1. Hi everyone, welcome to my presentation and of course a happy new year!
2. During this presentation I will give a short overview of the motivation and background of my research. Then I will talk about the research questions answered during this research, the methodology used for answering these questions and the results found. Then I will discuss the answers to the questions in the conclusion. Afterwards will be time for questions.
3. Lets start with the motivation and background!
4. Many horse owners and equine professionals make assumptions about the health of horses. Some of these assumptions claim that specific weather conditions are responsible for particular diseases. People believe, for example, that sudden high amounts of wind can cause colic, high humidity can cause respiratory problems and skin problems like fungal infections. Muddy pastures cause mud fever and laminitis is caused by high amounts of sugar in grass, due to cold nights. The goal of this research is to see which assumptions are true and which are false. This could help horse owners and professionals to more accurately diagnose horses when seeing symptoms.
5. During this research we focussed on Colic: severe abdominal pain, Laminitis: inflammation in the hoofs, respiratory disease: everything from coughing or snotty noses to pneumonia and skin disease: like fungal or bacterial infections or allergic reactions, such as mud fever; wounds just above the hoofs
6. The research questions of this research are based on a literature review of the relation of the diseases and the weather
7. The first question, answered in this research is “what is the influence of the Dutch weather on the health of horses?” This questions is divided into four sub questions: does temperature, barometric pressure and high amounts of wind influence the occurrence of colic?, is the development of laminitis dependent on stress in the grass, due to cold and drought? Does hot, humid, or cold weather worsen or induce respiratory disease? Do skin diseases occur more in periods of heavy rainfall and high humidity?
8. The second research question is “To what extend can the Dutch weather be used to predict the occurrence of Colic, laminitis, respiratory disease and skin disease?”
9. –
10. For this research, two datasets are merged: the data of animal clinic den ham and the weather data. For the weather data, the KNMI dataset of weather station Heino is used. Den ham is the weather station that is closest to den ham. The weather data contains one row for each day. For this research the period of 13 November 98 till 9 April 2020 is used to cover the data of the veterinarian, starting just a bit later, at 20 January 99. The data contains 34 different variables such as the daily main wind speed, maximum wind speed and the hour in which this maximum is measured. These variables are divided over 7 categories: wind, temperature, sunshine duration, precipitation, barometric pressure, humidity and evaporation. The data of the veterinarian consists of 144 thousand lines, each containing a explanatory text or medication, the date and the veterinarian in charge. For this research these lines are merged into consults, where a consult is specified as the lines of the same horse on the same day. This results into 60 thousand consults over the 21 years of data.
11. To work with the data, the consults all need a label with the diagnosis. At first a keyword search is used to label the consults. This approach did give multiple problems. A keyword could be a false positive when the word is used in a sentence like “this is not a case of …” or it could give a false negative when no text is given, typos or spelling mistakes are made or when there are terms used that are not considered as key words. Also the count of the number of consults containing a keyword is very low, this is also confirmed by the expert. To fix these problems, 10.000 consults are labelled by hand. Using a binary labelling approach, where for each of the diseases a 1 or 0 is given for each of the diseases, if the horse is, or is not diagnosed with the disease during the consult. With these labelled consults a stochastic gradient descent approach, with different classifiers and loss functions is used to label the rest of the consults. Using machine learning to label the consults allows some of the consults without written text to also be labelled since the medication also can give an indication of the disease. In the 10.000 hand labelled consults roughly the same count for each of the diseases was found as the count of the keyword search over the whole data set. The hand labelled data is used to calculate the accuracy of the keyword search and the automatically labelled data. The accuracy of the keyword search is probably less than calculated since the hand labelled data is mainly the data that contains text and therefore was already quite accurately labelled with the keyword search.
12. To get more insight into the data, and the influence of the weather on the diseases, visualization are made. These first four visualization show the count of the consults concerning the disease over the years for each month, to see the seasonal influence on the diseases.

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.

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 kept 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. The higher number of cases of laminitis during January, compared to December and February can be explained by the low temperatures

Most of the consults that concern respiratory disease are seen in spring other literature found winters as even more risk full. The expert explains the high numbers in the spring and lower numbers in the autumn by the fact that most foals experience respiratory problems in their first months

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. Changes in the weather are mostly considered to cause diseases in horses. If this is the case cannot been concluded from these figures.

