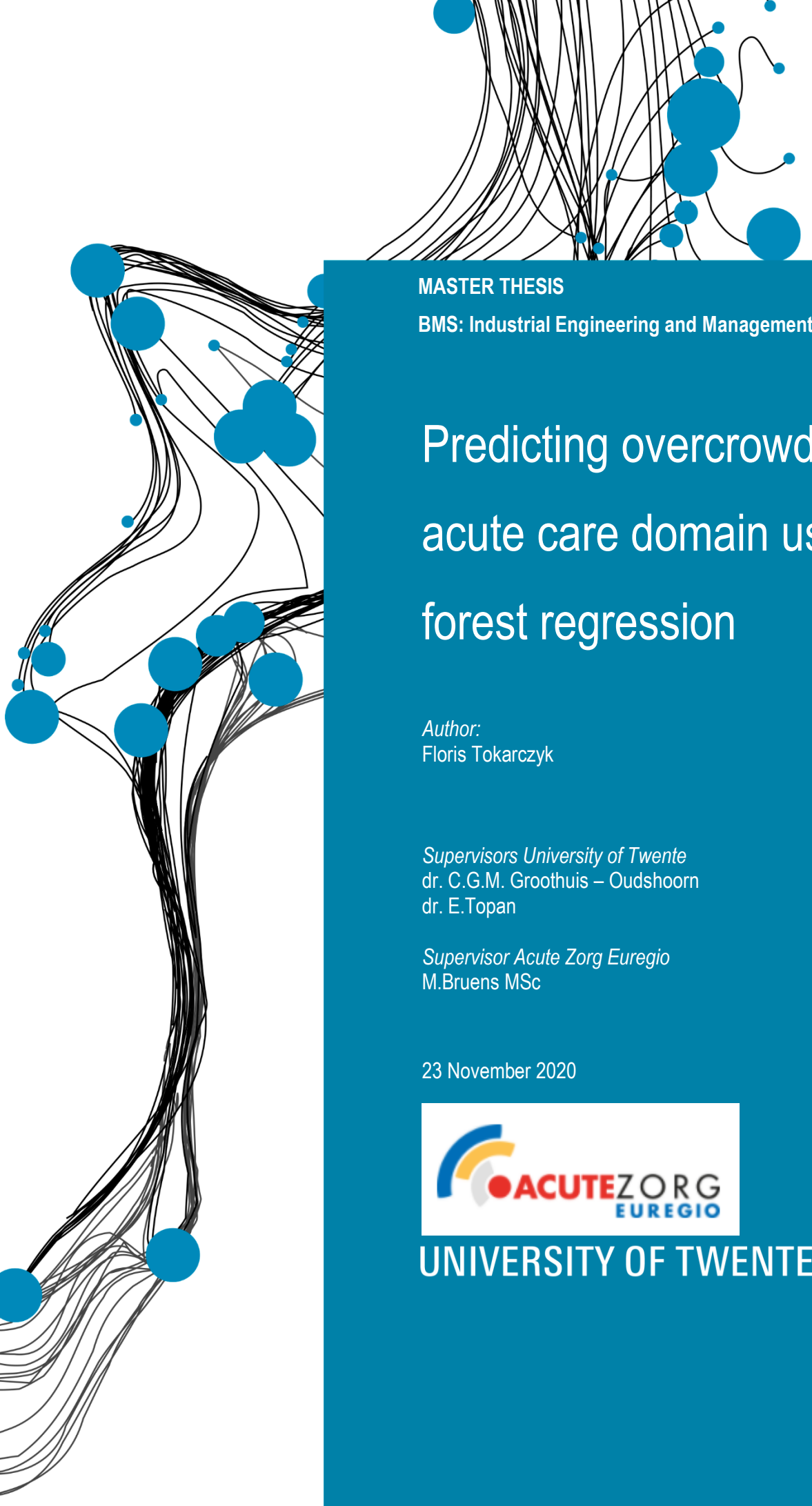

MASTER THESIS

BMS: Industrial Engineering and Management

Predicting overcrowding in the acute care domain using random forest regression

Author:
Floris Tokarczyk

Supervisors University of Twente
dr. C.G.M. Groothuis – Oudshoorn
dr. E.Topan

Supervisor Acute Zorg Euregio
M.Bruens MSc

23 November 2020

ACUTEZORG EUREGIO

UNIVERSITY OF TWENTE.

Management Summary

Crowding is a phenomenon that occurs more frequently within the facilities of the acute care domain. Crowding causes patient waiting times to increase and overall satisfaction levels to decrease. Additionally, medical employees experience stress caused by the increased workload. Although crowding is perceived by patients and medical employees, there is no clear definition or measure for crowding. Similarly, there are also no clear indicators of crowding, which means that facilities cannot prepare for this. In this research, we focus on the day to day crowding in three acute care facilities in the region of Oost-Achterhoek. These are;

- The ambulance service in north and east Gelderland, Witte Kruis NOG.
- The ED of the hospital in Winterswijk, Streekziekenhuis Koningin Beatrix (SKB).
- The GP-post of Oost-Achterhoek, which is part of the general practitioners' care of eastern Achterhoek (HZOA).

For these partners, we want to quantify crowding, find potential predictors of crowding, and create a machine learning model that can predict crowding. We aim to answer the following research question in this report:

> ***What machine learning model can be used as an adequate early warning system for overcrowding and what is its performance in the acute care domain in the region of Oost-Achterhoek?***

We started the research by getting familiar with the processes and the degree of overcrowding at the different ED, GP-post, and ambulance services. Right at this point, we were also hit by the Covid-19 pandemic, which resulted in the withdrawal of the ambulance services from the project. Additionally, we had to use old datasets for the ED and GP-post as we were no longer able to retrieve new data from them. The dataset of the ED contained data from 2012-2018 of 85048 patients. The dataset of the GP-post contained data from 2013-2017 of 149725 patients.

We proceeded by doing literature research on measures and predictors for crowding in the acute care domain. The literature described various measures commonly used for crowding in the acute care domain but unfortunately, not all of these were appliable to our situation. Mainly caused by the fact that we could not retrieve any new data from the partners. We decided to use total daily visitors as a measure of crowding since these were determinable with the old datasets. The idea is then to predict this measure one-day-ahead, allowing planners to adjust schedules of employees if crowding is expected and predicted within reasonable boundaries. The predictors of crowding that we investigated were chosen based on findings in the literature and opinions of experts in the field. We used the following predictors:

- Visitors of the day before in the acute care domain
- Date related data (day of the week, the month of the year, etc)
- German and Dutch Holidays
- Pollen data (allergenic and non-allergenic)
- Events data (music festivals, sports events, etc)
- Weather data (temperature, amount of rainfall, etc)

After we had defined our measure for crowding and potential predictors of crowding we had to find the best machine learning algorithm applicable to our situation. Literature research was conducted to find related research that tried to forecast daily patients within the acute care domain. Based on that research we did additional literature research on machine learning algorithms to find a method

that best fits our situation. We decided to use random forest regression to predict the daily visitors at the GP-post and the ED. Factors that contributed to this decision were the following:

- The understandability of the method for people unfamiliar with machine learning
- The accuracy and ability of built-in validation using out-of-bag data.
- The ability to quite easily tune the models for better performance
- The variable importance can easily be derived to find relevant predictors.
- The ability to deal with categorical and numerical values without transformations.

The available data was analyzed and put together into datasets specifically for the ED and GP-post. The total number of features in the datasets is 105 and we reduced these to 47, we refer to these sets as the full dataset and the reduced dataset respectively. The deleted features were in the pollen and weather datasets. For the pollen set, we selected only those that were allergenic to people as we expect the remainder not to affect the health of people. For the weather set, we selected the features that were expected to affect people's health or their decision-making of visiting the acute care domain and we deleted features that were closely related to another.

We then ran an optimization on four variations of the datasets with two different validation techniques (the bootstrap method and cross-validation) to find the best models. The performance of the models was determined with the MAE, RMSE, and MAPE. Additionally, the models validated by bootstrap also have an out-of-bag score, which is related to the RMSE. The four variations that we tested were; full dataset with no events, a full dataset with events, a reduced dataset with no events, and a reduced dataset with events. This was done because the period over which we had event data was very small. We found that the addition of events did not have an improvement in the performance of the models for both the GP-post and ED. We also found that the models made with the reduced datasets most of the time performed slightly better than the models created with the full datasets. The best model found for the GP-post and the ED are summarized in Tables 1-2 for both of the validation methods.

Table 1 Best Models out of all scenarios found by the Bootstrap method. The first entry is the best GP-post model and the second entry is the best ED model.

Dataset	Num of Trees	Max Features	Min Samples	Sample Size	Run Time (s)	MAE	RMSE	MAPE	OOB Score
GP Reduced (No Event)	1000	20	3	0.9	4.69	8.32	11.44	12.49	0.96
ED Reduced (No Event)	1000	5	5	0.8	2.05	5.02	6.29	15.94	0.13

Table 2 Best models out of all scenarios found by the cross-validation method. The first entry is the best GP-post model and the second entry is the best ED model.

Dataset	Num of Trees	Max Features	Min Samples	Run Time (s)	MAE	RMSE	MAPE
GP Reduced (No Event)	1000	29	5	4.12	8.39	11.51	12.57
ED Full (No Event)	1000	22	2	5.12	5.01	6.29	15.86

The models for the GP-post were mostly explained by time-related features. With one in particular, which is whether the day is on the weekend or not. Then of less importance are features such as the day of the week, whether it is a holiday or not, the number of GP-post visitors of the previous day, and some weather-related features. The models for the ED were mostly explained by the weather-related-features, albeit not so much, as the OOB score indicates.

In its current state, the one-day-ahead forecasts produced by the best models that we found will not be an adequate early warning system for overcrowding, since the degree of uncertainty is too large.

The range of the predictions still varies too much to be used for employee schedules. The models were also compared with some simplistic baseline predictions (the value of one day before, an average of the last 3 days, the value of one week before, and the average over the days from 1, 2, and 3 weeks ago). We found that the models for the GP-post were significantly better than the baseline predictions, but the ED only managed to perform slightly better. Further research and improvements, as well as newer data, are required to improve the performance of the models.

Preface

This thesis has been written to finish my Master's degree in Industrial Engineering and Management at the University of Twente. This concludes my student period after first finishing a bachelor in Engineering Physics and a bachelor in Industrial Engineering and Management at the Saxion University of Applied Sciences.

First I would like to thank Manon from Acute Zorg Euregio who was my supervisor during this project. Although, I was not able to conduct my research at the office due to the Covid-19 pandemic. I was still able to get a lot of feedback and quick responses from her whenever I needed them through e-mail or the telephone.

Secondly, I would also like to thank Karin as my first supervisor from the University of Twente for guiding me through this project and providing me with good feedback whenever needed. Also, I want to thank Engin for being my second supervisor and also for providing feedback.

Lastly, I want to thank my friends and family for supporting me and keeping me motivated.

Floris Tokarczyk
Enschede, November 2020

Table of Contents

List of Figures

List of Tables

1 Introduction

This chapter will function as an introduction to the company where the research will be conducted, the problem description, and the research design. First, a short introduction will be given about the company in section 1.1. We then proceed with the motivation behind this research and the problem description in section 1.2. In section 1.3 we will look at the research design that is proposed to solve the core problem. The research methodology that will be used during this research is explained in section 1.4. The scope of the research will be defined in section 1.5. The deliverables of this research project will be presented in section 1.6. Lastly, the thesis outline will be presented in section 1.7

1.1 Acute Zorg Euregio

Acute Zorg Euregio (AZE) is one of the eleven acute care networks in the Netherlands. The network of AZE consists of acute care facilities located in Twente, Oost-Achterhoek, and the German border, see Figure 1.

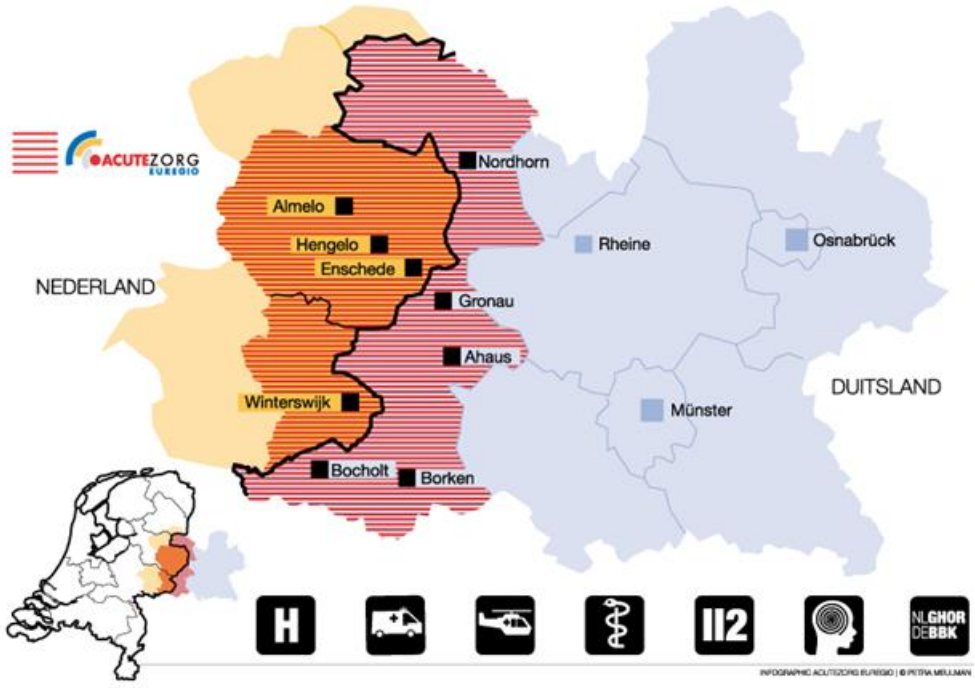

Figure 1 The region in The Netherlands and Germany that work together with Acute Zorg Euregio.[1]

The whole network has a coordinating function concerning optimizing acute care in their region. The importance of the patient is always paramount. AZE has a coordinating, stimulating, and facilitating role in the acute care chain to be able to carry outs its (legal) duties in coordination with their chain partners. Consultation with chain partners is conducted at different levels; with the directors of the acute care facilities, with the managers of the acute care facilities, and with professionals in the varying expert groups. This is all part of the so-called: 'Regionaal Overleg Acute Zorgketen (ROAZ)' which translates to regional consultation acute care chain. The activities of AZE are divided into the following subjects:

Acute care chain: Varying healthcare institutions and professionals work together when acute care is required for a patient. AZE ensures that within the network the spread and availability of acute care

[1] Illustration retrieved from https://www.acutezorgeuregio.nl/over-ons/

remain guaranteed. Expert groups have been created around emergency indications and focus areas to ensure and improve the quality of the provided care. This research will focus on this branch of their work field in collaboration with a hospital, general practitioner post, and an ambulance service.

Trauma care chain: Trauma care involves the whole acute care chain. All chain partners within the network aim to optimize the care of the trauma patient. This includes; general practitioners, general practitioner posts (GP-post), regional ambulance facilities, emergency departments (ED), departments in hospitals, and mobile medical teams. They make agreements about cooperation and monitor the quality of the care in different ways.

Crisis management & OTO: Certain regulations and procedures are important during crises. AZE is involved in preparing chain partners for certain disasters and crises. The procedure during large incidents is also called; scaled-up care. During this crisis, certain activities, such as triage, treatment, and allocation will be prioritized for wounded people. AZE also provides education, training, and practice opportunities (in Dutch (OTO): opleiden, trainen en oefenen) to prepare for certain events.

Knowledge center: One of the legal duties of AZE and its network is to share knowledge about acute care. One of the activities is the provision of training within the acute care domain. These are developed and implemented in collaboration with the partners. Research related to acute care and scaled-up care is also set up and carried out. Research projects are carried out in collaboration with chain partners, other acute care networks, Saxion university of applied Sciences, and the University of Twente. This thesis is one example of many pieces of research that have been (or will be) conducted at AZE to contribute to a more developed acute care network.

Cross-border acute care: AZE is the only network in the Netherlands that provides cross-border cooperation. They work with acute care facilities in the region of the German border. Both countries share information and patients and try to improve the care within their network just as the other networks in the Netherlands try to accomplish.

1.2 Problem Context

In this section, we will have a closer look at the problem in this research. The motivation behind the research will be given, the problem description will be provided and the core problems that will be addressed within this research will be listed.

1.2.1 Research motivation

Overcrowding is a phenomenon that occurs frequently in the acute care domain. More than a third of the EDs in the Netherlands experience overcrowding more than once a week (van Loghum, 2013). On top of that more than two-thirds of the managers of EDs experience overcrowding in their department multiple times per week (van der Linden C. , et al., 2014). Studies also reveal that overcrowding of the ED is associated with lower quality of care for the patient, in case of severe pain and normal situations (Hwang, et al., 2008) (Pines & Hollander, 2008). Acute care providers in the Netherlands (Oost-Achterhoek) are experiencing the same issues and want to gain insights into the causes of overcrowding and possibilities to predict this overcrowding. Preliminary research has been conducted by Arief Ibrahim, a former master's student (Business Information Technology) at the University of Twente. He created a forecasting model for the patient demand at the ED and the GP-post of Winterswijk using time series analysis with little machine learning applications. The model that he created had quite large errors and it requires further research to improve the model or propose another model to be useful for practical situations.

1.2.2 Problem description

An assignment has been created in collaboration with three partners of AZE in the eastern region of Achterhoek, which is located in the province of Gelderland in the Netherlands. The three collaborating partners are (From now on will be referred to as acute care domain):

- The ambulance service in north and east Gelderland, which is provided by Witte Kruis NOG.
- The ED of the hospital in Winterswijk, Streekziekenhuis Koningin Beatrix (SKB).
- The GP-post of Oost-Achterhoek, which is part of the general practitioners' care of eastern Achterhoek (HZOA).

These partners have indicated that they experience overcrowding within their work field regularly. The problem with overcrowding for these partners is that they do not have a quantifiable measure for overcrowding. There is still ambiguity in the definition of overcrowding. For example, it does not necessarily mean that a large number of patients causes the feeling of overcrowding. It occurs that the same number of patients are treated on two different days, but one day was experienced as extremely busy while the other was pretty calm. Things like the complexity of the required care and available resources also play an important role in the perception of overcrowding. Therefore, the partners want to know how overcrowding best can be measured and predicted such that appropriate actions can be taken beforehand.

1.2.3 Core problem

The main problem that exists is that there is overcrowding in the acute care domain. The effect of this overcrowding is that patients have increased waiting times. Their overall satisfaction decreases and they might suffer severe complications due to the long waiting. On top of that, the medical employees experience psychological as well as physical pressure during their shifts, as they are not able to keep up with the workload.

This perception of crowding for patients and medical employees is caused by the mismatch between the demand for acute care and the available capacity. It could be that there are insufficient available employees to see the patients or that there are no available rooms or resources at a given time. This mismatch can have two reasons; the demand for care is a lot higher than usual and is thus not expected or the overall number of employees/resources is insufficient. Both issues are a topic on its own, but in this research, we will only focus on the first one. The number of personnel and resources is assumed to be sufficient if the demand is known in advance.

The demand for acute care can fluctuate for several reasons. For example, it is expected that peak demand will occur more often during the day than at night. There is a dependency on time. It could also be the case that sudden peaks are caused by external factors such as; flu season, pollen allergies, or big events. Alternatively, it could also be caused by a lack of smooth transition within the facilities themselves. If patients for some reason spend a lot of time at the facilities, then over time the total number of patients will stack up. The sudden peak in demand could of course also just be random and have no particular reason at all. In this research, we will address the problem of patient demand, which is unknown. We want to find a method that can aid the facilities to get an indication of how many patients they can expect on a day.

1.3 Research Design

This section will give an overview of the research design that will aid in solving the core problem. In the first section 1.3.1, the objective of the research will be explained. The research question including its sub-questions will be defined and explained in 1.3.2.

1.3.1 Research objective

The objective of this research is to develop a method that grants insight into the unknown patient demand. The idea is to use predictors to train and validate a machine learning model that is capable of predicting crowding for the facilities in the acute care domain. The time window is set at one day in advance. This allows the facilities to have enough time to still adjust schedules if necessary. The models could be used to function as an introductory step towards an early warning system for crowding of the acute care facilities. Since the facilities in the acute care domain operate separately from each other with different tasks and patients, the goal is to create separate models for the partners.

1.3.2 Research questions

The main research question that will help to solve the core problems and aids in reaching the research objective is the following:

> ***What machine learning model can be used as an adequate early warning system for overcrowding and what is its performance in the acute care domain in the region of Oost-Achterhoek?***

Answering this question will provide us with a tool that encompasses the core problem and helps in predicting the overcrowding within the acute care domain. To answer the main research question we will have to answer several sub-questions. These questions are divided into several components and will be explained shortly below:

1. ***What is the current situation within the acute care domain concerning processes and overcrowding?***
   a) ***How does the acute care system work in the Netherlands?***
   b) ***What are the processes/tasks for the ED, GP-post, and ambulance service?***

The first question will address the current situation of the acute care system in the Netherlands and the processes and tasks that the contributing facilities have to fulfill. This is done to get an understanding of the facilities and the differences between them.

2. ***What are good measures to define overcrowding within the acute care domain?***
   a) ***What does the literature say about measures for overcrowding in the acute care domain?***
   b) ***Which measures are available and should be used to monitor the overcrowding for the facilities in the acute care domain?***

The second question is to acquire knowledge about the measures that are used in similar studies. The facilities currently have no clear definition of overcrowding and we hope to find information from other studies. Secondly, we decide which measures will be used for the facilities in this research depending on the findings in the literature and the available data.

3. ***What are the relevant predictors for overcrowding within the acute care domain?***
   a) ***What does the literature say about predictors for overcrowding in the acute care domain?***

   b) ***Which predictors are available and how should they be used for the prediction of overcrowding for the facilities in the acute care domain?***

The third question is about gaining insights into possible predictors that influence the overcrowding measures, defined in the previous question. The predictors are not limited to the internal data of the facilities. External sources will also be reviewed as potential predictors for overcrowding in the acute care domain.

4. ***What machine learning models are relevant for this research and how can they be evaluated?***
   a) ***What research has been done related to our research?***
   b) ***What machine learning models are there in literature that are commonly used for these kinds of problems?***
   c) ***Which machine learning model(s) is most suitable for this research?***
   d) ***What metrics can be used to evaluate the performance of the machine learning model(s)?***

The fourth question will aid in selecting suitable machine learning models. The first step is to acquire useful information from researches that have already been done. Secondly, we explain the possible machine learning models and identify the pros and cons. Based on these findings we choose a model for our problem. Lastly, we also have to determine how we will measure the performance of the models.

5. ***What is the performance of the proposed machine learning model(s) in the acute care domain?***
   a) ***What steps have to be taken to create the model(s)?***
   b) ***How do we obtain the best model(s)?***
   c) ***What are the relevant features of this model(s)?***
   d) ***What is the performance of this model(s)?***

Lastly, the fifth question will address the final model that we create. The first step is to review which steps have to be taken to model the problem in this way. This includes but is not limited to; the software that will be used, the data preparation that is required to model the situation, and the validation method that should be used. Ultimately, we want to find the best models, list the relevant features, and determine the performance of the best models.

1.4 Methodology

Over the years various methods have been developed to deploy certain types of research. A commonly used methodology for machine learning projects is the **Cr**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (CRISP-DM) (Shearer, 2000). This methodology was introduced in 1996 by Daimler Chrysler, SPSS, and NCR. The structure of this methodology is shown in Figure 2. This methodology guides the researcher from start to end of the project by completing different phases. Although the phases seem to occur iteratively, the whole system is a continuous flow of information and things are adapted as soon new information is known. How the different phases will be used during this research will be explained shortly below the Figures on the next page.

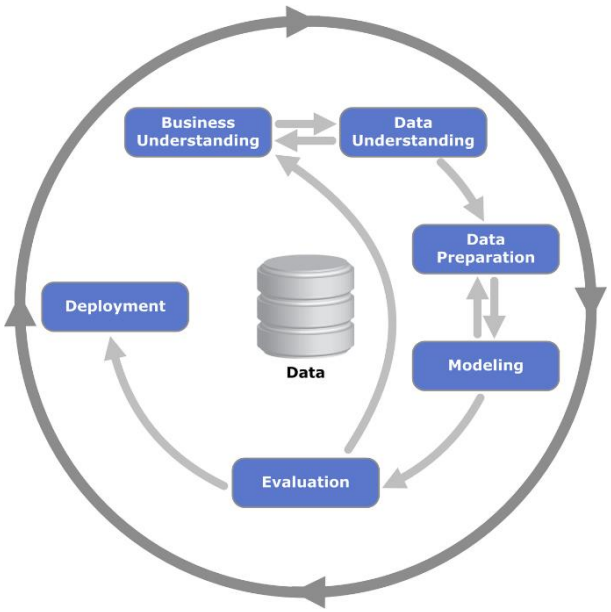

Figure 2 Phases within a project following CRISP-DM.[2]

1.4.1 Business understanding

This phase focuses on understanding the project objectives and requirements from a business perspective. This includes a thorough understanding of the underlying problems and how the current system works. The goal is then to translate this understanding into a data mining problem definition and a plan designed to achieve the objectives. This chapter functions as the first part of that phase, while the second chapter will give more information about the current situation in the acute care domain.

1.4.2 Data understanding and Preparation

In this phase, we get familiar with the data and make adjustments to use them for the modeling part. The data selection will mainly be dictated by literature research and expert opinions. A global description and exploration of the datasets will follow, in which we try to find anomalies or missing data as well as get a basic understanding of the data. We then make a selection of the data that we want to use and make sure that everything is constructed and formatted in the correct datasets suiTable to use for modeling. This phase will be reported in chapter 4 of this thesis.

1.4.3 Modelling

This phase focuses on the modeling of the data and includes the selection of the method based on literature review and available data. The tools to assess the performance of the model and how to validate the models will be explained. In combination with the prior information, we will develop a method to obtain the best models for our situation. This phase will be addressed in chapter 5 of this thesis.

1.4.4 Evaluation

In this phase, we will evaluate the results obtained in the previous phase and make sure that the main findings are correctly documented and presented, such that appropriate conclusions can be

[2] Illustration retrieved from: https://dzone.com/articles/machine-learning-in-a-box-week-2-project-methodolo-1

written. We will also compare our results with the research found during the literature review and discuss the differences. This will be addressed in chapter 6 of this thesis,

1.4.5 Deployment

The last step includes the documentation of the conclusions, limitations, and recommendations based on the findings during our evaluation. The whole process is of course documented in this thesis and will also be orally presented in a final presentation. The tools that have been developed will be accompanied by a user guide for future use.

1.5 Scope

This section will give the boundaries of this research thesis. Since the execution of the master thesis is limited to half a study year (30 EC), it's important to define certain limits to the research.

- Due to the COVID-19 virus pandemic, the acute care facilities work according to a crisis protocol. This means that they are not able to collect and distribute new data for the research. The result is that we only have old previously collected data for the ED and GP-post and unfortunately no data for the ambulance services. As a consequence, the ambulance service will only be included in the description of the acute care network in the Netherlands (Chapter 2).
- The models are limited to the acute care facilities in Oost-Achterhoek, involved with this research. These are the ED of SKB, the GP-post of Winterswijk.
- The models will try to predict the overcrowding and function as one of the beginning steps of an early warning system. We will not address the further allocation of personnel/resources to this demand (capacity planning).

1.6 Deliverables

At the end of the research assignment, the following will be delivered:

- A software application that can be used to create a prediction of the overcrowding in the acute care domain. This application can be one of the early steps of an early warning system in the acute care domain for crowding. In a later stadium, the goal would be that the application could be used by the planners of the ED and the GP-post to match resources to the predicted demand by creating suitable rosters for personnel.
- An instruction manual for the application, written such that workers unfamiliar with programming or data analysis can use it.
- A thesis (and presentation) that contains the decisions on model selection, creation, optimization, and performance. As well as providing recommendations for further research.

1.7 Outline of the Thesis

In Chapter 2 we will introduce the acute care system in The Netherlands and describe the processes of the facilities related to this research. In Chapter 3 we review the literature regarding the measures & predictors for overcrowding in the acute care domain, the related research that has been conducted, and the potential machine learning models. In Chapter 4 we will introduce the datasets and explain how we create the final datasets for the modeling part. In Chapter 5 we will discuss the method to obtain our best models to predict the crowding. The results of our best models will be presented in Chapter 6. Lastly, we will give the conclusions and recommendations in Chapter 7.

2 Acute Care System

In this chapter, we will give an overview of the acute care system in the Netherlands. For the relevant partners that are involved in this research, we will also explain their daily processes and look at some degrees of crowding over the years based on the aggregated numbers in the Netherlands. The goal of this chapter is to get an overall idea of the facilities involved in acute care, the different tasks that they deploy, and some numbers related to crowding in recent years.

2.1 Acute care system: The Netherlands

Acute care in the Netherlands consists of a network of several entities that co-operate to deliver the care that patients require. As the name says, this network aids the patient that requires acute care, which is care that should be treated as soon as possible. Among a few others, we can separate four entities that are involved with acute care. These are the GP / GP-post, the ambulance service, hospitals, and the nursing home (Kremers, Nanayakkara, Levi, Bell, & Haak, 2019). The system is built such that GPs take care of patients with urgent primary care and EDs provide care for patients who urgently need specialized care. Nursing homes are for patients who do not require specialized care but still require admission. These nursing homes may prevent unnecessary ED visits, especially in elderly patients (Kremers, Nanayakkara, Levi, Bell, & Haak, 2019). A general illustration of the network in The Netherlands (although not necessarily completely relevant for our situation) is presented in Figure 3.

