Master Thesis

The Impact of the Dutch Weather on the Health of Horses

J. van ‘t Padje

University of Twente
Faculty of Electrical Engineering, Mathematics and Computer Science
Data Management & Biometrics (DMB)

SUPERVISORS:
dr. M. Poel
dr. E. Mocanu
dr. C.G.M. Groothuis-Oudshoorn

22 December 2020

UNIVERSITY OF TWENTE.

ABSTRACT

Gut feeling and farm wisdom often attribute diseases in horses to specific weather conditions, which might lead to false assumptions. The goal of this research is to see if these assumptions are valid or not by answering the questions *What is the influence of the Dutch weather on the health of horses?* and *To what extend can the Dutch weather be used to predict the occurrence of colic, laminitis, respiratory disease and skin disease?*

To answer these questions the data of animal clinic Den Ham is used. This data required preprocessing. Duplicate horses are merged, measured horse temperatures are extracted and the data is grouped into consults. These consults are labelled with one or more of the previously mentioned diseases using the text description of the consult and the admitted medication. The labelling is performed with a bag of words approach using Stochastic Gradient Descent testing different classifiers, loss functions and other parameters. This data is merged with the weather data of Heino form the weather station of the KNMI (The Royal Netherlands Meteorological Institute), which needed imputation of some values and variables. The values are imputed using a k-Nearest neighbours approach. The missing variables are taken from the weather station in Hoogeveen. This weather station, most likely, has the least difference with Heino for the missing variables. Visualizations are made to find obvious correlations between the diseases and changes in the weather and to see the occurrence of the diseases over a year. To find correlations between the weather and the diseases, the weather variables are split into two groups: the weather on the days where the disease occurs and the weather variables on the remaining days. Permutation tests are performed for significance testing between the two groups of weather variables. When a significant difference is found between the weather conditions of those two groups, the weather variable is considered to be correlated to the weather variable. Predictions are made using Ensemble predictions, which are compared to four single classifiers: Logistic Regression, Support Vector Machine, Decision Tree and Neural network. The ensemble prediction methods Voting, Bagging and Boosting are tested. Voting combines Logistic Regression, Support Vector Machine, Decision Tree and Neural network. Bagging is performed once for each of these four classifiers and Boosting is performed using Decision Trees only.

The methods as described above produce the following results; The measured temperature of the horses can be obtained from the data with an accuracy of 0.99805. With ten Nearest neighbours, an R^2 score of 0.99556 is achieved for the imputation of the missing weather values. The surrounding weather stations did not have very different results for the missing variables, therefore the weather station of Hoogeveen is used as a donor for the missing variables since this weather station is closest to Heino and Den Ham. The visualizations of the changes in the weather do not show any obvious correlations. The correlation analysis does not show clear links between specific weather variables and one of the diseases. Roughly the same variables are correlated to each of the diseases. Laminitis has turned out to be the hardest to predict with an accuracy of 65%, obtained using a single Support Vector Machine or a single Neural Network. Colic and skin disease are both predicted best using the Bagging algorithm with Decision Trees with respectively 70% and 74% accuracy. The best result has been achieved for respiratory disease with an accuracy of 79.8%. This is achieved with the Voting algorithm, Bagging Support Vector Machines and with a single Support Vector Machine.

One can expect better results when better-structured veterinarian data is used since the labelling of the consults has proven to be challenging. From this data, we cannot conclude that the Dutch weather influences the health of horses. Neither is the weather a good predictor for diseases.

CONTENTS

List of Figures

List of Tables

1 INTRODUCTION

In horse care, a lot of assumptions are made regarding the causes of diseases. These assumptions are often based on gut feeling and farm wisdom. For example, white hooves are weaker than dark hoofs, so white hooves have more problems like cracks in the hooves. White horses are more prone to develop cancer. Muddy pastures cause mud fever. It would be interesting to see if it is possible to validate these assumptions regarding the health of horses, to obtain more insight into the causes of diseases.

For veterinaries, it is tempting to be guided by these assumptions. They can help a veterinarian to diagnose a horse faster but they can just as easily lead to wrong diagnoses. When the assumptions can be proven right or wrong by data analysis, they can be used more accurately. The specific data analyzes performed in this research take advantage of a wide range of Data Science and Artificial Intelligence methods, such as data statistics, data integration, pre-processing and interpolation, as well as classification and regression methods.

To do this, the data of Animal Clinic Den Ham will be used. This is a large animal clinic, treating all domestic animals, featuring a team of equine specialists. They are providing us with data of the past 21 years.

The objective of this research is to investigate the influence of the weather on the health of horses. During this research, we will focus on the influence of the weather on (1) the occurrence of colic, (2) the development of laminitis, (3) the occurrence of skin diseases and (4) the development of respiratory diseases.

This research provides a summary of related work, regarding the objectives as given above. It discusses the questions that still require an answer, as well as a method on how to address them. This is followed by the obtained results from following this method.

In section 2 a summary of studies is given, describing the impact of the weather on the health of horses. Section 3 contains unanswered questions for further research. Section 4 contains details about the methodology that will be used to: 4.2 prepare the data, 4.3 visualize the data, 4.4 find correlations between the weather and the diseases, and 4.5 predict diseases. Subsection 4.1 provides details about the data. Section 5 gives an overview of the results and findings. Section 6 gives insight into the assumptions that are made during the process and the consequences of these assumptions for this, and further research. The conclusions and answers to the research questions are given in section 7.

2 RELATED WORK

This section provides a summary of papers that look into health issues in horses related to the weather conditions.

2.1 Colic and weather

Colic is known to be the number one cause of death in horses [2]. Over the years, many researchers have investigated factors associated with an increased risk of colic. Changes in the weather are one of the factors.

An overview of studies to the correlation between weather and horse colic is given in Table 2.1. These studies range from 1970 to 2018. The data for the different studies is obtained either from one or more veterinarian practice(s) (v) or from the (stable) owner (o) of the horse. This can provide different results since not all horses, showing signs of colic, are examined by veterinarians [3].

Most of the studies find some correlation between horse colic and (specific or non-specific) weather types or changes, monthly or seasonal patterns. Despite this, some studies reviewing these papers are doubting the statistical significance of the findings [4, 5, 6]. One of the reasons for this could be the fact that most studies used data from only one or two years. Besides, the management of the horses (turnout, types of food, etc) changes due to changes in the weather and seasonal patterns, which can lead to colic as well. An example is given in [7], in which a group of horses experienced colic during a snowstorm, the horses were kept in the stable, while they normally would be turned out, but the feeding regime of the horses has not been adapted accordingly. So the management was more likely to have caused the colic than the snowstorm itself. Also, risk factors likely vary with the type of colic [8].

In the fifteen studies described above, logistic regression is used most, six times, to find correlations between the weather and the occurrence of Colic. Pearson's correlation coefficient is used three times and Spearman Correlation twice. In one of the papers, SPSS is used for statistical analysis, no further details were provided, and one used visualizations to draw their conclusions. For three of the papers, no method is given.

Temperature and barometric pressure are mentioned as possible risk factors for colic in many of the papers. Approximately half of them succeeded in proving this assumed correlation. Changes in the weather, monthly- and seasonal patterns are also often suggested being correlated to colic. These seem to be easier to prove, almost all the researchers succeeded in finding correlations. No papers were found studying a correlation between wind and colic, as suggested by a veterinary expert. For this research, it would be interesting to see if a correlation can be found with temperature and barometric pressure in this data. The correlation between colic and wind will be investigated as well.

	[9]	[10]	[7]	[3]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[19]	[20]	[21]
Publication year	’70	’92	’95	’97	’99	’01	’01	’04	’06	’08	’09	’14	’17	’17	’18
Source	v	v	o	o	v	o	o	v	v	v	v	v	o	v	v
Num of years	1	2	1	1	1	1	1	10	10	2	1	2	3	1	12
Temperature		×	×								✓	✓	✓		
Barometric Pressure			×								×	✓			✓
Humidity			✓										×		
Rainfall		×	×												
Snow			✓												
Weather changes	✓				✓		×							✓	
Months				✓				✓	✓						
Seasons						✓				✓			✓		
Pearson’s Corr. Coef.		●												●	
Logistic Regression			●		●			●	●	●		●			●
Visualization						●									
Statsitical Analysis											●				
Spearman Correlation													●		●

Table 2.1: Literature overview of relations between horse colic and the weather. For the results hold: ✓= found correlation, × = no correlation found, empty = not investigated. The source of the data is either a veterinarian practice (v) or the owner/stable (o)

2.2 Laminitis and weather

Although Polzer and Slater [22] failed to find a correlation between seasons and laminitis, the risk of developing laminitis in horses is found to be higher during the summer and winter months according to Wylie et al. [23]. In addition, Menzies-Gow et al. [24] found a positive association between the number of sunshine hours and the incidence of laminitis, this was assumed to be due to changes in the grass contents, and not the direct influence of the sunlight on the horses. No associations between rainfall or temperature and the occurrence of laminitis in horses was found by Menzies-Gow et al. [24].

Eating high sugar feed, can cause insulin resistance in horses [25, 26, 27] which is found to be associated with laminitis [28]. Overdosing insulin or oligofructose (a sweetener) can also induce laminitis in horses [29, 30]. During the day, through photosynthesis, grass produces sugars which is stored in the stems and leaves [31, 32, 33, 34]. This sugar is used by the grass to grow. The storage allows the grass to grow when photosynthesis is impaired, by shading or during the night [35, 36].

Although some grass species produce less sugar during cold periods [37], sugars are found to accumulate in the grass by low temperatures [38, 39, 40]. This phenomenon can be explained by the fact that the grass is unable to grow during cold but photosynthesis is possible.

Besides low temperatures, grass can experience stress from water deficit as well. Drought stress is another cause of sugar storage in grass [41, 42]. Silva and Arrabaca [43] showed that sudden water deficit reduced the levels of sucrose and starch in the grass, while gradual water deficit indeed raised sugar levels, except for starch. This supports the assumption of Menzies-Gow et al. [24] that grass grown under certain weather conditions can cause laminitis.

Taking this into account, an increase of laminitis can be expected during autumn and spring, when the days are warm, allowing the grass to produce sugar, and nights are cold, preventing the grass from growing and therefor using the sugars. During periods of drought, the number of cases of laminitic horses also is supposed to be higher.

Laminitis can be induced by diets with high sugar. The sugar levels in the grass will raise when the grass cannot grow, due to drought or low temperatures. In this data, the number of cases of laminitis is therefore expected to be higher when the temperatures are low at night and high during the day, and during periods of drought.

	[50]		[44]	[45]	[46]	[51]	[47]	[48]	[52]	[53]	[54]	[55]	[49]
Public. year	1976		1981	1988	1994	1996	2002	2003	2003	2005	2006	2010	2016
	EHV-1 EAV	ERV-1											
High temp.				✓	✓		✓	✓					✓
Low temp.			✓			✓				✓	✓	✓	
High humid.		✓	✓	✓	✓	✓	✓	✓					✓
Dry cond.	✓												
Winter									✓				
Spring									✓				

Table 2.2: Literature overview of relations between respiratory disease and weather conditions. For the results hold: ✓= affects, empty = not mentioned

2.3 Respiratory disease and weather

A correlation between respiratory disease and high humidity is suggested by a veterinary expert. Table 2.2 shows an overview of papers and the weather conditions mentioned by those papers, responsible for different types of respiratory diseases.

According to Sainsbury [44] damp stables in combination with low temperatures can cause respiratory problems. Warm, humid weather can worsen some respiratory disorders like: laryngeal stridor [45], tracheal collapse [46], and Inflammatory Airway Disease [47, 48]. Increasing the bronchial temperature by breathing hot, humid air can cause bronchospasm, especially when the airway was already inflamed [49]. Bullone [49] concluded that spore concentrations are higher during warm, humid weather, which can lead to irritation in the respiratory tract.

In the research of Donaldson, to the survival of airborne viruses [50], including the equine herpesvirus type 1 (EHV-1), equine arteritis virus (EAV) and the equine rhinovirus (ERV-1), only ERV-1 survived well in high humidity. It did poorly in dry conditions. The survival rate of the other viruses was the same or lower in high humidity compared to dry conditions.

In contrast, according to Robinson et al. [51] COPD (Chronic Obstructive Pulmonary Disease or equine asthma) in horses is rare in countries like California and Australia where the climate is warm and dry, while COPD is most common in Northern Hemisphere. Laurent et al. [52] investigated the risk factors of Recurrent Airway Obstruction (RAO) and concluded that the diagnosis of RAO is given more often during winter (1.6x) and spring (1.5x) compared to the summer. The occurrence of RAO in autumn was significantly less. Exercising in cold air can result in asthma-like airway disease [53] and lower airway disease [54]. Even being outdoors during the winter can increase the number of inflammatory cells [55].

Even though viruses, except for ERV-1, do not thrive well in humid conditions, horses seem to be more prone to develop respiratory diseases during humid and either hot or cold weather. Knowing this, it is expected to see more cases of respiratory disease in the data during high humidity and extreme temperatures.

2.4 Skin disease and weather

High humidity and rainfall are one of the most mentioned causes of skin diseases like fungal infections. A veterinary expert suggested a possible correlation between skin diseases and high humidity.

Fungi and bacteria are often causes of skin diseases in horses. Table 2.3 shows weather conditions for two bacteria (Dermatophilus congolensis and Staphylococcus) and two fungi (Histoplasma farciminosum and Hyphomyces destruens) that, according to the reviewed literature, are associated with skin diseases in horses.

Dermatophilus congolensis causes mud fever and rain rot or rain scald. The appearance and

	DC					HF		HD		S	
	[56]	[58]	[59]	[60]	[61]	[62]	[63]	[64]	[65]	[66]	[67]
publication year	1980	1990	1996	2005	2010	1983	2006	1978	1982	1995	2005
Rainfall	✓	✓	✓	✓				✓	✓		
Humidity		✓			✓		✓			✓	✓
Wet pastures			✓	✓							
Low temperatures				✓		✓	×				
High temperatures							✓		✓	✓	✓
Dry conditions							×				

Table 2.3: Literature overview of relations between skin disease and weather conditions. The following fungi and bacteria are considered: Dermatophilus congolensis (DC, bacterium), Histoplasma farciminosum (HF, fungus), Hyphomyces destruens (HD, fungus) and Straphylococcus (S, bacterium). For the results hold: ✓= affects, ×= does not affect, empty = not mentioned

spread of these diseases increases with rainfall [56, 57, 58, 59]. According to Hyslop [56] and Mollins [58], the skin barrier is damaged by high amounts of rain. Therefore, the intensity of rainfall is the main problem, not the annual rainfall. The mobility of the infective zoospores can be increased by rain [56]. In Israel, a herd of horses was infected with rain scald and mud fever four weeks after heavy rainfall, which led to muddy pastures. Both the heavy rainfall and the muddy pastures are associated with the onset of the disease [59]. Muddy pastures are also mentioned as a problem by White [60], in addition to autumn and winter weather which is associated with heavy rainfall. Colles et al. [61] states that the association between wet or damp conditions is plausible, but not always the case.

Gabal and Hennager [62] discovered that histoplasma farciminosum survived longer (18 weeks) at -15°C, compared to warmer temperatures, up to 26°C. This is in contrast to the findings of Armeni [63], that discovered many cases on locations with a hot, humid climate and only a few in cold, dry or windy climate. Hyphomyces destruens is considered more common during wet periods [64, 65]. Miller and Campbell [65] discovered a rise in the occurrence during heavy rainfall in the summer, this is considered to help the fungi grow. Flooding will help spread the fungi to other individuals.

As with dermatophilus congolensis, the staphylococcus benefits form weakened skin barriers. Warm humid weather compromises the skim barriers and is therefore a risk factor for contracting this bacterium [66, 67].

Skin diseases are expedited to appear more in the data after long periods of rain. High humidity and high or low temperature could also be an indication for skin diseases.

3 RESEARCH QUESTIONS

In section 2, an overview of the related work concerning the objectives of this paper is given. Based on the related work, the following questions arise which need further research.

Q1: What is the influence of the Dutch weather on the health of horses?

Based on the findings in the related work, the input of veterinary experts and the available data, the following sub-questions are constructed to answer this research question:

1.1 Does the temperature, barometric pressure and high amount of wind influence the occurrence of colic?

1.2 Is the development of laminitis dependent on stress in the grass, due to cold and drought?

1.3 Does hot, humid or cold weather worsen or induce respiratory disease?

1.4 Do skin diseases occur more in periods of heavy rainfall and high humidity?

Q2: To what extent can the Dutch weather be used to predict the occurrence of ...

a. colic?

b. laminitis?

c. respiratory disease?

d. skin disease?

4 MATERIALS AND METHODS

This section starts with an extensive overview of the two data sets that are used for this research. The data sets used for this research are the medical data of animal clinic Den Ham, containing descriptions of the consults, and the weather data of weather stations Heino of the KNMI [68]. This will be followed by the methods used to answer the research questions and the choices that have been made in this research. The methodology is split into the different stages of the research: preparation and visualization of the data, finding correlations between weather values on days with and without disease and prediction of the diseases based on the correlating weather variables.

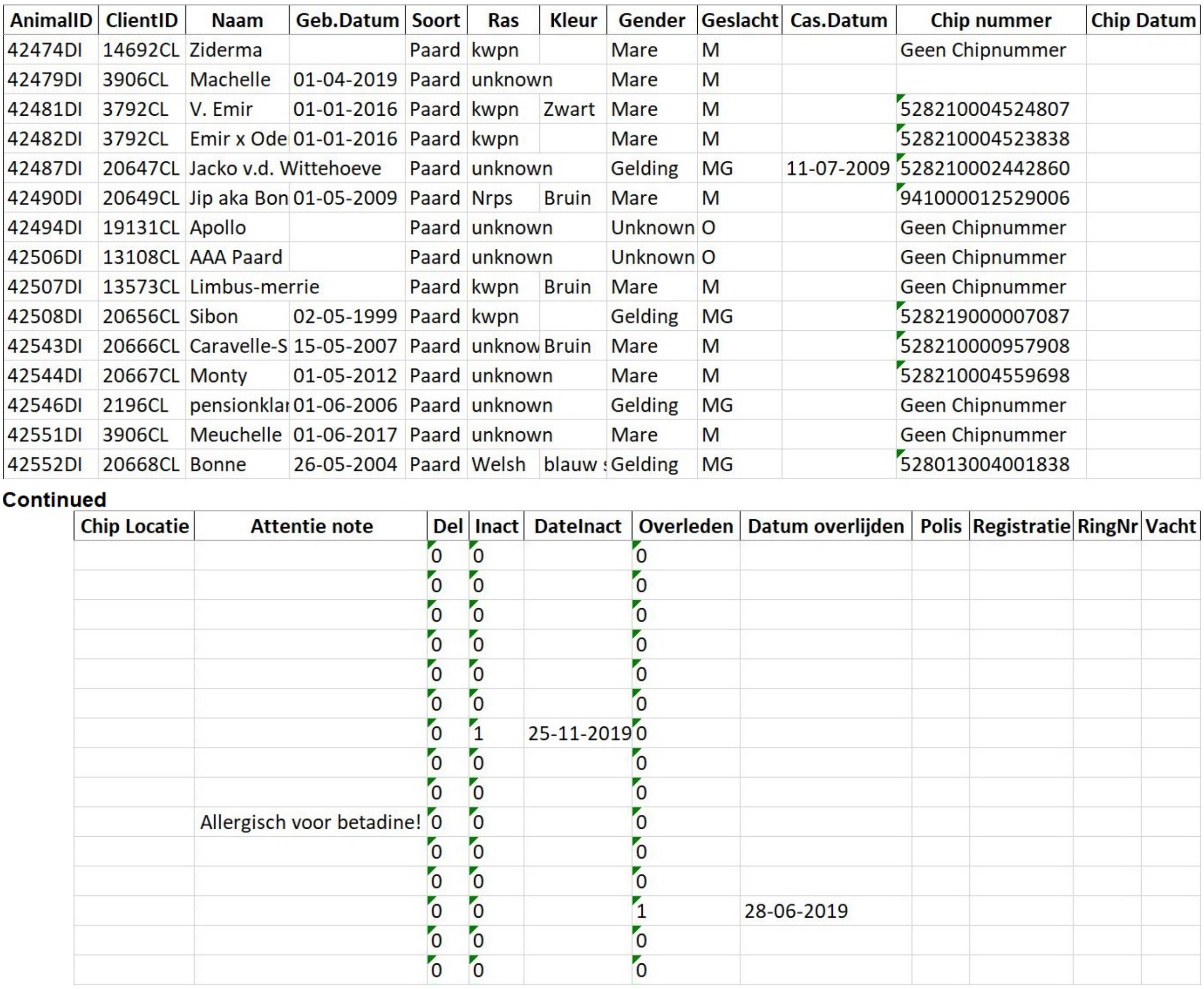
4.1 Used data sets

In this section, an overview of the data is given. It explains the used data and choices for the use of this specific data.

4.1.1 Veterinarian data

For this research, the medical data of animal clinic Den Ham is used. The medical data consists of a summary for each consult. At the clinic, these summaries are used to create the invoices. A combination of the summaries of one specific horse gives an overview of the health of that horse. All summaries of all horses can be combined to provide an overview of the performance of the treatments.
Animal clinic Den Ham works with the software of Viva Veterinary [69], a web application. To conduct this research, access was granted to the web application and an export of the data. This export consists of two separate CSV files, containing information about the horses and the medical files. The data shown in the web application differs slightly in format from the data in the export files. The web application contains clients, animals, and medical files. The clients are the owners of the animals. Each owner has one or more animals. For each animal, a medical file exists.
As described below, there are different types in the medical files. Some of these types are very structured, while others are not. With this combination, we can create a basic overview with the structured types, and make them more specific using the information from the unstructured types. First, the different entities in the data will be described.

Clients Each client has a client code, consisting of the first three letters of the last name and a number. Further, personal information is saved, like the first name, last name, address, telephone number. Viva also contains fields for birthdate, social security number, billing address, etc. but those fields often are blank.
There is no information available on the clients in the exported data set. In the exported data, a ClientID is available in both tables. This ID is not the same as the client code that is used in the web application. The ClientID is a number, ending with CL.



AnimalID	ClientID	Naam	Geb.Datum	Soort	Ras	Kleur	Gender	Geslacht	Cas.Datum	Chip nummer	Chip Datum
42474DI	14692CL	Ziderma		Paard	kwpn		Mare	M		Geen Chipnummer	
42479DI	3906CL	Machelle	01-04-2019	Paard	unknown		Mare	M			
42481DI	3792CL	V. Emir	01-01-2016	Paard	kwpn	Zwart	Mare	M		528210004524807	
42482DI	3792CL	Emir x Ode	01-01-2016	Paard	kwpn		Mare	M		528210004523838	
42487DI	20647CL	Jacko v.d. Wittehoeve		Paard	unknown		Gelding	MG	11-07-2009	528210002442860	
42490DI	20649CL	Jip aka Bon	01-05-2009	Paard	Nrps	Bruin	Mare	M		941000012529006	
42494DI	19131CL	Apollo		Paard	unknown		Unknown	O		Geen Chipnummer	
42506DI	13108CL	AAA Paard		Paard	unknown		Unknown	O		Geen Chipnummer	
42507DI	13573CL	Limbus-merrie		Paard	kwpn	Bruin	Mare	M		Geen Chipnummer	
42508DI	20656CL	Sibon	02-05-1999	Paard	kwpn		Gelding	MG		528219000007087	
42543DI	20666CL	Caravelle-S	15-05-2007	Paard	unknow	Bruin	Mare	M		528210000957908	
42544DI	20667CL	Monty	01-05-2012	Paard	unknown		Mare	M		528210004559698	
42546DI	2196CL	pensionklar	01-06-2006	Paard	unknown		Gelding	MG		Geen Chipnummer	
42551DI	3906CL	Meuchelle	01-06-2017	Paard	unknown		Mare	M		Geen Chipnummer	
42552DI	20668CL	Bonne	26-05-2004	Paard	Welsh	blauw s	Gelding	MG		528013004001838	

Continued

Chip Locatie	Attentie note	Del	Inact	DateInact	Overleden	Datum overlijden	Polis	Registratie	RingNr	Vacht
		0	0		0					
		0	0		0					
		0	0		0					
		0	0		0					
		0	0		0					
		0	0		0					
		0	1	25-11-2019	0					
		0	0		0					
		0	0		0					
	Allergisch voor betadine!	0	0		0					
		0	0		0					
		0	0		0					
		0	0		1	28-06-2019				
		0	0		0					
		0	0		0					

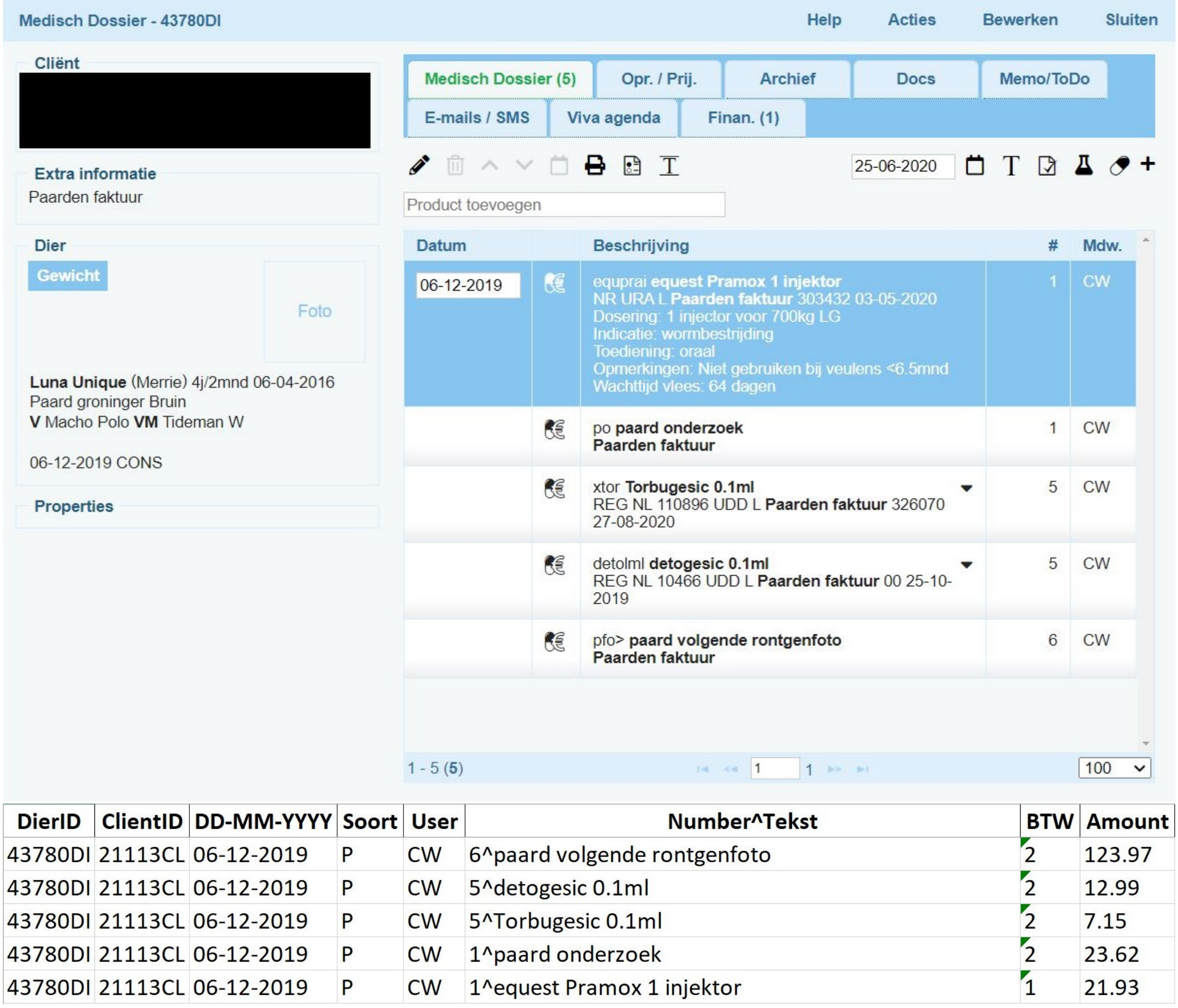
Figure 4.1: The information of a horse, in the export file.

Animals For the animals, the name, species (this study focuses on horses), breed, colour, birth date, sex, chip number, and whether the horse is insured, can be stored. If the sex is female sterilized or male castrated, the date of the procedure can be given. When a horse dies or becomes inactive, the corresponding box can be checked and the corresponding date can be entered into the system. Figure 4.1 shows the information about some horses in the export file.

In the animal export file, all information described above is available. Each of the animals has an AnimalID. This ID consists of a number, followed by DI. Also, the ClientID of the owner is available. This gives the possibility to group animals of the same owner.

Medical files All animals have a medical file. In this file, an overview of the treatments is given. The first date in the medical files is 19-08-1999. The file contains one entry before that, from the 20th of January 1999 but that only reads ”Opgenomen in dit bestand”, freely translated to ”added to this file”.

The medical file contains the columns date, type, description, count, and the veterinarian who conducted the consult. For horses, 4 types are used in the medical files: product, treatment, text, and lab results. Figure 4.2 shows the medical file of one horse in both the web application and the export file. This horse has been in the clinic once, on 6 December 2019. The horse is treated by veterinarian CW. The horse has had products (two types of sedation and dewormer)



Cliënt

[redacted]

Extra informatie
Paarden faktuur

Dier

Luna Unique (Merrie) 4j/2mnd 06-04-2016
Paard groninger Bruin
V Macho Polo **VM** Tideman W

06-12-2019 CONS

25-06-2020

Datum		Beschrijving	#	Mdw.
06-12-2019		equprai **equest Pramox 1 injektor** NR URA L **Paarden faktuur** 303432 03-05-2020 Dosering: 1 injector voor 700kg LG Indicatie: wormbestrijding Toediening: oraal Opmerkingen: Niet gebruiken bij veulens <6.5mnd Wachttijd vlees: 64 dagen	1	CW
		po **paard onderzoek** **Paarden faktuur**	1	CW
		xtor **Torbugesic 0.1ml** REG NL 110896 UDD L **Paarden faktuur** 326070 27-08-2020	5	CW
		detolml **detogesic 0.1ml** REG NL 10466 UDD L **Paarden faktuur** 00 25-10-2019	5	CW
		pfo> **paard volgende rontgenfoto** **Paarden faktuur**	6	CW

DierID	ClientID	DD-MM-YYYY	Soort	User	Number^Tekst	BTW	Amount
43780DI	21113CL	06-12-2019	P	CW	6^paard volgende rontgenfoto	2	123.97
43780DI	21113CL	06-12-2019	P	CW	5^detogesic 0.1ml	2	12.99
43780DI	21113CL	06-12-2019	P	CW	5^Torbugesic 0.1ml	2	7.15
43780DI	21113CL	06-12-2019	P	CW	1^paard onderzoek	2	23.62
43780DI	21113CL	06-12-2019	P	CW	1^equest Pramox 1 injektor	1	21.93

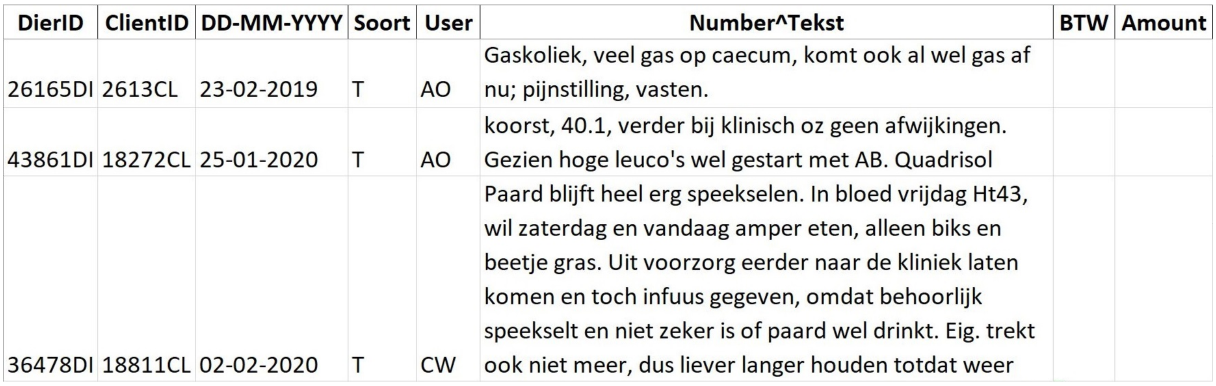
Figure 4.2: The medical file of a horse, as shown in the web application (top) and the export file (bottom).

and treatment (six x-rays and general examination). There is no text about this visit, so we don't know the reason or the outcome. As shown in figure 4.2 the layout of the export file (bottom) differs slightly from that of the web application (top).

For some products, the pharmacy has entered the standard information about the product. These are standard fields like dosage, indication (why it can be used), administration, comment (like shake before use or doping), and the waiting time between the application and slaughter for consumption. This standard information is not included in the export file.

