Predicting Car Accidents in Buenos Aires City, Argentina

1. INTRODUCTION

- Argentina
  - Area of 2,780,400 km²
  - Population of approx. 45,000,000 inhabitants
- Large quantity of people moving to and out of Buenos Aires City
- The plan is to analyze data sets of car accidents, of car accidents

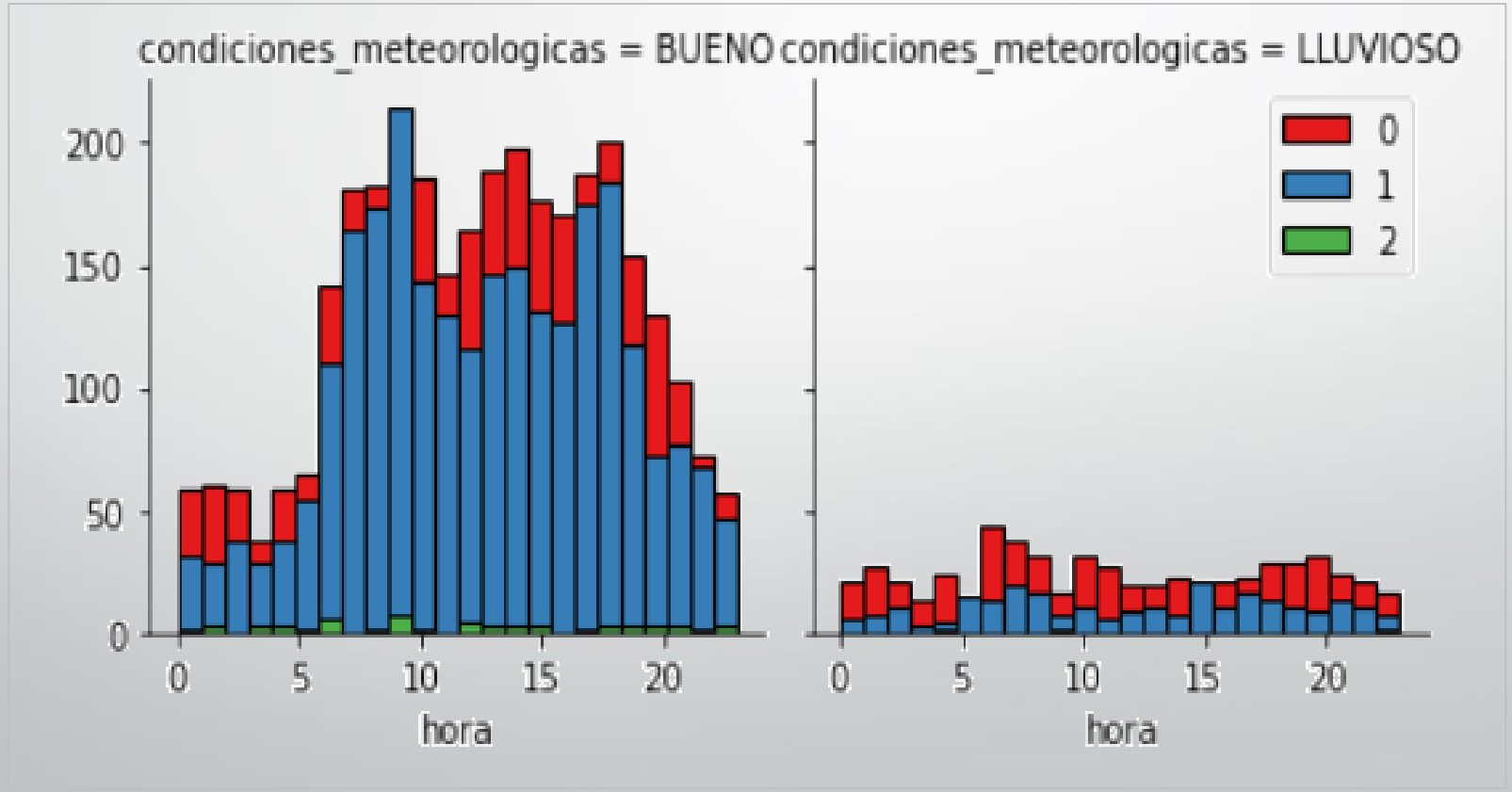
DATA

- An official Argentine site of data sets
- Intervenciones de Seguridad Vial.csv
- Whenever an accident requires attention from Road Safety, it is recorded in the data set

