CLASSIFICATION OF SEVERITY FOR NEW TRAFFIC EVENTS

LEONARDO VELA

APPLIED DATA SCIENCE CAPSTONE

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

BUSINESS UNDERSTANDING

DO WE ABLE TO PREDICT THE SEVERITY OF A NEW TRAFFIC ACCIDENT?

BY MEANS OF MACHINE LEARNING ALGORITHS WE CAN ESTIMATE OR PREDICT INFORMATION.

BUSINESS UNDERSTANDING

WHO CAN BE INTERESTED IN THIS NEW INFORMATION?

GOVERNMENT ENTITIES AND EMERGENCY SERVICES FOR INSTANCE

BUSINESS UNDERSTANDING

CAN PREDICTIONS HELP TO IMPROVE THE QUICK ANSWER FROM EMERGENCY SERVICES?

PREDICTIONS CAN HELP TO IDENTIFY WHAT KIND OF TEAMS AND EQUIPMENTS WILL BE REQUIRED MOST OFTEN

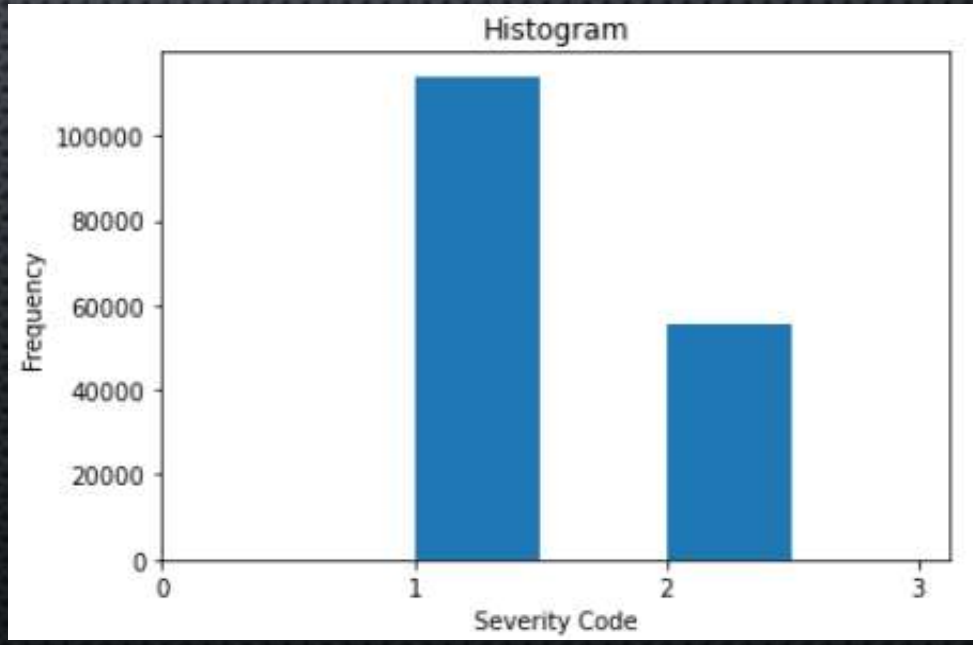
DATA UNDERSTANDING

- DATA RECORDS OF TRAFFIC EVENTS SINCE 2004 TO PRESENT, WITH 194673 RECORDS AND 38 FEATURES OF SEATTLE CITTY WAS USED TO TRAIN PREDICT MODELS

DATA UNDERSTANDING

- THE TARGET TO BE PREDICT HAS 5 POSSIBLE CODES:
- -3. FATALITY.
- -2B. SERIOUS INJURY.
- -2. INJURY.
- -1. PROP DAMAGE.
- -0. UNKNOWN.

SEVERITY CODES PRESENT IN DATASET



DATA PREPARATION

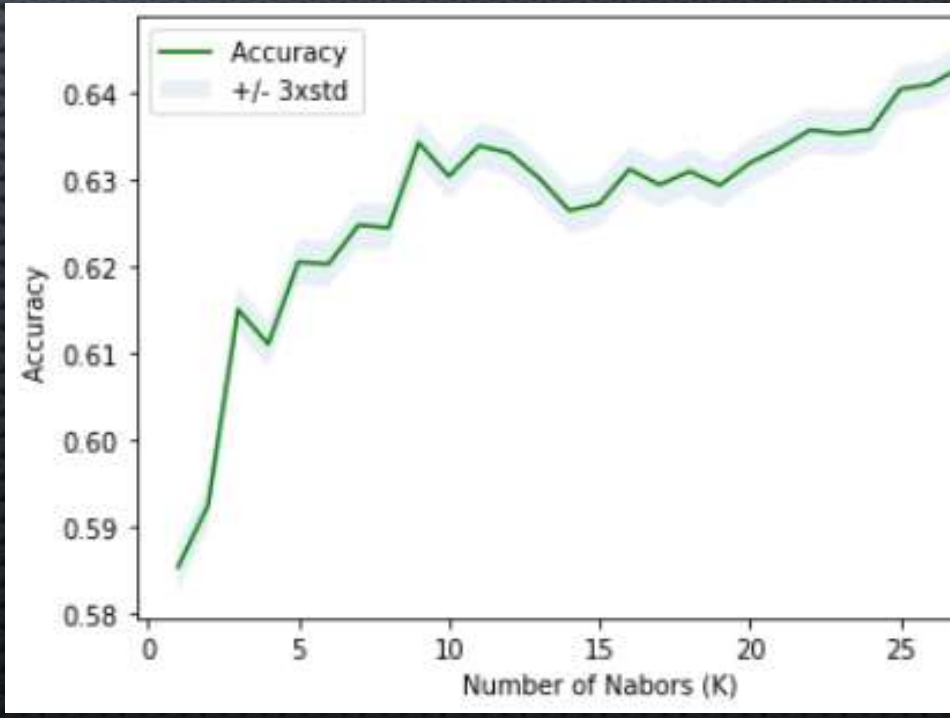
1. DEAL WITH NO DATA
2. FORMAT FEATURES
3. TURN INTO NUMERICAL VALUES CATEGORICAL FEATURES
4. EXTRACT HIGHEST TRAFFIC HOUR FEATURE FROM DATE
5. BALANCE DATAFRAME