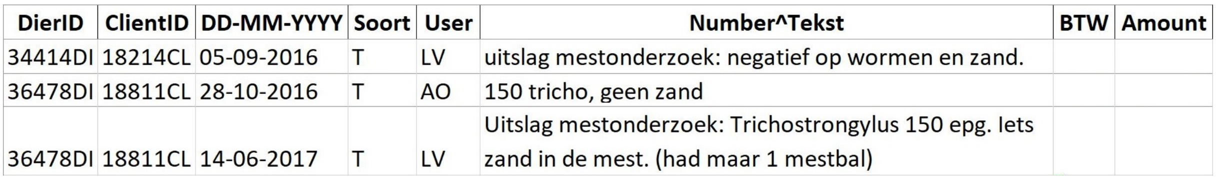
In some cases, a veterinarian added some additional information about the consult. This information is stored as text in both the export file and the web application. The text is unstructured, written text. Often, at least when the horse shows discomfort, the temperature of the horse was measured and is then given in the text, mainly indicated by "temp.", followed with a number. The text is very descriptive of the symptoms and/or the followed procedure. In many cases, the conclusion or diagnosis is missing. Figure 4.3 shows three text fields in the medical file of horses, as given in the web application (top) and the export file (bottom). In the export file, Texts are indicated in the column "Soort" (species) with a T, other treatments given are indicated with a P.

Some text types indicate the results of the faecal test. In this case, the field text starts with something like "uitslag mestonderzoek:" (results faecal test). This is followed by whether worm



DierID	ClientID	DD-MM-YYYY	Soort	User	Number^Tekst	BTW	Amount
26165DI	2613CL	23-02-2019	T	AO	Gaskoliek, veel gas op caecum, komt ook al wel gas af nu; pijnstilling, vasten.		
43861DI	18272CL	25-01-2020	T	AO	koorst, 40.1, verder bij klinisch oz geen afwijkingen. Gezien hoge leuco's wel gestart met AB. Quadrisol		
36478DI	18811CL	02-02-2020	T	CW	Paard blijft heel erg speekselen. In bloed vrijdag Ht43, wil zaterdag en vandaag amper eten, alleen biks en beetje gras. Uit voorzorg eerder naar de kliniek laten komen en toch infuus gegeven, omdat behoorlijk speekselt en niet zeker is of paard wel drinkt. Eig. trekt ook niet meer, dus liever langer houden totdat weer		

Figure 4.3: Three texts, as shown in the export file.

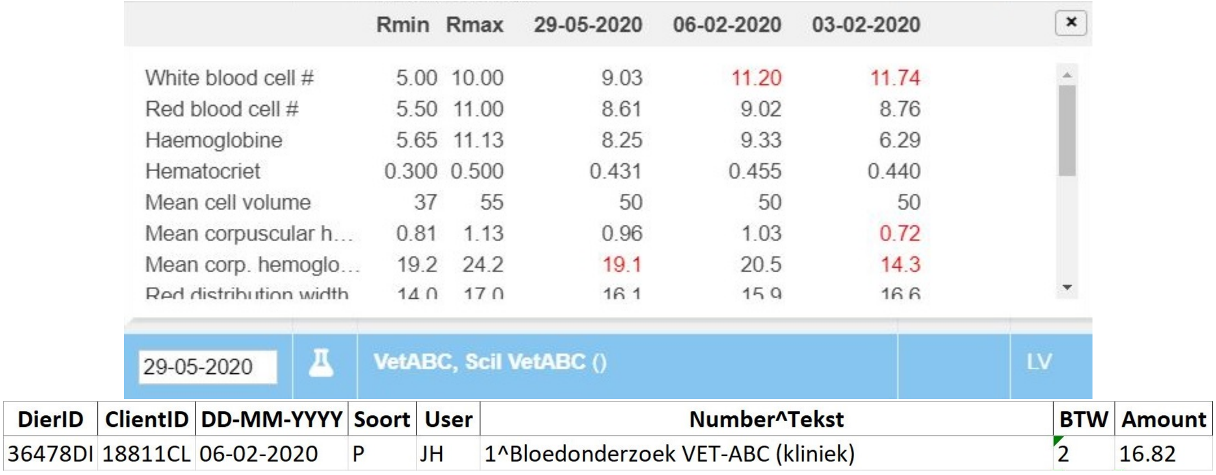


DierID	ClientID	DD-MM-YYYY	Soort	User	Number^Tekst	BTW	Amount
34414DI	18214CL	05-09-2016	T	LV	uitslag mestonderzoek: negatief op wormen en zand.		
36478DI	18811CL	28-10-2016	T	AO	150 tricho, geen zand		
36478DI	18811CL	14-06-2017	T	LV	Uitslag mestonderzoek: Trichostrongylus 150 epg. Iets zand in de mest. (had maar 1 mestbal)		

Figure 4.4: Three results of faecal tests, as shown in the export file.

eggs were found. If worm eggs are found, the species of worm is given. This information is necessary to determine what type of dewormer must be prescribed. It also states if sand is found in the faeces. Since July 1st of 2008, dewormers are not freely available in the Netherlands[70]. Before this date, horse owners dewormed their horse without knowing if the horse had a worm infection and what worms were present inside the horse. This has lead to resistance in worms found in horses. To prevent further resistance of worms against the dewormers, dewormers are only available via veterinarians. To reduce the resistance even further, veterinarians often do a faecal test first to check if deworming is necessary and if so, what dewormer would be best to use. Checking for sand in the faeces is important because the sand in the intestines of horses can cause sand colic, which can lead to death. If sand is found in the faeces, the horse should be treated.
Because of this, more faecal examination results can be expected after July 2008. The number of sold dewormer kits should rise from this date onward as well. Figure 4.4 shows three results of faecal tests. The first two start with ”uitslag mestonderzoek”, the third one does not. The results are obtained by two different vets (LV and AO).
In the web application, the lab results are displayed in a table, the first column of the table are the substances that are measured. The second and third columns contain the minimum and maximum value for a normal blood sample for each substance. The fourth column contains the measured values of the most recent test. If the test has been done before for this horse, the values of the previous tests are shown in the following columns.
The export file does not contain the actual results of the lab test. It only states that a blood test is done. Figure 4.5 shows the lab results of a horse as shown in the web application (top) and the export file (bottom).

Statistics and overview of the veterinarian data In the horse data, there are 15094 unique AnimalID’s and 4679 unique clientID’s. Table 4.1 shows the number of clients that have, or have had, a certain amount of horses, registered at the Animal Clinic. For example, there exist 124 clients in the database that have between 11 and 20 horses registered in the Animal Clinic Den Ham. The maximum amount of horses, registered under one client is 176. Clients with large amounts of horses registered are probably horse breeders, traders, or training farms.



	Rmin	Rmax	29-05-2020	06-02-2020	03-02-2020
White blood cell #	5.00	10.00	9.03	11.20	11.74
Red blood cell #	5.50	11.00	8.61	9.02	8.76
Haemoglobine	5.65	11.13	8.25	9.33	6.29
Hematocriet	0.300	0.500	0.431	0.455	0.440
Mean cell volume	37	55	50	50	50
Mean corpuscular h…	0.81	1.13	0.96	1.03	0.72
Mean corp. hemoglo…	19.2	24.2	19.1	20.5	14.3
Red distribution width	14.0	17.0	16.1	15.9	16.6

DierID	ClientID	DD-MM-YYYY	Soort	User	Number^Tekst	BTW	Amount
36478DI	18811CL	06-02-2020	P	JH	1^Bloedonderzoek VET-ABC (kliniek)	2	16.82

Figure 4.5: Lab results, as shown in the web application (top) and the export file (bottom).

# clients	# Horses
2703	1
786	2
379	3
217	5
273	6-10
124	11-20
32	21-30
22	31-50
16	51-100
9	100+

Table 4.1: The number of clients that has a given amount of horses registered at Animal Clinic Den Ham.

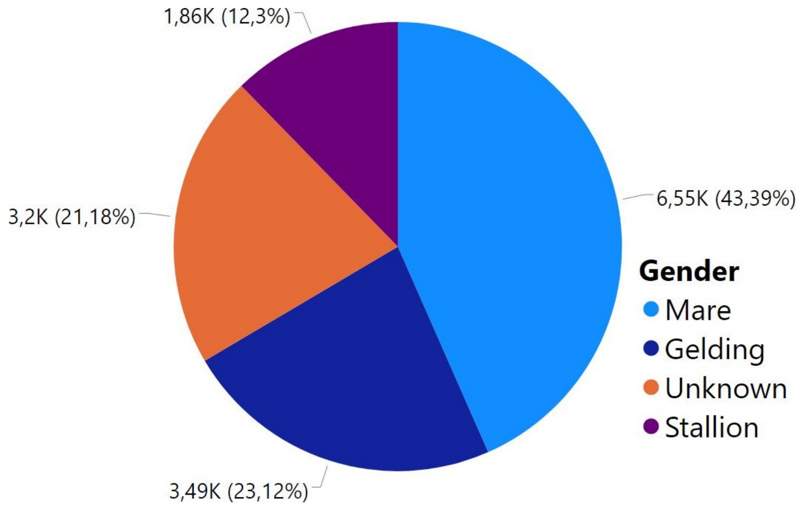


Figure 4.6: Distribution of the genders, including unknown

Looking at the names of horses, 1456 (less than 10%) of the horses have a name that is probably not the real name of the horse. In some cases, the name of the horse is a description of the horse, like the colour, age, breed, gender, or the diagnosis. In other cases the horse has a special character as name, there are also nine horses called X, of which eight have no other information. There is no certain way of knowing how many of these horses are duplicates of the other horses. A horse with a valid name may also be entered into the system twice.

In the data, 47 different breeds are counted. To reduce duplicates, the breeds are all set to lowercase and the white spaces are removed, but there are still some duplicates. For example, one horse is given the breed WPN, which probably needs to be KWPN (Royal Dutch Sport Horse). And, ”new forrest” and ”new forrest pony” both exist in the dataset. Those are most likely also the same breeds. 6300 of the 15094 horses is KWPN. For 5352 horses, no breed is given. The Frisian Horse is the next most existing breed with 502 horses in the dataset. Pie charts of the breeds, including and excluding unknown, is given in the Appendix, Figure A.1.

The genders of the horses are distributed as shown in figure 4.6. In this figure, all horses are shown, including the horses for which the gender is unknown. 43.39% of the horses are female and 35.42% is male (gelding and stallion combined). If the amount of males and females in the data set is equal, the unknown contains more males than females.

There are 5075 horses without a date of birth. Some of those horses have a date of death. This gives a range in which the horse must have been born. The horses are all treated by a veterinarian, which indicates when they were alive. Figure 4.7 (top) gives an overview of the months in which the horses are born. As expected, most horses are born in the spring months. The month with most births is January. This seems odd since this is a winter month and December and February are not as popular. When the date of birth of a horse is unknown, the age is guessed and the first of January is given as substitute date of birth. In that case, the age of the horse is guessed and the horse is given a date of birth which is therefore often the first of January. Also, the first of April, May, and June are popular estimation dates. To get a more realistic overview of the births of horses, in figure 4.7 (bottom) the first days of each month are not added. By leaving the first of the month out of the count, we also removed some horses that are born on the first of a month, therefore, each of the bars should be approximately a thirtieth higher.

The first birth date is 16 December 1967 and the youngest horse is born 3 April 2020. The distribution of the birth dates of the horses is given in the Appendix, Figure A.2.

The horses registered as death are distributed quite evenly over the months. An overview of the months in which horses are registered as death is shown in the Appendix, Figure A.4, with

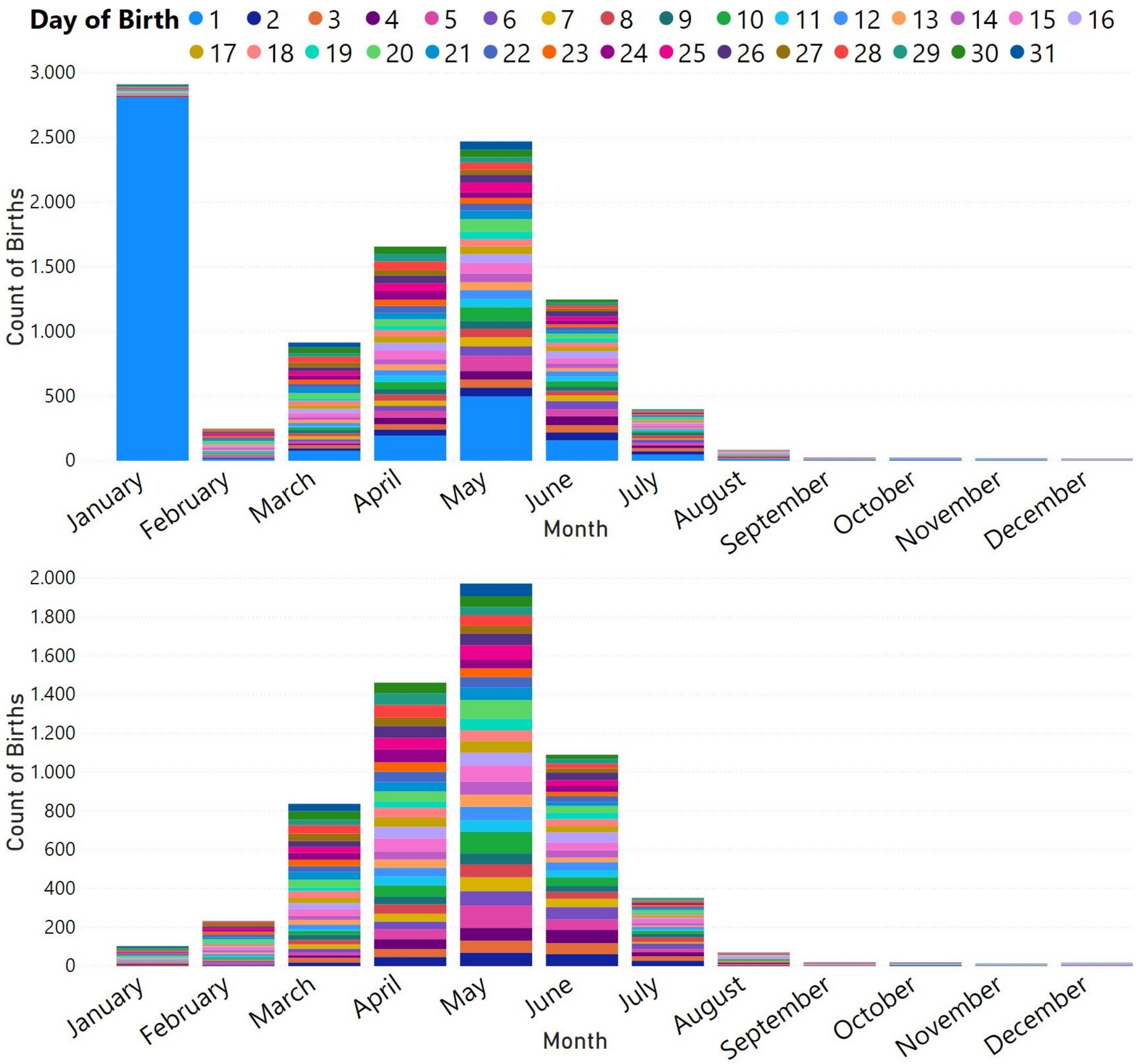


Figure 4.7: Distribution of births over the months, with (top) and without (bottom) the first day of each month.

and without the first day of the month. None of the months seems to be extremely more popular than others. Also, when comparing top and bottom, there is no phenomenon such as shown in figure 4.7.
The first horse that is registered to be dead, died on 1 February 2000. The last horse to die died on 26 March 2020. In total, 928 horses are registered dead in the database. For these horses, the variable ”Overleden” (passed away) is one and the date of death is given. The distribution for the dates of death can be found in the Appendix, Figure A.5. The horses that are registered dead are the horses that are euthanized, or examined postmortem by the veterinarians of animal clinic Den Ham. This is why only so few horses are registered to be dead.
In the data, 680 horses have a birthdate and a date of death. For all horses that have a birthdate and a date of death, the age is calculated. These ages of the horses are shown in the Appendix, Figure A.3. This shows, the oldest horse was 35, and the majority of horses does not reach the age of 25.
5978 horses seem to have correct chip numbers, yet only 5354 are unique. 22 of the horses occur 3 times in the file, and 279 horses are duplicated. For 7158 horses, no number is given. The phrase ”Geen chipnummer” (No chip number) is given for 1861 of the horses. There are 34

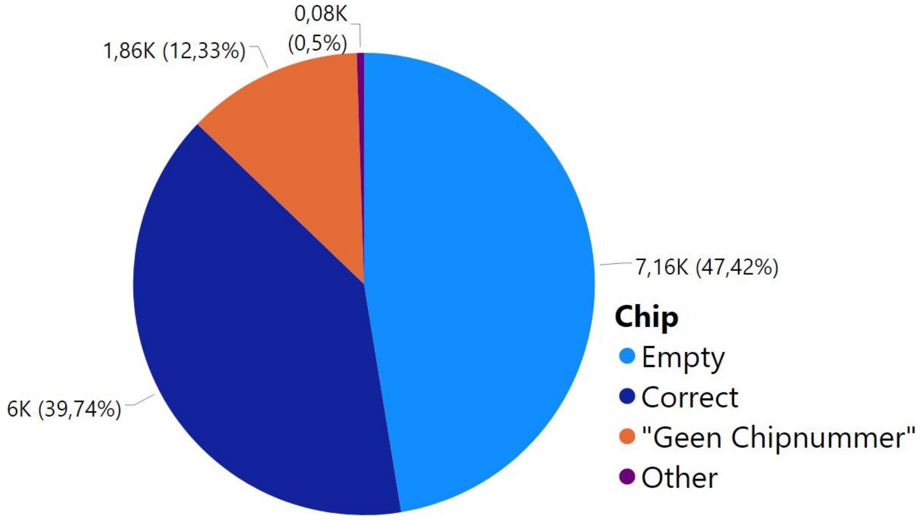


Figure 4.8: Distribution of the chip numbers as given. 'Correct' includes all numbers that seem to be correct, including duplicates.

horses for which the chip number starts with "DE". These may be valid German chip numbers. The rest of the horses do have a different type of ID, like the text "Brandmerk" (brand), "DNA" with some number, or just something random like "Onbekend" (unknown), "NVT" (does not apply), "Manegepony" (riding pony) or a very short number. An overview of the distribution of the chip numbers is given in Figure 4.8.
When looking at the names of horses, some horses have a number, very similar to a chip number as name, 9 of which have the same number stated at the column as chip number.
The medical file consists of 144399 lines. A consult can be specified as a unique combination of AnimalID and date. Animal Clinic Den Ham has had 58927 consults over the years. Each consult takes an average of almost 2.5 lines in the data set. The number of consults has been growing throughout the years to over 15000 in last year. The month's March to August seems to be busier months at the clinic. The distribution of consults over the months and years is shown in Appendix, Figure A.6.

Data quality The data that is entered into the system is probably correct, but we miss a lot of data from the horses, and it is unknown how many of the horses are duplicates. When looking at individual consults, this will not be a problem, since this will not affect the ability of the veterinarian to make a diagnosis. However, it does have an impact on the analysis of the full live-span of the horses. Because of the missing values, the overall quality of the data is not very good. The biggest problems with the data are the identification of a horse but since this research focuses on the occurrence of specific diseases and not on the diagnosis of individual horses this is no problem. Identifying the consults that concern the diseases used in this paper will be challenging when no descriptive text is given.

4.1.2 Weather data

To investigate the influence of the weather on the health of horses, the weather data and the veterinarian data will be merged. For this research, the focus will be on the short term and long term impact of the weather on the horses, for example, a very hot or wet week or cold or mild winter.
The KNMI [71], the Royal Dutch Meteorologic Institute provides data sets, obtained at weather stations throughout the Netherlands, on its website. This data is collected at 35 weather stations spread over the Netherlands. According to the KNMI, the data of four of these weather stations

are homogenized and therefore suitable for trend analysis. The other weather stations are not suitable for trend analysis since it is possible that the weather stations have been moved or the observation methods have changed [72].

Locations The KNMI has 35 weather stations spread over the Netherlands. Figure 4.9 shows the different stations and their locations, as well as the location of the animal clinic in Den Ham. The red pointers are the weather stations of De Kooi, Eelde, De Bilt and Vlissingen, these are the homogenized ones. The homogenized weather stations are all far away from the horse clinic in Den Ham. As mentioned above, the homogenized weather stations are suitable for trend analysis and the other ones are not. Since this is not an analysis on the weather data only, but an analysis used to support the veterinarian data, it is more suitable to use the data of a weather station closer to the animal clinic Den Ham.

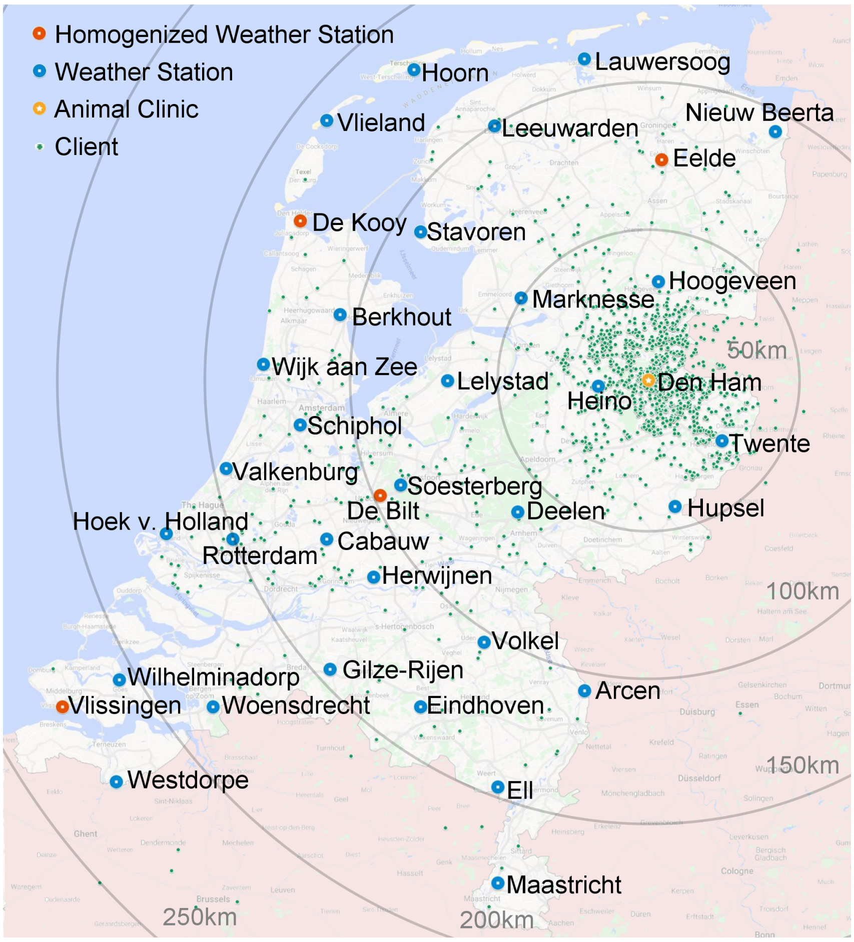
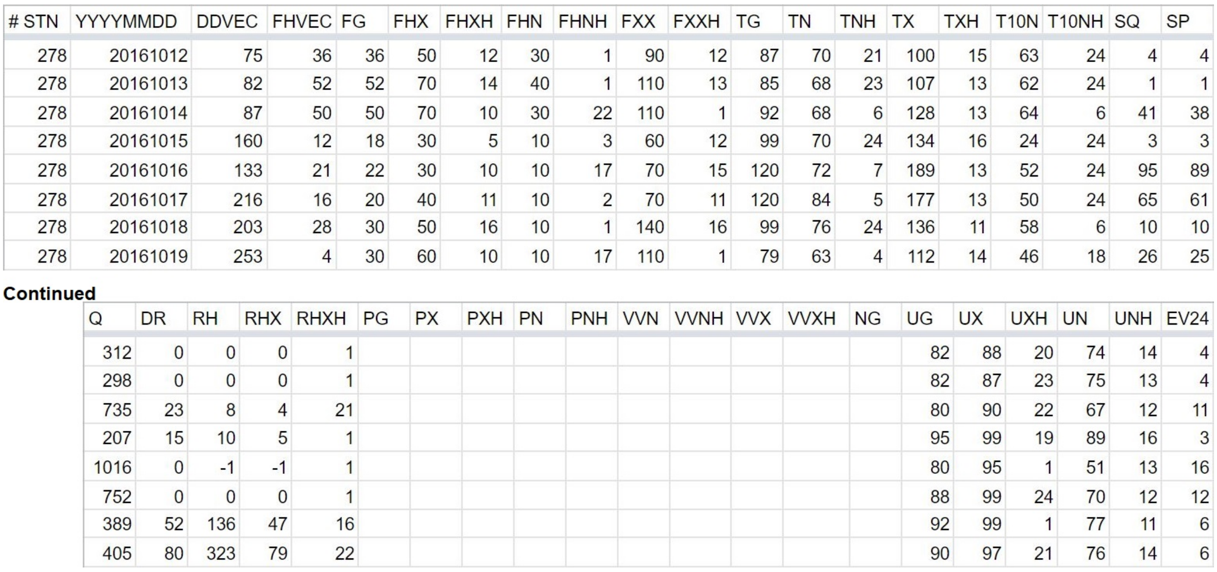


Figure 4.9: The locations of different weather stations in the Netherlands, the animal clinic Dan Ham and the location of their clients

To choose a weather station, the locations of the clients with horses are plotted on the map shown in image 4.9. The locations are the billing addresses of the clients. When the horse is kept at home, this address is the location of the horse. In some cases, the horse is located at a stable. Most horse owners will keep their horses close to their home. Therefore, the location of the horse and the location of the horse owner are probably not far apart. The majority of clients with horses is located around Den Ham but there are clients anywhere in the Netherlands and even in other countries. According to the expert, this is due to purchase inspections where the horse is inspected in at Animal Clinic Den Ham but the person who pays the inspection lives far away. The weather stations of Heino, Hoogeveen and Twente are all laying at the outside



# STN	YYYYMMDD	DDVEC	FHVEC	FG	FHX	FHXH	FHN	FHNH	FXX	FXXH	TG	TN	TNH	TX	TXH	T10N	T10NH	SQ	SP
278	20161012	75	36	36	50	12	30	1	90	12	87	70	21	100	15	63	24	4	4
278	20161013	82	52	52	70	14	40	1	110	13	85	68	23	107	13	62	24	1	1
278	20161014	87	50	50	70	10	30	22	110	1	92	68	6	128	13	64	6	41	38
278	20161015	160	12	18	30	5	10	3	60	12	99	70	24	134	16	24	24	3	3
278	20161016	133	21	22	30	10	10	17	70	15	120	72	7	189	13	52	24	95	89
278	20161017	216	16	20	40	11	10	2	70	11	120	84	5	177	13	50	24	65	61
278	20161018	203	28	30	50	16	10	1	140	16	99	76	24	136	11	58	6	10	10
278	20161019	253	4	30	60	10	10	17	110	1	79	63	4	112	14	46	18	26	25

Continued

Q	DR	RH	RHX	RHXH	PG	PX	PXH	PN	PNH	VVN	VVNH	VVX	VVXH	NG	UG	UX	UXH	UN	UNH	EV24
312	0	0	0	1											82	88	20	74	14	4
298	0	0	0	1											82	87	23	75	13	4
735	23	8	4	21											80	90	22	67	12	11
207	15	10	5	1											95	99	19	89	16	3
1016	0	-1	-1	1											80	95	1	51	13	16
752	0	0	0	1											88	99	24	70	12	12
389	52	136	47	16											92	99	1	77	11	6
405	80	323	79	22											90	97	21	76	14	6

Figure 4.10: The weather data, as provided by the KNMI

of where the majority of the clients are located. The weather station of Heino is the best to use for this analysis because it is closest to Den Ham and surrounded by most of the clients.

Types of data For the different weather stations, the KNMI has several groups of measurements and observations:

- Hourly weather station values of multiple weather conditions
- Daily weather station values of multiple weather conditions
- Daily precipitation values
- Month- and year values of multiple weather conditions
- Ground temperatures
- Parallel temperature measurements

For this research, an overview of the weather before an appointment is needed to get insight into the influence of the weather on the horses. To analyse the weather of the past days, weeks and months, the hourly data provides too much and the weekly, monthly and yearly too little detail. Therefore, the daily values would suit best. The precipitation is included in the daily weather station values also, therefore there is little need for data separation.

Explanation of the data The daily weather data sets of the KNMI all have the same layout, containing 40 variables. The explanation of these variables is given in table 4.2. Of these 40 variables, 12 values are the maximum and minimum measured that day and 12 are the hour in which these extremes are measured. For example, The TXH is the hour in which the TX is measured. The maximum temperature (TX) of 12-10-2016 is 10°C, this maximum is measured between 14:00 and 15:00 hour. So the TXH at 12-10-2016 is 15.
Figure 4.10 shows the weather data as provided by the KNMI. The columns of the data sets from all the weather stations are in the same order. The first column (#STN) is the number of the weather station. The second column is the day of the observations, formatted: year, month, day. Followed by the observations of the rest of the days.

Category	Variable name	Explanation
	YYYYMMDD	Date (YYYY=year MM=month DD=day)
Wind	DDVEC	Vector mean wind direction in degrees (360=north, 90=east, 180=south, 270=west, 0=calm/variable)
	FHVEC	Vector mean wind speed (in 0.1 m/s)
	FG	Daily mean wind speed (in 0.1 m/s)
	FHX	Maximum hourly mean wind speed (in 0.1 m/s)
	FHXH	Hour in which FHX was measured
	FHN	Minimum hourly mean wind speed (in 0.1 m/s)
	FHNH	Hour in which FHN was measured
	FXX	Maximum wind gust (in 0.1 m/s)
	FXXH	Hour in which FXX was measured
Temperature	TG	Daily mean temperature (in 0.1 degrees Celsius)
	TN	Minimum temperature (in 0.1 degrees Celsius)
	TNH	Hour in which TN was measured
	TX	Maximum temperature (in 0.1 degrees Celsius)
	TXH	Hour in which TX was measured
	T10N	Minimum temperature at 10 cm above surface (in 0.1 degrees Celsius)
	T10NH	6-hour in which T10N was measured; 6=0-6 UT, 12=6-12 UT, 18=12-18 UT, 24=18-24 UT
Sunshine duration	SQ	Sunshine duration (in 0.1 hour) calculated from global radiation (-1 for <0.05 hour)
	SP	Percentage of maximum potential sunshine duration
	Q	Global radiation (in J/cm2)
Precipitation	DR	Precipitation duration (in 0.1 hour)
	RH	Daily precipitation amount (in 0.1 mm) (-1 for <0.05 mm)
	RHX	Maximum hourly precipitation amount (in 0.1 mm) (-1 for <0.05 mm)
	RHXH	Hour in which RHX was measured
Barometric Pressure	PG	Daily mean sea level pressure (in 0.1 hPa) calculated from 24 hourly values
	PX	Maximum hourly sea level pressure (in 0.1 hPa)
	PXH	Hour in which PX was measured
	PN	Minimum hourly sea level pressure (in 0.1 hPa)
	PNH	Hour in which PN was measured
Visibility	VVN	Minimum visibility; 0: <100 m, 1:100-200 m, 2:200-300 m,..., 49:4900-5000 m, 50:5-6 km, 56:6-7 km, 57:7-8 km,..., 79:29-30 km, 80:30-35 km, 81:35-40 km,..., 89: >70 km)
	VVNH	Hour in which VVN was measured
	VVX	Maximum visibility; 0: <100 m, 1:100-200 m, 2:200-300 m,..., 49:4900-5000 m, 50:5-6 km, 56:6-7 km, 57:7-8 km,..., 79:29-30 km, 80:30-35 km, 81:35-40 km,..., 89: >70 km)
	VVXH	Hour in which VVX was measured
Cloud coverage	NG	Mean daily cloud cover (in octants, 9=sky invisible)
Humidity	UG	Daily mean relative atmospheric humidity (in percents)
	UX	Maximum relative atmospheric humidity (in percents)
	UXH	Hour in which UX was measured
	UN	Minimum relative atmospheric humidity (in percents)
	UNH	Hour in which UN was measured
Evaporation	EV24	Potential evapotranspiration (Makkink) (in 0.1 mm)

Table 4.2: The explanation of the variables given by the KNMI

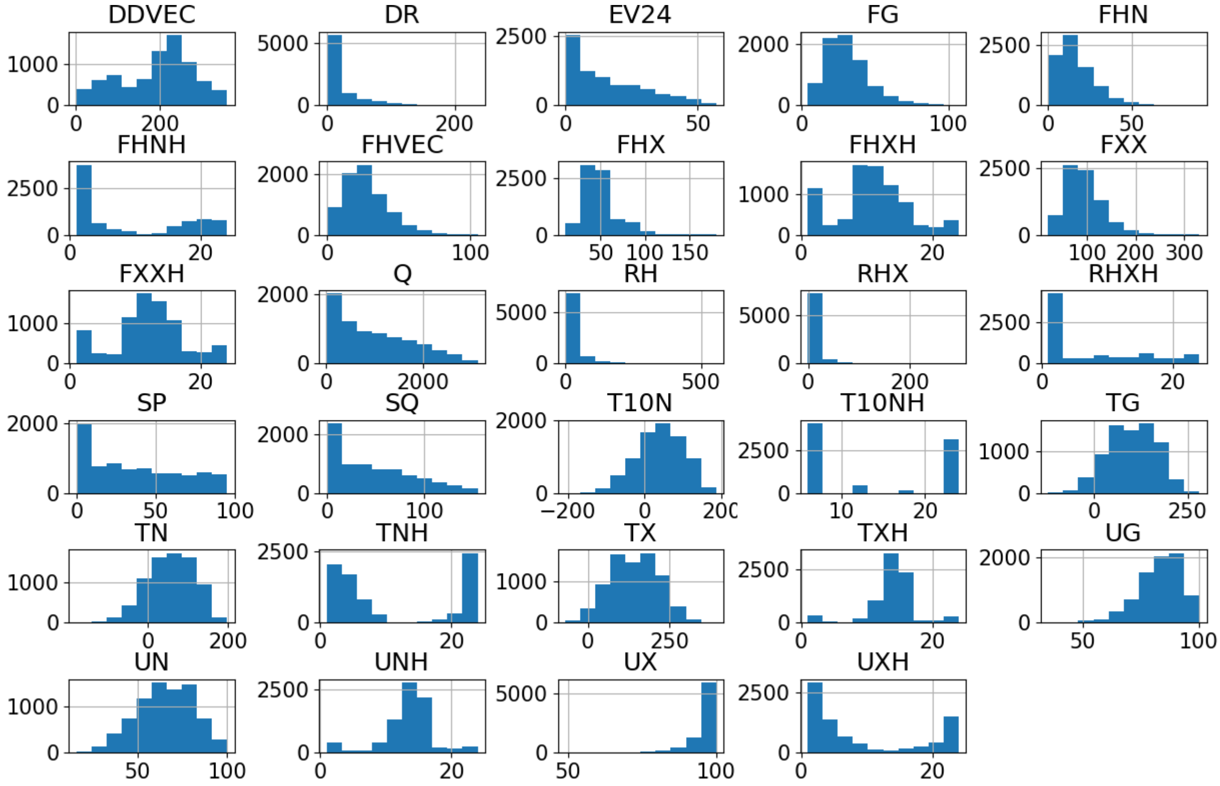


Figure 4.11: The distribution of the variables of the weather data of Heino (13-11-1998 to 09-04-2020), the missing variables are not plotted.