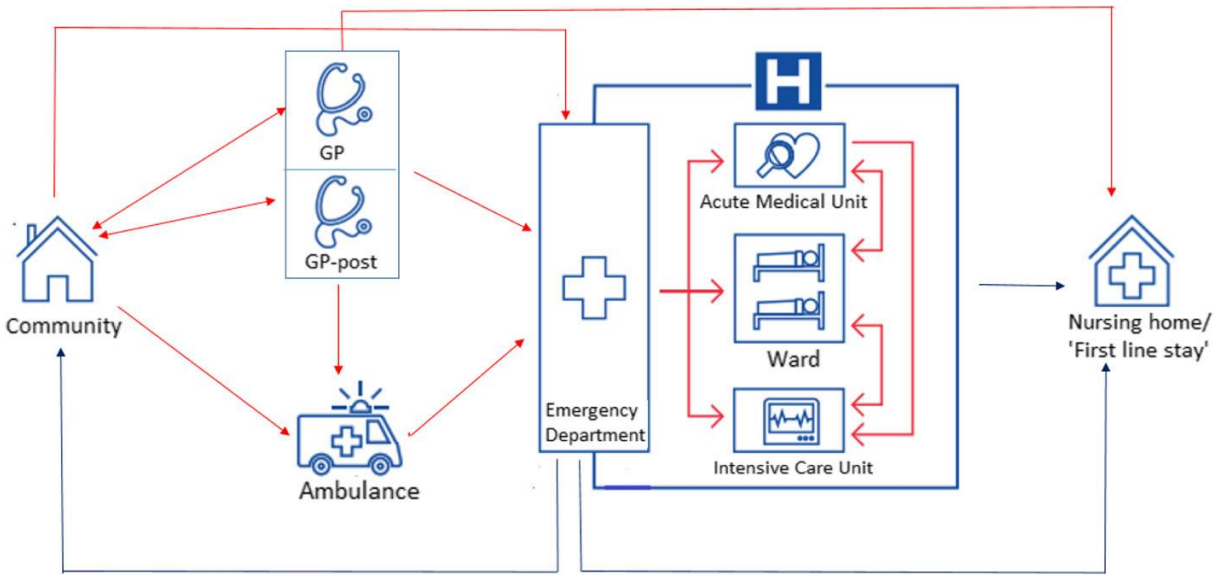

Figure 3 A global overview of the acute care system in the Netherlands.[3]

We will not focus on all the entities named in the above picture. In this research, we will focus on the GP-post, the ambulance service, and the ED. The general practitioner, the different departments in the hospital, and the nursing home are out of the scope. For each of the relevant entities, we will give an explanation of their function and processes in the following sections.

[3] Illustration retrieved from: Strengths and weaknesses of the acute care systems in the United Kingdom and the Netherlands: What can we learn from each other? (Kremers, Nanayakkara, Levi, Bell, & Haak, 2019)

2.2 The GP-post

The GP-post functions as a gatekeeper in the Netherlands. Patients should first contact their usual GP if that's possible. Once the patients' usual GPs are closed for the day (due to closing times) a group of GPs will take over located in a central post. GP-posts operate on; the evening (17:00 - 24:00), night (0:00 – 8:00), weekends, and on national holidays. This system ensures a 24/7 availability for the patient that requires immediate attention. The GP-post is intended for non-life-threatening acute care. They are not meant for care that can be dealt with the next day. People with mild complications should wait for the next opportunity to contact their usual GP. In some hospitals, the GP-post and the ED work in close collaboration with each other. Self-referrals to the ED can then be seen by a GP, which lowers the volume at the ED, but increases it at the GP-post.

Patients are supposed to call the GP-post, where a triage-assistant will indicate the urgency of the required care. This is done according to the Nederlandse Triage Standaard (NTS), which is a method to divide the required care into six distinct groups (U0-U5). U0 is care that requires immediate actions and U5 has the least priority, see Table 3 for a description of the urgencies and Figure 4 for the distribution of urgencies from the years 2012-2018 at the GP-posts in the Netherlands. We can see that the consults with more urgent patients are increasing over the years.

Table 3 NTS urgency levels translated from Dutch[4].

Code	Colour	Title	In words	In time
U0	Red	Resuscitation	Failing vital function	Immediately
U1	Orange	Life-threatening	Instable vital function	As soon as possible
U2	Yellow	Urgent	A threat to a vital function	Within an hour
U3	Green	Fairly urgent	A real risk of damage	Within a couple of hours
U4	Blue	Not urgent	Negligible damage	Within a day
U5	White	Advice	No chance of damage	Next day

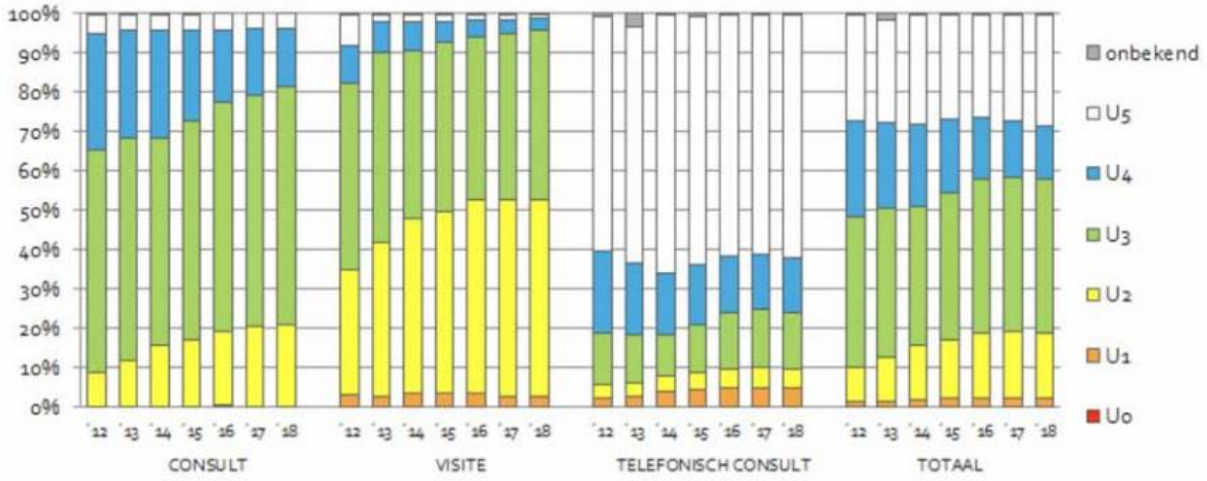

Figure 4 The distribution of urgencies (onbekend = unknown) for consults, visits, phone consults, and the combined total for the years 2012-2018.[5]

There are certain field regulations for the GP-post concerning urgencies. In case of emergency, they should answer the telephone within 30 seconds in 98% of the cases. Another rule states that 90% of the citizens living within the catchment area of the GP-post should be able to reach this post within

[4] Table translated and retrieved from: https://de-nts.nl/nts/basisprincipes-nts/

[5] Illustration retrieved from: Ineen Benchmark Huisartsenpost 2018 https://ineen.nl/wp-content/uploads/2020/02/InEen-Benchmarkbulletin-Huisartsenposten-2018.pdf

30 minutes by car. In case of urgency U0 or U1, the GP should arrive at the patient within 20 minutes in 90% of the cases and 30 minutes in 98% of the cases. In case of urgency U2 they need to arrive within 60 minutes in 90% of the cases and within 120 minutes in 98% of the cases (Nederlandse Zorgautoriteit, 2019).

Once the urgency is determined either a phone consult, a consult, or a visit follows. During a phone consult the instructions will be given through the phone to the patient. In the case of a normal consult, the patient comes to the GP-post where they will be consulted by a GP. Sometimes the patient is not able to visit the GP-post and the GP will visit the patient at their home. It is also possible that the GP-post will advise the patient to see the ED or that they will call an ambulance for the patient.

The total amount of consults in the Netherlands for 2009 – 2018 are illustrated in Table 4. It shows a decrease from 2009-2013 but from 2013-2018 it seems to increase again. The main increase is the number of phone consults which were only 94 in 2013 and 105 in 2018. The consults and visitations remain somewhat stable with small fluctuations.

Table 4 Number of consults at the HAP per 1000 citizens in The Netherlands.[6]

# Consults	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Phone	110	101	102	99	94	91	94	97	99	105
Consults	124	120	123	121	121	123	128	131	126	126
Visits	26	25	24	24	24	23	23	22	21	21
Total	260	245	249	244	239	237	245	251	246	251

In recent years the average time for a phone call (before consult) and consult time have also increased slightly. The phone time has increased to 5 minutes and 56 seconds (5:40 and 5:50 in 2016 and 2017). The average consult time has increased to 14 minutes and 22 seconds (13:49 in 2017). This in combination with an increasing number of total consults, means that the total workload has increased for the GP-posts in The Netherlands (Ineen, 2019).

2.3 The emergency medical service

The emergency medical service is for patients that require immediate care at an incident and transportation (not necessarily both). The ambulance care in the Netherlands is regionally divided into 25 emergency medical services called RAV (in Dutch: regionale ambulancevoorziening). The global process at the RAV for an incident is illustrated on the next page in Figure 5. The citizens of the Netherlands can contact the national dispatch center through the alarm number '112', from there they will be connected to the local dispatch center. Once a call is made a nurse operator will conduct triage to determine the urgency of the situation. This occurs according to the NTS, see Table 3 in the previous section. The triage is then translated to the urgency for the ambulance which is divided into three categories (A1, A2, and B), see Figure 6 for the distribution of the urgencies for the years 2014-2017 in the Netherlands. Medical professionals can also request an ambulance for a patient if they deem this necessary.

- A1 urgency requires an ambulance as fast as possible; a life-threatening situation or permanent disability for the patient can occur. The ambulance uses optical and sound signals on its way to the patient. The response time has to be under 15 minutes for 95% of the cases in the region (Volksgezondheidenzorg, 2020).

[6] Table retrieved from: Benchmark Huisartsenpost 2018 https://ineen.nl/wp-content/uploads/2020/02/InEen-Benchmarkbulletin-Huisartsenposten-2018.pdf

- A2 urgency requires an ambulance as fast as possible as well, there is no life-threatening situation but a fast response is desired. Usually, the ambulance won't use optical or sound signals on its way to the patient. The response time has to be under 30 minutes in 95% of the cases in the region (Volksgezondheidenzorg, 2020).

- B urgency is for the planned ambulance care. Some regions in the Netherlands further divide this to give more clarity. There is no defined response time for this type of urgency (Volksgezondheidenzorg, 2020).

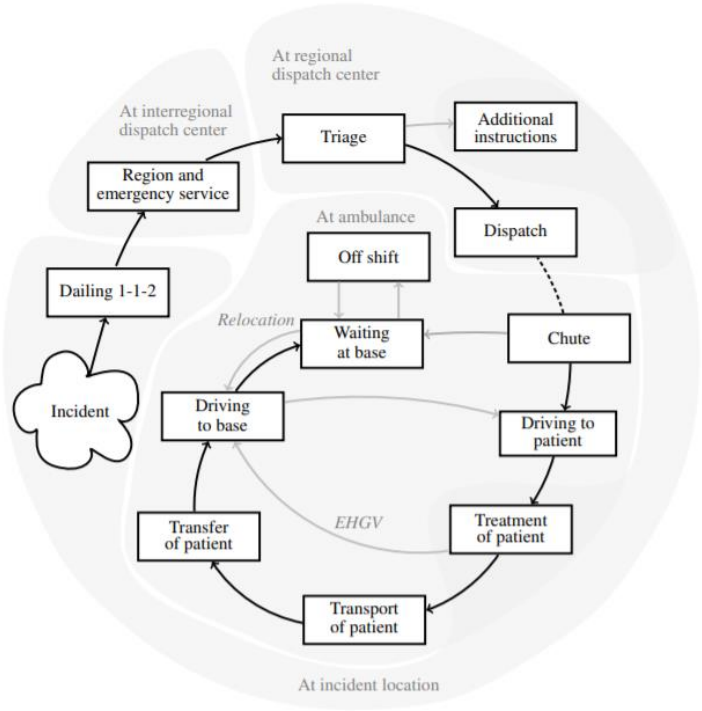

Figure 5 The ambulance service process as it usually takes place at an incident. (EHGV stands for first-aid no transport).[7]

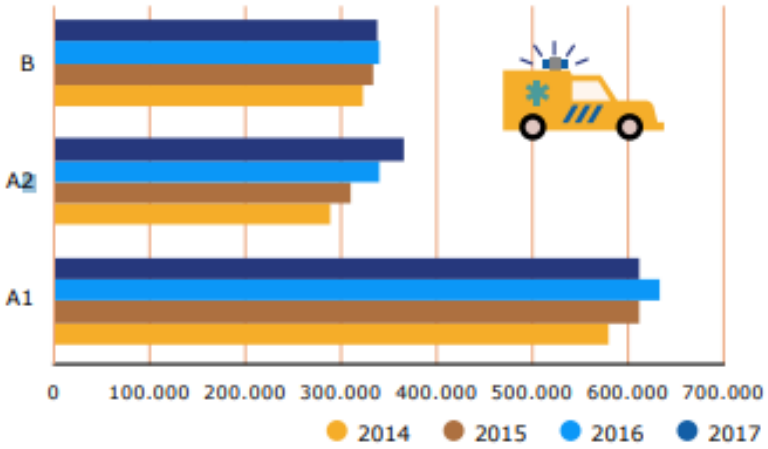

Figure 6 The distribution of A1, A2, and B urgencies for the years 2014-2017 in The Netherlands.[7]

[7] Illustration retrieved from: Monitor acutezorg 2018 https://puc.overheid.nl/nza/doc/PUC_260889_22/1/

Figure 6 shows an increase in the total amount of urgent (A1 + A2) ambulance uses in recent years. However, the average response time does not seem to increase and the exceedance of the A1 norm remains constant (Nederlandse Zorgautoriteit, 2019). Exceedance of the A1 and A2 norm could be caused by various reasons (Volksgezondheidenzorg, 2020);

- insufficient distribution of ambulances
- not enough ambulances available
- force majeure like; weather, closed roads, and untraceable addresses.
- Processes of the dispatch center and ambulance services require time.

When the urgency and location are known, the information will be sent to the local dispatch center. Here an ambulance and personnel will receive a signal to dispatch, for which they will prepare. The time it takes to leave the base since the signal was given for dispatch is called the chute time.

The ambulance then takes an amount of time to drive to the patient. The total time between the first call and the arrival of the patient is called the response time. It is also possible that whilst the ambulance is traveling to the patient that the centralist decides to cancel the trip because it is no longer necessary. This incident is denoted as an interrupted trip.

Once the ambulance arrives at the incident there are a few possible options; there is no patient, the patient requires treatment but no transport or the patient requires transport to a hospital. In case the patient is for whatever reason no longer present, the call will be classified as a loss and the ambulance returns to the base. If the patient only receives treatment but is not transported then it is referred to as an EHGV (Dutch: Eerste hulp geen Vervoer), which translates to first aid no transfer and then returns to the base. When the patient also requires transportation then the call is referred to as declarable. The treatment time of the patient is the difference between the arrival time and the departure time of the ambulance from the incident. Figure 7 shows the different scenarios for A1 and A2 trips in the Netherlands. It shows that most patients are transferred to the ED (or related) or receive first aid and do not require transportation.

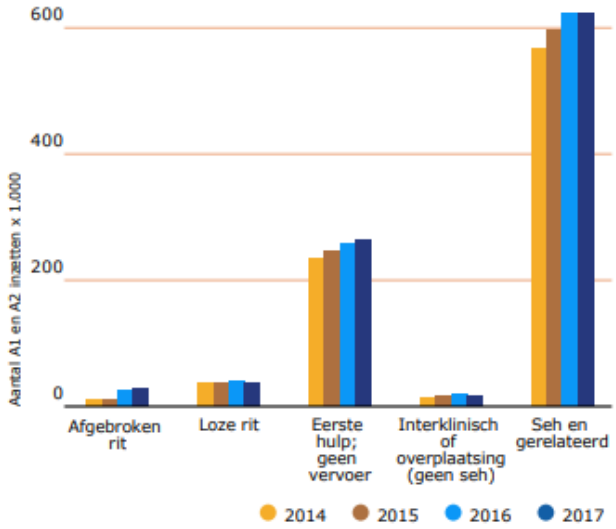

Figure 7 The number of the labeled trips (x1000) for an ambulance, in order: Interrupted trip, loss, EHGV, interclinical or transfer, and ED or related in The Netherlands[8].

The time between leaving the location of the patient and arriving at the destination is called the transportation time. Lastly, the ambulance needs to prepare for new usage. This means that it has to

[8] Illustration retrieved from: Monitor acutezorg 2018 https://puc.overheid.nl/nza/doc/PUC_260889_22/1/

return to its base and have to be cleaned and prepared for a new trip. The time between arriving at the destination and being ready for deployment again is called the release time.

2.4 The emergency department

The emergency department provides specialized acute care to patients that GPs or ambulances cannot provide. A patient is normally referred by a GP/GP-post, an ambulance, or another hospital. However, some patients decide to show up at the ED without any referral, the so-called self-referrals. The patient is always advised to first contact a GP and if necessary they can direct them further to the ED. In most cases, the GP or GP-post can offer the care that they require and an ED visit is unnecessary. The result is that EDs often waste useful resources on these types of patients. Another reason is that it is financially better for the patients. The costs that are associated with an ED visit are covered by healthcare insurance, but they would still have to pay their deductible part. In recent years there is luckily a clear decrease visible in the number of self-referrals to the ED and an increase in the referrals by GP/GP-posts in The Netherlands, as illustrated in Figure 8. However, this is mainly caused by the fact that GP-post and EDs are now working together more. This way the ED automatically sends the self-referrals to the GPs (Nederlandse Zorgautoriteit, 2019).

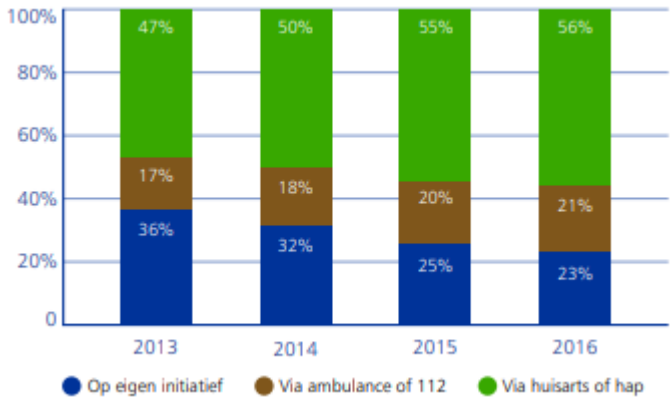

Figure 8 The percentage of patients entering the ED referred by (blue = self, brown = ambulance or 112 and green = GP or GP-post in The Netherlands.[9]

Once the patient arrives at the ED they will be registered and take place in the waiting room. A nurse will be appointed to the patient and they will do the first anamnesis and triage. The nurse will direct the patient to an empty room and will conduct some standard tests and assess the priority of the patient. There are different systems to classify the urgency, as mentioned earlier the GP-post and emergency services use NTS. Some EDs also use this system, but the ED in Winterswijk uses the Manchester triage system (MTS) to determine the priority of the patient. Each priority is associated with a target time in which the patient should be seen by a doctor. Table 5 illustrates the five priorities of the MTS. It shows a similar structure and logic as the NTS.

Table 5 Triage priorities given to patients according to the MTS.

Priority	Colour	Triage category	Target time to be seen by a doctor (min)
1	Red	Immediate	0
2	Orange	Very urgent	10
3	Yellow	Urgent	60
4	Green	Standard	120
5	Blue	Non-urgent	240

[9] Illustration retrieved from: Marktscan acute zorg 2017 https://puc.overheid.nl/nza/doc/PUC_3650_22/1/

Most issues that patients have can be treated immediately. In the EDs in the Netherlands, the patient can often go back to their home after a few hours (almost 2 out of 3 times) (Nederlandse Zorgautoriteit, 2019). They sometimes have to come back at a later time for a check-up but are discharged from the hospital for now. However, a fairly big portion, a little bit more than one third requires admission to the hospital. This is mainly dominated by older people (65+ years old) and young children (0-4 years old). The remainder, a really small portion (±2%) go to any of the other options, for example; intensive care, first-line stay, etc. In Figure 9 we see that there seems to be an increasing trend of patients that require admission after they visit an ED in The Netherlands. These are also the patients that often require the most intense care and contribute to crowding.

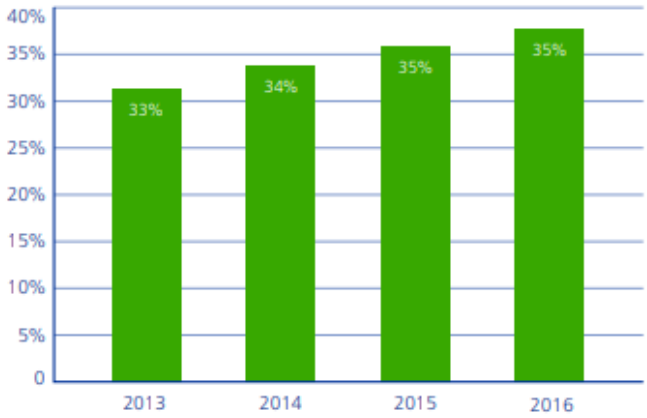

Figure 9 The percentage of ED visits that require clinical admission for the years 2013-2016 in The Netherlands.[10]

[10] Picture retrieved from: marktscan acute zorg 2017 https://puc.overheid.nl/nza/doc/PUC_3650_22/1/

3 Literature review

This chapter will address the literature review that has been conducted. In section 3.1 we will list our findings of crowding measures that are used in the acute care domain. In section 3.2 we will list predictors of crowding that were found in the literature. Section 3.3 will show researches that have been conducted on predicting crowding in the acute care domain. Lastly, section 3.4 will give an overview of machine learning methods and decide which model we will use in this research.

3.1 Measures used for Crowding in the acute care domain

This section will give an overview of the measures of overcrowding found in the literature. The conclusion of this section will determine which measure will be used during this research.

3.2.1 Measures to identify crowding in the acute care domain

All over the world healthcare institutions have researched measures to quantify crowding in the acute care domain. There is no global consensus about a golden rule to measure crowding. Several measures have been proposed and tested in research. This section will give an overview of different techniques to track crowding in the acute care domain.

One metric that is used and is pretty straightforward is the total amount of patients at the facility within a certain frame (Ospina, et al., 2007). A larger amount of patients at the facility creates an increased demand for care and thus resources from the facilities. Taking this a step further and we get the occupancy rate, which is the ratio between the number of patients and the total amount of resources. This ratio can be calculated at any moment and gives an impression of the overall saturation of resources (Ospina, et al., 2007) (van der Linden, et al., 2016). The system is considered to be crowded as soon as the ratio exceeds 1. This means that there are more patients present than there are available resources. This metric is often calculated with the available number of beds in the ED but can be used with other forms of resources as well, for example, available employees. Another measurement that is often used to express crowding is the length of stay (LOS) / total duration, which is the time that a patient spends within the acute care facility. This can be seen as an indirect measure of crowding. When the average time spent on a patient increases it suggests that the outflow of patients is stagnating. This will result in an increase in overall patients as time flows (Ospina, et al., 2007) (van der Linden, et al., 2016). There are also several other time-related variables, such as time from bed request to bed assignment, time from triage to examination by an emergency physician, and time from the bed being ready to transfer to the ward. Where a longer time indicates a more busy ED (Ospina, et al., 2007).

Previously named metrics are based on one metric, but there are also scores developed that are based on multiple criteria. One of these scores for crowding in the ED is the National Emergency Department Overcrowding Score (NEDOCS). This is a score based on a 23-question site-sampling based on input from academic physicians at eight medical schools representative of academic EDs nationwide. Based on these results and the assessment of the charge nurse and ED physicians of the crowding on the ED at randomly selected times, a model was created for predictive purposes. Although the model considering all variables was the most accurate, it was not practical for EDs. Therefore a five-question reduced model was calculated used a backward step-down procedure. The results of a five-question reduced model are valid and accurate in predicting the degree of overcrowding in academic centers (Weiss, et al., 2004). Although there are doubts about whether the NEDOCS measure works outside the USA and whether it's too complex (Boyle, et al., 2015). There are also doubts about whether the NEDOCS tool might be accurate for an extremely high-volume ED setting (Wang, et al., 2014). An alternative version, the mNEDOCS was tested in the Netherlands and

found a strong correlation between the score and perceived crowding by ED staff in both a low volume and high volume ED (van der Linden, et al., 2018).

Similar to the idea behind the NEDOCS there is also the Emergency Department Work Index (EDWIN). Which is calculated by a formula based on five variables, related to the patient and it's urgency and resources of the ED. The formula was tested during a setting over 35 consecutive days at 225-time points in which 2647 patients aged 18 and older were assessed. The measurement of crowding was estimated by the charge attending physician and nurse using a Likert scale. The EDWIN exhibits face and content validity and at one institution was associated with nurse and physician assessment of ED crowding. The score may be programmed into patient tracking software for use as a real-time measurement of ED activity (Bernstein, Verghese, Leung, Lunney, & Perez, 2003).

Lastly, another method of assessing the crowding at the ED is an eight-point measure, called the International Crowding Measure in Emergency Departments. The idea is based on eight rules which can be violated. An increase in violations is associated with increased crowding perception of the personnel. A combination of violations, probably three, predicts clinician concerns better than individual violations. However future work is required to validate this (Boyle, et al., 2015).

The formulas and checklist behind the NEDOCS, EDWIN, and ICMED can be found in Appendix A. All measures mentioned in the text above as well as the literature associated with them are summarized in Table 6.

Table 6 Summary of metrics found in the literature to quantify crowding.

Measure for crowding	**Description**	**Literature**
Total number of patients	The total number of patients that are present at the facility	(Ospina, et al., 2007)
Occupancy ratio	The ratio between the number of patients and the available resources	(Ospina, et al., 2007) and (van der Linden, et al., 2016)
Length of stay	The total length of stay of a patient at the facility	(Ospina, et al., 2007) and (van der Linden, et al., 2016)
Time-related variables	Different times are associated with how long the patient must wait for transfers or consults.	(Ospina, et al., 2007)
NEDOCS	A score based on five variables to score the crowding of an ED	(Weiss, et al., 2004), (Boyle, et al., 2015), (Wang, et al., 2014) and (van der Linden, et al., 2018)
EDWIN	A tool that calculates crowding based on five variables	(Bernstein, Verghese, Leung, Lunney, & Perez, 2003)
ICMED	An eight-point evaluation system for crowding	(Boyle, et al., 2015)

3.1.2 Conclusion on measures to identify crowding in the acute care domain

In the previous section, we listed several measures commonly used to measure crowding in the acute care domain. It must be noted that these are all based on research done in EDs. The reason being that the GP-post as it is used in the Netherlands is quite a unique concept. The result is that we were not able to find any research in which they propose measures for crowding at the GP-post. However, some of the metrics for EDs can of course directly be used for GP-post as well, such as the total number of patients, occupancy ratio, and certain times between actions. The NEDOCS, EDWIN, and

ICMED would be more difficult as the GP-post lacks certain criteria that are required to calculate these measures.

Unfortunately, we are also not able to use most of the measures that we've found in the literature for our situation. The reason being that we have to use an old dataset in which certain variables are not collected. Since we cannot obtain new data from the facilities due to the reasons explained earlier (Covid-19). For that reason, we've decided to use a variation of the first metric in Table 6. We will look at the total number of patients that visit the GP and ED on a given day as a metric for crowding. Using this metric we will be able to create a one-day-ahead forecast for the total number of visitors to the ED and GP-post. Predictions on the same day for a certain time slot is not possible with our datasets, since the arrival times are not always (correctly) registered. Additionally, a prediction that spans more days for example one-week-ahead would be less effective. Since it would be hard to determine on which days of that particular week more patients are expected since the values would be aggregated. Predictions one-day-ahead would still allow planners to schedule rosters of employees to the predicted demand, whilst keeping the uncertainty of when the peak demand will occur relatively low.

3.2 Predictors of crowding in acute care domain

This section will give an overview of predictors of overcrowding found in the literature. The conclusion of this section will determine which predictors will be used during this research.

3.2.1 Predictors for crowding in acute care domain

In section 3.1 we found different measures that quantify the crowding in the acute care domain. In this section, we want to find predictors for these measures. Knowing these can help the facilities to identify and notice crowding such that hopefully, they can react before it occurs. A summary of the predictors can be found in Table 7.

The ED is open the entire day but studies show that there are time-related variables that influence the daily volume and LOS. Daily demand for ED services is characterized by seasonal and weekly patterns (Calegari, et al., 2016). Similarly, an increase of LOS has also been associated with days of the week, months of the year, and even the time of arrival, where mornings seemed to be the most noTable (Hofer & Saurenmann, 2017) (Weiss, Rogers, Maas, Ernst, & Todd, 2014).

We also found that there are weather-related variables that influence the daily volume of ED visitors. A study found that temperature may be a sensitive marker for total ED patient volume (Tai, Lee, Shih, & Chen, 2007). Another study found that some climatic factors displayed a significant correlation with demand series but did not increase the accuracy of prediction when incorporated in the model (Calegari, et al., 2016).

There are also characteristics associated with the patient that appears to influence crowding. The urgency given to a patient (triage level 5 most severe) is associated with increased LOS (Hofer & Saurenmann, 2017). Similarly, patients that are referred by general practitioners often require more intense care and have a serious impact on crowding (van der Linden C. , et al., 2013) (Hofer & Saurenmann, 2017). Patients with multiple trauma (comorbidity) also have an impact on crowding (van der Linden C. , et al., 2013). A study in the Netherlands found that crowding at the GP-post occurs due to parents with young children with non-urgent problems (Keizer, et al., 2018).

One less expected predictor for crowding is using the road traffic flow data. A study has shown that road traffic as an external overall covariate can contribute to an improvement in forecasting crowding in emergency departments (Rauch, Hübner, Denter, & Babitsch, 2019).

Table 7 Summary of metrics found in the literature to quantify crowding.

