Datlas Challenge

Team: Enigma

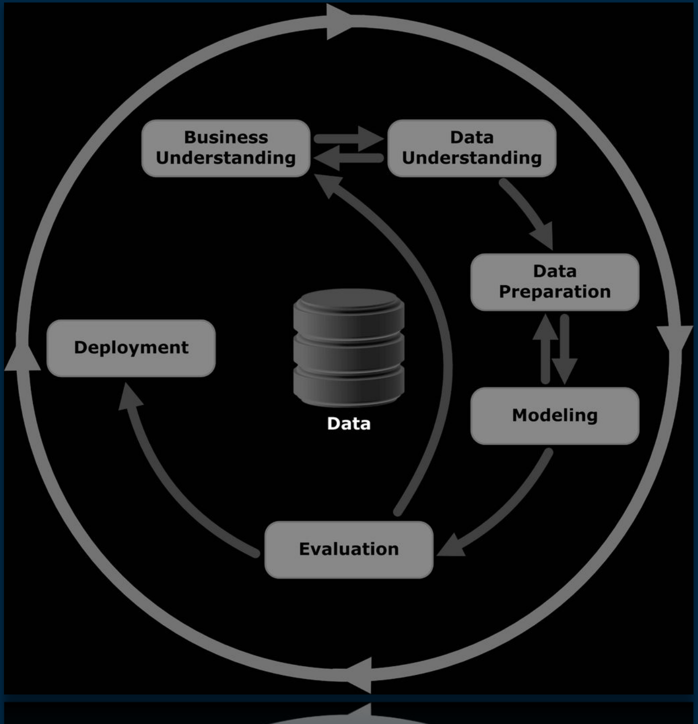
Ramon Díaz
Daniela Gómez
Michael Zenkl
Jorge Ayala

Objective

To obtain insights from Nuevo Leon’s car accidents with the use of data analysis and machine learning methods.

Methodology

- CRISP-DM Methodology (Cross-industry standard process for data mining)
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deploying



Data Cleaning

- Missing Values (statistical imputation)
- Binned Variables with Multiple Labels (Model Years)
- Removed Special Characters (Car Colors)

Feature Engineering

- Climate conditions
- Type of road identification
- Nearby pedestrian crossings
- Geohashing of coordinates in different areas (from 5 to 8 hash characters)
- Identification of holidays

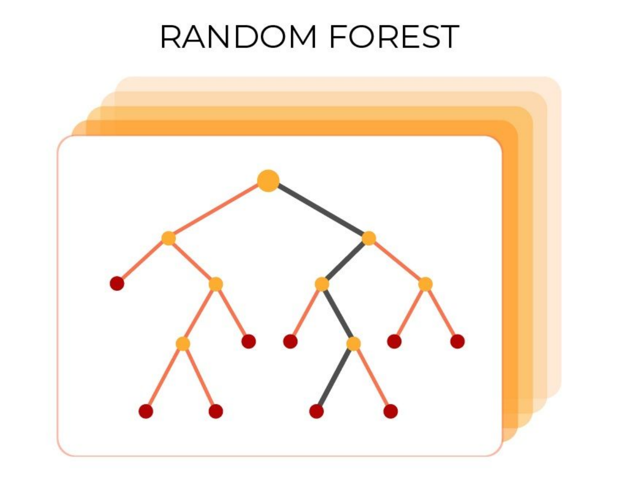
Data Preprocessing

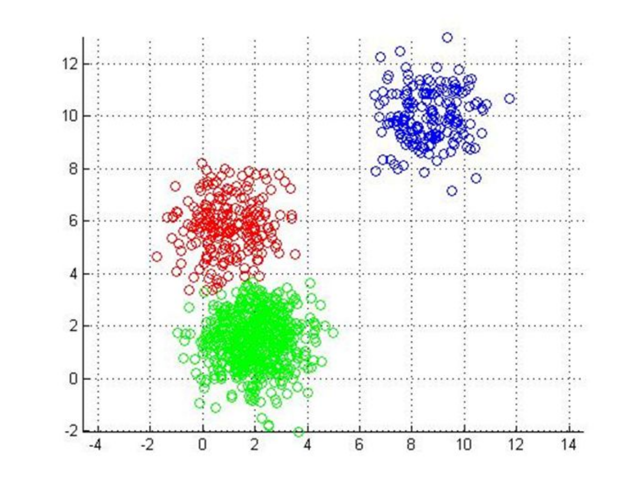
- Standardization : MaxMin method
- Recursive Feature Elimination for Feature Selection

