1. Business problem.

You probably have a familiar, friend (or even yourself) who is afraid of air travel. The speed, the height, the feeling when taking off and landing... Well, you should know that the probabilities of suffering a plane accident are very low if we compare them, for example, with the possibilities of suffering a car accident.

Source: thaviationbusiness.com

Car travelling is more than present in many people's lives. We use the car for many and even many times in one day. Even when the use of public transport is increasing, there are not always public transport options available in all places.

So, how can we face the risk of traveling by car?

- You should listen to the authorities
- Use your belt
- Do not drive if you are drunk.
- Do not exceed speed limits

But that's not all. There are many other aspects and variables about the possibilities of having an accident (and the severity of it) that must be taken into account.

In this sense, the aim of this project is to help stakeholders reduce exposure to the risk of suffer a fatal car accident based on a prediction of the severity of a car accident based on different conditions.

1.1 Stakeholders

The main stakeholders in this work would be:

- **Car drivers**. Probably this is the group that is directly affected, knowing the conditions that influence the severity of an accident can be of great value to them. Moreover, if at the beginning of a trip the conditions for a severe accident are present, that information could be useful for the driver, for example, to reduce speed, change the route or wait for another time to make the trip.
- **The authorities.** The authorities would benefit in the same way by being able to mitigate or prevent those indicators most predominant in serious accidents
- **The insurance companies.** They could dynamically adapt policies to suit actual risk exposure.

2. Data Understanding and preparation

We will use for our purpose the following dataset:
"All collisions provided by SPD and recorded by Traffic Records " This Dataset includes a record of all collisions ocurred in the city of Seattle (Washingthon) from 2004 to 2020.
The dataset includes the feature "SEVERITY CODE" which indicates the severity of the car accident. This variable will be our dependent variable (Y), and our goal will be to predict it. The variable can have the following values per accident:

- 2—injury: physical harm referred to a person
- 1—prop damage: material damage.

Listed below, we can see a detailed description with all the features included in the dataset with a description of the meaning of their values. Please, noticed that not all these features will be useful or at least, will need processing or transformation

Feaute name	Description
SEVERITYCODE	A code that corresponds to the severity of the collision
X	Longitude
Y	Latitude
OBJECTID	ID
INCKEY	A unique key for the incident
COLDETKEY	Secondary key for the incident
REPORTNO	na
STATUS	
ADDRTYPE	Collision address type
INTKEY	Key that corresponds to the intersection associated with a collision
LOCATION	Description of the general location of the collisio
EXCEPTRSNCODE	na
EXCEPTRSNDESC	na
SEVERITYCODE.1	
SEVERITYDESC	A detailed description of the severity of the collision
COLLISIONTYPE	Collision type
PERSONCOUNT	The total number of people involved in the collision
PEDCOUNT	The number of pedestrians involved in the collision. This is entered by the state
PEDCYLCOUNT	The number of bicycles involved in the collision. This is entered by the state.
VEHCOUNT	The number of vehicles involved in the collision. This is entered by the state.
INCDATE	The date of the incident.
INCDTTM	The date and time of the incident.
JUNCTIONTYPE	Category of junction at which collision took place
SDOT_COLCODE	A code given to the collision by SDOT.
SDOT_COLDESC	A description of the collision corresponding to the collision code
INATTENTIONIND	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.
WEATHER	A description of the weather conditions during the time of the collision.
ROADCOND	The condition of the road during the collision.
LIGHTCOND	The light conditions during the collision
PEDROWNOTGRNT	Whether or not the pedestrian right of way was not granted. (Y/N)
SDOTCOLNUM	A number given to the collision by SDOT.
SPEEDING	Whether or not speeding was a factor in the collision. (Y/N)
ST_COLCODE	A code provided by the state that describes the collision. For more information about these codes, please see the State Collision Code Dictionary.
ST_COLDESC	A description that corresponds to the state's coding designation.
SEGLANEKEY	A key for the lane segment in which the collision occurred.
CROSSWALKKEY	A key for the crosswalk at which the collision occurred.
HITPARKEDCAR	Whether or not the collision involved hitting a parked car. (Y/N

2.1 Feature selection

Based on the Metadata information, we select the useful and potential useful columns:

'SEVERITYCODE','ADDRTYPE','PERSONCOUNT','PEDCOUNT','PEDCYLCOUNT','VEHCOUNT','INCDATE','INCDTTM','JUNCTIONTYPE','INATTENTIONIND','UNDERINFL','WEATHER', 'ROADCOND', 'LIGHTCOND', 'PEDROWNOTGRNT','SPEEDING','HITPARKEDCAR'

We will back at this process in further steps if needed.

