IBM Data Science Certificate

Applied Data Science Capstone Project

Predict Car Accident in France

Laurent Cesari

October 2020

PLAN

01 The problem

02 The data

03 Methodology

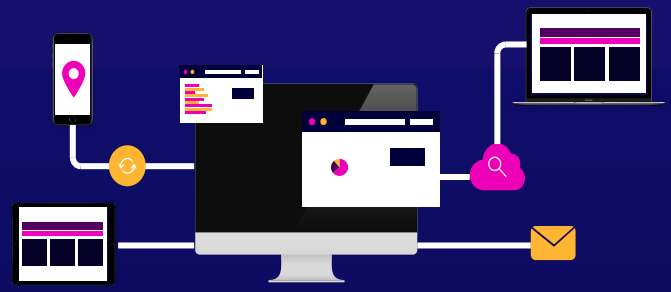
04 Results and discussion

ROAD ACCIDENTS – CONTEXT AND STAKES

- Worldwide, approximately 1.35 million people die each year as a result of road traffic crashes
- In 2019, nearly 3,500 people died on the roads of mainland France or overseas
- Overall cost of road accidents is estimated at €50.9 Billions representing 2.2% of the France GDP
- Build a model to predict the severity of road accident in France can be useful:
  - For government or local institutions to prevent accidents through a variety of measures: prevention, communication, renovation or enhancements of infrastructure...
  - For hospitals or healthcare services:
    - To anticipate the needs for medical services for certain periods of time and locations
    - When they are about to handle accident injuries and receive the accident detail from the emergency services

THE DATA COLLECTED

- The dataset is provided by the French Government with details on personal accident from 2005 to 2018 (data.gouv.fr)
- For each of these years, the data is split in 4 files:
  - Accident details
  - Victim details
  - Vehicle details
  - Location details



- We will use for our prediction model the severity of the accident that is provided with the victim details
- This dataset does not include any personal data that could allow us to link the accident information to the actual person involved in these accidents (no “human factor” such as alcohol / dugs consumption, vehicle speed,… no vehicle details such as brand, model,…)
- Due to a change in change in 2018 data impacting the classification of the severity, we will exclude 2018 sets
- In order to try to improve the model, we will include additional information such as holidays, weekend days…

METHODOLOGY

Explore and Prepare the Data

Build the Model(s)

Evaluate the model(s)

Collect the Data

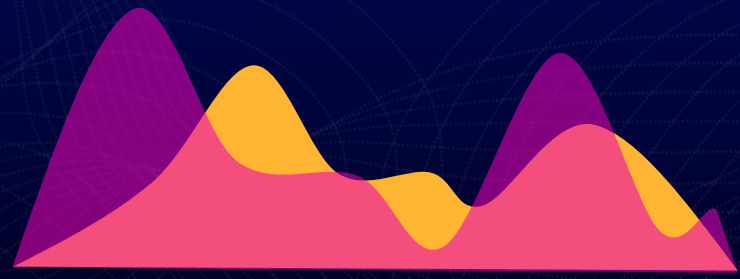
Select the Features

Fine-tune and validate the Model(s)

METHODOLOGY

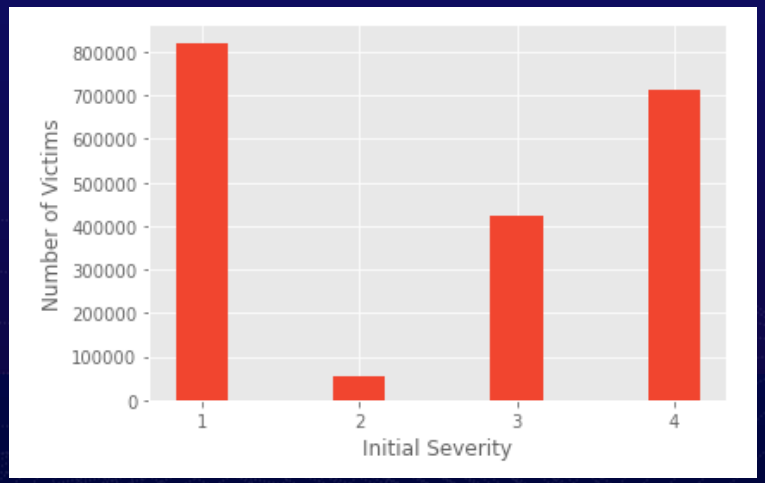
- DATA COLLECTION: after the concatenation and merge of the different files (51), I ended up with a dataset of 51 columns and 2,012,026 rows