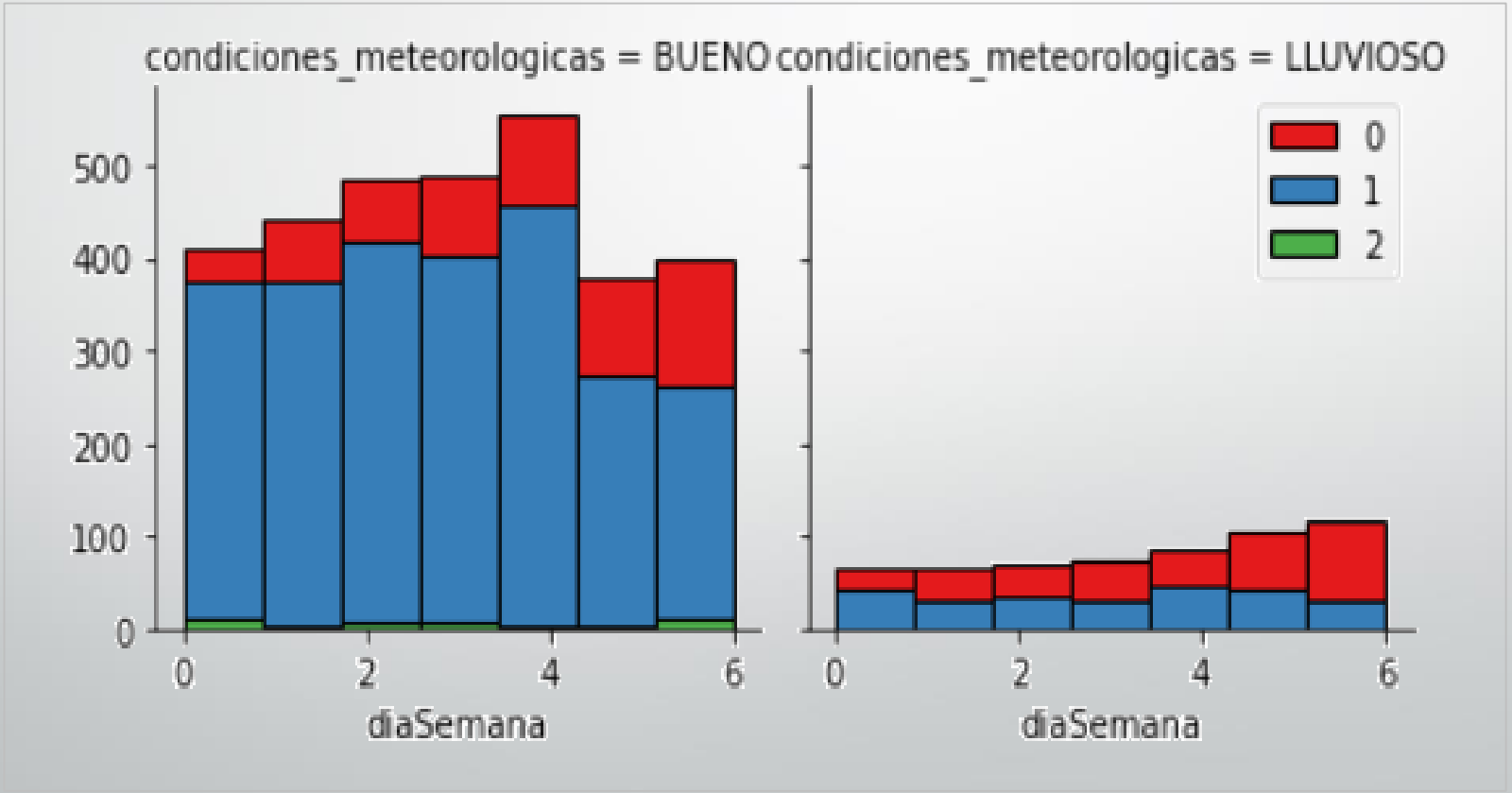
METHODOLOGY

- 6,663 rows x 15 columns
- The most likely hours for a car accident to occur are 9 am, 2 pm and 6 pm
- The most likely day of the week for a car accident to occur is Friday
- The most likely months for a car accident to occur are December and July
- Variable “severidad” (severity) was created

