MACHINE LEARNING MINI PROJECT

Kevin Jetten

318843

Inhoud

[Step 1 – Describe the data 3](#_Toc497915332)

# 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/>