# Capstone Project – Predicting accidents with victims on Brazilian federal roads

Udacity - MACHINE LEARNING ENGINEER NANODEGREE

Carlos Eduardo de Souza
São Paulo, Brazil
https://github.com/SouzaCadu
https://www.linkedin.com/in/carlos-eduardo-de-souza/?locale=en_US

June 17, 2020

**ABSTRACT**

This project analyses accident data on Brazilian federal roads, collected by the Federal Highway Police, from 2017 to 2020 to develop a machine learning able to predict multi-class accident type. This dataset contains accidents recorded along the more than 70,000 km of monitored highways, every day of the year; also, are recorded the weather conditions, runway conditions, type of accident, number of people involved and victims, among other information. Will be performed data cleaning, treatment of categorical variables, exploratory analysis and selection of the information that will go through the model to predict if given these characteristics an accident result in “no victims”, “injured victims” or “fatal victims”. A regression model will be defined to estimate the probability of accidents occurring to victims. This information can assist in planning the distribution of checkpoints, improving road actions or triggering an alert system so that drivers drive more carefully.

**Keywords:** Exploratory Data Analysis · Supervised Learning

## 1 Problem Statement

In Brazil, the large number of deaths on federal highways is alarming, in the period 2010 - 2017 there were 62,120 deaths, with an average of 21 per day, almost one victim per hour; in addition to 201,006 seriously injured and 578,954 lightly injured as a result of 2,392,205 accidents.

The goal is from the analysis of the three main relevant factors in the study of road safety on federal highways: the element of infrastructure and environment (conditions of the roads - geometry, conservation, signalling and number of lanes; and weather and time conditions); human factor (physical, emotional states and compliance with traffic regulations); and the vehicle factor (maintenance conditions and safety equipment), to classified accidents with injured victims, dead victims or no victims.

This project is a classification problem in which the model will use as input variables the information recorded at the time of the accident, such as the direction of the road, the layout of the way, climatic condition, number of vehicles involved, type of accident, place where the

accident happened, and it will produce as an expected output the classification of the kind of accident, without victims, with injured victims, with dead victims.

## 2 Datasets and Inputs

There are four data files associated with this project:

- datatran2017.csv – Traffic accident data on Brazilian national roads in 2017 arranged by occurrences; 89.563 (accidents) x 30 features
- datatran2018.csv – Traffic accident data on Brazilian national roads in 2018 arranged by occurrences; 69.295 (accidents) x 30 features
- datatran2019.csv – Traffic accident data on Brazilian national roads in 2019 arranged by incidents; 67.446 (accidents) x 30 features
- datatran2020.csv – Traffic accident data on Brazilian national roads in 2020 arranged by occurrences; 19.573 (accidents) x 30 features

Bellow has presented a brief description of these features:

| Variable | Description |
|---|---|
| **id** | Variable with numerical values, representing the accident identifier. |
| **data_inversa** | Date of occurrence in dd / mm / yyyy format. |
| **dia_semana** | Day of the week of the occurrence. |
| **horario** | Time of occurrence in hh: mm: ss format. |
| **uf** | Federation Unit. |
| **br** | Variable with numerical values, representing the BR identifier of the accident. |
| **km** | Identification of the kilometre where the accident occurred, with a minimum value of 0.1 km and with a decimal point separated by a point. |
| **município** | Name of the municipality where the accident occurred. |
| **causa_acidente** | Identification of the main cause of the accident. In this data set, accidents with the main cause variable equal “No” are excluded. |
| **tipo_acidente** | Identification of the type of accident. E.g., frontal collision, lane departure, etc. In this data set, types of accidents with an order greater than or equal to two are excluded. The order of the accident shows the chronological sequence of the types present in the same occurrence. |
| **classificação_acidente** | Classification according to the severity of the accident: Without Victims, With Injured Victims, With Fatal Victims and Ignored. |
| **fase_dia** | The phase of the day at the time of the accident. |

| | |
|---|---|
| **sentido_via** | The direction of the road considering the collision point: Ascending and decreasing |
| **condição_meteorologica** | Meteorological condition at the time of the accident. |
| **tipo_pista** | Track type considering the number of tracks |
| **tracado_via** | Description of the route layout. |
| **uso_solo** | Description About the characteristics of the accident site: Urban = Yes; Rural = No. |
| **latitude** | Latitude of the accident site in decimal geodetic format. |
| **longitude** | Longitude of the accident site in decimal geodetic format. |
| **pessoas** | Total people involved in the incident. |
| **mortos** | Total dead people involved in the occurrence. |
| **feridos_leves** | The total number of people with minor injuries involved in the incident. |
| **feridos_graves** | Total of people with serious injuries involved in the occurrence. |
| **feridos** | The total number of injured persons involved in the incident (this is the sum of the light and the serious injured). |
| **ilesos** | Total unscathed people involved in the incident. |
| **ignorados** | Total of people involved in the occurrence and the physical state was not known. |
| **veiculos** | Total vehicles involved in the occurrence. |

