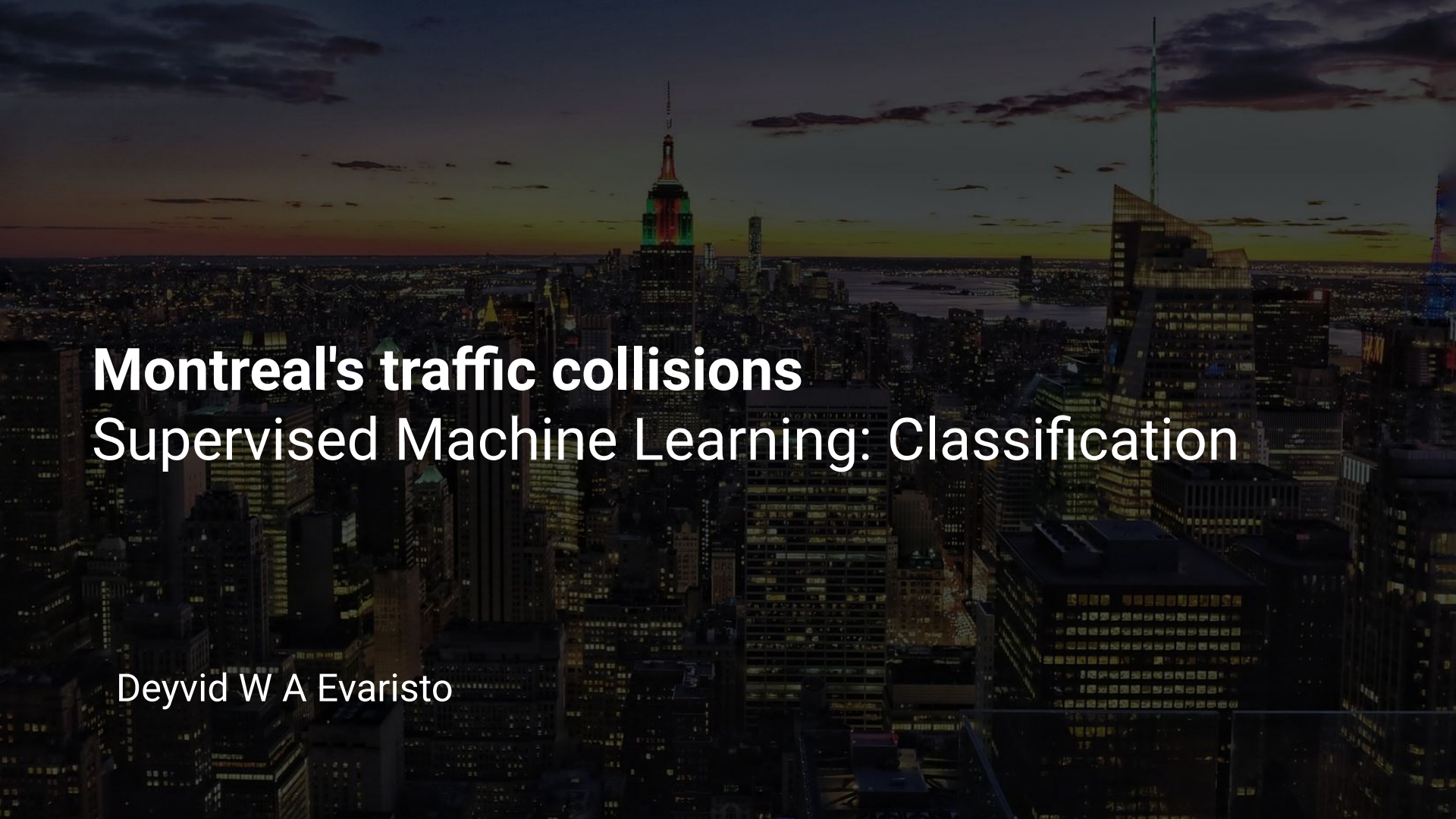


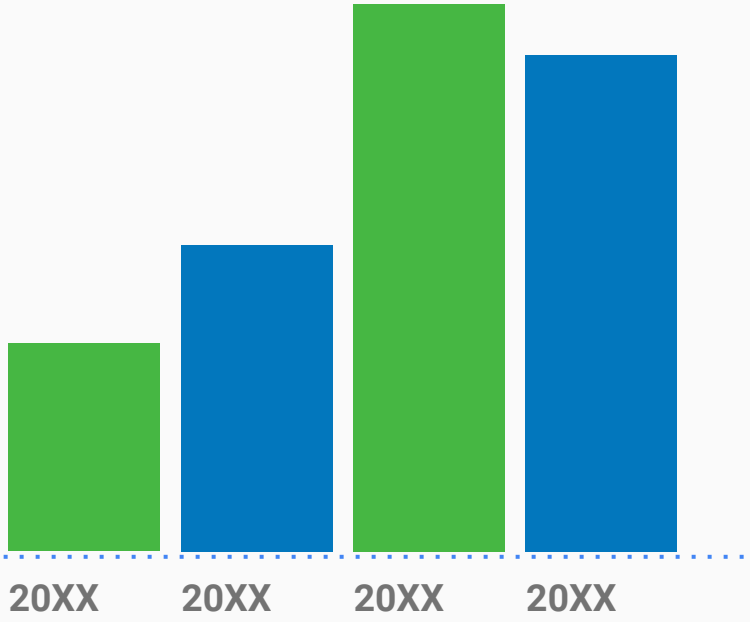
Montreal's traffic collisions

Supervised Machine Learning: Classification

Deyvid W A Evaristo

Objective

To apply a supervised machine learning algorithm that analyzes a public dataset containing Montreal's traffic collision records to predict the severity (Major/Minor) of a new accident.



