We think it would be interesting to have a software capable to predict the category of a crime just analysing the content of the its summary, for the capital. To do so, we had to clean and restructure the data in order to obtain better results. First, we just got the incidences that occur in Kabul and, using a glossary table it has been possible to replace the acronyms for their respectively real meaning. This is an important step, but not the only one to clean the data, since it contains many stop words that can give us a non-reliable result. Later, we create a matrix that represents the incidences on the rows and every unique word on the columns. With this matrix, it’s really easy to represent the words that every summary contains.

We have used 4 classifiers in this project, to be able to see the differences between them. The obtained scores are, respectively:

Gaussian NB 🡪 39.44%

K-Nearest Neighbours 🡪 61.03%

Decision Tree 🡪 69.64%

Random Forest 🡪 77.15%

As we can appreciate, the best score is obtained using the Random Forest classifier, which doesn’t surprise us having in mind that is the most complex algorithm.

On the other hand, we also get a reasonably good score applying the Decision Tree algorithm. It’s normal that is lower than the Random Forest, since this one tries to improve the Decision Tree classifier.