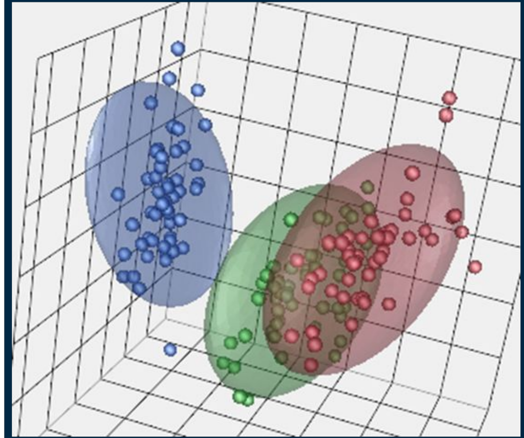
Tools

https://github.com/danisha20/project_DATLAS

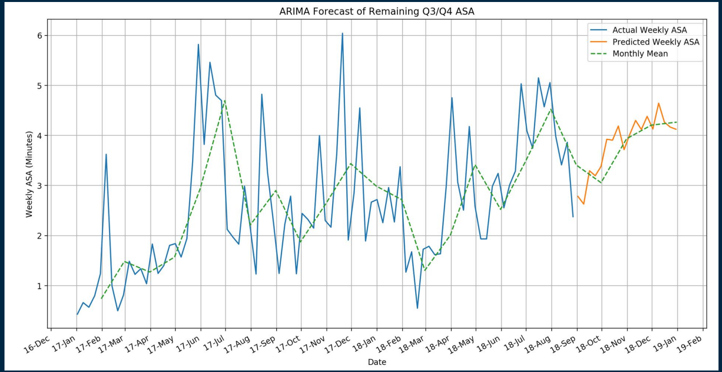
Methods



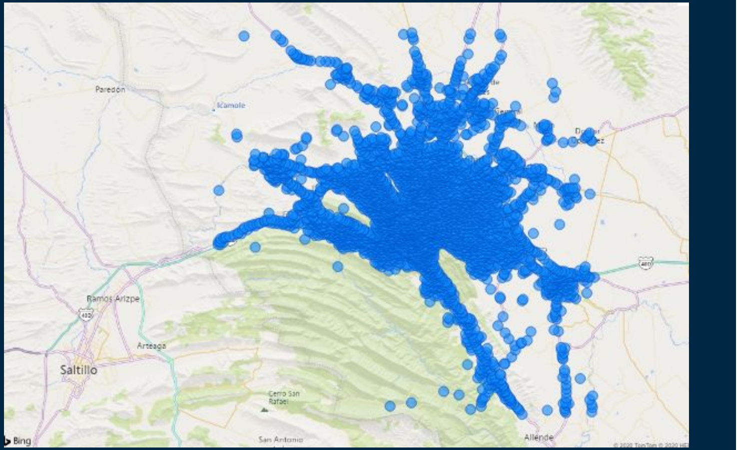




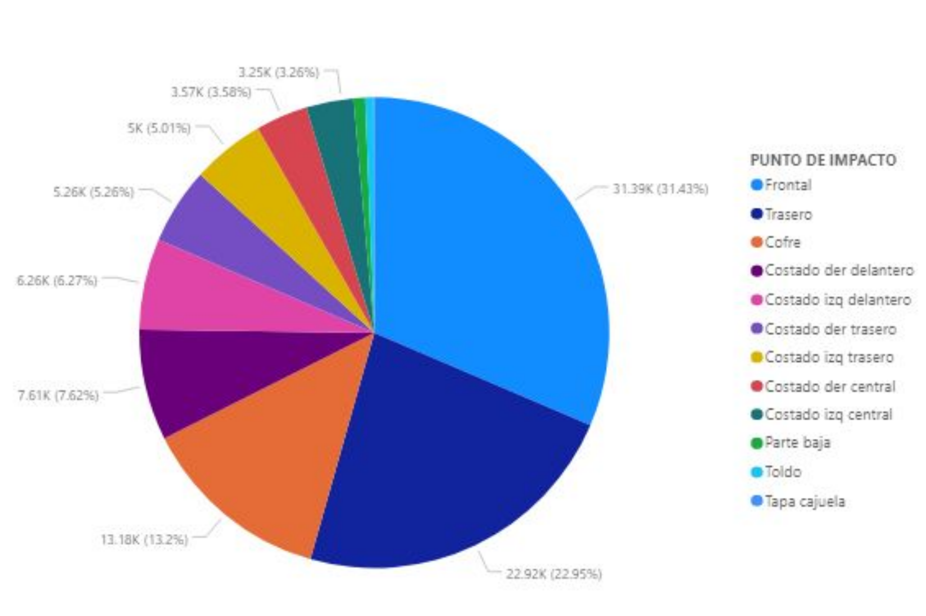
K-means
K-modes



EDA

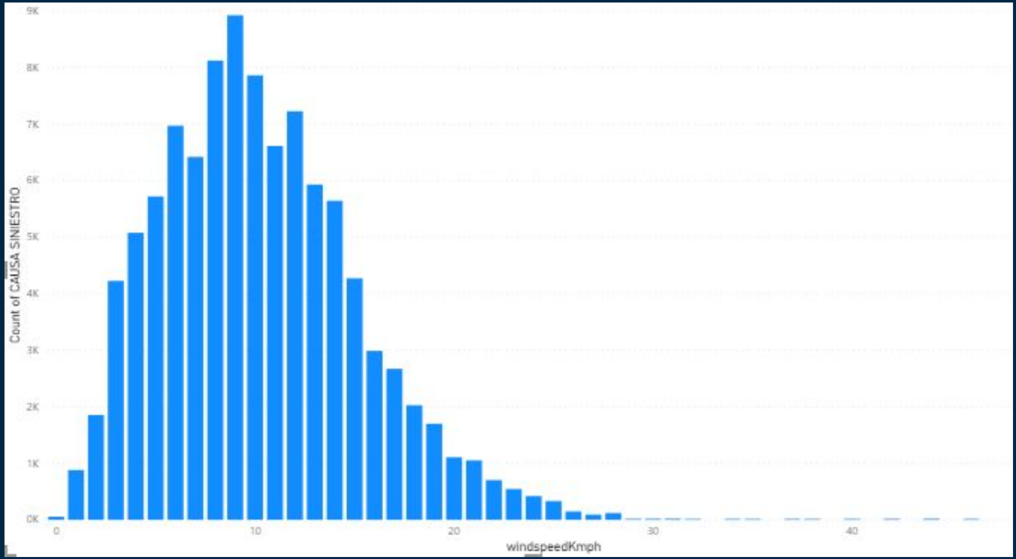


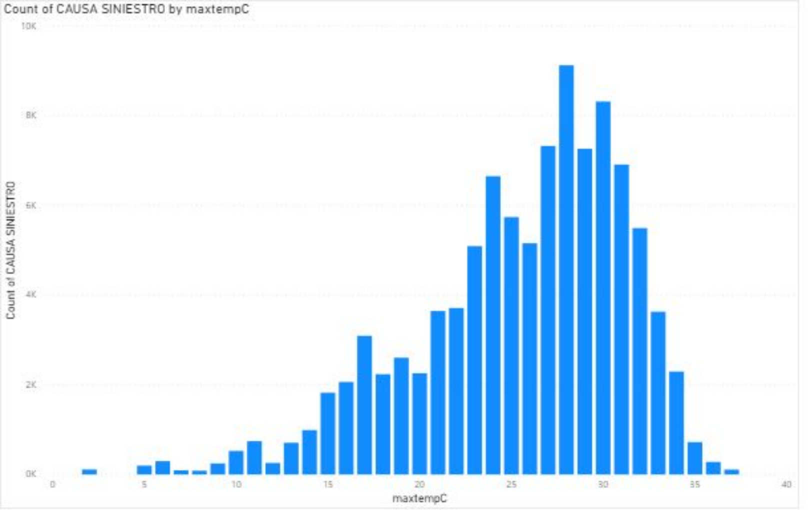
Paredón
Icamole
Ramos Arizpe
Saltillo
Arteaga
Cerro San Rafael
San Antonio
Allende
Bing



3.25K (3.26%)
3.57K (3.58%)
5K (5.01%)
5.26K (5.26%)
6.26K (6.27%)
7.61K (7.62%)
13.18K (13.2%)
22.92K (22.95%)
31.39K (31.43%)
PUNTO DE IMPACTO
● Frontal
● Trasero
● Cofre
● Costado der delantero
● Costado izq delantero
● Costado der trasero
● Costado izq trasero
● Costado der central
● Costado izq central
● Parte baja
● Toldo
● Tapa cajuela