The dataset

The dataset is a list of traffic accidents in Montreal registered by the SAAQ (Société de l'assurance automobile du Québec) and the SPVM (Service de Police de la Ville de Montréal) from 2012 to 2019.

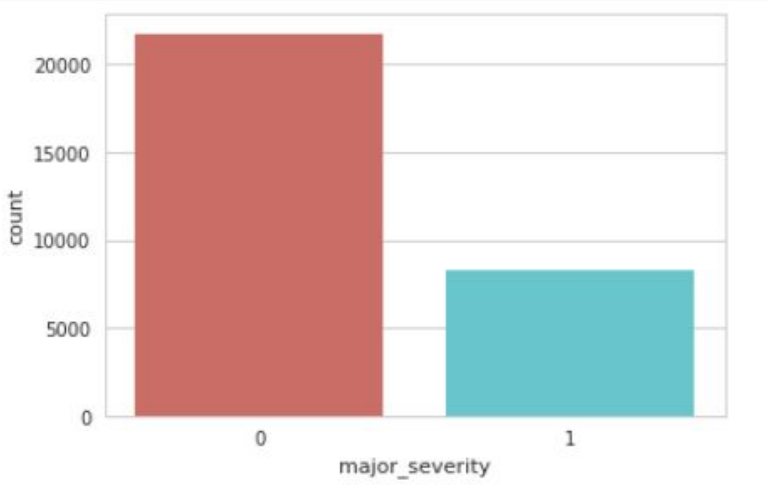
It is available on the City of Montreal's Open Data website and accessible using a public web API with query support.

