MACHINE LEARNING MINI PROJECT

Kevin Jetten

318843

Inhoud

[Step 1 – Describe the data 3](#_Toc500161285)

[Step 2 – Explore the data 4](#_Toc500161286)

[Faulty records 4](#_Toc500161287)

[Findings 4](#_Toc500161288)

[Step 3 – Do the Analysis 7](#_Toc500161289)

# Step 1 – Describe the data

The name of the dataset is “Diabetes 130-US hospitals for years 1999-2008 Data Set”. It contains 101767 records of hospital admissions about people with diabetes. It contains 55 attributes but not all the attributes are available for all of the records. It is very clear that the domain of this dataset is medical.

I would like to use supervised machine learning to predict the amount of days a patient is likely to stay in the hospital based on the other attributes. I think the most useful attributes for this prediction are: age, weight, admission type and admission source but I am certain there will be other attributes that are very important in predicting this which I will find out about by making all kinds of different plots.

It’s not necessary to clean the dataset, however because there are records that have missing attributes I have to determine the importance of these attributes after I made the plots. Then I can decide what to do with the missing data. Because there are so many records the best choice might be to remove the records from which the crucial attributes are missing.

The dataset can be found at the following link:

<https://archive.ics.uci.edu/ml/machine-learning-databases/00296/>

# Step 2 – Explore the data

## Faulty records

While exploring the data with different kinds of plots I found some faulty data. There are about 3 records with their weight set on ‘20’ while every other person which has their weight filled in has it put in a category like ‘25-50’. I suspect that these people should be in ‘200+’ but I can’t know for sure. Because there are only so few of these I decided to remove them from the dataset.

There were also a few records in which the gender was unknown, I will again remove these because they are not useful in any way and would only confuse the machine learning algorithm.

## Findings

While starting the project I was suspecting that the weight of a patient would have a great impact on the amount of days he would spent in the hospital but this seems untrue.

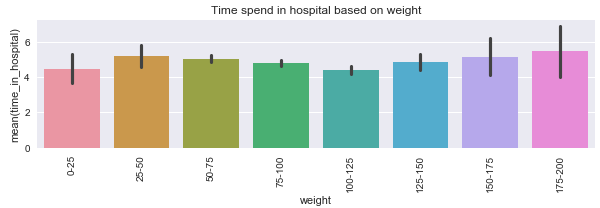


Image 1: Hospital stay duration and weight

Most notable are the last two weight groups in which the records jump out the most. Even though there seems to be some sort of correlation, I don’t want to make the weight field an requirement and remove the records that have it empty. Reason being that there are only 3000 entries where weight is filled in while there are otherwise 100k and it doesn’t seem to be overwhelmingly important.

Because there are so many different plots that I’ve made I will only talk about the ones that stand out.

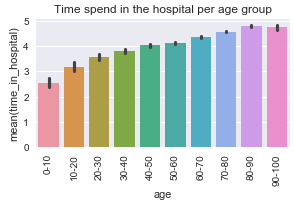


Image 2: Hospital stay duration and age

The difference between the first and the second group is the biggest but the others are still consistently climbing up over time. You could roughly draw a line that goes upward connecting all of the bars.

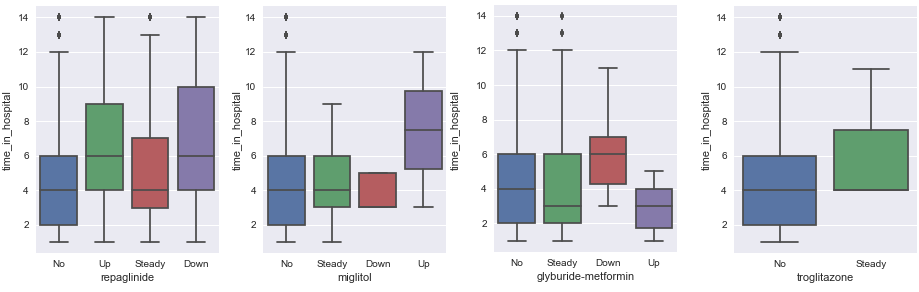


Image 3: Hospital stay duration and various medicines

I’ve made plots for all the medicines (22 different kinds) and these are the most notable. It is hard for me to say something meaningful about this because I have no clue what these medicines do, but there seem to be a correlation with the stay period.

There might also be a correlation with the stay period and a certain combination of medicines but I could not think of a proper way to plot this and find out. However, machine learning should be able to catch these correlations if they exist.

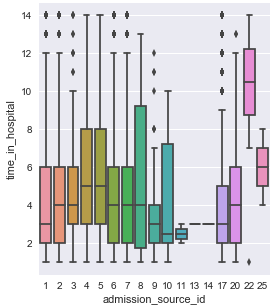


Image 4: Hospital stay duration and admission sources

A few of the admission sources stand out compared to the rest. Let’s take id 22 as example, the description states that this group is transferred from the hospital in which the patient already had stayed overnight. It’s quite a safe bet that they are going to stay even more days in the hospital in this case.

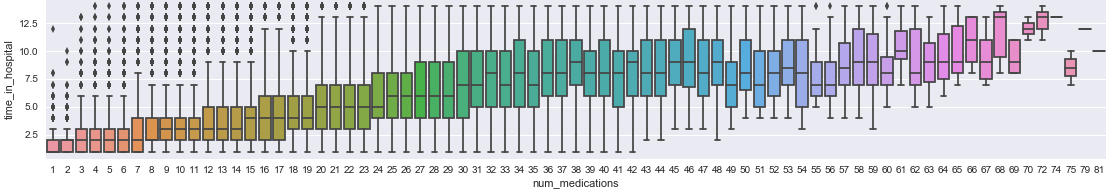


Image 5: Hospital stay duration and number of medications

The last graph I want to talk about this also the largest one, the amount of different medicines a patient uses. The correlation between these two fields is very obvious and you don’t have to be a doctor to get a feeling on why this can be the case.

I would like to draw some initial conclusions from the graphs that I talked about in this chapter. First of all that the older you are the longer your stay will probably be. Also some medication will impact your stay duration more than others and some medicine won’t have an impact at all. The number of medications you are on will have a very big impact, the more medications you are on the longer you will stay in the hospital.

# Step 3 – Do the Analysis

I would first like to think about what in theory would be the best algorithm. There are some key questions I can answer about the data and what I want to predict now which can help me choose what algorithm to use.

* I have more than 100k samples but not by that many
* A lot of features can be important
* It’s textual data
* It’s labeled but it’s not totally clear which features are important

The top algorithms I have to look at based on these points are:

1. Decision Tree
2. Mini batch K-means
3. SGD Classification