Predictor of crowding	Description	Literature
Time-related variables	Variables related to the month, day, and time of the day	(Hofer & Saurenmann, 2017), (Calegari, et al., 2016) and (Weiss, Rogers, Maas, Ernst, & Todd, 2014)
Weather-related variables	Variables related to weather conditions such as; temperature, rain fail, etc.	(Tai, Lee, Shih, & Chen, 2007) and (Calegari, et al., 2016)
Patient-related variables	The urgency of the required care, referral by the general practitioner, comorbidity, and age	(Hofer & Saurenmann, 2017), (van der Linden C. , et al., 2013) and (Keizer, et al., 2018)
Road traffic flow	The activity on the roads measured as a flow	(Rauch, Hübner, Denter, & Babitsch, 2019)

3.2.2 Conclusion on predictors for crowding in acute care domain

In the previous section, we've listed variables that have been associated with crowding in the acute care domain. Similar to the measures of crowding, most of the research was related to EDs, although we also managed to find one paper that addressed young children as a factor for crowding in GP-posts. However, we once again think that the most relevant predictors for the ED will also work for the GP-post, for that reason we will not make a difference in the selection of predictors for the facilities.

We will try to use all the variables listed in Table 7 as predictors for our models, except for the patient-related variables. Some of the variables are not or only partially collected, such as the comorbidity and age. However, since we use aggregated daily volumes as a measure for crowding, it would also not make much sense to include the details of the patients that visited the facilities a day prior (which are used to predict for today). We also include events (festivals, gatherings, etc.) and pollen data in our predictors. We did not find any literature about these predictors, but these were suggested as potential predictors by experts in the acute care domain.

3.3 Literature related to the research goal

In the previous sections, we've found metrics and predictors for crowding and we've chosen the daily volume of patients as a measure for crowding. In this section, we will try to list research in which they have tried to predict the daily volumes of patients (or similar) in the acute care domain. We will list which techniques were applied and what kind of results were found.

3.3.1 Literature on related research

There have been multiple attempts at predicting crowding in the acute care domain. (Calegari, et al., 2016) tried to forecast the daily volume and acuity of patients in the emergency department. They used exponential smoothing (ES), multiplicative holt-winters (SMHW), seasonal autoregressive integrated moving average (SARIMA), and multivariate autoregressive integrated moving average (MSARIMA). When all types of patients were jointly considered, the ES performed best. The SARIMA

performed better for the very urgent and urgent patient. The MSARIMA did not improve the performance over SARIMA.

(Jones, et al., 2008) also looked at forecasting the daily patient volumes in the emergency department. In this study, they used linear regression (as a benchmark), SARIMA, ES, Time series regression, and an Artificial neural network (ANN). They found that all the methods performed better than the benchmark, but the gain in accuracy was very small even for the best model (Time Series regression).

(Marcilio, Hajat, & Gouveia, 2013) looked at forecasting daily emergency department visits using calendar variables and ambient temperature readings. They tested three different methods; generalized linear models (GLM), generalized estimating equations (GEE), and SARIMA. All models were built with and without the effect of temperature. The GLM and GEE models showed the best performance and the inclusion of temperature did not improve the forecasting accuracy

(Zlotnik, Gallardo-Antolín, Alfaro, Pérez Pérez, & Martínez, 2015) attempted to forecast emergency department visits and dynamic nursing staff allocation using machine learning techniques. For the visits, they used two different regression models in a free software tool (WEKA). They used support vector regression (SVR) and M5P tree and found that the performance of both models was superior to the stratified average model with a 95% confidence interval.

(Volmer, et al., 2020) applied machine learning techniques and time series algorithms to forecast the demand at emergency departments. In this paper, they compared traditional time series algorithms like ES and ARIMA with machine learning algorithms such as; GLM, random forests (RF), gradient boosting machines (GBM), and k-nearest neighbors (k-NN). They found that the performance of both methods was more or less equal but the predictions were more diverse so that staked predictions are more robust and accurate.

(Khaldi, El Afia, & Chiheb, 2019) forecasted weekly patient visits to the emergency department. Here they look at ANN combined with a signal decomposition technique, Ensemble Empirical Mode Decomposition (EEMD). They benchmark this versus a normal ANN, an ANN with a discrete wavelet transform (DWT), and an ARIMA model. They found that the ANN-EEMD outperforms the benchmarking models for approximation and generalization capabilities, thus the model can be employed to forecast efficiently ED arrivals.

(Nas & Koyuncu, 2019) used ten different machine learning algorithms to predict the ED arrivals, which they then use for a simulation study. They found that the use of a long short-term memory (LSTM) model performed the best out of all models, followed by the RF and decision tree (DT) models.

Last but not least (Ibrahim, 2019) did a study for the same acute care facilities as we investigate in our research. In this study, they tried to predict the daily volume of patients at a GP-post and ED with SARIMAX (a SARIMA model that can incorporate exogenous data) and a hybrid model using SARIMAX and gradient tree boosting. They found that the hybrid model gave the best performance

3.3.2 Conclusion on related research

Quite some literature was available about forecasting daily visitors to the ED. Most of the research was done using time series models such as SARIMA or variations of that. The performance of these models was often good. Some research also addressed different machine learnings techniques. These methods had similar performances and in some cases added additional benefits. Prior research has been conducted at the ED of SKB and the GP-post of HZOA in which SARIMAX models have been

applied. For that reason, we are not going to approach it in the same way again. We are more interested in the machine learning approach. The literature that we found mainly used time-related variables and weather variables as predictors for the machine learning prediction models. In our research, we want to expand on this literature by incorporating pollen data, events data, national holidays, and the visitors of the GP and ED respectively in our prediction models. We hope that the addition of these predictors will have a positive effect on the performance of the models.

3.4 Machine learning models

Machine learning models are a relatively new set of tools that utilize data for practical problems. The idea behind machine learning is that a computer creates its program using existing data. The program can then be used with new data for; classification, regression, clustering, etc. This is different in comparison with the traditional way of programming. Traditionally a program is created by the data scientist and used with the input data to say something about the output (Bishop, 2009). A comparison between traditional programming and machine learning is illustrated in Figure 10.

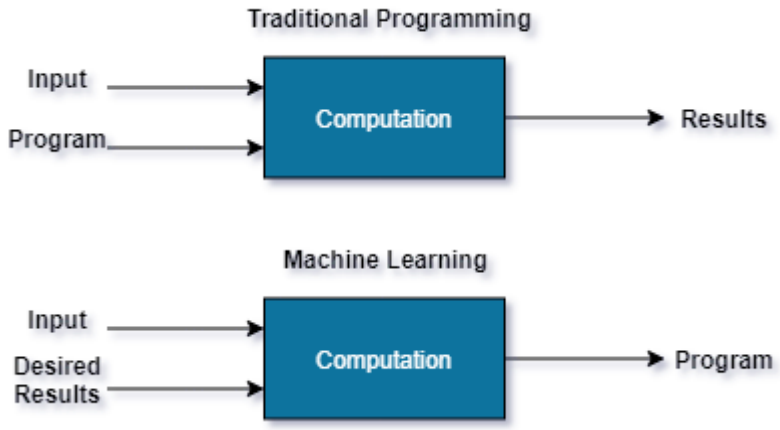

Figure 10 Traditional programming versus Machine learning.

As specified in the previous sections we want to predict daily ED and GP-post visitors, based on certain predictors. Since we want to link a numerical dependant variable to independent variables, we require regression methods. Machine learning algorithms that can deal with regression are a subset of the supervised methods. Supervised machine learning means that the computer is fed input data and the corresponding output data. The algorithm then learns to map the inputs to the output. The learning phase of the algorithm is also called model training and is done by a training set (a subset of the total dataset). Once the model is trained an additional test set (unseen data which was not used for training) can be used to determine the performance of the algorithm. This does however mean that the quality of the model depends on which data is chosen as a training and test set (Bishop, 2009).

3.4.1 Machine learning models for regression

In this section, we will address different machine learning algorithms that deal with regression and which we have also seen back in the literature about demand forecasting in acute care (section 3.3). A short description of the method and the characteristics will be given for each algorithm. The algorithms that will be considered are the following; artificial neural networks, decision trees (random forest and gradient boosting as well), and support vector machines.

3.4.1.1 Artificial Neural Networks

Artificial neural networks (ANN) is a method that is inspired by biological neural networks that constitute animal brains. An ANN consist of several nodes and connections between the input data and the output (Bishop, 2009). These nodes are divided into distinct layers; the input layer, hidden layer, and output layer. The first type of layer is the input layer and contains the data that will be used for training. Every data category that will be considered is a separate node in the first layer. The second layer type is the hidden layer. A simple model can have one hidden layer but more advanced models can have several hidden layers next to each other. The last layer is the output layer and contains the associated output for a given input determined by the model. A simple illustration is presented in Figure 11.

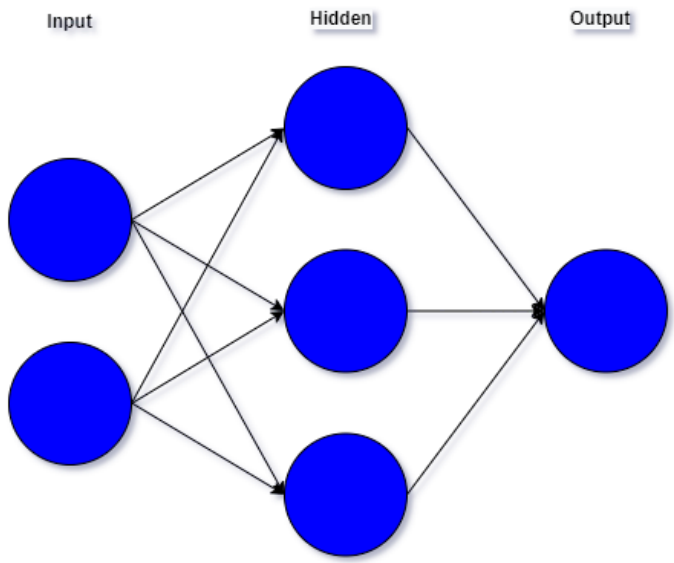

Figure 11 Schematic overview of a simple 2-3-1 ANN. We have 2 input variables, 3 nodes in the hidden layer, and 1 output value.

All nodes that are connected have a certain weight that determines how much the value of a preceding node will influence the receiving node. A receiving node is a combination of all outputs of the preceding nodes, as we can see in Figure 11. The sum is taken over the combination of the output value times the weight. This value is then often transformed with the use of an activation function, such as; sigmoid, hyperbolic tangent, etc (Karlik & Olgac, 2010). This forward propagation continues until the output nodes have a value. The network is usually initialized with random weights between the nodes. The model is then trained with a training set to adjust the weights in such a way that the model finds satisfying output results. The performance of the output value is monitored by a so-called loss function. This function determines the difference between the output value of the model and the actual output value. A commonly used loss function is the residual sum of squares for regression problems. The error is then backpropagated through the network to find the new weights of the nodes using gradient descent. This process is repeated until the model has performed a certain number of iterations or the error has decreased under a certain threshold value.

The advantage of ANNs is that the structure of the design allows a big variety of possibilities to model certain situations. The model contains lots of parameters that can be twisted to form the model to your design, such as; the use of different (nonlinear) activation functions, loss function, number of hidden layers, and number of nodes within a layer. This allows the model to often produce very good results if sufficient data is available. The ANN also has some drawbacks in comparison with other algorithms. The flexibility of the algorithm has a downside. By utilizing different layers and functions

within the model it will be really hard to grasp what is happening within the model. The model more or less becomes a black box where data is inserted and an output value comes out. The complex structure of the model can also cause computation times to increase fast for larger problems. Another disadvantage is that the algorithm often requires a large amount of data to give good results. Lastly, the algorithm uses numerical values as input and outputs within the nodes. This means that categorical variables require some sort of transformation.

3.4.1.2 Decision Trees

A decision tree (DT) algorithm is essentially a really simple algorithm. The algorithm constructs a tree with branches based on the most important features (Kingsford & Salzberg, 2008). When you follow the structure of the tree you will end up in a leaf node where a prediction is made based on the mean values within the node (in case of regression). A DT is built sequentially. Each split considers all available features and branches further on the feature with the most information gain. The Gini index and the entropy are often used to measure the information gain of a certain split (Kingsford & Salzberg, 2008). Both measures indicate the distortion within a node, a Gini index or entropy value of 0 is therefore preferred. The information gain is determined by the amount of distortion before and after a split. The split that decreases the distortion the most has the most information gain. A simple example of a decision tree is illustrated in Figure 12.

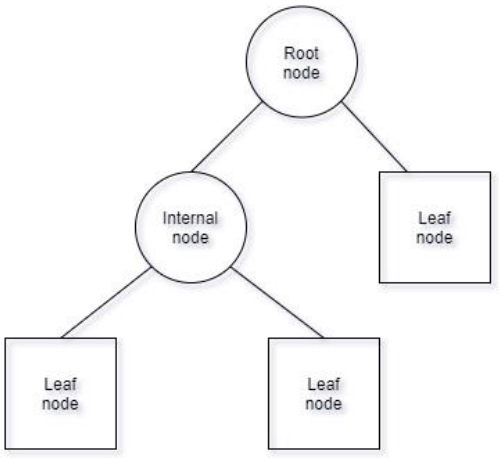

Figure 12 Simple illustration of a decision tree We start at the top node and branch until we reach a leaf node.

A big advantage of this model is that it has high interpretability for the user. The important features are the ones with the most information gain and the overall method is easy to grasp and visualize. The model can also use categorical data without any transformation, as the variables can simply be branched on. A disadvantage is however the vulnerability to overfitting (small changes in the data can result in other trees), techniques like early stopping or pruning of nodes are implemented to prevent this from happening. The algorithm often performs less than other algorithms because it is limited by its simplicity (Hastie, Tibshirani, & Friedman, 2017). However, extensions have been created for the DT algorithm. The algorithm is used in so-called ensemble learning. This method combines several decision trees to produce a better prediction than utilizing a single decision tree. These extensions are called random forest and gradient boosting.

Random forest

A random forest (RF) is an assemble of many decision trees. This means that the algorithm does not train one decision tree, but multiple trees as in a forest. The different trees are built by selecting random samples and features for every tree in the forest (Breiman, 2001). When we are working

with regression models, we simply take the average over all the different trees in the forest for a certain input to obtain an estimate. A major advantage is that the performance of the model increases significantly in comparison with a single decision tree. Additionally, the algorithm is less likely to overfit on the data, since estimates are now averaged over a multitude of trees. The algorithm also requires no additional transformation of categorical variables and can deal with missing data. The algorithm is also quite flexible with hyperparameter tuning. Lastly, the model has a built-in validation method that can be used, the so-called Out-Of-Bag (OOB) error. This is determined by the samples that were not chosen during the random selection. A disadvantage in comparison with the traditional DT is that the interpretability is more or less lost. When the number of trees in the forest is large it's hard to keep track of all the different variants. However, the basic idea behind the model is still easy to explain. Despite this loss in interpretability, the algorithm can still give the feature importances, which also grants some explainability to the users.

Gradient boosting

Gradient Boosting (GB) is also an ensemble of decision trees, which combines weak learners to create a strong learner (Friedman, 2002). However different from RFs, in GB trees are not built in parallel, but sequentially. Trees are added one at a time and existing trees in the model are not changed. Every new tree is trained on the residuals of the previous tree which are defined by a loss function. During every new step, the derivative of the loss function is used to adjust the weights to find a way that minimizes the total error (gradient descent approach) (Bishop, 2009). This method of sequentially adding parts to the solution while lowering the error by adjusting the weights is called boosting. An advantage of the gradient boosting method is that the prediction accuracy is often very high. The algorithm is quite flexible with hyperparameter tuning. Similar to the RF and DT the categorical variables do not need to be transformed to use this algorithm and it can deal with missing data. Some disadvantages are that the algorithm tries to minimize all errors which makes it prone to overfitting. Parameter tuning must be done correctly to prevent this. The algorithm is also computationally expensive, as it often requires many trees which can be time and memory exhaustive. The ensemble of smaller models makes it hard to interpret the results, similar to RF models. However, tools to determine the variable importances are also available.

3.4.1.3 Support vector machines

The support vector machine is an algorithm that is developed intentionally for classification problems. Later on, it was extended to also deal with regression problems, this extension is named support vector regression (SVR) (Drucker, Burges, Kaufman, Smola, & Vapnik, 1997). In SVM a hyperplane is constructed such that the data is separated linearly into distinct classes. A hyperplane is a subspace whose dimension is one less than its ambient space. The hyperplane is linear but the training samples are often not linear. However, by utilizing the so-called kernel trick, it is possible to map nonlinear data points in its current dimension to points that are linearly separable in higher dimensions. Graphical visualization of mapping is presented with an example in Figure 13.

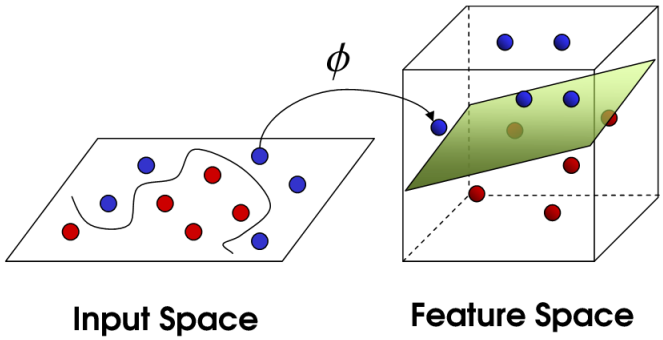

Figure 13 An example of data that is inseparable in the input space, but once it is mapped in a higher dimension a hyperplane can linearly separate the points.[11]

The algorithm tries to maximize the boundaries of the hyperplane such that the closest pair of data points belonging to opposite classes are still separated. These points are called the support vectors as they determine how the margins are set. During training, the algorithm tries to find the maximal margin hyperplane that optimally separates the different classes. In SVR the method is similar but the hyperplane is not used to separate the classes but to estimate the continuous prediction (Bishop, 2009). It is often not possible to find a perfect hyperplane and that is why some room for error is allowed and so-called soft margins are used. One of the main advantages of SVR is that its computational complexity does not depend on the dimensionality of the input space. Additionally, it has great generalization capability, with high prediction accuracy. A disadvantage is that the tuning of the kernel function and its parameters is quite hard. Additionally, the model requires the computation of the input data, which means that categorical variables must be transformed before they can be used. Lastly, the kernel function may map the data in dimensions greater than 3 which makes it hard to visualize what's happening.

3.4.2 Conclusion on the machine learning models

In the previous section, we looked at a few different machine learning techniques that are used for regression and which have also been used for predicting daily visitors in the acute care domain (section 3.3). For our research, we require a method that is easily tunable and has good performance. For that reason, decision trees on themselves are less relevant, as they perform less than the other methods. It is also important that we can derive the important variables within the model, such that we can translate these as predictors of crowding to the partners. The RF and gradient boosting methods can quite easily give us feature importances with the use of the information gain. It is also possible to determine the feature importances with ANN and SVM, but these require more work and additional techniques. A plus for the ensemble methods is also that they can deal with categorical variables as well. The other algorithms can deal with them but require some sort of transformation to numerical variables. Another benefit of the RF model is that it has a built-in validation method with the out-of-bag predictions that it can make. Lastly, the RF model is also based on a quite simple expansion of decision trees. This keeps the explainability behind the algorithm quite high, although the model itself might become some sort of black box. Based on the overall positive sides of the RF algorithm, we've decided to use that algorithm for our models.

[11] Illustration obtained from: https://towardsdatascience.com/the-kernel-trick-c98cdbcaeb3f

4 Data Understanding and Preparation

This chapter will introduce the internal and external datasets that will be used as input for the machine learning algorithm. The chapter will focus on understanding the basics of the datasets and explain which alterations must be made to construct the final dataset.

4.1 Data collection and description

The first step is to select different sources of data that could be of use for the final model. We will introduce the acquired datasets. These sets have been selected based on literature and expert opinions in the work field. We will also give a short explanation about sources that have been looked into but could unfortunately not be acquired during this research. Lastly, we will give a small overview of all the available datasets and their periods.

4.1.1 Data from the ED of SKB and the GP-post HZOA

The ED of SKB has provided a dataset that contains data collected from 2012-2018. The dataset contains the information of 85048 individual patients that have visited the ED in this period. This dataset will only be used to determine the daily visitors to the ED. The GP-post HZOA has provided a dataset that contains data collected from 2013-2017. The dataset contains the information of 149725 individual consults (telephone, home visit, or consult at the post) given in this period. Similarly, as for the ED, the dataset will only be used to obtain the daily visitors to the GP-post.

4.1.2 Time-related data

This is a dataset that contains information about time-related events. As literature has shown that these have a good influence on daily visitor prediction in the acute care domain. On top of this, we also include yearly recurring holidays in the Netherlands and some in Germany. The holidays in the Netherlands are relevant because the GP-post will be open 24 hours on weekdays contrary to normal weekdays if a holiday falls on them. We also include the regional holidays from Germany's Lower Saxony and North Rhine-Westphalia. These are the regions that are close to Winterswijk. Expert opinions told us that they experience an increase in German patients that visited the Netherlands during these holidays. Table 8-9 show the time-related variables and holidays that we consider.

Table 8 A description of the time-related data.

Variable name	**Variable definition**	**Variable type**
Day	The day of the week	Categorical (day)
Month	The month of the year	Categorical (month)
Holiday	Specifies whether the day is a national holiday in The Netherlands	Categorical (yes/no)
GermanHoliday	Specifies whether the day is a national holiday in Germany	Categorical (yes/no)
IsWeekend	Specifies whether the day falls on the weekend	Categorical (yes/no)

Table 9 List of the national holidays and German specific holidays we take into account.

National holidays		**German specific holidays**	
New Year	Ascension day	Labour day	All Saint's day
Easter	Pentecost	German Unity day	Corpus Christi
Queensday / Kingsday	Christmas	Reformation day	

4.1.3 Weather data

The first external dataset is a text file containing different weather-related measures. This dataset is obtained from The Royal Netherlands Meteorological Institute (Koninklijk Nederlands Meteorlogisch Instituut, KNMI). They have several weather stations distributed over the country and on the sea that accumulate daily values for a variety of weather-related measures. The station in Hupsel has been chosen as a reference for the weather conditions that apply to our region as this is the closest station.

Figure 14 KNMI weather stations distributed over The Netherlands. The Arrow points at the chosen station and the red cross is the location of Winterswijk[12].

The location of this weather station can be found in Figure 14. The dataset is freely available on the website of the KNMI and can easily be downloaded from their website for a specified period. The dataset that we have acquired contains data from the period 01-01-2012 – 19-12-2018. This is equal to the period for which we have data about the ED. The total dataset contains 2545 days of measurements for 29 (numerical) measures. Table 10 shows the description of the various measurements collected in that dataset.

Table 10 Measures in the KNMI dataset from Hupsel.

Weather code:	Description	Weather code:	Description
DDVEC	Vector average air direction (degrees)	T10NH	Hour when T10N was measured
FHVEC	Vector average air speed (0.1 m/s)	SQ	Duration of sunshine (0.1 hours)
FG	Daily average wind speed (0.1 m/s)	SP	Percentage longest possible sunshine

[12] Illustration retrieved from: https://www.meteolimburg.nl/knmi-weerstation-arcen-blijft-bestaan

FHX	Highest hourly average wind speed (0.1 m/s)	Q	Global radiation (J/cm2)
FHXH	Hour when FHX was measured	DR	Duration of rain (0.1 hours)
FHN	Lowest hourly average wind speed (0.1 m/s)	RH	The total amount of rain in the day (0.1 mm)
FHNH	The hour when FHN was measured	RHX	Highest hourly total rain (0.1 mm)
FXX	Highest wind gust (0.1 m/s)	RHXH	Hour when RHX was measured
FXXH	The hour when FXX was measured	UG	Daily average relative humidity (percentage)
TG	Average daily temperature (0.1 degrees Celcius)	UX	Maximum relative humidity (percentage)
TN	Minimum temperature (0.1 degrees Celsius)	UXH	Hour when UX was measured
TNH	The hour when TN was measured	UN	Minimum relative humidity (percentage)
TX	Maximum temperature (0.1 degrees Celsius)	UNH	Hour when UN is measured
TXH	Hour when TX was measured	EV24	Reference crop evaporation
T10N	Minimum temperature 10 cm above the ground (0.1 degrees Celsius)		

4.1.4 Pollen data

The second external dataset is about the amount of pollen in the air. These values are collected by the Elkerliek hospital located in Helmond, who has been doing this for over 25 years. They use a measurement system placed on top of the roof of their facility to collect this data. The website of Elkerliek contains the daily pollen values measured daily since 2012. An Excel file was obtained from them containing all the data from 2012 up to the point that the dataset was requested, mid-2019. The data after 19-12-2018 will not be used. The file contains the daily measurements of 44 different pollen types. They've also specified which pollen is non-, mildly-, strongly- or very strongly allergenic to people. Tables 11-12 show the allergenic and non-allergenic pollen that are listed in the dataset.

Table 11 Allergenic pollen collected in the dataset.

Mildly allergenic	**Strongly allergenic**	**Very strongly allergenic**
Corylus	Betula	Poaceae
Alnus	Artemisia	Ambrosia
Rumex	Cladosporium	Alternaria
Plantago		
Cedrus libani		

Table 12 Non-allergenic pollen collected in the dataset.

Non-allergenic		
Cupressaceae	Acer	Rosaceae
Ulmus	Platanus	Asteraceae
Populus	Pinus	Ranunculaceae
Fraxinus	Ilex	Apiaceae
Salix	Sambucus	Brassicaceae
Carpinus	Castanea	Urtica

Hippophae	Tilia	Chenopodiaceae
Fagus	Ligustrum	Fabaceae
Quercus	Juncaceae	Humulus
Aesculus	Cyperaceae	Filipendula
Juglans	Ericaceae	Indet

4.1.5 Events data

The third external dataset is an excel file containing events that occurred in the region of north and eastern Gelderland provided by the safety region north and eastern Gelderland (Veiligheidsregio Noord-en Oost- Gelderland, VNOG). This file contains information about events that occurred from 2016 to 2019. For every event, five variables are collected. Table 13 shows the variables that are collected with their definitions.

Table 13 Variables that are filled in for every event in the dataset.

Variable name	**Variable definition**	**Variable type**
Date start of Event	The date (+ sometimes time) when the event starts	Date (day/month/year) + time
Description Event	A short description of the event	Text description
Municipality	The municipality where the event took place	Categorical (municipality names)
Visitors	The number of visitors (if known)	Numerical if known otherwise 'Onbekend'
Risk classification	The risk classification associated with the event	Categorical (different options for risks)

The risk classification is a measure given to registered events, to indicate the level of safety measures that are required. If an event does not need to be registered it will suffice to simply report that there is an event (Kennisgeving, in the file). All other events are classified in either class: A, B, or C

- Class A events are regular events with no special risks. They require little or no extra cooperation from police, fire brigade, and medical services. For example a block party or small gathering.
- Class B events are events that require extra attention, these events often require advice from police, fire brigade, and medical services. For example parties in the city or sports events.
- Class C events are the events with the most associated risks and always require advice from police, fire brigade, and medical services to ensure safety during the events. For example music festivals or big gatherings.

Several factors influence the risk that is associated with an event. For example, the number of visitors is an indicator of the risk, but also the target audience (young people with alcohol/drugs) and location (difficult to reach or not), and various more factors. It's up to the municipality to decide which classification a certain event will obtain.

4.1.6 Flu season data

Another dataset that we wanted to acquire was about the flu season in the Netherlands. Nivel is an institution for research in healthcare and they collect and distribute information on different diseases in the Netherlands. They do this to create an active picture of the current cases and distribution of healthcare problems. They report their results weekly on their website. The flu is one of the healthcare problems that they provide information on. Nivel mentions on their website that

they provide data for research and that they can be contacted about possibilities[13]. After contacting them we had to, unfortunately, exclude this option. The reason being that the datasets were only buyable per year. The total costs to obtain the data over the period that we are interested in would cost several thousand euros, which are funds that we do not have available for this research project.

4.1.7 Traffic intensity data

Lastly, a dataset that we wanted to acquire was about the traffic intensity on big roads close to the region of interest. The national database traffic information (Nationale Databank Wegverkeersgegevens, NDW) collects and distributes data about traffic intensity and speed at certain roads in the Netherlands.

Figure 15 The sections of interest are denoted by the pink spots within the black circle.

Unfortunately, after further investigation, we found that there was only limited data available for the roads of our interest (A18/N18), visible in Figure 15. The available data contained the daily average over 2018, which was only one measure. For all other years, there was no data available. Unfortunately, this means that the data is not useful for our applications.

4.1.8 Graphical overview of datasets

In sections 4.1.1-7, we introduced the different datasets that are available during this research and the ones that were unfortunately not. In Figure 16 we've displayed all the different datasets that we will use and the period over which they have data available. The leading datasets are these of the ED and the GP-post since they contain the values that we want to predict. The GP-post is the smallest dataset of the two and thus is the leading set for the period over which the models could be made.

[13] https://www.nivel.nl/nl/nivel-zorgregistraties-eerste-lijn/griep-centraal-weekcijfers-en-meer

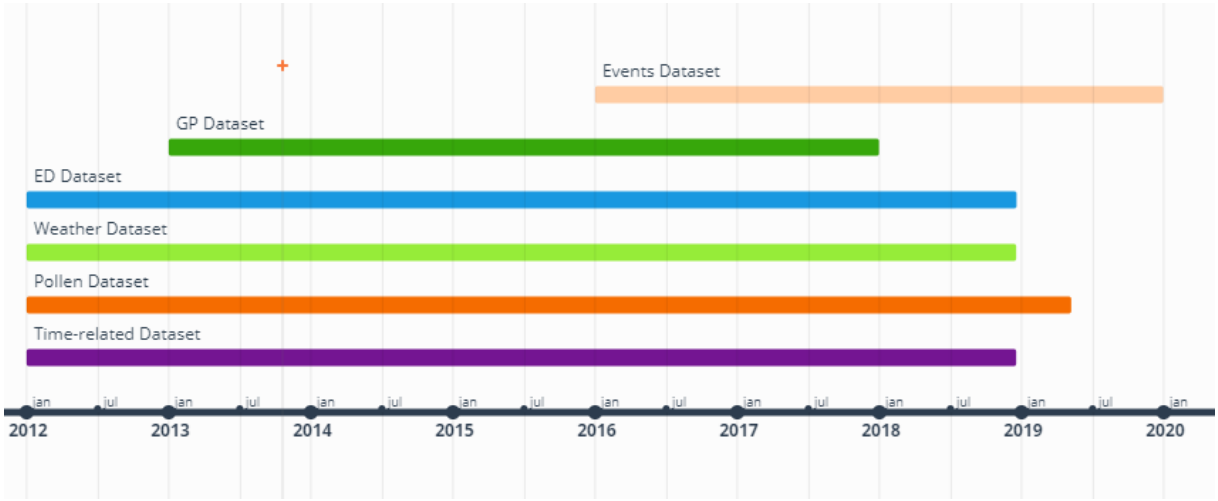

Figure 16 A display of all the available datasets and the periods over which they contain data.