2.2 Identifying missing values

Before starting the explanatory data analysis, we have to identify missing and NaN values. Since we are working mainly with categorical data, there are to principal options to handle this situation:

- **Replace the missing value with the most common v**alue of the feature. Important to understand the potential negative implications, especially in features with a huge number of different values without a clear most common value.
- **Delete de rows with missing values.**

We decided to go with the first approach since the columns with the missing values are categorical and there is not column with a high number of missing values.

2.2 Encoding Categorical variables

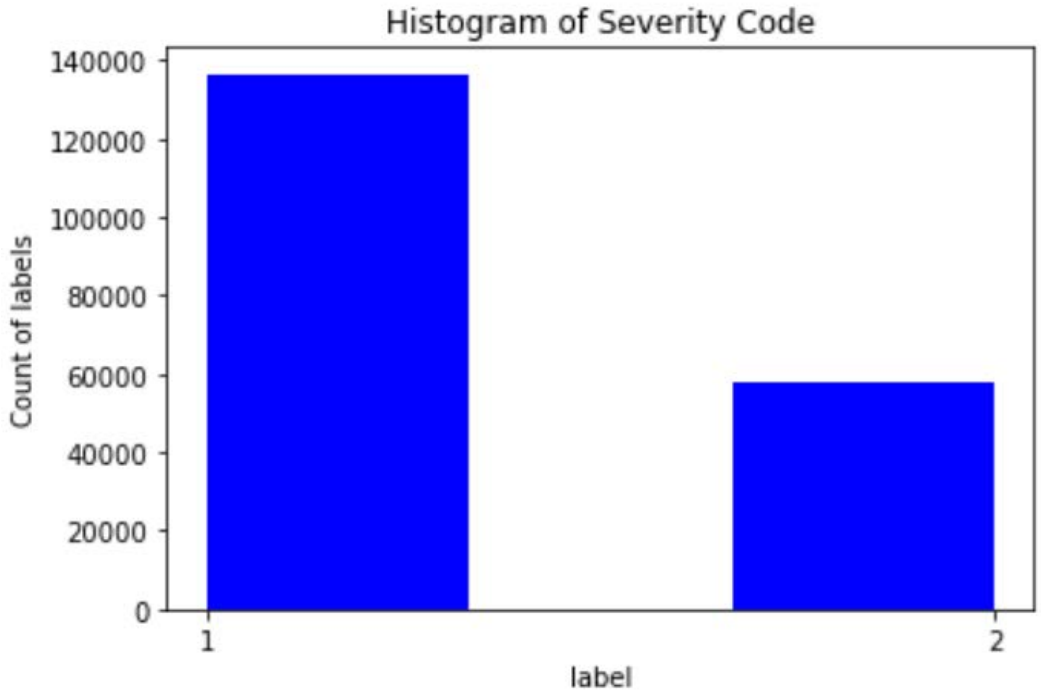
Our machine learning algorithm can only read numerical values. It is essential to encoding categorical features into numerical values. We used the "label encoder "from sklearn to convert the categorical variables to numerical.

3. Methodology

This phase represents the main component of the report where we are going to discuss and describe the exploratory data analysis that we did, and machine learning techniques applied.

3.1 Exploratory data analysis.

Our objective variable “SEVERITYCODE”, shows the following distribution.