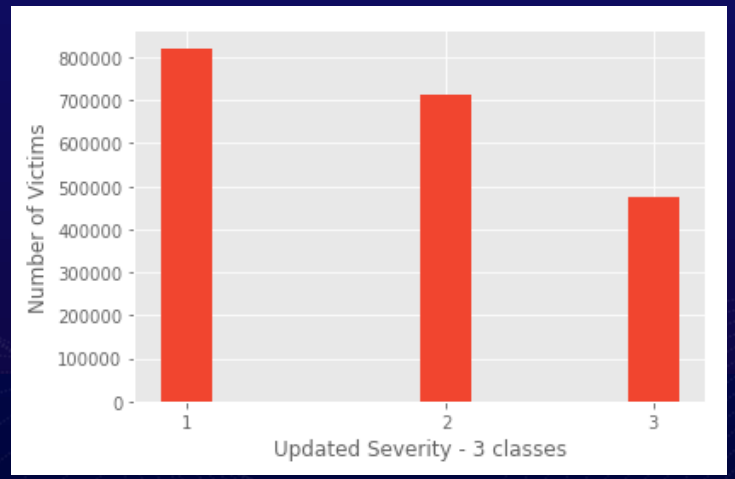
- DATA EXPLORATION AND PREPARATION:
  - 500 duplicate rows have been removed
  - Missing values have been replaced for 3 columns / features
  - Several columns have been removed considering:
    - They had too many missing information and / or inappropriate format (flow direct, number of persons in the vehicle, median strip width)
    - For some of them, they corresponded to location information (road details, gps code, longitude, latitude, road marker information, address) that we could replace with other existing features (city, county, road id)
  - In addition, 39,000 rows having missing values have also been dropped
  - Data for 4 columns / features had to be simplified and updated

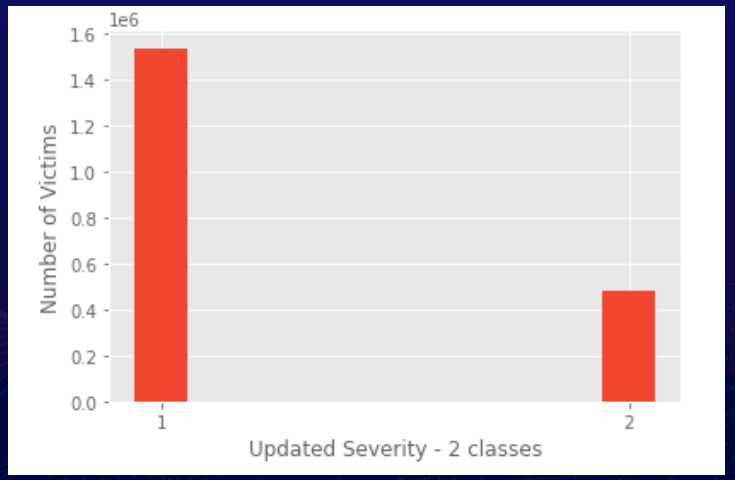


METHODOLOGY

- FEATURES SELECTION:
  - The classification for the severity has been reviewed and updated:
    - From 4 classes:
      1. Unarmed
      2. Killed
      3. Wounded – hospitalized
      4. Light injury
    - To a model of 3 classes:
      1. Unarmed
      2. Light injury
      3. Wounded – hospitalized or killed
    - And a model of 2 classes:
      1. Unarmed or Light injury
      2. Wounded – hospitalized or killed







- The goal being to test these classifications on our model(s) and keep the most providing the best performance

METHODOLOGY

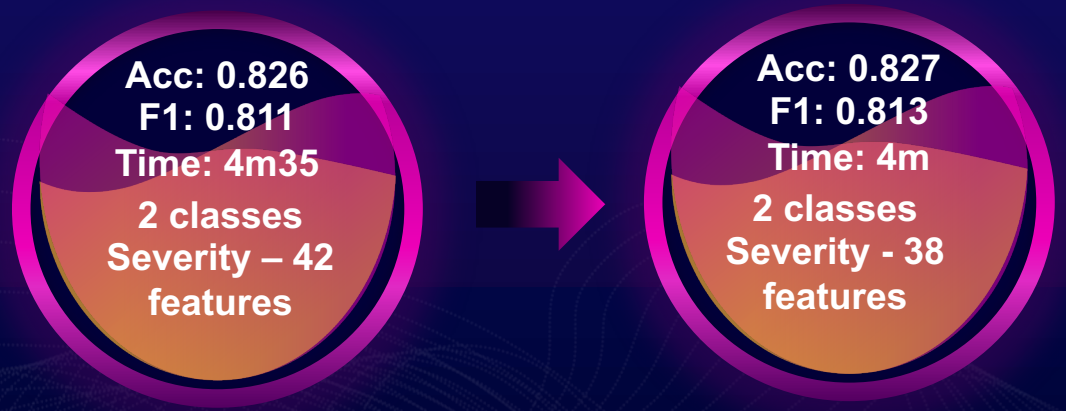
- FEATURES SELECTION:
  - Aimed at improving the accuracy of the model(s), 5 features have been added:
    - Linked to the time of the accident: hour, period of the day
    - Linked to the day of the accident: holidays, weekday, weekend
    - Linked to the birth date: age
  - After analysis of the remaining features and their correlation matrix, we decided:
    - To remove some additional columns which were not relevant to predict the severity of a future accident (year, date, accident ID, vehicle ID)
    - To keep the other features and use the modeling and validating phase to select the most relevant ones
  - After this phase, we ended up with:
    - 1,972,656 rows or observations
    - 3 potential classifications to train and select from
    - 42 potential features to train and select from

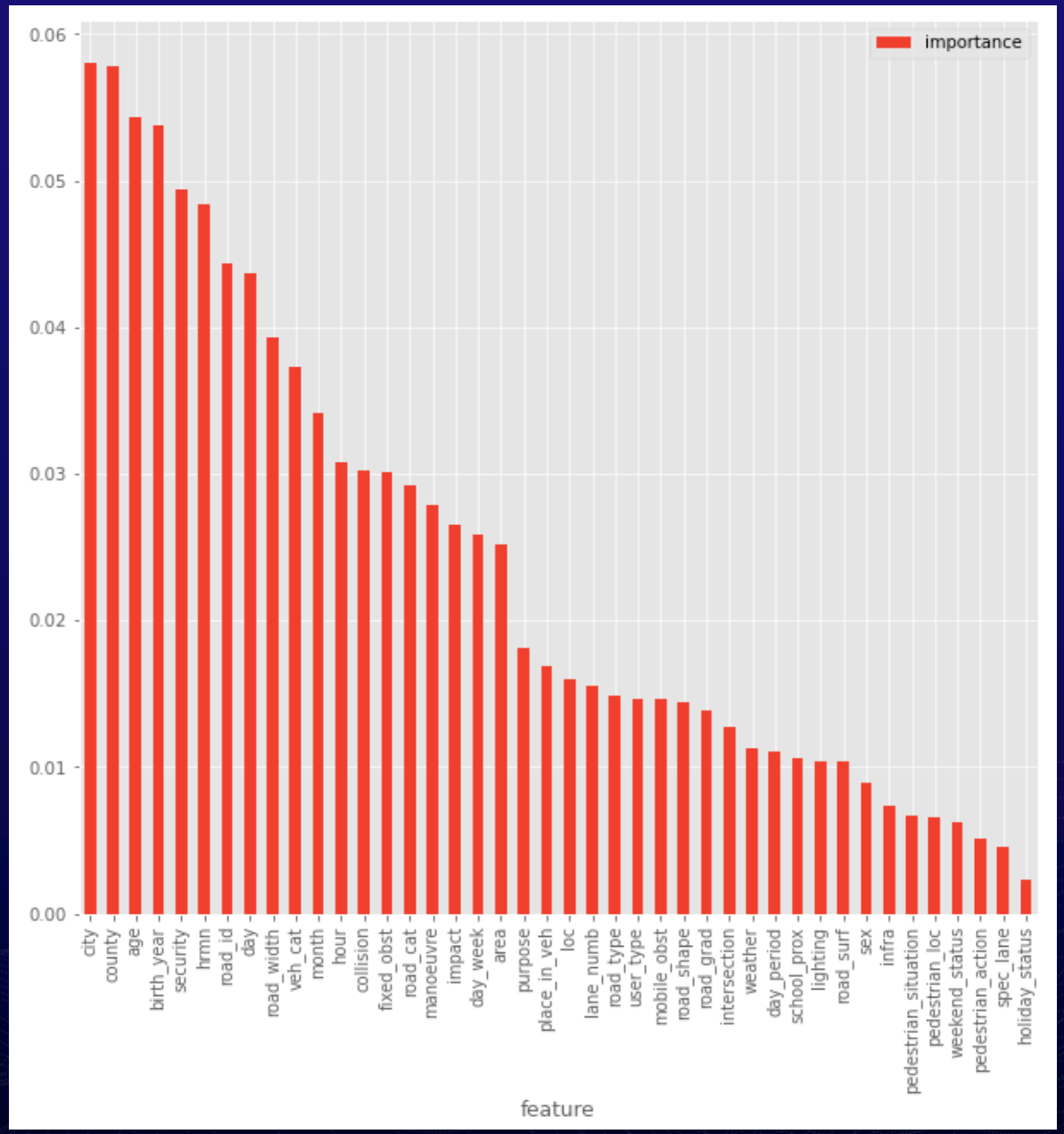
METHODOLOGY

- Model building & Validation :
  - I selected 2 models based on Decision Tree which, known to perform quite well for similar classification of structured categorical data :
    - Random Forest
    - XGboost Classifier
  - The dataset was split following the 80/20 rule into:
    - One training set, which was then split into a set dedicated to the training and a set for the validation of the model
    - A test set to evaluate the models
  - The Random Forest model with 2 classes severity showed a significantly better performance with a number of trees set to 100:

METHODOLOGY

- Model building & Validation:
  - The review of the features importance for the Random Classifier allowed to select some potentially redundant features (age over birth year, time of the accident over hour,…)
  - With these features removed, accuracy and F1 score slightly improved:





METHODOLOGY

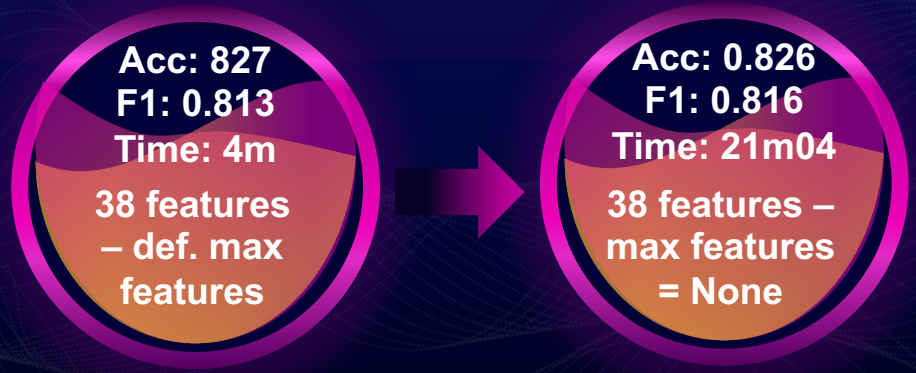
- Model building & Validation :
  - Reduction of the number of features had a slight impact on the processing time and performance

  - Depending on the goals and information available to predict the severity of an accident, we can adapt the selection of our features to maximize the accuracy of our model
  - Considering the gain in performance from 5 to 10 features, we might recommend to start with a model having at least 10 or 15 features
  - If we can have the information, we can use the model with all the the features (no significant different in the processing time)

METHODOLOGY

- Model building & Validation :
  - Due to the size of the dataset and the technical limit of my environment, I could not implement a proper fine tuning of the model hyperparameters (grid search function / validation curves)
  - Though, some tests showed some possibilities of slight improvements notably with the increase of number of trees but required a significant increase of the processing time (especially for 300 trees)

  - While other parameter change did not improve the performance (max features = None)



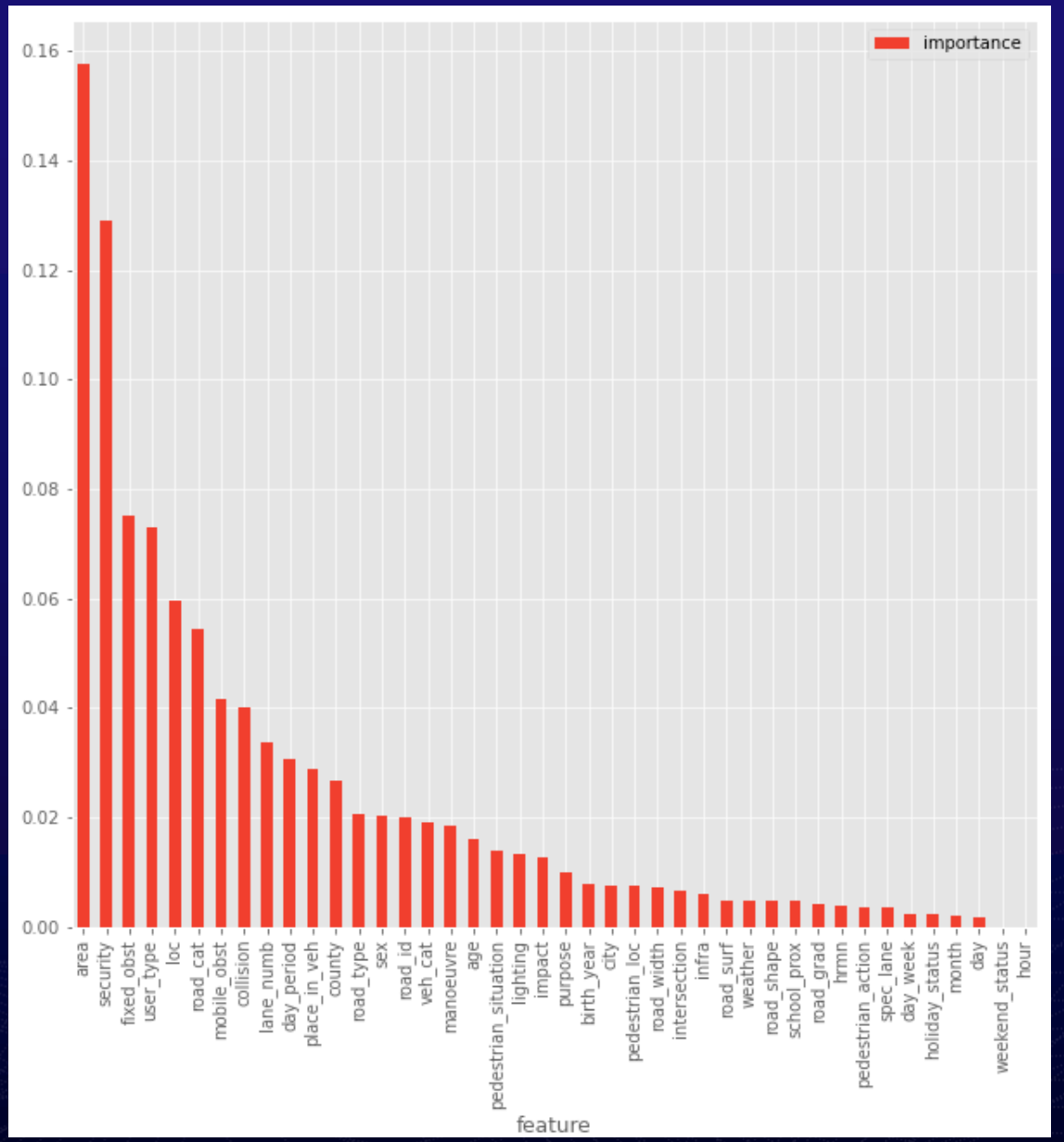
METHODOLOGY

- Model building & Validation :
  - As with the Random Forest, the XGBoost Classiffier model with 2 classes severity showed a significantly better performance with a number of trees set to 100 and all the features:

  - Which were actually slightly better than the Random Classifier with a lower processing time

METHODOLOGY

- Model building & Validation:
  - The review of the performances for this model revealed a different ranking than with the Random Forest
  - The review of the features importance for the Random Classifier allowed to select as well some potentially redundant features different from the ones of the Random Forrest (age over birth year, time of the accident over hour,…)
  - With these features removed, accuracy and F1 score slightly improved:



METHODOLOGY

- Model building & Validation :
  - Reduction of the number of features had as well a slight impact on the processing time and performance

  - Similarly to the Random Forest model:
    - We can choose the model that fits the amount of information we have to predict the severity of the accident
    - Considering the gain in performance from 5 to 15 features, we might recommend to start, if possible, with a model having at least 15 features
    - Depending on the goals in terms of accuracy, we can use whether the model with 25 features or the model with all the features if we can have all the information

METHODOLOGY

- Model building & Validation :
  - As well as for the Random Forest Model, technical environment did not allow to implement a proper fine tuning of the model hyperparameters (grid search function / validation curves)
  - The increase of number of trees showed an increase of the performance but this time with a more limited increase of the processing time

  - Keeping the 300 trees as parameter, a last test between the model with all features vs a model with 25 features showed a decrease of the accuracy and F1 Score

METHODOLOGY

- Evaluation of the model:
  - For the evaluation, considering the trade-off performance – processing time, models were set with the following parameters:
    - Features selection: all the features (38) for both models
    - 200 trees for the Random Forest and 300 for the XGBoost Classifier
  - Using the test sets, we got similar results if not slightly better scores:

  - As we could expect, the XGBoost Classifier offers slightly better performance both in terms of score and processing time.

DISCUSSION

- We managed to get a model that performed fairly well:
  - With the reduction of the number of classes for our severity classification from 4 to 2,
  - While keeping some room for improvement
- The best performing model(s) were the one(s) including a significant number of features
- The importance of these factors were not always the same for the different algorithms
- Depending on the goal and use of this model, we might not have the information for all the features (emergency service receiving the information of an accident to handle the potential victims)
- Fortunately, a more limited number of features (from 5 or 10) was already providing an accuracy or around 80%
- For anticipation purpose (prevention / forecast of accidents by governments, institutions…), we should be able to provide the model with more information and to reach a higher accuracy.

DISCUSSION

- How we could improve this model:
  - Perform a proper proper fine-tuning of the hyperparameters, which might require to limit our dataset to the last 10 or even 5 years,
  - Implement some clustering techniques on some features that had an important number of unique values such as the location, days, time of the day,…
  - Adapt the severity classification to include the number of victims involved in one single accident
  - Include some factors which are often considered as very important in terms of impact on road accidents:
    - Human factors such as alcohol / dugs consumption, vehicle speed vs maximum speed limit, use of mobile…
    - Vehicle characteristics such as car type, power engine, age of the vehicle…
- Although the models selected provided some fairly good scores in terms of performance, we could also try other classification models not relying on Decision Tree (such as logistic Regression, support vector machine, neural networks….)

CONCLUSION

- Based on a dataset of road accidents in France from 2005 to 2017, we have built a model to predict the severity of a road accident
- This model could be used by government or local institutions to prevent such accidents or for hospitals or healthcare services to anticipate the needs for medical services
- Using models based on Tree Decision models, we managed to reach 83.15% of accuracy while reducing the percentage of false positives / negatives to less than 18% (precision and recall scores being both quite close to the F1 score)
- The best of the 2 models we tested was the XGBoost Classifier
- These results leave some room for improvements:
  - When it comes to the data:
    - By adding human factors or vehicle characteristics
    - By including in the severity the weight of the number of victims per accident
  - Regarding the model:
    - By performing a appropriate fine-tuning of the hyperparameters, which might require to limit the number of years of our dataset
    - By implementing some clustering techniques to handle some features with an important number of unique values
    - And maybe by testing other types of algorithms to see if they provide better results

THANK YOU