4.2 Explore data & verify the quality

In this section, we will take a closer look at the available data. We will look at the quality of the data and check for anomalies within the datasets. We also try to find certain patterns and correlations within the data. Note that the descriptive statistics and correlation analysis in this section are not used for future selection for the models. They are purely determined to get an overall idea and expectation of the datasets we are dealing with.

4.2.1 ED and GP-post data

Before we can explore the data in these sets, we have to make a few alterations to the two datasets. The datasets from the ED and the GP-post contain unique patient arrivals. In this research, we want to say something about daily volumes. This means that we have to aggregate the datasets such that we obtain the total sum of patient arrivals per day. For all patients, the arrival dates were known, which also means there was no missing data (some other features had missing data, but these features are not used) to construct the aggregated values. After this alteration, we can plot the daily arrivals over the years and compute some descriptive statistics.

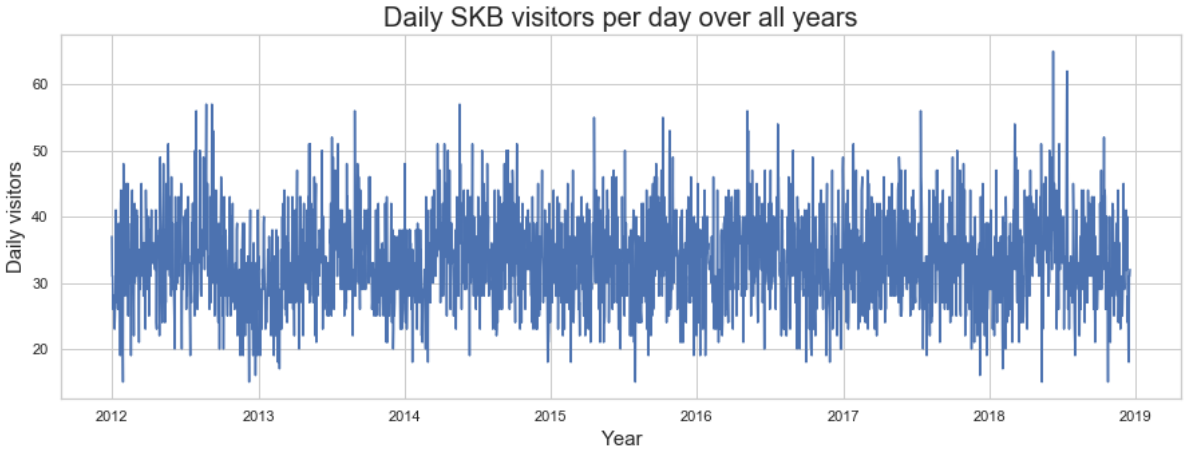

Figure 17 Daily ED visitors from 2012 to the end of 2018

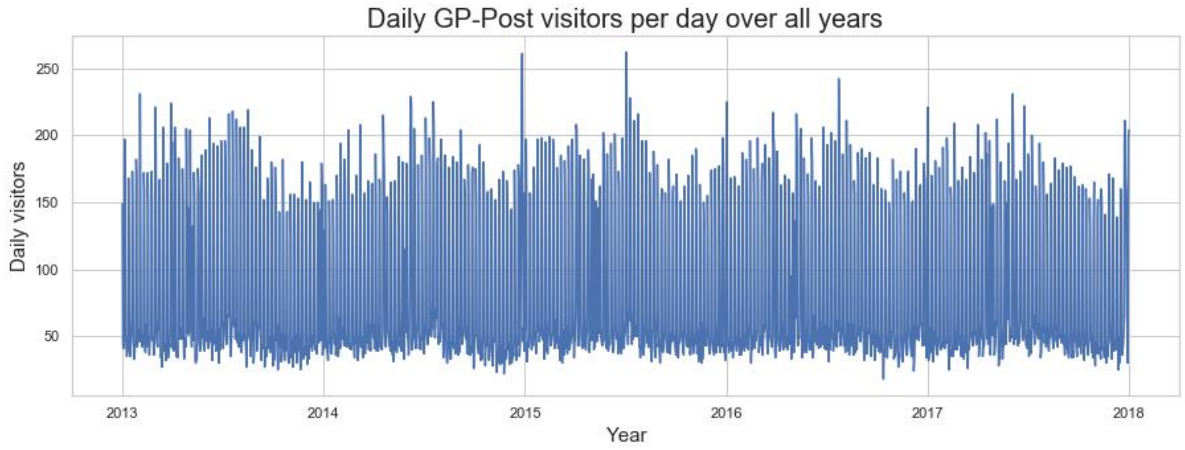

Figure 18 Daily GP-post visitors from 2013 to the end of 2017.

Table 14 Descriptive statistics for GP-post and ED.

	Count	Mean	Standard deviation	Min	25%	50%	75%	Max
GP-post	1826	82.00	57.95	18	42	50	75	262
ED	2545	33.42	6.92	15	28	33	38	65

When we take a look at Figures 17-18 and Table 14, we can see that the variation in patient arrivals is a lot bigger for the GP-post in comparison with the daily arrivals of the ED. This is caused by two reasons; the first reason is that the GP-post sees more patients overall than the ED and the second reason is that the daily arrivals of the GP-post are heavily dependent on the day. During normal workdays, they are only open outside regular office hours, but at the weekend they are open 24 hours a day. To illustrate the differences between the days in a week we have constructed two boxplots for both datasets that distinguish the patient arrivals per day in Figures 19-20 and created Table 15 with descriptive statistics for weekdays and weekends.

Table 15 Descriptive statistics for ED and GP-post separately for weekdays and weekends.

	Count	Mean	Stdev	Min	25%	50%	75%	Max
GP-post (weekday)	1304	47.91	19.81	18	40	45	51	224
GP-post (weekend)	522	167.13	24.75	112	150	165	182.75	262
ED (weekday)	1818	34.29	6.65	15	30	34	39	57
ED (weekend)	727	31.24	7.09	15	26	31	36	65

Looking at Table 15, we can see that there is only a small difference for the arrivals at the ED during the weekends and the weekdays. It appears that they see slightly fewer patients at the weekend. For the GP-post we find that the mean arrivals are more than three times as large during the weekends. Looking at the boxplots in Figures 19-202, we can confirm that Saturday is the busiest day, followed by Sunday. It also seems like the ED has slightly more arrivals on Monday and Friday in comparison with the other days.

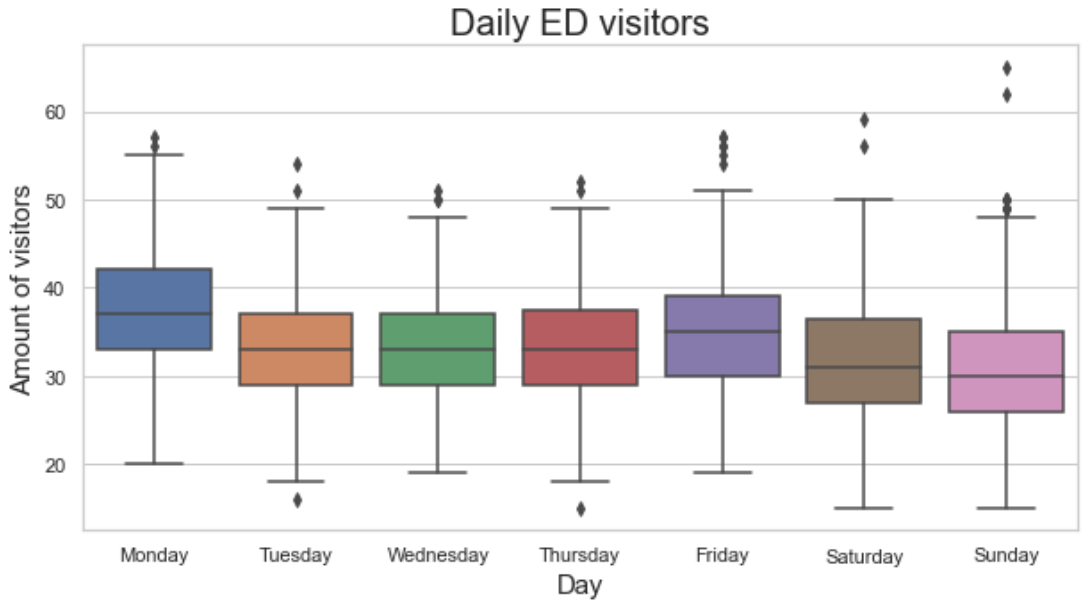

Figure 19 Boxplot for daily ED Visitors divided by day of the week.

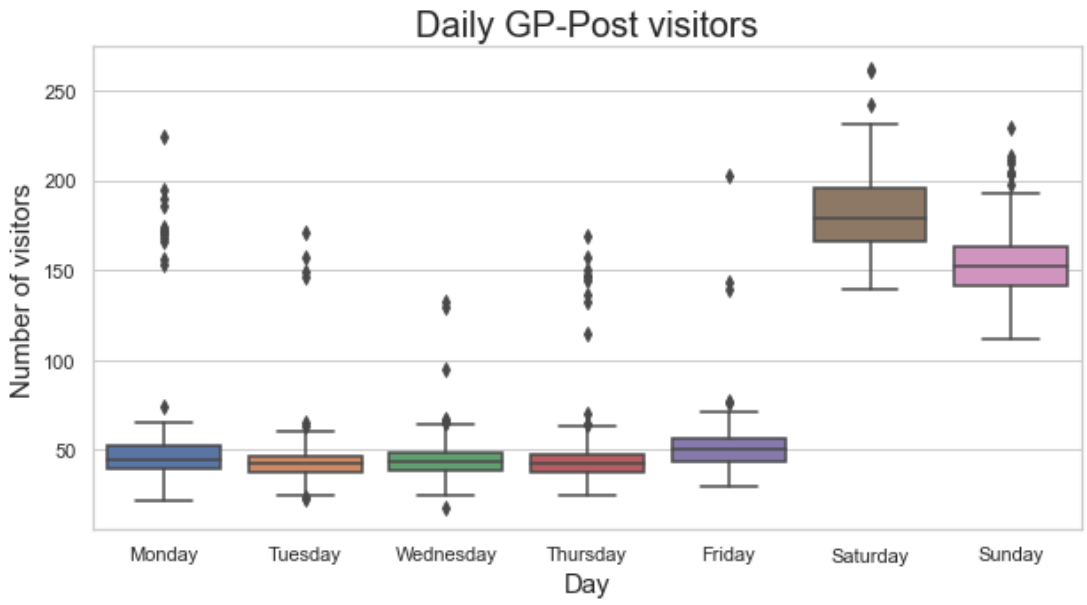

Figure 20 Boxplot for daily GP-post visitors divided by day of the week.

Figure 20 suggests that there are quite some significant outliers for the number of visitors during the weekdays for the GP-post. This effect is caused by holidays in the Netherlands. On these days the GP-post operates 24 hours contrary to normal weekdays. We obtain the boxplot displayed in Figure 21 by temporarily removing the holiday weekdays to illustrate that the big outliers disappear once we look at normal weekdays only.

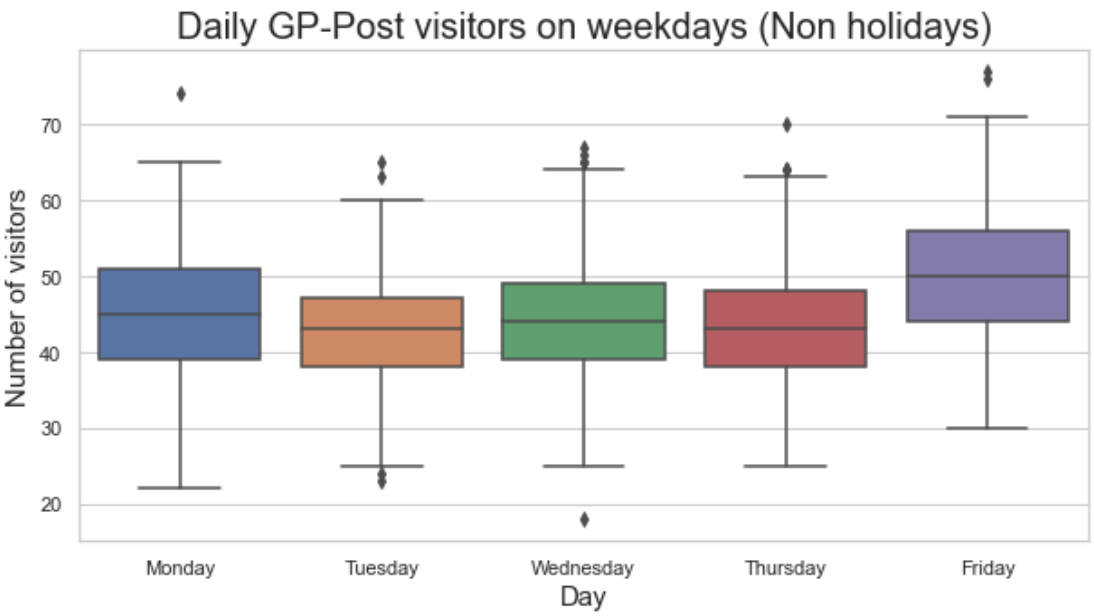

Figure 21 Boxplot for daily GP-post visitors divided by day of the week (non-holidays).

We also wanted to check whether we find differences in the number of visitors if we separate the dataset by months. We created boxplots again for both scenarios. A slight adaption is made for the GP-post in which we take the difference between weekdays and weekends into account. The boxplots are presented in Figures 22-23.

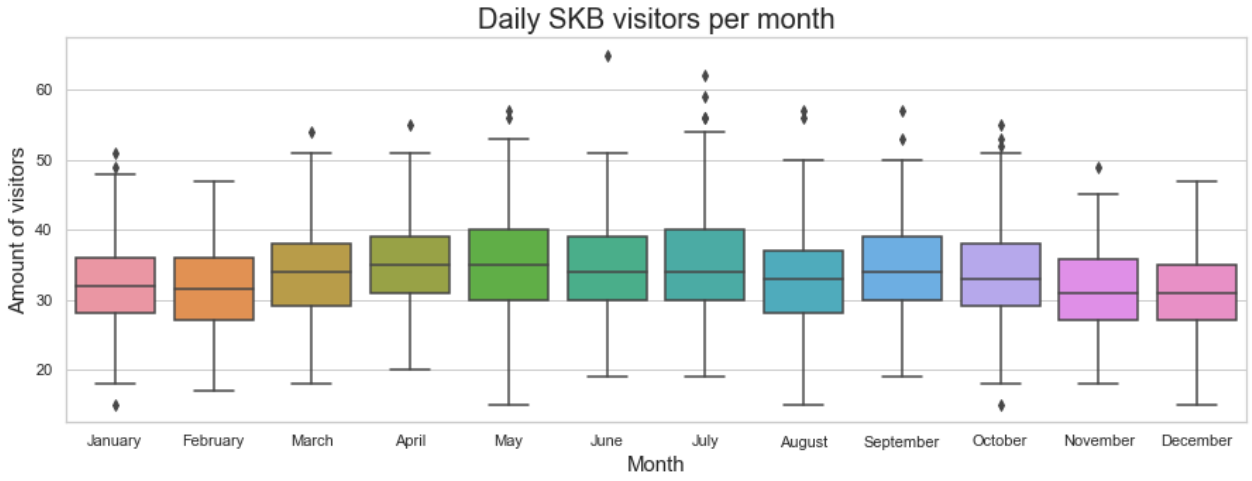

Figure 22 Boxplot for daily ED visitors divided per month.

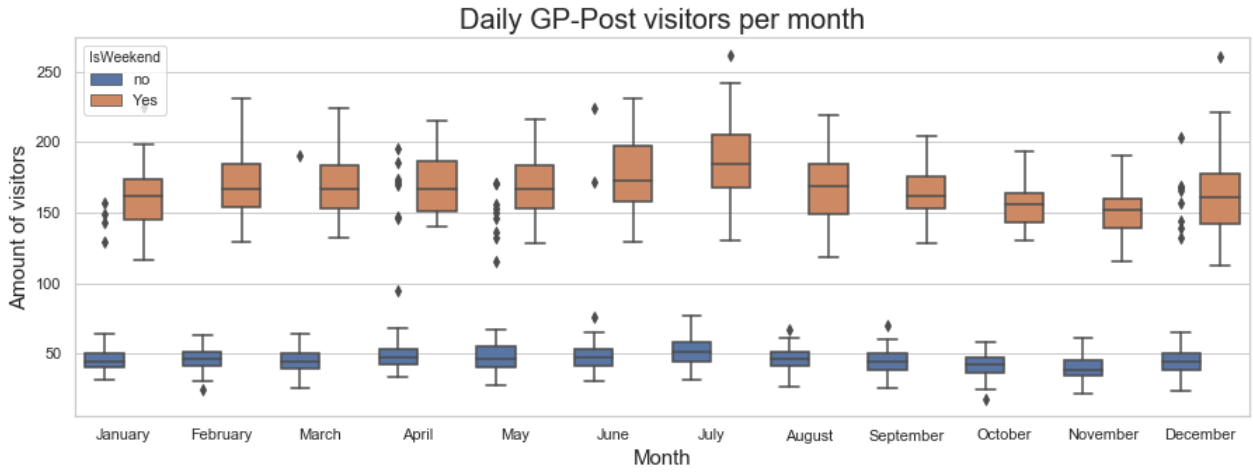

Figure 23 Boxplot for daily GP-post visitors divided per month and grouped by weekend or non-weekend.

The daily arrivals at the ED for the different months seem to be pretty stable. The slight variation may simply be caused by some randomness in the arrival pattern. The arrivals at the GP-post at the weekends do seem to fluctuate a bit more. Figure 23 suggests that July is the busiest month and that November seems to be the least busy. The high outliers are once again due to the holidays within these months.

Lastly, we are also interested in whether there is a linear correlation between the arrivals at the ED and the GP-post. To also account for the difference in weekdays and weekends, we checked for three different scenarios and determined the correlation coefficients. The first scenario is simply the normal correlation coefficient between the two arrivals, the second scenario is comparing only weekend data and the last scenario compares only week data. We found the following correlation coefficients:

- Scenario 1: -0.164
- Scenario 2: 0.307
- Scenario 3: -0.038

Although all three are quite low it is interesting to see that there appears to be a fair correlation between the arrivals at the weekends according to the correlation relationship listed in appendix B.

4.2.2 Weather data

The weather dataset contains numerical values for all 29 different measures. We found that in the entire file there appears to be missing data for only one day in row 1787. This row misses data for; TG, UX, UXH, UN, UNH, and EV24. Since there is only one line with missing values in the file we decided to not use any in-depth missing data techniques to alter this. We simply imputed these values by taking the average over the three values before that missing entry and the three values after that missing entry. We've decided to create histograms of the different measures, to get an idea of the behavior and distribution of the variables. All these plots are combined into one big plot illustrated in Figure 24 to keep it tidy and clear.

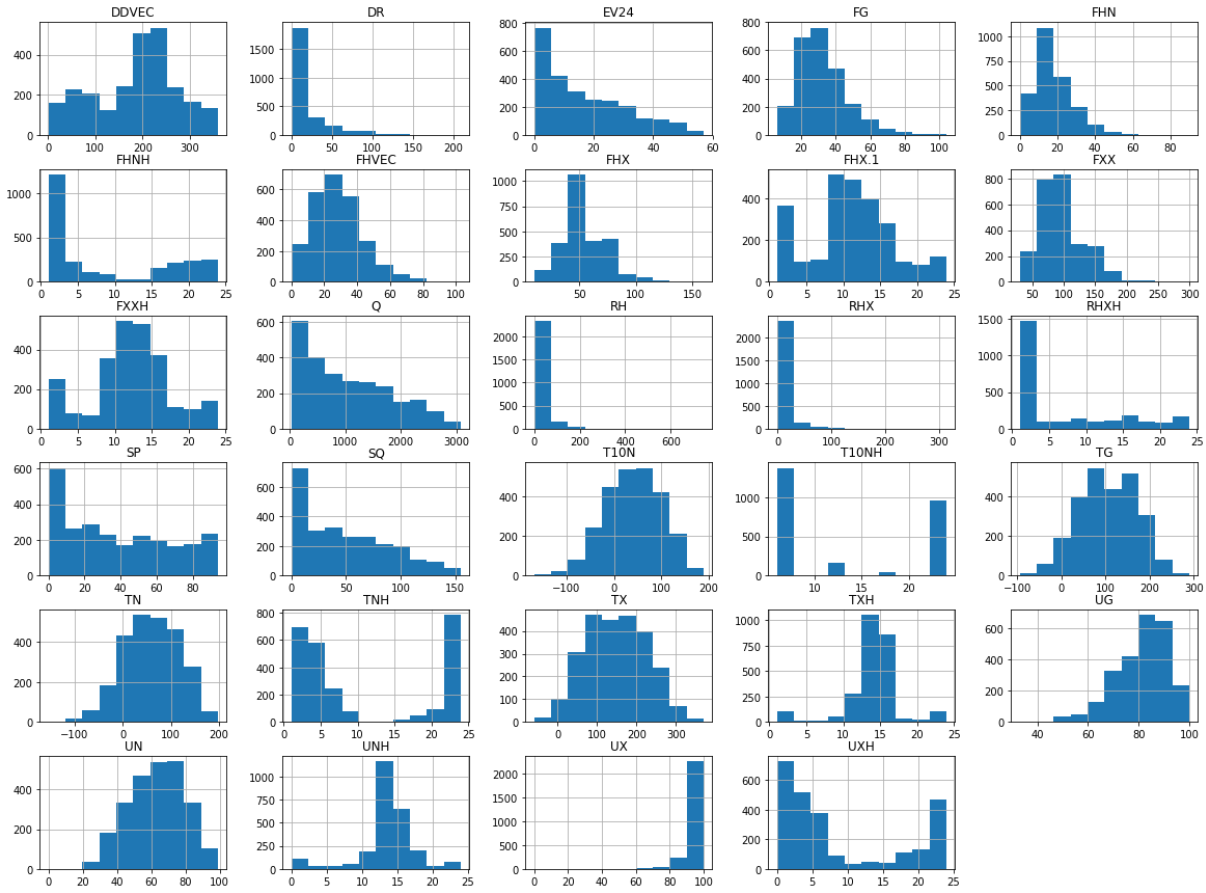

Figure 24 Histograms of the 29 numerical variables in the weather dataset.

We are also interested in whether there are any linear correlations between the daily weather data and the daily arrivals at the ED and GP-post. These correlation coefficients can be found in Tables 16-17. The descriptions of the variables are given in section 4.1.3.

Table 16 Correlation coefficients between weather variables and ED arrivals.

Variable	Correlation	Variable	Correlation	Variable	Correlation	Variable	Correlation
DDVEC	-0.062	**FXXH**	0.012	**SQ**	0.184	**UX**	-0.024
FHVEC	-0.066	**TG**	0.183	**SP**	0.150	**UXH**	-0.054
FG	-0.080	**TN**	0.128	**Q**	0.208	**UN**	-0.181
FHX	-0.071	**TNH**	-0.050	**DR**	-0.082	**UNH**	0.028
FHX.1	0.047	**TX**	0.207	**RH**	-0.036	**EV24**	0.208
FHN	-0.085	**TXH**	0.021	**RHX**	-0.018		
FHNH	0.002	**T10N**	0.094	**RHXH**	-0.070		
FXX	-0.066	**T10NH**	-0.045	**UG**	-0.147		

Table 17 Correlation coefficients between weather variables and GP-post arrivals.

Variable	Correlation	Variable	Correlation	Variable	Correlation	Variable	Correlation
DDVEC	0.023	**FXXH**	0.021	**SQ**	0.036	**UX**	-0.034
FHVEC	0.023	**TG**	0.036	**SP**	0.017	**UXH**	-0.017
FG	0.037	**TN**	0.029	**Q**	0.064	**UN**	-0.055
FHX	0.034	**TNH**	0.023	**DR**	-0.016	**UNH**	0.020
FHX.1	0.013	**TX**	0.038	**RH**	-0.020	**EV24**	0.061
FHN	0.025	**TXH**	0.036	**RHX**	0.003		
FHNH	-0.018	**T10N**	0.023	**RHXH**	-0.032		
FXX	0.036	**T10NH**	0.011	**UG**	-0.057		

We find that none of the daily weather variables have a correlation coefficient of at least (+ or -) 0.3 with either the daily ED or GP-post arrivals. Therefore we conclude that there is no correlation according to the correlation relationship listed in appendix B.

4.2.3 Pollen data

The dataset of the pollen seems to contain three pollen types that have no values recorded over the entire period. These are the following: Fabaceae, Ranunculaceae, and Hippophae, which all are non-allergenic. Additionally, it appears that two of the pollen types were only observed during 2012, 2013, and a small portion (first 1.5 months) of 2014. These were the strongly allergenic Cladosporium and the very strongly allergenic Alternaria. Figure 25 presents the pollen count per day of all the allergenic pollen types.

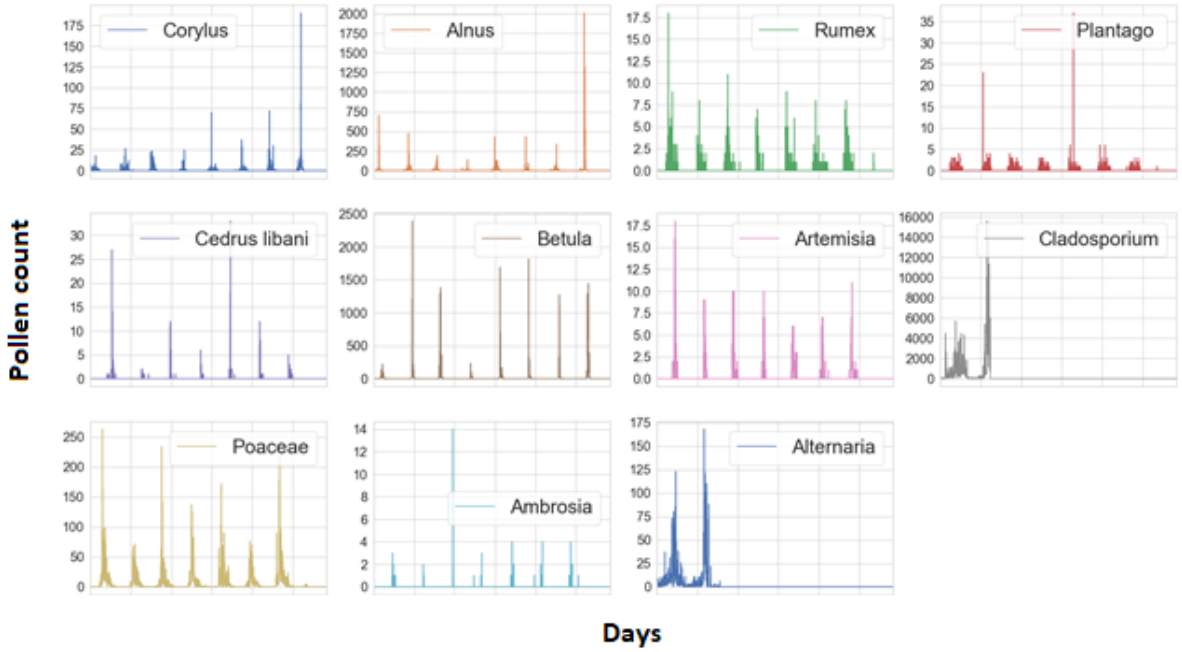

Figure 25 The data of all the allergenic pollen. Note that the pollen Alternaria and Cladosporium appear to be no longer collected after a certain period.

The allergenic pollen that contained data over the full period was also compared with the daily ED and daily GP-post visitors to see if there are any linear correlations between them. We had to adjust the datasets for both scenarios because the pollen data had to be aligned with the ED and GP-post respectively since all three datasets have a different amount of data (different periods). The final correlation coefficients that were found are given in Table 18.

Table 18 Correlation coefficient for allergenic pollen with full data versus ED visitors and GP-post visitors.

Pollen type	ED visitors	GP-post visitors
Corylus	-0.012	0.034
Alnus	0.008	0.003
Rumex	0.114	0.031
Plantago	0.044	-0.024
Cedrus libani	0.051	-0.009
Betula	0.044	-0.017
Artemisia	0.000	0.002
Poaceae	0.100	0.055
Ambrosia	0.014	-0.006

Table 18 suggests that there are no relevant linear correlations between the allergenic pollen in the air and the daily visitors to the ED or GP-post according to the correlation relationship listed in appendix B.

Lastly, we cannot determine whether there are missing values in the dataset. The zero measurement values are listed as empty entries in the dataset. For this reason, we are not able to distinguish between missing values and a zero measurement. Therefore, we assume that all empty entries are zero measurements.

4.2.3 Events data

The Events dataset contains a total of 13,671 events distributed over the years 2016-2019. The number of visitors is known for 6,142 of the events, the remainder is unknown. There seems to be a mistake in the dataset as the minimum number of visitors for one of the events is -100, which should most likely be 100. The maximum number of visitors for a certain event is 250,000. The frequency of the risks that are associated with the events is illustrated in Table 19, with their average visitors (if known and not 'onbekend').

Table 19 Counts for different risk groups for events.