For each of the locations, the starting date differs. As well as the start date of the individual variables. This is due to the fact that the weather stations grow over time and more instruments are added. Heino is still missing some instruments, and the associated values. The variables correlated to the missing instruments are PG, PX, PXH, PN, PNH, VVN, VVNH, VVX, VVXH and NG. Table 4.3 gives an overview of the first time the data appeared and after which date the variable is always available for the data set of Heino, Twente and Hoogeveen.
As shown in table 4.3 no data set contains all variables from the beginning of 1999 onwards. The data set of Heino has a gap in 2001 for the variable DR. The values are missing from 27-02-2001 to 19-07-2001. Also, misses the variable PG to NG all together. Hoogeveen is missing data for FXX and FXXH, and both miss data for T10N, T10NH and VVX to NG.

Statistics and overview of the weather data In this paragraph, some statistics about the weather data of weather station Heino will be given. The data used covers the period 13 November 1998 until 9 April 2020, after which the veterinarian data ends. Data beyond that date is superfluous. The given period counts 7819 days. The value DR misses 143 days and the values PG, PX, PXH, PN, PNH, VVN, VVNH, VVX, VVXH and NG are missing for all days. Table 4.4 shows the mean, median, standard deviation, maximum and minimum value of each of the variables of the data set. To get more insight into the weather data, Figure 4.11 shows the histograms of the variables of the weather data. Most of the weather variables are not normally distributed. No histograms are created for the variables without data. The missing values of DR are not imputed either.

Data quality From Figure 4.11 we can conclude that the vector mean wind speed (FHVEC) and the daily mean wind speed (FG) are very similar, both have most values around 25 and their maximum values around 100. They are somewhere between the maximum (FHX) and

	Heino		Twente		Hoogeveen	
	first time available from	always available from	first time available from	always available from	first time available from	always available from
DDVEC	19891013	19981113	19510101	19840717	19891013	19980425
FHVEC	19891013	19981113	19510101	19840717	19891013	19980425
FG	19891013	19981113	19510101	19510430	19891012	19980425
FHX	19891013	19981113	19510101	19600823	19891013	19980425
FHXH	19891013	19981113	19510101	19600823	19891013	19980425
FHN	19891013	19981113	19510101	19600823	19891013	19980425
FHNH	19891013	19981113	19510101	19600823	19891013	19980425
FXX	19891013	19981113	19710101	19950510	19891013	20050219
FXXH	19891013	19981113	19710101	19950510	19891013	20050219
TG	19891013	19981113	19510101	19510430	19891013	19900515
TN	19891013	19981113	19510101	19600823	19891013	19900515
TNH	19891013	19981113	19510101	19600823	19891013	19900515
TX	19891013	19981113	19510101	19600823	19891013	19900515
TXH	19891013	19981113	19510101	19600823	19891013	19900515
T10N	19930610	19981113	19701231	20110102	19891013	20110102
T10NH	19930611	19981113	19900101	20110102	19900111	20110102
SQ	19930801	19981113	19630101	19931228	19910227	19910414
SP	19930801	19981113	19630101	19931228	19910227	19910414
Q	19930716	19981113	19870416	19900426	19891013	19900515
DR	19930202	20010720	19740601	19740532	19930301	19940125
RH	19920305	19981113	19740601	19740532	19930301	19940125
RHX	19920305	19981113	19740601	19760108	19930301	19940125
RHXH	19920311	19981113	19800816	19931231	19930301	19940125
PG	-	-	19510101	19541227	19891013	19900515
PX	-	-	19510101	19641112	19891013	19900515
PXH	-	-	19800816	19931231	19900111	19931231
PN	-	-	19510101	19641112	19891013	19900515
PNH	-	-	19800816	19931231	19900111	19931231
VVN	-	-	19550101	20190918	19921222	20170909
VVNH	-	-	19550101	20190918	19921222	20170909
VVX	-	-	19550101	20190918	19921222	20170909
VVXH	-	-	19550101	20190918	19921222	20170909
NG	-	-	19510101	20190125	20001108	20170909
UG	19930610	19981113	19710101	19701232	19891013	19900515
UX	19930610	19981113	19710101	19701232	19891013	19900515
UXH	19930610	19981113	19710101	19701232	19891013	19900515
UN	19930610	19981113	19710101	19701232	19891013	19900515
UNH	19930610	19981113	19710101	19701232	19891013	19900515
EV24	19930716	19981113	19870416	19900426	19891013	19900515

Table 4.3: For each of the variables, the first date and the date after which no missing values occur for the three weather stations closest to the animal clinic in Den Ham (the dates have the format YYYYMMDD, red cells indicate missing values between 01-01-1999 and 09-04-2020)

	Mean	Median	STD	Max	Min
DDVEC	193.49	213.0	86.72	360	1
FHVEC	27.40	25.0	15.37	105	0
FG	30.96	28.0	14.62	107	4
FHX	49.70	50.0	19.69	180	10
FHXH	10.75	11.0	5.44	24	1
FHN	13.58	10.0	12.51	90	0
FHNH	8.86	4.0	8.66	24	1
FXX	96.39	90.0	35.97	330	20
FXXH	12.00	12.0	5.40	24	1
TG	101.69	102.0	63.37	280	-121
TN	57.10	59.0	57.91	198	-179
TNH	11.09	6.0	9.41	24	1
TX	143.66	143.0	74.54	393	-69
TXH	13.72	14.0	3.77	24	1
T10N	40.07	42.0	60.42	186	-207
T10NH	13.80	6.0	8.61	24	6
SQ	47.15	39.0	40.65	155	0
SP	36.64	33.0	29.46	95	0
Q	1010.24	854.0	774.85	3120	11
DR	17.78	2.0	29.20	236	0
RH	21.27	1.0	42.33	556	-1
RHX	8.62	1.0	18.26	296	-1
RHXH	7.19	2.0	7.72	24	1
UG	82.68	84.0	9.40	100	35
UX	96.04	97.0	4.62	100	49
UXH	9.20	5.0	8.65	24	1
UN	65.52	66.0	15.64	100	16
UNH	13.35	14.0	3.98	24	1
EV24	15.95	12.0	13.63	57	0

Table 4.4: The mean, median, standard deviation (STD), maximum (max) and minimum (min) value of the weather data of weather station Heino, from the time period from 13-11-1998 to 09-04-2020

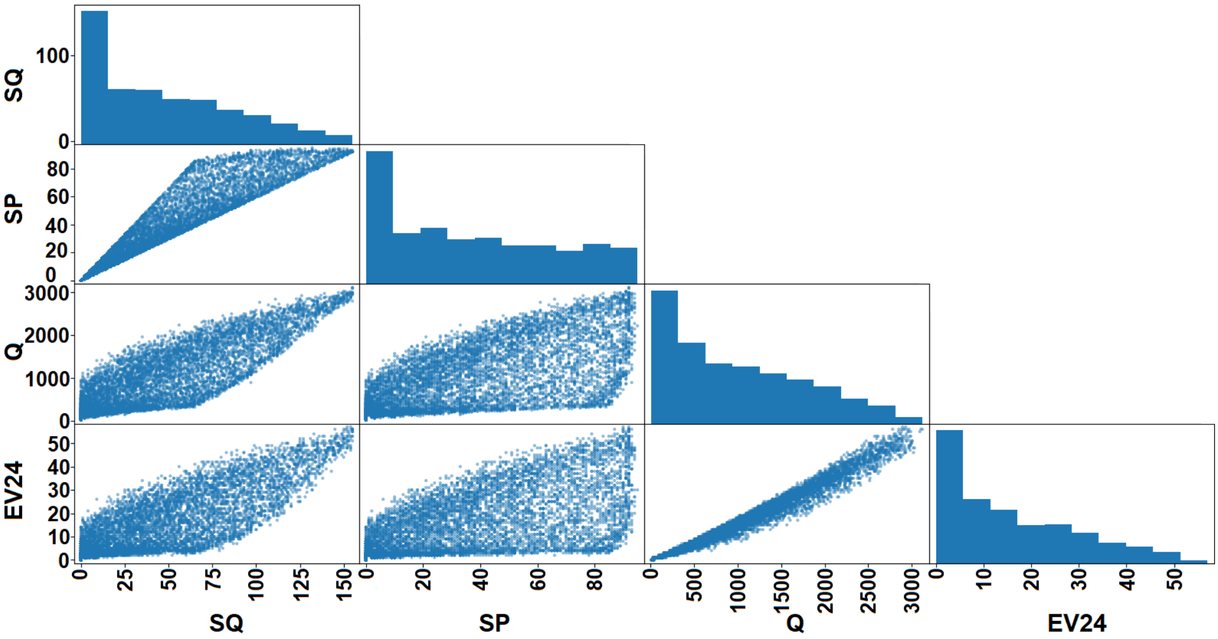


Figure 4.12: Correlation between variables SQ (sunshine duration), SP (percentage of maximum potential sunshine duration), Q (global duration) and EV24 (potential evaporation), in the period 13-11-1998 to 09-04-2020

minimum (FHN) wind speed which have most values around 50 and 10, with maximum values of 150 and 60 respectively. This is as expected. The max wind gust (FXX) lies higher then the rest, with most values around 100, and values up to 250 the maximum wind gust (FXX) and the maximum hourly wind speed (FHX) both are measured mostly during the day (FHXH and FXXH), while the minimum wind speed is measured more often during the nights (FHNH).

The daily mean temperature (TG) is normally distributed, with most values around 100 The minimum (TN) and maximum (TX) daily temperatures have most values around 50 and 150 respectively which, again, is as excepted. The minimum temperatures are mostly measured during the night (TNH) and the maximum temperatures (TXH) are measured between 12 and 2 PM. The minimum temperature at 10cm above the ground (T10N) is quite similar to the minimum temperature (TN). This minimum is measured mainly around midnight (T10NH).

The three values for the sunshine duration (SQ, SP and Q) and the potential evapotranspiration (EV24) seem quite similar to each other when looking at the histograms in figure 4.11. This could indicate a correlation between these values but as shown in figure 4.12, this is only the case for EV24 and Q. All three have most values at 0. The maximum values are quite different. The percentage of maximum potential sunshine duration (SP) stays somewhat above 500 after the peek at zero. While the other two are dropping from 1000 to 0. The percentage of maximum potential sunshine duration (SP) stops, obviously, a 100. The sunshine duration (SQ) has values up to 150 and the global radiation (Q) up to 2500. The EV24 has values up to 60.

The precipitation duration (DR), daily precipitation amount (RH) and the maximum hourly precipitation amount (RHX) Show the same figure, almost all values are between 0 and 20. The precipitation duration (DR) and the maximum hourly precipitation amount (RHX) have values up to 250. The daily precipitation amount (RH) has values up to 500. The hour in which the maximum hourly precipitation amount is measured (RHXH) has most values on 0. This is probably because, when the RHX is 0, the RHXH also is 0.

The humidity is given in percentages. This means that all values are somewhere between 0 and 100. Most of the values of the daily mean relative humidity (UG) lay somewhere between 75 and 100. Most of the time, the maximum value (UX) is 100. The minimum value (UN) lies

most of the time between 50 and 75. The minimum value is mostly measured during the day (UNH), while the maximum value is mostly measured during the night (UXH).
The Vector mean wind direction in degrees has two peaks, one smaller peak around 100 and a higher peak around 250. This histogram does not look like any of the others.
Some plots look alike, when looking at the values measured, this is explainable. From this, it can be concluded that the quality of the data is good. Many plots do look alike, meaning not all variables will be used.

4.2 Prepare data

The data from the veterinarian includes all animals instead of only horses and need to get more structured before it is possible to create the visualizations, find correlations and do predictions on the data. The weather data is missing some variables and values that are important for this research. Therefore the data from the veterinarian and the weather data must be prepared before starting the visualization of the data.

4.2.1 Veterinarian data

Preparation of the veterinarian data is needed; The data consists of two files, containing all animals, the horses need to be extracted from this. The data contains duplicates, those should be identified and merged as much as possible. And the text needs to be interpreted. The methodology for how this will be done is given below.

Horse information data The veterinarian data consists of two files, the animal information file and the animal medical file. Since this research focuses on horses only, the other animals need to be removed from the files. The type of animal is given in the animal information file but not in the medical file. All animals labelled as a horse will be selected in the animal information file to create a horse information file.
Both the information file and the medical file contain the AnimalID's. The rows in the medical file that have an AnimalID that does not exist in the newly created horse information file will be removed from the medical file, resulting in the horse medical file.

Removal of duplicate horses Finding duplicate horses is very challenging. Many of the horses with a generic name like "paard" (horse) are probable to occur somewhere in the data with their real information but it is not in the scope of this research to identify these if it is even possible. When the chip numbers of different identities are the same the entries are, most likely, the same horse.
In the horse information file, 22 chip numbers occur three times, and 279 chip numbers that occur two times. When the chip number of the horse is known, a lot of the information of the horse is often known as well; like the date of birth, the gender, the colour, and the breed of the horse. These characteristics are not likely to be different in the different entries, except for the gender, since a horse can be gelded at a point in time. Some of the entries, however, are differing a lot from each other. The entries of the horses with identical chip numbers will be merged.
To find the duplicate chip numbers, all chip numbers that occur 2 or 3 times and have a length of 6 or longer collected. For merging the rows of the duplicate chip numbers, different strategies are used for the different columns:

- For the name, colour and breed, when a string is more then 75% alike or one fits into the other, the shortest of the two will be discarded. When this is not the case, both strings will be preserved.

- Each horse in the database has one of the following options for gender: 'O' (unknown), 'M' (mare), 'MG' (gelding) or 'H' (Stallion). When the different entities of one horse have different genders there is no way to know what the real gender is. When a horse is gelded it is unclear when this has happened. Therefore a set of all genders will be preserved. The 'O' (unknown) will be removed if one or more other options are available.

- When one or more of the entries is registered dead, inactive or deleted, the merged entry of the horse will also be registered accordingly.

- The birth dates that start with 01-01 (first of January) will be removed when other birth dates are available. All other birth dates will be preserved.

- The AnimalID's and ClientID's will be concatenated with a comma and space between them in the new entry. the AnimalID's and ClientID's in the medical file will be changed accordingly.

Creating consults The medical file will be grouped into consults. Consults, in this research, are specified as the unique combination of AnimalID and date. A horse can have more than one consult in one day, for example when the situation of the horse deteriorates. But it is impossible to guarantee that all consults are separated. Likely, a second or third consult on the same day to the same horse is for the same problem as the previous one. Therefore separate consults of the same horse on the same day are treated as one consult.
After merging the rows of the medical file into consults, the data rows will contain the following: The fields 'soort' (type), 'Number^Tekst' (description), BTW (VAT), and 'Amount' (amount) will contain a list of the previously separate item of the identical names. The other fields will contain a string when all items in the consult were identical for that column. Otherwise, a list is created containing all unique items. For example, when different veterinarians visited the horse on the same day, a list, containing the different veterinarians is created. the index of the consults data set is the AnimalID and the data of the consult.

Get temperatures When the temperature of a horse is measured, this will be entered in the text. To obtain the temperatures all consecutive numbers, commas and dots are collected from the texts. Dots and commas at the beginning and ends of the collected strings are removed and the commas are replaced by dots. The strings are now parsed to floats. All floats between 36 and 43 are considered as potential temperatures. The potential temperatures are checked by hand to see if the numbers are actual temperatures. The accuracy of this method of collection of temperatures will be measured. When specific numbers are often found to be false positives, the number will be excluded as potential temperature to get the temperatures as accurate as possible. These temperatures are added to the consults data set.

Medical data text Preparing the medical data consists of labelling the consults with the cause of the consult. This way, the consults labelled with the diseases can be used for answering the research questions. To label the consults a keyword search will be performed. To find the consults concerning colic, the keyword "Koliek" will be used. to find the consults concerning laminitis, the keyword "hoefbevangen" will be used. To find consults concerning skin diseases, the keywords ”schimmel” (fungal) and ”mok” (mud fever) will be used. Consults concerning respiratory problems will be found using the keywords ”snot” (snot), ”luchtweg” (airway), ”bronchitis” (bronchitis) and ”longontsteking” (pneumonia).

In addition to the keyword search over 10.000 consults, that is over one-sixth of the total number of consults, will be labelled by hand for seven different verities of the same label sets. These label sets will be used for supervised learning. The different sets make it possible to find the best way of labelling. One label set will contain all the reasons for the consult separated by commas. A second label set (Reduced) contains the main reason for the consult, this is one of the reasons given in the first label set. The third label set (Simple) will be the same label set as the second but consults that have a label that occurs less the 30 times will be relabelled to ”overig” (other). The remaining four label sets are binary label sets, one for each of the diseases used for this research. Containing a ”1” when the diseases is a cause for the consult and a ”0” when this is not the case. Keyword search is expected to have false-positive results when the keywords are used to indicate that it is not this disease and false negatives when other words are used to describe the illness, type mistakes are made or if no text is given. When no text is given by the veterinarian, it is possible to still identify an illness by the medication that is given. Therefore the hand labelled consults also will be used to train machine learning models to see if these models can predict the cause of the consult better then keyword search.

Since the binary sets will have a lot more zeros than ones, the model will learn to classify everything as zero since this will still give a good accuracy. To prevent this from happening the label sets will be balanced using Random Over Sampling. This means that random positive samples will be duplicated until there are as many positive as negative labels in the label set.

For labelling the data using Machine Learning the tutorial ”Working With Text Data” [73] from scikit-learn will be used. Before we can start training models the texts of the consults will be turned into numerical feature vectors using a bag of words representation, in this research the scikit-learn function CountVectorizer [74] will be used to perform this pre-processing. For the pre-processing and tokenization of the texts the use of unigrams, bigrams or both will be tested (ngram_range), as well as whether stop-words should be filtered from the texts or not (stop-words).

To normalize the count matrix, created before, Term Frequency (tf) or Term Frequency times Inverse Document Frequency (tf_idf) will be used. In this research TfidfTransformer [75] from scikit-learn will be used. Whether tf or tf-idf works better for this data will be tested (use_idf). Also, the application of sublinear tf scaling i.e. replacing with 1 + log(tf) (sublinear_tf) and soothing of idf weights by adding one to the document frequencies (smooth_idf) are tested.

Scikit Learn has the following classification models to use: a multinomial Naive Bayes classifier, a linear Support Vector Machine with 3 different loss functions (hinge, squared_hinge and modified_huber), a Logistic Regression Classifier (log), a Neural Network (perceptron), an Ordinary Least Regression (squared_loss), a Robust Regression (huber) and a Linear Support Vector regressions with two different loss functions (epsilon_insensitive and squared_epsilon_insensitive). All models will be tested to see which works best for the label sets of this data. These last six models are all minimized using first-order Stochastic Gradient Descent learning routine. In this research scikit-learn, MultinomialNB [76] will be used for the multinomial Naive Bayes classifier and SGDClassifier [77] will be used to train the six machine learning models with the different loss functions. In the implementation of scikit-learn, the model used is given with the loss function. All ten combinations of models and loss function given above are tested to determine which works the best for each of the labels sets. For the classifier some other variables will be tested as well; For the Naive Bayes classifier an additive laplace/lidstone smoothing parameter will be tested (alpha) with the values 1, 1e-1, 1e-2 and 1e-3 and whether to learn class prior probabilities or not (fit_prior). For the other ten classifiers the following parameters will be tested: a constant that multiplies the regularization term (alpha) with the values 1, 1e-1, 1e-2, 1e-3, whether or not the intercept should be estimated (fit_intercept), the stopping criterion (tol)

SGDClassifier	ngram_range:	(1,1), (1,2), (2,2)
	stop_words:	True, False
	use_idf:	True, False
	smooth_idf:	True, False
	sublinear_tf:	True, False
	loss:	hinge, squared_hinge, modified_huber, log, perceptron, squared_loss, huber, epsilon_ insensitive squared_epsilon_insensitive
	alpha:	1, 1e-1, 1e-2, 1e-3
	fit_intercept:	True, False
	power_t:	0.9, 0.7, 0.5, 0.3, 0.1
	shuffle:	True, False
	tol:	1, 1e-1, 1e-2, 1e-3, 1e-4
MultinomialNB	ngram_range:	(1,1), (1,2), (2,2)
	stop_words:	True, False
	smooth_idf:	True, False
	sublinear_tf:	True, False
	use_idf:	True, False
	alpha:	1, 1e-1, 1e-2, 1e-3
	fit_prior:	True, False

Table 4.5: The parameters used for testing the loss functions of the Stochastic Gradient Descent and Naive Bayes

with values 1, 1e-1, 1e-2, 1e-3, 1e-4 where training stops when loss > best_loss - tol, whether the training data will be shuffled after each epoch or not (shuffle), the exponent for inverse scaling leaning rate (power_t) with values 0.9, 0.7, 0.5, 0.3, 0.1. An overview of the parameters and the values for each parameter is given in table 4.5. To test these parameters, a grid search is performed. In this research GridSearchCV[78] from scikit-learn is used. The grid search will always be performed with a 5-fold cross validation. Afterwards the best way of labelling will be determined: keyword search, the label set 'Reduced', the label set 'Simple' or the binary label sets. To determine the best way of labelling the accuracy, precision, recall and for the binary sets the precision at 100% recall will be calculated and compared.

The accuracy, precision and recall of imbalanced data can be very misleading due to the large difference in positive and negative labels. Therefore the precision at 100% recall will also be taken into account. The precision at 100% recall is a method where the precision and recall are calculated for different thresholds. Each threshold will have a different precision and recall. For some threshold, the recall will be 100%, and therefore the there will be no false negatives. This means that at that threshold all consults that are labelled positive for a disease contain all consults that contain the disease (100% recall). The precision at 100% recall is then the percentage of positively labelled consults that actually are positive for that disease. The precision at 100% recall is a good metric when working with imbalanced data.

4.2.2 Weather data

The weather data is already quite structured but some values are missing. In some cases, the values are missing because the stations miss the equipment. In other cases, the values are missing for a couple of days or months, probably due to a malfunction in one of the measurement equipment. More about this can be found in 4.1.2.

The medical files of the veterinarian data start on 20 January 1999, but as shown in table 4.2 the majority of variables for Heino are stable from 13 November 1998. This means that there are only a few months of weather data before the first entry in the medical file. In the year 1999,

there are four entries, one in January which only states ”Opgenomen in dit bestand” (added to this document), two in August (belonging to the same consult), and one in November. One or more years of weather data exist before the appointments, for all appointments except one.
As shown in table 4.3 some of the values from precipitation duration and all of the values form barometric pressure, visibility, and cloud coverage are missing. This is probably because the equipment for these measurements are not available at weather station Heino. To assess whether or not the variables are crucial, an equine specialist ranked the categories of table 4.2 from most to least impact on the health of horses in his experience. Dr. G. Kampman ranked the variables as follows:

1. Humidity
2. Wind speed
3. Temperature
4. Precipitation duration
5. Wind direction
6. Sunshine duration
7. Precipitation amount
8. Barometric Pressure
9. Cloud coverage
10. Visibility
11. Evaporation

The precipitation duration is ranked as number 4, therefore it would be valuable for this research to impute the missing values of DR. Barometric pressure is proven to play an important role in colic in horses [18, 21]. These variables will be imputed into this data set as well.
For the imputation of missing values, removal of the rows with missing data points is a technique that is used often [79]. This is not desirable for this data since this will cause a gap of a few months in which we cannot do any predictions. Another approach is filling in the mean, median, or zero for missing values but this also might cause disruption in time series data [1]. El-Nesr [1] therefor suggests using imputation methods depending on the time. Consequently, all 11 methods proposed in [1] will be explored in this study.
Linear interpolation is used by Yu et al. [80] for four or fewer consecutive missing data points. If more are missing, the rows will be removed. Since there are a couple of months missing in a row, linear interpolation will not give an accurate imputation. Rodriguez et al. [81] tested whether k-Nearest neighbours (kNN) or Artificial Neural networks performed better on their data. No combination was found superior compared to the others for the reconstruction of the missing data. Troyanskaya et al [82] also used kNN for prediction of the missing values which, they show, is more robust than Singular Value Decomposition (SVD) as imputation method. KNN could be a good option for weather data, since it can be expected that, with similar weather, most weather conditions will be similar. Therefore, a kNN approach will be tested for the imputation of this data. For this, the KNNImputer of skikit-learn [83] of Python will be used, as explained in [84].
Imputing the values of the barometric pressure is more challenging since all values for Heino are missing, meaning no test can be done to determine how well any approach works. But since we have access to the data of weather stations surrounding Heino, that have values for the barometric pressure, these data sets can be used to get an indication of the barometric

pressure to work with. First, the values will be compared by calculating the R^2 between all combinations of the four data sets to see how much the data differs between the weather stations. If the data does not differ much between the weather stations, one of the other weather stations can be used. After imputation, this weather data consists of 34 variables. The 34 variables can be divided into nine main categories.
11 of the variables are the hour in which X is measured, where X is an other variable, are not considered to be interesting for this study since this does not give any information about the weather. These variables are FHXH, FHNH, FXXH, TNH, TXH, T10NH, RHXH, PXH, PNH, UXH and UNH. In the rest of this research, the 23 other variables will be considered as the weather variables.
Since one can assume that longer periods of a certain kind of weather affect the health of horses, it is undesirable to link the weather of the day of the appointment to the appointment. Instead, an investigation into the weather before the appointment should be done, focusing on the seasons before the appointments. For example, the average temperature during the winter, or the average amount of precipitation during the summer. It would also be interesting to focus on the periods closer to the appointments, for example, the average precipitation in the two weeks before the appointment. Thus, for each of the 23 weather variables, the following variables are added to the table:
The four averages of the prior seasons, the average of the past 30 days, the average of the past 14 days, the prior four separate days. This way, the weather before each appointment is already available in the table and does not have to be calculated for each appointment separately. This gives a total of 161 weather variables.

4.3 Visualization of the data

To get a better insight into the diseases, visualizations for the investigated diseases will be created. The percentage of consults concerning the disease will be plotted against the years and the actual number of consults concerning the disease will be plotted against the months. For the year, the percentage of consults are plotted since the total number of consults grows over the years. This normalization is not desirable for the overview over the months since the purpose of the figures is to show differences in the appearance of disease over the months and figure A.6 shows that there are more consults during the summer months. As the weather conditions change over the months, it can be expected to see this reflected in the visualizations. This especially holds for laminitis, since this is expected to increase during spring and autumn. To visualize the impact of changes in the weather, the values of the weather variable for daily mean temperature (TG), daily mean sea level pressure PG, daily mean wind speed FG, daily mean precipitation duration DR and daily mean humidity UG are plotted against themselves on different days $d_{t-1}, d_{t-2}, d_{t-3}, d_{t-4}$ and the average value over 14 and 30 days. The y-axis then contains the value on $d - t$ and the x-axis contains the value of the later date or the average. The dots in the plot are coloured red when one or more consults concerned the disease on d_t and blue when non of the consults concerned the disease on d_t. If a correlation exists between an increase or decrease of the variable, the lower right or upper left corner will contain more red coloured points, respectively.

4.4 Correlations between horse health and weather

This section describes the method to address the first research question RQ1. To answer RQ1 the correlation between the weather conditions and the diseases will be investigated. According to Dr. G. Kampman, humidity, wind speed, temperature, precipitation duration and wind direction are the five categories that have the most impact. The literature suggests different categories for each of the diseases; Colic is mostly associated with (low) temperature, (high) barometric pressure, (high) humidity and snow. For laminitis, this is (lack of) precipitation and (low) temperatures. Skin diseases seem to be increased by (long-term) precipitation, (high) humidity and (high or low) temperatures. Respiratory diseases do mostly occur in (high or low) temperatures and (high) humidity.

Many different strategies are used in the related work when it comes to the period of weather data used. The most common options are the day of the appearance of the disease, a few days before the appearance of the disease and average of the month in which the disease occurred. The best option for the number of days before the disease is probably depending on the disease. According to Dr. G. Kampman, colic occurs by a change in the weather, mainly at the transition of summer to autumn. The sugar content of grass can vary per day, therefore laminitis should be seen more at days after cold nights. For a higher occurrence of skin disease, there should be high humidity for approximately 14 days. For the occurrence of respiratory disease, it will depend on the cause of the problem. Influenza will occur more during autumn, winter and early spring. While COPD often occurs in winters due to dust in the stables or summer due to drifting of sand. Therefore, the following periods will be tested: The four averages of the four seasons prior before the disease, the average of 30 days before the disease, the average of 14 days before the disease, the four separate days before the disease and the day of the disease.

In the related works, logistic regression [7, 11, 14, 15, 16, 18, 21], Pearson's correlation coefficient [10, 20] and Spearman correlation [19, 21] are used most often to find correlations between the number of cases of a disease and the weather in a given time frame. In this data, there are few cases per disease, with keyword search only 671 colic cases over 18 years and even fewer cases of the other diseases. A large time frame should be used to have one or more consults concerning the disease. But these large time frames do not correspond well with the weather data as these will take an average over a long period.

Given the data of this study Students T-Test, as used in [85] to find significant differences when studying the effects of weather conditions on acute laryngotracheitis, could work well. One of the assumptions for T-Tests is normally distributed data. The distributions of the weather variables are shown in figure 4.11. Most of the weather variables are not normally distributed as clearly shown in the histograms. Since the data is not normally distributed, a non-parametric test is needed. Roussel et al. [86] use permutation tests to find correlations between the weather and seasonal influenza. A permutation test is a non-parametric statistical test that makes few assumptions about the data [87] but a lot of computations are required. Modern computers can handle the required number of computations therefore this is no problem. A permutation test calculates the difference in mean between the two given sample sets. In this study, the one sample set consists of the weather values for each date on which the disease occurs and the other sample set consists of the weather values of the remaining days. This means that whether values of days on which the disease occurs multiple times are used multiple times as well. The samples of the groups will be shuffled to create two random groups, again the difference in mean will be calculated. After shuffling the groups a set amount of times, the probability will be calculated that the difference in means of the original groups actually is significant or that it emerged by chance. In this research, the data will be shuffled 1000 times.

The null hypothesis of the permutation test is that there is no significant difference between the means of the original groups. The permutation test gives a P-value and the difference in mean between the groups. The P-value is the probability that the null hypothesis indeed is true and therefore the groups have no significant difference in mean. If the P-value is small enough the null hypothesis can be rejected, meaning that the means of the groups are significantly different. When performing multiple tests on the same data the number of false positives, the false discovery rate (FDR), can go up but permutation-based methods are known for controlling the false-positive rate. A standard threshold can be used. In this research, a threshold of 0.01 will be used.