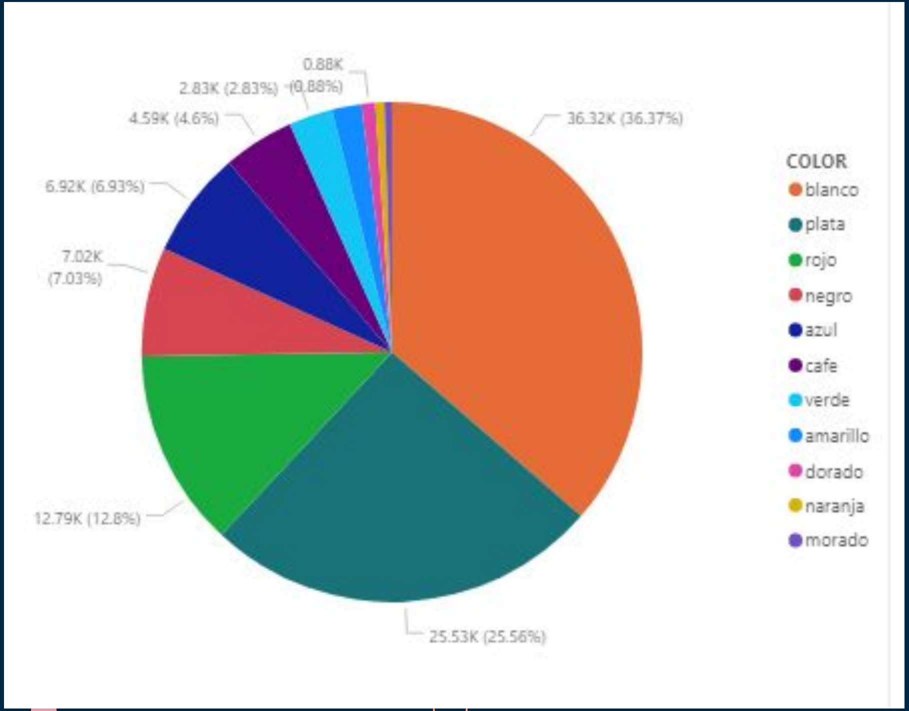
EDA

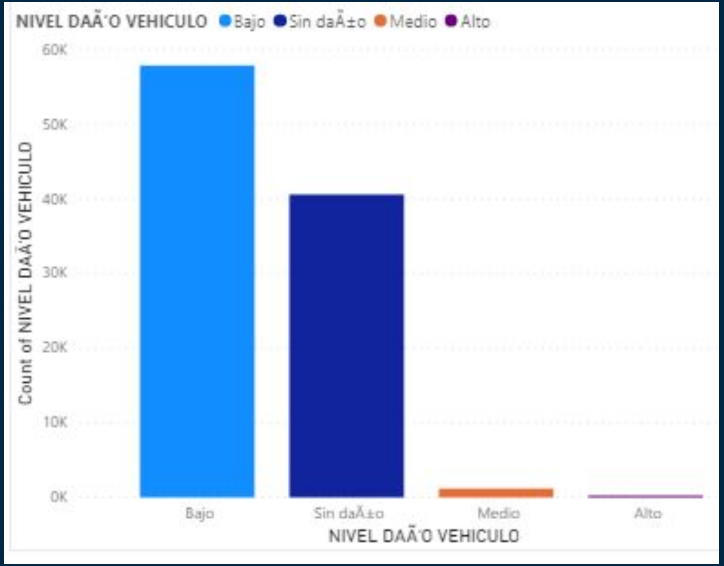




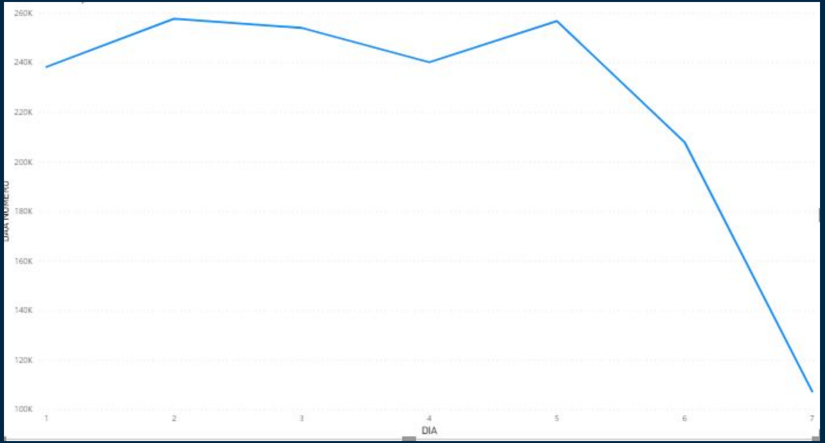
Mean max temp = 25.219°C
SD max temp = 5.745°C
Mean winds peed = 10.258 kmph
SD wind speed =5.029 kmph

EDA

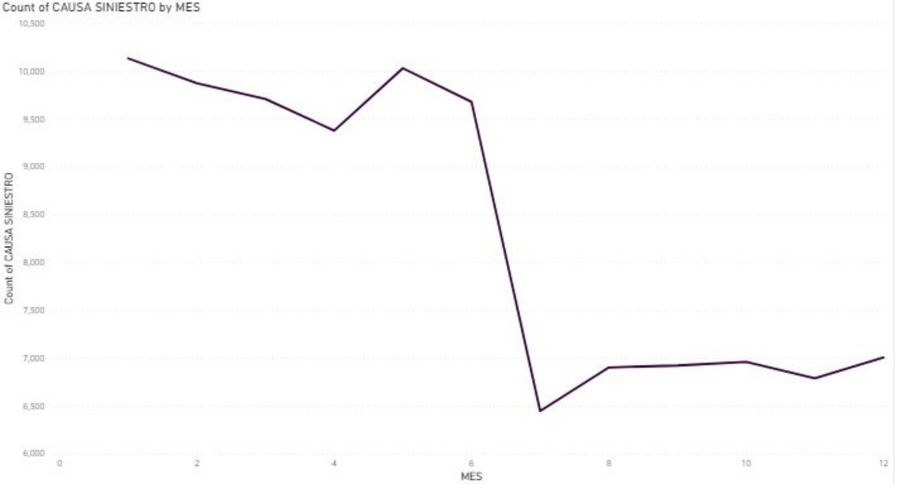

0.88K
(0.88%)
2.83K (2.83%)
4.59K (4.6%)
36.32K (36.37%)
6.92K (6.93%)
7.02K
(7.03%)
12.79K (12.8%)
25.53K (25.56%)
COLOR
● blanco
● plata
● rojo
● negro
● azul
● cafe
● verde
