Final report Capstone Project

Tom van Eijk

September 2020

1 Introduction

The Netherlands is a well developed country. However, there are also relatively many traffic accidents. Often vulnerable traffic attendants such as pedestrians and cyclists. It would be interesting for the government to know in which region many accidents occur.

1.1 Problem Description

This is a clear problem for the government, since citizens of these specific areas do often not realize the danger that traffic brings with it. So first, these regions should be located and then the government could invest in advertisement to warn these citizens.

1.2 Target Auditions

To solve this problem, a data science team led by myself has been engaged by the government. The political leaders expect us to locate these specific regions with decent machine learning models.

1.3 Success Criteria

The success criteria of this project will be a good recommendation of the regions with many traffic accidents based on 1 key factor; amount of deadly accidents. It should allow easy replication for future traffic accident data.

2 Data Description

As we need to explore, segment, and cluster the regions of the Netherlands. The coordinates of accidents and the kind of accidents are key in this project. I retrieved a data set of traffic accidents that happened in 2004. This data set contains information about:

- Coordinates of accidents
- Accidents IDs
- Kind of accidents
- Cause of accidents

Following data sources will be used to get the required information:

- centers of areas will be generated algorithmic-ally and approximate addresses of centers of those areas will be obtained using
- number of accidents and their type and location in every neighborhood will be obtained
- coordinates of regions will be obtained

The file was hosted by the Dutch government in 2004. The first 5 rows can noticed in Figure 1. In total, it contains 116 columns and 146813 rows.

	Longitude	Latitude	OngevalID	Communicatie_Ref	ProcesverbaalOpgem	Afloop3	AantalPartijen	Aard	GekoppledNiveau	Wegsituatie	...
0	3.364700	51.313763	120041872038	04-176360	NaN	Uitsluitend materiele schade	2	Vast voorwerp	Ongeval gekoppeld op straat niveau	Rechte weg	...
1	4.310351	51.527046	120041856048	04-177712	NaN	Letsel	1	Eenzijdig	Ongeval exact gekoppeld aan BN	Bocht	...
2	4.048437	51.499148	120041879850	04-186040	NaN	Uitsluitend materiele schade	2	Kop/staart	Ongeval exact gekoppeld aan BN	Rechte weg	...
3	3.364700	51.313763	120041871785	04-178227	NaN	Uitsluitend materiele schade	2	Vast voorwerp	Ongeval gekoppeld op straat niveau	Rechte weg	...
4	3.371984	51.313218	120041921852	04-209663	NaN	Uitsluitend materiele schade	2	Vast voorwerp	Ongeval exact gekoppeld aan BN	Rechte weg	...

Figure 1: Traffic accident raw data

2.1 Data Features

We will be leveraging on features in a reliable location information provider such as the government open source data platform to explore the various types of attributes. We will also need to understand the impact of these attributes

nearby on deadly accidents. The information obtained per region will be as such like below and has to be in a structured format so to allow for further computation:

1. Longitude
2. Latitude
3. GemeenteNaam
4. OngevalID
5. Afloop3
6. Aard
7. GekoppledNiveau
8. Wegsituatie

	Longitude	Latitude	GemeenteNaam	OngevalID	Afloop3	Aard	GekoppledNiveau	Wegsituatie
0	3.364700	51.313763	Sluis Z	120041872038	Uitsluitend materiele schade	Vast voorwerp	Ongeval gekoppeld op straat niveau	Rechte weg
1	4.310351	51.527046	Bergen op Zoom	120041856048	Letsel	Eenzijdig	Ongeval exact gekoppeld aan BN	Bocht
2	4.048437	51.499148	Reimerswaal	120041879850	Uitsluitend materiele schade	Kop/staart	Ongeval exact gekoppeld aan BN	Rechte weg
3	3.364700	51.313763	Sluis Z	120041871785	Uitsluitend materiele schade	Vast voorwerp	Ongeval gekoppeld op straat niveau	Rechte weg
4	3.371984	51.313218	Sluis Z	120041921852	Uitsluitend materiele schade	Vast voorwerp	Ongeval exact gekoppeld aan BN	Rechte weg

Figure 2: Separated features

3 Methodology

I have made a distinction between non-deadly accidents and deadly incidents with the help of dummies. Only deadly accidents are used for the data analysis. After the distinction, the amount of deadly accidents per region is calculated.

Furthermore, we need to know the coordinates and locations of this regions, and therefore a JSON file with coordinates of all regions has been used for achieving this objective. This is important so that we can input this information into the chloropeth to obtain correct information in these regions, and this is precisely what we have done for it in this project.

We will also use machine learnings techniques such as SVM to test our prediction model for future accidents. This is critical as we need to recommend to the government the regions with most deadly accidents. These regions need more advertisement about safety issues among traffic participants.

