Data Mining report

Mohamed Moudjahed

April 2023

Contents

1 Problem Understanding (Introduction)

This data mining work aims to analyze road accident data in France between 2005 and 2021, in order to understand the characteristics of these accidents and determine the factors that influence them. We are particularly interested in the classification of the severity of accidents. The problem is to know if it is possible to predict the severity of an accident?
To answer this question, we will study data on the characteristics of the accident, its location, the vehicles involved and the users. We will also analyze the statistics on the number of accidents, weather conditions, light conditions, gender, etc. Finally, we will use the features of the accidents to classify their severity using machine learning algorithms.

2 Data Description

The available data are the annual databases of road traffic accidents in metropolitan France, in the overseas departments and in the other overseas territories. The data were collected between 2005 and 2021. The databases are extracted from the national file of traffic accidents (BAAC file), administered by the National Interministerial Observatory of Road Safety (ONISR). The data includes information on the characteristics of the accident and its location, the vehicles involved and the users. The files are in CSV format and consist of four files: Characteristics, Locations, Vehicles and Users. However, certain specific data relating to users and vehicles are hidden in order to respect the privacy of individuals.

3 Data Preparation

The format of the data changes from year to year. A work has been done so that the format of the data is the same in order to make an analysis on the years. For the classification part of our study we were only interested in the year 2021, we merged the tables corresponding to characteristics, users and vehicles in order to work with the characteristics that interest us.

4 Data Understanding

We notice that the number of accidents is decreasing every year because cars are more and more modern with various equipment keeping better alive their occupants as well as a more rigorous driving. Interestingly we notice an exception in 2020, the year when COVID started, a significant decrease of accidents is recorded. We can also look at the average number of accidents per month, we notice that at the beginning of the year there are more accidents, this could be due to the weather conditions.

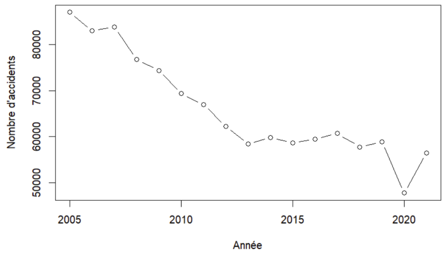

Figure 1: Number of accidents from 2005 to 2021

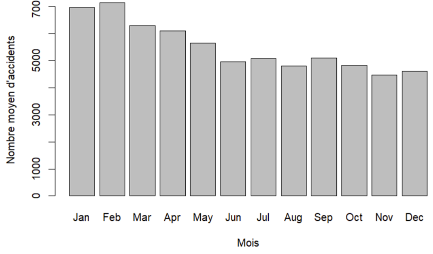

Figure 2: Average number of accidents per month from 2005 to 2021

It is also interesting to find out in which geographical area there are the most accidents. The data provide us with the latitude and longitude of each recorded accident. Thus we could locate each accident in a map of the world (France and overseas). The data does not provide us with information on the region of the accident but only the department of the accident. So with a csv which lists the correspondences departments/regions, we could display the regions having the most accidents.

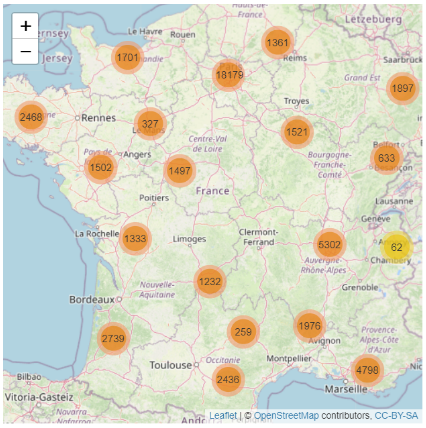

Figure 3: Dynamic map of accidents in France

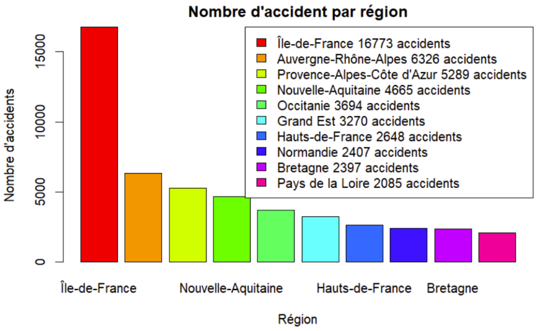

Figure 4: Number of accidents by region

We can also look at other interesting statistics:

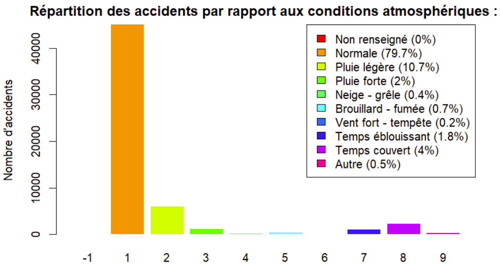

Figure 5: Distribution of accidents by weather conditions

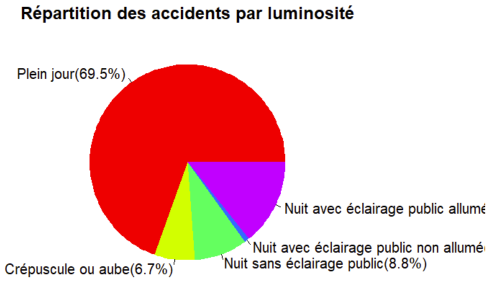

Figure 6: Distribution of accidents by light intensity

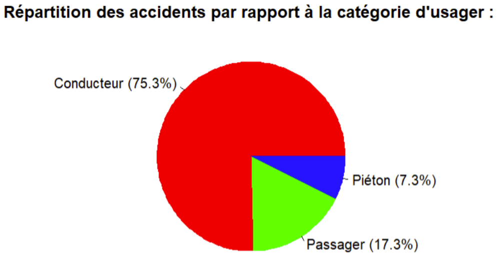

Figure 7: Distribution of accidents by user category

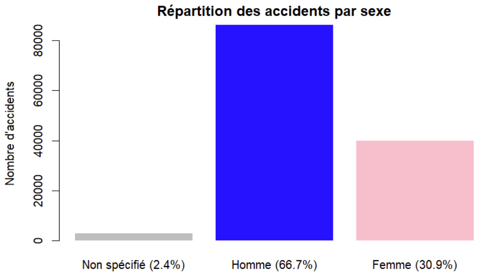

Figure 8: Distribution of accidents by gender

5 Classification of the severity of an accident

5.1 Preparation of the Dataframe

In this study we will classify the severity of an accident. We will try to predict the following labels:

- 0 - The accident is light
- 1 - The accident is serious

From the column “grav” which is composed of 4 different values (Unharmed (0), Killed, Injured (1), Hospitalized (1), Slightly injured (0))

In order to classify our data we will use the following characteristics:
lum, agg, int, atm, col, catr, circ, surf, situ, infra, vma, shock, manv, obs, catv, place, secu1, secu2, secu3, locp, actp

As indicated in the description of the databases, those related to user behavior are not given as the disclosure of these data would violate the protection of privacy, which is quite unfortunate.

5.2 Cleaning our data

5.2.1 Missing values

As a first step, we checked if we need to delete rows that give us insufficient information in order to improve the quality of the data by eliminating potential errors and avoiding bias in our classification results. The rows with missing data were judiciously removed by distinguishing between pedestrians or vehicular victims of accidents. We kept secu1, secu2, secu3, because the missing values correspond to the fact that a user has no security. After cleaning we removed 10210 values.