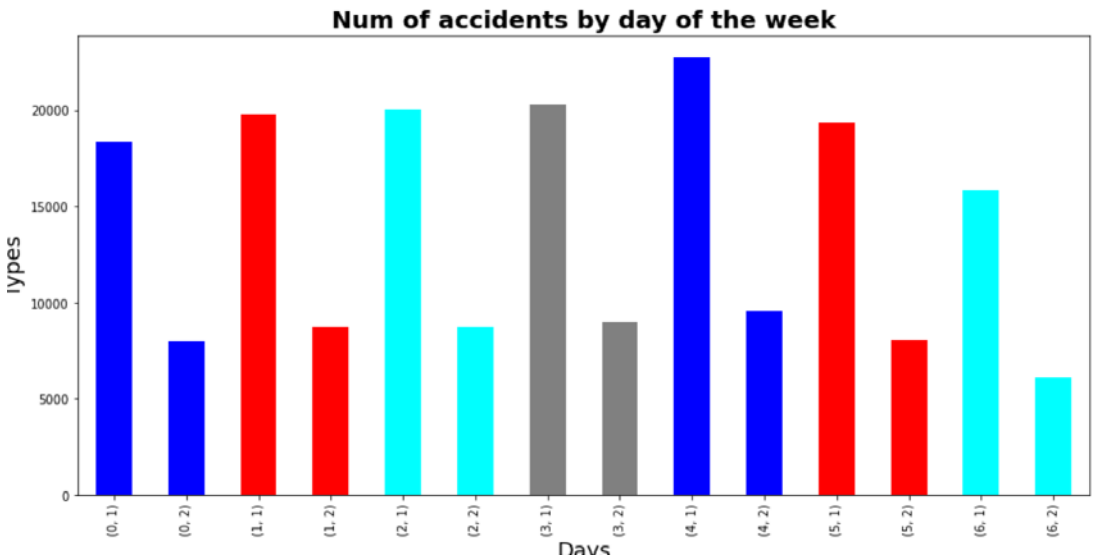


The value 1 "Property damage" it represented almost the double than the value 2 "injury". That means that the data needs to be balance in order to the model perform correctly. Imbalanced classification involves developing predictive models on classification datasets that have a severe class imbalance.

Perhaps the most widely used approach to synthesizing new examples is called the Synthetic Minority Oversampling TEchnique, or SMOTE for short. This technique was described by Nitesh Chawla, et al. in their 2002 paper named for the technique titled “SMOTE: Synthetic Minority Over-sampling Technique.”

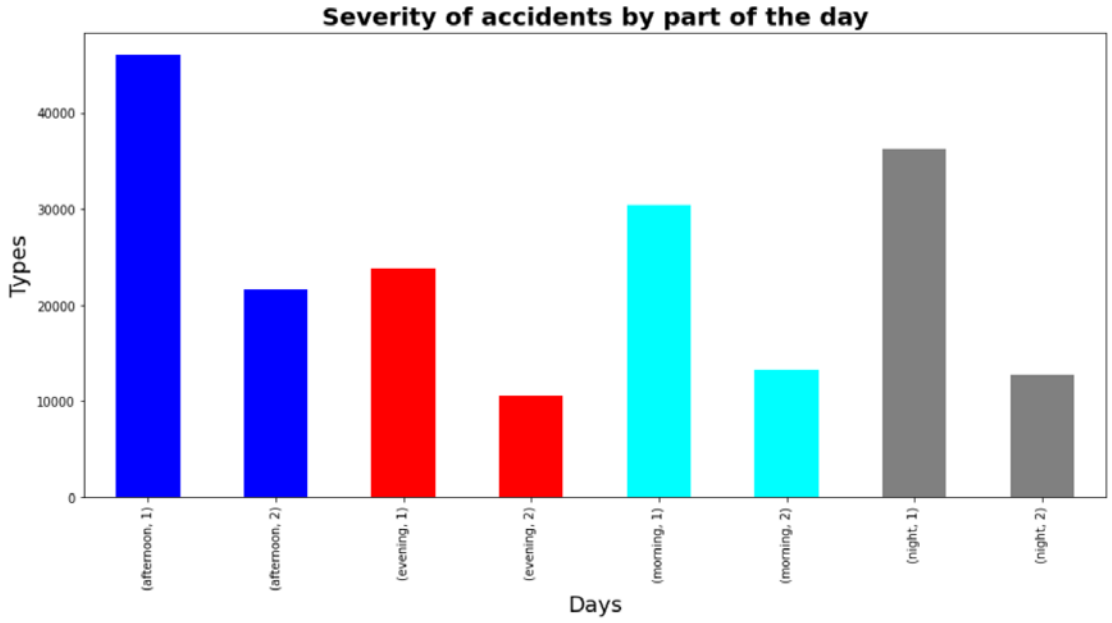
We will back to this but let’s keep exploring the data.

One hypothesis that could make sense is that the day of the week could be relevant in order to predict the severity of the accident. The problem is that we do not have a column with this value, but we can create a new column with the day of the wake with some transformations using de column 'INCDATE'.