Additionally, is show a brief descriptive report of the variables:

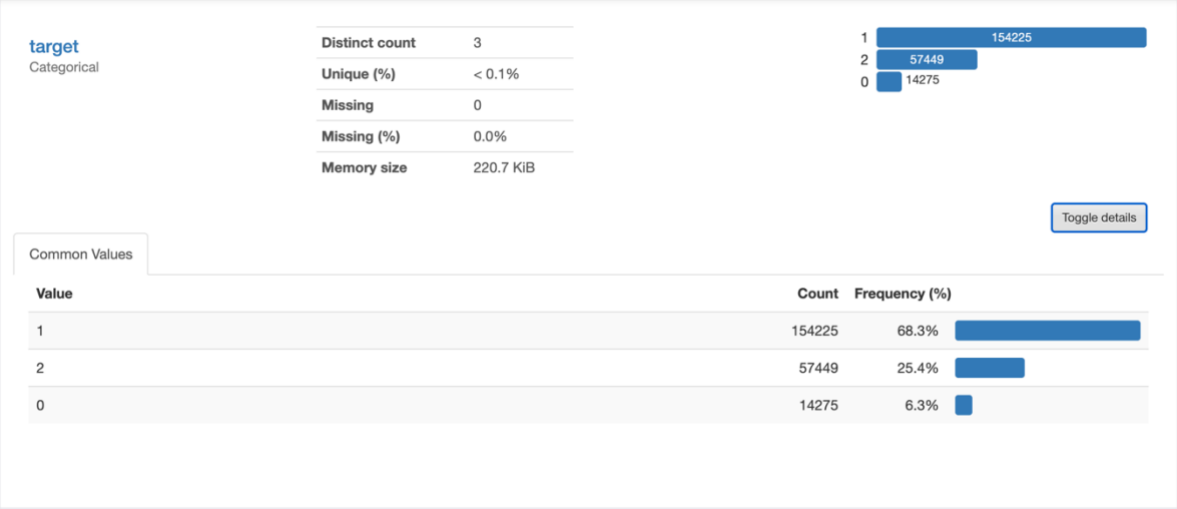


**Figure 1 - Number of cases per class (0: With dead victims, 1: With injured victims, 2: No victims)**

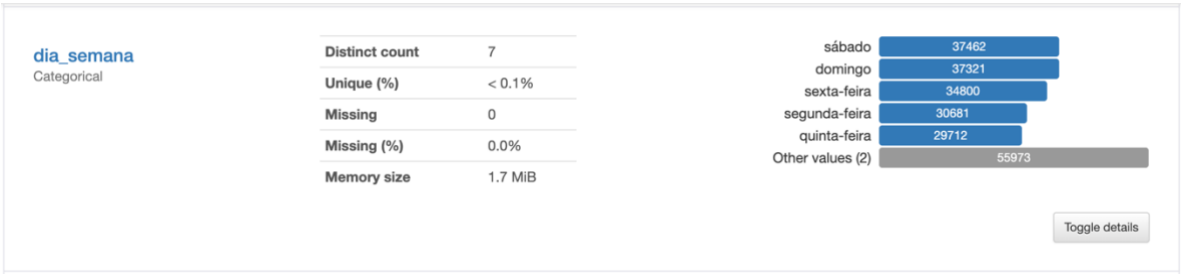


Figure 2 - Number of accidents per day of the week

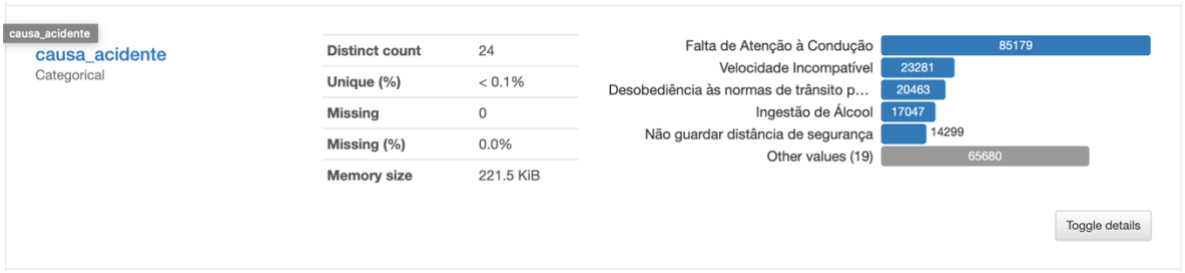


Figure 3 - Cause of the accident

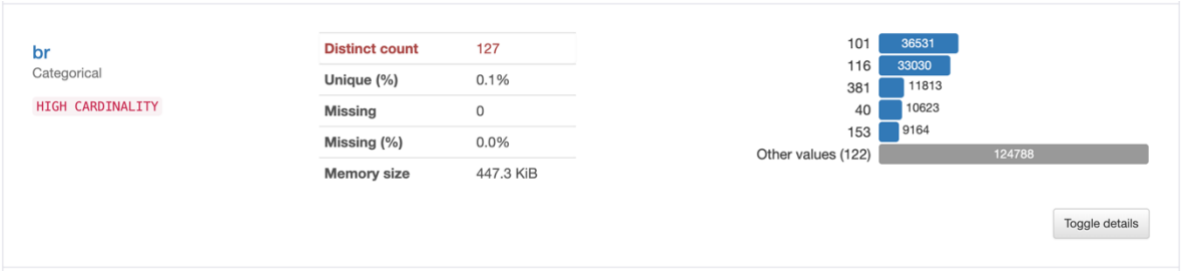


Figure 4 - Accidents on each road

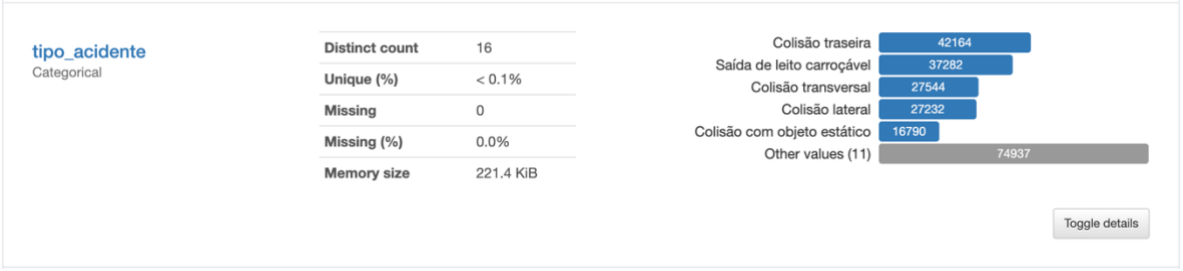


Figure 5 - Type of accident

