**Feature Selection**

This first notebook contains all the steps and transformations I performed for the feature selection. The datasets contain an extended descriptions of different aspect of the accidents, thus I've selected the most relevant and useful data for my analysis.

**Predicting Traffic Accident Severity**

* Data Description -
* Data Cleaning -
* EDA -
* Model Development -
  + Random Forest -
  + Logistic Regression -
  + KNN -
  + SVM -
* Results -:

| **Algorithm** | **Jaccard** | **f1-score** | **Precision** | **Recall** | **Time(s)** |
| --- | --- | --- | --- | --- | --- |
| Random Forest | 0.722 | 0.72 | 0.724 | 0.591 | 6.588 |
| Logistic Regression | 0.661 | 0.65 | 0.667 | 0.456 | 6.530 |
| KNN | 0.664 | 0.66 | 0.652 | 0.506 | 200.58 |
| SVM | 0.659 | 0.65 | 0.630 | 0.528 | 403.92 |

For this specific problem precision means the % of predicted severe accidents that were truly severe. The recall instead, is the % of truly severe accidents that were properly predicted. For this specific problem, the recall is more important than the precision as a high recall will favor that all required resources will be equipped up to the severity of the accident. In this case, the recall is more important than the precision as a high recall will favor that all required resources will be equipped up to the severity of the accident. The logistic regression, KNN, and SVM models have similar accuracy, however the computational time from the regression is far better than the other two models. With no doubt the Random Forest is the best model, in the same time as the log. res. it improves the accuracy from 0.66 to 0.72 and the recall from 0.45 to 0.59.