1. For the following visualizations, the weather data is split into 2 groups. This is done for each of the diseases. This table is an example on how these groups are created. Group A consists of the days where colic is diagnosed and group B are the remaining days. When we consider the days before, denoted with dt-1, the colour of the rows will be transposed to one row above for each row to create the new groups, for dt-2 the colour will be transposed two rows above etc.
2. To visualize the influence of changes in the weather on the different diseases, plots are made for each of the diseases and the 5 most mentioned weather variables: daily mean temperature, daily mean wind speed, daily mean barometric pressure, precipitation duration and daily mean relative atmospheric humidity. In these visualizations, the group A’s of the current day, so the dt-0 is plotted against the other time options. The same holds for group B. When a disease would correlate strongly to a change in one of these weather variables, one would expect many red dots on the upper left or lower right corner and less blue ones. As we can see the red and blue dots mostly follow the same patterns.
3. To make this more clear I have made two of these visualizations a bit bigger. And as you can see there are a lot more dots, That is because I only plotted 100 dots per class in the previous image since the two classes do override each other. To show that both classes indeed override each other histograms are added at the sides of the plot. The histograms show roughly the same and therefore we can conclude that both classes indeed have the same shape. I only showed you colic but the rest of the images are very similar to the ones plotted in these images. From this we can conclude that there are no obvious correlations between changes in the weather and the diseases
4. To find further correlations, permutation tests are used for significance testing. As with the visualizations, the weather variables are split into groups A and B. With a permutation test, the difference in mean on the groups is calculated. Then the groups are shuffled randomly, the difference in mean is calculated again, and this shuffling is done 1000 times for each of the weather variables and for each of the diseases. Using all those differences in mean, the permutation test returns the probability that the difference in mean between group A and B is created by chance, the p-value.
5. This results in four tables, from which one is shown here. The gray cells indicate a p-value smaller than 0.01, which means that there is only 1% chance that the difference in mean in by chance and it is therefore 99% sure that the difference in mean is a actual difference. The green cells indicate a positive correlation and the red cells indicate a negative correlation. So for example, TX has a green cell for dt-4, meaning that mean of group A is 4 bigger than the mean of group B.
6. I know that you can cannot read this, but this are the four tables, for the four diseases. What stands out here is that, when a correlation is found for multiple diseases, there is always a positive correlation or a negative correlation for all of them, for example wind is always negatively correlated. and that there are a lot of correlations found. I would have expected to see more specific weather variables for the different diseases and not “a bit of all”.
7. Lastly I wanted to see to what extend it is possible to predict the diseases using the weather. According to the literature ensemble predictions work well when predicting with weather. Therefore I tested the 3 ensemble predictions: bagging, boosting and voting and compared the results to a single classifier. Ensemble predictions work by combining one or more classifiers. Therefore I tested linear regression, support vector machine, decision tree and neural network since these are also found to give good results when working with weather data. Per disease, the single classifier and bagging algorithm are performed four times, once for each of the classifiers. The boosting and voting algorithm are performed just once, the boosting algorithm only with decision trees and the voting algorithm combines the four classifiers.
8. Here is an overview of the accuracies of the classifiers. For the single classifiers and the bagging algorithm, only the best accuracy is shown, and within the brackets the classifier or classifiers are given. The bagging algorithm, using decision trees is most popular since it works best or each of the diseases except for laminitis.
9. Now we have all results to answer the research questions
10. Just a short recap of the first question.
11. All three weather categories do have some correlation with colic but the wind has a negative correlation, while a positive correlation was expected. Also the literature suggests that a high barometric pressure is correlated while we also find a low pressure to be correlated.
12. The temperature is correlated with laminitis, but the correlation found is positive, which suggests that a higher temperature is correlated with laminitis. Drought is indeed correlated with laminitis.
13. For respiratory disease we indeed see a positive correlation between the temperature and the disease but also a lower humidity.
14. For skin disease we see that less rain and a lower average, higher maximum and lower minimum humidity is correlated to the disease
15. Even though a lot of correlations are found, in general the diseases do not seem to correlate to specific weather conditions.
16. The second research question was to what extend the diseases could be predicted using the weather. Respiratory disease is most accurate to predict with almost 80% accuracy. Laminitis is the hardest to predict with an accuracy of 65%. Since 50% accuracy is compete randomness I would say that the prediction of the diseases in general is not very good.
17. Thank you for listening to my presentation. Are there any questions?