4.5 Predictions on horse health

The methodology for addressing the second research question RQ2 is given in this section. For addressing RQ2 occurrence of the different diseases, described in this paper, will be predicted given the current weather conditions. Research that has made predictions based on weather conditions are investigated to determine what prediction methods are suitable for this type of predictions. The results of this investigation are shown in table 4.6. Ensemble predictions are used four times out of nine [88, 89, 90, 91], of which one is combined with an artificial neural network [88]. Artificial neural networks are also used in one other paper [92]. Also different types of regression models are used four times as prediction models [93, 94, 95, 96]. In [95], a special type of regression model is used: Time-varying periodic splines. In the search to the relation between the weather and colic, logistic regression is used seven times to find correlations [7, 11, 14, 15, 16, 18, 21].

Since ensemble prediction models seem to perform well on similar data [88, 89, 90, 91], an ensemble prediction model will be trained for this data to predict the occurrence of diseases on horses, given the current weather. The models in the ensemble prediction can be combined in multiple ways. Brownlee [97] describes the most popular three: Boosting, Bagging and Voting. For boosting a gradient boosting classifier is trained, using decision trees. Bagging and Voting are performed using Logistic Regression, Support Vector Machines, Random Forest and Neural Networks as estimators. Bagging is performed four times per disease, each time with an other estimator. Voting is performed once with the four estimators to create one combined model. These four models are selected as these methods are commonly used for binary classification and therefore are expected to give good results for this data es well [98]. Logistic Regression is already proven to be successful on similar data [93, 94, 95, 96]. The Artificial Neural Networks did not perform well in the related work, Brace et al. [92] contribute this to the fact that artificial neural networks are not designed for time series data. Support Vector Machines and Random Forest are not used in the related work.

Separate models will be trained for the four diseases. For addressing RQ2, the results of the models will be used. The results of the models will be compared to the results of a single classifier Linear Regression, Support Vector Machine, Decision Tree and Neural Network.

	Year	Predicts	Method	Features	Results
[93]	1978	Corn yields	4 regression models	Mean temperature and precipitation	Model 1 did not perform well. Models 2, 3 and 4 had a prediction error of ∼10 bushels/acre
[94]	1983	Migration of insects	Linear regression	Sunshine, rainfall, max, min and mean temperature	The start of the migration was predicted well. The end not that well.
[95]	1993	Electricity load	Time-varying periodic splines	Temperature (others are not specified)	Predicted better by warm temperatures
[92]	1993	Electricity load	Artificial neural network	Temperature (others are not specified)	Did not perform well compared to other models
[96]	1997	Electricity load	Multiple regression models (one for each hour)	Temperature (others are not specified)	Performed very well
[88]	2002	Electricity load	Neural network with weather ensemble predictions	Temperature (others are not specified)	Performed better than single point forecasting
[89]	2003	Electricity demand	Ensemble predictions	Temperature, wind speed and cloud cover	Performed better than single point forecasting
[90]	2008	Flood	Grand ensemble predictions	Rainfall	Performed better than single point forecasting
[91]	2009	Flood	Grand ensemble predictions	Rainfall	Performed better in dry compared to wet months

Table 4.6: Results of the literature study to methods used for predictions on weather.

5 RESULTS

This section gives an overview of the results and findings of this research.

5.1 Prepare data

The data is prepared as described in 4.2. The results are given in this section

5.1.1 Imputation missing values weather data

The missing values in the weather data set are found using the imputation methods as given in the article form El-Nesr [1]. In this article, a few different imputation methods are tested. The proposed methods, Rolling mean, and Rolling median do not work for this problem because these methods work with a sliding window and this data set misses 143 consecutive data points. Meaning that the window must be larger than that. The results of the different imputation methods are given in the Appendix, Table B.1. Filling in the median of the data gives an R^2 of 0.99099, the highest value for R^2 for theses imputation methods.
In addition, K-Nearest neighbours is used for the imputation of the missing values as described by Abdul Majed Raja [84]. When using two or more neighbours, the FillMedian method from El-Nesr [1] is outperformed. When ten nearest neighbours are used the R^2 score is 0.99556, this is the highest R^2 score achieved. The results are shown in the Appendix, Table B.2.

5.1.2 Imputation missing variables weather data

The barometric pressure is not available in the weather data set of Heino. But, as mentioned in section 4.2.2 access to this data would be preferred. Four of the weather stations around Heino do have data on the barometric pressure. These are the weather stations of Twente, Deelen, Lelystad, and Hoogeveen. The data sets of these weather stations have no missing data for the barometric pressure from 1993 onwards, or even far before this date.
To get more insight into the variation of barometric pressure between the weather stations, the R^2 values are calculated for each combination of weather stations for the different variables. The values of R^2 are given in the Appendix, Table B.3. For the variables PG, PX and PN we have an R^2 between 0.985 and 0.996. The PXH and PNH have an R^2 value between 0.813 and 0.892. This means that the barometric pressure, measured at one day is very similar for these four weather stations. The hour on which the maximum and minimum pressure is measured varies a bit more. The lowest values for R^2 are given for combinations Lelystad-Twente and Deelen-Hogeveen these are the weather stations with the biggest distance between them. Taking this into account, the barometric pressure measured at weather station Hoogeveen will be imputed into the weather data set. This decision is made because the weather station of Hoogeveen is the closest to Heino and Den Ham.

Colic		Predicted	
		1	0
Actual	1	726	9
	0	39	10065

Laminitis		Predicted	
		1	0
Actual	1	176	32
	0	20	10611

Respiratory		Predicted	
		1	0
Actual	1	246	45
	0	27	10521

Skin		Predicted	
		1	0
Actual	1	577	39
	0	61	10162

Table 5.1: Confusion matrices for keyword search of the diseases

5.1.3 Occurrence of diseases

After merging the medical file into consults, the data contains 58927 consults. The occurrence of the four diseases used in this paper, using keywords search is as follows; ”koliek” (colic) occurs in 765 consults, ”hoefbevangen” (laminitis) occurs 196 times. To detect the occurrence of skin disease, ”schimmel” (fungal) and ”mok” (mud fever) were used which occurs 324 and 326 times respectively in 638 consults. For respiratory problems, the following keywords were used: ”snot” (snot): 150, ”luchtweg” (airway): 71, ”bronchitis” (bronchitis): 60, ”longontsteking” (pneumonia): 7, one or more of these keywords where found in 273 consults. The context of these words is not known, therefore 10839 consults are labelled by hand. Resulting in seven label sets. The distribution of the labels can be found in the Appendix, Table C.1. The label set ’All’ contains multiple terms per consult for some of the consults, this results in a higher total number of labels. A total list of labels can be found in the Appendix, Table C.2. There is a total of 121 (combinations of) labels in the label set. The accuracy of the keyword search is tested using the hand-labelled data. For this, the binary label sets are used to create a confusion matrix for each of the four diseases. The confusion matrices are shown in Table 5.1 and the accuracy, precision and recall for keyword search are shown in Table 5.3. For the classification, a naive Bayes classifier and a Stochastic Gradient Descent with six different classifiers and a selection of parameters is tested to see what classification method and parameters would work best for each of the label sets. For the selection of the parameters, a grid search is used to tune the parameters. The grid search returns the combination of parameters that performs best. The parameters that are found to be the best for the different label sets are given in Table 5.2.
For the different label sets, the accuracy, precision, recall and precision at 100% recall, obtained with the parameters given in Table 5.2, are given in Table 5.3. The confusion matrices for the label sets ”colic”, ”laminitis”, ”respiratory” and ”skin” label sets are shown in Table 5.4. Table 5.5 shows the number of occurrences of the diseases in the different label sets. The keyword search has a very low count compared to the others. Also the expert confirms that the counts of the keyword search are to low to be realistic. The binary label sets do have a higher count for almost all diseases than the ”simple” and ”reduced” label set.

5.1.4 Temperature

In total, 2974 potential temperatures are found. These temperatures are found over 2308 consults. For each of these found temperatures, the text is checked by hand to ensure that the number indeed is the given temperature. This was true for 1698 of the found temperatures.
The number of correctly labelled temperatures per potential temperature are given in the Appendix, Table D.1. Potential temperatures with two decimal numbers and whole numbers are labelled wrong more often then others. Mainly 40.0 is found often in the texts and almost always wrong. This is because, apparently, many medicines are given in 40 cc dose. When selecting the numbers between 36 and 43 out of the text, an accuracy of 0.98344 can be obtained, as

	Colic	Laminitis	Respiratory	Skin	Reduced	Simple
ngram_range	(1,1)	(1,1)	(1,1)	(1,1)	(1,1)	(1,1)
stop_words	True	True	False	False	True	True
use_idf	False	False	False	False	True	True
smooth_idf	True	False	False	True	True	True
sublinear_tf	True	True	True	True	True	True
Classifier	SH	MH	H	SL	MH	MH
alpha	1e-3	1e-2	1e-3	1e-2	1e-3	1e-3
fit_intercept	False	True	True	True	True	True
power_t	0.5	0.5	0.5	0.5	0.5	0.5
shuffle	True	True	True	False	False	True
tol	1e-4	1e-4	1e-2	1e-4	1	1

Table 5.2: For each label set the classifier and parameters that have given the best results and will be used for labelling the data. The used classifiers and loss functions are: squared_hinge (SH), modified_huber (MH), hinge (Hi) and squared_loss (SL)

		Accuracy	Precision	Recall	Precision at 100% Recall
Keyword	Colic	0.99557	0.94902	0.98776	
	Laminitis	0.99077	0.90439	0.93669	
	Respiratory	0.99336	0.90110	0.84536	
	Skin	0.99520	0.89796	0.84615	
Binary	Colic	0.98893	0.88432	0.95822	0.47676
	Laminitis	0.97509	0.42009	0.92000	0.23866
	Respiratory	0.98727	0.71341	0.84173	0.12225
	Skin	0.84613	0.26667	0.97709	0.17053
Reduced	Colic	0.64225	0.92394	0.93181	0.45774
	Laminitis		0.85227	0.76530	0.04239
	Respiratory		0.73684	0.91304	0.07360
	Skin		0.93727	0.83828	0.10670
Simpel	Colic	0.91605	0.91375	0.96307	0.46255
	Laminitis		0.84466	0.87000	0.16501
	Respiratory		0.78808	0.86232	0.06565
	Skin		0.91753	0.87541	0.08660

Table 5.3: Accuracy, precision, recall and precision at 100% recall for the different label sets and the keyword search. Binary is a combination of the results form the ”colic”, ”laminitis”. ”respiratory” and ”skin” label sets. Using the parameters and classifiers from table 5.2

Colic		Predicted	
		1	0
Actual	1	343	16
	0	46	5015

Laminitis		Predicted	
		1	0
Actual	1	93	7
	0	124	5195

Respiratory		Predicted	
		1	0
Actual	1	116	23
	0	44	5237

Skin		Predicted	
		1	0
Actual	1	300	9
	0	826	4285

Table 5.4: Confusion matrices for prediction of the label sets colic, laminitis, respiratory and skin, using the parameters and classifiers from table 5.2

	reduced	simple	binary	keyword
koliek (colic)	1859	2082	4196	765
hoefbevangen (laminitis)	1321	1934	5037	196
luchtweg (respiratory)	2439	1825	2007	273
huid (skin)	1560	2205	18135	638

Table 5.5: The number occurrence of the diseases in the different label sets. Binary is the combination of the ”colic”, ”laminitis”, ”respiratory” and ”skin” label sets. Keyword is the count of the keyword search. Using the parameters and classifiers from table 5.2

shown in the Appendix, Table D.2. When specific numbers, that are labelled incorrectly often, are removed as potential temperature the accuracy goes up to 0.99805. Table 5.6 shows the accuracy for different approaches. Removing numbers that are correctly labelled 0% of the time, as given in the Appendix, Table D.1, gives us an accuracy of 0.98474. When only the number 40.0 is removed, the accuracy is 0.99652. This can be explained by the fact that 40.0 is labelled 1026 times but only 17 of these are correct. Therefore this has a big impact on the overall accuracy.

5.2 Visualization of the data

To get more insight into the data and possible correlations between the seasons or the weather and the diseases, visualizations are made. For each of the diseases, a figure is created giving an overview of the percentage of consults per year concerning the disease. As well as a figure of the number of consults concerning the disease for each month, combining all years.
Figure 5.1 shows the percentage of consults concerning the different diseases over the years. Colic and respiratory disease have a high percentage in the early years. This is due to the low total number of consults in these years.

Removal of ...	Accuracy
... none	0,98344
... 0% correct	0,98474
... less then 10% correct	0,99783
... less then 20% correct	0,99805
... less then 50% correct	0,99803
... the number 40.0	0,99652

Table 5.6: The accuracy for the temperatures when specific numbers are considered as not relevant. The percentages, and numbers that are removed, are shown in table D.1

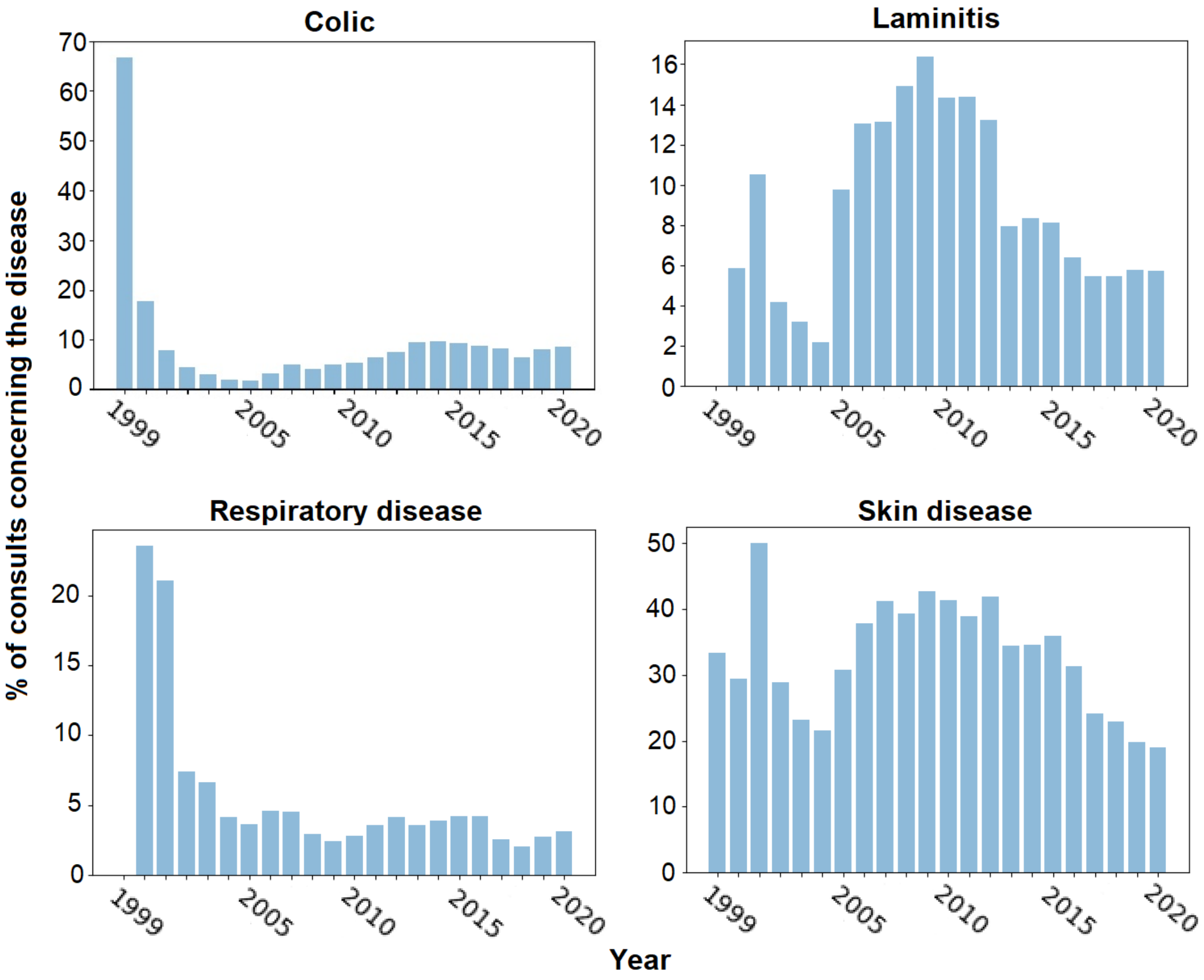


Figure 5.1: The percentage of consults concerning the disease over the years compared to all consults of that year

Figure 5.2 shows the count of consults concerning the different diseases over the months. The numbers of consults concerning the diseases do not differ a lot trough out the year but they mostly seem to be in accordance with the related work.

The number of consults concerning colic lower during the summer months and highest from March to June and October to December. This is as expected by the expert. The literature pointed to a higher risk of colic in autumn and spring [12, 16, 19] compared to the rest of the year. Archer et al [15] observed more colic cases during December, January and February, but the results show that there is no rise in the number of consults in January and February in this data.

The number of consults concerning laminitis goes up from February and peaks in July and October. The expert explains the peek at the end of the year by the fact that horses nowadays are capt on the grass till late in the year due to climate change but the quality of the grass deteriorates throughout the year. The rising number of cases of laminitis during the spring and beginning of summer can be explained by the higher amounts of sunshine, as Menzies-Gow et al [24] also observed. The higher number of cases of laminitis during January, compared to December and February can be explained by the low temperatures [37, 38, 39, 40]. This all is in accordance with Wylie et al. [23] who found higher risks in summer and winter months.

Some studies [44, 51, 53, 54, 55] found that cold weather could cause respiratory disease, this data shows a somewhat higher count of consults concerning respiratory disease during January, compared to December and February. Most of the consults that concern respiratory

disease are seen in spring as Laurant et al. [52] also observed in their data, but they found winters as even more risk full. High temperatures also are found to cause respiratory and skin disease [45, 46, 47, 48, 49, 63, 65, 66, 67] this is not supported by this data but this can be due to the mild climate of the Netherlands. The expert also was surprised at first by the high numbers in the spring and lower numbers in the autumn but he could explain this by the fouls that are born in the spring that are more susceptible for respiratory diseases.
During the winter months, the number of consults concerning skin disease is lowest. These numbers rapidly climb to the highest in May. The low number of skin diseases during the winter is in contrast with the related work suggesting that low temperatures could cause skin disease [60, 62]. Changes in the weather are mostly considered to cause diseases in horses. If this is indeed true, a scatter plot with the weather values on the day of the disease d_t and one of the days before the disease would show a different pattern than the weather values on days without the disease. Figure 5.3 shows scatter plots of the weather values of DT on d_t against the average weather values of DR over the 14 days before d_t for the days with and without laminates and the weather values for TG on d_t against the weather values on d_{t-1} for the days with and without colic. The dots corresponding to the days with the disease are plotted later than the class without the disease and since there are so many points, the first plotted dots are not visibly anymore. The histograms on the top and right side of the plot show the count of the dots for d_t and d_{t-1}/avg14. Since the histograms have the same size for the days with and without the disease, the shapes of the blue and orange dots are most likely the same as well. For each combination of disease and weather variable, such a plot is made with only 100 randomly selected dots per class. These plots are shown in the Appendix, Figures E.1, E.2, E.3 and E.4.

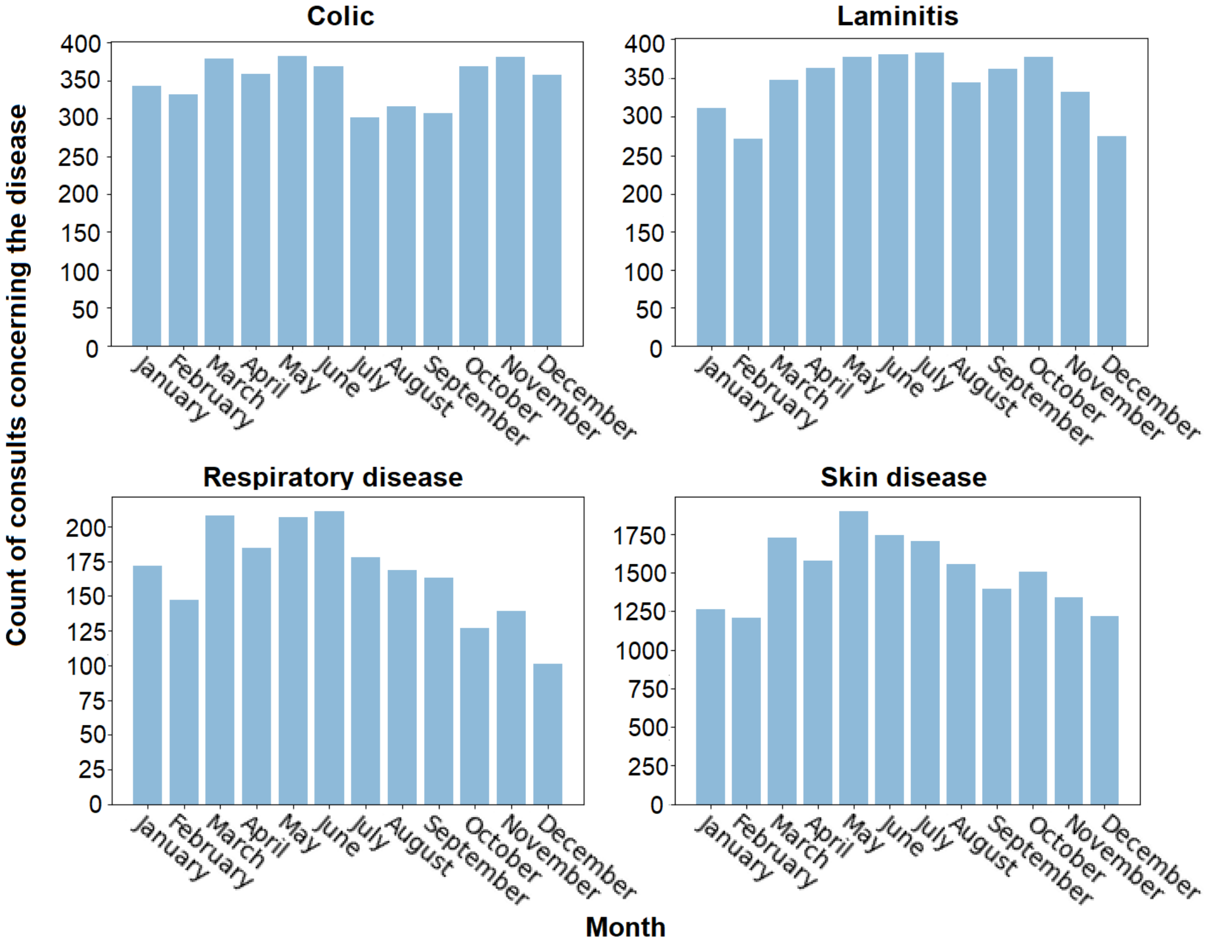


Figure 5.2: The number of consults concerning the disease over the months for all years

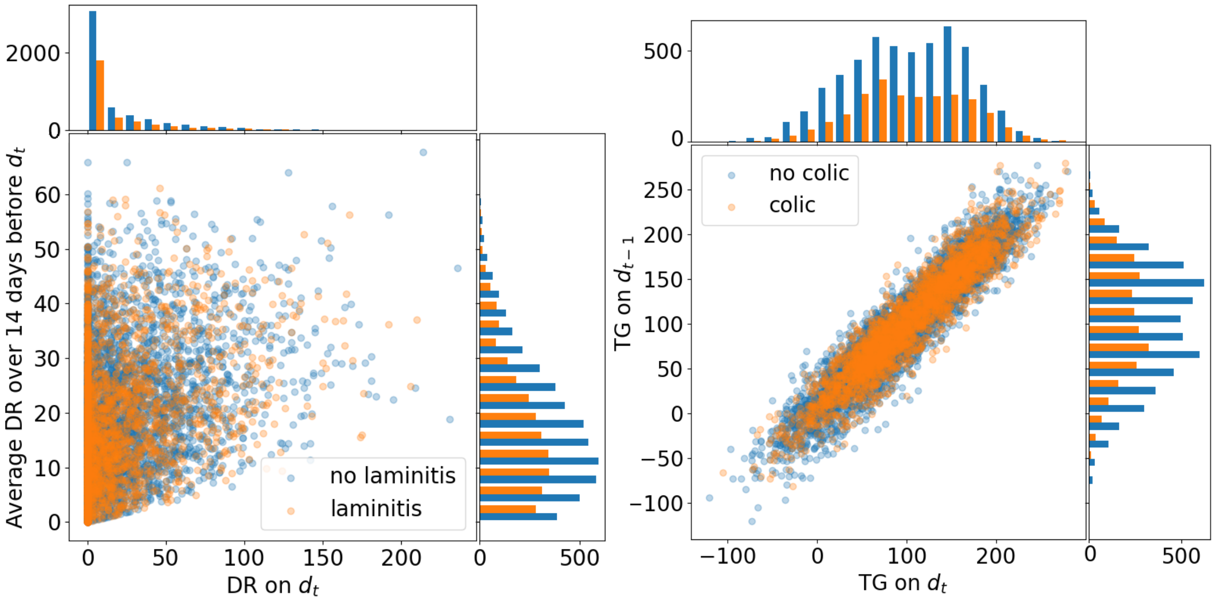


Figure 5.3: Scatter plots with histograms of the weather values of DR on d_t plotted against the average of the weather values DR over 14 days before t_t for days with and without laminitis and the weather values of TG on d_t plotted against the weather values of TG on d_{t-1} for days with and without colic.

These figures also show a similar shape for the different classes. The columns contain (left to right) the weather variables that are chosen for this, these are the daily mean temperature (TG) in 0.1 degrees Celsius, the daily mean sea level pressure (PG) in 0.1 hPa, the daily mean wind speed (FG) in 0.1 m/s, the precipitation duration (RH) in 0.1 hours and the daily mean relative atmospheric humidity (UG) in percentage. The rows (top to bottom) are for each of the weather values above the value of day d_t against the day before (d_{t-1}), two days ago (d_{t-2}), three days ago (d_{t-3}), four days ago (d_{t-4}), the average over 14 days prior to d_t and 30 days prior to d_t. Dots in the plots are coloured red when one or more consult on d_t concerns the given disease. When a drop or raise in one of the weather values is associated with the disease, the majority of red dots are expected in the upper left corner or lower right corner of the plot respectively. The locations of the red dots in the plots of all the figures are distributed very similarly to the locations of the blue dots. This means that there are no obvious correlations between changes in these variables and the diseases.

5.3 Correlations between horse health and weather

One of the goals in this research is to find out whether the observed diseases do correlate to one or more of the measured weather variables. For the correlation analysis, permutation tests are used. The results of the correlation analysis can be found in the tables 5.7, 5.8, 5.9 and 5.10. The cells that are coloured are variables that are considered correlated to the disease. The green cells indicate that the difference in mean is higher than zero and therefore the weather variable is significantly higher when the disease occurs. The red cells indicate a difference in mean is lower than zero and so the weather variable is significantly lower when the disease occurs, for a p-value smaller 0.01 on the permutation test. For most of the variables that have a negative correlation to one or more diseases the difference in mean is very small. The variables that are positively correlated to one or more diseases generally show bigger differences in mean. Another thing that does stand out is that the weather variables all are either positively or negatively correlated to the different diseases. The wind is for example always negatively correlated to the diseases. Sometimes there is a positive difference in mean but then the variable is not correlated.

For each of the diseases a correlated weather variable is chosen and for that variable 300 data points (150 with and 150 without disease) are plotted as a swarm plot and box plot. These plots are is shown in Figure 5.4. Less wind during the two weeks and month before the occurrence of colic, more sunshine duration two days and more before the occurrence of colic, less rain during the month before colic, lower barometric pressure and lower humidity are correlated to colic. The lower humidity is measured for all different time options for the daily mean and the minimum humidity. The lower barometric pressure is measured 2 days, up to the average of 14 days before the colic, again for the daily mean and minimum pressure. Also, the maximum barometric pressure is lower in the average of 14 days before the colic occurs.

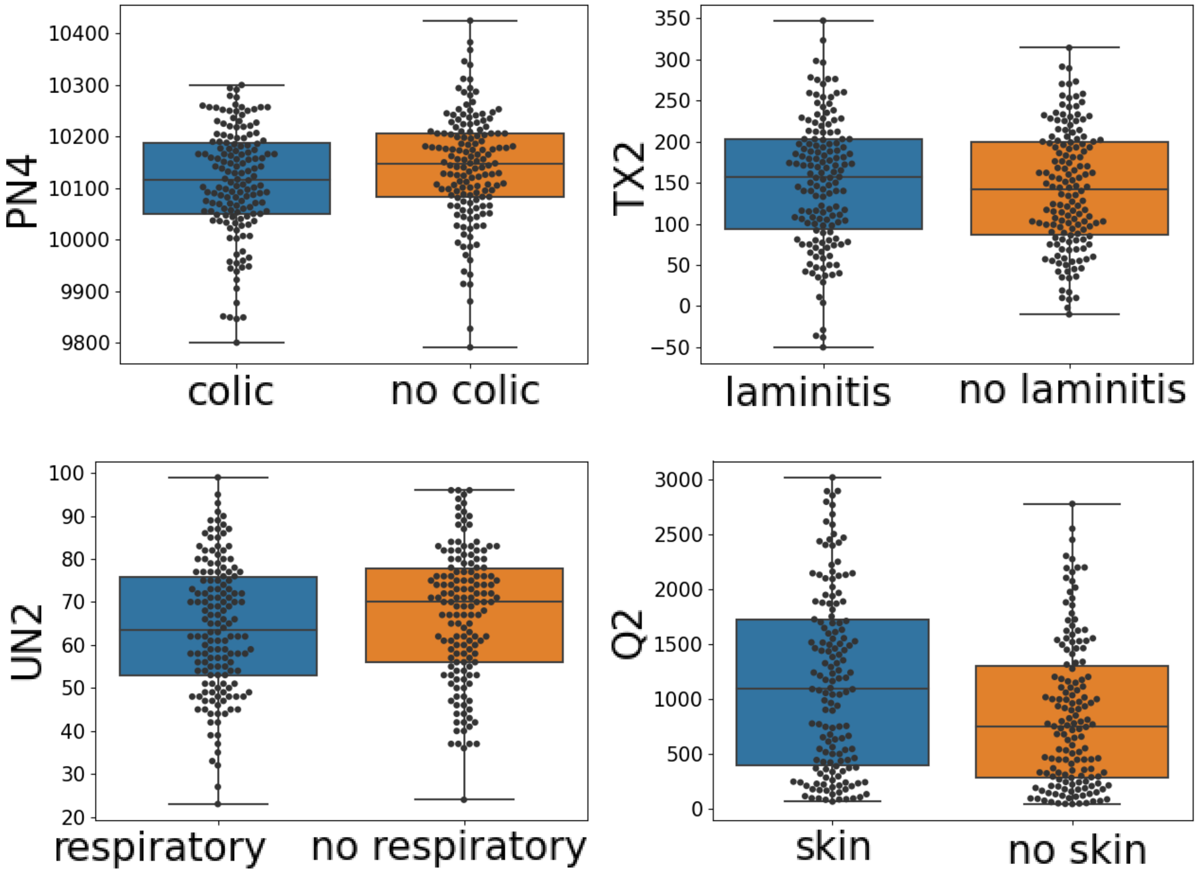


Figure 5.4: For each of the diseases, swarm plots and box plots of the values of a correlated weather value.

Less wind, higher temperatures, more sunshine duration, less rain, lower mean and minimum humidity, higher maximum humidity and more potential evapotranspiration are correlated to the occurrence of laminitis. Except for the precipitation and pressure, which have only a view variables correlated to laminitis, all weather categories have mostly all or non of the time options (the whole row) correlating to the disease for the specific variable. This indicates stable weather.

Except for the barometric pressure, all categories have one or more variables correlating to respiratory disease. For the correlations hold less wind on the current day, a view days in advance and during the two weeks/month, higher temperatures, more sunshine duration, less rain a view days in advance and during the two weeks/month and evapotranspiration are correlated to the occurrence of respiratory disease. As with laminitis, in general, all options of a weather variable of a category do correlate or none at all. Again, indicating stable weather.