● amarillo
● dorado
● naranja
● morado


NIVEL DAÃ'O VEHICULO
● Bajo
● Sin daÃ±o
● Medio
● Alto
60K
50K
40K
30K
20K
10K
0K
Count of NIVEL DAÃ'O VEHICULO
Bajo
Sin daÃ±o
Medio
Alto
NIVEL DAÃ'O VEHICULO

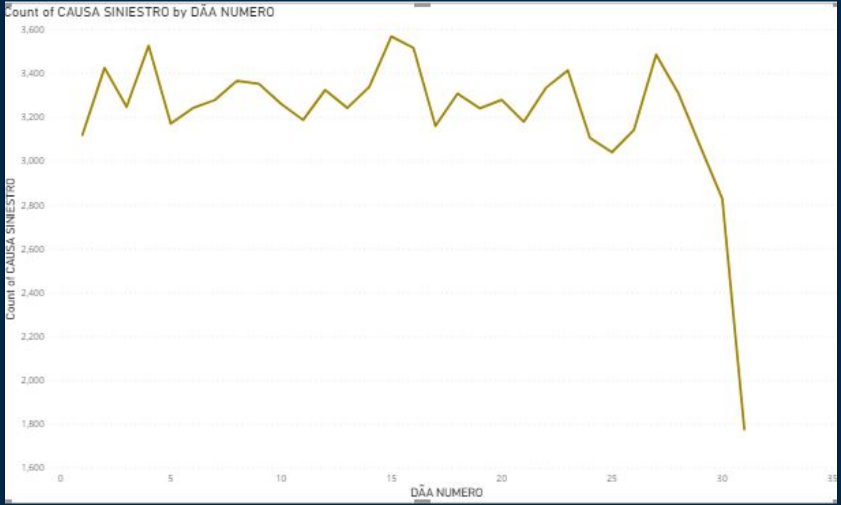
EDA



260K
240K
220K
200K
180K
160K
140K
120K
100K
DÃA NUMERO
1
2
3
4
5
6
7
DIA



Count of CAUSA SINIESTRO by MES
10,500
10,000
9,500
9,000
8,500
8,000
7,500
7,000
6,500
6,000
Count of CAUSA SINIESTRO
0
2
4
6
8
10
12
MES



Count of CAUSA SINIESTRO by DÃA NUMERO
3,600
3,400
3,200
3,000
2,800
2,600
2,400
2,200
2,000
1,800
1,600
Count of CAUSA SINIESTRO
0
5
10
15
20
25
30
35
DÃA NUMERO

Machine Learning - Clustering

Clustering was made with kMeans and kModes (a variant for categorical datasets).

It considered:

- Location of accident
- Type of accident
- Car data
- Climate data
- Visibility
- If it was a holiday
- Type of road where it happened
- If it happened near a pedestrian crossing
- If it happened at night

Machine Learning - Classification

Validity Index using supervised Classifiers.

k-Means clustering was validated with 0.999 AUC with Random Forest classification.

The importance of the attributes in the construction of the random forest were as follows:

NIVEL DAÑO VEHICULO_2	Bajo
NIVEL DAÑO VEHICULO_1	Sin daño
TIPO VEHICULO_1	Camión
TIPO VEHICULO_2	Auto
MODEL_YEAR_5	older
NIVEL DAÑO VEHICULO_3	Medio
PUNTO DE IMPACTO_2	Frontal

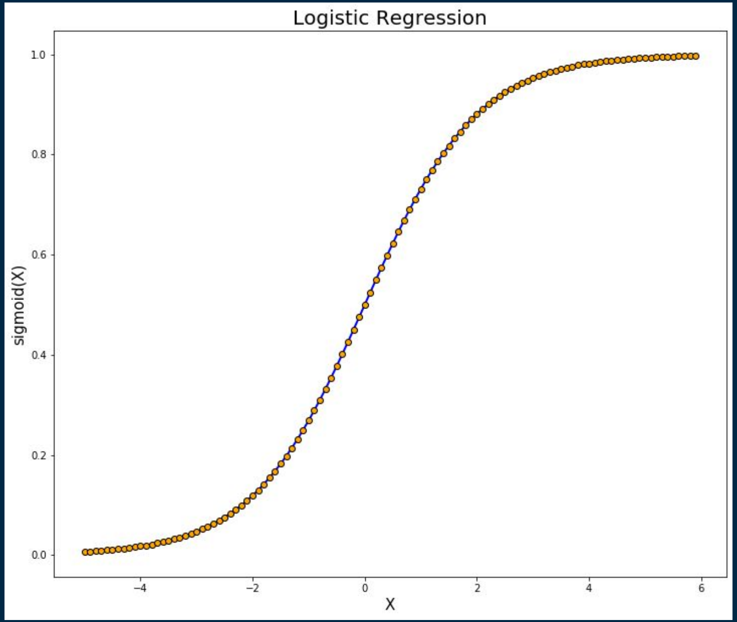
Machine Learning - Classification

What does this mean?
The best way to segment the dataset is by accident severity.
Cars are more likely to get damaged than trucks.
Older cars are more likely to get damaged than newer ones.

This information could be useful for car insurance companies in determining how likely it is for the car to be damaged considering its make, it’s age, and where it mostly drives.

Machine Learning - Logistic Regression

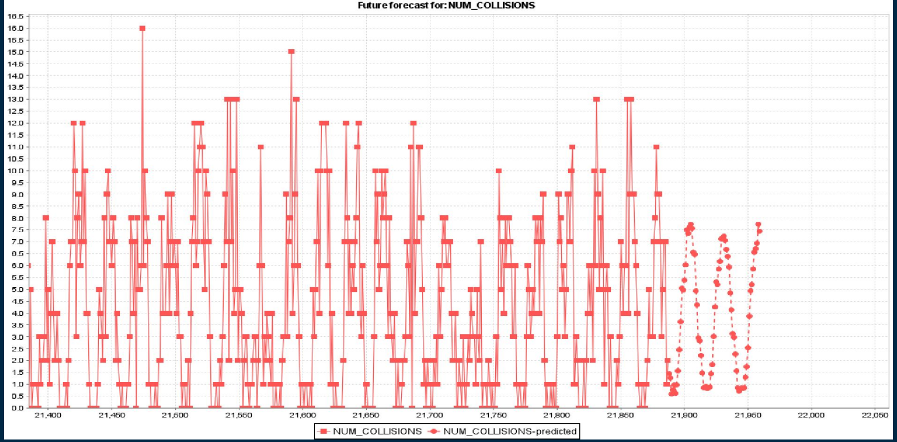
- Logistic regression not entirely suitable for this task
  - Relationship between variables not linear.
- Recursive Feature Elimination employed, but low accuracy achieved
- Adjusted R-Squared > 0.3



