Side	Shows the relative side of the street (Right/Left) in address field.	Yes
14	City	Shows the city in address field.	Yes
15	County	Shows the county in address field.	Yes
16	State	Shows the state in address field.	Yes
17	Zipcode	Shows the zip code in address field.	Yes
18	Country	Shows the country in address field.	Yes
19	Timezone	Shows time zone based on location of the accident (eastern, central, etc.).	Yes

[2] The data set can be found in the following link: https://www.kaggle.com/sobhanmoosavi/us-accidents

20	Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.	Yes
21	Weather_Timestamp	Shows the timestamp of weather observation record (in local time).	Yes
22	Temperature(F)	Shows the temperature (in Fahrenheit).	Yes
23	Wind_Chill(F)	Shows the wind chill (in Fahrenheit).	Yes
24	Humidity(%)	Shows the humidity (in percentage).	Yes
25	Pressure(in)	Shows the air pressure (in inches).	Yes
26	Visibility(mi)	Shows visibility (in miles).	Yes
27	Wind_Direction	Shows wind direction.	Yes
28	Wind_Speed(mph)	Shows wind speed (in miles per hour).	Yes
29	Precipitation(in)	Shows precipitation amount in inches, if there is any.	Yes
30	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)	Yes
31	Amenity	A POI annotation which indicates presence of amenity in a nearby location.	No
32	Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.	No
33	Crossing	A POI annotation which indicates presence of crossing in a nearby location.	No
34	Give_Way	A POI annotation which indicates presence of give way in a nearby location.	No
35	Junction	A POI annotation which indicates presence of junction in a nearby location.	No
36	No_Exit	A POI annotation which indicates presence of no exit in a nearby location.	No
37	Railway	A POI annotation which indicates presence of railway in a nearby location.	No
38	Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.	No
39	Station	A POI annotation which indicates presence of station in a nearby location.	No
40	Stop	A POI annotation which indicates presence of stop in a nearby location.	No
41	Traffic_Calming	A POI annotation which indicates presence of traffic calming in a nearby location.	No
42	Traffic_Signal	A POI annotation which indicates presence of traffic signal in a nearby location.	No
43	Turning_Loop	A POI annotation which indicates presence of turning loop in a nearby location.	No
44	Sunrise_Sunset	Shows the period of day (i.e., day or night) based on sunrise/sunset.	Yes
45	Civil_Twilight	Shows the period of day (i.e., day or night) based on civil twilight.	Yes
46	Nautical_Twilight	Shows the period of day (i.e., day or night) based on nautical twilight.	Yes
47	Astronomical_Twilight	Shows the period of day (i.e., day or night) based on astronomical twilight.	Yes

Table 7: Data Sources - Dataset attributes

7. Bibliography

2021. *CDC. National Center for Health Statistics.* October 14. https://www.cdc.gov/nchs/fastats/injury.htm.

Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, Rajiv Ramnath. 2019. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." *arXiv.org.* September 19. https://arxiv.org/abs/1909.09638.

Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Rajiv Ramnath. 2019. "A Countrywide Traffic Accident Dataset." *arXiv.org.* Jun 12. https://arxiv.org/abs/1906.05409.