All categories have one or more variables that are correlated to skin disease. Less wind, higher temperatures, more sunshine, a higher minimum barometric pressure, less precipitation, lower main and minimum humidity and higher maximum humidity and higher potential evapotranspiration all are related to the occurrence of skin disease. As with laminitis and respiratory disease, most of the rows are completely coloured, indicating stable weather.

	d_t		d_{t-1}		d_{t-2}		d_{t-3}		d_{t-4}		$avg14$		$avg30$	
	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$
DDVEC	0.82917	0.35608	0.63337	0.90222	0.79121	0.41347	0.63936	-0.72608	0.19181	-2.30417	0.28172	-0.87379	0.38961	-0.51013
FHVEC	0.01399	-0.76387	0.20979	-0.39589	0.50749	-0.21324	0.38561	-0.27110	0.79920	0.08339	2.00E-03	-0.60492	2.00E-03	-0.82950
FG	0.01998	-0.68584	0.31568	-0.29360	0.45754	-0.23566	0.37762	-0.26128	0.72128	0.10616	2.00E-03	-0.59902	2.00E-03	-0.78888
FHX	4.00E-03	-1.17543	0.27972	-0.44238	0.22577	-0.49253	0.18781	-0.55380	0.51548	0.27358	2.00E-03	-0.81869	2.00E-03	-1.04504
FHN	0.05594	-0.50122	0.92507	0.01793	0.93307	0.01902	0.75724	0.10035	0.60739	0.12914	0.03796	-0.30806	7.99E-03	-0.38434
FXX	0.36763	-0.64850	0.36963	0.66358	0.69331	0.28593	0.55544	0.46497	0.03996	1.50386	0.52348	-0.23007	0.02398	-0.68720
TG	0.23576	1.54996	0.16783	1.77569	0.14585	1.97262	0.10390	2.22266	0.06394	2.44107	0.09391	1.98827	0.10789	1.86580
TN	0.86913	0.22227	0.61139	0.58979	0.59540	0.58710	0.68332	0.44145	0.37762	0.97807	0.60539	0.49658	0.64336	0.43663
TX	0.08991	2.77860	0.09391	2.65047	0.04995	3.08160	0.01598	3.78279	7.99E-03	4.00532	0.01598	3.37241	0.02597	3.21569
T10N	0.03197	-2.73724	0.14386	-1.86777	0.07792	-2.11051	0.02398	-2.60678	0.15984	-1.69085	0.01199	-2.37851	5.99E-03	-2.41088
SQ	0.03996	1.76681	0.18382	1.17527	2.00E-03	2.37073	0.01199	2.19483	4.00E-03	2.89358	2.00E-03	1.89383	2.00E-03	1.81896
SP	0.07393	1.11729	0.20380	0.79168	2.00E-03	1.79751	9.99E-03	1.59284	4.00E-03	2.16013	2.00E-03	1.36862	2.00E-03	1.38121
Q	0.13387	24.58175	0.34765	14.97431	0.08392	28.99317	0.07992	29.22613	0.04196	33.35421	0.10589	22.14958	0.15584	18.46423
DR	0.03996	-1.25179	0.18182	-0.81915	0.02597	-1.34612	0.90909	0.05580	0.99101	-0.01271	0.01199	-0.73090	2.00E-03	-0.98614
RH	0.02597	-2.04794	0.96903	-0.05926	0.24775	-1.05909	0.82917	0.19626	0.83317	-0.19471	5.99E-03	-0.86904	2.00E-03	-1.17424
RHX	0.07193	-0.66611	0.67732	0.13538	0.94705	-0.04517	0.41758	0.28984	0.51149	0.23741	0.14785	-0.18624	4.00E-03	-0.31075
UG	2.00E-03	-0.92264	2.00E-03	-1.10100	2.00E-03	-1.33326	2.00E-03	-1.47911	2.00E-03	-1.52943	2.00E-03	-1.09409	2.00E-03	-1.01837
UX	0.23976	0.10784	0.80320	-0.02862	0.61538	-0.05179	0.42957	-0.08358	0.15385	-0.13969	0.18781	0.06583	2.00E-03	0.13225
UN	2.00E-03	-1.52685	2.00E-03	-1.66415	2.00E-03	-1.96612	2.00E-03	-2.42351	2.00E-03	-2.36496	2.00E-03	-1.87380	2.00E-03	-1.81880
PG	0.21379	-2.45405	0.02597	-4.51918	9.99E-03	-5.58359	5.99E-03	-6.21809	7.99E-03	-6.64415	5.99E-03	-3.68130	0.41159	-0.65011
PX	0.29570	-1.92431	0.02997	-4.08816	0.01399	-4.52851	0.01399	-4.98560	0.02198	-4.86161	9.99E-03	-3.03114	0.85115	-0.10605
PN	0.15584	-2.92185	0.02398	-5.02700	5.99E-03	-6.78208	2.00E-03	-7.60243	4.00E-03	-8.15551	4.00E-03	-4.35857	0.17383	-1.27217
EV24	0.16384	0.40344	0.38561	0.24001	0.11389	0.47538	0.09990	0.47608	0.05395	0.55513	0.14386	0.36106	0.20579	0.29607

Table 5.7: The P-values and difference in mean. where $\overline{X}_a$ is the mean of the days with colic and $\overline{X}_b$ is the mean of the days without colic. for the different weather values on days with and without colic. Gray cells have a p-value that is considered correlated. The corresponding differences in mean are colored red and green. depending on a negative or positive correlation.

	d_t		d_{t-1}		d_{t-2}		d_{t-3}		d_{t-4}		$avg14$		$avg30$	
	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$
DDVEC	0.07592	-3.26807	0.04196	-3.54738	0.12587	2.59529	0.01199	4.67398	0.77722	0.51509	0.34965	-0.74608	0.92707	0.06351
FHVEC	4.00E-03	-0.95194	7.99E-03	-0.86365	2.00E-03	-1.03007	2.00E-03	-1.71063	2.00E-03	-1.39857	2.00E-03	-1.36719	2.00E-03	-1.13249
FG	4.00E-03	-0.92094	5.99E-03	-0.80574	4.00E-03	-0.95467	2.00E-03	-1.53571	2.00E-03	-1.25912	2.00E-03	-1.28536	2.00E-03	-1.03917
FHX	0.03397	-0.84029	0.07393	-0.68092	0.04795	-0.81406	2.00E-03	-1.54427	2.00E-03	-1.20859	2.00E-03	-1.24295	2.00E-03	-0.98435
FHN	4.00E-03	-0.90092	2.00E-03	-0.84519	2.00E-03	-1.07761	2.00E-03	-1.32450	2.00E-03	-1.18656	2.00E-03	-1.16396	2.00E-03	-0.97722
FXX	0.30569	-0.69664	0.36563	-0.60272	0.94106	-0.06918	0.01399	-1.59388	0.06593	-1.29359	2.00E-03	-1.19265	4.00E-03	-0.75927
TG	2.00E-03	8.96700	2.00E-03	8.53544	2.00E-03	8.72333	2.00E-03	8.31940	2.00E-03	9.01164	2.00E-03	7.73463	2.00E-03	7.52751
TN	2.00E-03	6.03729	2.00E-03	5.80932	2.00E-03	5.12304	2.00E-03	5.06298	2.00E-03	5.58235	2.00E-03	4.82460	2.00E-03	4.88957
TX	2.00E-03	11.64580	2.00E-03	10.81523	2.00E-03	11.89521	2.00E-03	11.21403	2.00E-03	11.83533	2.00E-03	10.50553	2.00E-03	10.04014
T10N	7.99E-03	3.19876	5.99E-03	3.26003	0.10589	1.90693	0.08392	1.90660	0.01998	2.68962	0.03596	1.93601	0.01998	2.08089
SQ	2.00E-03	4.09395	2.00E-03	3.57599	2.00E-03	5.71260	2.00E-03	4.59882	2.00E-03	4.05770	2.00E-03	4.31941	2.00E-03	4.09992
SP	4.00E-03	1.64525	0.01998	1.46750	2.00E-03	3.10365	2.00E-03	2.45364	4.00E-03	1.82083	2.00E-03	2.05785	2.00E-03	1.90323
Q	2.00E-03	106.22960	2.00E-03	99.76362	2.00E-03	125.38723	2.00E-03	104.57881	2.00E-03	102.14283	2.00E-03	107.00402	2.00E-03	104.87470
DR	0.07992	-1.02911	0.16783	-0.78939	2.00E-03	-1.72423	0.02597	-1.34237	0.09790	-0.95528	2.00E-03	-1.31512	2.00E-03	-1.01238
RH	0.14186	-1.17799	0.24575	-0.98538	0.03397	-1.79669	0.21179	-1.06688	0.99700	0.00629	5.99E-03	-0.70724	0.01399	-0.51234
RHX	0.52747	-0.22723	0.31568	-0.36569	0.42358	-0.28375	0.70330	-0.13713	0.29171	0.38878	0.79121	0.03305	0.66334	0.04449
UG	2.00E-03	-1.36374	2.00E-03	-1.40549	2.00E-03	-1.78219	2.00E-03	-1.52247	2.00E-03	-1.49821	2.00E-03	-1.48839	2.00E-03	-1.47832
UX	0.02997	0.18762	0.71129	0.02856	0.14585	0.14183	0.08991	0.15916	0.25774	0.10543	2.00E-03	0.13990	2.00E-03	0.11236
UN	2.00E-03	-2.76012	2.00E-03	-2.54213	2.00E-03	-3.32745	2.00E-03	-2.90082	2.00E-03	-2.79798	2.00E-03	-2.91268	2.00E-03	-2.81641
PG	0.08392	-3.51491	0.03796	-4.28841	0.07393	-3.64778	0.05195	-3.67582	0.02198	-4.36648	0.25974	-1.44340	0.33167	-0.88823
PX	0.02398	-4.36529	0.01798	-4.82841	0.06993	-3.40272	9.99E-03	-4.67366	0.01598	-4.36960	0.04396	-2.26276	0.04196	-1.63168
PN	0.40559	-1.90712	0.13387	-3.47640	0.09990	-3.51450	0.17982	-2.81684	0.05994	-3.96272	0.79321	-0.40976	0.98501	-0.06746
EV24	2.00E-03	1.91494	2.00E-03	1.77218	2.00E-03	2.16418	2.00E-03	1.82057	2.00E-03	1.81709	2.00E-03	1.85813	2.00E-03	1.81789

Table 5.8: The P-values and difference in mean. where $\overline{X}_a$ is the mean of the days with laminitis and $\overline{X}_b$ is the mean of the days without laminitis. for the different weather values on days with and without laminitis. Gray cells have a p-value that is considered correlated. The corresponding differences in mean are colored red and green. depending on a negative or positive correlation.

	d_t		d_{t-1}		d_{t-2}		d_{t-3}		d_{t-4}		$avg14$		$avg30$	
	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$
DDVEC	0.68332	0.99963	0.62937	1.03204	0.44555	1.59177	0.69530	-0.87269	0.57542	1.39391	0.97502	0.04850	0.63736	-0.36285
FHVEC	2.00E-03	-1.60210	0.01598	-0.92115	0.01798	-0.84112	0.44156	-0.27576	0.06993	-0.74727	2.00E-03	-0.99222	2.00E-03	-0.87446
FG	2.00E-03	-1.46041	0.02198	-0.84379	0.01399	-0.80630	0.72128	-0.12382	0.16184	-0.52334	2.00E-03	-0.83679	2.00E-03	-0.73776
FHX	9.99E-03	-1.40937	0.09391	-0.85673	0.32967	-0.47985	0.84116	0.11713	0.18182	-0.66166	0.01199	-0.67557	4.00E-03	-0.61737
FHN	2.00E-03	-1.65112	7.99E-03	-0.86376	2.00E-03	-1.11679	0.24975	-0.37319	2.00E-03	-1.03508	2.00E-03	-0.97175	2.00E-03	-0.78872
FXX	0.03996	-1.76024	0.62937	-0.36615	0.90110	0.12672	0.38561	0.81961	0.59141	-0.48174	0.31369	-0.45388	0.17982	-0.44918
TG	2.00E-03	8.52992	2.00E-03	7.82240	2.00E-03	7.45474	2.00E-03	6.88147	2.00E-03	7.23046	2.00E-03	7.35517	2.00E-03	6.24172
TN	2.00E-03	4.32979	4.00E-03	4.44834	0.01199	3.85551	4.00E-03	4.22468	2.00E-03	4.30363	2.00E-03	4.06207	4.00E-03	3.12171
TX	2.00E-03	12.31797	2.00E-03	11.01053	2.00E-03	10.81129	2.00E-03	9.79062	2.00E-03	9.49560	2.00E-03	10.42815	2.00E-03	9.07429
T10N	0.50949	1.11121	0.40559	1.32320	0.49550	1.14817	0.22378	2.01371	0.25375	1.83271	0.30370	1.42126	0.67133	0.53733
SQ	2.00E-03	5.03541	2.00E-03	4.97738	2.00E-03	7.14560	2.00E-03	5.76194	2.00E-03	4.26927	2.00E-03	4.85564	2.00E-03	4.68911
SP	4.00E-03	2.07533	2.00E-03	2.44914	2.00E-03	4.14940	2.00E-03	3.21934	0.03197	1.66084	2.00E-03	2.30269	2.00E-03	2.38366
Q	2.00E-03	140.37891	2.00E-03	132.05512	2.00E-03	153.83697	2.00E-03	134.12079	2.00E-03	124.01567	2.00E-03	128.96279	2.00E-03	119.04022
DR	0.01399	-1.83572	0.01399	-1.82948	2.00E-03	-2.58433	0.45954	0.55924	0.98901	0.00226	2.00E-03	-1.29379	2.00E-03	-1.39548
RH	0.03796	-2.22396	0.20979	-1.30973	0.08991	-1.66588	0.09590	1.84806	0.79720	0.25743	0.02797	-0.80249	2.00E-03	-1.18823
RHX	0.02597	-1.02641	0.76923	0.14137	0.66334	-0.18767	0.11588	0.76950	0.88312	0.06723	0.90310	-0.01772	0.03796	-0.23197
UG	2.00E-03	-1.56112	2.00E-03	-1.70178	2.00E-03	-2.20593	2.00E-03	-2.01202	0.00200	-1.61016	0.00200	-1.76412	0.00200	-1.71951
UX	0.02597	0.27864	0.86513	0.01820	0.68531	-0.04533	9.99E-03	-0.31558	0.20979	0.15115	0.77123	0.01975	0.72328	0.01997
UN	2.00E-03	-3.39189	2.00E-03	-3.03723	2.00E-03	-3.80596	2.00E-03	-3.63216	2.00E-03	-3.09680	2.00E-03	-3.24331	2.00E-03	-3.09241
PG	0.55544	-1.42041	0.47752	-1.65851	0.44555	-1.87276	0.08192	-4.34328	0.06993	-4.11286	0.35365	1.32836	0.09391	1.79315
PX	0.30170	-2.32862	0.16983	-3.10523	0.19181	-3.01224	0.10390	-3.80203	0.13986	-3.39753	0.51149	0.86189	0.15584	1.43897
PN	0.93706	-0.11149	0.85914	-0.50535	0.63337	-1.21839	0.07992	-4.71376	0.04595	-4.98379	0.25774	1.76621	0.04595	2.23335
EV24	2.00E-03	2.44873	2.00E-03	2.25875	2.00E-03	2.61720	2.00E-03	2.24614	2.00E-03	2.16436	2.00E-03	2.21734	2.00E-03	2.00110

Table 5.9: The P-values and difference in mean. where $\overline{X}_a$ is the mean of the days with respiratory disease and $\overline{X}_b$ is the mean of the days without respiratory disease. for the different weather values on days with and without respiratory disease. Gray cells have a p-value that is considered correlated. The corresponding differences in mean are colored red and green. depending on a negative or positive correlation.

	d_t		d_{t-1}		d_{t-2}		d_{t-3}		d_{t-4}		$avg14$		$avg30$	
	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$	P-value	$\overline{X}_a - \overline{X}_b$
DDVEC	0.61339	-1.00635	0.62138	-0.79231	0.46354	1.39700	0.81119	0.44251	0.93107	-0.01224	0.27572	-0.84293	0.06793	-1.07064
FHVEC	2.00E-03	-1.89937	2.00E-03	-1.93026	2.00E-03	-2.13818	2.00E-03	-1.91030	2.00E-03	-1.78916	2.00E-03	-1.98978	2.00E-03	-1.91542
FG	2.00E-03	-1.76458	2.00E-03	-1.69187	2.00E-03	-1.94089	2.00E-03	-1.72293	2.00E-03	-1.59646	2.00E-03	-1.83670	2.00E-03	-1.75899
FHX	2.00E-03	-1.96519	2.00E-03	-1.76690	2.00E-03	-2.20544	2.00E-03	-1.57482	2.00E-03	-1.37767	2.00E-03	-1.83831	2.00E-03	-1.73899
FHN	2.00E-03	-1.64685	2.00E-03	-1.61689	2.00E-03	-1.62977	2.00E-03	-1.75014	2.00E-03	-1.79628	2.00E-03	-1.72204	2.00E-03	-1.63817
FXX	4.00E-03	-2.19696	0.01399	-1.75273	4.00E-03	-2.19173	0.07592	-1.29235	0.07193	-1.35063	2.00E-03	-1.91539	2.00E-03	-1.84315
TG	2.00E-03	13.56052	2.00E-03	13.32466	2.00E-03	12.66883	2.00E-03	12.35755	2.00E-03	13.06435	2.00E-03	12.15348	2.00E-03	10.72479
TN	2.00E-03	9.08091	2.00E-03	8.79280	2.00E-03	7.31058	2.00E-03	7.13564	2.00E-03	8.07982	2.00E-03	7.34097	2.00E-03	6.06338
TX	2.00E-03	17.64260	2.00E-03	17.62742	2.00E-03	17.53799	2.00E-03	17.14324	2.00E-03	17.62220	2.00E-03	16.54893	2.00E-03	14.97312
T10N	2.00E-03	4.82495	2.00E-03	4.42302	0.03596	2.64073	0.03596	2.62545	4.00E-03	3.92453	9.99E-03	2.89283	0.08991	1.63449
SQ	2.00E-03	6.56390	2.00E-03	6.58729	2.00E-03	8.20271	2.00E-03	7.59952	2.00E-03	7.35143	2.00E-03	6.83054	2.00E-03	6.73269
SP	2.00E-03	2.72811	2.00E-03	3.00988	2.00E-03	4.24912	2.00E-03	3.99669	2.00E-03	3.45767	2.00E-03	3.13026	2.00E-03	3.15841
Q	2.00E-03	180.02619	2.00E-03	182.36102	2.00E-03	203.09665	2.00E-03	192.09764	2.00E-03	188.11195	2.00E-03	182.76076	2.00E-03	176.84148
DR	2.00E-03	-1.66241	0.05994	-1.15529	2.00E-03	-2.89286	2.00E-03	-2.06187	0.07792	-1.06043	2.00E-03	-1.76062	2.00E-03	-1.68144
RH	0.02198	-1.81470	0.92907	-0.12103	0.05594	-1.61615	0.11588	-1.20488	0.91109	-0.08371	2.00E-03	-0.92011	2.00E-03	-0.91047
RHX	0.21978	-0.46364	0.32168	0.38294	0.81518	0.08639	0.83317	0.08259	0.20380	0.45812	0.33367	0.12296	0.72527	0.03541
UG	2.00E-03	-2.09654	2.00E-03	-2.23934	2.00E-03	-2.66566	2.00E-03	-2.66248	2.00E-03	-2.45419	2.00E-03	-2.37983	2.00E-03	-2.38332
UX	2.00E-03	0.34940	0.01598	0.25814	0.11389	0.15047	0.41159	0.08147	0.02398	0.23168	2.00E-03	0.21490	2.00E-03	0.19513
UN	2.00E-03	-4.22428	2.00E-03	-4.48520	2.00E-03	-5.04837	2.00E-03	-5.06212	2.00E-03	-4.90932	2.00E-03	-4.65803	2.00E-03	-4.63906
PG	0.25574	-2.17790	0.24176	-2.15853	0.59341	-0.96515	0.28971	-2.01152	0.09191	-3.35950	0.94505	-0.09669	0.25175	0.99891
PX	0.02398	-3.80384	0.05395	-3.46286	0.23976	-2.21697	0.07792	-3.40627	0.03397	-4.03235	0.12388	-1.63219	0.60340	-0.45077
PN	0.95504	-0.07640	0.83716	-0.42911	0.79121	0.54475	0.73926	-0.67797	0.21778	-2.51291	0.17782	1.67632	7.99E-03	2.62148
EV24	2.00E-03	3.17689	2.00E-03	3.17941	2.00E-03	3.47281	2.00E-03	3.30292	2.00E-03	3.28678	2.00E-03	3.16063	2.00E-03	3.00490

Table 5.10: The P-values and difference in mean. where $\overline{X}_a$ is the mean of the days with skin disease and $\overline{X}_b$ is the mean of the days without skin disease. for the different weather values on days with and without skin disease. Gray cells have a p-value that is considered correlated. The corresponding differences in mean are colored red and green. depending on a negative or positive correlation.

5.4 Predictions on horse health

The second research question of this research is to what extent one can predict the diseases of horses using the weather data. The weather variables that are found to be correlated to the disease are used as the input vector. Ensemble predictions are used for the predictions of the diseases, bagging, boosting and voting algorithms are compared. The bagging algorithm is performed four times, each time with one of the following estimators: Linear Regression, Support Vector Machine, Decision Trees and Neural network. The boosting algorithm is tested Decision Trees only. The voting algorithm uses these four estimators, Linear Regression, Support Vector Machine, Decision Tree and Neural Network, to create one combined model. The accuracy, precision and recall for each of the predictions are shown in the Appendix, Table F.1 for the ensemble predictions and Table F.2 for the single classifiers. The confusion matrices of the different models can be found in the Appendix, Table F.3 and Table F.4. There is not much difference between the results of the different classification methods. Respiratory disease can be predicted best with the highest accuracy of 0.798, using a voting algorithm, the bagging algorithm using SVM or the SVM single classifier. Followed by skin disease with an accuracy of 0.742 with a bagging algorithm using decision trees, the same combination of algorithms predicts colic best with an accuracy of 0.701. Laminates is the most challenging to predict with an accuracy of 0.652, obtained with both a single SVM and a single Neural Network.

6 DISCUSSION

6.1 Weather data

6.1.1 Choice of the weather station and data set

The weather data set of the KNMI of the weather station of Heino was used for this research. The weather data of weather station Heino is in-homogeneous, which means that the station can be relocated or the observation techniques can be changed, and therefore it is not suitable for trend analysis. The KNMI has 4 homogenized weather data sets, Eelde, De Kooy, Vlissingen en De Bilt. These 4 homogenized weather data sets are all more than 50 km away from Animal Clinic Den Ham, the animal clinic that provides the medical data for this project. This research works to determine whether correlations can be found between (changes in) the weather and the medical data of horses to see if the weather influences the health of horses. More accurate correlation may be found in an in-homogeneous data set of a weather station as close to the animals as possible, rather than a data set of a weather station that is more than 50 km from most of the clients. It was therefore decided to use the data set of Heino. But this decision can cause noise in the data and therefore influence the results of this research.

The weather stations of Twente and Hoogeveen are located close to the majority of the clients as well, but looking closely at figure 4.9 shows that moving further from Den Ham, the density of clients decreases. Therefore, the weather station closest to Den Ham will be closest to most of the clients. The weather stations of Twente and Hoogeveen are older than the weather station of Heino. This means that it would be possible to look back further with those two data sets. However, with weather data from November 1998, and veterinarian data from August 1999, there is almost one year of weather data available before the first medical data. Furthermore, the data set of Hoogeveen misses values from wind speed from 2005, and both miss values for temperature in 2011, second and third most important values according to Dr. G. Kampman. Both data sets contain all values from barometric pressure and miss values for Visibility and cloud coverage. The values for the barometric pressure will be imputed into the weather data set of weather station Heino. This will grant the most accurate data for the majority of the cases.

In this research, daily weather station values will be used. Because the majority of the health issues remain consistent are not acute, they will emerge over time. This means that the weather condition has to hold on for a while before the weather condition will cause problems. therefore, the daily weather will give us plenty of information.

6.1.2 Imputation values DR

When imputing the missing values of DR, a R^2 value of 0.991 was obtained when filling in the median value. When using kNN as an imputation method, a R^2 value of 0.996 was found when 10 neighbours were used. These results are very good so no other methods have to be tested.

6.1.3 Imputation variables barometric pressure

For the imputation of the barometric pressure, the R^2 values were calculated for the different variables of the barometric pressure of the weather stations around Heino. The maximum and minimum barometric pressure measured for each day is very similar for each of the data sets, the R^2 values lay between 0.996 and 0.985. The hour on which maximum and minimum barometric pressure is measured throughout the day is a little more diverse, with R^2 values between 0.89 and 0.81. Since the exact hour, is not important for this research, we can use the values from one of the other data sets. Hoogeveen and Twente have an equal distance to Animal Clinic Den Ham but Hoogeveen is closer to Heino. Therefore, the data of Hoogeveen shall be used.

6.2 Horse data

In this section, the choices, findings, and initial results of the horse data will be discussed.

6.2.1 Reliability of horses data

Figure 4.7 shows many horses are registered to be born on the first of January. According to Dr. G. Kampman, horses for whom the age is estimated are given the birth date of the first of January of the presumed birth year. When the name of a horse is unknown at the moment of the visit, the code ”AAA paard”, or a random name, like ”paard”, ”pony” or some description of the horse is given as a name. These horses will not provide reliable data when looking at the life events of one horse. The medical data of these horses can be used when looking at the occurrence of illnesses since this will not affect the ability of the veterinarian to make a prognosis.

6.2.2 Number of consults

In this paper, consults are defined as the combination of AnimalID and the date of the consult. A horse can have separate consults on one day, for example when the situation of a horse deteriorates. These separate consults will be combined into one consult. This decision is made since it is not possible to be sure that all different consults on one day for the same horse are separated because the exact time of the consult is not given and the horse can be seen by the same veterinarian multiple times.

6.2.3 Occurrence of diseases

To count the occurrence of diseases, the number of consults mentioning specific keywords are counted. The context in which these words are used is not considered in this technique of counting. This can introduce false positives when a consult states something like ”this is not a case of ...”. On the contrary, false negatives can occur when conjunctions or synonymous words are used or in the case spelling-/type mistakes are made by the veterinarian. Also, many consults do not have explanatory texts, and therefore no keywords. The medication given during these consults can still give an indication of the reason for the consult. To obtain a more precise count, pre-processing of the texts is needed.

Using test processing gives a higher number of consults concerning the different diseases. The precision at 100% recall lays between 0.47676 (for colic) and 0.12225 (for respiratory) for the binary sets. For the label sets reduced and simple this is between 0.46255 (for simple, colic) and 0.04239 (for reduced, laminitis). The precision at 100% recall is, for each of the diseases higher for the binary label sets compared to the reduced and simple sets, therefore the binary label set will be used during this research. The results of the binary label sets are poor, further research could be conducted to find other machine learning approaches to create better results for labelling of the data.

6.2.4 Hand labeled data

Over 10000 consults are labelled by hand to train classifiers for labelling the remaining consults. For this labelling, random consults with a clear reason for the consult were used. This way of labelling creates a bias in the data which will affect the results of this research.

6.2.5 Choices of the diseases

Colic is a collective term for severe abdominal pain and respiratory disease and skin disease also are collections of illnesses and diseases. These different forms of the diseases may have many different causes. This results in noisy data. Laminates is very straight forward, the disease is very well known and the causes are very clear. It would have been better to work with more specified diseases like laminitis but, even though the data of 21 years is used, it is hard to find a disease that occurs often and is very specific.

6.3 Correlation

6.3.1 Correlation vs. causation

In this research, the influence of the weather on the occurrence of diseases in horses is investigated by performing correlation analysis on veterinarian- and weather data. Finding a correlation does not imply a causal relationship in this case since not all factors are included in this research. The correlation could have been caused by indirect factors, such as management changes due to the weather. Proving a causal relationship not feasible with the current data and therefore out of the scope of this research.

6.3.2 The found correlations

Almost all weather categories have one or more variable correlated to each of the diseases. This may be caused by incorrect labelling of the data. Since so many weather categories are correlated to the diseases it is hard to draw conclusions with regards to correlations between diseases and weather.

6.3.3 Positive and negative correlations

When a weather value is correlated to two or more diseases it has, almost always, a positive or negative correlation to all diseases it is correlated to. For example, the category wind has for all correlating variables a negative correlation to the diseases. The categories temperature and sunshine duration are, for all diseases and all correlating variables always correlated positively. It is not clear what causes this phenomenon.
But when we take UX on d_{t-3}, for example, it has a positive difference in mean for laminitis and skin but negative for respiratory and colic. It is only correlated with respiratory. Therefore the variables can have positive and negative differences in mean for the different diseases.

6.3.4 Change in management

Change in management of horses can cause problems like laminitis due to overload from working or to much grass to quickly, or colic due to less movement or switching from grass to hey or visa versa. These changes can be made at random or due to the weather. For example a long trail ride on a beautiful day or keeping the horses inside due to the heavy rain or storms. Also, many horses are kept on grass during the summer and kept in a paddock during the winter. This makes spring, fall and the weather in general also an indirect cause of diseases in horses.

6.4 Predictions

6.4.1 Achieved accuracy

The maximum achieved accuracy for the predictions is almost 80%, achieved for respiratory disease. Laminitis is hardest to predict with a maximum achieved an accuracy of almost 65%. These results are not very good. This is probably because it already proved difficult to label the consults and no obvious correlations between the diseases and weather variables where found.

6.4.2 Building on previous results

All results of this research build on each other and are at the end dependent on the labelling of the consults. The consults are labelled using a classifier and hand labelled consults. Hand labelling the consults has created a bias in the data and the results of the classification were not great. This means that a part of the consults is incorrectly labelled. This creates noise in the correlation analysis and the predictions and therefore influences the of the results of this research.

6.5 Recommendations for the veterinarian

For research like this, it would be helpful if the data contained standardized consult reports. For example specific keywords like ”T” or ”Temp” when the temperature is stated and ”Diagnosis” or ”Symptoms” followed by the diagnosis or symptoms observed for that consult. This would also allow the veterinarians to have a more clear overview of the medical situation of the horse. This especially holds of the diagnosis. The diagnosis is missing in the texts quite often which makes it more difficult for a veterinarian to have a quick insight into the history of the horse. This could be encouraged by Viva, the software company, by making multiple text fields with a clear header or question that can be answered by the veterinarian. Since this will cost the veterinarians more time it is unlikely that this will be used in full extend, especially since it will not make the work of the veterinarian better for or easier for the current consult but, as mentioned above, the veterinarian could benefit from this during a follow-up consult.

6.6 Further research

In the visualizations, plotting weather variables on different times for disease and no disease days, no obvious correlations between changes in the weather and the diseases were found. Further research could be conducted to see if this observation is correct by using a better-labelled data set, especially since the expert is surprised that sudden worsening of the weather is not found to be correlated to colic.

Since the labelling of the consults has proven to be difficult for this data and the hand labelling of the training data has created a bias, it could be interesting to repeat this research with a better labelled or more structured data of the diseases of horses since this is the foundation of this research.

7 CONCLUSIONS

7.1 Data preparation

The number of consults labelled with a disease using keyword search is much lower than the number of consults labelled with a disease using machine learning. This is probably due to the number of consults without text, and therefore without keywords. The medication given during these consults can still give an indication of the disease treated. The keyword search has a higher accuracy than the predicted labels but the accuracy of the keyword search is calculated over the hand labelled data, which contains all consults that contain the keywords. Therefore the accuracy is probably lower since a lot of consults without texts actually are about the diseases. When considering the ”Reduced”, ”Simple” and the binary label sets, the binary label sets allow multi labelling classification i.e. allow consults to be labelled with more than one disease. This is preferred since this is the case for some of the consults. Also, the precision at 100% recall for the binary label sets is higher than for the other two label sets. The data will be labelled using binary labelling, this data will be used to carry out the rest of the research.

7.2 Answering the research questions

The answers to the research questions, as stated in Chapter 3 are given in the following subsections.

7.2.1 Q1: What is the influence of the Dutch weather on the health of horses?

This research question is divided into four sub-questions which will be answered below in more detail. Even though there are a lot of correlations between the weather variables and the diseases in general, the diseases do not seem to correlate to specific weather conditions. Therefore actual the influence of the Dutch weather on the health of horses is not found.

1.1 Does the temperature, barometric pressure and high amount of wind influence the occurrence of colic? For answering this question, the results shown in table 5.7 are used. A higher maximum temperature (TX) 4 days before colic (d_{t-4}) and a lower 30-day average minimum temperature at 10cm above the ground (T10N) is correlated to colic. A lower daily mean (PG) and minimum barometric pressure (PN) are related to colic on d_{t-2}, d_{t-3}, d_{t-4} where d_t is the day of colic and the 14 day average of PG and PN is correlated to colic, as well as a lower 14 day average of the maximum barometric pressure (PX). The 14 and 30 day average of the vector mean wind speed (FHVEC), the daily mean wind speed (FG) and the maximum hourly mean wind speed (FHX) are correlated to colic, as well as the 30 day average of the minimum hourly mean wind speed (FHN). The wind speed and the barometric pressure are negatively correlated to colic. Higher maximum temperature and/or a continued lower ground temperature are related to colic. Against the expectations, less wind and lower barometric pressure are correlated to colic.

1.2 Is the development of laminitis dependent on stress in the grass, due to cold and drought? The results shown in tables 5.8 are used to answer this question.
Low temperatures and low amounts of rain would be expected to be correlated to the occurrence of laminitis. When looking at the results of the permutation tests we see that higher temperatures for all variables except for some of the temperatures at 10cm above the ground (T10N) are positively correlated to laminitis. A shorter duration of precipitation 3 days before the occurrence of laminitis and shorter 14 and 30-day average precipitation duration (DR) are correlated to laminitis, as is a lower 14-day average daily precipitation amount (RH).
Drought is indeed correlated to laminitis but against the expectations, higher temperatures are correlated to laminitis as well.

1.3 Does hot, humid or cold weather worsen or induce respiratory disease? This question is answered using the results of the permutation tests in table 5.9. All variables of the daily mean temperature (TG) and the maximum temperature (TX) do correlate with respiratory disease. As well as almost all minimum temperatures (TN). For the temperatures, the difference in means between the groups are positive, so higher temperatures are correlated with the respiratory disease.
All variables for the daily mean relative atmospheric humidity (UG) and the minimum relative humidity (UN), and the maximum relative humidity on d_{t-3} are correlated to the occurrence of respiratory disease. All with negative differences in means. Therefore a lower humidity correlates to the occurrence of respiratory disease.
High temperatures and low humidity do correlate to the occurrence of respiratory disease.

