Universiteit Leiden
The Netherlands

Opleiding Informatica

Using datamining for predicting

traffic jams on Dutch highways

Jeroen Holtes

Supervisors
Mitra Baratchi & Matthijs van Leeuwen

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

19/07/2019

Abstract

Many people travel by car in the Netherlands, this leads to traffic jams. The amount and length of traffic jams differ for each day. These differences could be due to various factors, like whether or not it is a holiday period or what the weather is like. This thesis will be about testing those factors and using them to predict traffic jams to answer the following research questions. The first research question is *How much influence do the following factors have on traffic jam length in the Netherlands: time, holiday periods and the weather?* and the second one is *How precise can we predict traffic jam length on the Dutch highways by using the factors: time, holiday periods and the weather?*. The influence of these factors is tested with hypothesis testing and feature selection. For predicting the traffic jams multiple regression algorithms are used. First a baseline is constructed based on related work. Next, a regression model is built on the hypothesis testing results. Then three types of regression are used (linear regression, regression trees and a neural network) and compared to each other, before and after feature selection has been used. The conclusion is that there was an decrease of 30% in regard to the RMSE score for the mean when using the regression tree. The regression tree algorithms scored the best of all the algorithms. Another conclusion is that time and holiday periods have the greatest influence on traffic jam length.

Contents

Chapter 1

Introduction

More than half of the Dutch adults own a car [Acea19] and many of them use those cars to get to their work. Because work times are the same for many people, there will be traffic jams just before and after work. There may also be other factors which impact traffic jams which may help to predict them. A problem with current traffic jam prediction models is that they are not precise enough, it could be that adding those other factors will make a more precise model.
But why study traffic jams? Because traffic jams have negative consequences for the environment. They increase local air pollution and contribute to global warming due to increased carbon dioxide emissions. On a personal level, time is also wasted that could have been spent more productively and some parts of the car could break down earlier due to a driving style that requires more braking and accelerating [ALKADI14]. Since the mid 1950s there have been studies about traffic [Helbing01], but the underlying factors for a traffic jam are still poorly understood. Multiple papers, [Nagel13] & [Orosz09] , give insight in *how* traffic jams form but they do not give insight in *why* or *when* they form. If there is a better understanding of the factors that have an influence on traffic jams in the Netherlands, these consequences can be reduced.

1.1 The problem

The problem with existing traffic jam prediction models is that they are not precise enough, so the usefulness of these models is limited. The difficulties for predicting traffic jams are that traffic peaks are highly volatile. The traffic intensity on the road differs day by day. By using hypothesis testing we will try to show the factors and their impact on the traffic jams and help explain why the length of the traffic jams differ day by day. Another difficulty is that it is possible that over time the use of the highway changes and that the highways are changed or maintained almost constantly, thus resulting in inconsistencies in the results. This could also result in a model that will perform well on the test data, but not so well on new data were there is no maintenance. A way to try to avoid this difficulty, which is used in this thesis, is to use a large time period to prevent

a temporary 'blocked' highway changing the overall picture. However, this does not take into account the changes that are made to the roads. If for example at the start of the data the road is a two lane highway with loads of traffic jams and after some road works it becomes a four lane highway with far fewer traffic jams, the models will not be able to take this into account.
The specific problem that this thesis tries to solve is, if by adding weather and holiday data we can get an improved model compared to an existing traffic jam prediction model.

1.2 Research questions

There are a lot of reasons why traffic jams exist at a particular date and particular time, some people blame the weather, others the time of day and some people also accuse the holidays. However, it is not clear what the decisive factor is or if it is a combination of factors. This research is partly about finding that out and the first question to help with that is:
How much influence do the following factors have on traffic jam length in the Netherlands: time, holiday periods and the weather?
Just knowing what influences the traffic jam length is not sufficient, but using those factors and predicting what the traffic jam length will be, is more useful for the drivers, police and Rijkswaterstaat. This is why the second research question is stated as:
How precise can we predict traffic jam length on the Dutch highways by using the factors: time, holiday periods and the weather?

1.3 Hypotheses

We expect that when the data is complete, when the weather and holiday data has no missing or incomplete data, our model with added weather and holiday factors will be more precise than an existing traffic jam prediction model without those factors.
The second hypothesis is that the time of day has the most influence on traffic jam length, followed by the weather, and that holiday periods have the least influence on the traffic jam length. It is thought that time has the most influence due to the rush hour, which is in our belief an important factor for traffic jams. The weather is expected to be the next most important, because it is shown that in America the weather has a big influence on the traffic [USDO15]. The holiday periods should also have an influence, but less than the other factors, because even though fewer people will travel to their work there will be additional holiday traffic, which we think will mostly even each other out.

1.4 Thesis Contribution

The novelty of this thesis is that for the Netherlands, according to the literature review for this thesis, the influence of weather on traffic has not been properly studied. Furthermore, the Netherlands has a high traffic density in comparison to other countries [CBS15], so the results of studies in other countries will not completely apply to the Netherlands. Another novelty is that we study traffic jams, while most other researches study traffic flow or travel time and use smaller time periods while we use a time period of an hour. In short:

1. We study the topic of the influence of weather on traffic, which according to the literature review for this thesis, has not been properly studied for the Netherlands.
2. We study traffic jams, while most other researches study traffic flow or travel time.
3. We use a larger time interval than most other studies.
4. We developed methods to test important factors for traffic jam length.

1.5 Thesis Overview

This chapter contained the introduction, Chapter 2 discusses related work. Chapter 3 explains the methodology used for the experiments, while Chapter 4 details the outline of the experiments and the results of those experiments. Finally Chapter 6 concludes this thesis.

Chapter 2

Related Work

Two of the more commonly used ways to build a regression model, for this kind of problem, are machine learning and time series based model building. Both methods have their advantages and disadvantages.

2.1 Machine Learning

Machine learning is the science of getting the computers to learn from data and information and have them act on it [Emerj19]. So for this thesis we want a model to learn to predict traffic jams by giving it traffic and weather data. The biggest advantage of machine learning is that it enables you to learn from a large volume of data and discover trends and patterns that would not be easy to discover for a human. Also machine learning models keep improving themselves accuracy and efficiency wise when they gain experience, to let them make better decisions. A disadvantage is that if the data is biased or incomplete the model will also be incomplete and make predictions that will not necessarily be correct. We think that this will not be a big problem for this thesis, because we use a large training set.
Machine learning is already used in some researches in the Netherlands to predict travel time, [Zeng13] and [Linto6]. These studies used neural network to network to predict short term travel time with a fairly high accuracy. These studies show that neural networks are an interesting avenue to explore for traffic predictions in the Netherlands. The methods they used can not be implemented for this thesis, because they had the travel times from earlier on the road as inputs and the logic for their models was specific to predict travel times.
In [Nikovskio5], the authors have tested how well linear regression, a neural network and a regression tree predict travel time for short term predictions. They used as predictor the previous travel times with their corresponding time and date. The authors observed that linear regression gave the best and most stable results. The neural network was behaving erratic, one run having the same error score as the linear regression model while the next run scoring was worse. The regression tree had a worse error score than the linear regression model. Their data was collected with 5 minute intervals, this is a difference with this thesis because the data here is collected with intervals of an hour. Another difference with this thesis is that they used only the travel

times with the corresponding time and date while this thesis uses also other factors as predictors, such as weather factors and if it is a holiday. In this thesis these regression methods will be implemented and compared to see if we get the same observations as the authors of [Nikovski05] and to see if by adding weather features we get better results. So the results of [Nikovski05] will be the baseline to compare our results with.

2.2 Time series

Time series models are predictive models that use time as the independent variable and then a value y, in our case traffic jam length. The output from such a model is the value y for a particularly time. A lot of time series models are used to predict the traffic flow [Ermagun18]. These models only predict the flow and not if there is a traffic jam or the length of it. In 1999 there were already studies where they used time series forecasting to predict the traffic flow [Williams99]. In [Jia17] the authors used time series forecasting methods to forecast the amount of traffic on a road around Beijing. They not only compared normal forecasting methods but also forecasting methods with a rainfall impact. Their conclusion was that the time series forecasting models with rainfall impact performed better than the models without the rainfall impact. Also in [Gosh09] multivariate time series forecasting was used to predict the traffic conditions in Dublin. The results they had were promising, but this forecasting model has only added value if the time period is around 15 minutes. The advantage of time series models is that they are great to identify seasonality and use this to help forecast. A disadvantage is that the models become more inaccurate how farther in the future they have to predict. In this thesis time series models will not be used, because there are a lot of missing instances in the data, when we plot the data over time. This makes it really difficult to predict accurately with a time series model.

Chapter 3

Methodology

To test if we get a more precise model by adding weather and holiday data, we have to do experiments. The experiments are performed according to the steps in Figure 3.1. Firstly, the data is collected in hourly intervals from multiple data sources, the weather data comes from the KNMI and the traffic data from Rijkswaterstaat. This data is online and for free available. Secondly, the data is preprocessed. Then a baseline will be constructed based on [Nikovski05] to compare our models with, this is a model without weather data, because then we can see if by adding the data we can make models that are more precise. The models that we compare the baseline with are then constructed, V1 is a self constructed regression model based on hypothesis testing done with T-tests, V2 are three standard regression models with all the features and V3 are the same three regression models as V2, but with selected features. Both V1 and V3 test which features are important, but the difference is that V1 only tests features based on literature or common beliefs and also shows their effect, while V3 tests all the features and ranks them on importance. Then the V1, V2 and V3 models will be compared to the baseline to determine their usefulness.