Well, it looks like the day of the week does not have much impact on the severity of the accident, even when it looks like the friday the number of type 1 (property damage accidents), is higher, the proportion keeps constant. So, this will not be useful for our model.

Maybe the moment of the day has more impact, it is known that the drivers could be affect by the moment of the day in their behavior, for example getting sleepier after certain meals. Let's repeat the transforming exercise now with 'INCDTTM'.



The moment of the day does not have much impact on the severity of the accident. Even when in afternoons we can find a higher number of accidents (including type 1 and 2), the

proportion keeps constant and in the line of the rest of the parts of the day. As we didn't found relevant impact, **we drop time variables.**

3.2 Exploratory data analysis.

As we discovered in the point 3.1, our data needs to be balanced. We will use Synthetic Minority Oversampling Technique in order to solve this.

To see the difference between performing this action we will split the dataset appliying SMOTE and without SMOTE:

```
from collections import Counter

print("Before smote:", Counter(y_train))
print("After smote:", Counter(y_train_smote))
```

```
Before smote: Counter({1: 95402, 2: 40869})
After smote: Counter({2: 95402, 1: 95402})
```

3.3 Classification.

We are going to use the training set to build an accurate model with the help of different classification techniques in order to find the most accurate approach:

- **Decision Trees**: it is a Supervised Machine Learning technique where the data is continuously split according to a certain parameter.
- **K-Nearest Neighbor**: is a supervised classification method that serves to estimate the density function of predictors for each class
- **Logistic regression**: Logistic regression is named for the function used at the core of the method, the logistic function.

- **SVM:** In machine learning, support-vector machines (SVMs)are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis

4. Results

Decision Trees evaluation

	precision	recall f	1-score	support
Prop damage	0.74	0.99	0.85	41083
Injury	0.89	0.19	0.31	17319
accuracy			0.75	58402
macro avg	0.82	0.59	0.58	58402
weighted avg	0.79	0.75	0.69	58402

accuracy score	0.7527310708537379

K-Nearest Neighbor

	precision	recall f1	1-score	upport
Prop damage	0.79	0.79	0.79	41083
Injury	0.50	0.49	0.49	17319
accuracy			0.70	58402
macro avg	0.64	0.64	0.64	58402
weighted avg	0.70	0.70	0.70	58402

accuracy score	0.7037772679017842

Logistic Regression

	precision	recall f	1-score	support
Prop damage	0.75	0.97	0.85	41083
Injury	0.76	0.24	0.36	17319
accuracy			0.75	58402
macro avg	0.76	0.60	0.61	58402
weighted avg	0.76	0.75	0.70	58402

accuracy score	0.7526112119447964

SVM

	precision	recall f	1-score	support
Prop damage	0.75	0.99	0.85	41083
Injury	0.87	0.20	0.32	17319
accuracy			0.75	58402
macro avg	0.81	0.59	0.59	58402
weighted avg	0.78	0.75	0.69	58402

accuracy score	0.7538269237354885

		precision	recall	accuracy score	weighted avg F1
Desision Tree	Prop damage	0.74	0.99		
	Injury	0.89	0.19	0,75	0,69
K nearest	Prop damage	0.79	0.79		
	Injury	0.50	0.49	0,70	0,70
L Regression	Prop damage	0.75	0.97		
	Injury	0.76	0.24	0,75	0,70
SVM	Prop damage	0.75	0.99		
	Injury	0.87	0.20	0,75	0,69

Evaluating the resulting metrics of the 4 classification algorithms used, we can see how there are no major differences in performance between them. In any case, since both the accuracy score (0.75) and the Weighted avg F1 (0.70) are slightly higher than the rest of the models. However, given the nature of the problem to be solved, the most accurate model when predicting accidents that cause physical damage is the K nearest, with an accuracy for that value of 0.49.

5. Discussion

Although the model is not highly accurate, it can help to predict on the basis of the necessary conditions with 70% accuracy the nature of the accident, i.e. whether the damage to be suffered will be material or whether there could also be physical damage.

On this basis, drivers and authorities could vary routes, reduce speeds or delay trips according to the prediction obtained on the basis of the model built.

6. Conclusions

In relation to the dataset, in addition to the tasks of cleaning and preparation whose very nature makes them the most time-consuming phases, it has not been easy to work with an unbalanced dataset I believe that this fact influences the accuracy of the model in spite of having applied SMOT.

Nevertheless, it has been my first complete project applying ML and the phases of the