1.4 Do skin diseases occur more in periods of heavy rainfall and high humidity? The results shown in tables 5.10 are used to answer this question. Higher precipitation duration (PG) on the day of disease (d_t), on d_{t-3}, d_{t-4} and on the 14 and 30 day average are correlated to skin disease. As is the 14 and 30 day average of the daily precipitation amount. The difference in means of these variables is negative, therefore less rain correlated to skin disease.
All variables for the daily mean relative atmospheric humidity (UG) and minimum relative atmospheric humidity (UN) are correlated to the negatively are correlated to skin disease. The maximum relative atmospheric humidity (UX) is positively correlated to skin disease on d_t and the 14 and 30 day average of the disease.
Less rain and a lower average and minimum and higher maximum humidity are correlated to skin disease.

7.2.2 Q2: To what extent can the Dutch weather be used to predict the occurrence of ...

The data used for the predictions are the variables that are correlated to the disease, according to the permutation tests. The prediction algorithms used are Bagging, Boosting and Voting. Bagging is performed with the estimator's Logistic regression, Support Vector Machines, Decision Trees and Neural Networks. The Voting algorithm combines these four estimators. The accuracies of the models can be found in Table F.1. The accuracies of the single Linear Regression, Support Vector Machine, Decision Tree and Neural Network classifiers can be found in Table F.2.

a. colic? Colic can be predicted with an accuracy of 70% when using the Bagging algorithm with Decision trees.

b. laminitis? Laminitis can be predicted with an accuracy of 65% using a single SVM classifier or a single Neural Network.

c. respiratory disease? With an accuracy of almost 80%, respiratory disease can be predicted best of the four diseases. This is achieved with the Voting algorithm, combining the estimators, but also with the bagging algorithm using SVM and a single SVM classifier.

d. skin disease? Skin disease can be predicted with an accuracy of 74% using the Bagging algorithms again with Decision trees as the estimator.

ACKNOWLEDGEMENTS

I would like to express my appreciation to dr. Kampman, a veterinarian at Animal Clinic Den Ham. After my supervisor reached out to him, he invited us into the clinic to talk about possible Data Science projects using horse data. He provided us with the data that is collected in the clinic over the past 21 years. Dr. Kampman and I also discussed possible correlations between the weather and diseases in horses. Dr. Kampman provided some assertions that have anecdotal evidence, that wind could cause colic and high humidity can cause respiratory and skin problems. These assertions helped a lot with finding the related work and by formulating the Research Questions. When my supervisors or I had questions about the data, he also offered help.

REFERENCES

[1] Dr Mohammad El-Nesr. Filling gaps of a time-series using python., December 2019. medium.com/@drnesr/filling-gaps-of-a-time-series-using-python-d4bfddd8c460. Accessed: 5/22/2020.

[2] Erica Larson. Equine Postoperative Ileus Insights, July 2013. thehorse.com/116386/equine-postoperative-ileus-insights/. Accessed: 4/29/2020.

[3] Mary K. Tinker, N. A. White, P. Lessard, C. D. Thatcher, K. D. Pelzer, Betty Davis, and D. K. Carmel. Prospective study of equine colic incidence and mortality. *Equine Veterinary Journal*, 29(6):448–453, 1997.

[4] II White, A Nathaniel, and VV Tech. Prevalence, Demographics, and Risk Factors for Colic, 2005.

[5] D. C. Archer and C. J. Proudman. Epidemiological clues to preventing colic. *The Veterinary Journal*, 172(1):29–39, July 2006.

[6] Nathaniel A. White. Colic prevalence, risk factors and prevention. 2009. Advances in Equine Nutrition, Volume 4.

[7] Mary Kay Tinker. A farm-based prospective study for equine colic risk factors and risk-associated events. August 1995.

[8] Noah D. Cohen. Epidemiology of Colic. *Veterinary Clinics of North America: Equine Practice*, 13(2):191–201, August 1997.

[9] A. G. Limont. IV Observations on Colic and Abdominal Surgery in the Horse. *Equine Veterinary Journal*, 2(2):59–60, 1970.

[10] C. J. Proudman. A two year, prospective survey of equine colic in general practice. *Equine Veterinary Journal*, 24(2):90–93, 1992.

[11] Noah Cohen, Pete Gibbs, and April Woods. Dietary and Other Management Factors Associated with Equine Colic. page 4, 1999.

[12] M. H. Hillyer, F. G. R. Taylor, and N. P. French. A cross-sectional study of colic in horses on Thoroughbred training premises in the British Isles in 1997. *Equine Veterinary Journal*, 33(4):380–385, 2001.

[13] Josie L. Traub-Dargatz, Christine A. Kopral, Ann Hillberg Seitzinger, Lindsey P. Garber, Kim Forde, and Nathaniel A. White. Estimate of the national incidence of and operation-level risk factors for colic among horses in the United States, spring 1998 to spring 1999. *Journal of the American Veterinary Medical Association*, 219(1):67–71, July 2001. Publisher: American Veterinary Medical Association.

[14] D. C. Archer, C. J. Proudman, G. Pinchbeck, J. E. Smith, N. P. French, and G. B. Edwards. Entrapment of the small intestine in the epiploic foramen in horses: a retrospective analysis of 71 cases recorded between 1991 and 2001. *Veterinary Record*, 155(25):793–797, December 2004. Publisher: British Medical Journal Publishing Group Section: Papers & Articles.

[15] Debra C. Archer, Gina L. Pinchbeck, Christopher J. Proudman, and Helen E. Clough. Is equine colic seasonal? Novel application of a model based approach. *BMC Veterinary Research*, 2(1):27, August 2006.

[16] D. C. Archer, G. L. Pinchbeck, N. P. French, and C. J. Proudman. Risk factors for epiploic foramen entrapment colic in a UK horse population: A prospective case-control study. *Equine Veterinary Journal*, 40(4):405–410, 2008.

[17] G. Kaya, I. Sommerfeld-Stur, and C. Iben. Risk factors of colic in horses in Austria. *Journal of Animal Physiology and Animal Nutrition*, 93(3):339–349, 2009.

[18] J.E. Dechant, Z. Davidson, and P.H. Kass. An investigation into the association between changes in air temperature and barometric pressure on the incidence of colic in horses. 2014.

[19] N Diakakis and P Tyrnenopoulou. Correlation between equine colic and weather changes. *Journal of the Hellenic Veterinary Medical Society*, 68(3):455–466, 2017. Number: 3.

[20] Simon Bizhga, Ilir Dove, Sulo Kotkrri, and Rezart Postoli. Risk factors of colic episodes in the horses in Albania. April 2017.

[21] Justine Cianci. Determining the Relationship Between Barometric Pressure Changes and the Incidence of Equine Colic. May 2018.

[22] John Polzer and Margaret R. Slater. Age, breed, sex and seasonality as risk factors for equine laminitis. *Preventive Veterinary Medicine*, 29(3):179–184, January 1997.

[23] Claire E. Wylie, Simon N. Collins, Kristien L. P. Verheyen, and J. Richard Newton. Risk factors for equine laminitis: A case-control study conducted in veterinary-registered horses and ponies in Great Britain between 2009 and 2011. *The Veterinary Journal*, 198(1):57–69, October 2013.

[24] N. J. Menzies-Gow, L. M. Katz, K. J. Barker, J. Elliott, M. N. De Brauwere, N. Jarvis, C. M. Marr, and D. U. Pfeiffer. Epidemiological study of pasture-associated laminitis and concurrent risk factors in the South of England. *Veterinary Record*, 167(18):690–694, October 2010.

[25] K. H. Treiber, R. C. Boston, D. S. Kronfeld, W. B. Staniar, and P. A. Harris. Insulin resistance and compensation in Thoroughbred weanlings adapted to high-glycemic meals. *Journal of Animal Science*, 83(10):2357–2364, October 2005.

[26] R. M. Hoffman, R. C. Boston, D. Stefanovski, D .S. Kronfeld, and P. A. Harris. Obesity and diet affect glucose dynamics and insulin sensitivity in Thoroughbred geldings. 2003.

[27] Kibby H. Treiber, David S. Kronfeld, Tanja M. Hess, Bridgett M. Byrd, Rebecca K. Splan, and W. Burton Staniar. Evaluation of genetic and metabolic predispositions and nutritional risk factors for pasture-associated laminitis in ponies. *Journal of the American Veterinary Medical Association*, 228(10):1538–1545, May 2006.

[28] Simon R. Bailey, Nicola J. Menzies-Gow, Patricia A. Harris, Jocelyn L. Habershon-Butcher, Carol Crawford, Yoel Berhane, Raymond C. Boston, and Jonathan Elliott. Effect of dietary fructans and dexamethasone administration on the insulin response of ponies predisposed to laminitis. *Journal of the American Veterinary Medical Association*, 231(9):1365–1373, November 2007.

[29] A. W. Eps and C. C. Pollitt. Equine laminitis induced with oligofructose. *Equine Veterinary Journal*, 38(3):203–208, 2006.

[30] Katie E. Asplin, Martin N. Sillence, Christopher C. Pollitt, and Catherine M. McGowan. Induction of laminitis by prolonged hyperinsulinaemia in clinically normal ponies. *The Veterinary Journal*, 174(3):530–535, November 2007.

[31] M. Stitt, R. Gerhardt, I. Wilke, and H. W. Heldt. The contribution of fructose 2,6-bisphosphate to the regulation of sucrose synthesis during photosynthesis. *Physiologia Plantarum*, 69(2):377–386, February 1987.

[32] JohnE. Lunn and MarshallD. Hatch. Primary partitioning and storage of photosynthate in sucrose and starch in leaves of C4 plants. *Planta*, 197(2), September 1995.

[33] D. M. Bowden, D. K. Taylor, and W. E. P. Davis. Water-soluable carbohydrates in orchardgrass and mised forages. *Canadian Journal of Plant Science*, 48(1):9–15, January 1968.

[34] V. L. Lechtenberg, D. A. Holt, and H. W. Youngberg. Diurnal Variation in Nonstructural Carbohydrates, In Vitro Digestibility, and Leaf to Stem Ratio of Alfalfa1. *Agronomy Journal*, 63(5):719–724, 1971.

[35] T. A. Ciavarella, R. J. Simpson, H. Dove, B. J. Leury, and I. M. Sims. Diurnal changes in the concentration of water-soluble carbohydrates in Phalaris aquatica L. pasture in spring, and the effect of short-term shading. *Australian Journal of Agricultural Research*, 51(6):749, 2000.

[36] D. M. Burner and D. P. Belesky. Diurnal Effects on Nutritive Value of Alley□Cropped Orchardgrass Herbage. *Crop Science*, 44(5):1776–1780, September 2004.

[37] Anne M. Borland and J. F. Farrar. The influence of low temperature on diel patterns of carbohydrate metabolism in leaves of Poa Annual I. and Poa X Jemtilandica (Almq.) Richt. *New Phytologist*, 105(2):255–263, February 1987.

[38] N.J. Chatterton, P.A. Harrison, J.H. Bennett, and K.H. Asay. Carbohydrate Partitioning in 185 Accessions of Gramineae Grown Under Warm and Cool Temperatures. *Journal of Plant Physiology*, 134(2):169–179, March 1989.

[39] A J Parsons, S Rasmussen, H Xue, J A Newman, C B Anderson, and G P Cosgrove. Some ‘high sugar grasses’ don’t like it hot. page 9, 2004.

[40] C. J. Pollock and T. Jones. Seasonal patterns of fructane metabolism in forage grasses. *New Phytologist*, 83(1):9–15, July 1979.

[41] William C Spollen and Curtis J Nelson. Response of Fructan to Water Deficit in Growing Leaves of Tall Fescue’. page 8, 1994.

[42] R Munns, Cj Brady, and Ewr Barlow. Solute Accumulation in the Apex and Leaves of Wheat During Water Stress. *Functional Plant Biology*, 6(3):379, 1979.

[43] Jorge Marques da Silva and Maria Celeste Arrabaça. Contributions of soluble carbohydrates to the osmotic adjustment in the C4 grass Setaria sphacelata: A comparison between rapidly and slowly imposed water stress. *Journal of Plant Physiology*, 161(5):551–555, January 2004.

[44] D. W. B. Sainsbury. Ventilation and environment in relation to equine respiratory disease. *Equine Veterinary Journal*, 13(3):167–170, July 1981.

[45] W. R. Cook, R. M. Williams, C. A. Kirker-Head, and D. J. Verbridge. Upper airway obstruction (partial asphyxia) as the possible cause of exercise-induced pulmonary hemorrhage in the horse: An hypothesis. *Journal of Equine Veterinary Science*, 8(1):11–26, January 1988.

[46] Bruce Mcgorum. *Differential diagnosis of chronic coughing In the horse*. 1994.

[47] M Mazan, A hoffman, J F Wade, and M Mazan. Inflammatory airway disease: effect of athletic discipline. page 109, 2002.

[48] Melissa R. Mazan and Andrew M. Hoffman. Clinical techniques for diagnosis of inflammatory airway disease in the horse. *Clinical Techniques in Equine Practice*, 2(3):238–257, September 2003.

[49] M. Bullone, R. Y. Murcia, and J-P. Lavoie. Environmental heat and airborne pollen concentration are associated with increased asthma severity in horses. *Equine Veterinary Journal*, 48(4):479–484, July 2016.

[50] A.I. Donaldson and N.P. Ferris. The survival of some air-borne animal viruses in relation to relative humidity. *Veterinary Microbiology*, 1(4):413–420, December 1976.

[51] N.E. Robinson, F.J. Derksen, M.A. Olszewski, and V.A. Buechner-Maxwell. The pathogenesis of chronic obstructivepulmonary disease of horses. *British Veterinary Journal*, 152(3):283–306, May 1996.

[52] Laurent L. Couetil and Michael P. Ward. Analysis of risk factors for recurrent airway obstruction in North American horses: 1,444 cases (1990-1999). *Journal of the American Veterinary Medical Association*, 223(11):1645–1650, December 2003.

[53] Michael S. Davis, Jerry R. Malayer, Lori Vandeventer, Christopher M. Royer, Erica C. McKenzie, and Katherine K. Williamson. Cold weather exercise and airway cytokine expression. *Journal of Applied Physiology*, 98(6):2132–2136, June 2005. Publisher: American Physiological Society.

[54] M. S. Davis, C. M. Royer, E. C. McKENZIE, K. K. Williamson, M. Payton, and D. Marlin. Cold air-induced late-phase bronchoconstriction in horses. *Equine Veterinary Journal*, 38(S36):535–539, August 2006.

[55] N. E. Robinson, W. Karmaus, S. J. Holcombe, E. A. Carr, and F. J. Derksen. Airway inflammation in Michigan pleasure horses: prevalence and risk factors. *Equine Veterinary Journal*, 38(4):293–299, January 2010.

[56] N. St. G Hyslop. Dermatophilosis (streptothricosis) in animals and man. *Comparative Immunology, Microbiology and Infectious Diseases*, 2(4):389–404, January 1979.

[57] Josephine M. Kingali, I. D. Heron, and A. N. Morrow. Inhibition of Dermatophilus congolensis by substances produced by bacteria found on the skin. *Veterinary Microbiology*, 22(2):237–240, April 1990.

[58] Pauline J Mollison. Equine Welfare: a Study of Dermatophilosis and the Management of Data Relevant to the Healt and Wellbeing of Horses. page 406, 1990.

[59] I. Yeruham, D. Elad, and O. Egozi. Outbreak of Dermatophilosis in a Horse Herd in Israel. *Journal of Veterinary Medicine Series A*, 43(1-10):393–398, February 1996.

[60] Stephen D. White. Equine Bacterial and Fungal Diseases: A Diagnostic and Therapeutic Update. *Clinical Techniques in Equine Practice*, 4(4):302–310, December 2005.

[61] C. M. Colles, K. .M Colles, and J. R. Galpin. Equine pastern dermatitis. 2010.

[62] M. A. Gabal and S. Hennager. Study on the Survival of Histoplasma Farciminosum in the Environment/Experimentelle Untersuchungen zur Lebensfähigkeit von Histoplasma Farciminosum. *Mycoses*, 26(9):481–487, 1983.

[63] Gobena Armeni. Epidemiology of equine histoplasmosis (epizootic lymphangitis) in carthorses in Ethiopia. 2006.

[64] Denise R. Murray, P. W. Ladds, and R. S. F. Campbell. Granulomatous and Neoplastic Diseases of the Skin of Horses. *Australian Veterinary Journal*, 54(7):338–341, July 1978.

[65] R. I. Miller and R. S. F. Campbell. Clinical Observations on Equine Phycomycosis. *Australian Veterinary Journal*, 58(6):221–226, June 1982.

[66] Valerie Faldok. An Overview of Equine Dermatoses Characterized by Scaling and Crusting. 1995.

[67] Stephen White. What's New in Equine Dermatology. 2015.

[68] KNMI Klimatologische Dienst - Informatie over verleden weer. projects.knmi.nl/klimatologie/metadata/. Accessed: 2/20/2020.

[69] Viva Veterinary. www.vivaveterinary.nl/. Accessed: 3/18/2020.

[70] Natuur en Voedselkwaliteit Ministerie van Landbouw. Wijziging Diergeneesmiddelenregeling, July 2007.

[71] KNMI - Gehomogeniseerde tijdreeksen dagwaarden. www.knmi.nl/kennis-en-datacentrum/achtergrond/gehomogeniseerde-tijdreeksen-dagwaarden. Accessed: 2/18/2020.

[72] KNMI - Koninklijk Nederlands Meteorologisch Instituut. knmi.nl/home. Accessed: 1/10/2020.

[73] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Working With Text Data: Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html Accessed: 09/29/2020.

[74] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. CountVectorizer: Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html Accessed: 10/07/2020.

[75] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. TfidfTransformer: Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html Accessed: 10/07/2020.

[76] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. MultinomialNB: Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html Accessed: 10/07/2020.

[77] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. SGDClassifier: Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html Accessed: 10/07/2020.

[78] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. GridSearchCV: Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html Accessed: 10/24/2020.

[79] John W. Graham. Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*, 60(1):549–576, 2009. _eprint: https://doi.org/10.1146/annurev.psych.58.110405.085530.

[80] Liu Yu, Lei Yu, Yi Qi, Jianquan Wang, and Huimin Wen. Traffic Incident Detection Algorithm for Urban Expressways Based on Probe Vehicle Data. *Journal of Transportation Systems Engineering and Information Technology*, 8(4):36–41, August 2008.

[81] Hector Rodriguez, Juan J. Flores, Luis A. Morales, Carlos Lara, Armando Guerra, and Giovanni Manjarrez. Forecasting from incomplete and chaotic wind speed data. *Soft Computing*, 23(20):10119–10127, October 2019.

[82] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, June 2001. Publisher: Oxford Academic.

[83] scikit-learn: machine learning in Python — scikit-learn 0.23.1 documentation. scikit-learn.org/stable/index.html. Accessed: 5/29/2020.

[84] A. M. Raja. KNNImputer for Missing Value Imputation in Python using scikit-learn. datascienceplus.com/knnimputer-for-missing-value-imputation-in-python-using-scikit-learn/. Accessed: 5/25/2020.

[85] C. P. Fielder. Effect of weather conditions on acute laryngotracheitis. *The Journal of Laryngology & Otology*, 103(2):187–190, February 1989.

[86] Marion Roussel, Dominique Pontier, Jean-Marie Cohen, Bruno Lina, and David Fouchet. Quantifying the role of weather on seasonal influenza. *BMC Public Health*, 16(1):441, May 2016.

[87] Christopher R. Genovese, Nicole A. Lazar, and Thomas Nichols. Thresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate. *NeuroImage*, 15(4):870–878, April 2002.

[88] J.W. Taylor and R. Buizza. Neural network load forecasting with weather ensemble predictions. *IEEE Transactions on Power Systems*, 17(3):626–632, August 2002. Conference Name: IEEE Transactions on Power Systems.

[89] James W. Taylor and Roberto Buizza. Using weather ensemble predictions in electricity demand forecasting. *International Journal of Forecasting*, 19(1):57–70, January 2003.

[90] Florian Pappenberger, Jens Bartholmes, Jutta Thielen, Hannah L. Cloke, Roberto Buizza, and Ad de Roo. New dimensions in early flood warning across the globe using grand-ensemble weather predictions. *Geophysical Research Letters*, 35(10), 2008.

[91] Y. He, F. Wetterhall, H. L. Cloke, F. Pappenberger, M. Wilson, J. Freer, and G. McGregor. Tracking the uncertainty in flood alerts driven by grand ensemble weather predictions. *Meteorological Applications*, 16(1):91–101, 2009.

[92] M.C. Brace, V. Bui-Nguyen, and J. Schmidt. Another look at forecast accuracy of neural networks. In *[1993] Proceedings of the Second International Forum on Applications of Neural Networks to Power Systems*, pages 389–394, April 1993.

[93] William L. Nelson and Robert F. Dale. A Methodology for Testing the Accuracy of Yield Predictions from Weather-Yield Regression Models for Corn ^{1}. *Agronomy Journal*, 70(5):734–740, September 1978.

[94] M. G. Thomas and G. K. Goldwin. Associations between weather factors and the spring migration of the damson□hop aphid, Phorodon humuli, 1983. Annals of Applied Biology.

[95] Andrew Harvey and Siem Jan Koopman. Forecasting Hourly Electricity Demand Using Time-Varying Splines. *Journal of the American Statistical Association*, 88(424):1228–1236, 1993. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

[96] Ramu Ramanathan, Robert Engle, Clive W. J. Granger, Farshid Vahid-Araghi, and Casey Brace. Short-run forecasts of electricity loads and peaks. *International Journal of Forecasting*, 13(2):161–174, June 1997.

[97] Jason Brownlee. Ensemble Machine Learning Algorithms in Python with scikit-learn, June 2016. https://machinelearningmastery.com/ensemble-machine-learning-algorithms-python-scikit-learn/ Accessed: 6/24/2020.

[98] Steven Hurwitt. Classification in Python with Scikit-Learn and Pandas. https://stackabuse.com/classification-in-python-with-scikit-learn-and-pandas/ Accessed: 5/24/2020.

A USED DATA

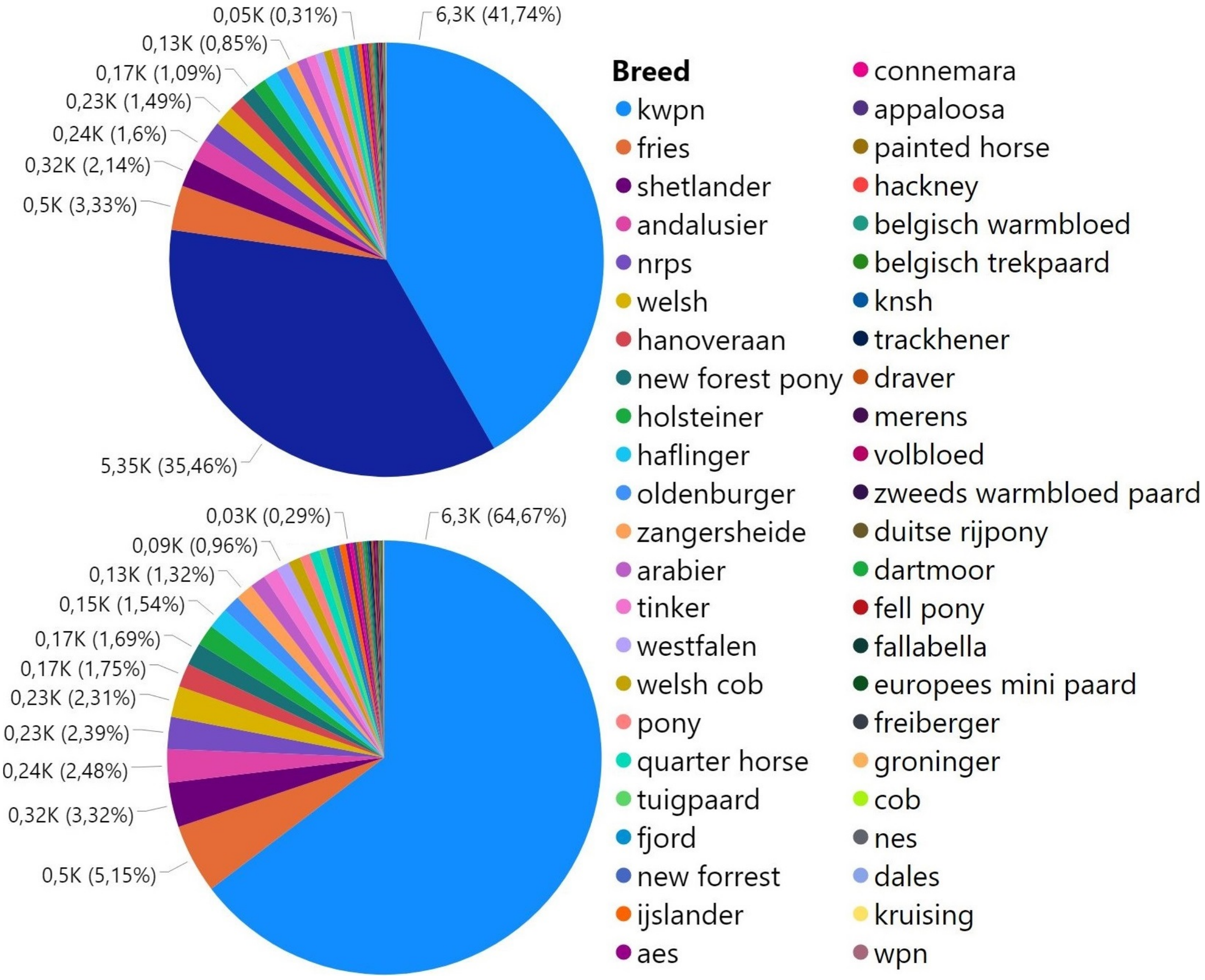


Figure A.1: Distribution of the breeds, including unknown (top) and excluding unknown (bottom)

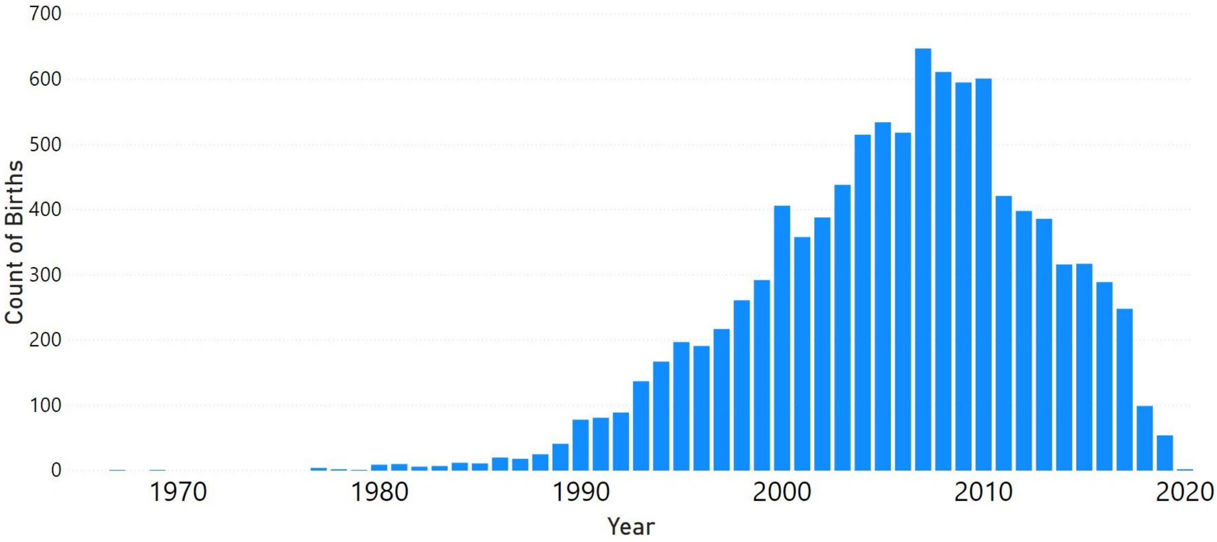


Figure A.2: Distribution of number of births over the years.

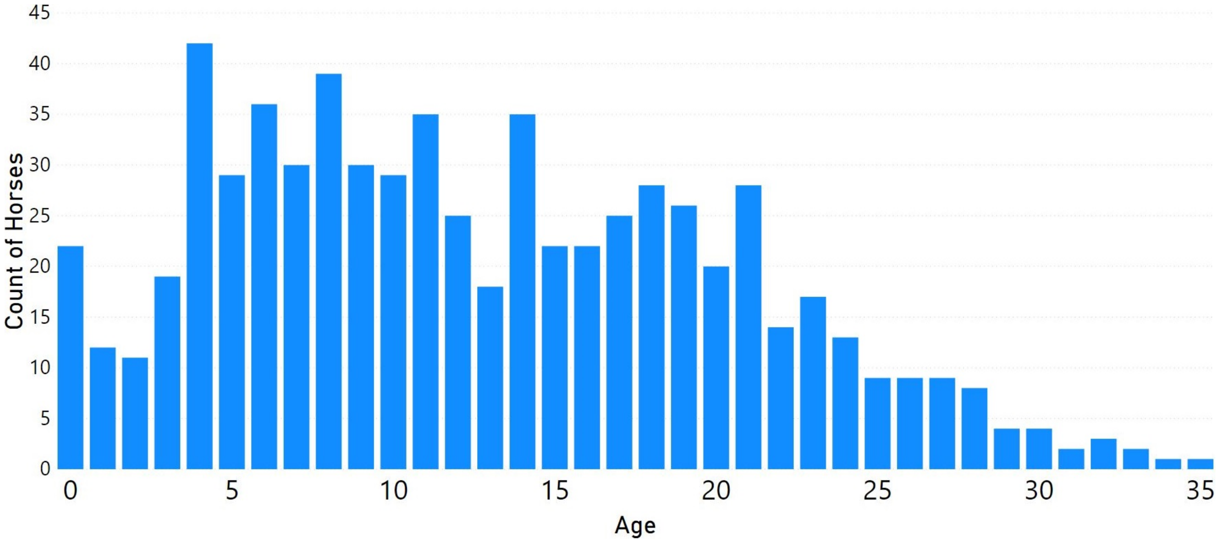


Figure A.3: Histogram of the age of the horses with a birth and death date.

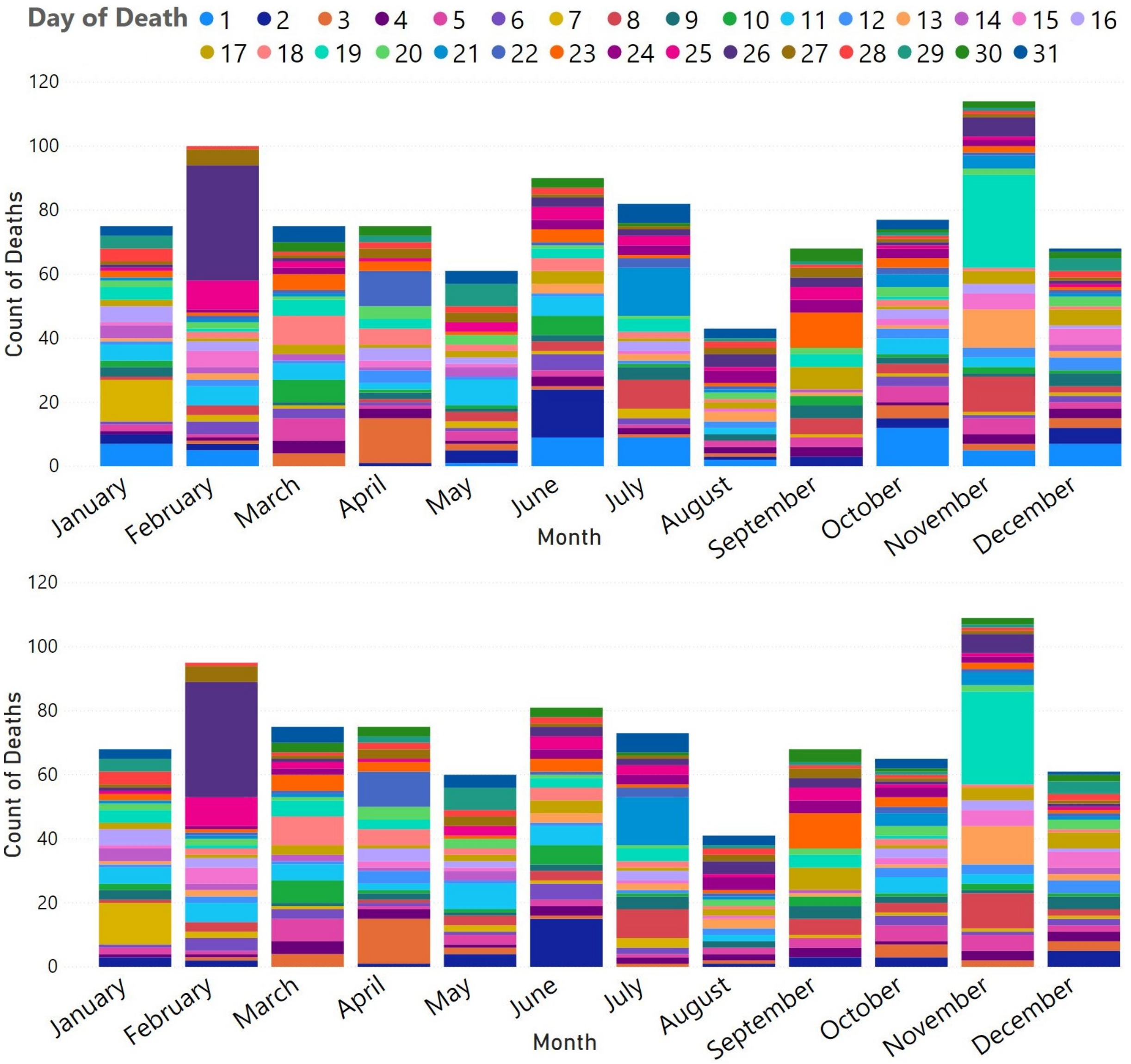


Figure A.4: Distribution of deaths over the months, with (top) and without (bottom) the first day of each month.