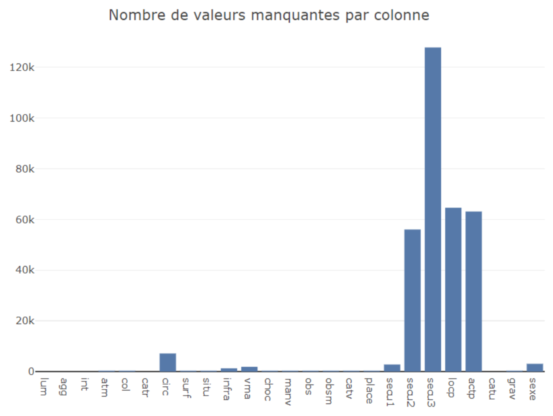

Figure 9: Number of missing values per features

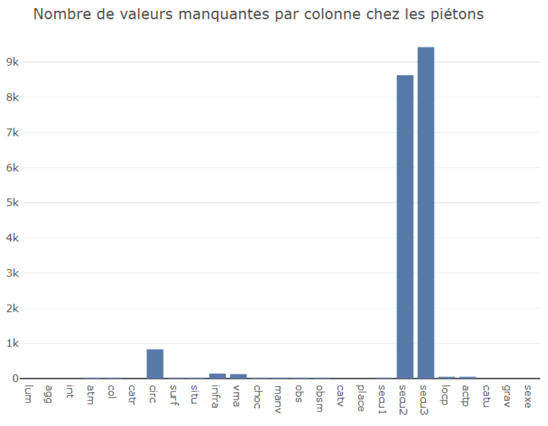

Figure 10: Number of missing values per features for pedestrians

5.2.2 Reduction of features

The presence of correlated data can lead to over- or under-representation of certain features. This can distort the results and make the models more complex. By removing them, we can improve the accuracy and generalization ability of the models.

We notice a strong correlation between ”place” and ”locp”. Indeed, ”place” tells us precisely where a user is located in a vehicle and if the user is a pedestrian. We can therefore remove the ”catu” column

The Gini index measures the importance of features in terms of their ability to reduce the impurity of the decision tree nodes, thus identifying the most important features for prediction. We retained only the 15 features with the most important Gini indices.

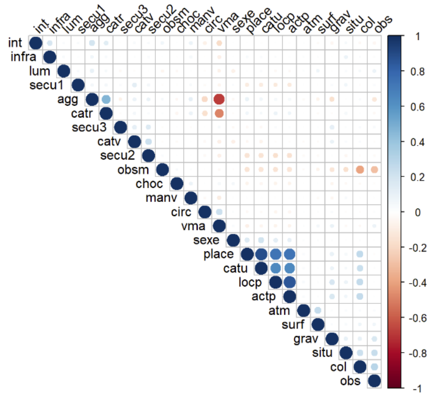

Figure 11: Correlation matrix

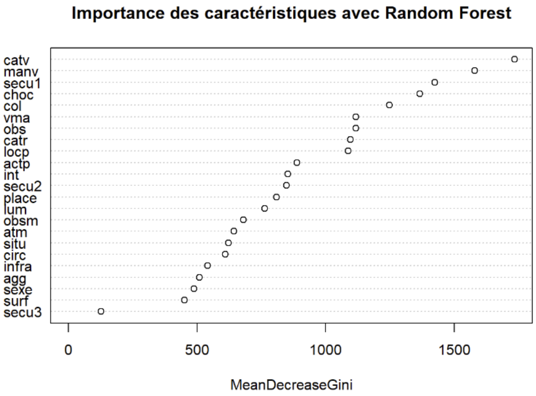

Figure 12: Importance of features with random forest

5.3 PCA Visualization

After PCA visualization on two principal components, we find that there is no clear clustering of data although there is a predominance of light accidents on the right side of our graph. It is therefore preferable to use non-linear classification approaches adapted to high data volume, such as Random Forest, Xgboost, Decision Tree and SVM with a non-linear kernel