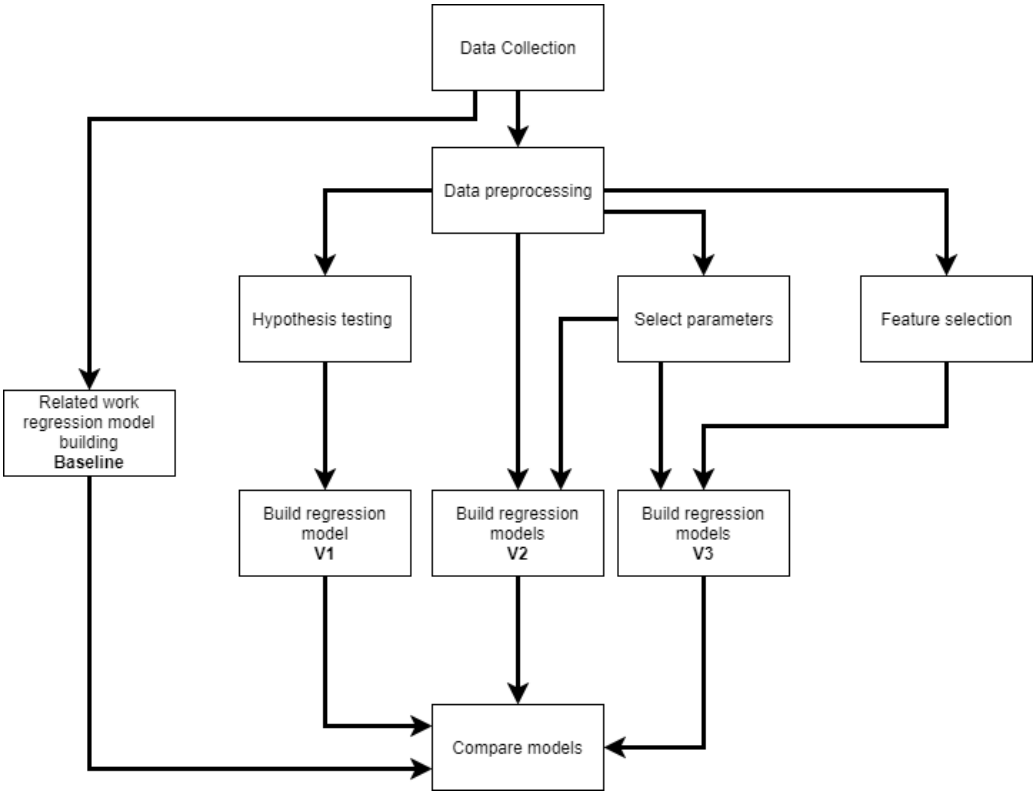

Figure 3.1: Research steps with corresponding names for the regression models implementations

3.1 Hypothesis testing

The following hypotheses will be used to give insight in how the factors influence the traffic jam length. We will compare the average traffic jam length of two groups for which one factor is different. Then we will determine if the means are statistically significantly different from each other and if so in what way. This is done because it will tell us what the influence of a factor is for; in this case the traffic jams in the time period 2012-2016, but also, for the whole population, the traffic jams after this time period.
To test the hypotheses, hypothesis testing will be done. Because the H_0^i are all in the format $\mu_1^i = \mu_2^i$, where i is the number of the hypothesis, we test if $\mu_1^i - \mu_2^i = 0$. Thus the tests were done according to the formula:

$$t = \frac{(\bar{y}_2 - \bar{y}_1) - 0}{se}$$

$$se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

And the confidence interval was constructed using the formula:

$$\bar{y}_2 - \bar{y}_1 \pm 1.96(se)$$

where the 1.96 means a 95% confidence interval. So that we can say, with 95% certainty, what the interval is that the difference between the two groups will be in for the population.
This is essentially a test to measure if the means of two different groups are the same [Agresti14]. Where $\bar{y}_1$ and $\bar{y}_2$ are the means of respectively group 1 and group 2. The standard deviations of each group are s_1 or s_2. And the number of instances in a group 1 is n_1 and for group 2 is that n_2. se stands for the standard error. t, the test statistic, is the number of standard errors that the means differ from each other. The bigger this is, the greater the evidence against H_0^i.

Many offices begin their day around 8 am and end their work day around 5 pm and according to [Hilberso6] traffic jams consist mainly of commuters traveling to or from work. Therefore, it is expected that during those times there is an increase in traffic jam length. Thus, the first hypothesis is:

H1: *Does the rush hour have an effect on traffic jam length?*
Let μ_1^1 be the average traffic jam length per hour outside the rush hour, where rush hour means the hours: {7, 8, 9, 16, 17, 18, 19}. And let μ_2^1 be the average traffic jam length per hour during the rush hour.

$$H_0^1 : \mu_1^1 = \mu_2^1$$

$$H_a^1 : \mu_1^1 \neq \mu_2^1$$

Most people do not work in the weekends so it would be expected that it is less busy on the highways during

those days. Therefore, it would not be busy during the rush hour either, this gives the second and third hypotheses:

H2: *Does the weekend have an effect on traffic jam length?*
Let μ_1^2 be the average traffic jam length per hour on a weekday and let μ_2^2 be the average traffic jam length per hour in the weekends.

$$H_{02} : \mu_1^2 = \mu_2^2$$

$$H_{a2} : \mu_1^2 \neq \mu_2^2$$

H3: *Does the rush hour in the weekend have an effect on traffic jam length?*
Let μ_1^3 be the average traffic jam length per hour in the weekends outside the rush hour and let μ_2^3 be the average traffic jam length per hour in the weekends during the rush hour.

$$H_{03} : \mu_1^3 = \mu_2^3$$

$$H_{a3} : \mu_1^3 \neq \mu_2^3$$

In holiday periods many people will not have to travel to work so the fourth hypothesis is:

H4: *Do holiday periods have an effect on traffic jam length?*
Let μ_1^4 be the average traffic jam length per hour in a nonholiday period and let μ_2^4 be the average traffic jam length per hour in a holiday period.

$$H_{04} : \mu_1^4 = \mu_2^4$$

$$H_{a4} : \mu_1^4 \neq \mu_2^4$$

According to the US department of Transportation traffic jams have three main sources, traffic−influencing events, traffic demand and transportation infrastructure [USDO15]. The traffic−influencing events include bad weather and is said that 15% of US traffic congestions come from this bad weather. So we would expect that in the Netherlands, traffic jams are also influenced by bad weather. When it rains the roads are more slippery,

thus the likelihood of accidents is higher. Therefore, more traffic jams are expected:

H5: *Does rain have an effect on traffic jam length?*
Let μ_1^5 be the average traffic jam length per hour while it does not rain and let μ_2^5 be the average traffic jam length per hour while it rains.

$$H_{05} : \mu_1^5 = \mu_2^5$$

$$H_{a5} : \mu_1^5 \neq \mu_2^5$$

When the sun shines or the temperature is high, people will be more inclined to take an alternative transportation method instead of the car. This results in less traffic on the highways and gives the sixth and seventh hypotheses:

H6: *Does the amount of sunshine have an effect on traffic jam length?*
Let μ_1^6 be the average traffic jam length per hour when there is no sunshine and let μ_2^6 be the average traffic jam length per hour when there is sunshine.

$$H_{06} : \mu_1^6 = \mu_2^6$$

$$H_{a6} : \mu_1^6 \neq \mu_2^6$$

H7: *Does the temperature have an effect on the traffic jam length?*
Let μ_1^7 be the average traffic jam length per hour when the temperature is below average and let μ_2^7 be the average traffic jam length per hour when the temperature is above the average of the data set.

$$H_{07} : \mu_1^7 = \mu_2^7$$

$$H_{a7} : \mu_1^7 \neq \mu_2^7$$

.

V1 Regression model

The results from the hypothesis testing are used to help build a regression model. This is done because the results from the hypothesis testing give an indication of the influence of a factor on the traffic jam length. For example the traffic jam length during the rush hour is on average greater than for the other hours. So for this model we assume that the traffic jam length during the rush hour is longer than when it is not a rush hour. We also assume that for the same circumstances the traffic jam length is the same. The approach for this model was invented. For all the hypotheses except **H3** a rule was defined:

1. If the alternative hypothesis is accepted, go to the next step, else do not use the hypothesis in the model.

2. Use least square estimate to estimate the coefficient for the hypothesis.

The reason that **H3** did not get included in the algorithm is that the combination of the rules for **H1** and **H2** should already include **H3**. **H3** is still tested, because it was thought to be interesting to see if this combination would hold. The rules are defined with the assumption that the traffic jam length increases or decreases with the found coefficient from the least square estimate. We do not use the values found with the hypothesis testing, because these values have to much noise from the other hypotheses in them. So we start the algorithm with a starting value, the average traffic jam length. Then we use the rules to predict the traffic jam lengths. This is the initial run. After this initial run, the algorithm is run again with a slightly different starting value to try to optimize the algorithm. After the optimizing the fitted algorithm is run on unseen data, to validate the results.
In formula this algorithm can be presented as:

$$E[y] = \alpha + \beta_1 H_1 + \beta_2 H_2 + \beta_4 H_4 + \beta_5 H_5 + \beta_6 H_6 + \beta_7 H_7$$

Where $E[y]$ is the prediction for this instance, α is the starting value, the intercept and β_i is the coefficient of the ith hypothesis and H_i is whether or not we should reduce or increase the prediction in regard to the ith hypothesis for this instance.

3.2 Regression

Regression is the process of predicting a numeric quantity. In this case we want to predict traffic jam length, a numeric value, so regression is a way to do this. Regression can be implemented using different methods. Three of those methods will be used in this paper: regression by using a neural network, linear regression and a regression tree. These three methods are used because in [Nikovski05] they used the same methods so we can compare the results they got with our own results for the V2 and V3 models.

3.2.1 Linear regression

Linear regression is the simplest of the three regression methods used in this research, but has as drawback that it is less flexible. It is a very static method. The idea of it is to represent the class as a linear function of the variables with their determined weights [Witten11]. It predicts a variable, the dependent variable, by using another variable, the independent variable. This is done by making a linear function with the formula:

$$E[y] = \beta_1 * i_1 + \beta_2 * i_2 + \beta_3 * i_3 + ... + \beta_n * i_n$$

Where $E[y]$ is expected value for the dependent variable, β_i is the coefficient for the ith independent variable and i_i is the ith independent variable.

3.2.2 Regression tree

A regression tree has instead of a classification in the leaf node a number that represents the average value of the instances that reach that leaf, we mean true or false or if we tried to predict a categorical variable one of its categories with classification . For problems that have different types of variables, a regression tree is more accurate for most of the time due to the fact that it can represent data in multiple ways. For example if a dataset with a boolean variable A plus a few more variables, reacts completely different on the independent variables when $A = true$ then if $A = false$, a simple linear model would have trouble to cope with this, but in a regression tree it would simply be a split. The downside of a regression tree instead of a linear model is that it is more difficult to understand when the size gets larger. [Witten11]

3.2.3 Neural network

A neural network consists of layers of nodes, the perceptrons, which get weighted inputs and have an activation function in it. If the weighted inputs activate the function, the perceptron fires to the next layer of nodes or the output. In case of regression, the neural network gives a numeric value as outcome instead of 0 or 1 . This numeric value is made up from the outputs of the last layer of perceptrons. Of the three models used this one is the least understandable due to the many nodes, which all have many inputs with different weights, thus making it difficult to understand how the value is determined. The upside of this method is that neural networks have a high tolerance to noisy data, so if the data is noisy the neural network should have the best predictions. [Aalst16]

3.3 Feature selection

Feature selection is used to remove the irrelevant features from the data and make the models perform and train better. It reduces the number of dimensions while keeping the loss of information as small as possible.

It has three main purposes, it reduces the training time, it reduces the chance of overfitting the model and most importantly it is used to remove irrelevant features such that the model is not based on those features. This is done because a model that is based on irrelevant features will be less accurate. Two feature selection algorithms will be used in this thesis, the algorithms 'mutual information regression' and 'f regression'. Both algorithms are of the Univariate feature selection kind. An Univariate feature selection algorithm examines each feature individually to determine the strength of the relationship between it and the dependent variable. The Univariate feature selection algorithms are in general good in gaining an understanding of the data. Both algorithms find the best features based on statistical tests and rank them. This will give an overview of the influence of the factors and help to answer the first research question.