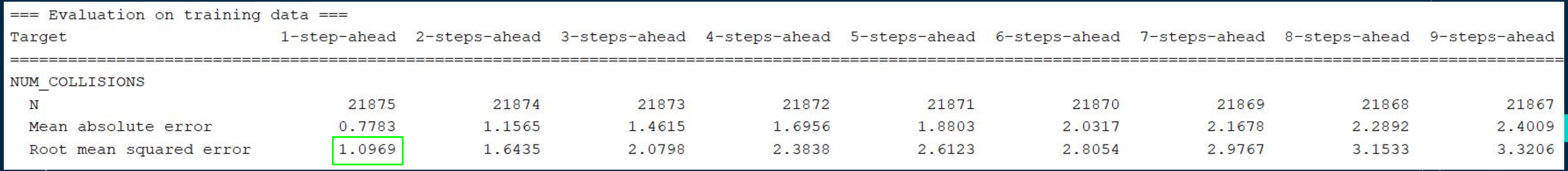
Machine Learning - Scikitlearn Random Forest

- Given the large amount of variables, Random Forest was suitable for the task
- 95.43% percent of accuracy
- Misprediction of ~5 accidents per day
- Geohash zones of 39.1km×19.5km (4 hash characters)

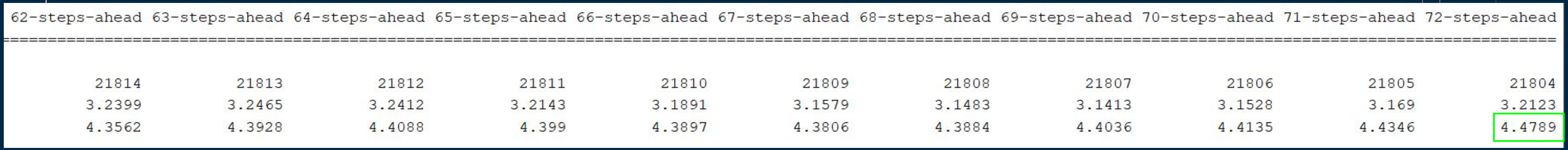
Machine Learning – Time-Series Analysis



Random Forest with 72 hours ahead.



```
=== Evaluation on training data ===
Target                    1-step-ahead  2-steps-ahead  3-steps-ahead  4-steps-ahead  5-steps-ahead  6-steps-ahead  7-steps-ahead  8-steps-ahead  9-steps-ahead
=================================================================================================================================================================
NUM_COLLISIONS
  N                              21875          21874          21873          21872          21871          21870          21869          21868          21867
  Mean absolute error           0.7783         1.1565         1.4615         1.6956         1.8803         2.0317         2.1678         2.2892         2.4009
  Root mean squared error       1.0969         1.6435         2.0798         2.3838         2.6123         2.8054         2.9767         3.1533         3.3206
```



```
62-steps-ahead 63-steps-ahead 64-steps-ahead 65-steps-ahead 66-steps-ahead 67-steps-ahead 68-steps-ahead 69-steps-ahead 70-steps-ahead 71-steps-ahead 72-steps-ahead
=================================================================================================================================================================

         21814          21813          21812          21811          21810          21809          21808          21807          21806          21805          21804
        3.2399         3.2465         3.2412         3.2143         3.1891         3.1579         3.1483         3.1413         3.1528          3.169         3.2123
        4.3562         4.3928         4.4088          4.399         4.3897         4.3806         4.3884         4.4036         4.4135         4.4346         4.4789
```

Future Work

- Evaluate RNN for time-series analysis.
- Integrate additional attributes regarding the accident location (the count of stop signals and traffic lights, whether it is a parking lot or not ,etc...)
- Train Random Forest with smaller geohash zones, more specific predictions.