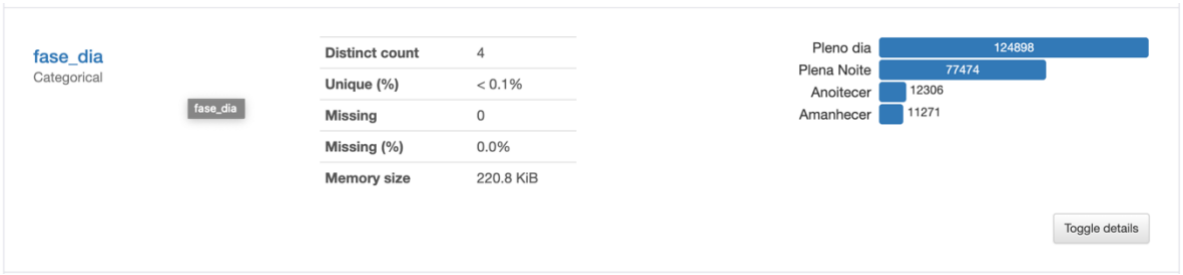


Figure 6 - Day period

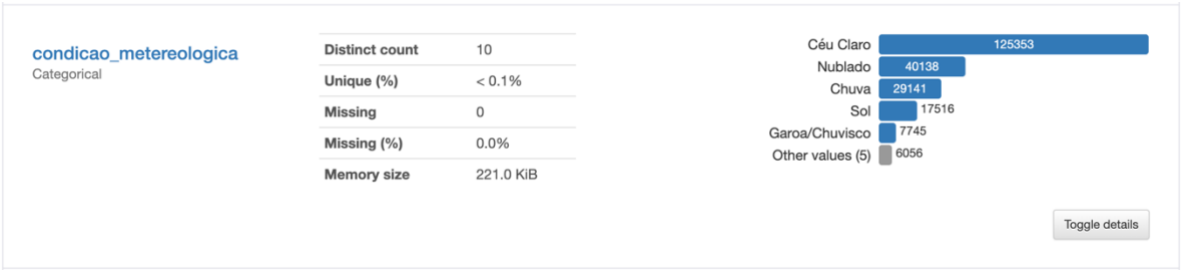


Figure 7 - A weather condition

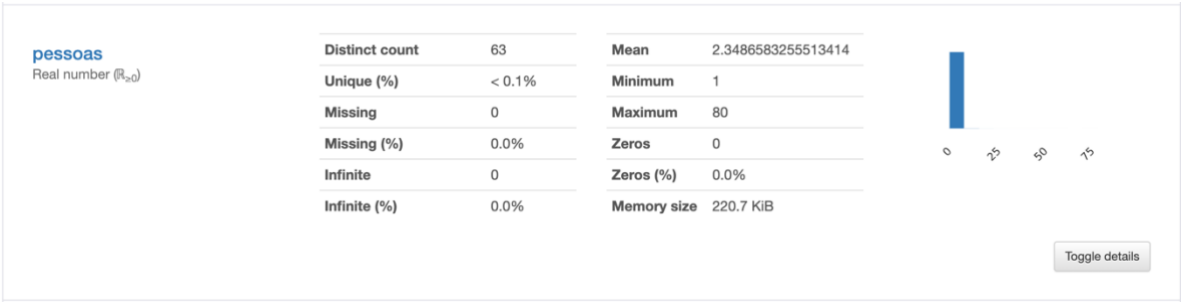


Figure 8 - People involved

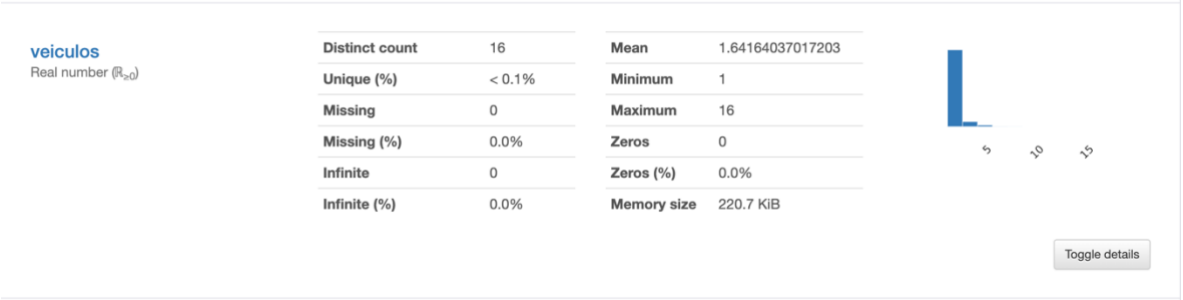


Figure 9 - Vehicles involved

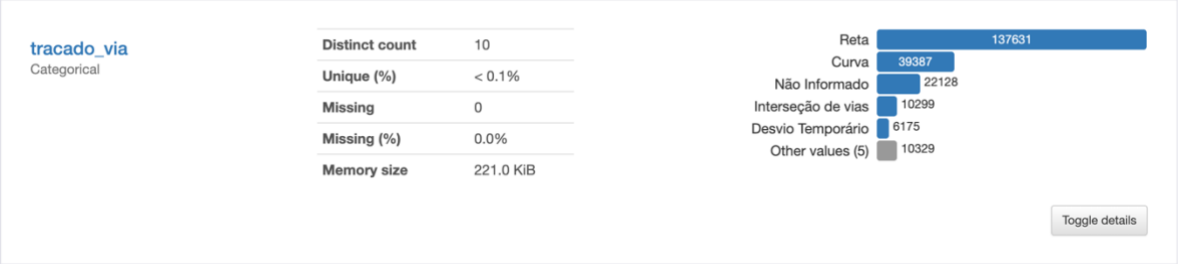


**Figure 10 - Track layout**

In the first analysis, we are noting the model's output un balancement between classes and some categorical variables have high cardinality; the transformation process of these variables can lead to the creation of sparse vectors, which can impact the performance of models sensitive to the dimensionality of the data.

## 3 Solution Statement

The analysed information from the years 2017 to 2019 will serve as a basis for training and testing the model. The database for 2020 will help for final validation of the model.

## 4 Benchmark Model

It will be used a simple logistic regression as a benchmark model. This model will serve as a basis, from which other classification models should be tested together with the adjustment of hyperparameters.