3.3.1 F regression

F regression, also called f regressor, is a test used to score the features. This is done by performing Univariate linear regression tests. It is a linear model for the testing of each feature. This is done in two steps. In the first step is for every feature and target the correlation computed. Secondly, the correlation is converted to a F-score and then a p-value. This method in regard to mutual info regression only estimates the linear dependency between two variables, while mutual info regression can capture more kinds of statistical dependencies. The upside of using f regressor is that it computes faster than mutual info regression. The downside is that it only finds linear dependencies. [Sklearnf18]

3.3.2 Mutual information regression

Mutual info regression is a test that estimates the mutual information for a continuous target variable. Mutual information is the measure of the dependency between two variables. It can also be expressed as the amount of information one can get about one variable from observing another variable. The mutual information can not be negative and also not be greater than one, which means you can get all the information for the dependent variable from the independent variable. If the two variables have no relationship the mutual information value will be zero, while a higher value means a higher dependency. The values are computed by using non-parametric methods based on the entropy estimation from k-nearest neighbours distances. The benefit of using this method in regard to f regressor is that it gives a better insight into the dependency of two variables, but at the cost of using more time. This time can be slightly reduced by using fewer neighbours to compute the dependencies. [Sklearnm18]

Chapter 4

Evaluation

4.1 Experiment setup

We want to test if the added weather and holiday data help us construct models that have a smaller error and explain more of the variance than the already existing models. This is done by first collecting and preproccessing the data and then making a baseline model. After that we make a model based on the results from the hypothesis testing, which is the V1 model. The V2 model consist of three separate models:

1. DecisionTreeRegressor, an algorithm that makes a regression tree.
2. MLPRegressor, an algorithm that builds a neural network with a numeric value as output.
3. LinearRegression, an algorithm that makes a linear function from the input variables.

For the V2 models, parameters are selected and used to make the models. These same parameters will be used in the V3 models, which are the same three models as the V2 models but the input data is different as that the input data depends on the results from the feature selection. After the V1, V2 and V3 models are constructed they are compared to the baseline. All the models are constructed and run four times, because some of the models use a random initialization and by running them four times the randomness will be reduced.

4.1.1 Error metrics

To compare the accuracy of the constructed models we need to have error metrics. We have chosen the following two metrics, the first is the root mean square error(RMSE) and the second is R^2. RMSE is the standard deviation of the prediction errors. It has the formula:

$$RMSE = \sqrt{\overline{(p-y)^2}}$$

Where p is the predicted value and y is the real value. The benefit of the RMSE is that it gives an absolute fit instead of a relative fit.
The second scoring function is the R^2 scoring function. R^2 is defined as $1 - U/V$ where U and V are respectively:

$$U = \sum_i (y_i - p_i)^2$$

$$V = \sum_i (y_i - y_{mean})^2$$

y_i is the real value for y on the ith row, p_i is the predicted value for the same row and y_{mean} is the mean of all the real values. The R^2 has at best a score of 1 which is perfect predictions for all the values, a score of zero means that the value predicted is always as good as just using the mean of y. If R^2 has a score of 0, the model has no added value. R^2 was chosen because we can immediately see if the model has added value and with RMSE we can then determine what the exact added value is.

4.1.2 Data collection

The data used for the experiments is extracted from different sources. All the data is from the time period 01-01-2012 to 31-12-2016. The traffic data comes from the public accessible records of the Dutch agency Rijkswaterstaat [Rijkswaterstaat18], a slightly simplified example of the traffic data is shown in Table 4.1. This data is collected with the online web application of Rijkswaterstaat.

Date	Hm	Start time	End time	Heaviness	Average length	Duration	Day	Province	Route
20160102	35.2	1314	1514	422.267	3499.931	120.65	4	6	A1
20160102	60.8	1400	1416	41.183	2429.695	16.95	4	6	A12

Table 4.1: Example of traffic jam data, source Rijkswaterstaat

The Date is the date of the traffic jam in the format yyyymmdd. Hm is the highway location marker where the traffic jam started. The Start and End time are the times that the traffic jam started and ended. Average length is the total length of the traffic jam divided by the duration. Duration is how long the traffic jam lasted. Heaviness is the combination of the factors length and duration, see Table 4.4. Day is the day of the week that the traffic jam started. Province is the province where the traffic jam started and Route is the highway where the traffic jam originated.
The weather data comes from public data from the Dutch weather institute, the KNMI [KNMI18]. They provide the data per weather station on their website. Per Dutch province, one weather station that has no missing data has been used to provide the information. A list of those stations with their corresponding province is shown in Table 4.2. One thing to bear in mind is that in Table 4.2 the number 12 is missing. This is done to match the province numbers used by the traffic data. These numbers are the same numbers as used in Table 4.2.

Nr	Station	Province
1	Lauwersoog	Groningen
2	Leeuwarden	Friesland
3	Hoogeveen	Drenthe
4	Twenthe	Overijsel
5	Deelen	Gelderland
6	De Bilt	Utrecht
7	Schiphol	Noord-Holland
8	Voorschoten	Zuid-Holland
9	Vlissingen	Zeeland
10	Eindhoven	Brabant
11	Valkenburg	Limburg
13	Marknesse	Flevoland

Table 4.2: Used weather stations with corresponding provinces

All the weather data that is used is shown in Table 4.3.

Code	Definition
Station	The weather station the data is from
YYYYMMDD	The date where YYYY is the year, MM the month and DD the day
HH	Hour that is measured, HH = 5 means the time starting 4 AM till 5 AM
DD	Wind direction in degrees, coming from the north is 360 degrees and 270 degrees means the wind is coming from the west
FH	Average windspeed in the timeframe, measured in 0,1 m/s
FF	Average windspeed in the last ten minutes, measured in 0,1 m/s
FX	Maximal windspeed in the timeframe, measured in 0,1 m/s
T	Average temperature at 1,5 metres above the ground measured in 0,1 degrees Celcius
T10N	Minimum temperature at 10 centimetres from the ground in the last 6 hours, measured in 0,1 degrees Celcius
SQ	Duration of sunshine in the timeframe, measured in 0,1 hours
Q	Global radiation in the timeframe, measured in 0,1 J/cm2
DR	Duration of rain in the timeframe, measured in 0,1 hours
RH	Amount of rain in the timeframe, measured in 0,1 mm
P	Average airpressure in the timeframe, measured in 0,1 hPa
VV	Visibility in timeframe, measured in 0,1 km
N	How strong is the cloud cover, 0 no clover and 9 means that the upper air is invisible
U	Relative moisture in % at 1,5 metres above the ground
M	Fog, 0 = no fog, 1 = fog in the timeframe
R	Rain, 0 = no rain, 1 = rain in the timeframe
S	Snow, 0 = no snow, 1 = snow in the timeframe
O	Thunderstorm, 0 = no thunderstorm, 1 = thunderstorm in the timeframe
Y	Icing, 0 = no icing, 1 = icing in the timeframe

Table 4.3: Used weather data, source KNMI

All information regarding the School holidays came from the Internet. [Landen18] The instance is classified as in a holiday when at least one of the Dutch school regions has a school holiday for the primary and secondary schools.

4.1.3 Data preprocessing

To make the data compatible and do the experiments, the weather data and traffic data needed to be combined. First the unused variables were deleted from the traffic data, such that only the information in Table 4.4 remained. The variables Duration, Average length and Heaviness are numeric variables. The Day and Province are categorical variables where for the Day 0 means Monday, 1 means Tuesday and so on till 6 means Sunday. For Province the numbers corresponded with the number in Table 4.2. The Starting hour is an ordinal variable where value 1 means 1 AM and 2 is 2 AM so so on, until 23 which means 23 PM.

Code	Definition
Date	The date of the start of the traffic jam
Day	Which day of the week the traffic jam started, 0 = Monday, 1 = Tuesday , etc
Duration	The duration of the traffic jam in hours
Average length	The average length of the traffic jam in metres
Starting hour	At which hour the traffic jam formed
Heaviness	Combined factor of duration and average length. Duration * average length = Heaviness
Province	The province where the start of the traffic jam is located

Table 4.4: Traffic jam data after removal of unused variables

First the unused variables were removed from the traffic data; next the data was grouped and summed by date, starting hour and province such that for every day and all the hours that there was a traffic jam there was an entry to use in the experiments. Then the summed traffic data was combined with the holiday data, such that for every date there is a boolean variable that indicates whether or not it is a holiday. The weather data was then prepared for merging. The only thing that needed to change was that the weather data did not have a province. With the use of Table 4.2 that was easily solved, by replacing the station number with its corresponding province and changing the name of the station to its province. If the weather variable HH has value 4, it means the time starting at 3 AM until 4 AM, while if the traffic variable Starting hour has value 4 it means 4 AM until 5 AM. Thus the value 4 for HH corresponds with the value 3 for Starting hour, therefore all the HH values were decreased by 1 to match with the Starting hours, then the name HH was also changed to Starting hour to make the join easier. The weather and traffic data were then joined on Date, Starting hour and Province which resulted in Table 4.5. Then the Date was removed from the joined data, to not give the instance an unique identifier, the combination of the Date and the starting hour.

Date	Starting Hour	Province	Traffic Data	Weather Data	Holiday

Table 4.5: Combined data after preprocessing

4.1.4 Overview of the data

The Figures 4.1 to 4.6 give a quick overview of the data. They were all constructed in a very similar way. First of all the traffic data that was in the format of Table 4.5, was grouped by Weekday and starting hour and taken the mean of. Then the Figures 4.3 and 4.4 were constructed to show the average traffic jam length and duration for the whole period for every hour and each day. Then the holiday variable was used and the other figures were constructed to see the difference between a normal week and a holiday week. If for example a traffic jam started at 11 AM and had a duration of two hours with a length of 6 kilometres, the length and duration were only used for the 11 AM starting hour even tough the traffic jam also existed at 12 PM.
If we take a look at Figure 4.3 it shows for every day two peaks of traffic jam length between six am and nine am and also between four pm and six pm. This could mean that the rush hour definition used in 3.1 could be too narrow in the morning and too broad in the afternoon. Figures 4.1 and 4.2 use a completely different scale, this would suggest that in holiday periods there are fewer traffic jams, Section 4.1.6 will explore this observation further.

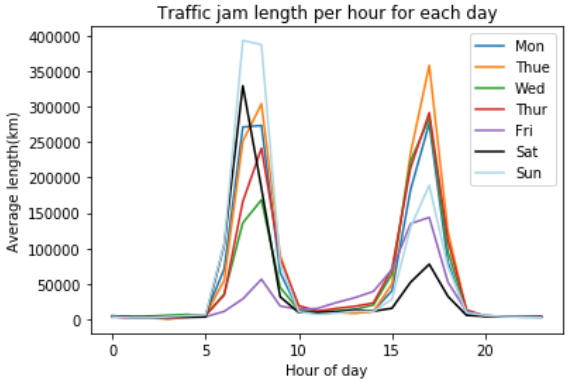

Figure 4.1: Traffic jam length in non holiday period

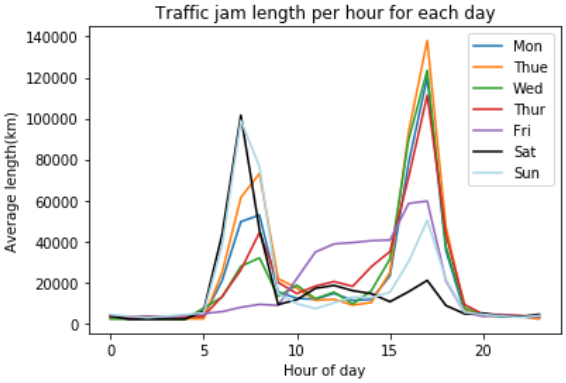

Figure 4.2: Traffic jam length in holiday period

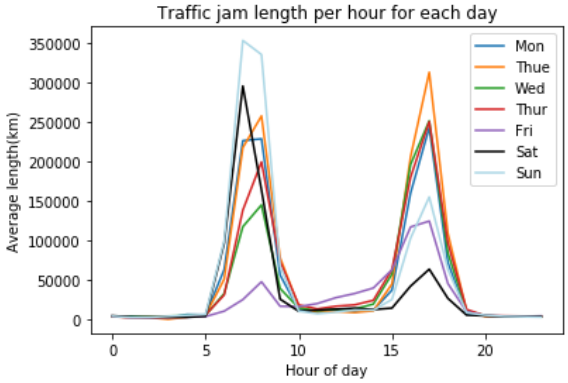

Figure 4.3: Traffic jam length over all periods

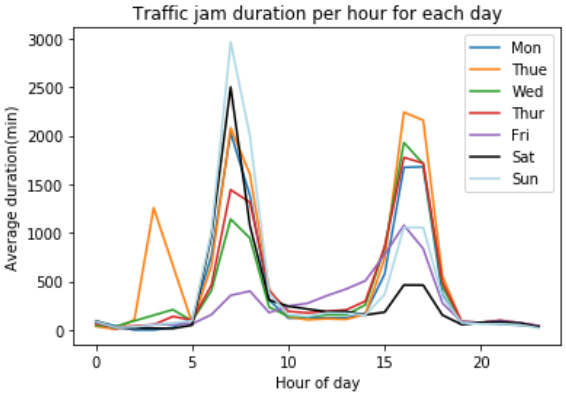

Figure 4.4: Traffic jam duration over all periods

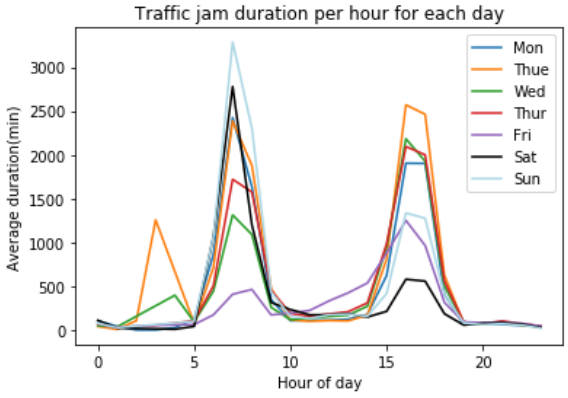

Figure 4.5: Traffic jam duration in non holiday period

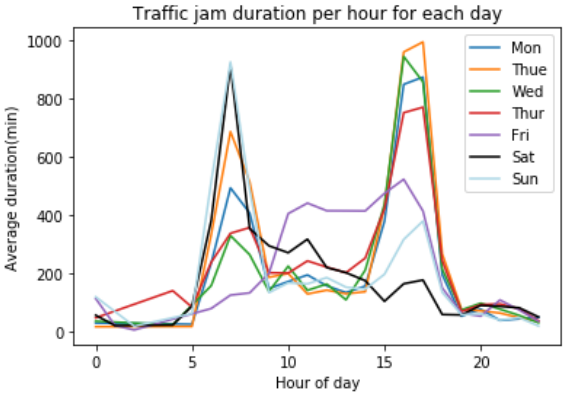

Figure 4.6: Traffic jam duration in holiday period

4.1.5 Baseline model

According to the research in [Nikovski05], three baseline regression models are built. The models are built twice, once on just the data for the province Noord-Holland and the other run is done on the whole data set. This is done because the V1 model is based only on the data from Noord-Holland and by making the baseline with the same data set we are then able to compare them. The input for the baseline models are only the time of day and the traffic jam length for that hour. The models are then tested in two ways, first data is split randomly in 10% testing data and 90% training data and then the models are built on this data. This method will give us a very optimistic result, because we overfit the data. It is possible that we predict for a point in time while we have trained with data points that are from further along. Secondly, the data is split more according to the research from [Nikovski05], where the first 12 months are training and the next 2 months are testing data. Here the first 4 years are used as training data and the last year is used as testing data. This second method will give us more realistic results, while the first method gives us optimistic results to compare them with.

4.1.6 V1 Hypothesis testing

All the hypothesis tests were done only on the biggest province, Noord-Holland with 14878 instances of the 70171 total instances. An instance is a data entry with the date, the time and the traffic jam length per province. The average traffic jam length per province is very variable per province so doing it this way gives a test with less noise than doing it on all the data. By using Python the mean, standard deviation and occurrence of every hypothesis was found.

H1: Does the rush hour have an effect on traffic jam length?
Let μ_1^1 be the average traffic jam length per hour outside the rush hours, where rush hours are the hours: {7, 8, 9, 16, 17, 18, 19}. And let μ_2^1 be the average traffic jam length per hour during the rush hours.
$H_0^1 : \mu_1^1 = \mu_2^1,\ H_a^1 : \mu_1^1 \neq \mu_2^1$
The mean, standard deviation and number of instances for these circumstances are:

- $\bar{y}_1^1 = 6819.12$ *metres*
- $\bar{y}_2^1 = 22675.917$ *metres*
- $s_1^1 = 7076.874$
- $s_2^1 = 24774.214$
- $n_1^1 = 6961$
- $n_2^1 = 7917$

$$se = \sqrt{\frac{7076.874^2}{6961} + \frac{24774.214^2}{7917}} = 291.066$$

$$t = \frac{22675.917 - 6819.12}{291.066} = 54.48$$

The n_1^1 and n_2^1 are really big. Therefore, t = 54.48 is enormous which gives a P-value that is 0 with more than five decimal places. Thus we can reject the H_0^1 and accept H_a^1 The confidence interval is:

$$Upperbound = 22675.917 - 6819.12 + 291.067 * 1.96 = 16427.286 \; metres$$

$$Lowerbound = 22675.917 - 6819.12 - 291.067 * 1.96 = 15286.31 \; metres$$

Thus on average in rush hours the traffic jam length per hour is 15 to 16 kilometres higher than in the normal hours.

H2: Does the weekend have an effect on traffic jam length?
Let μ_1^2 be the average traffic jam length per hour on a weekday and let μ_2^2 be the average traffic jam length per hour in the weekends. $H_0^2 : \mu_1^2 = \mu_2^2,\ H_a^2 : \mu_1^2 \neq \mu_2^2$
The mean, standard deviation and number of instances are:

- $\bar{y}_1^2 = 15721.557 \; metres$
- $\bar{y}_2^2 = 13383.196 \; metres$
- $s_1^2 = 20212.598$
- $s_2^2 = 20610.037$
- $n_1^2 = 11922$
- $n_2^2 = 2956$

This gives a standard error of 421.862 and a t-score of -5.54. This results in a 95% confidence interval of [-1511.512, -3165.21]. n_1^2 and n_2^2 are big. Therefore, the t-score of -5.54 gives a P-value that is 0 with more than five decimal places. Thus, we reject H_0^2 and we accept H_a^2. With the confidence interval we derive that in the weekends the average traffic jam length per hour is smaller than the average traffic jam length per hour for a weekday.

H3: Does the rush hour in the weekend have an effect on traffic jam length
Let μ_1^3 be the average traffic jam length per hour in the weekends outside the rush hour and let μ_2^3 be the average traffic jam length per hour in the weekends during the rush hour. $H_0^3 : \mu_1^3 = \mu_2^3,\ H_a^3 : \mu_1^3 \neq \mu_2^3$
The mean, standard deviation and number of instances are:

- $\bar{y}_1^3 = 6047.653 \; metres$
- $\bar{y}_2^3 = 21921.640 \; metres$
- $s_1^3 = 5533.917$

- $s_2^3 = 27354.646$
- $n_1^3 = 1590$
- $n_2^3 = 1366$

This gives a standard error of 753.025 and a t-score of 21.08. This results in a 95% confidence interval of [17349.917, 14398.06]. n_1^3 and n_2^3 are big. Therefore the t-score of 21.08 gives a P-value that is 0 with more than five decimals. Thus, we reject H_0^3 and we accept H_a^3. With the confidence interval it is derived that for the rush hour in the weekend the average traffic jam length per hour is greater than for the hours outside the rush hour.

H4: Do holiday periods have an effect on traffic jam length?
Let μ_1^4 be the average traffic jam length per hour in a nonholiday period and let μ_2^4 be the average traffic jam length per hour in a holiday periods. $H_0^4 : \mu_1^4 = \mu_2^4,\ H_a^4 : \mu_1^4 \neq \mu_2^4$
The mean, standard deviation and number of instances are:

- $\bar{y}_1^4 = 16824.823$ *metres*
- $\bar{y}_2^4 = 8781.172$ *metres*
- $s_1^4 = 21551.290$
- $s_2^4 = 12102.255$
- $n_1^4 = 11978$
- $n_2^4 = 2900$

This gives a standard error of 298.800 and a t-score of -26.92. This results in a 95% confidence interval of [-7458.005, -8629.297]. n_1^4 and n_2^4 are big. Therefore, the t-score of -26.92 gives a P-value that is 0 with more than five decimal places. Thus we reject H_0^4 and accept H_a^4. With the confidence interval we derive that the average traffic jam per hour in a holiday period is lower than in a nonholiday period.

H5: Does rain have an effect on traffic jam length?
Let μ_1^5 be the average traffic jam length per hour while it does not rain and let μ_2^5 be the average traffic jam length per hour while it rains. $H_0^5 : \mu_1^5 = \mu_2^5,\ H_a^5 : \mu_1^5 \neq \mu_2^5$
Rain is defined as when the value RH, see Table 4.3, is bigger than 0.
The mean, standard deviation and number of instances are:

- $\bar{y}_1^5 = 14745.798$ *metres*
- $\bar{y}_2^5 = 18647.871$ *metres*
- $s_1^5 = 19632.900$
- $s_2^5 = 24080.376$

- $n_1^5 = 12929$
- $n_2^5 = 1949$

This gives a standard error of 572.129 and a t-score of 6.820. This results in a 95% confidence interval of [5023.446, 2780.700]. n_1^5 and n_2^5 are big. Therefore the t-score of 6.820 gives a P-value that is 0 with more than five decimal places. Thus, we reject H_0^5 and accept H_a^5. The confidence interval shows that the traffic jam length for an hour is greater when it rains than for an hour when it does not rain.

H6: Does the amount of sunshine have an effect on traffic jam length?
Let μ_1^6 be the average traffic jam length per hour when there is no sunshine and let μ_2^6 be the average traffic jam length per hour when there is sunshine. $H_0^6 : \mu_1^6 = \mu_2^6,\ H_a^6 : \mu_1^6 \neq \mu_2^6$
Sunshine is defined when the value SQ is bigger than 0.
The mean, standard deviation and number of instances are:

- $\bar{y}_1^6 = 16836.530$ *metres*
- $\bar{y}_2^6 = 13669.311$ *metres*
- $s_1^6 = 22178.784$
- $s_2^6 = 18109.438$
- $n_1^6 = 7458$
- $n_2^6 = 7420$

This gives a standard error of 331.895 and a t-score of -9.54. This results in a 95% confidence interval of [-2516.705, -3817.733]. n_1^6 and n_2^6 are big. Therefore, the t-score of -9.54 gives a P-value that is 0 with more than five decimal places. Thus, we reject H_0^6 and accept H_a^6. With the confidence interval we derive that the traffic jam length is on average greater when there is no sunshine then when there is sunshine.

H7: Does the temperature have an effect on the traffic jam length?
Let μ_1^7 be the average traffic jam length per hour when the temperature is below average and let μ_2^7 be the average traffic jam length per hour when the temperature is above average. $H_0^7 : \mu_1^7 = \mu_2^7,\ H_a^7 : \mu_1^7 \neq \mu_2^7$
The average temperature is determined by taking the mean of the T feature, this gives a T value of 122.4 or an average temperature of 12.24 °c.
The mean, standard deviation and number of instances are:

- $\bar{y}_1^7 = 17342.071$ *metres*
- $\bar{y}_2^7 = 13242.400$ *metres*
- $s_1^7 = 22058.810$
- $s_2^7 = 18246.536$

- $n_1^7 = 7311$
- $n_2^7 = 8567$

This gives a standard error of 332.497 and a t-score of -12.33. This results in a 95% confidence interval of [-3447.976, -4751.366]. n_1^7 and n_2^7 are big. Therefore, the t-score of -12.33 gives a P-value that is 0 with more than five decimal places. Thus, we reject H_0^7 and accept the alternative hypothesis. The confidence interval shows that the traffic jam length is less when the temperature is above average compared to when it is below average.

4.1.7 Least square estimate

The results from the hypothesis testing give an indication what the influence from the hypotheses is on traffic jam length, but only for a single hypothesis, they do not have taken into account what the relation between the hypotheses is. That is why a least square estimate is also used for the same hypotheses, except H3, because H3 is a combination of two hypotheses itself. The results from the least square estimation are shown in Table 4.6. The meaning for the table is for example that for H1, does the rush hour have an effect on traffic jam length, the traffic jam length goes up by 13254 metres when it is a rush hour. The bias term is the intercept.

Hypothesis	**Least square estimate**
H1	13254
H2	-1250
H4	-5646
H5	4398
H6	-137
H7	-1230
Bias	8318

Table 4.6: The least square estimate for the hypotheses

4.1.8 Parameter selection for V2

Just as for the baseline three algorithms are used.

1. DecisionTreeRegressor, an algorithm that makes a regression tree.
2. MLPRegressor, an algorithm that builds a neural network with a numeric value as output.
3. LinearRegression, an algorithm that makes a linear function from the input variables.

The DecisionTreeRegressor is implemented with the default settings except for the max depth. The max depth is set to 10 to achieve the best result, see Figure 4.7. It was decided to use this algorithm for its ability to represent data in multiple ways and by being the only regression tree in Sklearn [Sklearn19]. Sklearn is a free machine learning library for Python, which includes algorithms for classification, clustering and regression. It is used because, it has algorithms which are easy to implement and is free to use.

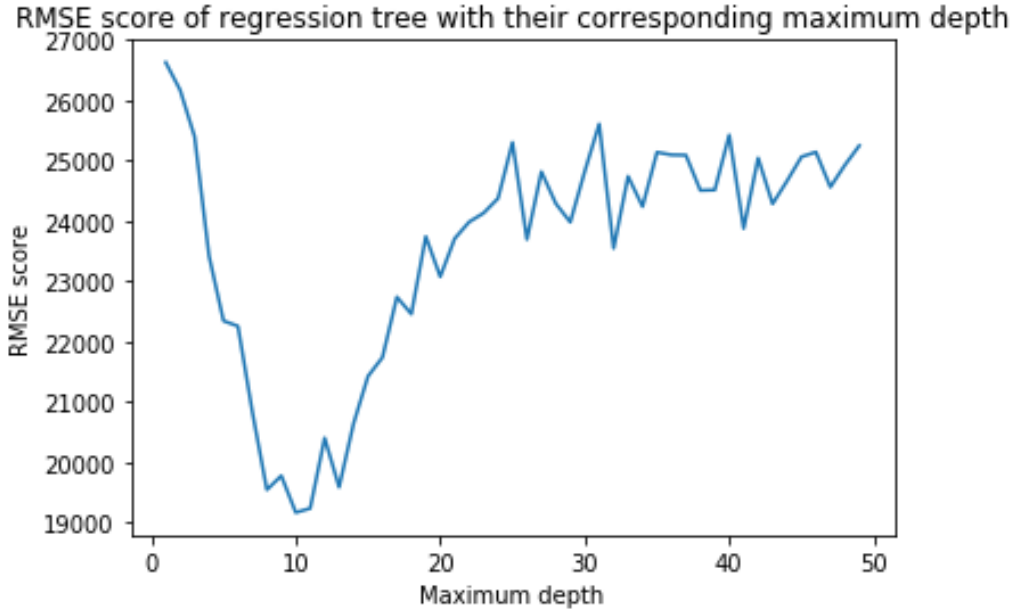

Figure 4.7: RMSE score for DecisionTreeRegressor on test data with corresponding maximum depth

The MLPRegressor is implemented with mostly the default settings, except for the hidden layer size. The hidden layer size is increased from one layer of one hundred perceptrons to three hidden layers of one hundred perceptrons each. This was done because it improved the test results of the early test runs. The MLPRegressor was chosen due to the amount of noise in the data, which originated partly from the gathering of traffic data. The LinearRegression algorithm was implemented with only the default settings. It was chosen because the features of the algorithm are easily analysed.
Due to not only running nominal variables it is expected that the LinearRegression algorithm scores lower than the other two.

4.1.9 V3

To reduce the number of features and thereby reducing the noise in the data, feature selection is used. The selected features will then be used in the three prediction algorithms used in chapter 4.1.8 and scored by RMSE. The used feature selection algorithms are:

1. F regression
2. Mutual information regression

These algorithms were used because they were the only algorithms suited in Sklearn for regression while doing feature selection. The six best features are then used to build regression models. The ten best ranked features from the F regression are shown in Table 4.7. The six highest ranked features are used to make the models. Therefore, these features were used for the regression models. The scores of the Mutual information regression are shown in Table 4.8. The differences between the six best scoring features from F regression

and Mutual information regression are quite big, for example the two highest scoring features from Mutual information regression are not even in the highest ten of the F regression.

Rank	Name
1	holiday
2	Q
3	U
4	T
5	T10
6	DR
7	VV
8	Day
9	SQ
10	R

Table 4.7: Ten highest ranked features for the F regression. Ranked from highest to lowest

Name	Mutual information	Name	Mutual information
Hour	0.175	R	0.002
Province	0.105	DD	0.002
Q	0.029	S	0.001
holiday	0.025	SQ	0.000
T	0.019	M	0.000
Day	0.011	Y	0.000
VV	0.010	U	0.000
T10	0.008	N	0.000
DR	0.006	FX	0.000
O	0.005	FH	0.000
P	0.005	FF	0.000
RH	0.003		

Table 4.8: Estimated mutual information per feature in regard to average traffic jam length

4.2 Results

4.2.1 Baseline results

The results of these baseline models are shown in Table 4.9. This table shows the results when run on the Noord-Holland data set, Table 4.10 shows the result for the whole data set. A baseline to compare the scores to is the RMSE score for the mean. If the RMSE is lower than RMSE score for the mean, then the model has added value. When we look at the results from Noord-Holland where the last year is the test data, we see that only the DecisionTreeRegressor predicts better than the mean. This does not correspond to the results of the other article, where the LinearRegression model scored better than the other two models and all the models scored better than the mean. An explanation for this phenomenon could be that they used smaller time periods and their models were based on only a few time periods earlier, for example the prediction for 8 am was based on the data for 7:55 am and 7:50 am. Here the prediction is based on all the earlier data. When we look at the randomly split data we see that all the models score better than the mean baseline of 20312 metres, but just as the other split only the DecisionTreeRegressor explains any of the variance in the data. The last year as testing data run scores worse than the run where the data is randomly split. This is logical, because when we randomly split the data the results are too optimistic and the last year as test data gives a more realistic picture. If we look at the results for the whole data set they are mostly the same as for the Noord-Holland build, the biggest difference is that all the models perform slightly worse compared to the mean.

		LinearRegression	MLPRegressor	DecisionTreeRegressor	Mean
Random split	RMSE	20023	19315	17674	20312
	R^2	0.00	0.01	0.24	0.00
Last year as test data	RMSE	27669	27654	24163	27476
	R^2	-0.07	-0.07	0.18	0.00

Table 4.9: Baseline for Noord-Holland

		LinearRegression	MLPRegressor	DecisionTreeRegressor	Mean
Random Split	RMSE	27244	26614	24919	27291
	R^2	0.00	0.05	0.16	0.00
Last year as test data	RMSE	34474	34427	31430	34438
	R^2	-0.04	-0.04	0.13	0.00

Table 4.10: Baseline for whole data set

4.2.2 V1 results

The results for the V1 model are shown in Figures 4.8 and 4.9. In Figure 4.8 the data is split with the last year as testing data for Noord-Holland. Then we get an RMSE score of 25099. This means that on average the prediction has an error of 25099 metres. If we use the simplest prediction algorithm (the mean) then we get a RMSE score of 27476. This means that we got an error decrease of 8.7%. This regression model decreased the average error by almost 9%, which is quite promising for further research and means that it added value.
The regression model has a R^2 score of 0.118 when the last year is used as test data, this means that almost 12% of the variance in the data is explained by this regression model. The V1 model therefore, explains more than a ninth of the variance in the data, so the model gives an insight in how much the tested factors influence traffic jam length and how precise these factors can predict traffic jam length when they are combined.
When we look at the Figure 4.8 we see that the model does not totally predict the extend of the peaks and valleys, it does not correctly predict the height of the peaks.
When the data is randomly split we get as result the Figure 4.9. The RMSE score for this model is 18561, thus on average the prediction has an error of 18561 meters. The simplest prediction has for this data set a RMSE score of 20321. This means that the error is decreased by 8.7%. This regression model decreased the average error by almost 9%, which is quite promising for further research and means that it has some added value.
The R^2 score is 0.169 when we split the data randomly, so we explain more of the variance of the data then when we use the last year as test data. Figure 4.9 can just like Figure 4.8 not predict the extend of the peaks and valleys.

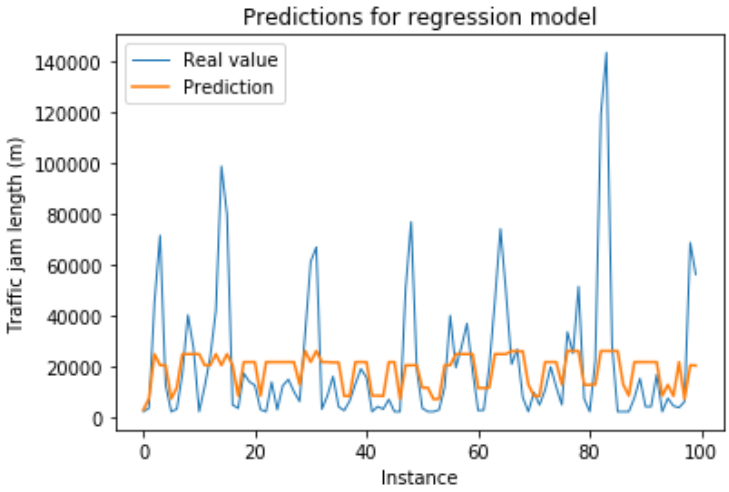

Figure 4.8: Result of V1 for Noord-Holland with last year as test data

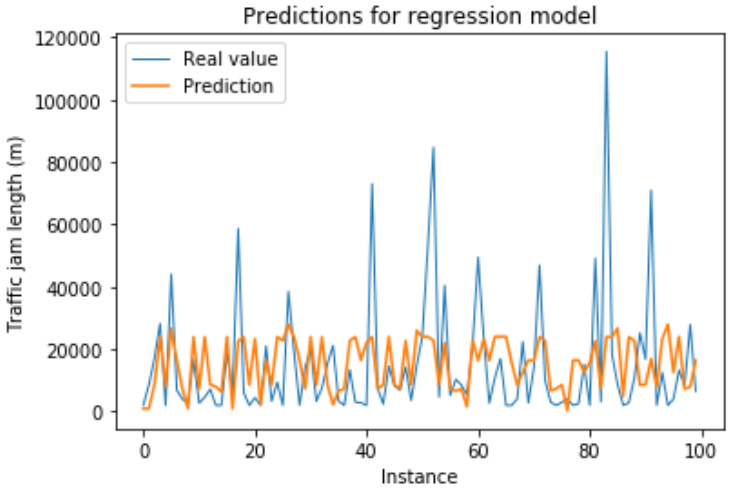

Figure 4.9: Result of V1 for Noord-Holland with data randomly split

4.2.3 V2 results

The result for the V2 models are shown in Table 4.11. It shows that the DecisionTreeRegressor easily outscores the other two algorithms. The scores for each algorithm are shown in Table 4.11, the MLPRegressor and

		LinearRegression	MLPRegressor	DecisionTreeRegressor	Mean
Random Split	RMSE	26233	25504	19197	27291
	R^2	0.07	0.12	0.50	0.00
Last year as test data	RMSE	33327	32738	26866	34438
	R^2	0.02	0.06	0.37	0.00

Table 4.11: Results of V2 models for whole data set

LinearRegression Algorithms have only little to no added value, but the DecisionTreeRegressor explains 50% of the variance in the data and predicts moderately accurate for the random split. Even with the last year as test data split, the DecisionTreeRegressor still explains 37% of the variance. For the LinearRegression this score was partly expected due to the use of not only numeric values, but also categorical values. The linear function is fitted to a non linear relation, which helps to explain the lower score.
The RMSE scores in Table 4.11 are derived in the same fashion as the average R^2 score, the average over four runs. If we compare these average scores with the mean we find that for MLPRegressor and LinearRegression the gain is only around four to six percent, but for the DesicionTreeRegressor the gain is 30% which is quite a good decrease in the overall error.
Figures 4.10, 4.11 and 4.12 give the hundred first predictions and their corresponding real value of the three algorithms. The first hundred were used because the graph became unreadable when using more instances. Interesting to note is that the LinearRegression algorithm and MLPRegressor are quite stationary. They move a bit around the mean and follow the real values slightly, but the DecisionTreeRegressor moves way more with the real predictions.
The three algorithms were then also tested on the same data set as for the hypothesis testing in Section 4.1.6. The scores are presented in Table 4.12. The RMSE score of 17267 for the DecisionTreeRegressor is better than the RMSE score of The MLPRegressor and the LinearRegression algorithms, respectively 19836 and 19537. In regard to the R^2 score the same pattern showed itself. The DecisionTreeRegressor had a score of 0.30, the LinearRegression algorithm had a score of 0.12 and the MLPRegressor had a score of 0.09. When we look at the scores for the last year as test data, we see that the RMSE scores are a lot closer and that the algorithms explain almost none of the variance.

		LinearRegression	MLPRegressor	DecisionTreeRegressor	Mean
Random Split	RMSE	19836	19537	17267	20312
	R^2	0.12	0.09	0.30	0.00
Last year as test data	RMSE	26352	26159	25313	27476
	R^2	0.03	0.04	0.10	0.00

Table 4.12: Results of V2 models for Noord-Holland

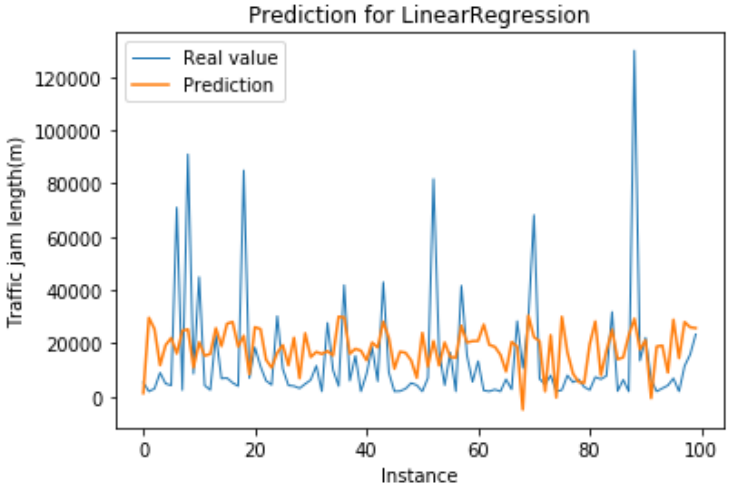

Figure 4.10: The 100 first predictions of the LinearRegression algorithm and their real values

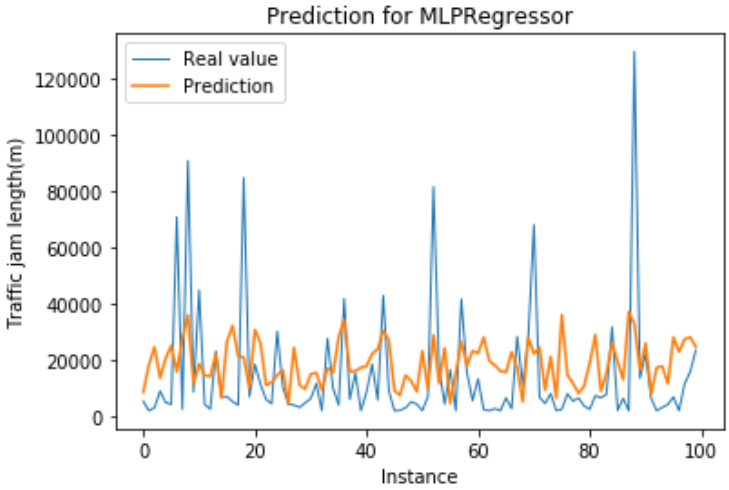

Figure 4.11: The 100 first predictions of the MLPRegressor algorithm and their real values

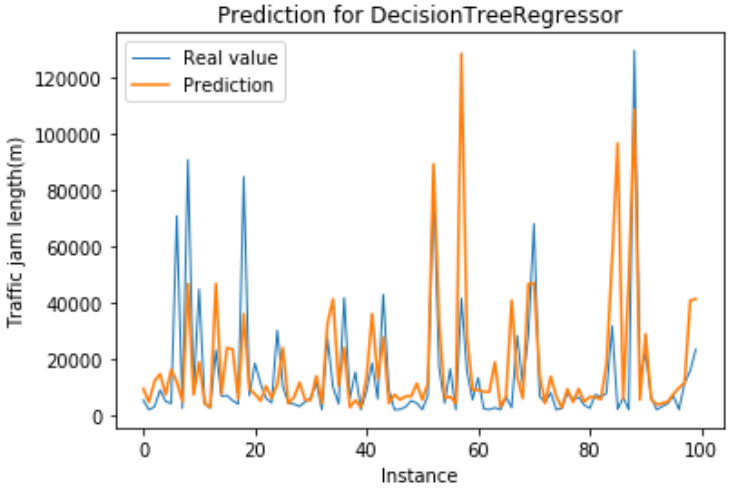

Figure 4.12: The 100 first predictions of the DesicionTreeRegressor algorithm and their real values

4.2.4 V3 results

The results of the V3 models are shown in Table 4.13. If we compare the F regression result with the results in Table 4.11, it shows that this method of feature selection did not improve the results of the models. In the case of the DecisionTreeRegressor it made the model a lot worse, which is logical because the F regression algorithm looks only at linear relation while the DecisionTreeRegressor is more versatile.
If we look at the RMSE results of Mutual information regression, see Table 4.13. The most important find is that the results of the MLPRegressor are 16%, with mutual information $>= 0.025$, or 12% with a mutual information of $>= 0.01$ better than the original results, see Table 4.11. Also the results from the DecisionTreeRegressor are where they were in the original model. The mutual information regression helps to reduce the number of features while the models lost almost no important information and the MLPRegressor functioned even better due to the fact that it lost some noise. Table 4.14 shows the RMSE results for the models when only applied for the province Noord-Holland. Here the algorithms run with slightly different features, because for the F regression the duration of the rain was ranked less important for the smaller dataset and replaced with the day of the week. While for the Mutual information regression we obviously had to take the province out of the features, because it has only one value left. These results follow the same trend as the results in Table 4.13, except for the DecisionTreeRegressor when the Mutual information threshold is >=0.025. There the results improved in comparison to the higher threshold instead of getting worse as in Table 4.13. A possible explanation is that in Noord-Holland the DecionTreeRegressor is more influenced by the top four scoring features than the whole data set. The results from the models with the features selected with Mutual information regression and a threshold >=0.025 were the best scoring. Therefore, these results will be used to compare with the other implementations in Section 4.3. Table 4.15 shows the RMSE results for the models when we do not use a random 10% as test data, but instead the last year as test data. Here also the DecisionTreeRegressor outscores the other two algorithms and the MLPRegressor and the DecisionTreeRegressor have the best score when we used Mutual information regression with a threshold >=0.025, this is why these results were also used in Section 4.3. The last Table 4.14 shows the RMSE results for the Noord-Holland subset of the data where the last year is used as test data. Here we see that only the DecisionTreeRegressor scores significantly better than the other algorithms for the Mutual information selections. We can also note that for this smaller dataset the F regression has a lower error than the Mutual inormation selections for the MLPRegressor. Figures 4.13, 4.14 and 4.15 show the first 100 predictions for the algorithms after using mutual information regression. If we compare the figures with each other the biggest difference is that while the MLPRegressor and the DecisionTreeRegressor follow roughly the real value, the LinearRegression algorithm stays around the mean. If we compare Figure 4.15 with Figure 4.10 they show the same pattern, both centered around the mean and no big movements. Also Figure 4.14 and Figure 4.12 show the same kind of movement, which the RMSE score would suggest. But when we compare Figure 4.13 with Figure 4.11, there is a great deal of difference and the biggest difference is that the predictions are not as much centered around the mean. The fact that the predictions also follow roughly the trend of the real values means that the RMSE has improved quite a bit.

	Mutual_information	**Mutual_information**	**F regression**
Treshold	>=0.01	>=0.025	
LinearRegression	26125	26728	26435
MLPRegressor	22761	21852	26281
DecisionTreeRegressor	19417	20329	26405

Table 4.13: RMSE scores for V3 models with corresponding feature selection method with the data random split

	Mutual_information	**Mutual_information**	**F regression**
Treshold	>=0.01	>=0.025	
LinearRegression	19700	19492	20306
MLPRegressor	19620	18180	19996
DecisionTreeRegressor	17935	17307	21177

Table 4.14: RMSE scores for V3 models with corresponding feature selection method for the province Noord-Holland with the data random split

	Mutual_information	**Mutual_information**	**F regression**
Treshold	>=0.01	>=0.025	
LinearRegression	33816	33830	33625
MLPRegressor	32126	30967	33555
DecisionTreeRegressor	26669	26190	33829

Table 4.15: RMSE scores for V3 models with corresponding feature selection method with the last year as test data

	Mutual_information	**Mutual_information**	**F regression**
Treshold	>=0.01	>=0.025	
LinearRegression	26917	27053	26368
MLPRegressor	27316	26962	26283
DecisionTreeRegressor	24867	24069	26711

Table 4.16: RMSE scores for V3 models with corresponding feature selection method for the province Noord-Holland with last year as test data

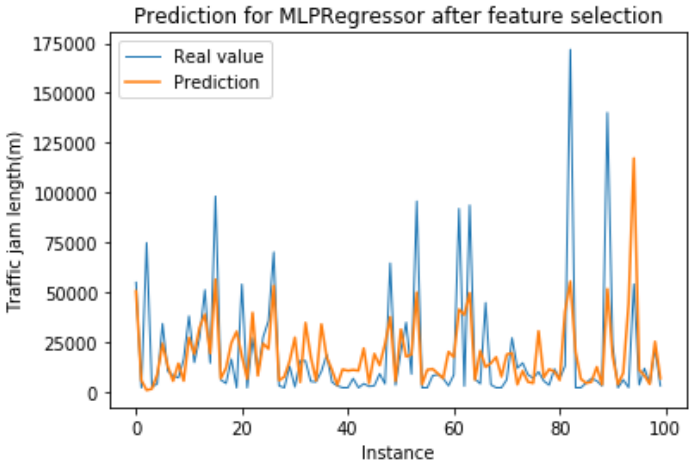

Figure 4.13: The 100 first predictions of the MLPRegressor algorithm and their real values after using mutual information feature selection

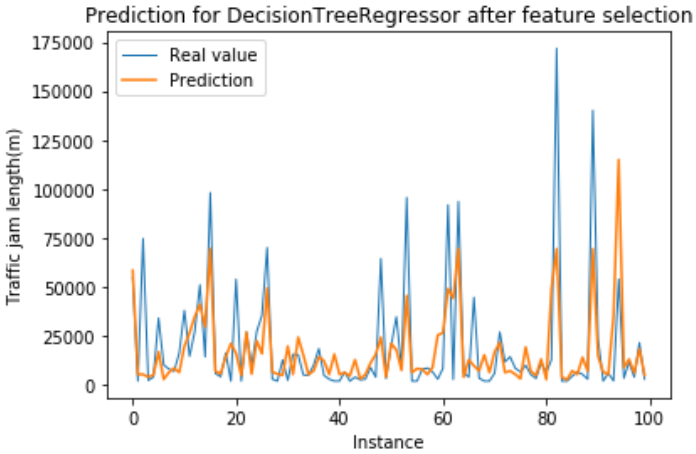

Figure 4.14: The 100 first predictions of the DecisionTreeRegressor algorithm and their real values after using mutual information feature selection

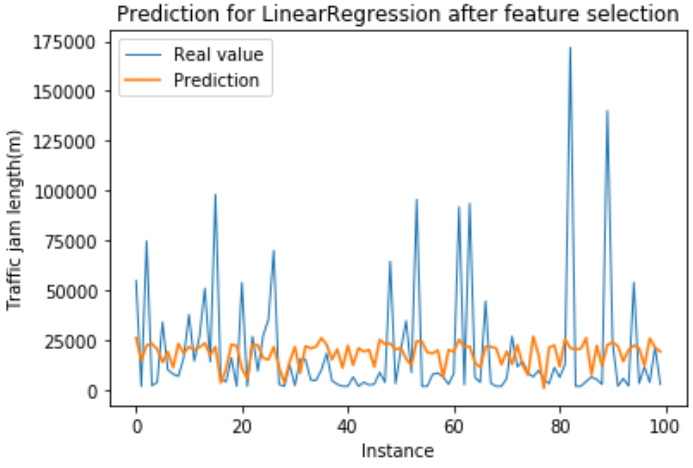

Figure 4.15: The 100 first predictions of the LinearRegression algorithm and their real values after using mutual information feature selection

4.3 Comparison of the models

When we look at the results for all the methods used for the province Noord-Holland in Table 4.17, the first conclusion is that all implementations score better than the mean, thus all models had added value. The second conclusion is that all the implementations of the DecisionTreeRegressor outscore all the other algorithms. Thirdly, the V1 model outscores the LinearRegression models and MLPRegressor models most of the time, except for the V3 version of the MLPRegressor. When we look at the results for the whole data set in Table 4.20. Both the V2 and V3 score better than the baseline models. The most precise model for traffic jam prediction is according to this research the DecisionTreeRegressor. And all implementations scored better than the mean. When we look at the result in graph form, Figures 4.16 and 4.17, the biggest conclusion is that when we look at a smaller part of the data set all the DecisionTreeRegressor models score almost the same and outscore all the models. When we look at the whole data set it is easy to see that the V2 DecisionTreeRegressor outscores all the other models by quite a margin and the baseline DecisionTreeRegressor scores a whole lot worse than the V2 and V3 implementations of the decision tree. When we look at the results of the data where the last year was used as test data, then we see the same pattern as for the randomly split data. Here also the DecisionTreeRegressor outscores all the other implementations. Interesting to note is that for the results of the Noord-Holland subset the V2 DecisionTreeRegressor scores worse than the V1 model and the DecisiontreeRegressor implementations of the Baseline and V3 models, while for the randomly split data the V2 DecisionTreeRegressor scored better. Another observation we can make is that for the whole dataset the V3 DecisionTreeRegressor scores the best when we use the last year as test data, while if we split the data randomly the V2 DecisionTreeRegressor scores the best. The improvement compared to the mean for the best model over the whole dataset for both splits is around the same, around 25%.

				Mean
V1	18561			
	LinearRegression	**MLPRegressor**	**DecisionTreeRegressor**	
Baseline	20023	19315	17674	**20321**
V2	19836	19537	17267	
V3	19492	18180	17307	

Table 4.17: Results of all the regression models for the province Noord-Holland with the data randomly split

	LinearRegression	**MLPRegressor**	**DecisionTreeRegressor**	**Mean**
Baseline	27244	26614	24919	
V2	26233	25504	19197	**27291**
V3	26727	21852	20329	

Table 4.18: Results of all the regression models for the whole data set with the data randomly split

				Mean
V1	25099			
	LinearRegression	**MLPRegressor**	**DecisionTreeRegressor**	
Baseline	27669	27654	24163	**27476**
V2	26352	26159	25313	
V3	27053	26962	24069	

Table 4.19: Results of all the regression models for the province Noord-Holland with the last year as test data

	LinearRegression	**MLPRegressor**	**DecisionTreeRegressor**	**Mean**
Baseline	34474	34427	31430	
V2	33327	32738	26866	**34438**
V3	33830	30967	26190	

Table 4.20: Results of all the regression models for the whole data set with the last year as test data

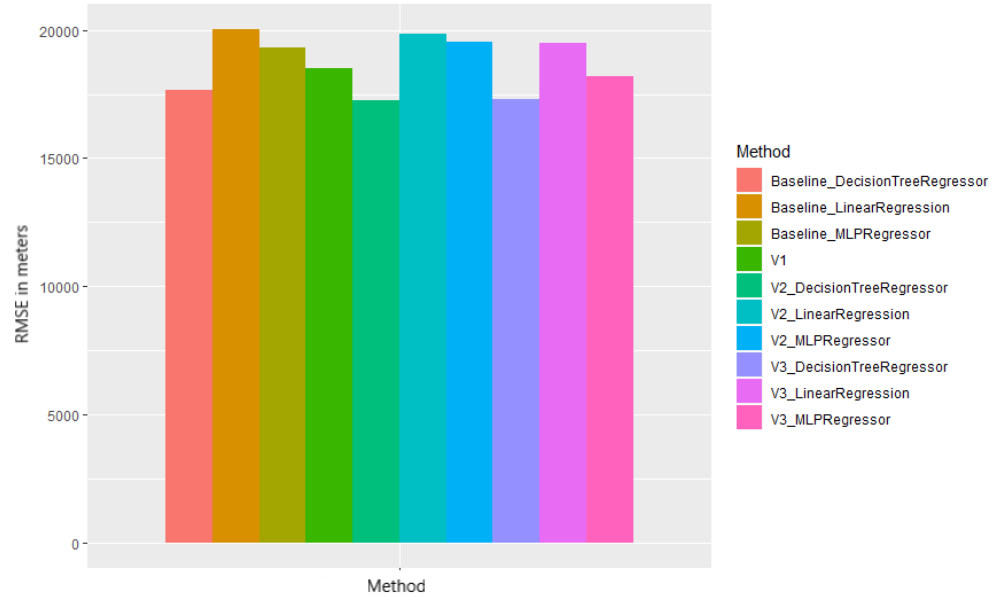

Figure 4.16: RMSE score in metres for the methods for the province Noord-Holland

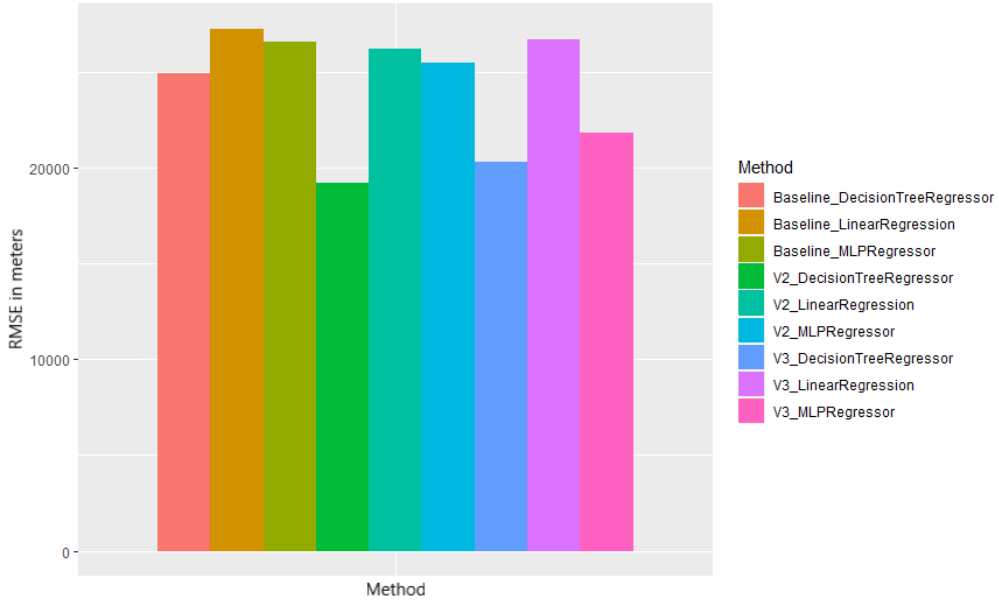

Figure 4.17: RMSE score in metres for the methods for the whole dataset

Chapter 5

Discussion

There are some limitations for this research. First of all, the time period that is used is large. This has as advantage that when a road is blocked for a small part of the time period, that it does not influence the results much and the overall image is correct. The downside is that when the road changes, for example the number of lanes becomes more, the used models will not be able to take this into account. So if there are fewer traffic jams due to road changes, the models will predict accurately, they will predict that there are more traffic jams than there are and have a higher error. A way how this could have been done different is by taking a smaller time period. A change in the highway would then not have mattered, because such a change takes quite a bit of time. Therefore with a small time period, the before and after situation of the highway will not be in the same data set, only one of the situations will be in a data set. So the models will predict better for the roads where the situation changes over the long term.
Another limitation is that the models give the same weight to all data points, the data from one day before the test data is just as important as the data from three years before. The advantage is that the model gives a good overall image, so when the data is stable over the long term, then the results will be fairly accurate. But the disadvantage is that when the data is not stable, that the results will not be as accurate. Traffic data is not very stable data, in the Netherlands the length of the traffic jams grows by the year. There are also made quite a lot of changes to the roads, which leads to different traffic situations and differences in the data. Giving weight to more recent data will help to take these changes into account. Using a small time period will make giving weight to the more recent instances redundant, because the changes that will be taken into account with the added weight will not exist for a small time period.

Chapter 6

Conclusions

To answer the first research question :*How much influence the following factors traffic jam length in the Netherlands: time, holiday periods and the weather?*. We use the information from Section 4.1.6 and Section 4.1.9. It is shown there that time and holiday are great influences on the traffic jam length. In Section 4.1.6 we can see that during rush hour the average traffic jam length increases by around 16 kilometres. This section also showed that the average traffic jam length decreases by around 8 kilometres during holidays. The weather factors also have influence on the traffic jam length but not to the same extent, if it rained the traffic jam length was only around 4 kilometres greater than when it was not raining. This was not as predicted with the general hypothesis, where it was expected that the weather would have more influence than the holiday periods.
The second question: *How precise can we predict traffic jam length on the Dutch highways by using the factors: time, holiday periods and the weather?* is answered in Sections 4.2.2, 4.2.3 and 4.2.4. By using the V2 or V3 LinearRegression models on this data set where the data is randomly split we can not predict the traffic jams precisely, but the V1 model can predict traffic jams slightly better. Its RMSE scores is 5% better than the score of the best scoring LinearRegression algorithm, the V3 LinearRegression model. However the V3 MLPRegressor can predict traffic jams more accurately, it has an increase of 16% for its accuracy score. In this case none of the algorithms can predict as accurately as the V2 DecisionTreeRegressor, not even the V3 MLPRegressor. Even when we take only the data for Noord-Holland when it is randomly split, the V2 DecisionTreeRegressor outperforms all the other implementations as shown in Section 4.3. However when we use the last year as test data, we see that for Noord-Holland the V3 DecisionTreeRegressor scores the best, but barely better than the Baseline model so the added value for the subset of the data is great. However, when we look at the whole data set we see that the V3 DecisionTreeRegressor scores much better than the Baseline model. According to the general hypothesis we could predict the data more precisely than the existing models when the data is complete, this is the case for when we look at the whole data set, but not when we look at a subset of the data.

6.1 Further research

For further research, the LinearRegression models could be optimised by making smaller data sets from only 1 province, 1 hour and 1 day to eliminate the categorical data and noise from the inputs, but for that the data set used needs to be bigger. Also not covered in this research is if the snow that already is on the road has influence on the traffic jam length. For the V1 model more factors can be inserted or use it to process other provinces to make it better and more versatile. Lastly the change over time is not researched, if possibly the influence of the factors is different in 2012 than it is for 2016. This could be an interesting and important factor to research.

Bibliography

[Hilbers06] H. Hilbers, D. Snellen, A. Hendriks. *Files en de ruimtelijke inrichting van nederland.* NAi Uitgevers. 2006.

[Helbing01] D. Helbing. *Traffic and related self-driven many-particle systems.* Reviews of Modern Physics 73. 2001.

[Nagel13] K. Nagel, P. Wagner and R. Woesle. *Still Flowing: Approaches to Traffic Flow and Traffic Jam Modeling.* Operations Research 51(5):681-710. 2013.

[Orosz09] G. Orosz, R. E. Wilson, R. Szalai and G. Stépán. *Exciting traffic jams: Nonlinear phenomena behind traffic jam formation on highways.* Physical Review 80. 2009.

[USDO15] US Department of Transportation. *Traffic congestion and reliability: trends and advanced strategies for congestion mitigation*. 2015. Url: http://www.ops.fhwa.dot.gov/congestion_report/chapter2.htm

[ALKADI14] O.ALKADI, O. ALKADI, R. ALSAYYED1, J. ALQATAWNA. *Road scene analysis for determination of road traffic density*. Frontiers of Computer Science. 8(4): 619628. 2014.

[Ermagun18] A. Ermagun, D. Levinson.*Spatiotemporal traffic forecasting: review and proposed directions*. Transport Reviews. 1-29. 2018.

[Rijkswaterstaat18] nis.rijkswaterstaat.nl *Openbare historische filegegevens*. Accessed: 2018. Url: https://nis.rijkswaterstaat.nl/SASPortal/main.do

[KNMI18] knmi.nl *Uurgegevens van het weer in Nederland*Accessed :2018. Url: https://www.knmi.nl/nederland-nu/klimatologie/uurgegevens

[Landen18] Landen.net *Schoolvakanties.* Accessed: 2018. Url: http://www.landen.net/schoolvakanties/

[Witten11] I. H. Witten, E. Frank, M. A. Hall. *Data Mining, Practical machine learning tools and techniques*. 2011. Third edition. Elsevier.

[Aalst16] W. van der Aalst. *Process Mining, Data Science in Action*. 2016. Second edition. Springer.

[Sklearnf18] scikit-learn.org *f regression.* Accessed: 2018. Url: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_regression

[Sklearnm18] scikit-learn.org *mutual info regression.* Accessed: 2018. Url: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_regression

[Sklearn19] scikit-learn.org *scikit-learn.* Accessed: 2019. Url: https://scikit-learn.org/stable/index.html

[Agresti14] A. Agresti, B. Finlay *Statistical methods for the Social Sciences*. 2014. Fourth Edition. Pearson. Pages 191-193.

[Jia17] Y. Jia, J. Wu, M. Xu *Traffic Flow Prediction with Rainfall Impact Using a Deep Learning Method*. 2017. Hindawi.

[CBS15] CBS *Transport and mobility 2015*. 2015. Statistics Netherlands.

[Nikovski05] D. Nikovski, N. Nishiuma, Y. Goto, H. Kumazawa *Univariate Short-Term Prediction of Road Travel Times*. IEEE Intelligent Transportation Systems. Pages 1074-1079. IEEE.

[Zeng13] F. Zeng, H. van Zuylen *Urban link travel time estimation based on sparse probe vehicle data*. Transportation Research Part C: Emerging Technologies. 2013. Volume 31. 145-157. Elsevier.

[Lint06] J. van Lint *Reliable Real-Time Framework for Short-Term Freeway Travel Time Prediction*. Journal of Transportation Engineering. 2006. Volume 132. 921932.

[Gosh09] B. Ghosh, B. Basu, M. OMahony *Multivariate short-term traffic flow forecasting using time-series analysis*. IEEE transactions on intelligent transportation systems. 2009. Volume 10. 246.

[Williams99] B. M. Williams, l. A. Hoel *Modeling and Forecasting Vehicular Traffic Flow as a Seasonal Stochastic Time Series Process*. 1999.

[Acea19] acea.be *Vehicles Per Capita, by Country.* Accessed: 2019. Url: https://www.acea.be/statistics/tag/category/vehicles-per-capita-by-country

[Emerj19] emerj.com *What is Machine Learning?* Accesed: 2019. Url: https://emerj.com/ai-glossary-terms/what-is-machine-learning/