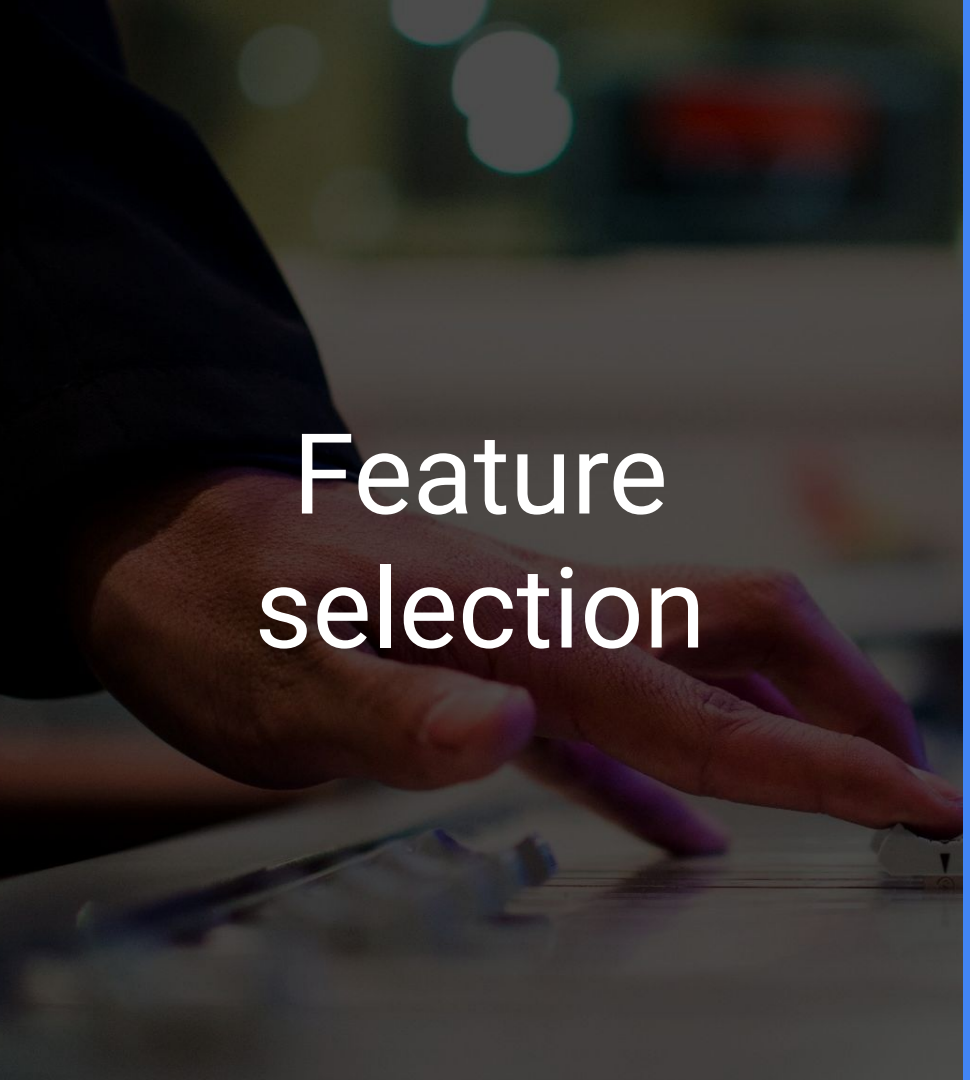
The approach

Through **supervised learning**, the computer program learns from the input variables (x) and an output variable (Y) and then uses this learning to classify new observations.

Binary **Classification** will be the approach used to predict the 2 classes to which a new data will fall under: severity (Major/Minor).

Databricks notebooks will be used with Python, Scikit-learn and other libraries.

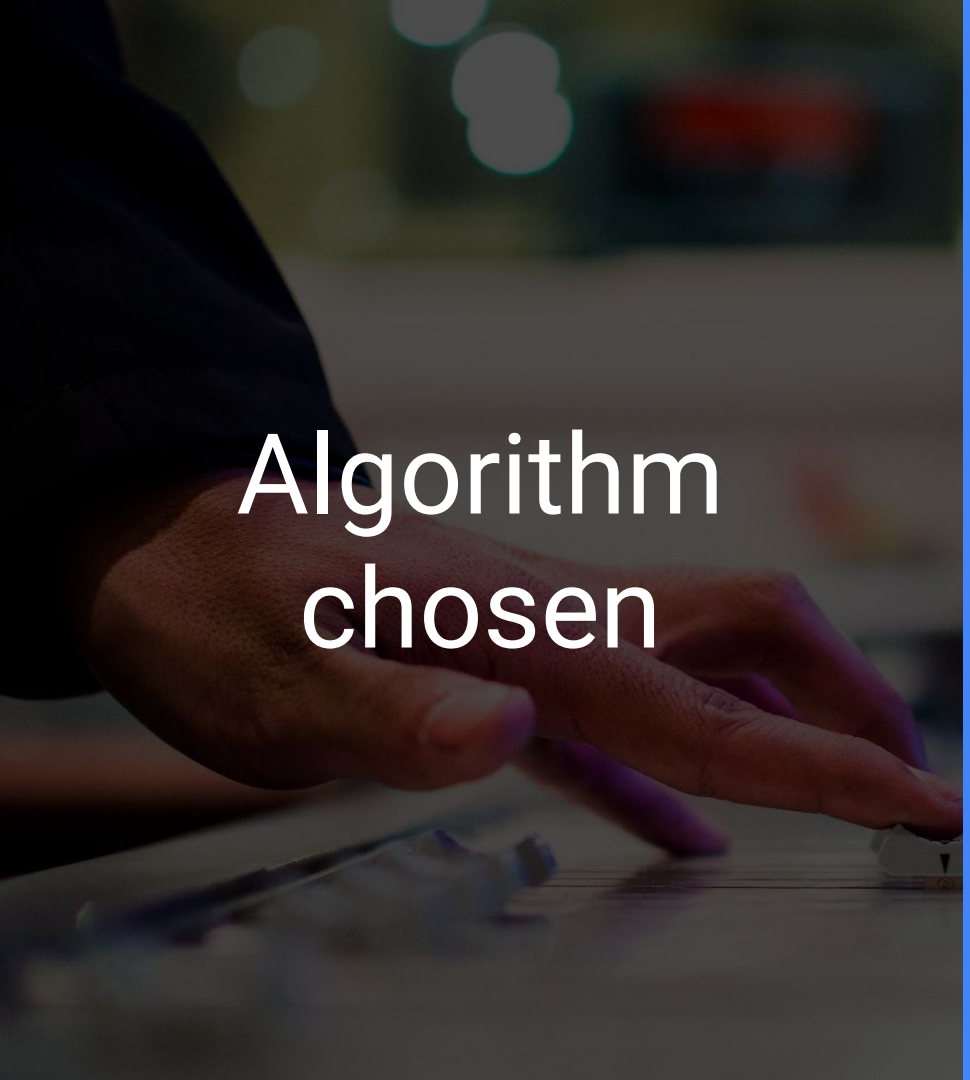




Feature selection

These were the seven elected variables after some exploratory data analysis:

- Month
- Day of the week
- Time of the day
- Max speed of the road
- Number of vehicles involved
- Condition of road surface
- Pedestrian/Cyclist involved

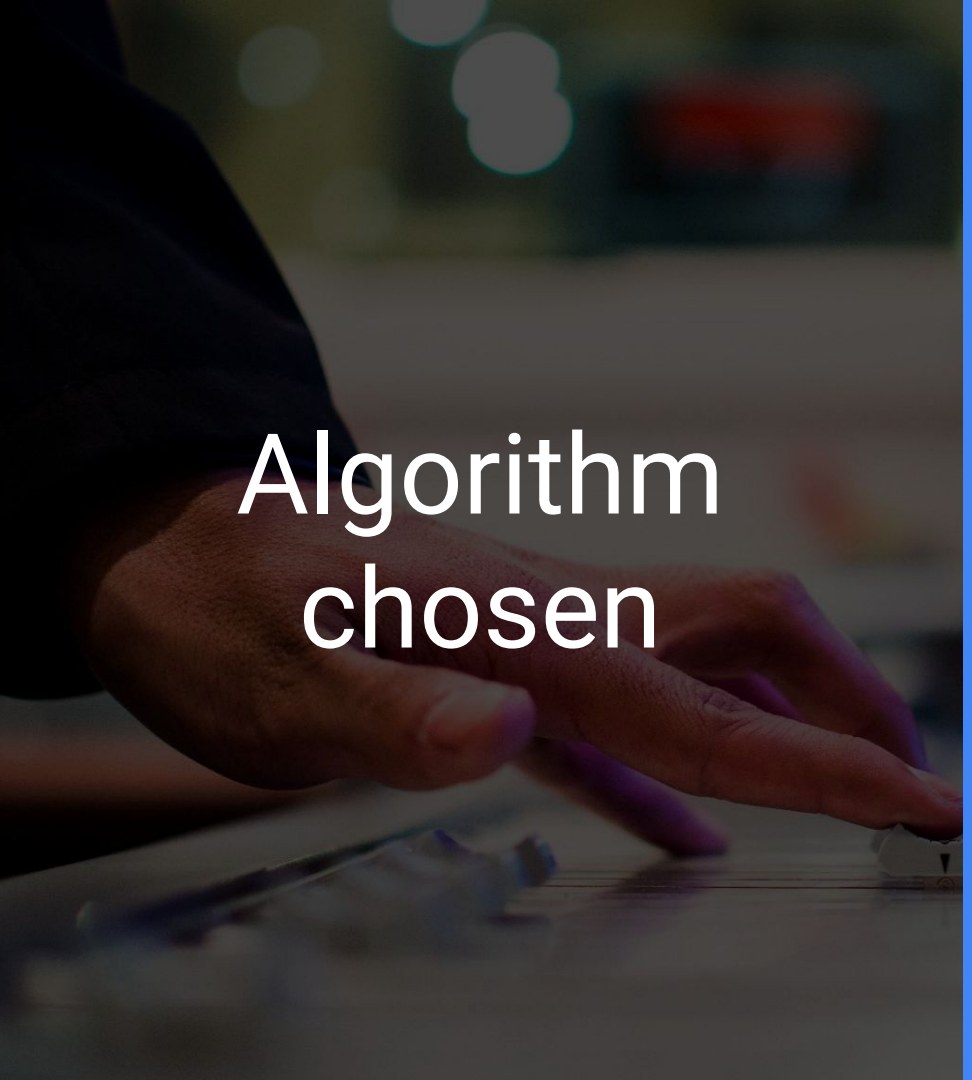


Algorithm chosen

After tests using different algorithms, Gradient boosting classifiers was the one which performed best and had the highest accuracy: 79%

Gradient boosting classifiers is a package from scikit learn library which combines a group of machine learning algorithm to create a strong predictive model.

Because it is based on Decision trees, it performs well on imbalanced datasets as its hierarchical structure allows them to learn signals from both classes (Major/Minor).



Algorithm chosen

Accuracy of Gradient Boosting classifier (Imbalanced): 0.79

	precision	recall	f1-score	support
0	0.79	0.97	0.87	11044
1	0.81	0.32	0.46	4297
accuracy			0.79	15341
macro avg	0.80	0.65	0.67	15341
weighted avg	0.79	0.79	0.76	15341

Feature Importance	
pedestrian_cyclist	0.895136
max_speed	0.046901
number_veh	0.036943
surface	0.015381
month	0.003927
hour	0.001588
week_day	0.000124

dtype: float64

Demo