Risk classification	Counts	Avg Visitors (if known)
Kennisgeving	3229	275
Regulier (A)	8692	931
Aandacht (B)	1580	4475
Risico (C)	69	24054
Empty	101	1229

Table 19 shows that the majority of the events are regular events (Class A: less risk associated) or events that do not require registration (Kennisgeving = only report that there is an event). It's also clearly visible that lower risk classifications are associated with fewer visitors. Unfortunately, a total of 101 events have no risk classification given to them.

4.3 Data selection

In the previous sections, we found that not all datasets have equal sizes and not all attributes are relevant or complete enough to use in the final model. In this section, we will determine which parts of each respective dataset will be used in the final dataset.

4.3.1 ED and GP-post dataset

We transformed all the unique arrivals into daily arrivals for the GP-post and ED. We will use all these daily values as our response variable in the models. We also want to use the values as a predictor for each other in the models. Unfortunately, the datasets were not collected over the same period. Which means that some data will be lost. The smallest of the two sets is that of the GP-post, which contains data from 01-01-2013 to 31-12-2017. That will be the time horizon that we will use in the base models.

4.3.2 Weather dataset

The weather station automatically collects the data from various measures. However, some of these measures are closely related to one another. This means that there is a very high correlation between them, which we want to prevent to counter collinearity. For example, the average daily temperature will be correlated with the minimum temperature measured 10 cm above the ground. Based on the prevention of this dependence and the used variables in literature we have decided to

include the following variables; daily average windspeed, daily average temperature, duration of sunshine, duration of rainfall, total amount of rainfall, and daily average relative humidity as predictors for the daily visitors in the models.

4.3.3 Pollen dataset

The majority of the pollen in the dataset are nonallergenic for humans, which means they are not expected to have any effect on the daily arrival rate at either the ED or GP-post. Unfortunately, we also found that two of the allergenic pollen were only collected (or no longer exist) for the first two years. For that reason, we have decided to not include these in the models as they would yield no benefit for future predictions if they are no longer registered. The remainder of the allergenic pollen will be used as predictors in the model.

4.3.4 Events dataset

The dataset contains data from the start of 2016 until the end of 2019, as mentioned earlier our relevant time horizon is from 01-01-2013 to 31-12-2017. This means that we cannot use the data from 2018-2019 to increase the performance of our final model. Additionally, this means that we lack any event data for the years before 2016. If we were to delete all data before 2016, our models would lose a big portion of the already scarce data. For that reason, it would be smarter to make two models, one that does contain the events and one that does not. This way we can still research whether the inclusion of events has relevance.

The dataset contains a lot of events in the region of North and Eastern Gelderland, however, not all of these events are relevant for our research. We will focus on events that take place in municipalities that are close to our region of interest. The municipalities that we will focus on are Aalten, Berkelland, Oost Gelre, and Winterswijk, since they are within the catchment area of the ED and GP-post. For these regions, we will sum all the events of a certain type per day and the total number of event visitors per day and use these as predictors for the model.

4.4 Integrating and constructing data

This section will show the steps that have been performed concerning the integration and construction of the data, such that we end up with datasets that can be used for the modeling part.

4.4.1 Data integration

All data is delivered in separate files, so the first thing that needs to be done is to integrate all data in four easy to use datasets, two for both the ED and GP-post models (with and without events). The main problem here is that all datasets have different lengths, so we have to make sure that we link the data from the correct dates with each other. Within these datasets, we still have to adapt some of the columns. Since we want to predict the daily visitors for the next day we need input values of the day before that. To do this we have to lag some of the predictors by 1 day. This means that the values of yesterday are used to predict the dependent variable of today. This is only done for variables that are not known on the day itself. For example, the day of the week or certain holidays are known for the future and don't need this.

4.4.2 Data construction

Our new datasets now contain all relevant data that was originally acquired. However, we also need to construct some new variables based on some of the others. we need to transform the categorical variables that we created into numerical values. Normally this wouldn't be the case for random forest regression, as the algorithm can work with categorical values. However, the random forest algorithm from the scikit-learn machine learning package in python requires the data to be

numerical. Fortunately, the data can easily be transformed into numerical values by using a method called one-hot encoding (Bishop, 2009). This method creates so-called 'dummy' variables for the different classes within a categorical variable. These dummy variables can then either have the value 1 (true) or 0 (false). An example is illustrated in Table 20.

Table 20 One-hot encoding example for variable; color.

Color	ColorRed	ColorBlue	ColorGreen
Red	1	0	0
Blue	0	1	0
Green	0	0	1

The example above shows that a single categorical variable 'color' with the values red, blue, and green can be transformed into three binary variables, which are either true (1) or false (0). Usually, one of the dummy variables is left out, since it can be predicted based on the values of the other dummy variables. For example, if the color is not red or blue it must be green. We can do this for our categorical variables as well to incorporate them into our models.

5 Method & experimental design

This chapter will address the proposed method that will aid in solving the research objective. The properties and design decisions of the method will be elaborated. Additionally, we will describe the experimental design that is used to assess the performance of the proposed method for the ED and the GP-post.

5.1 Proposed forecasting tool

In chapter 1 we introduced the idea of an all-inclusive tool. The main goal of this tool is that it can be used to create forecasts for the ED and GP-post. The forecasts are for the daily volume of patients arriving at the ED and the GP-post. We want to have separate models for these forecasts. The idea is to build a graphical user interface (GUI) that contains several functions that are required for the process from raw datasets to forecasts. The main reason for this is to keep the application simple for the end-users. All the operations can be done within one application, with the use of self-explanatory buttons. The functions that we want to add to this application are:

1) Use all individual raw datasets in combination with additional desired features to construct two separate main datasets that will function as input for the machine learning models of the ED and GP-post. This will allow the user to easily obtain the correct datasets and features for the actual training of the models. This also allows for easy adjustments in the future if new datasets are introduced or extra features need to be added since the function can simply be extended/altered.

2) Train user-specified (parameter settings) machine learning models for the ED and GP-post. These models can be saved on the computer and loaded into the application for later use. This will allow the user to easily try out different settings for the model and make predictions with earlier created models.

3) The trained machine learning models can be used to make predictions. These predictions can be on data for which the outcomes are known to test the performance of the model and also on real examples for which we have no verification (future forecasts).

4) There will be an option to optimize the parameters of the machine learning model. The user will be able to perform a grid search with user-specified parameters and the application will return the model with the best performance.

5) The performance of the machine learning model can be evaluated and shown within the GUI.

The application and its features will be build using the programming language Python. Python is a widely-used programming language for data science problems. It has several extensive and ready to use libraries for application building and machine learning algorithms. On top of that, the environment is completely free to use and new technologies are added/updated frequently.

A simplistic overview of the steps to be taken is illustrated in Figure 26 on the next page, the steps included will be explained more in-depth in the upcoming sections. We start with the raw datasets and combine these to create separate sets for the ED and GP-post. These sets contain all features and the labels (daily visitors) in the correct format to be used in the algorithm. We then pick our algorithm, set our hyperparameters, and give our datasets as input. The result will be a trained

model, with which we can create forecasts in real-life situations and also measure the performance on already known data.

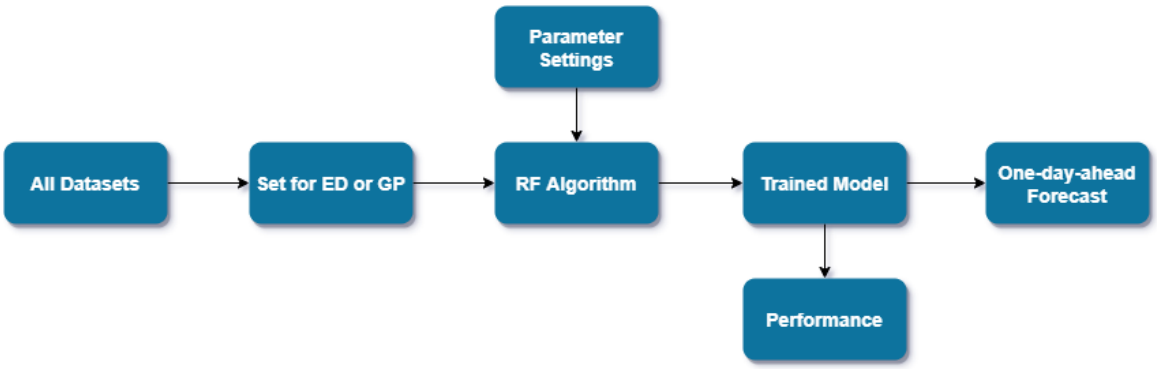

Figure 26 A simplified illustration of steps that need to be taken to go from the basic datasets to a useful forecast.

The first two steps in the Figure above have been addressed in the previous chapter about data understanding and preparation. There we've introduced our datasets and explained which features will be used for the final models. In this chapter, we will focus more on the algorithm itself, how we can properly train the model, and check the performance of the model.

5.2 Proposed model: random forest regression

In the literature study (chapter 3), we've found that the random forest algorithm would be the most suitable algorithm for our forecasting problem. We want to train random forest regression models on our data that can predict the daily volume of arrivals at the ED and the GP-post. We will first address the Python algorithm that we will use and it's required parameters. Secondly, we will elaborate on how we can determine the performance of the models that we train. We continue with the methods to validate our results and ensure that the produced errors are relevant and usable. Then we will address how we can find our model with the best parameters and performance. Lastly, we will explain what the forecasts for future situations should include.

5.2.1 Random forest regression algorithm and its parameters

The third block in the diagram in Figure 26 contains the algorithm and its predefined parameters that will be used to train a random forest model with the datasets. In Python, there is a machine learning library called scikit-learn which is free of use. This library provides functions for most common machine learning techniques, including a random forest regression algorithm. The function that will be used from this library is the 'RandomForestRegressor'. This is a meta estimator that fits many decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The function can take several parameters as input, which will influence the training of the random forest. Some of the parameters are not relevant for us to change and will be used with their default value. The parameters listed below are the ones which we will use to find the best models:

1) N_estimators: This is the number of decision trees that we will train in our random forest model. The default value is set at 100 trees.
2) Max_features: The number of features to consider when looking for the best split in a tree. This can be an integer that specifies the number of features or a float which will specify a ratio of the total number of features. The default value is set to the total number of features.

3) Min_samples_leaf: The minimum number of samples required to be a leaf node. This can be an integer that specifies the minimum number of samples to be in a leaf node or a float which is the ratio of the total number of samples. The default value is set at 1 sample.
4) Max_samples: The number of samples to draw (with replacement) from the dataset to train each base estimator, only if bootstrap is true.
5) Bootstrap: This specifies whether we want to use bootstrapped samples to build the trees or use the entire dataset. This can be turned on or off by setting the value to true or false respectively. The default value is set at true.
6) Oob_score: This specifies whether the out-of-bag data (non-chosen samples during bootstrap) must be used to estimate the R^2 on unseen data. This can also be specified as true or false. The default value is set at false.
7) Random_state: This controls the randomness of the bootstrapping of the samples used when building the trees and the sampling of the features to consider when looking for the best split at each node. An integer value can be entered which will specify a certain random seed. The default is set at None, which means we get different results every time we run the same settings.

A random forest model that contains more trees will always perform better than a model with fewer trees (Probst & Boulesteix, 2018). However, the improvement gained per tree decreases as the number of trees become larger and larger. For this reason, the number of trees should be set sufficiently high, whilst still maintaining reasonable running times. When the number of trees is chosen high enough, the randomness of the random forest is mainly influenced by max_features, min_samples_leaf, and max_samples (Probst, Wright, & Boulesteix, 2019). Max_features is shown to have the most effect, but experiments have shown that the latter two also are worth tuning. Section 5.3.2 will give more information about the parameters and within which ranges they will be tuned for our models.

In some of the models, the oob_score will also be activated to keep track of the performances of different models. We need to include the use of bootstrap samples to do this. More information about this method will be given in sections 5.2.3-4.

Lastly, the random_state is used to reproduce the same results when we rerun the algorithm. The actual number is not relevant, any seed will suffice. However, we must use the same random state for all different models (different parameter settings). This way we guarantee that we compare the performance of different models fairly and that the change in performance is not due to picking other features, samples, etc.

5.2.2 The performance of the model

Once the models have been successfully trained we need to determine the performance of such models. We are dealing with regression forecasts, therefore we'd like to express the error in terms of the predicted value and the actual value on known data. There are scale-dependant and percentage errors to assess the performance of a model (Hyndman & Athanasopolous, 2018) of which three will be introduced on the next page. These three metrics will be calculated and used to compare different models with each other on their performance.

For the performance metrics below (formulas 1-3) we define the error (e_t) as the actual value (y_t) minus the predicted value ($\hat{y}_t$). Where the actual value is a data point in the dataset and the predicted value is produced by the model. Since the objective is to create forecasts that are similar to the actual value it speaks for itself that the errors should be as low as possible.

Mean absolute error
The Mean Absolute Error (MAE) is a scale-dependent error. Forecast methods that minimize the MAE will lead to forecasts of the median. This method should not be used if you want to compare series that involve different units. The error has the same scale as the prediction, which in our case is the number of patients.

$$MAE = mean(|e_t|) \quad (1)$$

Root mean squared error
The Root Mean Squared Error (RMSE) is also a scale-dependent error. This metric is similar to the mean squared error but is easier to interpret because the scale is similar to that of the data. Similar to the MAE, this error also has the same scale as the prediction.

$$RMSE = \sqrt{mean({e_t}^2)} \quad (2)$$

Mean absolute percentage error
The Mean Average Percentage Error (MAPE) is a unit-free measure and is expressed as a percentage. The error is simply divided by the actual value Y_t. This measure is often used to compare the performance of different datasets and results where the scale is not the same. This method does not work well when the actual values are close to 0 since the error would then approach infinity.

$$MAPE = mean\left(\left|\frac{e_t}{Y_t}\right|\right) = mean(|p_t|) \quad (3)$$

5.2.3 Validation of the model

Now that we know how to evaluate the models on their predictions we need to ensure that metrics are calculated over a representative portion of the dataset. We must ensure that data that is used for training is not used to also evaluate the model, as this will give biased results. Secondly, we need to ensure that the set for evaluation is sufficiently large, whilst also keeping training samples large. We will use two different methods to validate our models. One of the methods is using the built-in function in the machine learning algorithm and is based on bootstrapping the data. The second method is a more commonly known method using k-fold cross-validation to estimate the error.

5.2.3.1 Bootstrapping

Earlier in this chapter, we mentioned that the machine learning algorithm had the option to include bootstrapping of the data. This is a technique to utilize the dataset for training and testing. Starting with the full dataset, for every iteration (every tree in the forest) we select a random amount of data points from the entire dataset for training. We do this with a replacement of already chosen points, which means that data points can be selected more than once. The unselected data points within an iteration will be used to determine the performance and are the so-called out-of-bag data points. Figure 27 illustrates this principle with an example.

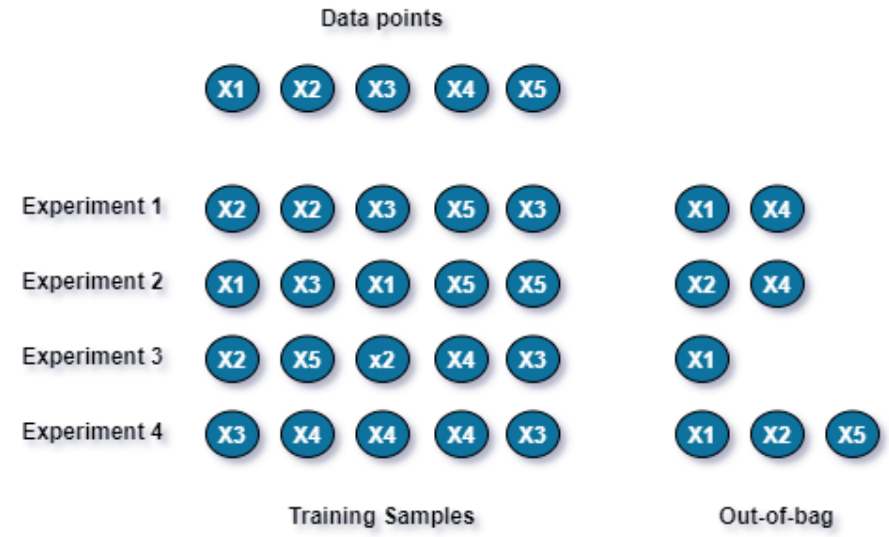

Figure 27 Bootstrapping illustration; we start with five data points and for every experiment, we select five random points with replacement for training, the remainder of the points will be used for testing.

Let us consider that we have (n) samples in our dataset and we pick (n) random samples with replacement. Since we pick them at random we have a $\frac{1}{n}$ chance to pick a sample and a $1 - \frac{1}{n}$ chance to not pick a sample. The total chance that after (n) picks we did not pick a certain sample is $(1 - \frac{1}{n})^n$ which for large n is approximately $\frac{1}{e} \approx 0.368$. In other words, we have around 36.8% of samples that will not be used for training in a certain tree of the random forest. Since the random forest consists of many trees which are all built with different training samples and have different unused samples, we can use this to our advantage. We check for every sample in our dataset in which tree they were not used for training. We then give this sample as input to these trees and take the average over all the outputs. This way we obtain a single value prediction for that sample. In the end, we will have an out-of-bag estimate for all our samples in the dataset. These predictions can then be used to determine our performance metrics. This principle is illustrated for a single sample in Figure 28.

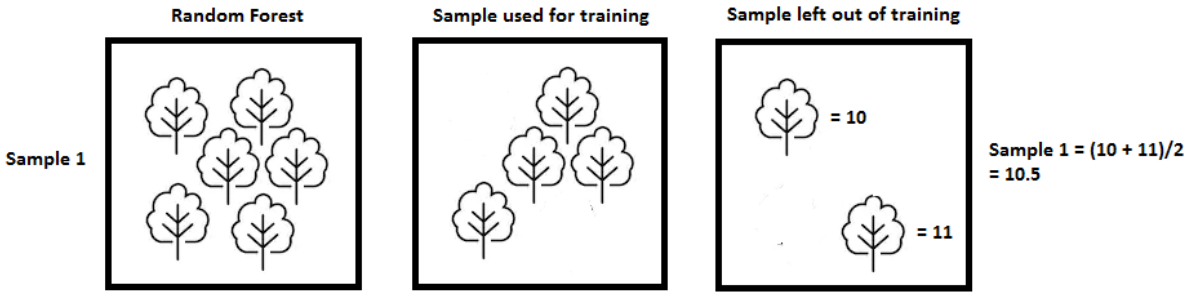

Figure 28 Example of a sample in a random forest consisting of six trees where the sample was out-of-bag in two of them. We then take the prediction of this sample from the two trees in which it was not used for training. The test prediction for Sample 1 will be the average of 10 and 11.

5.2.3.2 K-fold cross-validation

This method divides the entire dataset into different folds. The 'K' stands for the number of folds in which the total dataset will be partitioned. K-models are then trained with (k-1) training folds and tested with one testing fold. The testing fold and training folds are different for each of the K-models, although the training set will have overlap (Hastie, Tibshirani, & Friedman, 2017). An example of 4-fold cross-validation is illustrated in Figure 29.

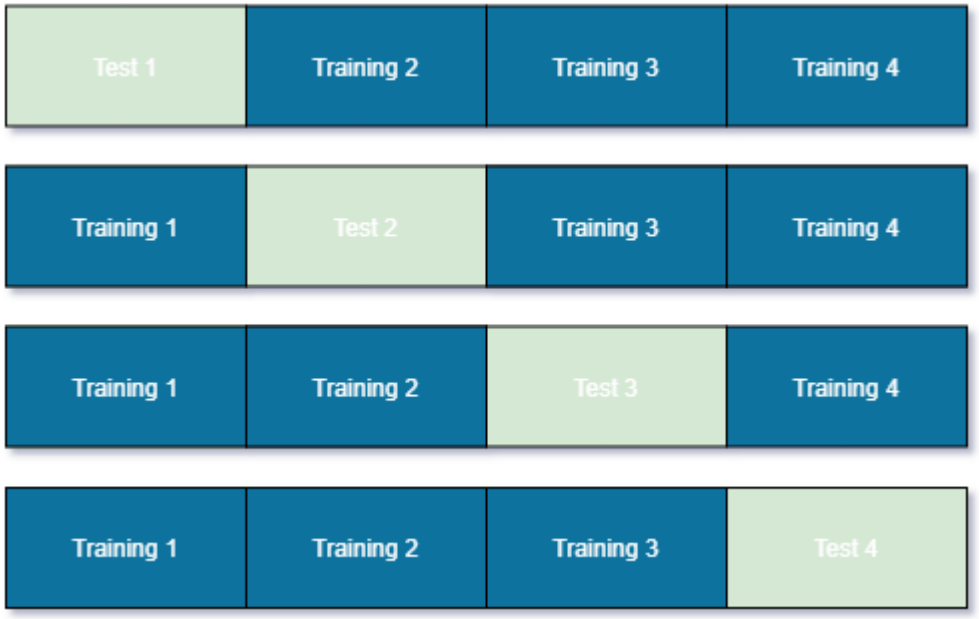

Figure 29 4-fold cross-validation; the testing fold is denoted by the yellow area and is different for every model, the training fold changes similarly but has overlap with the training set in the other models.

The performance of the final model is then estimated as the average over the error over the different testing folds. In the example above we'd take the four test errors and calculate the average of these to find the model performance.

5.2.4 Best model selection

Now that we know the important parameters of our model and how we can test the performance of our models. We can combine this knowledge to find the best parameters for our model such that we obtain the model with the best performance. To do this we need to train many different models with varying parameters and compare their performances with each other. We propose to apply a grid search on the parameters that influence the performance of the random forest model. The process to find the model with the best parameters would work as depicted in Figure 30.

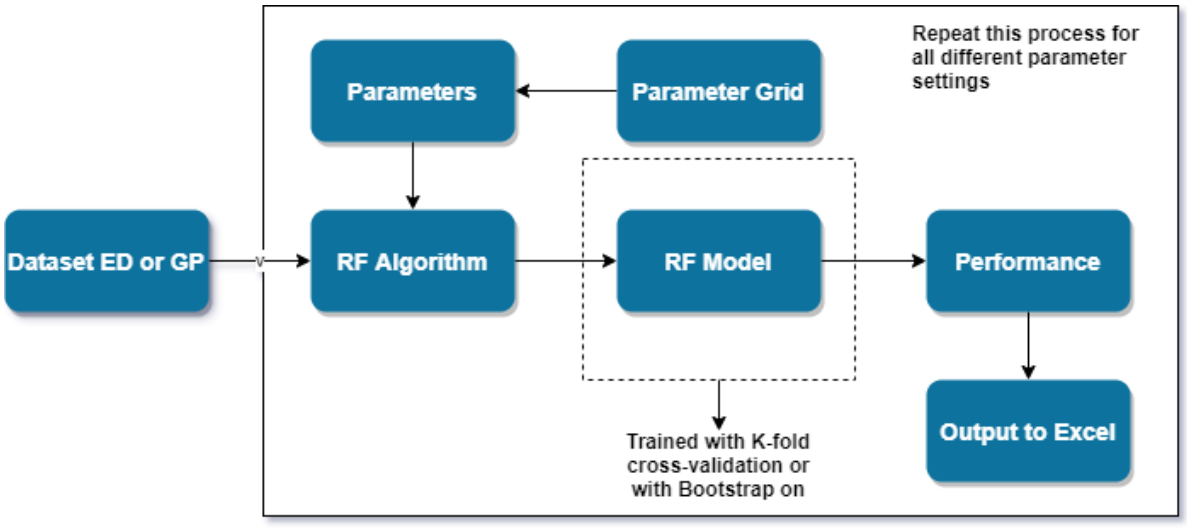

Figure 30 Process to obtain the model with the best parameter settings for optimal results.

The parameters in the grid search will be; the max features to be considered per split, the minimum number of samples in a node, and the number of samples to draw as mentioned earlier. The actual values that we will consider for these parameters will be introduced in section 5.3 about the proposed experiments.

The optimization is performed twice, once for each of the validation methods. The option to use bootstrap samples within the random forest algorithm will be turned on during the bootstrap experiments and off during the K-fold cross-validation method. Once all experiments are performed we will have the results in Excel. Depending on the validation method we can then find the best model as follows:

Bootstrap Method: The models are trained and tested with all available data points. The best model will be selected based on the out-of-bag error estimates of the models.

Cross-validation Method: We will have K-models for every parameter setting that we try. We can estimate the model performance by taking the average over the (K) performances. The final model is then fitted to all the data with the corresponding parameter settings (Hastie, Tibshirani, & Friedman, 2017).

We also want to know which of the features are most important within the best models. We want to create an overview of the most important features in the final models. These features are important for the acute care domain to get an idea of which factors influence the crowding on the ED or GP. We do this using the feature_importances attribute of the random forest algorithm in Python. The documentation of the algorithm specifies that the importance of the feature is computed as the total reduction of the criterion brought by that feature. The standard measure of the criterion is the mean squared error (MSE, 4) for random forest regression models (Sklearn Ensemble RandomForestRegressor, 2020). In formula (4) the mean squared error is given in terms of the actual value y_i and the prediction $\hat{y}_i$.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (4)$$

This means that for every split it checks how much the overall error improves for the predictions. The variables can then be ranked based on their improvement, where a larger improvement means a more important variable.

5.2.5 One-day-ahead forecast

The final model that we obtain can be used to create one-day-ahead forecasts for the number of visitors at the ED and the GP-post. This forecast will be an average over all the outputs of every individual tree within the forest. In addition to the forecast, we want to provide a prediction interval for this value. The model consists of multiple individual decision trees. So instead of only presenting the average value, the idea is to also register all the individual tree forecasts. These can then be ordered from smallest to largest. We then pick the 5^{th} and 95^{th} percentile of these numbers as a lower and an upper limit of the prediction. A good model would preferably have a small prediction interval for a prediction.

5.3 Proposed experiments

This section introduces the experiments that will be conducted to answer the research questions of this project. We first explain the experiments that we will do for the final models of the ED and the GP-post. Secondly, we will explain the variety of experiments that will be done to find the best

parameters for these models. Lastly, we address the comparison that we will make between the results of the models and some baseline predictions.

5.3.1 Unique models for ED and GP-post

As mentioned before we will create separate models for the ED and the GP-post. The reason being that they both operate independently of each other and also show great variation in the number of daily visitors, caused by the different opening times. In addition to that, we will also separate the dataset into one with events and one without. The model without the events will include all data from 2013 to 2017. The model with events however will only include data from 2016 to 2017. We decided to separate this because the period over which we have event data is so small that it would be a big waste to delete all other data for the period before that. We'd hope to see that the addition of events will have a beneficial effect on the model. Although a comparison between the two models will be difficult as there is a large gap in the available data. We also want to make the comparison between the dataset which was formed by ourselves after selecting data in chapter 4 and the full dataset including all features, without any preselecting. Lastly, we'd have to run the above-mentioned models all twice to account for the two validation methods. The total number of models that we will find are shown in Table 21.

Table 21 Total number of models for ED and GP-post. Note that they should be multiplied by two to account for both validation methods.

Model	**Years**	**Labels**	**Number of features**
ED (All data - Event)	2013-2017	Daily ED Visitors	100
ED (All data + Event)	2016-2017	Daily ED Visitors	105
ED (Selected data - Event)	2013-2017	Daily ED Visitors	42
ED (Selected data + Event)	2016-2017	Daily ED Visitors	47
GP-post (All data - Event)	2013-2017	Daily GP-post Visitors	100
GP-post (All data + Event)	2016-2017	Daily GP-post Visitors	105
GP-post (Selected data - Event)	2013-2017	Daily GP-post Visitors	42
GP-post (Selected data + Event)	2016-2017	Daily GP-post Visitors	47

This means that with all models mentioned in the Table above and the two validation methods that we have we will have to find 16 optimal models. The parameters that we will change to find the best models will be explained in the next section.

5.3.2 Experiments for finding the best model parameters

In the previous section, we mentioned that we will create 16 final models. Eight for each, the ED and GP-post. To obtain these final models we will have to perform a grid search on the model parameters as explained in section 5.2.4. This grid search will be done over three parameters; the maximum features per split, the minimum samples per node, and the sample size. We earlier mentioned that the number of trees should be sufficiently large to obtain good models. Research on 29 datasets with the random forest model has found that there were no significant differences between random forests build-out of 128 trees and more (Oshiro, Perez, & Baranauskas, 2012). However, since the running time of our algorithm is pretty fast we can take some safety for our models. We decided to use a total of 1000 trees in all the models. This should be more than enough to guarantee sufficient decision trees in every model.

For the actual parameters, we will try different values within the allowed limits. The maximum features considered per split can be any number between 1 and the total number of features. Since the total number of features of our datasets is not large we can easily explore all numbers within that

range. Typical values for maximum features per split are 1 and 5 used for classification and regression problems respectively (Probst, Wright, & Boulesteix, 2019). We can try to find differences in the performance of the models by changing that value from 1 to 5. Lastly, the sample size influences the diversity of the individual trees which has the effect of producing less correlating trees. On average this could increase the accuracy but the individual trees become less accurate. The optimal value for the sample size is found to be less than the total number of samples, which is often suggested to use (Martínez-Muñoz & Suárez, 2010). We will check for fractions between 0.5 and 1 of the total number of samples. This will only be done for the models validated by the bootstrap method, as the other models validated by the cross-validation method will have their respective folds for training and testing. This means that no samples will be drawn for these models and therefore this parameter is not relevant. All values for the different parameter settings can be found in Table 22.

Table 22 Parameter grid search for the best models. Note: The sample size is only used for bootstrap models.

Parameter	**Minimum value**	**Step size**	**Maximum value**
Number of trees	1000	-	1000
Maximum features per split	1	1	Total number of features
Minimum samples per node	1	1	5
Sample size (fraction of total)	0.5	0.1	1

5.3.3 Comparison model results versus baseline results

Lastly, we want to address the comparison that we want to make between the performance of the best models versus some fairly simplistic baseline results. The expectation is that the models will improve a lot on the results in comparison with these baseline results. If this is not the case then it means that the models are not able to outperform relatively easy methods of prediction.