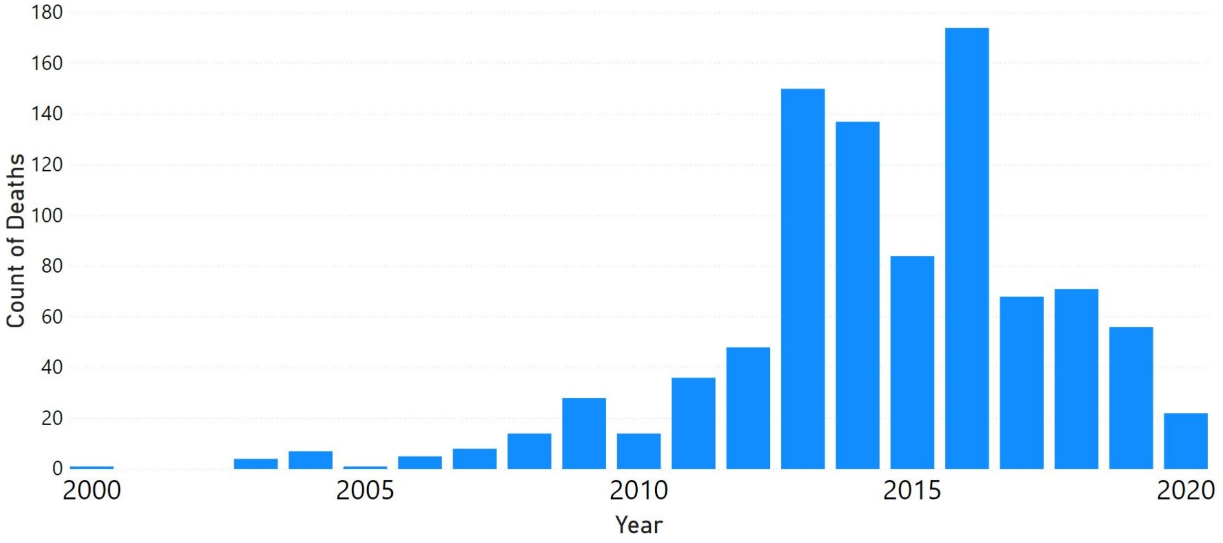


Figure A.5: Distribution of number of deaths over the years.

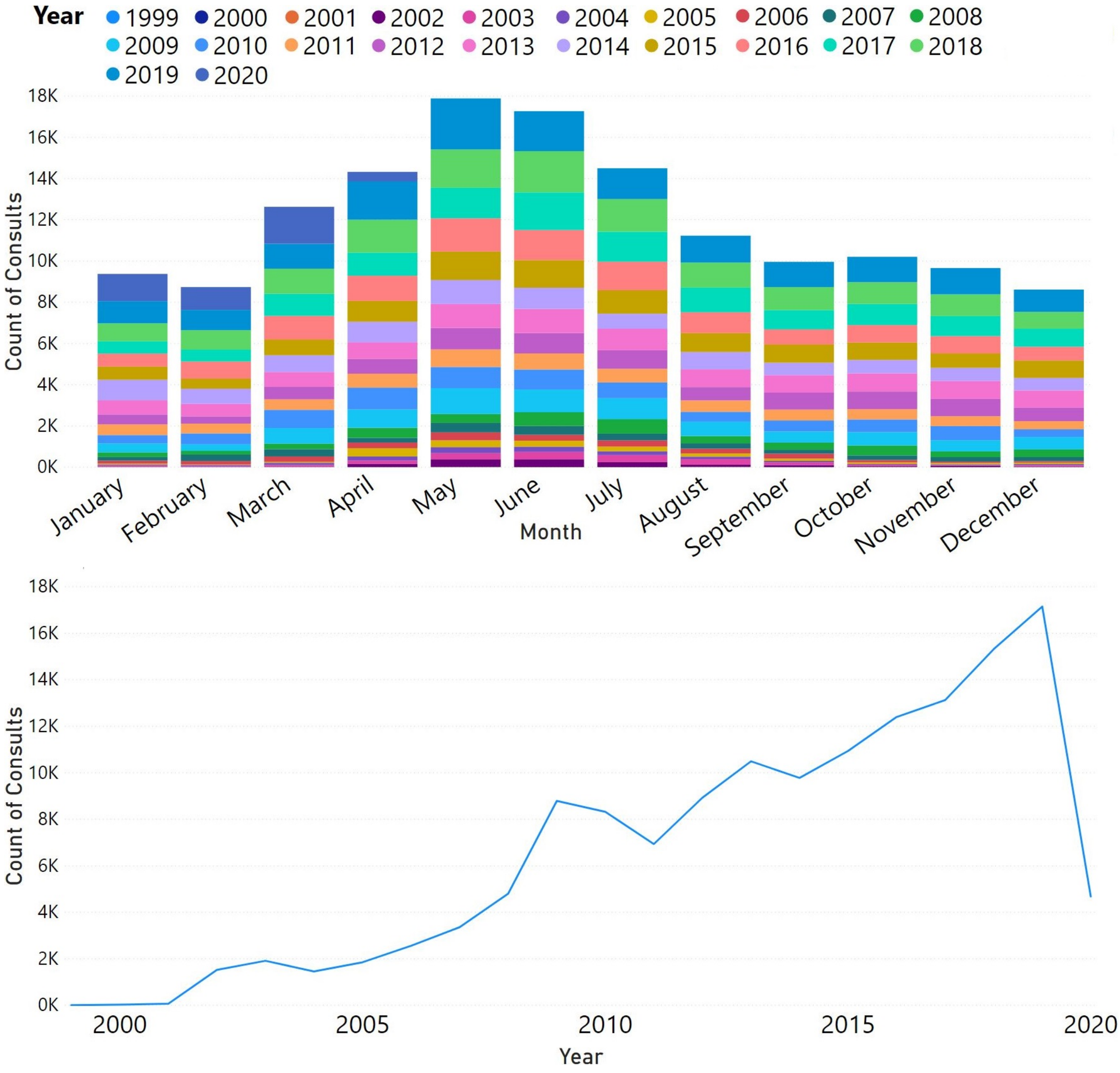


Figure A.6: The distribution of the number of consults over the months (top) and the years (bottom)

B IMPUTATION MISSING WEATHER VALUES AND VARIABLES

Method	R^2
FillMedian	0.99099
FillMean	0.99094
InterpolateLinear	0.99040
InterpolateTime	0.99040
InterpolateSLinear	0.99040
InterpolateSpline3	0.93563
InterpolateQuadratic	0.87843
InterpolateCubic	0.74874
InterpolatePoly7	-253371.34788
InterpolatePoly5	-1594.74889
InterpolateSpline5	-1565.61449
InterpolateSpline4	-1437.43624
InterpolateAkima	-1.72344e+21

Table B.1: Results of imputation methods for our data, using the methods given in [1]

k	R^2
1	0.99016
2	0.99361
3	0.99454
4	0.99480
5	0.99523
6	0.99520
7	0.99535
8	0.99535
9	0.99555
10	0.99556
11	0.99543
12	0.99554
13	0.99551
14	0.99544
15	0.99548
16	0.99545
17	0.99532
18	0.99526
19	0.99522

Table B.2: Results for imputation using kNN with k as the number of nearest neighbours used for the prediction

Twente	PG	0.99386		
	PX	0.99318		
	PXH	0.88714		
	PN	0.99253		
	PNH	0.87038		
Deelen	PG	0.98769	0.99586	
	PX	0.98636	0.99525	
	PXH	0.82229	0.86785	
	PN	0.98534	0.99451	
	PNH	0.81337	0.87131	
Lelystad	PG	0.99408	0.99126	0.99496
	PX	0.99371	0.99078	0.99431
	PXH	0.84432	0.82887	0.89193
	PN	0.99246	0.98892	0.99390
	PNH	0.83665	0.81400	0.87607
		Hoogeveen	Twente	Deelen

Table B.3: The calculated R^2 for the barometric pressure for combinations of weather data sets

C LABELING CONSULTS

Labelsets	Colic	Respiratory	Laminitis	Skin	All	Reduced	Simple
sum	10839	10839	10839	10839	11899	10839	10839
0	10104	10548	10631	10223	0	0	0
1	735	291	208	616	0	0	0
dracht	0	0	0	0	2155	2138	2138
eczeem	0	0	0	0	6	5	0
enting	0	0	0	0	4792	3753	3751
euthanasie	0	0	0	0	24	47	47
geendiagnose	0	0	0	0	56	57	56
genezen	0	0	0	0	19	19	0
hoefbevangen	0	0	0	0	204	205	208
hoefsmid	0	0	0	0	82	53	77
hoefzweer	0	0	0	0	36	29	0
huid	0	0	0	0	606	602	607
keuring	0	0	0	0	789	787	787
kliniek	0	0	0	0	4	4	0
koliek	0	0	0	0	735	724	724
kreupel	0	0	0	0	357	320	319
luchtweg	0	0	0	0	291	286	286
maagzweer	0	0	0	0	31	31	31
oog	0	0	0	0	50	48	48
operatie	0	0	0	0	51	60	60
overig	0	0	0	0	0	22	59
ppid	0	0	0	0	48	47	48
tandarts	0	0	0	0	588	648	647
tumor	0	0	0	0	37	37	37
verstopping	0	0	0	0	5	5	0
wond	0	0	0	0	94	93	93
wormen	0	0	0	0	832	816	816
zand	0	0	0	0	7	3	0

Table C.1: The distribution of the the labels in the different label sets.

	Colic	Respiratory	Laminitis	Skin	All	Reduced	Simple
	10839	10839	10839	10839	10839	10839	10839
0	10104	10548	10631	10223	0	0	0
1	735	291	208	616	0	0	0
dracht	0	0	0	0	1963	2138	2138
enting	0	0	0	0	3849	3753	3751
enting, dracht	0	0	0	0	165	0	0
enting, dracht, tandarts	0	0	0	0	1	0	0
enting, dracht, wormen	0	0	0	0	1	0	0
koliek, dracht, enting, wormen	0	0	0	0	1	0	0
eczeem	0	0	0	0	3	5	0
enting, eczeem	0	0	0	0	2	0	0
enting, diaree	0	0	0	0	1	0	0
enting, luis	0	0	0	0	1	0	0
enting, luis, tandarts	0	0	0	0	1	0	0
enting, euthanasie	0	0	0	0	2	0	0
euthanasie	0	0	0	0	45	47	47
enting, zand	0	0	0	0	4	0	0
geendiagnose	0	0	0	0	55	57	56
tandarts, geendiagnose	0	0	0	0	1	0	0
genezen	0	0	0	0	16	19	0
genezen, dracht	0	0	0	0	1	0	0
genezen, wormen	0	0	0	0	2	0	0
enting, hoefbevangen	0	0	0	0	8	0	0
hoefbevangen	0	0	0	0	184	205	208
hoefbevangen, hoefzweer	0	0	0	0	2	0	0
hoefbevangen, hoefzweer, koliek	0	0	0	0	1	0	0
hoefbevangen, wormen, enting	0	0	0	0	1	0	0
huid, enting	0	0	0	0	31	0	0
huid, hoefbevangen	0	0	0	0	4	0	0
koliek, dracht, hoefbevangen	0	0	0	0	1	0	0
luchtweg, hoefbevangen	0	0	0	0	3	0	0
enting, hoefsmid	0	0	0	0	27	0	0
enting, hoefsmid, wormen	0	0	0	0	1	0	0
hoefsmid	0	0	0	0	25	53	77

	Colic	Respiratory	Laminitis	Skin	All	Reduced	Simple
enting, hoefzweer	0	0	0	0	2	0	0
hoefzweer	0	0	0	0	5	29	0
hoefzweer, hoefsmid	0	0	0	0	22	0	0
dracht, huid	0	0	0	0	1	0	0
enting, huid	0	0	0	0	3	0	0
hoefzweer, huid	0	0	0	0	4	0	0
huid	0	0	0	0	491	602	607
huid, dracht	0	0	0	0	8	0	0
huid, dracht, wormen	0	0	0	0	1	0	0
huid, enting, hoefsmid	0	0	0	0	1	0	0
huid, hoefsmid	0	0	0	0	3	0	0
huid, keuring	0	0	0	0	2	0	0
huid, kreupel	0	0	0	0	33	0	0
huid, tandarts	0	0	0	0	3	0	0
huid, wormen	0	0	0	0	11	0	0
keuring, huid	0	0	0	0	1	0	0
keuring, huid, hoefsmid	0	0	0	0	1	0	0
koliek, huid	0	0	0	0	1	0	0
koliek, huid, enting	0	0	0	0	1	0	0
luchtweg, huid	0	0	0	0	2	0	0
oog, dracht, huid	0	0	0	0	1	0	0
oog, huid	0	0	0	0	1	0	0
schimmel, koliek	0	0	0	0	2	0	0
wond, huid	0	0	0	0	1	0	0
enting, keuring	0	0	0	0	51	0	0
enting, keuring, wormen	0	0	0	0	1	0	0
keuring	0	0	0	0	731	787	787
keuring, dracht	0	0	0	0	1	0	0
keuring, enting	0	0	0	0	1	0	0
kliniek	0	0	0	0	4	4	0
koliek	0	0	0	0	694	724	724
koliek euthanasie	0	0	0	0	1	0	0
koliek, dracht	0	0	0	0	2	0	0
koliek, enting	0	0	0	0	5	0	0
koliek, euthanasie	0	0	0	0	15	0	0
koliek, wormen	0	0	0	0	7	0	0
enting, kreupel	0	0	0	0	64	0	0
kreupel	0	0	0	0	255	320	319
kreupel, dracht	0	0	0	0	1	0	0
enting, luchtweg	0	0	0	0	9	0	0
koliek, luchtweg	0	0	0	0	4	0	0

	Colic	Respiratory	Laminitis	Skin	All	Reduced	Simple
kreupel, huid, luchtweg	0	0	0	0	1	0	0
luchtweg	0	0	0	0	253	286	286
luchtweg, dracht	0	0	0	0	4	0	0
luchtweg, eczeem	0	0	0	0	1	0	0
luchtweg, enting	0	0	0	0	7	0	0
luchtweg, kreupel	0	0	0	0	3	0	0
luchtweg, tandarts	0	0	0	0	3	0	0
luchtweg, wormen	0	0	0	0	1	0	0
enting, maagzweer	0	0	0	0	3	0	0
maagzweer	0	0	0	0	28	31	31
enting, oog	0	0	0	0	17	0	0
oog	0	0	0	0	31	48	48
enting, castratie	0	0	0	0	6	0	0
operatie	0	0	0	0	51	60	60
abces	0	0	0	0	1	0	0
allergie	0	0	0	0	2	0	0
artrose	0	0	0	0	1	0	0
blaasontsteking	0	0	0	0	1	0	0
enting, artrose	0	0	0	0	1	0	0
enting, bacteria	0	0	0	0	6	0	0
enting, bacteria, dracht	0	0	0	0	1	0	0
enting, bacteria, wormen	0	0	0	0	1	0	0
enting, virus	0	0	0	0	1	0	0
sloom	0	0	0	0	4	0	0
spierbevangen	0	0	0	0	2	0	0
virus	0	0	0	0	1	0	0
enting, ppid	0	0	0	0	1	0	0
ppid	0	0	0	0	46	47	48
dracht, tandarts	0	0	0	0	2	0	0
enting, tandarts	0	0	0	0	246	0	0
enting, tandarts, ppid	0	0	0	0	1	0	0
enting, tandarts, wormen	0	0	0	0	3	0	0
tandarts	0	0	0	0	323	648	647
tandarts, enting	0	0	0	0	1	0	0
tumor	0	0	0	0	37	37	37
enting, slokdarmverstopping	0	0	0	0	1	0	0
slokdarmverstopping	0	0	0	0	2	0	0
verstopping	0	0	0	0	2	5	0
enting, wond	0	0	0	0	26	0	0

	Colic	Respiratory	Laminitis	Skin	All	Reduced	Simple
wond	0	0	0	0	67	93	93
enting, wormen	0	0	0	0	230	0	0
enting, wormen, tandarts	0	0	0	0	3	0	0
enting, wormen, zand	0	0	0	0	3	0	0
hoefsmid, wormen	0	0	0	0	2	0	0
wormen	0	0	0	0	562	816	816
wormen, sloom	0	0	0	0	1	0	0
zand	0	0	0	0	0	3	0
overig	0	0	0	0	0	22	59

Table C.2: Results Prepare data: Occurrence of diseases

label set	classifier	accuracy	alpha	smooth_idf	sublinear_tf	use_idf	ngram_range	stop_words	fit_intercept	power_t	shuffle	tol	fit_prior
Reduced	Hi	0.89750	1e-3	T	F	T	(1,1)	T	T	0.5	F	1	
	L	0.86750	1e-3	T	T	T	(1,1)	F	T	0.5	T	1e-2	
	MH	0.90000	1e-3	T	T	T	(1,1)	T	T	0.5	F	1	
	SH	0.88000	1e-1	T	T	F	(1,1)	T	F	0.5	F	1	
	P	0.86500	1e-3	T	F	T	(1,1)	T	F	0.5	T	1	
	SL	0.89750	1e-3	T	F	T	(1,1)	T	T	0.5	F	1	
	Hu	0.87500	1e-2	T	T	T	(1,2)	T	F	0.5	F	1e-3	
	EI	0.88750	1e-3	T	T	T	(1,1)	T	T	0.5	F	1e-2	
	SEI	0.86750	1e-2	T	T	T	(1,1)	T	F	0.5	T	1	
	NB	0.88750	1e-1	T	T	T	(1,2)	T	F	0.5			F
Simple	Hi	0.89000	1e-3	T	F	T	(1,1)	T	T	0.5	T	1e-2	
	L	0.86750	1e-3	T	T	T	(1,1)	F	T	0.5	T	1	
	MH	0.90000	1e-3	T	T	T	(1,1)	T	T	0.5	T	1	
	SH	0.88500	1e-2	T	T	T	(1,1)	F	F	0.5	F	1e-3	
	P	0.87000	1e-3	T	T	T	(1,1)	T	F	0.5	T	1	
	SL	0.89500	1e-3	T	F	T	(1,1)	T	F	0.5	T	1e-1	
	Hu	0.87500	1e-2	T	T	T	(1,1)	T	F	0.5	F	1e-2	
	EI	0.88500	1e-3	T	F	T	(1,2)	T	T	0.5	T	1	
	SEI	0.87250	1e-2	T	T	T	(1,1)	T	F	0.5	F	1	
	NB	0.88250	1e-1	F	T	T	(1,1)	T	F	0.5			F

Table C.3: Results of parameter tuning for the label sets Reduced and Simple, to test each loss function: hinge (Hi), log (L), modified_huber (MH), squared_hinge (SH), perceptron (P), squared_loss (SL), huber (Hu), epsilon_insensitive (EI), squared_epsilon_insensitive (SEI) and the naïve Bayes classifier (NB)

label set	classifier	accuracy	alpha	smooth_idf	sublinear_tf	use_idf	ngram_range	stop_words	fit_intercept	power_t	shuffle	tol	fit_prior
Colic	Hi	0.97000	1e-3	T	F	F	(1,1)	T	T	0.1	T	1e-1	
	L	0.96250	1e-3	T	F	F	(1,1)	T	T	0.1	T	1	
	MH	0.97000	1e-3	F	F	T	(1,1)	T	T	0.1	F	1e-4	
	SH	0.97250	1e-3	T	T	F	(1,2)	T	T	0.1	T	1e-3	
	P	0.96250	1e-2	T	T	F	(1,1)	T	T	0.1	T	1e-4	
	SL	0.96500	1e-2	T	F	F	(1,1)	T	T	0.1	T	1	
	Hu	0.97000	1e-3	T	T	F	(1,1)	T	T	0.1	T	1e-4	
	EI	0.97000	1e-2	T	T	F	(1,1)	T	T	0.1	T	1e-4	
	SEI	0.96750	1e-2	T	F	T	(1,1)	T	T	0.1	T	1	
	NB	0.95250	1e-1	T	T	F	(1,1)	T	F	0.1			F
Laminitis	Hi	0.95679	1e-3	T	F	F	(1,1)	F	T	0.1	T	1	
	L	0.92811	1e-3	T	F	F	(1,1)	T	T	0.1	F	1e-2	
	MH	0.94239	1e-3	T	T	F	(1,1)	T	T	0.1	T	1e-1	
	SH	0.94228	1e-2	T	T	F	(1,1)	F	T	0.1	F	1	
	P	0.94715	1e-2	T	F	F	(1,2)	T	T	0.1	F	1e-3	
	SL	0.92811	1e-2	T	F	F	(1,1)	T	T	0.1	T	1e-3	
	Hu	0.93275	1e-3	T	T	F	(1,1)	T	T	0.1	F	1e-4	
	EI	0.94715	1e-3	T	T	F	(1,1)	F	T	0.1	T	1e-4	
	SEI	0.93275	1e-2	T	T	F	(1,1)	T	T	0.1	T	1e-1	
	NB	0.90894	1e-1	T	T	F	(1,1)	T	F	0.1			F
Respiratory	Hi	0.92764	1e-3	T	T	F	(1,1)	T	T	0.1	F	1e-4	
	L	0.91040	1e-3	T	F	F	(1,1)	T	F	0.1	T	1e-1	
	MH	0.92764	1e-3	T	T	F	(1,1)	T	T	0.1	F	1e-3	
	SH	0.91730	1e-2	T	T	F	(1,1)	T	T	0.1	T	1e-3	
	P	0.92420	1e-2	T	F	F	(1,1)	T	T	0.1	T	1e-4	
	SL	0.90695	1e-1	T	T	F	(1,1)	T	F	0.1	T	1	
	Hu	0.90695	1e-3	T	T	F	(1,1)	T	F	0.1	T	1e-3	
	EI	0.91730	1e-3	T	T	F	(1,1)	T	T	0.1	T	1e-1	
	SEI	0.92075	1e-2	T	T	F	(1,1)	T	T	0.1	F	1e-1	
	NB	0.90006	1	T	F	F	(1,1)	T	T	0.1			T
Skin	Hi	0.95750	1e-3	T	T	F	(1,1)	T	T	0.1	F	1e-3	
	L	0.94500	1e-3	F	F	T	(1,1)	T	F	0.1	T	1	
	MH	0.95500	1e-3	T	T	F	(1,2)	T	T	0.1	F	1e-3	
	SH	0.95250	1e-2	T	T	F	(1,1)	T	T	0.1	T	1e-3	
	P	0.94500	1e-2	T	T	F	(1,2)	T	T	0.1	F	1e-4	
	SL	0.94250	1e-2	T	T	F	(1,1)	T	F	0.1	T	1	
	Hu	0.92750	1e-3	T	T	F	(1,1)	T	F	0.1	T	1e-4	
	EI	0.95000	1e-3	T	T	F	(1,1)	F	T	0.1	T	1	
	SEI	0.95000	1e-2	T	T	F	(1,1)	T	T	0.1	T	1e-2	
	NB	0.94250	1e-1	T	T	F	(1,1)	T	F	0.1			F

Table C.4: Results of parameter tuning for the label sets Colic, Laminitis, Respiratory and Skin, to test each loss function: hinge (Hi), log (L), modified_huber (MH), squared_hinge (SH), perceptron (P), squared_loss (SL), huber (Hu), epsilon_insensitive (EI), squared_epsilon_insensitive (SEI) and the naïve Bayes classifier (NB)

		Colic	Laminitis	Respiratory	Skin	Reduced	Simple
classifier	Hi	x	x	x	x	x	x
	MH	x	x	x	x	x	x
	SH	x	x	x	x	x	x
	P		x	x			x
	Hu	x					
	EI	x	x	x	x	x	x
	SEI			x	x		
	NB					o	
alpha	1e-1					o	
	1e-2	x	x	x	x		x
	1e-3	x	x	x	x	x	x
smooth_idf	True	x	x	x	x	x o	x
	False	x					
sublinear_tf	True	x	x	x	x	x o	x
	False	x	x	x		x	x
use_idf	True	x				x o	x
	False	x	x	x	x		
ngram_range	(1,1)	x	x	x	x	x	x
	(1,2)	x	x		x	o	x
stop_words	True	x	x	x	x	x o	x
	False		x		x		x
fit_intercept	True	x	x	x	x	x	x
	False						x
power_t	0.5					x	x
	0.1	x	x	x	x		
shuffle	True	x	x	x	x		x
	False	x	x	x	x	x	x
tol	1	x	x		x	x	x
	1e-1	x	x	x			x
	1e-2				x	x	x
	1e-3	x	x	x	x		x
	1e-4	x	x	x			
fit_prior	False					o	

Table C.5: For each label set the classifiers and parameters used. x = used in Stochastic Gradient Descent and o = used in naive Bayes. Parameters that are not used for any of the label sets are not shown. The used classifiers are: hinge (Hi), modified_huber (MH), squared_hinge (SH), perceptron (P), huber (Hu), epsilon_insensitive (EI), squared_epsilon_insensitive (SEI) and the naive Bayes classifier (NB)

		Hi	MH	SH	Hu	EI
count		37	26	6	0	31
alpha	1e-2	0	6	0	0	2
	1e-3	37	20	6	0	29
smooth_idf	TRUE	37	26	6	0	31
	FALSE	0	0	0	0	0
sublinear_tf	TRUE	26	16	5	0	21
	FALSE	11	10	1	0	10
use_idf	TRUE	0	0	0	0	0
	FALSE	37	26	6	0	31
ngram_range	(1, 1)	31	22	4	0	30
	(1, 2)	6	4	2	0	1
stop_words	TRUE	37	26	6	0	31
fit_intercept	TRUE	37	26	6	0	31
power_t	0.1	37	26	6	0	31
shuffle	TRUE	21	16	5	0	20
	FALSE	16	10	1	0	11
tol	1	9	12	1	0	8
	1e-1	11	6	3	0	10
	1e-3	9	2	2	0	4
	1e-4	8	6	0	0	9
avg accuracy		0.98189	0.98240	0.98000	-	0.98331
max accuracy		0.99000	0.99250	0.98500	-	0.99250

Table C.6: The results of parameter tuning 100 times for the ”colic” label set, the number of times each classifier is considered best (count) and the number of times each parameter is used. Classifiers used: hinge (Hi), modified_huber (MH), squared_hinge (SH), huber (Hu), epsilon_insensitive (EI)

		Hi	MH	SH	P	EI
count		17	27	1	38	17
alpha	1e-2	0	2	0	31	3
	1e-3	17	25	1	7	14
smooth_idf	TRUE	17	27	1	38	17
sublinear_tf	TRUE	14	23	1	31	16
	FALSE	3	4	0	7	1
use_idf	FALSE	17	27	1	38	17
ngram_range	(1, 1)	17	25	1	29	16
	(1, 2)	0	2	0	9	1
stop_words	TRUE	15	22	1	30	15
	FALSE	2	5	0	8	2
fit_intercept	TRUE	17	27	1	38	17
power_t	0.1	17	27	1	38	17
shuffle	TRUE	11	12	1	14	11
	FALSE	6	15	0	24	6
tol	1	3	13	0	16	0
	1e-1	8	6	0	0	6
	1e-3	1	5	1	8	6
	1e-4	5	3	0	14	5
avg accuracy		0.95910	0.96150	0.97131	0.95619	0.95911
max accuracy		0.97607	0.98084	0.97131	0.97143	0.98072

Table C.7: The results of parameter tuning 100 times for the ”laminitis” label set, the number of times each classifier is considered best (count) and the number of times each parameter is used. Classifiers used: hinge (Hi), modified_huber (MH), squared_hinge (SH), perceptron (P), epsilon_insensitive (EI)

		Hi	MH	SH	P	EI	SEI
count		23	39	4	11	17	6
alpha	1e-2	0	9	4	5	2	6
	1e-3	23	30	0	6	15	0
smooth_idf	TRUE	23	39	4	11	17	6
sublinear_tf	TRUE	12	23	0	7	12	1
	FALSE	11	16	4	4	5	5
use_idf	FALSE	23	39	4	11	17	6
ngram_range	(1, 1)	23	39	4	11	17	6
stop_words	TRUE	23	39	4	11	17	6
fit_intercept	TRUE	23	39	4	11	17	6
power_t	0.1	23	39	4	11	17	6
shuffle	TRUE	18	22	4	4	11	4
	FALSE	5	17	0	7	6	2
tol	1e-1	10	20	3	6	6	0
	1e-3	7	16	1	1	5	4
	1e-4	6	3	0	4	6	2
avg accuracy		0.95303	0.95415	0.95529	0.95063	0.95694	0.95189
max accuracy		0.97943	0.97247	0.97241	0.96908	0.97247	0.96908

Table C.8: The results of parameter tuning 100 times for the ”respiratory” label set, the number of times each classifier is considered best (count) and the number of times each parameter is used. Classifiers used: hinge (Hi), modified_huber (MH), squared_hinge (SH), perceptron (P), epsilon_insensitive (EI), squared_epsilon_insensitive (SEI)

		Hi	MH	SH	EI	SEI
count		38	44	2	8	8
alpha	1e-2	0	4	2	0	8
	1e-3	38	40	0	8	0
smooth_idf	TRUE	38	44	2	8	8
sublinear_tf	TRUE	38	44	2	8	8
use_idf	FALSE	38	44	2	8	8
ngram_range	(1, 1)	33	42	2	7	5
	(1, 2)	5	2	0	1	3
stop_words	TRUE	32	40	2	5	8
	FALSE	6	4	0	3	0
fit_intercept	TRUE	38	44	2	8	8
power_t	0.1	38	44	2	8	8
shuffle	TRUE	29	27	2	5	1
	FALSE	9	17	0	3	7
tol	1	8	22	2	2	2
	1e-2	14	8	0	2	2
	1e-3	16	14	0	4	4
avg accuracy		0.95612	0.95545	0.95625	0.95156	0.94906
max accuracy		0.97500	0.97250	0.95750	0.95750	0.95750

Table C.9: The results of parameter tuning 100 times for the ”skin” label set, the number of times each classifier is considered best (count) and the number of times each parameter is used. Classifiers used: hinge (Hi), modified_huber (MH), squared_hinge (SH), epsilon_insensitive (EI), squared_epsilon_insensitive (SEI)

		Hi	MH	SH	EI	NB
count		0	100	0	0	0
alpha	1e-1	0	0	0	0	0
	1e-3	0	100	0	0	0
smooth_idf	TRUE	0	100	0	0	0
sublinear_tf	TRUE	0	100	0	0	0
	FALSE	0	0	0	0	0
use_idf	TRUE	0	100	0	0	0
ngram_range	(1, 1)	0	100	0	0	0
	(1, 2)	0	0	0	0	0
stop_words	TRUE	0	100	0	0	0
fit_intercept	TRUE	0	100	0	0	0
power_t	0.5	0	100	0	0	0
shuffle	FALSE	0	100	0	0	0
tol	1	0	100	0	0	0
	1e-2	0	0	0	0	0
avg accuracy		-	0.90000	-	-	-
max accuracy		-	0.90000	-	-	-

Table C.10: The results of parameter tuning 100 times for the ”reduced” label set, the number of times each classifier is considered best (count) and the number of times each parameter is used. Classifiers used: hinge (Hi), modified_huber (MH), squared_hinge (SH), epsilon_insensitive (EI), Naïve Bayes (NB)

		Hi	MH	SH	P	EI
count		0	100	0	0	0
alpha	1e-2	0	0	0	0	0
	1e-3	0	100	0	0	0
smooth_idf	TRUE	0	100	0	0	0
sublinear_tf	TRUE	0	100	0	0	0
	FALSE	0	0	0	0	0
use_idf	TRUE	0	100	0	0	0
ngram_range	(1, 1)	0	100	0	0	0
	(1, 2)	0	0	0	0	0
stop_words	TRUE	0	100	0	0	0
	FALSE	0	0	0	0	0
fit_intercept	TRUE	0	100	0	0	0
	FALSE	0	0	0	0	0
power_t	0.5	0	100	0	0	0
shuffle	TRUE	0	100	0	0	0
	FALSE	0	0	0	0	0
tol	1	0	100	0	0	0
	1e-1	0	0	0	0	0
	1e-2	0	0	0	0	0
	1e-3	0	0	0	0	0
avg accuracy		-	0.90000	-	-	-
max accuracy		-	0.90000	-	-	-