Two other works related to road accidents will be considered. Although these studies used different methodologies and techniques, they can provide a basis for comparison for the final model.

The first is the study by Chong, Abraham & Paprzycki, 2005. In their study, these researchers used a neural network for trying to classify victim's degree of injuries on traffic accidents. They had five classes of injures, including no harm, possible injury, non-incapacitating injury, disabling damage, and fatal injury. The features were based on both on-road and driver's characteristics. The road characteristics were similar to what we found in the datasets of National Road Police. The authors found a ~60% accuracy in predicting the type of victims.

The other study, which also cites the first work as a reference, is Oudemans, 2016. Oudemans proposed the use of machine learning techniques to predict the severity of traffic accidents in England, in this study, the proposed evaluation metric was AUROC, and the best model presented reached a value of 0.723.

## 5 Evaluation Metrics

For this project, the metric used to measure the quality of the model will be Precision. The reason for choosing this metric is to minimize the type I error, which occurs when the model predicts an event as accurate while the observation is false.

This type of error is captured by Precision, given its formulation: True Positives / Total Positives, in other words, Precision talks about how precise/accurate the model is out of those predicted positive, how many of them are positive.

For this problem, it is particularly important to correctly predict real events, accidents than with victims, as the costs associated with false positives are quite high.

One cannot calculate what the loss of human life represents or the psychological damage and traumatic stresses to which traffic victims and their families are subjected after events of this nature. On the other hand, there is also the formation of economic and financial costs that directly impact families, as well as society in general, and that can be estimated using specific calculation methodologies.

The details of this methodology can be found in the document ***Impactos sociais e econômicos dos acidentes de trânsito nas rodovias brasileiras*** (Social and economic impacts of traffic accidents on Brazilian highways), prepared by competent bodies of the Brazilian government to assist in the elaboration of public policies on the subject.

# 6 Project Design

## 6.1 Preprocessing

### 6.1.1 Assess Missing Data

In this step the demographics data is assessed in terms of missing data at the column level, row-level and data level (some values in columns could also represent unknown data, and we need to re-encode them to NumPy NaN).

### 6.1.2 Select and Re-Encode features

Checking for missing data isn't the only way in which one can prepare a dataset for analysis. Since the unsupervised learning techniques that we will use, will only work on data that is encoded numerically, we need to make a few encoding changes or additional assumptions to be able to make progress. Even though almost all of the values in the dataset are encoded using numbers, not all of them represent numeric values:

- For numeric and interval data, these features can be kept without changes.

- Most of the variables in the dataset are ordinal. While ordinal values may technically be non-linear in spacing, we can make the simplifying assumption that the ordinal variables can be treated as being an interval (that is, kept without any changes).

- Special handling may be necessary for the remaining two variable types: categorical, and' mixed.'

## 6.2 Oversampling

Adjust the number of cases between classes to correct data imbalance, it's a technique that helps increase the model's ability to predict each of the categories correctly, contribute to reducing overfitting by expanding the model's generalizability.

6.3 Supervised Learning Model

In the modelling stage, in addition to separating the bases in training, testing and validation, any imbalance between classes that cannot be adjusted by some data augmentation technique must be passed as a parameter to the model so that it can adapt to the situation

After creating the model, final validation will be made on the 2020 data, and the last error of the model will be calculated.

References:

OUDEMANS, Eva Elisabeth – PREDICTING TRAFFIC ACCIDENT INJURY SEVERITY IN GREAT BRITAIN USING MACHINE LEARNING TECHNIQUES.

Vasconcellos, Paulo – Como saber se seu modelo de Machine Learning está funcionando mesmo
URL https://paulovasconcellos.com.br/como-saber-se-seu-modelo-de-machine-learning-est%C3%A1-funcionando-mesmo-a5892f6468b

Shung, Koo Ping – Accuracy, Precision, Recall or F1?
URL https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

Estimativa dos Custos dos Acidentes de Trânsito no Brasil com Base na Atualização Simplificada das Pesquisas Anteriores do Ipea
URL http://repositorio.ipea.gov.br/bitstream/11058/7456/1/RP_Estimativa_2015.pdf

Prediction of accident victims on federal roads in Brazil
URL https://github.com/leportella/federal-road-accidents#prediction-of-accident-victims-on-federal-roads-in-brazil

Impactos sociais e econômicos dos acidentes de trânsito nas rodovias brasileiras
URL http://infraestrutura.gov.br/images/Educacao/Publicacoes/custos_acidentes_transito.pdf

Polícia Rodoviária Federal
URL https://portal.prf.gov.br/dados-abertos-acidentes