The number of visits in the past that are known can be used in some way to create predictions for the one-day-ahead forecasts. The easiest would be to just assume that today's number of visits will be the number of visits for tomorrow. We can also slightly extend this by taking the average over the last three days as our prediction for the next day. This should provide better results for the ED than for the GP, the reason being that the GP is not always open for an equal number of hours (weekdays and weekends for example) in contrast to the ED. Therefore the arrivals of one day could differ greatly from the next day.

To compensate for that effect we also want to add predictions based on the known value of one week ago and the average of that same day over three weeks. For example, the prediction of Monday 28-5 would be based on Monday 21-5 for the first prediction and the average over Monday (7-5, 14-5, and 21-5) for the second prediction. Table 23 gives an overview of the four different baseline methods that we will take into account and compare with the results of our best models.

Table 23 Prediction methods for the baseline results, where $\hat{y}_t$ is the prediction and y_t are known values. the period t is given in days, such that a difference of seven equals one week.

Prediction method	**Prediction**
One day prior	$\hat{y}_t = y_{t-1}$
Average of three days prior	$\hat{y}_t = \frac{y_{t-1} + y_{t-2} + y_{t-3}}{3}$
One week prior	$\hat{y}_t = y_{t-7}$
Average of three weeks prior	$\hat{y}_t = \frac{y_{t-7} + y_{t-14} + y_{t-21}}{3}$

6 Experimental Results & Discussion

This chapter will present the results that were acquired by performing the experiments proposed in the previous chapter. The results for the ED and the GP-post will be divided into separate sections. This chapter will conclude with a discussion about the results.

6.1 Results for the GP-post

Experiments were performed with four different scenarios for the GP-post. The scenarios either include the Events data or not and are done with the full dataset and the reduced dataset (four scenarios in total). We also used two validation methods to find the optimal models. The result is that we obtained four optimal models based on the Bootstrap method and four optimal models based on the cross-validation method. The models in the Bootstrap method are ranked by their out-of-bag score which is automatically calculated by the random forest algorithm when bootstrap samples are included. This score is closely related to the RMSE which we've also calculated for all models. For that reason, the models validated by the cross-validation method will be ranked by their RMSE score. Additionally, the performance of these models is the average over the performance of the five different folds. The best models that we found for these two validation methods are summarized (numbers rounded to two decimals) in Tables 24-25. Note that the running time is in seconds, the MAE and RMSE are expressed in patients and the MAPE is expressed in a percentage error. The top ten models for every scenario can be found in Appendix C.

Table 24 Best models found for the GP-post, validated with the Bootstrap method.

Dataset	Num of Trees	Max Features	Min Samples	Sample Size	Run Time (s)	MAE	RMSE	MAPE	OOB Score
Reduced (No Event)	1000	20	3	0.9	4.69	8.32	11.44	12.49	0.96
Reduced (with Event)	1000	35	3	1.0	3.11	8.45	11.72	12.48	0.96
Full (No Event)	1000	71	3	0.9	14.47	8.31	11.47	12.35	0.96
Full (with Event)	1000	72	3	1.0	5.86	8.57	11.83	12.62	0.96

Table 25 Best models found for the GP-post, validated with the Cross-validation method.

Dataset	Num of Trees	Max Features	Min Samples	Run Time (s)	MAE	RMSE	MAPE
Reduced (No Event)	1000	29	5	4.12	8.39	11.51	12.57
Reduced (with Event)	1000	36	2	2.64	8.48	11.67	12.53
Full (No Event)	1000	65	5	9.73	8.35	11.55	12.41
Full (with Event)	1000	79	3	4.82	8.58	11.76	12.67

Looking at the results above we see that the performance of the models with no events is slightly better than the models with events for both of the validation methods. However, we must realize that the available data for the models with events is much smaller. Similarly, we've also found that the reduced models perform slightly better on the ranking criteria than the models containing all features. However, the differences are minimal in both cases. This could be due to the fact a lot of useless variables are still included in the full dataset, which could cause a random selection to only have bad picks. We've also determined the top 10 most important features for each of the models, these are displayed in tables 26-27. Interestingly it appeared that all models were greatly dominated by the same features. The most important feature in all of the models was whether the day was on the weekend or not. After this feature, a few features were fluctuating around in position a bit. These features were;

- The days; Saturday and Sunday
- Whether the day is a holiday or not
- The daily GP visitors of the day before

The importance of the features that came after these were all very small and almost neglectable on the total importance.

Table 26 Top 10 features for GP-post models validated with the Bootstrap method.

	Reduced (No Event)	**Reduced (with Event)**	**Full (No Event)**	**Full (with Event)**
Feature 1	IsWeekend_Yes (0.39)	IsWeekend_Yes (0.44)	IsWeekend_No (0.42)	IsWeekend_Yes (0.42)
Feature 2	IsWeekend_No (0.37)	IsWeekend_No (0.41)	IsWeekend_Yes (0.42)	IsWeekend_No (0.42)
Feature 3	Day_Saturday (0.09)	Day_Saturday (0.03)	Day_Saturday (0.03)	Day_Saturday (0.04)
Feature 4	Day_Sunday (0.04)	Holiday_Yes (0.03)	Holiday_No (0.03)	Holiday_Yes (0.03)
Feature 5	Holiday_Yes (0.03)	Holiday_No (0.03)	Holiday_Yes (0.03)	Holiday_No (0.03)
Feature 6	Holiday_No (0.03)	Day_Sunday (0.01)	Day_Sunday (0.02)	Day_Sunday (0.02)
Feature 7	Daily-GP-1 (0.02)	Daily-GP-1 (0.01)	Daily-GP-1 (0.01)	Daily-GP-1 (0.01)
Feature 8	TG (0.01)	TG (0.01)	TX (0.00)	Event_K (0.00)
Feature 9	UG (0.00)	Event_K (0.00)	TG (0.00)	TX (0.00)
Feature 10	SQ (0.00)	FG (0.00)	DDVEC (0.00)	TG (0.00)

Table 27 Top 10 features for GP-post models validated with the Cross-validation method.

	Reduced (No Event)	**Reduced (with Event)**	**Full (No Event)**	**Full (with Event)**
Feature 1	IsWeekend_Yes (0.42)	IsWeekend_Yes (0.44)	IsWeekend_No (0.42)	IsWeekend_Yes (0.43)
Feature 2	IsWeekend_No (0.42)	IsWeekend_No (0.41)	IsWeekend_Yes (0.41)	IsWeekend_No (0.42)
Feature 3	Day_Saturday (0.04)	Holiday_Yes (0.03)	Day_Saturday (0.04)	Holiday_No (0.03)
Feature 4	Holiday_Yes (0.03)	Holiday_No (0.03)	Holiday_Yes (0.03)	Holiday_Yes (0.03)
Feature 5	Holiday_No (0.03)	Day_Saturday (0.03)	Holiday_No (0.03)	Day_Saturday (0.03)
Feature 6	Day_Sunday (0.02)	Day_Sunday (0.01)	Day_Sunday (0.02)	Day_Sunday (0.01)
Feature 7	Daily-GP-1 (0.01)	Daily-GP-1 (0.01)	Daily-GP-1 (0.01)	Daily-GP-1 (0.01)
Feature 8	TG (0.01)	TG (0.01)	TX (0.00)	Event_K (0.00)
Feature 9	UG (0.00)	Event_K (0.00)	TG (0.00)	TX (0.00)
Feature 10	FG (0.00)	FG (0.00)	Day_Friday (0.00)	TG (0.00)

We were also able to plot the results of the out-of-bag predictions accompanied by their prediction intervals. This was only available for the models trained with the bootstrap samples. We've also added the actual values as an indication of the performance of the models. The predictions of the best model can be found in Figure 31 and the other plots can be found in the first three Figures of Appendix D. Note that the predictions are sorted from smallest to largest, which clearly shows the difference in predictions during the normal working days and when they are open 24 hours. We also see that the gap between the lower and upper bound of the prediction intervals are quite big, which suggests that the model is not certain about the prediction. On top of that, quite some of the actual values fall outside of the prediction interval bounds, which also suggests that the model is not performing optimally. We've also determined the residuals of the prediction and constructed a Q-Q plot to compare the quantiles of the residuals to that of a normal distribution. These plots are illustrated in Figure 32. Similar plots can be found for the other models in the first three Figures in Appendix E.

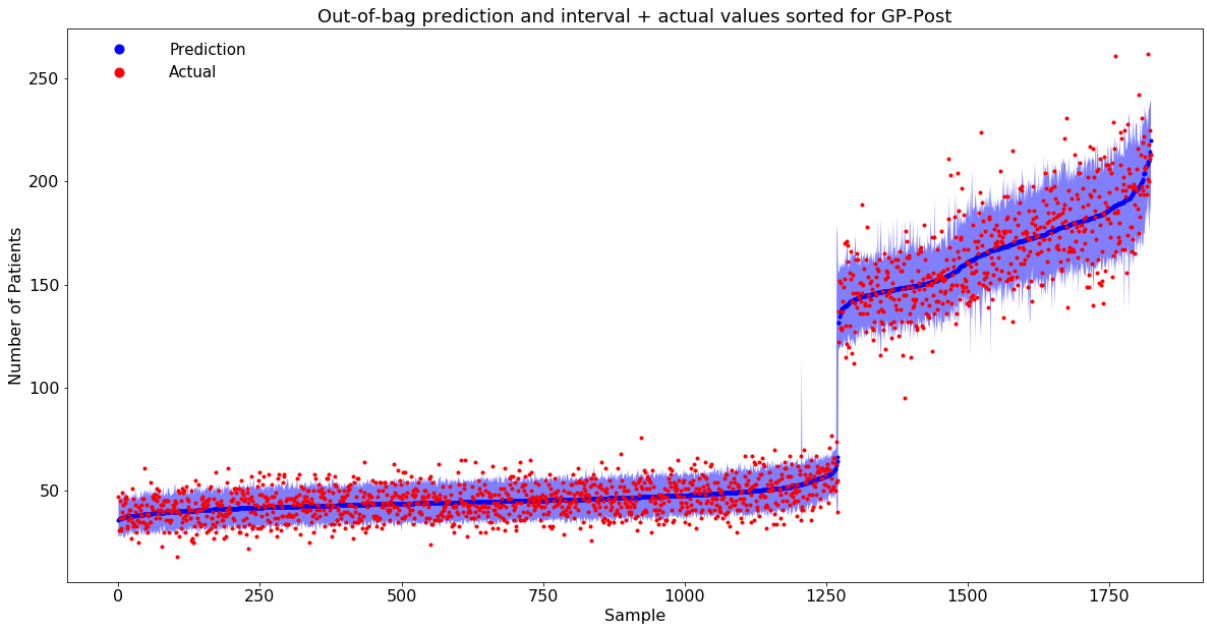

Figure 31 GP-post out-of-bag predictions + intervals for the scenario with the reduced dataset and no events sorted from small to large. The gap illustrates the difference between normal working days and days when the GP is open 24 hours.

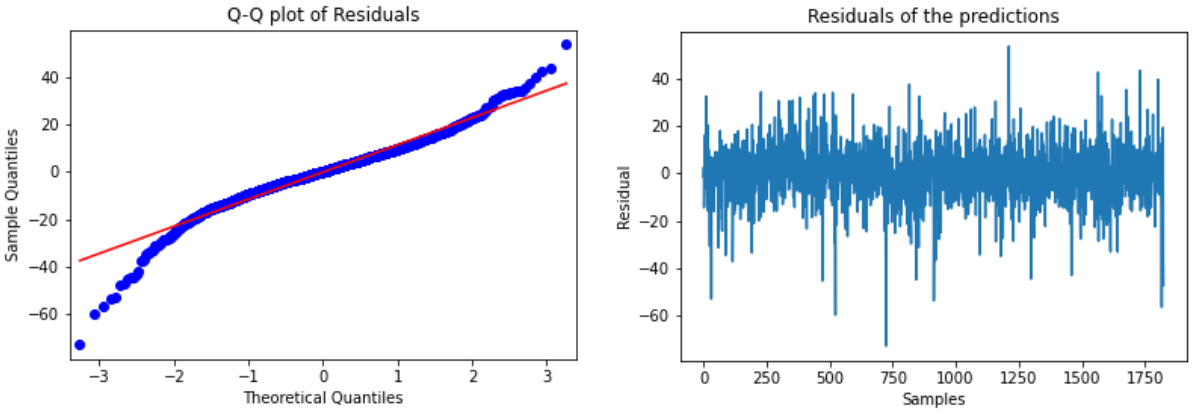

Figure 32 Q-Q plot and residuals of out-of-bag predictions for the scenario with the reduced dataset and no events.

We see that the points in the Q-Q plot in Figure 32 mostly follow the theoretical line. However, there are some deviations at the beginning and the end of the line. Note that this is a somewhat unfair comparison, as we compare the residuals of the predictions of the longer days with the shorter days. The deviation is less prominent when the residuals are split for the shorter days and longer days as shown in Figure 33.

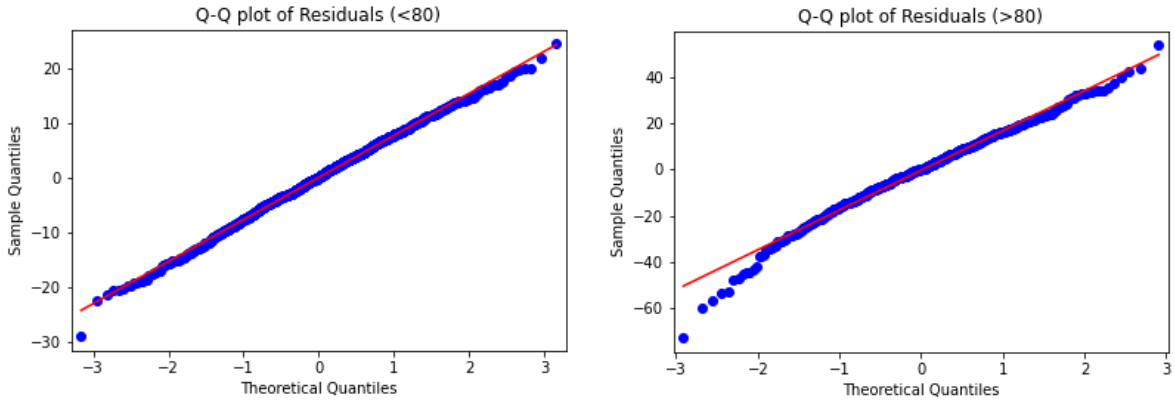

Figure 33 Q-Q plots of residuals of out-of-bag predictions adjusted for values smaller than 80 and larger than 80 for the scenario with the reduced dataset and no events.

Lastly, we also want to compare the results of the best models with the baseline results as proposed in section 5.3.3. The results of the baseline performances are displayed in Tables 28-29.

Table 28 Baseline results for GP models with no events.

Baseline Prediction	**MAE**	**RMSE**	**MAPE**
One day prior	44.03	67.03	62.91
Average over 3 days prior	60.03	74.63	82.48
One week prior	14.41	25.60	21.21
Average over 3 weeks prior	12.85	21.83	19.20

Table 29 Baseline results for GP models with events.

Baseline Prediction	**MAE**	**RMSE**	**MAPE**
One day prior	43.24	66.17	61.48
Average over 3 days prior	59.53	73.91	81.47
One week prior	14.11	24.86	20.52
Average over 3 weeks prior	12.27	20.90	18.55

As expected the first two predictions perform very badly since the demand per day varies a lot for the GP. We see that the best baseline performances are obtained by taking the average value over the three days that are respectively 1, 2, and 3 weeks ago for both the models with events and without events. The performance is however significantly worse in comparison with the performance of the best random forest models. This shows that the model has an advantage over simple prediction methods.

6.2 Results for the ED

Similarly as for the GP-post, a total of four scenarios were tested for the ED and validated with the Boostrap and Cross-validation method. The models were respectively ranked by their out-of-bag score and the RMSE again. The best models that we found for these two validation methods are summarized (numbers rounded to two decimals) in Tables 30-31. Note that the running time is in seconds, the MAE and RMSE are expressed in patients and the MAPE is expressed in a percentage error. The top ten models for every scenario can be found in Appendix C.

Table 30 Best models found for the GP-post, validated with the Bootstrap method.

Dataset	Num of Trees	Max Features	Min Samples	Sample Size	Run Time	MAE	RMSE	MAPE	OOB Score
Reduced (No Event)	1000	5	5	0.8	2.05	5.02	6.29	15.94	0.13
Reduced (with Event)	1000	5	1	0.6	1.66	5.21	6.44	16.57	0.11
Full (No Event)	1000	45	3	0.5	6.69	5.02	6.30	15.91	0.13
Full (with Event)	1000	8	3	0.7	1.67	5.23	6.49	16.69	0.10

Table 31 Best models found for the GP-post, validated with the Cross-validation method.

Dataset	Num of Trees	Max Features	Min Samples	Run Time	MAE	RMSE	MAPE
Reduced (No Event)	1000	5	5	1.60	5.03	6.30	15.94
Reduced (with Event)	1000	5	5	1.11	5.21	6.46	16.61
Full (No Event)	1000	22	2	5.12	5.01	6.29	15.86
Full (with Event)	1000	5	4	1.14	5.27	6.52	16.79

Similarly, as for the GP-post, we find that the models with no Events perform slightly better than the models with Events. Additionally, the models created with the reduced dataset seem to outperform the models with the full dataset, except for the cross-validated models with no events. The differences are once again very minimal, similar to the previous section. We also determined the top 10 most important features of the models again. However, this time we found that there was no clear winner for the most important feature. As most of the features seemed to be equally (low) important in value. Only in the models with the reduced dataset, validated by cross-validation and Bootstrapping method. We found that the daily visitors of the GP-post of the day before and the temperature of the day before scored an importance value of greater than 0.1. All other importances that were found were around the same value but in comparison to the GP-post models, mostly not time-related (weekend or day). Most of the important values are weather-related. the top 10 important features for all of the best ED models can be found in Tables 32-33.

Table 32 Top 10 features for ED models validated with the Bootstrap method.

	Reduced (No Event)	Reduced (with Event)	Full (No Event)	Full (with Event)
Feature 1	Daily GP (0.11)	SQ (0.08)	Daily GP (0.05)	SQ (0.04)
Feature 2	TG (0.11)	TG (0.08)	DDVEC (0.04)	Q (0.03)
Feature 3	SQ (0.09)	Daily GP (0.08)	TX (0.04)	Daily GP (0.03)
Feature 4	FG (0.07)	FG (0.07)	Day_Monday (0.04)	TX (0.03)
Feature 5	UG (0.06)	UG (0.07)	Q (0.04)	SP (0.03)
Feature 6	Daily-ED-1 (0.06)	Daily-ED-1 (0.07)	TG (0.04)	TG (0.03)
Feature 7	Day_Monday (0.06)	RH (0.05)	Daily-ED-1 (0.03)	UN (0.03)
Feature 8	Poaceae (0.05)	Visitors (0.05)	UN (0.03)	DDVEC (0.03)
Feature 9	RH (0.04)	DR (0.04)	T10N (0.03)	TN (0.03)
Feature 10	DR (0.04)	Poaceae (0.04)	TN (0.03)	T10N (0.03)

Table 33 Top 10 features for ED models validated with the Cross-validation method.

	Reduced (No Event)	Reduced (with Event)	Full (No Event)	Full (with Event)
Feature 1	Daily GP (0.11)	SQ (0.09)	Daily GP (0.05)	SQ (0.04)
Feature 2	TG (0.10)	Daily GP (0.08)	TX (0.04)	Q (0.03)
Feature 3	SQ (0.09)	TG (0.08)	DDVEC (0.04)	Daily GP (0.03)
Feature 4	FG (0.07)	UG (0.07)	Q (0.04)	SP (0.03)
Feature 5	UG (0.07)	Daily-ED-1 (0.06)	TG (0.03)	TG (0.03)
Feature 6	Daily-ED-1 (0.06)	FG (0.06)	Day_Monday (0.03)	UN (0.03)
Feature 7	Day_Monday (0.06)	Poaceae (0.05)	T10n (0.03)	TX (0.03)

Feature 8	Poaceae (0.04)	Visitors (0.05)	TN (0.03)	T10n (0.03)
Feature 9	RH (0.04)	RH (0.05)	UN (0.03)	EV24 (0.03)
Feature 10	DR (0.04)	DR (0.04)	Daily-ED-1 (0.03)	TN (0.03)

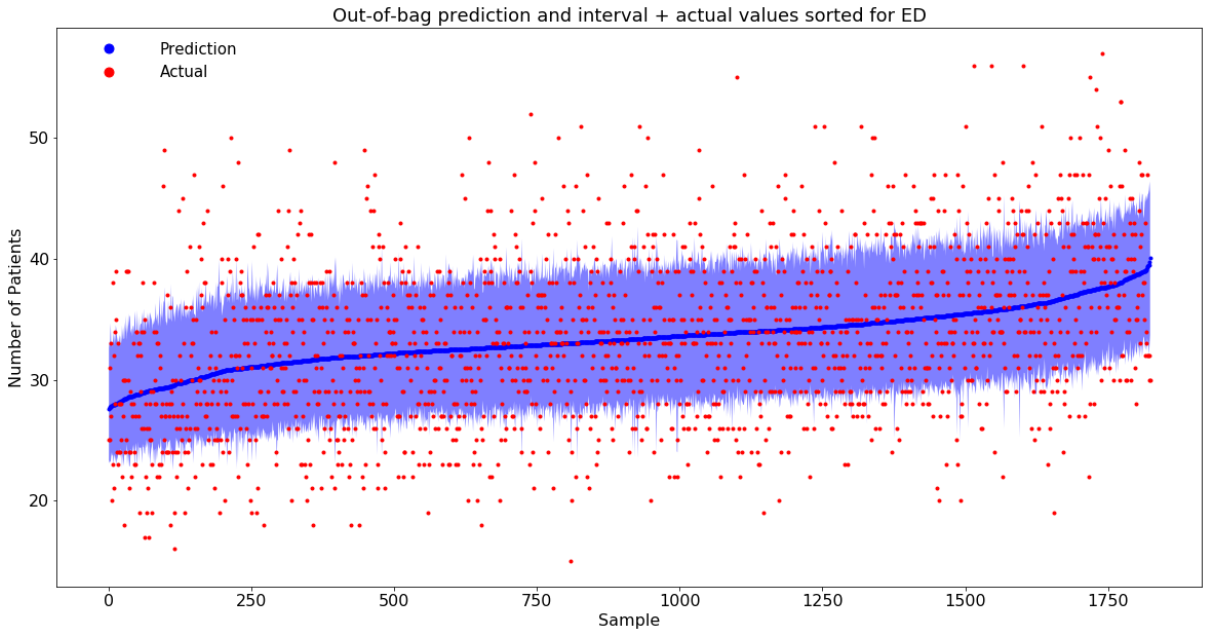

Figure 34 ED out-of-bag predictions + intervals for the scenario with the reduced dataset and no events.

We also made plots for the out-of-bag predictions obtained by the models that were validated by the bootstrap method. The model with the best performance is plotted in Figure 34. The others can be found in the last three Figures in Appendix D. Looking at the predictions and their interval it is very clear that this is not a good model. The interval is quite broad and many of the observations still fall outside of the region. This means that the actual values are not predicted within any of the trees in the 5th and 95th percentile. We've also created a plot of the residuals and Q-Q plot, illustrated in Figure 34. The quantiles of the residuals seem to fit the theoretical normal quantiles pretty well, with a few small deviations at the beginning and end of the line. Similar plots for the other models can be found in the last three Figures of Appendix E.

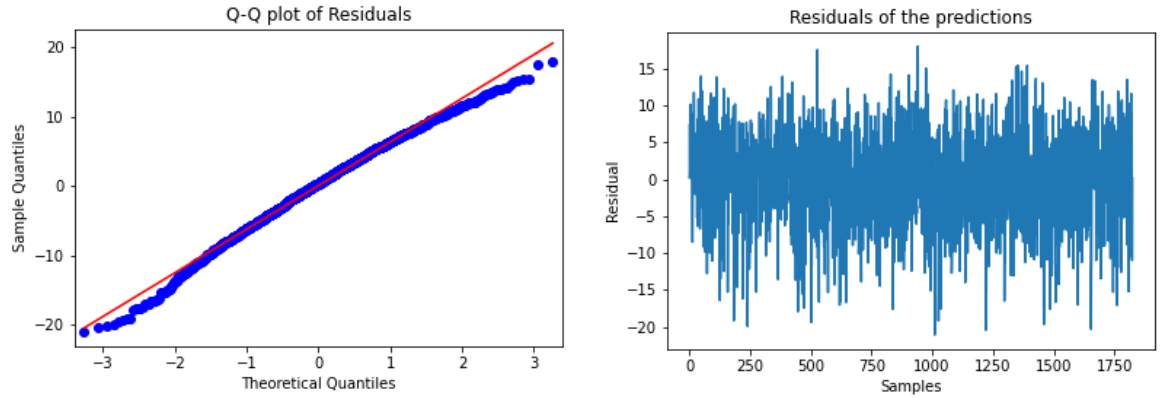

Figure 35 Q-Q plot and residuals of out-of-bag predictions for the scenario with the reduced dataset and no events.

Lastly, we also determined the baseline performances for the ED with and without events, presented in Tables 34-35.

Table 34 Baseline results for ED models with no events.

Baseline Prediction	MAE	RMSE	MAPE
One day prior	7.36	9.22	23.1
Average over 3 days prior	6.04	7.52	18.34
One week prior	7.05	8.88	22.13
Average over 3 weeks prior	5.82	7.37	17.70

Table 35 Baseline results for ED models with events.

Baseline Prediction	MAE	RMSE	MAPE
One day prior	7.30	9.18	23.05
Average over 3 days prior	6.24	7.69	19.08
One week prior	7.34	9.19	23.04
Average over 3 weeks prior	5.95	7.53	18.08

We see that the average over the days respectively 1, 2, and 3 weeks ago provide the best performance. In contrast to the baseline results of the GP models, we now find that these are similar to those of the random forest model. This suggests that the random forest model is only able to slightly perform better than simple prediction methods.

6.3 Discussion of the Results

In the previous two sections, we demonstrated the main findings of our experimental design. In this section, we will take a closer look at the results and compare them with other findings we discussed in our literature review. Besides we will also discuss what we could've done differently and how the models should improve before they can be useful.

To our knowledge, this research was the first time in which a random forest regression model was used with GP-post/ED connected data, pollen data, event data, and holidays in The Netherland and Germany. Some literature addresses random forests as a means for predicting patients, whilst using weather and time-related variables. This research is therefore an extension of the available literature on random forest models for acute care prediction with external predictors.

Starting with the results for the ED models we see that the performance is quite bad. The out-of-bag score is varying between (0.10 and 0.13). This suggests that the predictors do not work well for the predictions in the ED setting. We see this back in the performance on the MAE, RMSE, and MAPE. Looking at the RMSE, which is the metric we used to sort the models, we find values ranging from 6.29 to 6.52. Although that does not look too high, we have to keep in mind that the ED in our research only sees 33 patients per day on average. We also found that the performance of the models was only slightly better than the simple baseline prediction results, which suggests that the complex model is not a great addition. The models for the GP-post are better and we find out-of-bag scores of 0.96 for all of the models. This indicates that the predictors do a better job in the GP-post setting in comparison with the ED. The RMSE for these models is still quite high, they range from 11.44 to 11.83 but the GP-post in this research sees on average 82 visitors per day (48 on weekdays and 167 on weekends). These models did however show a great improvement in comparison with the baseline predictions, which suggests that a more complex model is better for prediction.

Comparing our results with our findings in the literature research it seems strange that our performance seems to lack in comparison to the performance of the other two papers that applied RF models (Volmer, et al., 2020) (Nas & Koyuncu, 2019). It could be that the ED and GP-post considered in our setting was too small or the data (and utilization of the data) was not sufficient to

get an accurate model. The EDs in their settings had arrivals of 75, 106, and 208 patients per day. Which is significantly bigger than our ED. However, in comparison with a similar study for the same GP-post and ED, we find results that are somewhat equal in performance. A hybrid model of SARIMAX and Gradient Tree boosting found MAE, RMSE, and MAPE of 5.25, 6.56, and 16.5% for the ED and 9.09, 13.19, and 13.26% for the GP-post (Ibrahim, 2019). Our models were slightly better but the difference is neglectable. This suggests that the model might not be the problem, but that the available data is simply not able to perform better.

Looking at the important features we do not match the findings of the expert opinions in the acute care domain. They've stated that schedules are adjusted for some known big events but we do not find any improvements in the models with events, on the contrary, they are even slightly worse. Although a comparison between the models is not completely fair as the period is not equal for these models. The decrease in model performance could be caused by the lower amount of data, or that these years were simply less predictable. Additionally, experts have also stated that on warm days they can see an increase in demand in the facilities. Although the average temperature does seem to be in the top 10 of some of the models, the contribution is rather small. Lastly, We also did not find any significant influence of the pollen data on the prediction models. While that was also suggested by experts as a potential predictor of crowding.

We do find some similarities with predictors which we found in the literature. We mentioned temperature shortly above, but other weather-related variables have appeared mostly in the ED models. This suggests that they do explain some of the variability although it is very small. This is in agreement with some of the research that we have found (Tai, Lee, Shih, & Chen, 2007) (Calegari, et al., 2016). We also find that date-related variables such as holidays and weekdays play an important role in the models for the GP-post. This is in agreement with research done by others (Hofer & Saurenmann, 2017) (Calegari, et al., 2016) (Weiss, Rogers, Maas, Ernst, & Todd, 2014). Although the biggest contribution is easily explained by the fact that the GP-post is opened longer on weekends and holidays. It was therefore more or less expected that this would be the biggest predictor for these models. The daily visitors on the weekends are significantly higher than on weekdays. Lastly, the German holidays did not seem to play a significant role in any of the models. Although many Germans seem to visit the Netherlands on their holidays it does not seem to contribute to a significant effect on the acute care demand.