Table C.11: The results of parameter tuning 100 times for the ”simple” label set, the number of times each classifier is considered best (count) and the number of times each parameter is used. Classifiers used: hinge (Hi), modified_huber (MH), squared_hinge (SH), perceptron (P), epsilon_insensitive (EI)

	dracht	enting	euthanasie	geendiagnose	hoefbevangen	hoefsmid	huid	keuring	koliek	kreupel	luchtweg	maagzweer	oog	operatie	overig	ppid	tandarts	tumor	wond	wormen
dracht	1054	1	0	4	1	1	1	2	2	4	2	1	3	2	0	1	1	0	1	6
enting	11	1713	1	1	7	10	8	16	1	18	11	1	3	4	7	3	48	2	6	23
euthanasie	0	0	24	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0
geendiagnose	0	0	0	17	2	0	2	0	5	0	3	1	0	0	0	0	0	0	0	1
hoefbevangen	1	0	0	1	87	1	1	0	2	4	2	0	0	0	0	1	0	0	0	0
hoefsmid	0	0	0	0	0	29	0	0	0	2	1	0	0	0	0	0	0	0	0	0
huid	3	2	0	3	0	1	267	3	3	12	4	0	1	0	0	0	2	0	1	3
keuring	0	0	0	0	0	1	2	368	0	4	0	0	0	0	0	0	1	0	3	0
koliek	2	0	0	3	1	0	1	1	339	1	1	0	0	0	1	1	1	0	0	0
kreupel	1	1	1	0	3	4	3	4	1	149	2	0	0	0	1	0	3	0	2	0
luchtweg	2	2	0	1	0	0	1	0	2	4	119	0	0	0	2	1	0	0	3	1
maagzweer	0	0	0	0	1	0	0	0	3	0	0	10	0	0	0	0	0	0	0	0
oog	1	1	0	0	0	0	1	0	0	0	0	0	22	0	0	0	1	0	0	0
operatie	0	0	0	0	0	0	0	0	0	0	2	0	0	26	0	0	0	0	1	0
overig	0	0	0	0	0	2	2	0	9	2	2	0	0	0	6	0	0	0	1	3
ppid	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	13	1	0	0	0
tandarts	1	3	0	2	0	1	0	3	0	2	0	0	0	0	0	2	286	0	0	9
tumor	0	0	0	0	1	0	0	0	1	0	1	0	1	1	1	0	0	11	0	0
wond	0	2	0	0	0	0	1	0	0	2	0	0	0	2	0	0	1	0	29	0
wormen	0	4	0	3	0	0	1	1	1	2	0	1	0	0	12	0	5	0	1	394

Table C.12: The confusion matrix for prediction of the ”simple” label set

	dracht	eczeem	enting	euthanasie	geendiagnose	genezen	hoefbevangen	hoefsmid	hoefzweer	huid	keuring	kliniek	koliek	kreupel	luchtweg	maagzweer	oog	operatie	overig	ppid	tandarts	tumor	verstopping	wond	wormen	zand
dracht	1045	0	0	1	1	0	1	0	0	1	3	0	2	4	4	1	3	2	1	2	1	1	1	4	9	0
eczeem	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
enting	12	2	250	1	2	5	9	21	0	3	18	3	0	31	13	7	19	5	9	3	139	3	0	30	1311	0
euthanasie	0	0	0	24	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0
geendiagnose	0	0	0	0	15	0	2	0	0	2	0	0	5	0	4	1	0	0	1	0	0	0	0	0	1	0
genezen	0	0	0	0	0	0	0	0	0	2	0	0	4	0	1	0	0	0	0	0	0	0	0	0	1	0
hoefbevangen	1	0	0	0	1	0	75	0	1	0	0	0	3	9	2	1	0	0	0	3	0	0	0	1	1	0
hoefsmid	0	0	0	0	0	0	0	20	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0
hoefzweer	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
huid	3	0	0	0	2	0	0	0	0	254	3	0	3	21	4	0	1	0	0	0	2	0	0	5	5	0
keuring	0	0	0	0	0	0	0	0	0	2	368	0	0	4	1	0	0	0	0	0	1	0	0	3	0	0
kliniek	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
koliek	1	0	0	0	1	0	1	0	0	0	1	0	328	3	2	2	0	0	1	1	1	1	1	0	8	0
kreupel	0	0	0	0	0	0	0	0	1	3	4	0	1	149	4	1	1	0	1	0	4	1	0	5	0	0
luchtweg	1	0	0	0	0	0	0	0	0	1	0	0	1	3	126	0	0	0	0	1	0	0	0	4	1	0
maagzweer	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	12	0	0	0	0	0	0	0	1	0	0
oog	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	22	0	0	0	1	1	0	0	0	0
operatie	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	26	0	0	0	0	0	1	0	0
overig	0	0	0	0	0	0	0	0	0	0	0	0	2	2	1	0	0	0	4	0	0	0	0	0	1	0
ppid	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	13	1	0	0	0	0	0
tandarts	0	0	0	0	0	0	0	0	0	0	1	0	0	3	1	1	0	0	0	0	289	0	0	1	13	0
tumor	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	1	0	0	0	13	0	0	0	0
verstopping	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
wond	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	2	0	0	1	0	0	31	0	0
wormen	0	0	0	0	3	1	0	0	0	1	0	0	1	2	1	1	0	0	7	0	5	0	0	1	402	0
zand	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	0

Table C.13: The confusion matrix for prediction of the ”reduced” label set

D TEMPERATURE

Temperature	Incorrect Label	Correct label	Labeled	Actual	% Correctly labeled
42.4	2	0	2	0	0%
42.35	1	0	1	0	0%
42.33	1	0	1	0	0%
42.0	77	0	77	0	0%
41.8	1	0	1	0	0%
40.05	1	0	1	0	0%
39.95	1	0	1	0	0%
39.94	1	0	1	0	0%
39.12	1	0	1	0	0%
38.12	1	0	1	0	0%
40.0	1009	17	1026	17	2%
37.0	82	16	98	16	16%
41.0	7	6	13	6	46%
39.0	42	47	89	47	53%
38.0	24	109	133	108	82%
40.3	1	6	7	6	86%
39.7	1	12	13	12	92%
40.1	1	13	14	13	93%
38.5	5	76	81	76	94%
39.9	1	19	20	19	95%
37.1	1	23	24	23	96%
38.8	2	48	50	48	96%
37.2	1	25	26	25	96%
38.6	2	53	55	53	96%
39.4	1	29	30	29	97%
37.9	3	94	97	93	97%
38.4	2	66	68	65	97%
38.3	1	50	51	50	98%
38.2	1	103	104	103	99%
42.7	0	1	1	1	100%
41.3	0	1	1	1	100%
41.2	0	1	1	1	100%
41.1	0	1	1	1	100%
40.9	0	3	3	3	100%
40.8	0	3	3	3	100%
40.6	0	5	5	5	100%
40.5	0	10	10	10	100%
40.4	0	3	3	3	100%
40.2	0	13	13	13	100%
39.8	0	18	18	19	100%
39.6	0	21	21	21	100%
39.5	0	36	36	36	100%
39.3	0	27	27	27	100%
39.2	0	22	22	22	100%
39.1	0	30	30	30	100%
38.9	0	32	32	32	100%
38.7	0	43	43	43	100%
38.1	0	87	87	87	100%
37.8	0	116	116	117	100%
37.7	0	89	89	89	100%
37.6	0	102	102	101	100%
37.5	0	73	73	73	100%
37.4	0	88	88	88	100%
37.3	0	37	37	37	100%
36.9	0	9	9	9	100%
36.8	0	9	9	9	100%
36.7	0	3	3	3	100%
36.5	0	3	3	3	100%
36.3	0	1	1	1	100%
36.2	0	1	1	1	100%

Table D.1: For each temperature: the number of incorrect labels, correct labels, the total amount of labels, the number of times the temperature actually occurs in the texts of the medical file and the percentage of correctly labeled temperatures.

		Actual						
		0	1	2	3	4	5	6
Prediction	0	56619	0	0	0	0	0	0
	1	617	1220	0	0	0	0	0
	2	125	129	93	0	0	0	0
	3	20	37	5	15	0	0	0
	4	9	9	5	2	3	0	0
	5	4	3	2	3	1	1	0
	6	2	1	0	1	0	1	0

		Actual						
		0	1	2	3	4	5	6
Prediction	0	0,98646	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
	1	0,01075	0,87205	0,00000	0,00000	0,00000	0,00000	0,00000
	2	0,00218	0,09221	0,88571	0,00000	0,00000	0,00000	0,00000
	3	0,00035	0,02645	0,04762	0,71429	0,00000	0,00000	0,00000
	4	0,00016	0,00643	0,04762	0,09524	0,75000	0,00000	0,00000
	5	0,00007	0,00214	0,01905	0,14286	0,25000	0,50000	0,00000
	6	0,00003	0,00071	0,00000	0,04762	0,00000	0,50000	0,00000

Table D.2: Confusion matrix for the actual and predicted number of temperatures per consult for all temperatures found. The accuracy is 0.98344

		Actual						
		0	1	2	3	4	5	6
Prediction	0	57332	25	0	0	0	0	0
	1	57	1356	4	0	0	0	0
	2	5	15	100	1	1	0	0
	3	1	3	1	20	1	0	0
	4	0	0	0	0	2	0	0
	5	0	0	0	0	0	2	0
	6	1	0	0	0	0	0	0

		Actual						
		0	1	2	3	4	5	6
Prediction	0	0,99888	0,01787	0,00000	0,00000	0,00000	0,00000	0,00000
	1	0,00099	0,96926	0,03810	0,00000	0,00000	0,00000	0,00000
	2	0,00009	0,01072	0,95238	0,04762	0,25000	0,00000	0,00000
	3	0,00002	0,00214	0,00952	0,95238	0,25000	0,00000	0,00000
	4	0,00000	0,00000	0,00000	0,00000	0,50000	0,00000	0,00000
	5	0,00000	0,00000	0,00000	0,00000	0,00000	1,00000	0,00000
	6	0,00002	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000

Table D.3: Confusion matrix for the actual and predicted number of temperatures per consult for all temperatures that are labeled correctly more then 20% of the time, as shown in D.1. The accuracy is 0.99805

E VISUALIZATION

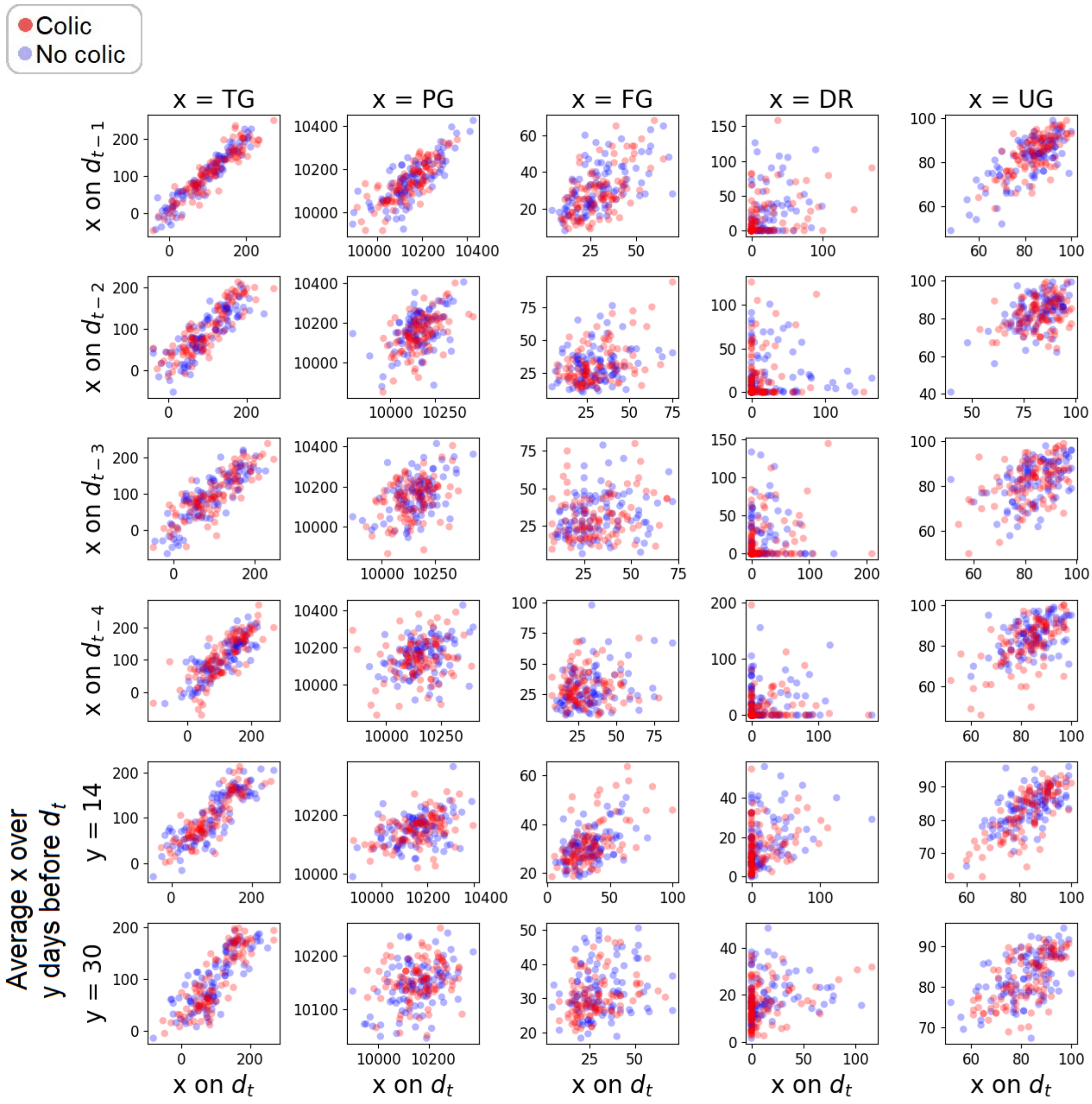


Figure E.1: 100 randomly selected values on d_t for the daily mean temperature (TG) in 0.1 degrees Celsius, the daily mean sea level pressure (PG) in 0.1 hPa, the daily mean wind speed (FG) in o.1 m/s, the precipitation duration (RH) in 0.1 hours and the daily mean relative atmospheric humidity (UG) in percentage (from left to right), against the value on d_{t-1}, d_{t-2}, d_{t-3}, d_{t-4}, the average over 14 and 30 days prior to d_t (top to bottom) for the same variable. d_t are all days in the weather data set, the red colored dots are the dates of a consult concerning colic.

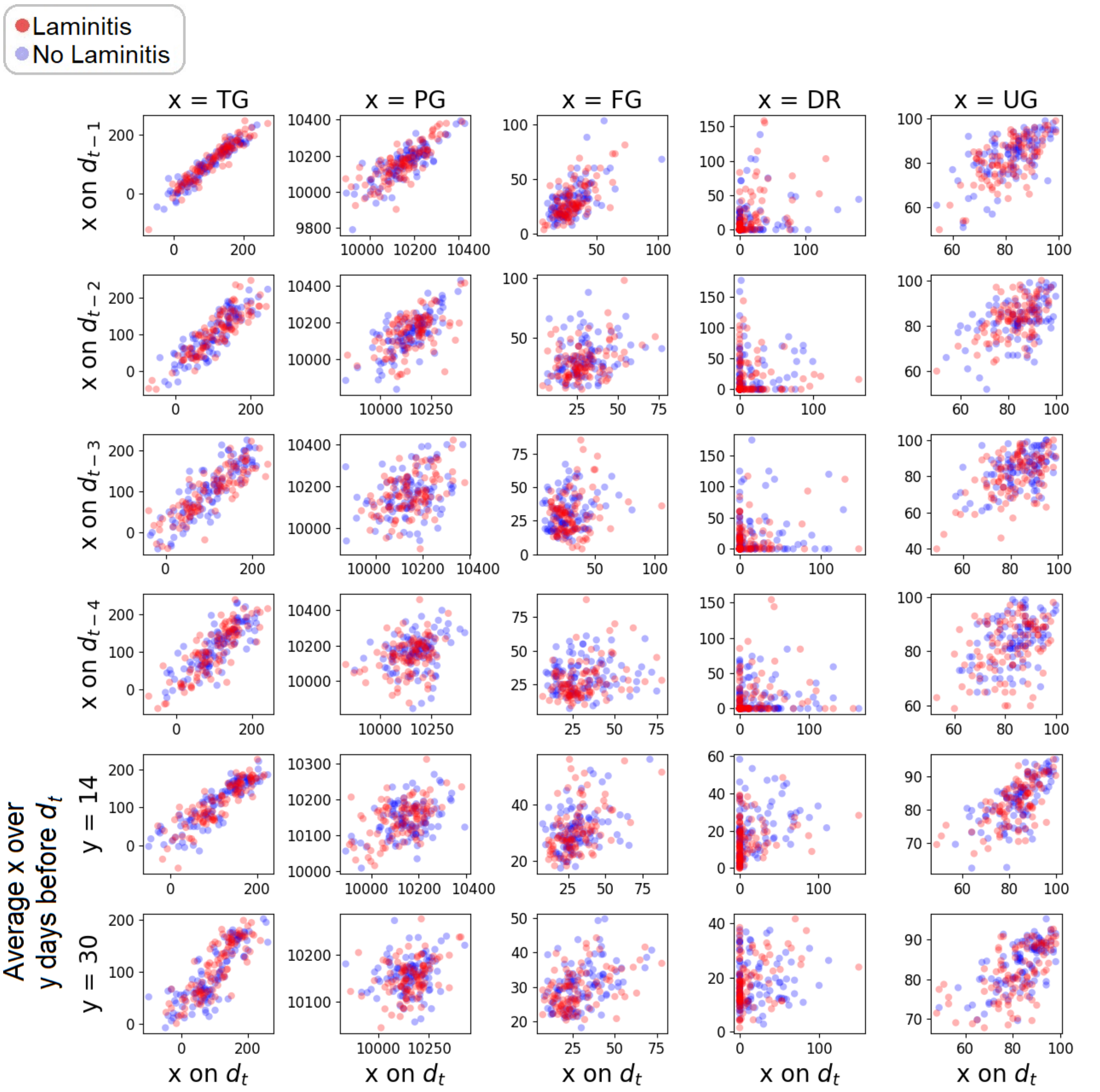


Figure E.2: 100 randomly selected values on d_t for the daily mean temperature (TG) in 0.1 degrees Celsius, the daily mean sea level pressure (PG) in 0.1 hPa, the daily mean wind speed (FG) in o.1 m/s, the precipitation duration (RH) in 0.1 hours and the daily mean relative atmospheric humidity (UG) in percentage (from left to right), against the value on d_{t-1}, d_{t-2}, d_{t-3}, d_{t-4}, the average over 14 and 30 days prior to d_t (top to bottom) for the same variable. d_t are all days in the weather data set, the red colored dots are the dates of a consult concerning laminitis.

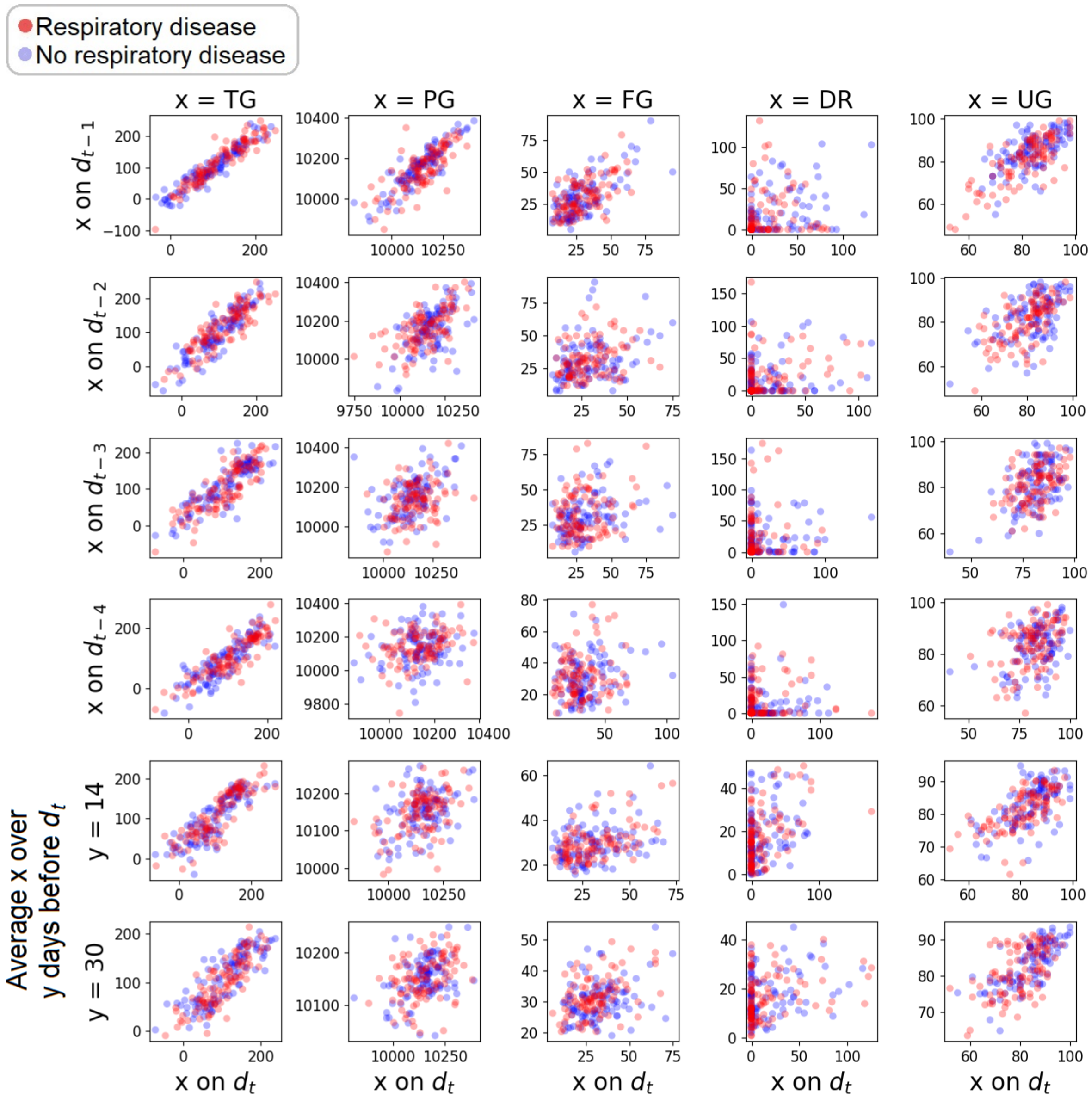


Figure E.3: 100 randomly selected values on d_t for the daily mean temperature (TG) in 0.1 degrees Celsius, the daily mean sea level pressure (PG) in 0.1 hPa, the daily mean wind speed (FG) in o.1 m/s, the precipitation duration (RH) in 0.1 hours and the daily mean relative atmospheric humidity (UG) in percentage (from left to right), against the value on d_{t-1}, d_{t-2}, d_{t-3}, d_{t-4}, the average over 14 and 30 days prior to d_t (top to bottom) for the same variable. d_t are all days in the weather data set, the red colored dots are the dates of a consult concerning respiratory disease

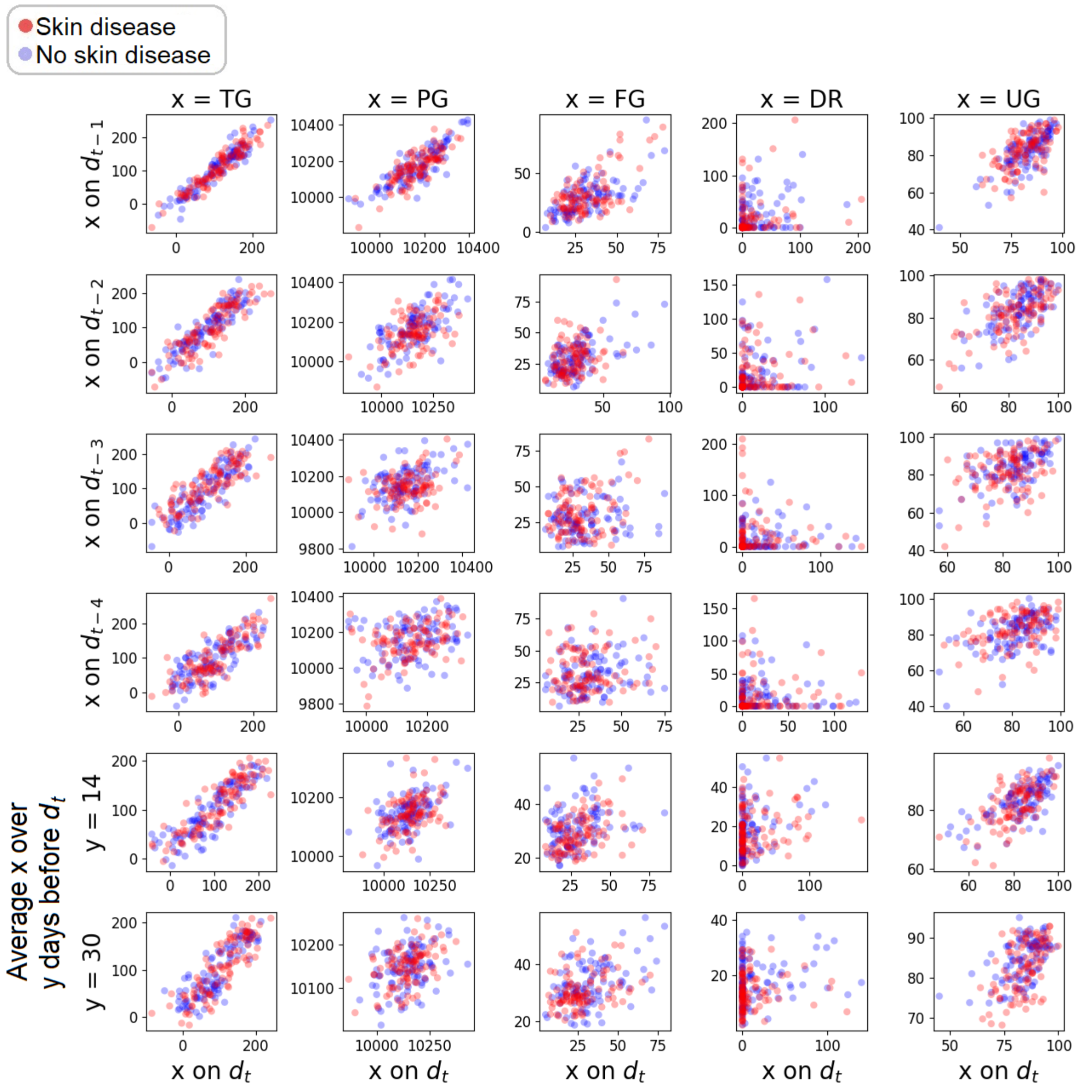


Figure E.4: 100 randomly selected values on d_t for the daily mean temperature (TG) in 0.1 degrees Celsius, the daily mean sea level pressure (PG) in 0.1 hPa, the daily mean wind speed (FG) in o.1 m/s, the precipitation duration (RH) in 0.1 hours and the daily mean relative atmospheric humidity (UG) in percentage (from left to right), against the value on d_{t-1}, d_{t-2}, d_{t-3}, d_{t-4}, the average over 14 and 30 days prior to d_t (top to bottom) for the same variable. d_t are all days in the weather data set, the red colored dots are the dates of a consult concerning skin.

F PREDICTIONS

			Accuracy	Precision	Recall
Colic	Voting		0.68338	0.61644	0.03586
	Boosting		0.69207	0.58824	0.13546
	Bagging	LR	0.67954	0.50595	0.06773
		SVM	0.67903	-	0.00000
		DT	0.70102	0.56869	0.28367
		NN	0.67826	0.20000	0.00080
Laminitis	Voting		0.63555	0.42553	0.01580
	Boosting		0.62583	0.46965	0.11611
	Bagging	LR	0.63376	0.46226	0.11611
		SVM	0.63529	-	0.00000
		DT	0.64680	0.49287	0.19115
		NN	0.63146	0.44091	0.07662
Respiratory	Voting		0.79847	-	0.00000
	Boosting		0.79463	0.33333	0.01904
	Bagging	LR	0.79821	0.42857	0.00381
		SVM	0.79847	-	0.00000
		DT	0.79361	0.32075	0.02157
		NN	0.79719	0.40741	0.01396
Skin	Voting		0.66215	0.69354	0.92027
	Boosting		0.72225	0.71759	0.93149
	Bagging	LR	0.66880	0.70104	0.83894
		SVM	0.63836	0.63836	1.00000
		DT	0.74194	0.89303	0.75025
		NN	0.64501	0.90785	0.66180

Table F.1: Accuracy, precision and recall of the predictions of diseases using the weather variables and bagging, boosting and voting. LR = Linear Regression, SVM = Support Vector Machine, DT = Decision Tree and NN = Neural Network

		Accuracy	Precision	Recall
Colic	LR	0.67749	0.48193	0.06375
	SVM	0.67903	-	0.00000
	DT	0.63683	0.4356	0.44462
	NN	0.46343	0.33373	0.6741
Laminitis	LR	0.64089	0.44413	0.12243
	SVM	0.65162	-	0.0
	DT	0.59659	0.40363	0.41817
	NN	0.65162	0.5	0.00079
Respiratory	LR	0.79821	0.50000	0.00381
	SVM	0.79847	-	0.0
	DT	0.67059	0.21065	0.23096
	NN	0.69463	0.22493	0.21066
Skin	LR	0.66803	0.7057	0.82332
	SVM	0.63836	0.63836	1.0
	DT	0.65115	0.72767	0.72476
	NN	0.62711	0.64103	0.94511

Table F.2: The confusion matrices and accuracy, precision and recall of the predictions of the diseases using the weather variables and single classifiers Linear Regression (LR), Support Vector Machine (SVM), Decision Tree (DT) and Neural Network (NN)

		Colic				Laminitis				Respiratory				Skin			
Voting				P				P				P				P	
				0	1			0	1			0	1			0	1
		A	0	2627	28	A	0	2341	27	A	0	3122,	0	A	0	399	1015
			1	1210	45		1	1246	20		1	788	0		1	199	2297
Boosting				P				P				P				P	
				0	1			0	1			0	1			0	1
		A	0	2536	119	A	0	2202	166	A	0	3092	30	A	0	499	915
			1	1085	170		1	1119	147		1	773	15		1	171	2325
Bagging	LR			P				P				P				P	
				0	1			0	1			0	1			0	1
		A	0	2572	83	A	0	2197	171	A	0	3118	4	A	0	521	893
			1	1170	85		1	1119	147		1	785	3		1	402	2094
	SVM			P				P				P				P	
				0	1			0	1			0	1			0	1
		A	0	2655	0	A	0	2368	0	A	0	3122	0	A	0	0	1414
			1	1255	0		1	1266	0		1	788	0		1	0	2496
	DT			P				P				P				P	
				0	1			0	1			0	1			0	1
		A	0	2385	270	A	0	2119	249	A	0	3086	36	A	0	672	742
			1	899	356		1	1024	242		1	771	17		1	267	2229
	NN			P				P				P				P	
				0	1			0	1			0	1			0	1
		A	0	2651	4	A	0	2245	123	A	0	3106	16	A	0	256	1158
			1	1254	1		1	1169	97		1	777	11		1	230	2266

Table F.3: The confusion matrices of the prediction of diseases using bagging, boosting and voting. P = Predicted, A = Actual, 0 = not disease, 1 = disease, LR = Linear Regression, SVM = Support Vector Machine, DT = Decision Tree, NN = Neural Network.

	Colic				Laminitis				Respiratory				Skin			
LR			P				P				P				P	
			0	1			0	1			0	1			0	1
	A	0	2569	86	A	0	2174	194	A	0	3119	3	A	0	557	857
		1	1175	80		1	1111	155		1	785	3		1	441	2055
SVM			P				P				P				P	
			0	1			0	1			0	1			0	1
	A	0	2655	0	A	0	2368	0	A	0	3122	0	A	0	0	1414
		1	1255	0		1	1266	0		1	788	0		1	0	2496
DT			P				P				P				P	
			0	1			0	1			0	1			0	1
	A	0	1932	723	A	0	1657	711	A	0	2440	682	A	0	737	677
		1	697	558		1	755	511		1	606	182		1	687	1809
NN			P				P				P				P	
			0	1			0	1			0	1			0	1
	A	0	966	1689	A	0	2367	1	A	0	2550	572	A	0	93	1321
		1	409	846		1	1265	1		1	622	166		1	137	2359

Table F.4: The confusion matrices of the prediction of diseases using single classifiers: Linear Regression (LR), Support Vector Machine (SVM), Decision Tree (DT) and Neural Network (NN). P = Predicted, A = Actual, 0 = not disease, 1 = disease.