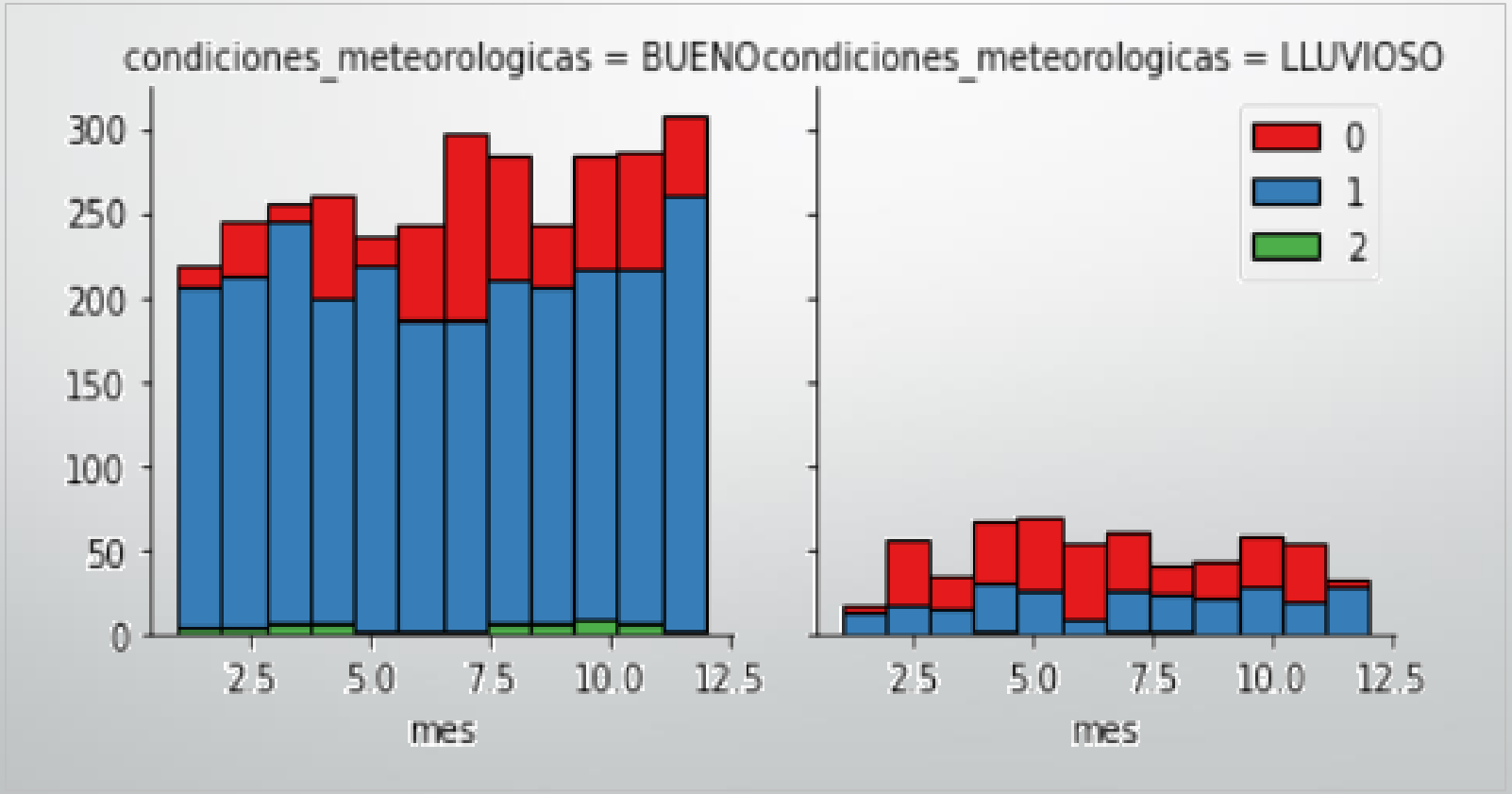
Relation between severity, weather and hour



Relation between severity, weather and day of week



Relation between severity, weather and month



MODELING

- Four classification models were used: K-Nearest Neighbor, Decision Tree, Support Vector Machine and Logistic Regression.
- K-Nearest Neighbor: best value is K = 20
- Decision Tree: criterion was set to “entropy”
- Support Vector Machine: kernel was set to “rbf”
- Logistic Regression: solver was set to “liblinear” and C = 0.01

RESULTS (ACCURACY)

	K-Nearest Neighbor	Decision Tree	Support Vector Machine	Logistic Regression	K-Nearest Neighbor	Decision Tree	Support Vector Machine	Logistic Regression
Data Set	Train				Test			
Jaccard Index	61.04%	75.76%	56.50%	55.78%	57.82%	54.27%	58.50%	58.28%
F1-score	58.45%	75.05%	45.53%	44.49%	54.12%	53.33%	47.94%	48.30%
Log loss	-	-	-	72.22%	-	-	-	75.50%

CONCLUSIONS

- The accuracy obtain was poor (below 60%)
- The models are not a good estimator for severity
- To improve accuracy more data is needed
- Quantity of data set available in Argentina is not too much