We've focused on one-day-ahead daily visitor predictions. The idea behind this was that the planners could still adjust the rosters of personnel for the next day. Predicting for different times on the day, was not possible due to the available data. Predictions over a longer period, for example, one week could've been possible. However, one loses the ability to adapt daily schedules as we'd lose information about which day would be the busiest. The predictions are mostly based on data that occurred one day before the prediction. It may be that some of these predictors don't show their full potential based on their 1-day effect only. For some variables such as weather-related, it would not make sense that they'll affect the demand one week later. However, for some types of events, it could be that the demand for care arises a few days later. Unfortunately, we didn't look into this aspect.

The current prediction tool provides the GP-post and ED with predictions that are not good enough to base rosters for personnel on it. As we saw the prediction interval on the predictions for both models was quite large. Considering that our ED and GP-post only see 33 and 82 (48 on weekdays and 167 on weekends and holidays) patients on average per day, the intervals shouldn't be too wide. When we provide the facilities with a prediction that has an interval of 10-15 patients, this would have no added value since that would be almost half the daily capacity for the ED. To have added

benefits the models would have to improve significantly. Since the daily visitors vary between the ED and GP, they would require different standards. It's difficult to give a definitive number for how much the models should improve. It's unknown how much personnel is currently used to deal with the patients. However, it is safe to say that the performance of the models should be such that it is clear for planners that extra personnel is required for a certain day, thus the models should be able to produce predictions with prediction intervals that correspond with -1 or +1 extra personnel. Any predictions that are more uncertain than that would not be able to function well as a basis for roster planning.

7 Conclusion and Recommendations

In this chapter, we will give the conclusions that were acquired during this research. Additionally, we will provide the limitations of this research. Lastly, we will give useful recommendations for further research.

7.1 Conclusion of the research

In chapter 1 we introduced the main research question and the corresponding sub-questions of this research. In this section, we will present the main findings of our research and answer the research question. The research question that we formulated in chapter 1 was the following:

> ***What machine learning model can be used as an adequate early warning system for overcrowding and what is its performance in the acute care domain in the region of Oost-Achterhoek?***

We were able to build several random forest regression models that predict the daily number of visitors for the GP-post and the ED in the region of Oost-Achterhoek. These models were built using daily visitor data from the acute care domain, date-related data, data related to german and dutch holidays, pollen data, events data, and weather-related data. These models were validated with two techniques; the bootstrap method and cross-validation. With both these validations techniques, a total of four models were created for the ED and GP-post. The best model found for the GP-post and the ED are summarized in Tables 36-37 for both of the validation methods.

Table 36 Best Models out of all scenarios found by the Bootstrap method. The first entry is the best GP-post model and the second is the best ED model.

Dataset	Num of Trees	Max Features	Min Samples	Sample Size	Run Time (s)	MAE	RMSE	MAPE	OOB Score
GP Reduced (No Event)	1000	20	3	0.9	4.69	8.32	11.44	12.49	0.96
ED Reduced (No Event)	1000	5	5	0.8	2.05	5.02	6.29	15.94	0.13

Table 37 Best models out of all scenarios found by the cross-validation method. The first entry is the best GP-post model and the second is the best ED model.

Dataset	Num of Trees	Max Features	Min Samples	Run Time (s)	MAE	RMSE	MAPE
GP Reduced (No Event)	1000	29	5	4.12	8.39	11.51	12.57
ED Full (No Event)	1000	22	2	5.12	5.01	6.29	15.86

The models for the GP-post were mostly explained by the time-related features. With one in particular which is whether the day is on the weekend or not. Then of less importance are features such as the current day, whether it is a holiday or not, the number of GP-post visitors of yesterday, and some weather-related features. The models for the ED were mostly explained by the weather-related-features, albeit not so much, as the OOB score indicates.

In its current state, the one-day-ahead forecasts produced by the best models that we found will not be an adequate early warning system for overcrowding, since the degree of uncertainty is too large. The range of the predictions still varies too much to be used for employee schedules. The models for the GP-post were able to outperform baseline predictions quite significantly, however, this was not the case for the ED models. Further research and improvements, as well as newer data, are required

to improve the performance of the models and decrease the prediction uncertainty before it can be used for personnel planning.

7.2 Limitations of the research

- The data that we used during this research was limited, which resulted in the fact that we had different datasets each spanning a different period. We were also not able to obtain any new data from the GP-post or ED. The data from the GP-post was recorded over 2013-2017, whilst the data from the ED was recorded over 2012-2018. Since we used each other as predictors for the models this meant that we lost the years 2012 and 2018 from the ED dataset. Unfortunately, the Events data was also only available over 2016-2019, which meant that we could only use the data from 2016-2017.

- The pollen data were collected in the Elkerliek hospital in Helmond. This means that they may not be representable for the same amount of pollen in the air in the region of Winterswijk.

- The Acute care domain in this research was first introduced as the ED, GP-post, and emergency services of Winterswijk. Unfortunately, the last one was not looked further into for the models as no data was obtainable due to the Covid-19 pandemic. Some potential valuable information may be lost by this.

- There are also limitations to the method that was used to create the model. A predefined algorithm in Python was used, which meant that some utilities were used as they were programmed by default. For example, a random forest model usually does not require data transformation for categorical variables, however, the algorithm that was used did require numerical values.

7.3 Recommendations for further research

- The ED and GP-post should try to increase the amount of data they collect. Currently, most of the measures for crowding in literature could not be calculated for them. Registering for example the available number of beds or number of present employees allows the calculation of occupancy rate, which is commonly used as a metric for crowding. Additionally, the registration of scores like the NEDOCS, ICMED, or EDWIN might provide more information about their overall daily crowding.

- In this research, we forecasted the daily volume of visitors to the ED and GP-post for the next day. Doing this we mainly used predictors of the day prior, which may not be a good representation of what will happen the next day. This could be extended by also including certain measures from 2,3, et cetera days ago for some of the variables. In addition to the above, it may also be interesting to log certain measures more frequently throughout the day, such that models could be created that try to identify the crowding on the same day.

- Further improvements may be obtained by including extra data to the model. In this research, we also looked at the flu season as a predictor and the volume of cars on the roads. However, we were not able to obtain relevant data on these matters, due to monetary reasons and lack of available data. Perhaps that further cooperation with Nivel or Rijkswaterstaat could help in sharing data or start collecting new data.

- The plan was to also include ambulance data from the start, but due to reasons mentioned before this was no longer an option. We found that the daily visitors of the day before to the GP was a predictor in both the GP and ED models. A similar relation might be possible for ambulance data, which could be beneficial for the models. The inclusion of ambulance data would also mean that models could be made for them. For that reason, the second recommendation does count for them as well, it might be beneficial to log certain measures throughout the day to try and capture crowding differences within a day.

- In this research, we chose to predict the crowding with random forest regression. One of the reasons was that the model is easy to understand for people unfamiliar with the subject, it has no 'black-box' mechanic. Perhaps that some other methods like neural networks or support vectors regression which are somewhat less understandable can provide an additional gain in performance while losing some of the information about which variables are important. However, it is suggested to do this in combination with the recommendations above. It would not be advised to do additional research with the current datasets, as two studies have been conducted with this data and the performance seems to be similar and not sufficient.

References

Bernstein, S. L., Verghese, V., Leung, W., Lunney, A. T., & Perez, I. (2003). Development and Validation of a New Index to Measure Emergency Department Crowding. *Academic Emergency Medicine 10 (9)*, 938-942.

Bishop, C. M. (2009). *Pattern Recognition and Machine Learning.* New York: Springer.

Boyle, A., Coleman, J., Sultan, Y., Dhakshinamoorthy, V., O'Keeffe, J., Raut, P., & Beniuk, K. (2015). Initial validation of the International Crowding Measure in Emergency Departments (ICMED) to measure emergency department crowding. *Emerg Med*(32), 105-108.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45,5-32.

Calegari, R., Fogliatto, F. S., Lucini, F. R., Neyeloff, J., Kuchenbecker, R. S., & Schaan, B. D. (2016). Forecasting Daily Volume and Acuity of Patients in the Emergency Department. *Computational and Mathematical Methods in Medicine*, vol. 2016, Article ID 3863268, 8 pages.

Chan, Y. H. (2003). Biostatistics 104: Correlational Analysis. *Singapore Medical Journal*, 44(12): 614-619.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support Vector Regression Machines. *Advances in Neural Information Processing Systems 9*, 155-161.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4): 367-378.

Griep centraal: weekcijfers en meer – Nivel Zorgregistraties Eerste Lijn. (2020). Retrieved from Nivel: https://www.nivel.nl/nl/nivel-zorgregistraties-eerste-lijn/griep-centraal-weekcijfers-en-meer

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of Statistical Learning.* Springer.

Hofer, K. D., & Saurenmann, R. K. (2017). Parameters affecting length of stay in a pediatric emergency department: a retrospective observational study. *European journal of pediatrics 176(5)*, 591-598.

Hwang, U., Richardson, L., Livote, E., Harris, B., Spencer, N., & Sean, M. R. (2008). Emergency department crowding and decreased quality of pain care. *Academic Emergency Medicine 15(12)*, 1248-1255.

Hyndman, R. J., & Athanasopolous, G. (2018). *Forecasting principles and practice.* Otexts.

Ibrahim, A. (2019). *Forecasting Patient Demand and Predicting Inpatient Admission via Machine Learning Techniques in Acute Care Domain.* Enschede: University of Twente.

Ineen. (2019). *Benchmark Huisartsenposten 2018.* Ineen.

Jones, S. S., Thomas, A., Evans, S. R., Welch, S. J., Haug, P. J., & Snow, G. L. (2008). Forecasting daily patient volumes in the emergency department. *Academic emergency medicine*, 15(2):159-70.

Karlik, B., & Olgac, A. V. (2010). Performance Analysis of Various Activation Functions in Generalized MLP Architectures of Neural Networks. *International Journal of Artificial Intelligence and Expert Systems*, 1(4): 111-122.

Keizer, E., Giesen, M.-J., van de Pol, J., Knoben, J., Wensing, M., & Giesen, P. (2018). Drukte op de HAP doour ouders met jonge kinderen. *Huisarts en wetenschap 61*, 34-38.

Khaldi, R., El Afia, A., & Chiheb, R. (2019). Forecasting of weekly patient visits to emergency department: a real case study. *Procedia Computer Science*, 148:532-541.

Kingsford, C., & Salzberg, S. L. (2008). What are decision trees? *Nature Biotechnology*, 26(9): 1011-1013.

Kremers, M. N., Nanayakkara, p. W., Levi, M., Bell, D., & Haak, H. R. (2019). Strengths and weaknesses of the acute care systems in the United Kingdom and the Netherlands: what can we learn from each other? *BMC emergency medicine*, 19(1):40.

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. *Neural Information Processing Systems. 26*.

Marcilio, I., Hajat, S., & Gouveia, N. (2013). Forecasting daily emergency department visits using calendar variables and ambient temperature readings. *Academic emergency medicine*, 20(8):769-777.

Martínez-Muñoz, G., & Suárez, A. (2010). Out-of-bag estimation of the optimal sample size in bagging. *Pattern Recognition Volume 43 Issue 1*, 143-152.

Nas, S., & Koyuncu, M. (2019). Emergency Department Capacity Planning: A Recurrent Neural Network and Simulation Approach. *Computational and Mathematical Methods in Medicine*, 13 pages.

Nederlandse Zorgautoriteit. (2019). *Monitor acute zorg 2018.* NZA.

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How Many Trees in a Random Forest? *Machine Learning and Data Mining in Pattern Recognition*, 154-168.

Ospina, M. B., Bond, K., Schull, M., Innes, G., Blitz, S., & Rowe, B. H. (2007). Key indicators of overcrowding in Canadian emergency departments: a Delphi study. *Canadian Journal of Emergency Medicine Volume 9(5)*, 339-346.

Pines, J. M., & Hollander, J. E. (2008). Emergency department crowding is associated with poor care for patients with severe pain. *Annals of Emergency Medicine 51(1)*, 1-5.

Probst, P., & Boulesteix, A.-L. (2018). To Tune or Not to Tune the Number of Trees in Random Forest. *Journal of Machine learning research 18*, 1-18.

Probst, P., Wright, M., & Boulesteix, A.-L. (2019). Hyperparameters and Tuning Strategies from Random Forest. *Wires Data Mining and Knowledge Discovery volume 9 issue 3*.

Rauch, J., Hübner, U., Denter, M., & Babitsch, B. (2019). Improving the prediction of Emergency Department Crowding: A time Series Analysis Including Road Traffic Flow. *Studies in health technology and informatics*, 260:57-64.

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, 5(4): 13-22.

Sklearn Ensemble RandomForestRegressor. (2020). Retrieved from Scikit-learn: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

Tai, C.-C., Lee, C.-C., Shih, C.-L., & Chen, S.-C. (2007). Effects of ambient temperature on volume, specialty composition and triage levels of emergency department visits. *Emergency Medical Journal 24 (9)*, 641-644.

van der Linden, C. M., van Loon, M., Gaakeer, M. I., Richards, J. R., Derlet, R. W., & van der Linden, N. (2018). A different crow, a different crowding level? The predefined thresholds of crowding scales may not be optimal for all emergency departments. *International Emergency Nursing 41*, 25-30.

van der Linden, C., Reijnen, R., Derlet, R. W., Lindeboom, R., van der Linden, N., Lucas, C., & Richards, J. R. (2013). Emergency department crowding in The Netherlands: managers' experiences. *International Journal of Emergency Medicine 6 (41)*.

van der Linden, C., Reijnen, R., Derlet, R., Lindeboom, R., van der Linden, N., Lucas, C., & Richards, J. R. (2014). Drukte op de Spoedeisende Hulpafdelingen in Nederland: Ervaringen van verpleegkundig managers. *Triage*.

van der Linden, N., van der Linden, C. M., Richards, J. R., Derlet, R. W., Grootendorst, D. C., & van den Brand, C. L. (2016). Effects of emergency department crowding on the delivery of timely care in an inner-city hospital in the Netherlands. *European Journal of Emergency medicine 23(5)*, 337-343.

van Loghum, S. B. (2013). Drukte op de SEH. *Critical Care*, 10,4.

Volksgezondheidenzorg. (2020, March 14). *Acute Zorg - Regionaal en internationaal - Ambulancezorg*. Retrieved from volksgezondheidenzorg: https://www.volksgezondheidenzorg.info/onderwerp/acute-zorg/regionaal-internationaal/ambulancezorg

Volmer, M. A., Glampson, B., Mellan, T. A., Mishra, S., Mercuri, L., Costello, C., . . . Bhatt, S. (2020). A unified machine learning approach to time series forecasting applied to demand at emergency departments.

Wang, H., Robinson, R. D., Bunch, K., Huggins, C. A., Watson, K., Jayswal, R. D., . . . Zenarosa, N. R. (2014). The inaccuracy of determining overcrowding status by using the national ED overcrowding study tool. *The American journal of emergency medicine*, 32(10):1230:1236.

Weiss, S. J., Derlet, R., Arndahl, J., Ernst, A. A., Richards, J., Fernández-Frankelton, M., . . . Todd, N. G. (2004). Estimating the Degree of Emergency Department Overcrowding in Academic Medical Centers: Results of the National ED overcrowding Study (NEDOCS). *Acad Emerg Med. 11*, 38-50.

Weiss, S. J., Rogers, B. D., Maas, F., Ernst, A. A., & Todd, N. G. (2014). Evaluating community ED crowding: the Community ED Overcrowding Scale Study. *American Journal of Emergency Medicine 32*, 1357-1363.

Zlotnik, A., Gallardo-Antolín, A., Alfaro, M. C., Pérez Pérez, M. C., & Martínez, M. M. (2015). Emergency Department Visit Forecasting and Dynamic Nursing Staff Allocation using Machine Learning Techniques with Readily Available Open-Source Software. *Computers, Informatics, Nursing*, 33(8): 368-377.

A Methods to measure crowding in ED

NEDOCS

The NEDCOS is based on five variables and two known parameters of the ED. The formula used to determine the score is presented below (5) and the different outcomes are presented in Table 38.

$$NEDOCS = 85.8 * \left(\frac{C}{A}\right) + 600 * \left(\frac{F}{B}\right) + 13.4 * D + 0.93 * E + 5.64 * G - 20 \quad (5)$$

A = Number of ED beds
B = Number of inpatient beds
C = Number of ED patients
D = Number of critical care patients (in ED)
E = Longest ED admit (in hours)
F = Number of ED admits
G = Last door-to-bed time (in hours)

Table 38 Different levels of NEDOCS.

NEDOCS score	0-20	21-60	61-100	101-140	141-180	181>
ED situation	Not busy	Busy	Extremely busy	Overcrowded	Severely overcrowded	Dangerously overcrowded

EDWIN

The EDWIN score is evaluated based on five variables. The formula used to determine the score is presented below (6) and the different outcomes are presented in Table 39.

$$EDWIN = \sum \frac{n_i t_i}{N_a(B_t - B_a)} \quad (6)$$

n_i = number of patients present in ED triaged with urgency i
t_i = the triage category ordinal scale 1-5 (ESI reversed so 1 least acute and 5 severe)
N_a = number of attending physicians at a given time
B_t = total number of beds in ED
B_a = number of admitted patients (holds in ED)

Table 39 Different levels of EDWIN.

EDWIN score	Score < 1.5	1.5 < Score < 2	Score > 2
ED situation	Active but manageable	Busy ED	Crowded ED

ICMED

The ICMED is based on the violation of eight rules, where more violations are linked to more crowding. The eight rules are listed below:

- The ability of ambulances to offload patients (90th centile > 15 min waiting)
- Patients who leave with being seen (>=5%)
- Time until triage (>5 min after arrival)
- ED occupancy rate (>100%)
- The patient total length of stay (90th centile > 4 hours)
- Time until physician first sees patient (>30 min)
- ED boarding time (if less than 90% left the ED 2 hours after admission decision)
- Number of patients boarding in the ED (occupancy is >10% by boarders)

B Strength of correlation coefficients.

In this research, we use the guideline presented in Table 40 on the strength of the linear relationship corresponding to the correlation coefficient value. (Chan, 2003)

Table 40 Strength of linear relationship.

Correlation coefficient		**Strength of linear relationship**
+ 1	- 1	Perfect
+ 0.9	- 0.9	Very Strong
+ 0.8	- 0.8	Very Strong
+ 0.7	- 0.7	Moderate
+ 0.6	- 0.6	Moderate
+ 0.5	- 0.5	Fair
+ 0.4	- 0.4	Fair
+ 0.3	- 0.3	Fair
+ 0.2	- 0.2	Poor
+ 0.1	- 0.1	Poor
0	0	None

C Settings for models found by optimization

Results GP-post: reduced dataset no events

A total of 1260 models were created using the bootstrap validation method and a total of 210 models were created using the cross-validation method (with 5-fold). The top 10 models based on the out-of-bag score (RMSE for cross-validation) of these models are presented in Tables 41 and 42.

Table 41 Top 10 models based on bootstrap validation (reduced dataset no events).

Num of Trees	Max Features	Min Samples	Sample Size	Run Time	MAE	RMSE	MAPE	OOB Score
1000	20	3	0.9	4.69	8.32	11.44	12.49	0.96
1000	23	4	0.7	4.14	8.31	11.44	12.46	0.96
1000	25	5	0.8	4.47	8.31	11.44	12.44	0.96
1000	19	3	0.8	4.20	8.32	11.44	12.50	0.96
1000	25	2	0.7	5.20	8.33	11.44	12.49	0.96
1000	20	3	0.8	4.41	8.33	11.44	12.51	0.96
1000	27	4	0.7	4.42	8.32	11.44	12.45	0.96
1000	25	5	0.9	4.67	8.33	11.45	12.49	0.96
1000	34	5	0.7	5.05	8.33	11.45	12.44	0.96
1000	24	3	0.5	3.84	8.31	11.45	12.43	0.96

Table 42 Top 10 models based on 5-fold cross-validation (reduced dataset no events).

Num of Trees	Max Features	Min Samples	Run Time	MAE	RMSE	MAPE
1000	29	5	4.12	8.39	11.51	12.57
1000	27	5	3.90	8.38	11.51	12.58
1000	25	4	3.95	8.38	11.51	12.58
1000	22	5	3.46	8.38	11.51	12.57
1000	23	4	3.66	8.37	11.51	12.56
1000	20	4	3.37	8.38	11.51	12.59
1000	22	4	3.67	8.38	11.51	12.57
1000	26	4	4.87	8.38	11.52	12.57
1000	18	3	3.37	8.39	11.52	12.60
1000	27	4	4.14	8.39	11.52	12.58

Results GP-post: reduced dataset including events

A total of 1410 models were created using the bootstrap validation method and a total of 235 models were created using the cross-validation method (with 5-fold). The top 10 models based on the out-of-bag score (RMSE for cross-validation) of these models are presented in Tables 43 and 44.

Table 43 Top 10 models based on bootstrap validation (reduced dataset including events).

Num of Trees	Max Features	Min Samples	Sample Size	Run Time	MAE	RMSE	MAPE	OOB Score
1000	35	3	1.0	3.11	8.45	11.72	12.48	0.96
1000	38	3	1.0	3.25	8.47	11.72	12.51	0.96
1000	36	3	1.0	3.13	8.47	11.73	12.50	0.96

1000	37	3	1.0	3.17	8.45	11.73	12.48	0.96
1000	39	3	1.0	3.27	8.45	11.73	12.45	0.96
1000	34	1	0.7	3.28	8.47	11.73	12.49	0.96
1000	32	3	1.0	2.95	8.46	11.73	12.49	0.96
1000	40	3	1.0	3.36	8.45	11.74	12.44	0.96
1000	39	2	1.0	3.63	8.47	11.74	12.51	0.96
1000	42	3	1.0	3.48	8.48	11.74	12.49	0.96

Table 44 Top 10 models based on 5-fold cross-validation (reduced dataset including events).

Num of Trees	Max Features	Min Samples	Run Time	MAE	RMSE	MAPE
1000	36	2	2.64	8.48	11.67	12.53
1000	30	2	2.41	8.48	11.68	12.52
1000	38	2	2.73	8.48	11.68	12.52
1000	42	2	2.92	8.49	11.69	12.52
1000	41	2	2.89	8.49	11.69	12.52
1000	42	3	2.62	8.46	11.69	12.46
1000	41	3	2.81	8.47	11.69	12.48
1000	37	2	2.75	8.49	11.69	12.53
1000	38	3	2.49	8.46	11.69	12.46
1000	32	2	2.48	8.49	11.70	12.53

Results GP-post: full dataset no events

A total of 3000 models were created using the bootstrap validation method and a total of 500 models were created using the cross-validation method (with 5-fold). The top 10 models based on the out-of-bag score (RMSE for cross-validation) of these models are presented in Tables 45 and 46.

Table 45 Top 10 models based on bootstrap validation (full dataset no events).

Num of Trees	Max Features	Min Samples	Sample Size	Run Time	MAE	RMSE	MAPE	OOB Score
1000	71	3	0.9	14.47	8.31	11.47	12.35	0.96
1000	50	2	0.9	12.00	8.33	11.49	12.40	0.96
1000	47	1	1.0	14.44	8.32	11.49	12.39	0.96
1000	55	3	0.8	10.83	8.32	11.49	12.37	0.96
1000	66	2	0.8	14.06	8.33	11.49	12.37	0.96
1000	43	2	0.9	10.66	8.31	11.49	12.40	0.96
1000	62	2	0.8	13.31	8.33	11.49	12.39	0.96
1000	63	1	0.9	17.03	8.32	11.49	12.38	0.96
1000	72	5	0.7	10.86	8.30	11.49	12.32	0.96
1000	86	2	0.9	19.02	8.32	11.49	12.34	0.96

Table 46 Top 10 models based on 5-fold cross-validation (full dataset no events).

Num of Trees	Max Features	Min Samples	Run Time	MAE	RMSE	MAPE
1000	65	5	9.73	8.35	11.55	12.41

1000	61	5	9.24	8.36	11.56	12.42
1000	73	5	10.68	8.35	11.56	12.40
1000	63	5	9.53	8.35	11.56	12.41
1000	70	5	10.28	8.36	11.56	12.41
1000	71	5	10.81	8.35	11.56	12.40
1000	68	5	10.35	8.36	11.56	12.40
1000	67	5	9.88	8.35	11.57	12.40
1000	62	4	9.83	8.36	11.57	12.41
1000	72	5	10.59	8.37	11.57	12.41

Results GP-post: full dataset including events

A total of 3150 models were created using the bootstrap validation method and a total of 525 models were created using the cross-validation method (with 5-fold). The top 10 models based on the out-of-bag score (RMSE for cross-validation) of these models are presented in Tables 47 and 48.

Table 47 Top 10 models based on bootstrap validation (full dataset including events).

Num of Trees	Max Features	Min Samples	Sample Size	Run Time	MAE	RMSE	MAPE	OOB Score
1000	72	3	1.0	5.86	8.57	11.83	12.62	0.96
1000	71	3	1.0	5.86	8.56	11.83	12.59	0.96
1000	81	3	1.0	6.47	8.56	11.84	12.62	0.96
1000	91	3	1.0	7.06	8.60	11.84	12.65	0.96
1000	74	3	1.0	6.06	8.54	11.84	12.59	0.96
1000	84	3	1.0	6.58	8.60	11.84	12.64	0.96
1000	74	2	1.0	6.97	8.58	11.84	12.61	0.96
1000	63	2	1.0	5.91	8.58	11.85	12.60	0.96
1000	78	3	1.0	6.28	8.57	11.85	12.60	0.96
1000	94	3	1.0	7.20	8.62	11.85	12.66	0.96

Table 48 Top 10 models based on 5-fold cross-validation (full dataset including events).

Num of Trees	Max Features	Min Samples	Run Time	MAE	RMSE	MAPE
1000	79	3	4.82	8.58	11.76	12.67
1000	81	3	4.91	8.59	11.76	12.67
1000	76	3	4.73	8.59	11.76	12.68
1000	89	3	5.24	8.58	11.76	12.65
1000	84	2	5.58	8.60	11.76	12.70
1000	82	2	5.54	8.59	11.77	12.70
1000	67	2	4.74	8.59	11.77	12.70
1000	94	3	5.51	8.61	11.77	12.69
1000	78	3	4.82	8.60	11.77	12.68
1000	65	2	4.67	8.60	11.77	12.69

Results ED: reduced dataset no events

A total of 1260 models were created using the bootstrap validation method and a total of 210 models were created using the cross-validation method (with 5-fold). The top 10 models based on the out-of-bag score (RMSE for cross-validation) of these models are presented in Tables 49 and 50.

Table 49 Top 10 models based on bootstrap validation (reduced dataset no events).

Num of Trees	Max Features	Min Samples	Sample Size	Run Time	MAE	RMSE	MAPE	OOB Score
1000	5	5	0.8	2.05	5.02	6.29	15.94	0.13
1000	5	3	0.7	2.03	5.02	6.29	15.93	0.13
1000	4	4	0.7	1.80	5.03	6.29	15.95	0.13
1000	4	3	0.8	1.95	5.02	6.30	15.93	0.13
1000	5	4	0.6	1.84	5.03	6.30	15.94	0.13
1000	4	4	0.6	1.73	5.03	6.30	15.96	0.13
1000	4	5	1.0	1.98	5.03	6.30	15.95	0.13
1000	3	3	0.7	1.94	5.03	6.30	15.96	0.13
1000	3	2	0.6	1.94	5.03	6.30	15.96	0.13
1000	5	4	0.8	2.09	5.03	6.30	15.94	0.13

Table 50 Top 10 models based on 5-fold cross-validation (reduced dataset no events).

Num of Trees	Max Features	Min Samples	Run Time	MAE	RMSE	MAPE
1000	5	5	1.60	5.03	6.30	15.94
1000	5	4	1.59	5.03	6.30	15.93
1000	7	4	1.87	5.03	6.30	15.93
1000	6	4	1.74	5.03	6.30	15.94
1000	5	3	1.68	5.03	6.30	15.93
1000	8	5	1.91	5.03	6.30	15.93
1000	8	4	1.98	5.03	6.30	15.93
1000	9	4	2.13	5.03	6.30	15.93
1000	4	5	1.43	5.03	6.30	15.97
1000	7	5	1.85	5.03	6.30	15.94

Results ED: reduced dataset including events

A total of 1410 models were created using the bootstrap validation method and a total of 235 models were created using the cross-validation method (with 5-fold). The top 10 models based on the out-of-bag score (RMSE for cross-validation) of these models are presented in Tables 51 and 52.

Table 51 Top 10 models based on bootstrap validation (reduced dataset including events).

Num of Trees	Max Features	Min Samples	Sample Size	Run Time	MAE	RMSE	MAPE	OOB Score
1000	5	1	0.6	1.66	5.21	6.44	16.57	0.11
1000	5	2	1.0	1.72	5.20	6.44	16.57	0.11
1000	6	5	0.5	1.38	5.18	6.44	16.52	0.11
1000	3	2	1.0	1.42	5.20	6.44	16.58	0.11
1000	5	3	0.5	1.41	5.19	6.44	16.55	0.11

1000	2	3	1.0	1.27	5.20	6.44	16.58	0.11
1000	8	2	0.7	1.67	5.20	6.44	16.55	0.11
1000	7	1	0.5	1.73	5.20	6.44	16.57	0.11
1000	5	5	0.5	1.33	5.18	6.44	16.53	0.11
1000	3	1	0.5	1.53	5.21	6.44	16.58	0.11

Table 52 Top 10 models based on 5-fold cross-validation (reduced dataset including events).

