Capstone Project - Car Accident Severity

Report

Ida Kusumawati

Introduction

Traffic accidents are bad events to experience. Frequent accidents cause many losses for many people. In addition to claiming lives, accidents also cause material losses because these events can result in traffic jams or damage to public infrastructure that have been the target of collisions so that national development is hampered. The Seattle government, through the Seattle Department of Transportation, has released data on accidents that have occurred in Seattle since 2004 and it is always updated every week. The uploading of the data shows how concerned the Seattle government is about this topic. Reporting from Colburn Law, it is known that almost every day, there are collisions in Seattle. In 2016, collisions severity occurred as follows: 58% properly damage only, 27% injury collisions, and 1% serious injury collisions. It is also known that the "common" type of collisions occurred were car collisions, with a total of 3,644 incidents in 2016. Of course this number is not a low number. Departing from this background, this project was created. Analysis and prediction of car accident severity from various variables in the dataset using machine learning.

Data

In order to fulfill my grade on this IBM Data Science capstone project, I used an example of data uploaded by the Seattle government through the Seattle Department of Transportation in .csv format. The dataset is available in both categorical and numerical data. The dataset has been downloaded via the Coursera course. In addition to documents in .csv format, there is also metadata. Metadata contains all the information related to that data. There is a code to categorize an accident severity. Based on what is written in the metadata, it can be seen that this dataset was recorded from 2004. The dataset contains variables related to collisions in Seattle. Some of the variables are weather, road conditions, location, type of collisions, and etc. The existing variables will be linked and analyzed to determine what factors affect the car accident severity. After analyzing it, a car accident severity prediction program using Machine Learning will be made.

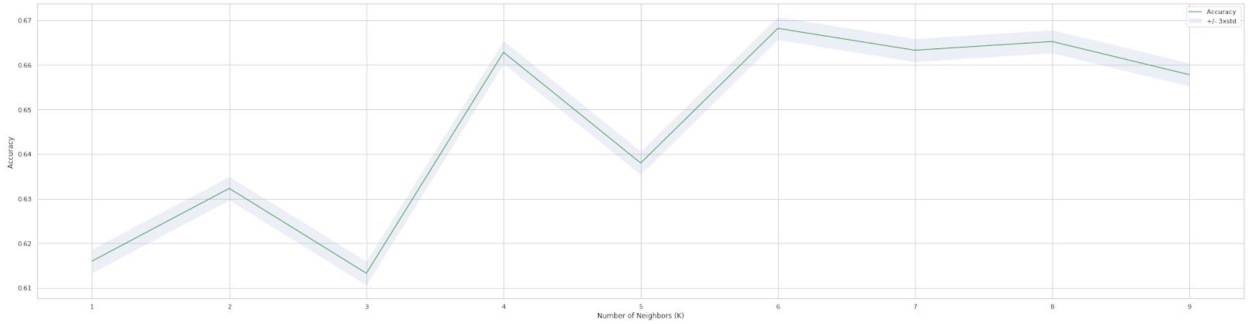
Methodology

Before modeling, csv data were plotted for each (several) variables with their severity for analysis. In this project, I chose 4 variables for modeling, including: Road Condition, Weather, Light Condition, and Junction Type using the K-Nearest Neighbors (KNN) algorithm. Therefore, after the data frame is defined, the value of each selected variable is converted from a string data type to numerical data. After the data was changed to numerical type, then modeling and preprocessing was done using sklearn. The existing data was carried out by train_test_split with test_size = 0.2 and random_state = 4. After that, the KNN algorithm was applied to obtain maximum accuracy from modeling with a certain k value used (in this project the k = 8 value is used).

Result

This dataset analyzed shows that accidents occur most frequently during the day with dry road conditions. This can happen because in both conditions it is a common time for people to have activities outside the home, such as school, work, and so on. Of course, it was in these two conditions that injury accidents occur most often. As for the junction type, accidents most often occurred in the mid-block (not related to intersection) and at intersection (related intersection).

It can be seen from the test_size written in the code, the data used for training is 80%, while for the test it is 20%. The modeling results obtained in this project indicate that the model can carry out car accident severity prediction (at k = 8) with a score of R^2 = 0.66 which shows a good enough number so that it can be said that this model can make predictions with the desired variables.



Discussion

From this project, the data shows the number of variables that can be linked to determine the factors of an accident. With the existence of data that has been published by the government and other authorities, it can make it easier for data scientists around the world to carry out tests so that problems like this can be more easily resolved. This project is not yet perfect and there are still many techniques in Machine Learning that can be used. Maybe in the next, regression techniques can be applied to make the car accident severity prediction program more real.

Conclusion

The conclusion from this project is that machine learning can be applied to real issues in the social environment, including in terms of predictions. KNN shows that this modeling has a fairly good level of accuracy.