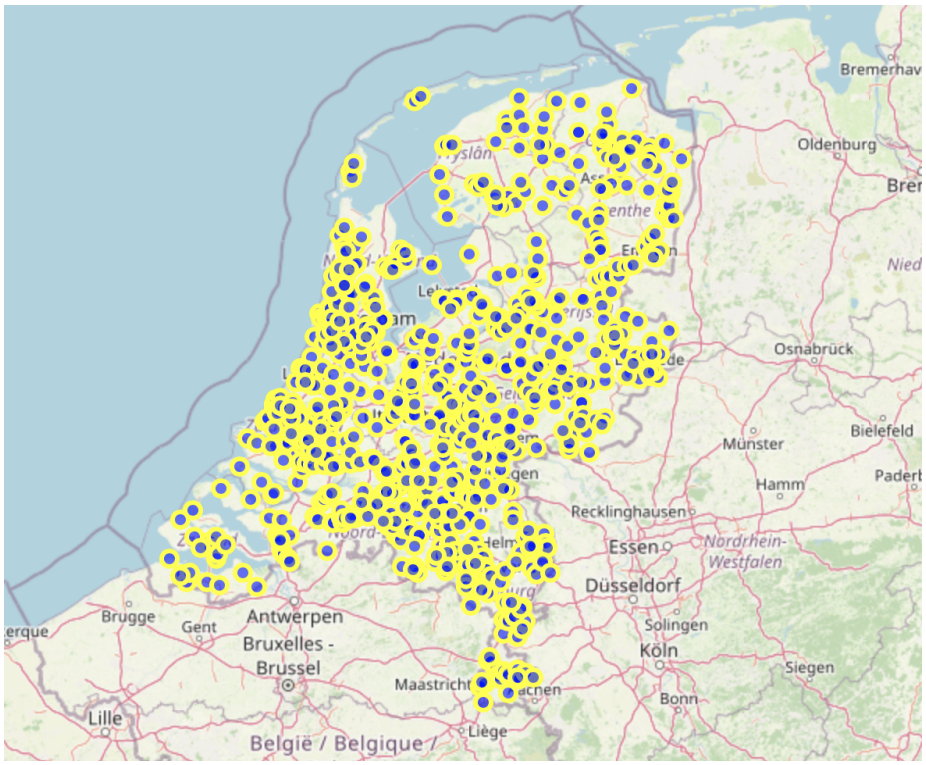

Figure 3: Locations of incidents

Finally, with all these methodologies, we will then be able to come up with a best recommendation to the government to their problem which is where the most deadly accidents happen. In other words, we will not want to recommend to the government to advertise in an area whereby there is already a low concentration of deadly accidents.

4 Results

With chloropeth, the highest concentrated deadly accident regions were identified. These regions are red together based on the similar nearby accidents in each of the regions. This information is critical so that we can target on the region where most deadly accidents occur, because the government is interested to invest in effective safety advertisement.

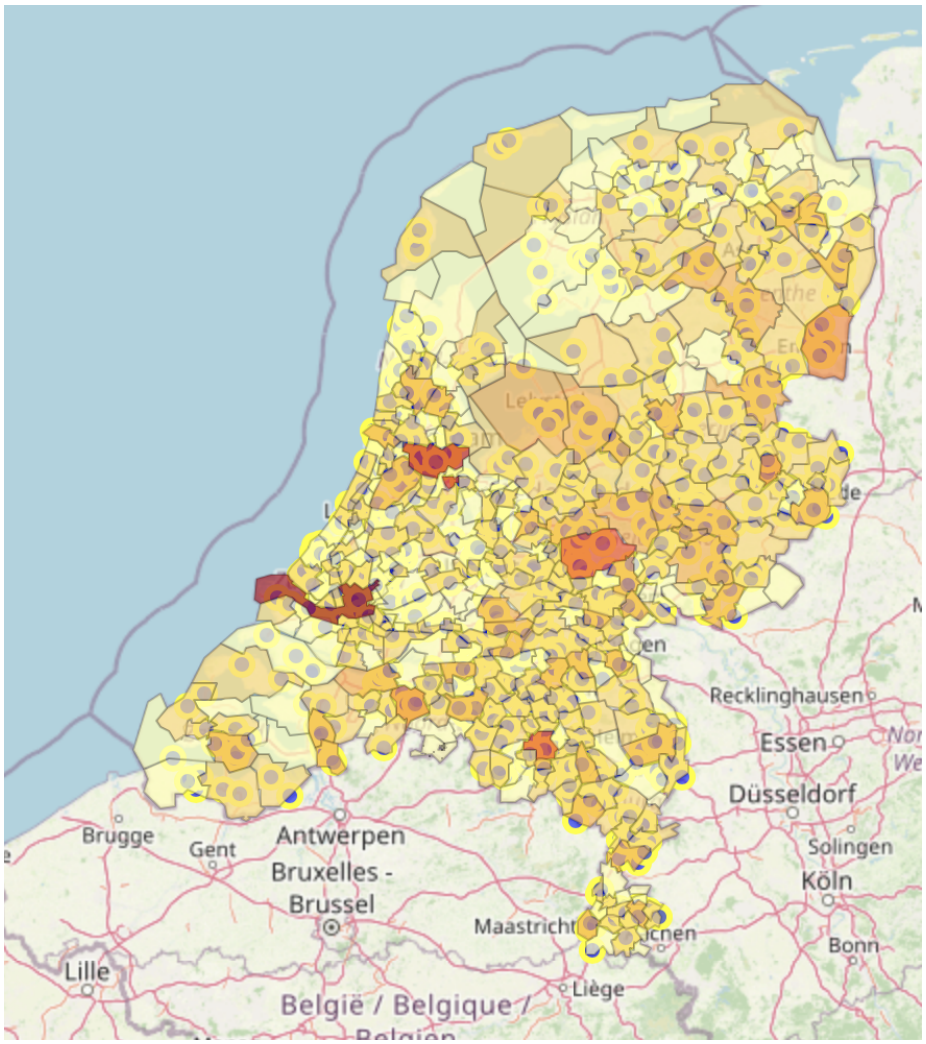

Figure 4: Most deadly accidents

With one-hot coding, all possible causes are separated. Logically, this can help with the prediction model for future deadly accidents.

	Longitude	Latitude	Dier	Eenzijdig	Flank	Frontaal	Geparkeerd voertuig	Kop/staart	Los voorwerp	Onbekend	Vast voorwerp	Voetganger
0	3.364700	51.313763	0	0	0	0	0	0	0	0	1	0
1	4.310351	51.527046	0	1	0	0	0	0	0	0	0	0
2	4.048437	51.499148	0	0	0	0	0	1	0	0	0	0
3	3.364700	51.313763	0	0	0	0	0	0	0	0	1	0
4	3.371984	51.313218	0	0	0	0	0	0	0	0	1	0

Figure 5: One-hot coding

When we get the data, after data cleaning, pre-processing and wrangling, the first step we do is to feed it to an outstanding model and of course, get output in probabilities. The SVM trained classifier predicts the result. This needs to be displayed and that is where the Confusion matrix comes into the limelight. Confusion Matrix is a performance measurement for machine learning classification.

```
                precision    recall  f1-score   support

     Dodelijk       0.00      0.00      0.00       162
Niet dodelijk       0.99      1.00      1.00     29201

     accuracy                           0.99     29363
    macro avg       0.50      0.50      0.50     29363
 weighted avg       0.99      0.99      0.99     29363

Confusion matrix, without normalization
[[    0   162]
 [    0 29201]]
```

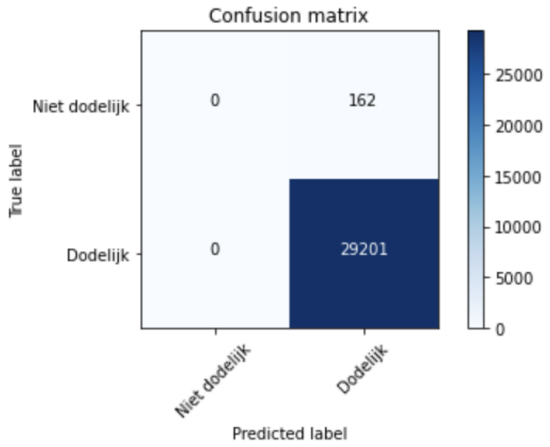

Figure 6: One-hot coding

5 Discussion

Based on the result above, the prediction model looks to offer a higher number of accuracy and allows the government to replicate this analysis in the future as part of their marketing plan.

Within the Rotterdam region, we will like to recommend a more advertisement to give the community a higher advantage and chance to avoid accidents. Hence, with this in mind, the Netherlands could be made much more safe for each traffic participant.

6 Conclusion

With that, we have concluded that the best recommendation for the government is to first offer their services in Rotterdam with the key factors to consider such as most deadly accidents. These regions are most dangerous for traffic participants:

1. Rotterdam
2. Amsterdam
3. Eindhoven

It is also recommended to the government to re-run this data science program to get the updated result and use the result into consideration as part of the marketing plan in selecting the next region to offer their services. This is critical not only to make sure that they got the updated result for better decision making, but also to make sure that they can re-validate the findings from this project. Finally, thank you for the opportunity in this project and we wish you the best success in your teaching the right people about safety.