Num of Trees	Max Features	Min Samples	Run Time	MAE	RMSE	MAPE
1000	5	5	1.11	5.21	6.46	16.61
1000	3	4	1.00	5.21	6.46	16.61
1000	4	5	1.03	5.21	6.46	16.62
1000	3	5	1.02	5.22	6.47	16.63
1000	2	3	1.02	5.22	6.47	16.65
1000	4	3	1.08	5.22	6.47	16.63
1000	4	4	1.08	5.22	6.47	16.64
1000	7	5	1.19	5.22	6.47	16.62
1000	6	5	1.16	5.22	6.47	16.62
1000	5	4	1.13	5.22	6.47	16.63

Results ED: full dataset no events

A total of 3000 models were created using the bootstrap validation method and a total of 500 models were created using the cross-validation method (with 5-fold). The top 10 models based on the out-of-bag score (RMSE for cross-validation) of these models are presented in Tables 53 and 54.

Table 53 Top 10 models based on bootstrap validation (full dataset no events).

Num of Trees	Max Features	Min Samples	Sample Size	Run Time	MAE	RMSE	MAPE	OOB Score
1000	45	3	0.5	6.69	5.02	6.30	15.91	0.13
1000	31	1	0.5	6.59	5.02	6.30	15.94	0.13
1000	31	3	0.9	7.41	5.02	6.30	15.90	0.13
1000	27	1	0.6	6.84	5.02	6.30	15.92	0.13
1000	19	2	0.5	4.14	5.03	6.31	15.94	0.13
1000	23	2	0.8	6.38	5.02	6.31	15.92	0.13
1000	18	2	0.6	4.45	5.03	6.31	15.94	0.13
1000	17	2	0.6	4.39	5.03	6.31	15.95	0.13
1000	22	2	0.5	4.63	5.02	6.31	15.93	0.13
1000	22	2	0.6	5.13	5.03	6.31	15.94	0.13

Table 54 Top 10 models based on 5-fold cross-validation (full dataset no events).

Num of Trees	Max Features	Min Samples	Run Time	MAE	RMSE	MAPE
1000	22	2	5.12	5.01	6.29	15.86
1000	33	2	7.14	5.01	6.29	15.87
1000	27	2	5.93	5.01	6.29	15.87

1000	41	2	8.45	5.01	6.29	15.87
1000	18	2	4.43	5.01	6.29	15.87
1000	29	2	6.29	5.01	6.29	15.87
1000	31	2	6.65	5.01	6.29	15.86
1000	32	1	8.15	5.01	6.29	15.88
1000	19	3	4.06	5.01	6.29	15.88
1000	28	2	6.08	5.01	6.29	15.87

Results ED: full dataset including events

A total of 3150 models were created using the bootstrap validation method and a total of 525 models were created using the cross-validation method (with 5-fold). The top 10 models based on the out-of-bag score (RMSE for cross-validation) of these models are presented in Tables 55 and 56.

Table 55 Top 10 models based on bootstrap validation (full dataset including events).

Num of Trees	Max Features	Min Samples	Sample Size	Run Time	MAE	RMSE	MAPE	OOB Score
1000	8	3	0.7	1.67	5.23	6.49	16.69	0.10
1000	17	1	0.5	2.28	5.24	6.49	16.68	0.09
1000	9	1	0.7	2.19	5.25	6.49	16.71	0.09
1000	9	1	0.6	2.08	5.24	6.50	16.70	0.09
1000	11	4	0.7	1.81	5.23	6.50	16.65	0.09
1000	16	1	0.7	2.56	5.25	6.50	16.71	0.09
1000	11	3	1.0	2.09	5.24	6.50	16.68	0.09
1000	16	3	0.6	1.98	5.25	6.50	16.70	0.09
1000	14	1	0.5	2.22	5.23	6.50	16.68	0.09
1000	8	2	0.6	1.70	5.24	6.50	16.69	0.09

Table 56 Top 10 models based on 5-fold cross-validation (full dataset including events).

Num of Trees	Max Features	Min Samples	Run Time	MAE	RMSE	MAPE
1000	5	4	1.14	5.27	6.52	16.79
1000	4	2	1.27	5.28	6.52	16.83
1000	7	1	1.72	5.29	6.53	16.83
1000	2	1	1.26	5.28	6.53	16.84
1000	4	1	1.48	5.28	6.53	16.82
1000	3	1	1.41	5.29	6.53	16.85
1000	8	4	1.31	5.27	6.53	16.79
1000	5	2	1.33	5.28	6.53	16.81
1000	5	3	1.18	5.27	6.53	16.82
1000	8	3	1.41	5.28	6.53	16.79

D Out-of-bag predictions versus actual values

For the models that were validated by the bootstrap method, we were able to generate out-of-bag predictions on the dataset. In addition to these predictions, we've also added prediction intervals based on the 5th and 95th percentile of predictions over all trees in the forest. The actual values are also plotted in the Figures. Figures 36-38 contain the predictions for the GP-post and Figures 39-41 for the ED.

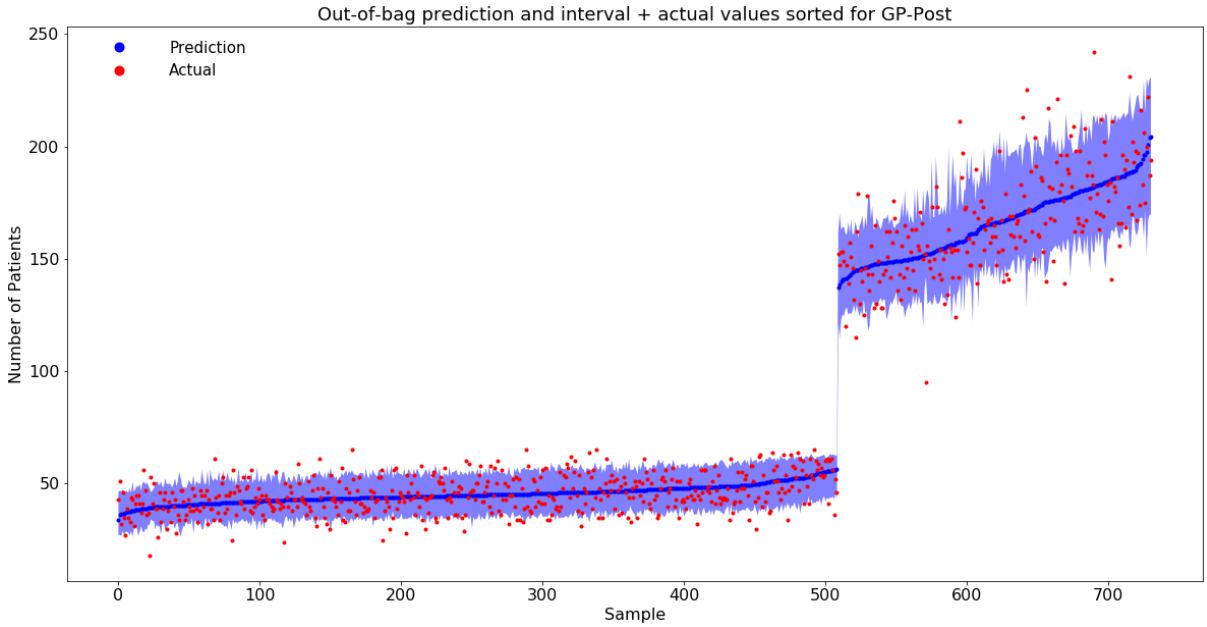

Figure 36 GP-post Out-of-bag predictions + intervals for the scenario with the reduced dataset and Events.

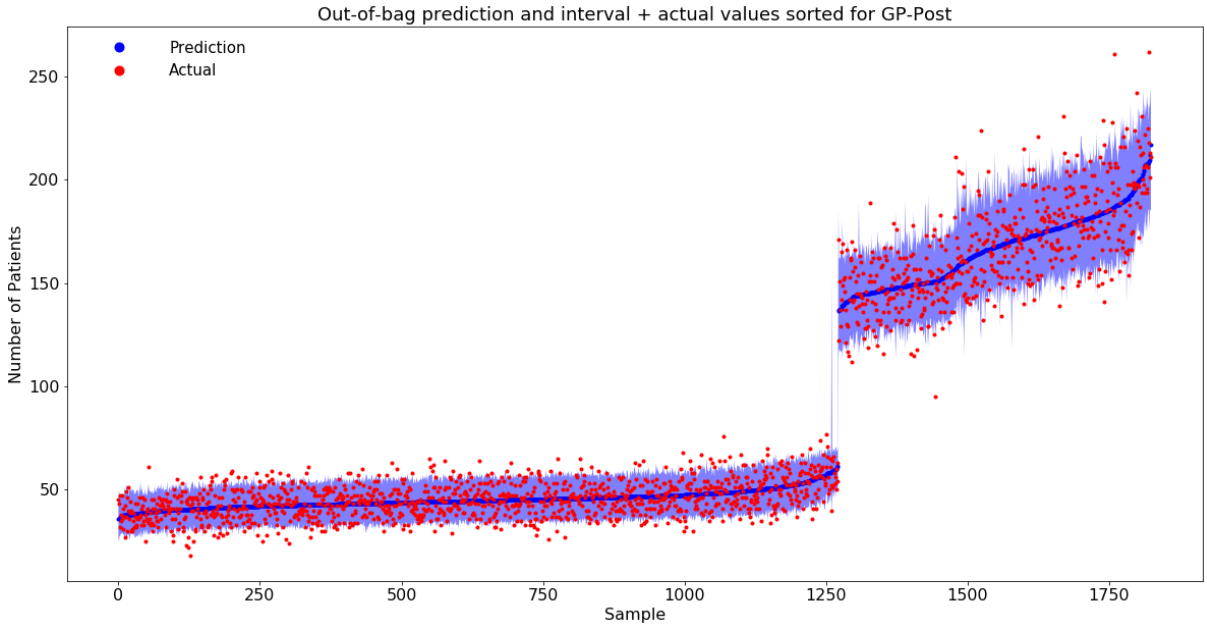

Figure 37 GP-post Out-of-bag predictions + intervals for the scenario with the full dataset and no Events.

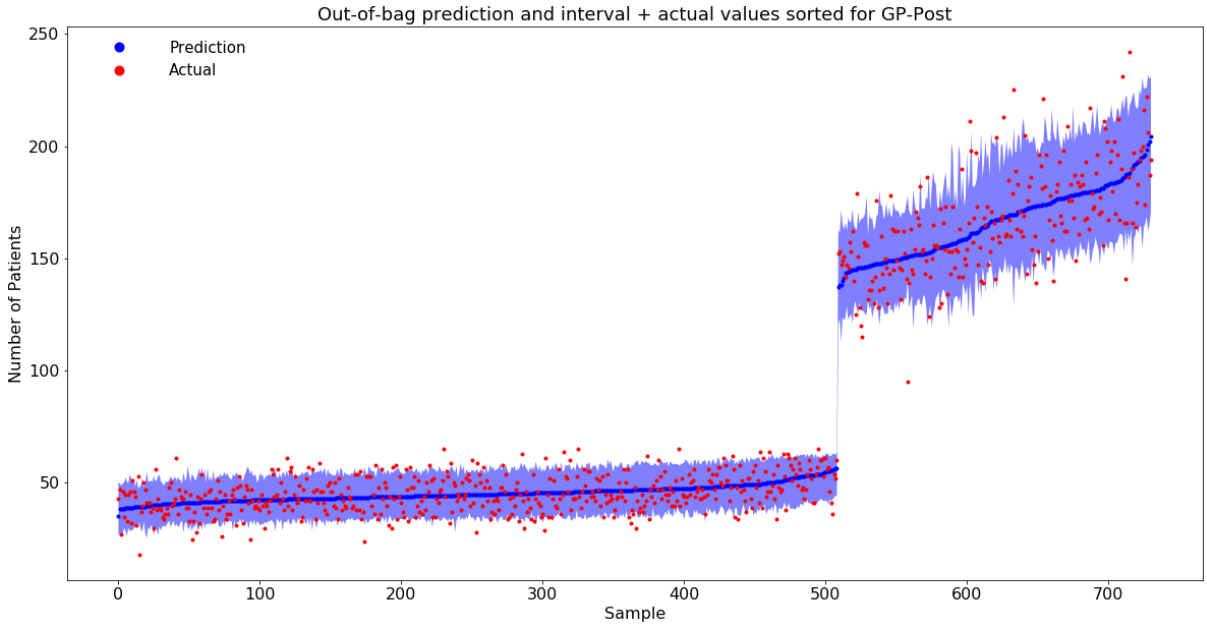

Figure 38 GP-post Out-of-bag predictions + intervals for the scenario with the full dataset and Events.

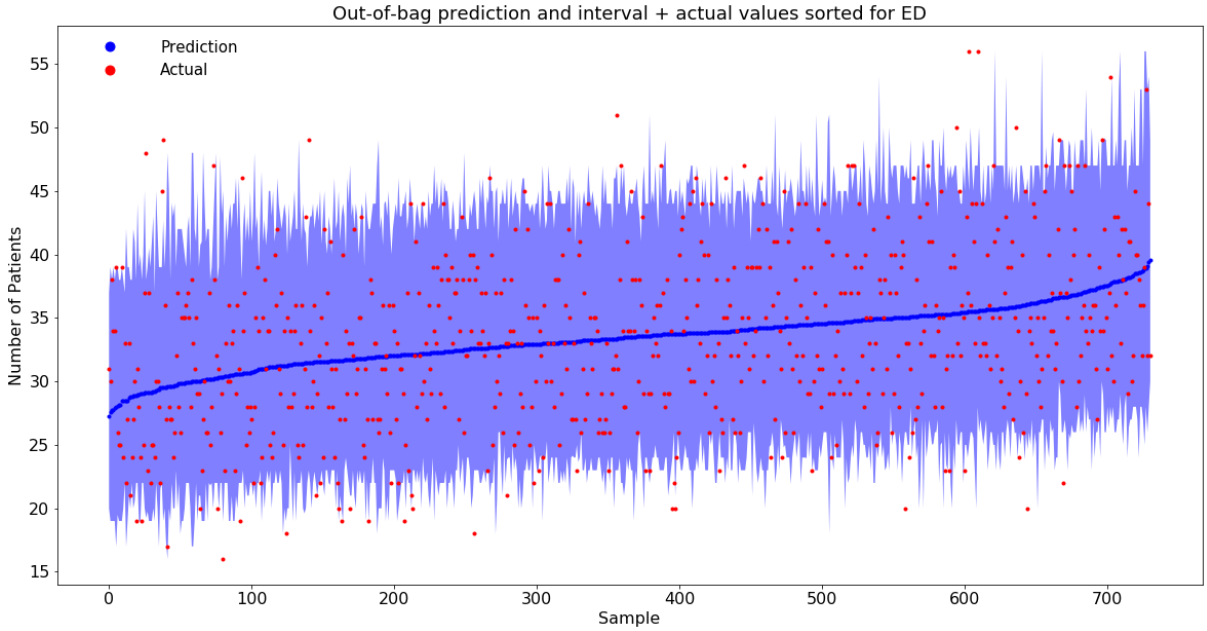

Figure 39 ED Out-of-bag predictions + intervals for the scenario with the reduced dataset and Events.

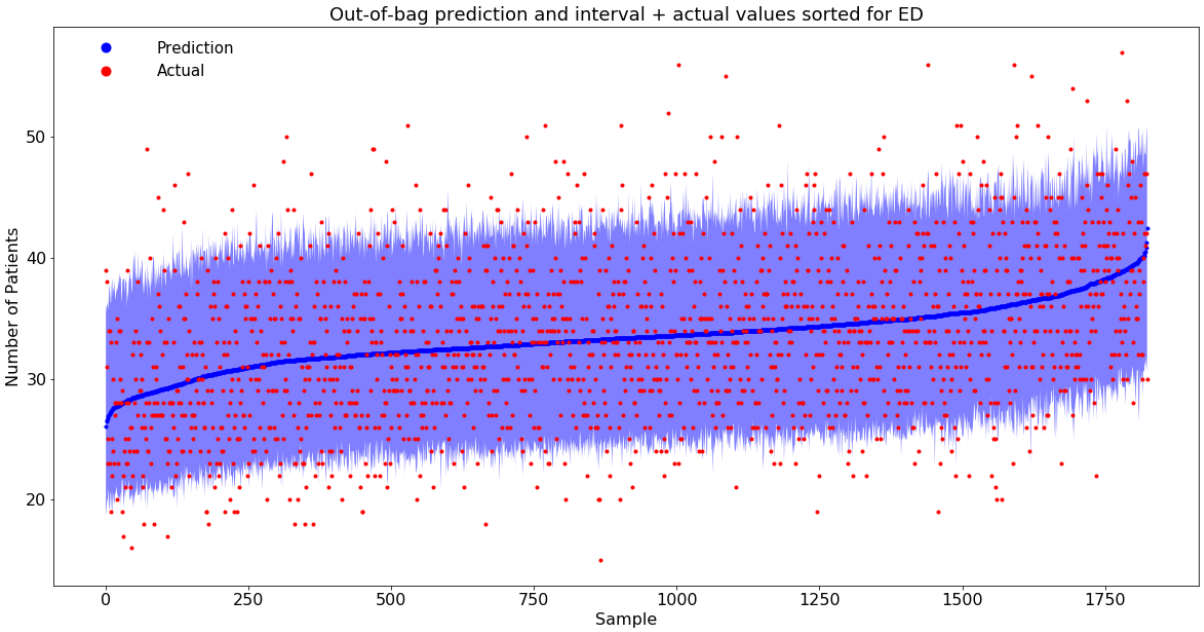

Figure 40 ED Out-of-bag predictions + intervals for the scenario with the full dataset and no Events.

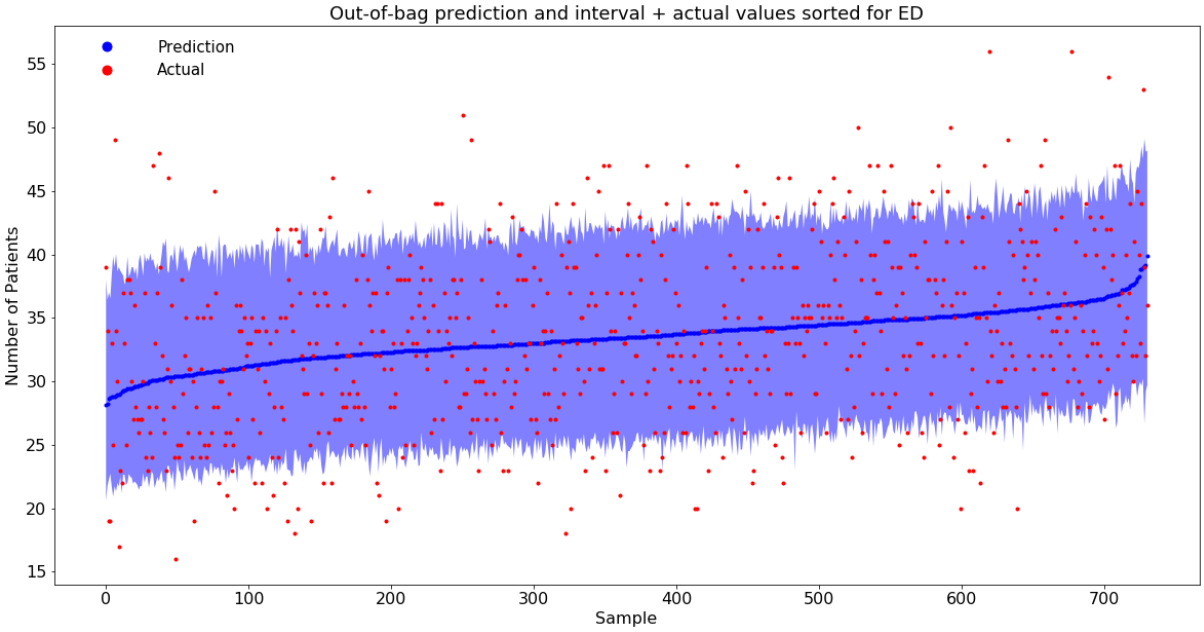

Figure 41 ED Out-of-bag predictions + intervals for the scenario with the full dataset and Events.

E Residual analysis plots of predictions

For the models that were validated with the bootstrap method, we've also created plots of the residuals and the q-q plots of these residuals testing against a normal distribution. The plots for the GP models are shown in Figures 42-44 and the plots for the ED are shown in Figures 45-47.

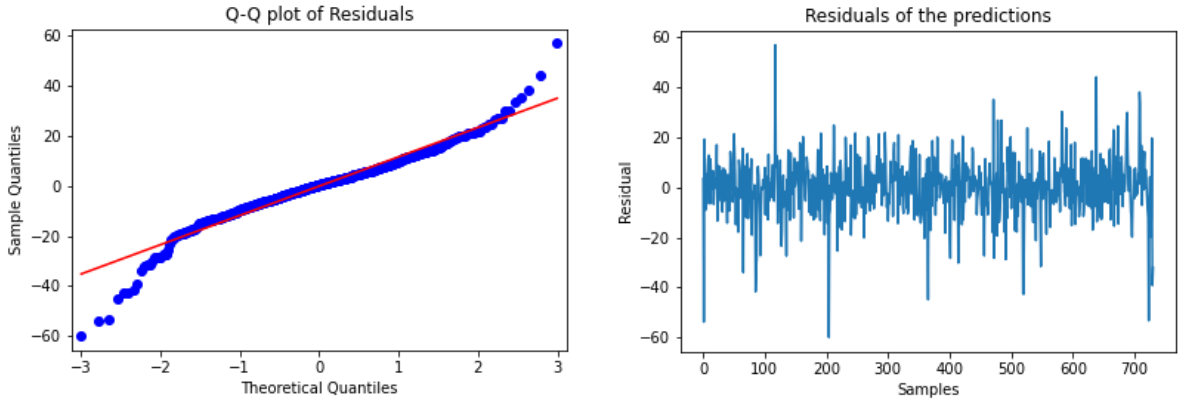

Figure 42 Q-Q plot and residuals of out-of-bag predictions for the scenario with the reduced dataset and events for the GP.

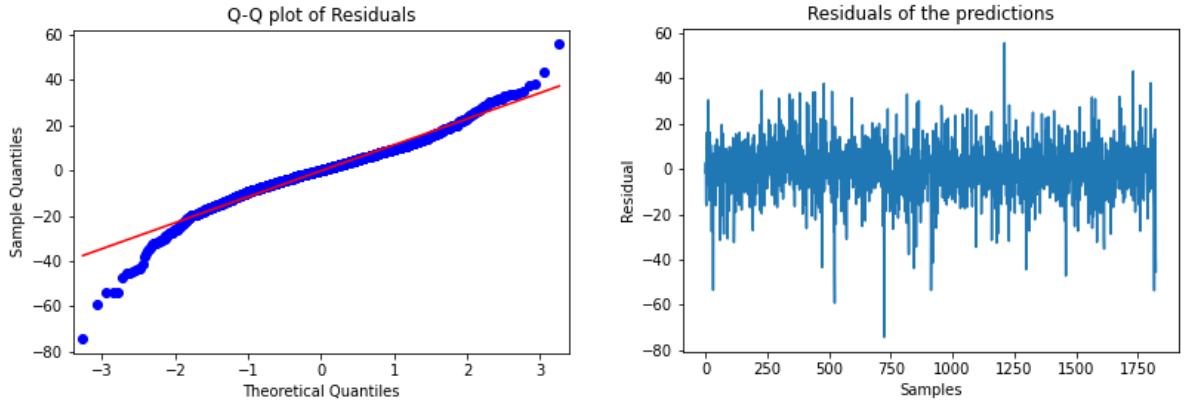

Figure 43 Q-Q plot and residuals of out-of-bag predictions for the scenario with the full dataset and no events for the GP.

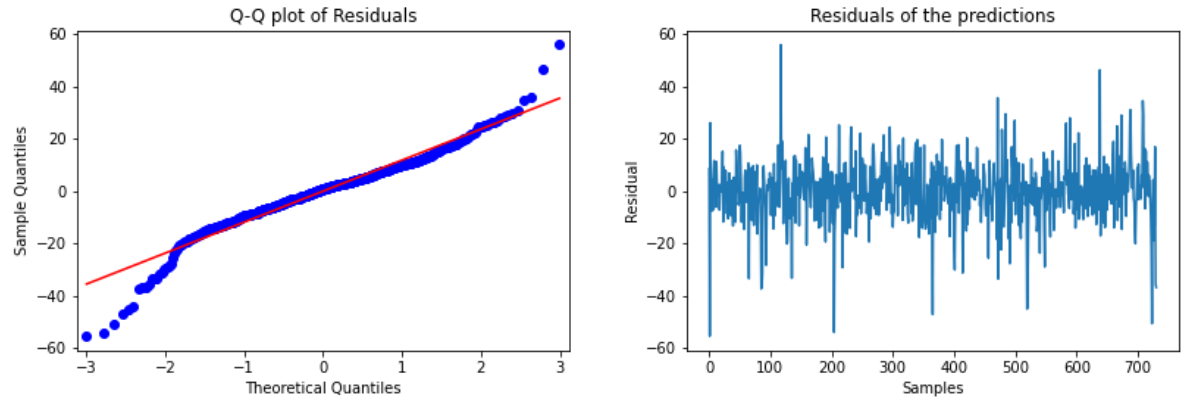

Figure 44 Q-Q plot and residuals of out-of-bag predictions for the scenario with the full dataset and events for the GP.

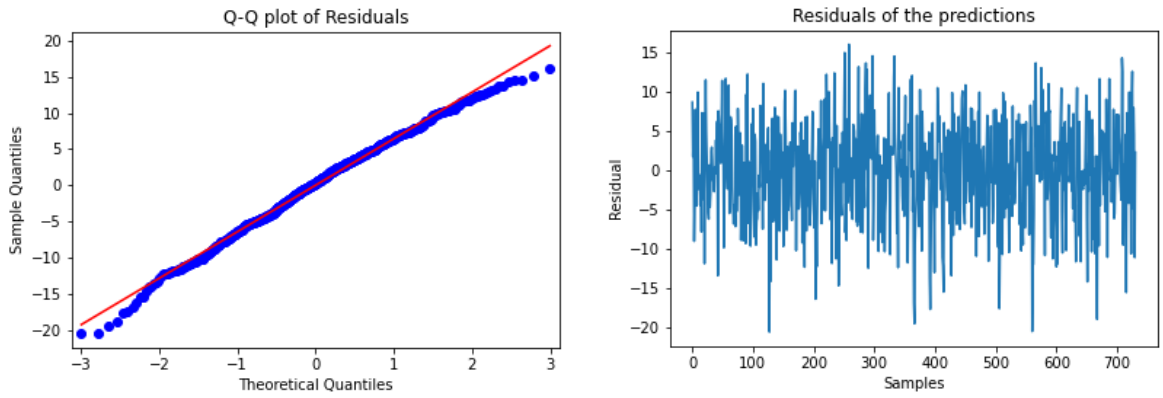

Figure 45 Q-Q plot and residuals of out-of-bag predictions for the scenario with the reduced dataset and no events for the ED.

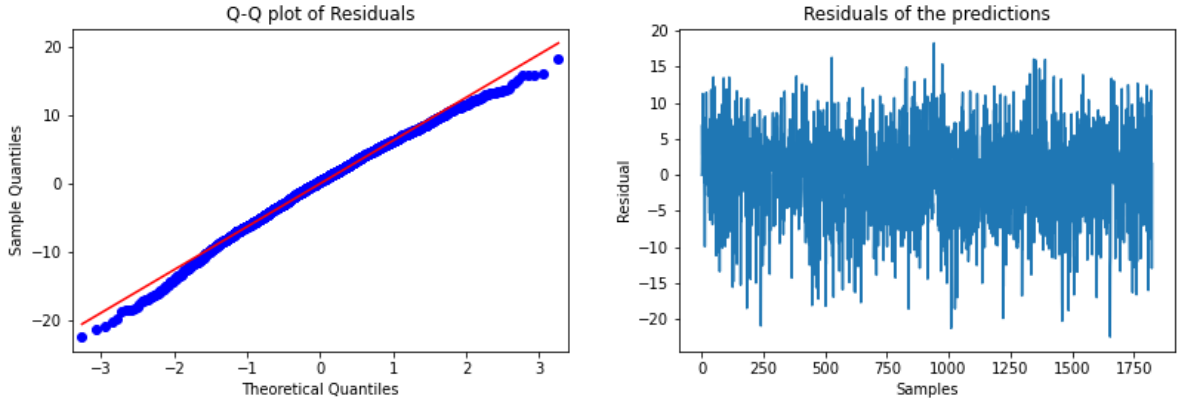

Figure 46 Q-Q plot and residuals of out-of-bag predictions for the scenario with the full dataset and no events for the ED.

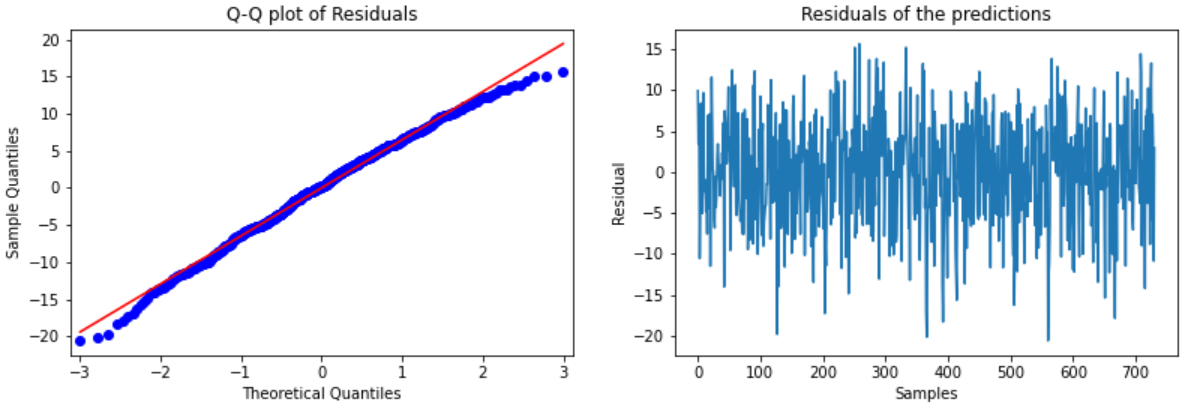

Figure 47 Q-Q plot and residuals of out-of-bag predictions for the scenario with the full dataset and events for the ED.