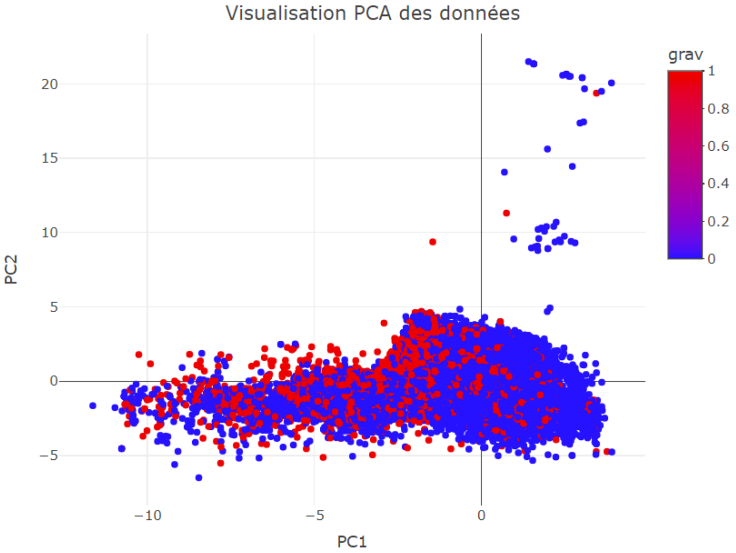

Figure 13: PCA visualization on two main components of our data

5.4 Use of ML algorithms

The results of the comparative performance analysis of the Random Forest, XGBoost, Decision Tree, and SVM models showed that all four algorithms had similar accuracies and F1 scores. All models were able to classify the data satisfactorily. With XGboost, we were able to achieve the best accuracy of 88%. However, it is important to note that the prediction time for each model was different. The SVM model needed more time to make predictions due to its more complex optimization algorithm that required more computation time. Random Forest and XGBoost made predictions faster than SVM, followed by Decision Tree.

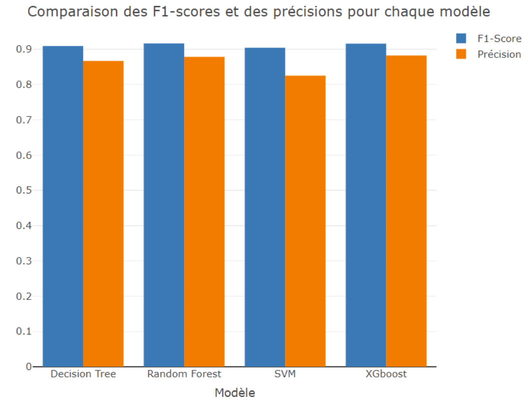

Figure 14: Accuracy and F1 score of each model

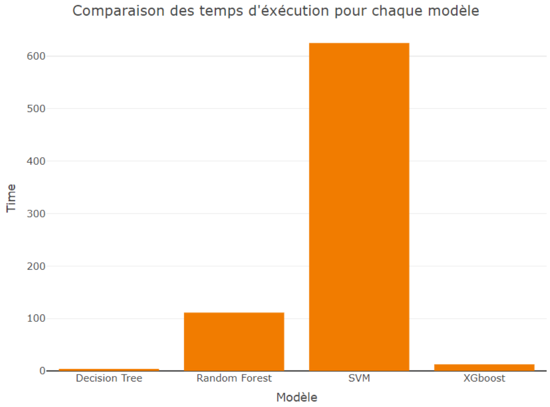

Figure 15: Training time of each model

6 Conclusion

The conclusions drawn from the analysis of these data indicate a significant relationship between the various factors that contribute to a crash and its severity. Data are proving to be a valuable resource for understanding complex phenomena such as traffic crashes. It is therefore essential to collect and examine data to improve our understanding of these phenomena, so that we can implement preventive measures to minimize the frequency and intensity of these crashes. Through this study, we used various data mining/analysis techniques which allowed me to get well acclimated with the R language. We can improve our models by tuning hyperparameters with cross validation with a good computer.