MODELLING

- BALANCED DATAFRAME 111.414 RECORDS
- TRAIN SET 77.990 RECORDS
- TEST SET 33.424 RECORDS (30%)

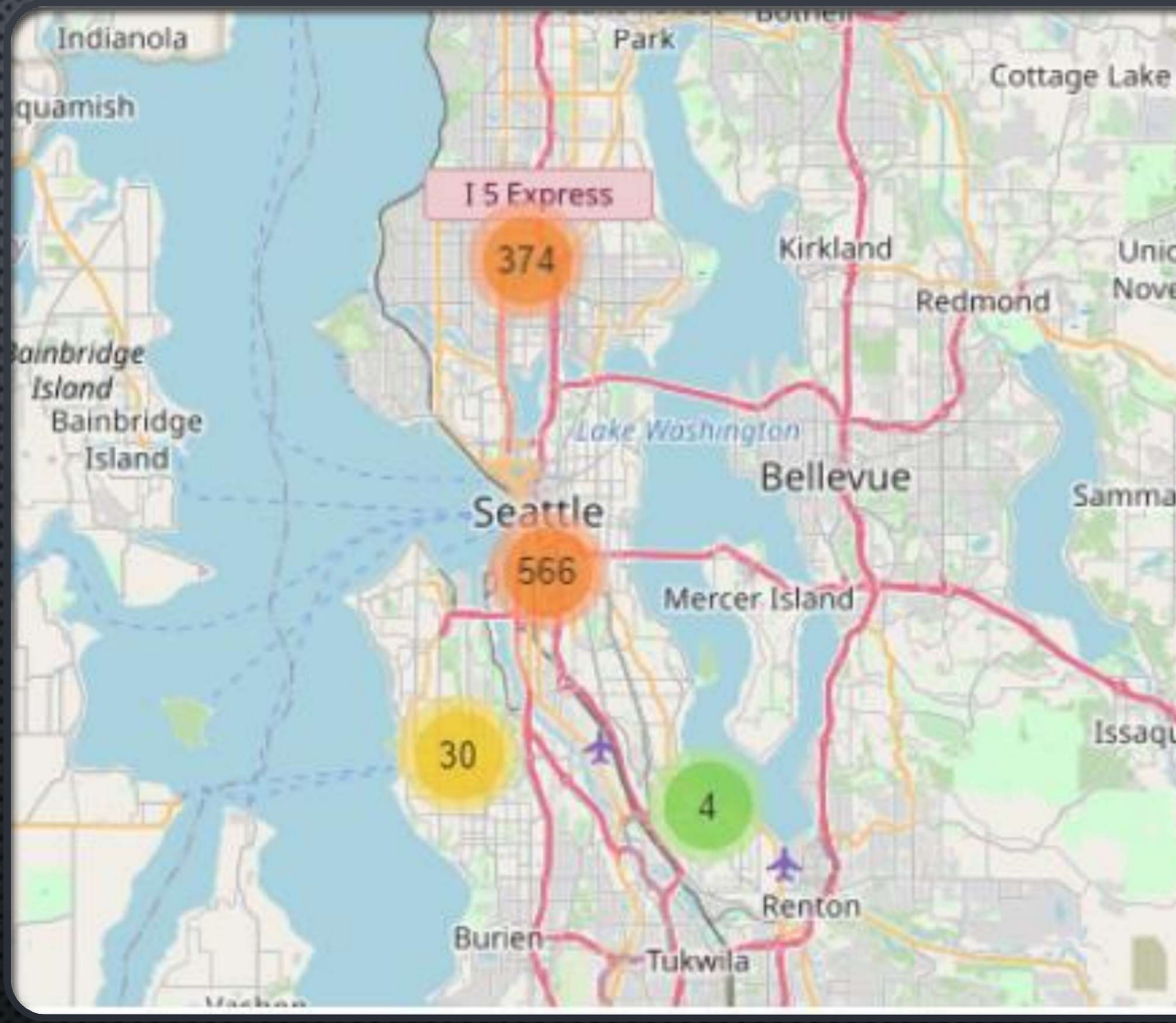
MODELLING

- K-NEAREST HIGHEST ACCURACY 0.6429 WITH 27 K-NEIGHBORS
- DECISION TREE ACCURACY 0.6413
- SUPPOR VECTOR MACHINE (SVM) 0.6474



RESULTS AND DISCUSSION

- UNBALANCED DATASET WITH RATIO 2:1, EVERY 2 PROP DAMAGE 1 INJURY
- HIGH FREQUENCY N DOWNTOWN AND SURROUNDING AREAS OF GREEN LAKE
- LOW MODEL´S ACCURACY, IT COULD BE DUE TO LOW CORRELATION AMONG FEATURES, HIGHEST CORRELATION 0.2153 AND LOWEST –0.00003
- HIGHEST MODEL'S ACCURACY WAS IN SVM WITH 0.6474



CONCLUSIONS

- Computer tools helps to save time in processing steps, this allows invest more time in analytics and creative thinking. However, pre-processing data takes time because requires understand very well the business problem and the available data

- Due to the data juts have 2 codes regarding to the 5 classificatory codes the machine learning algorithms will not be able to predict different codes.