# Seattle's collisions and severity recognition using machine learning algorithms

JUAN SEBASTIAN ALVARADO GALEANO

**ACM Reference Format:**
Juan Sebastian Alvarado Galeano. 2018. Seattle's collisions and severity recognition using machine learning algorithms. *ACM Trans. Graph.* 37, 4, Article 111 (August 2018), 2 pages. https://doi.org/10.1145/1122445.1122456

## 1 INTRODUCTION

Car collisions and its influence on pedestrians in a problem which has always focused the attention of goverments in an continuous effort to preserve life. In general, collision might occur for several reasons which include distractions, influence of alcohol, the road and light conditions, the date among others which can result in various types of injuries. In order to classify all incidents, Seattle has been using a coding scheme which classifies these events as a fatality (3), a serious injury (2b), an injury (2), a prop damage (1) or even unknown (0). The latter coding scheme help to find patterns and important factors to take into account for future prevention of accidents implemented in upcoming laws as well as a way of testing actual regulations.

The present work shows a Machine Learning approach to collision severity clasification via a supervised algorithm as a function of the relevant involved factors such as fatalities, property damage, injuries etc. to identify relevant collision factors and optimize the statistical data recognition to give statistical results more efficiently.

Author's address: Juan Sebastian Alvarado Galeano, jsalvaradog@outlook.com.


0730-0301/2018/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

## 2 DATA

A Seattle collisions data base recording all incidents from 2003 to present was obtained from Kaggle, which provides a suitable set for training the model prior to predict the severity of new incidents. The data set includes fatalities, collision type, number of involved persons their injuries and seriousness, the number of vehicles (and their speed), pedestrians, bicycles and parked cars involved, possible causes such as inattention, alcohol influence, whether or not the pedestrian right of way was not granted and environmental factors such as weather, road and light conditions and date among others. These features have been selected since I consider they are the main factors that helps to understand collisions and their severity, mostly influenced by the human affectation. However, it is worth to mention that the used data base includes more information which is not considered in this work.

First, the date information was splitted into three columns in order to verify the incidents per year, month and day. Then, data in come columns such as "UNDERINFL" has to be formatted in order to have the same convention for *yes*(Y) and *no*(N). Besides, numerical attributes have been re-scaled in such a way that represents the percentage over the total cases. In some graphs the sum of all values might not equal one, this happens because the difference is related to the unreported data.

Numerical values ensure the success of the machine learning algorithm since at first sight one

can see that a decision tree is suitable from the following table.

| SEVERITYCODE | PERSONCOUNT | PEDCOUNT | PEDCYLCOUNT | VEHCOUNT | INJURIES | SERIOUSINJURIES | FATALITIES |
|---|---|---|---|---|---|---|---|
| 0 | 0.008366 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.0 |
| 1 | 0.649969 | 0.086023 | 0.113542 | 0.699043 | 0.000000 | 0.000000 | 0.0 |
| 2 | 0.323344 | 0.780138 | 0.809879 | 0.286929 | 0.943309 | 0.000000 | 0.0 |
| 2b | 0.016202 | 0.116042 | 0.071772 | 0.012712 | 0.053206 | 0.969263 | 0.0 |
| 3 | 0.002115 | 0.017798 | 0.004641 | 0.001313 | 0.003486 | 0.030737 | 1.0 |

Nevertheless, it is important to find out if the algorithms succeeds with the environmental data, in this way predictions and decisions can be made before accidents. Since we have non-numerical attributes I take advantage of the *get_dummies*. KNN, decision tree, SVM and Logistic regression are going to be used for testing the 20% of the data.

To sum up, the considered features are:

- Numerical: SEVERITYCODE, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, INJURIES, SERIOUSINJURIES, FATALITIES, INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT, SPEEDING, HITPARKEDCAR, YEAR, MONTH, DAY
- Categorical: SEVERITYCODE,ADDRTYPE, COLLISIONTYPE, WEATHER, ROADCOND, LIGHTCOND

for more information about their meaning the metadata is available in https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf