Machine Learning Project

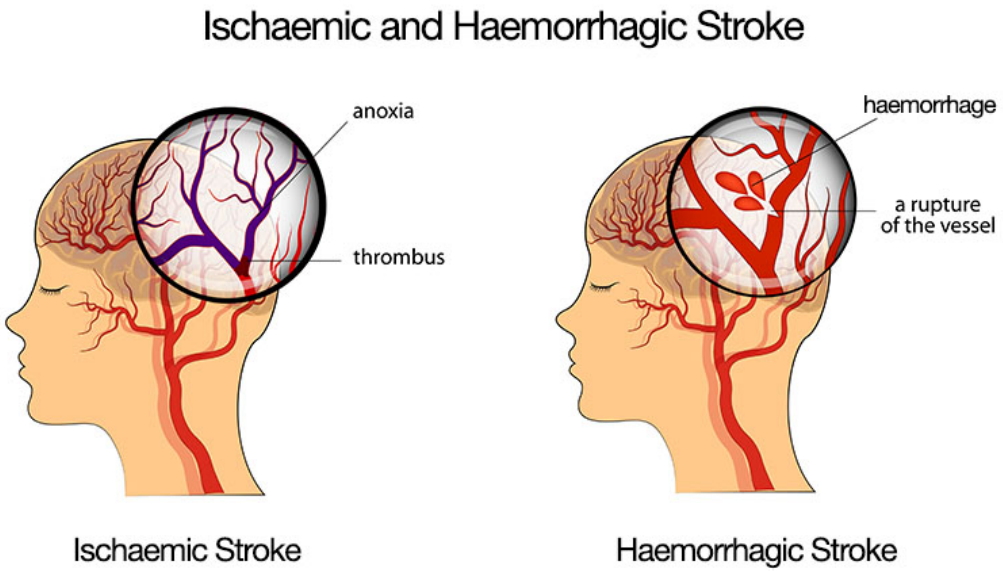
Stroke outcome prediction

Nazarbayev University – Dept. of Computer Science

Project Description

- The task of this project is a binary classification task. In particular, we are interested to find out, whether the patient survived the stroke or whether he/she did not.
- First you will need to spend some time on data cleaning/feature generation/normalization/etc
- As mentioned, the data contains a binary classification task on the outcome of the stroke
- In the second phase, you will be able to use your preferred algoirithm to classify the data.

Background information on stroke



- Two main types of stroke: ischaemic and haemorrhagic
- Stroke is the third leading cause of death in the United States. More than 140,000 people die each year from stroke in the United States.
- Each year, approximately 795,000 people suffer a stroke.
- Stroke is the leading cause of serious, long-term disability in the United States.

About the data

- We have anonymized data from the UMC at NU. This is real data from real patients.
- This data is for you and ONLY FOR YOU. Please do not share this data.
- This is a project in collaboration with Prof. Dmirty Videmann. He is MD & Asst. Prof. at NUSOM.
- You can be part of this collaboration.

The actual data

- have data of 150 subjects, who were all administered to the hospital with stroke
- The task of this project is to predict, wheter the patient will survive or not (column 'M' of 'basline info', i.e. a binary classification task.

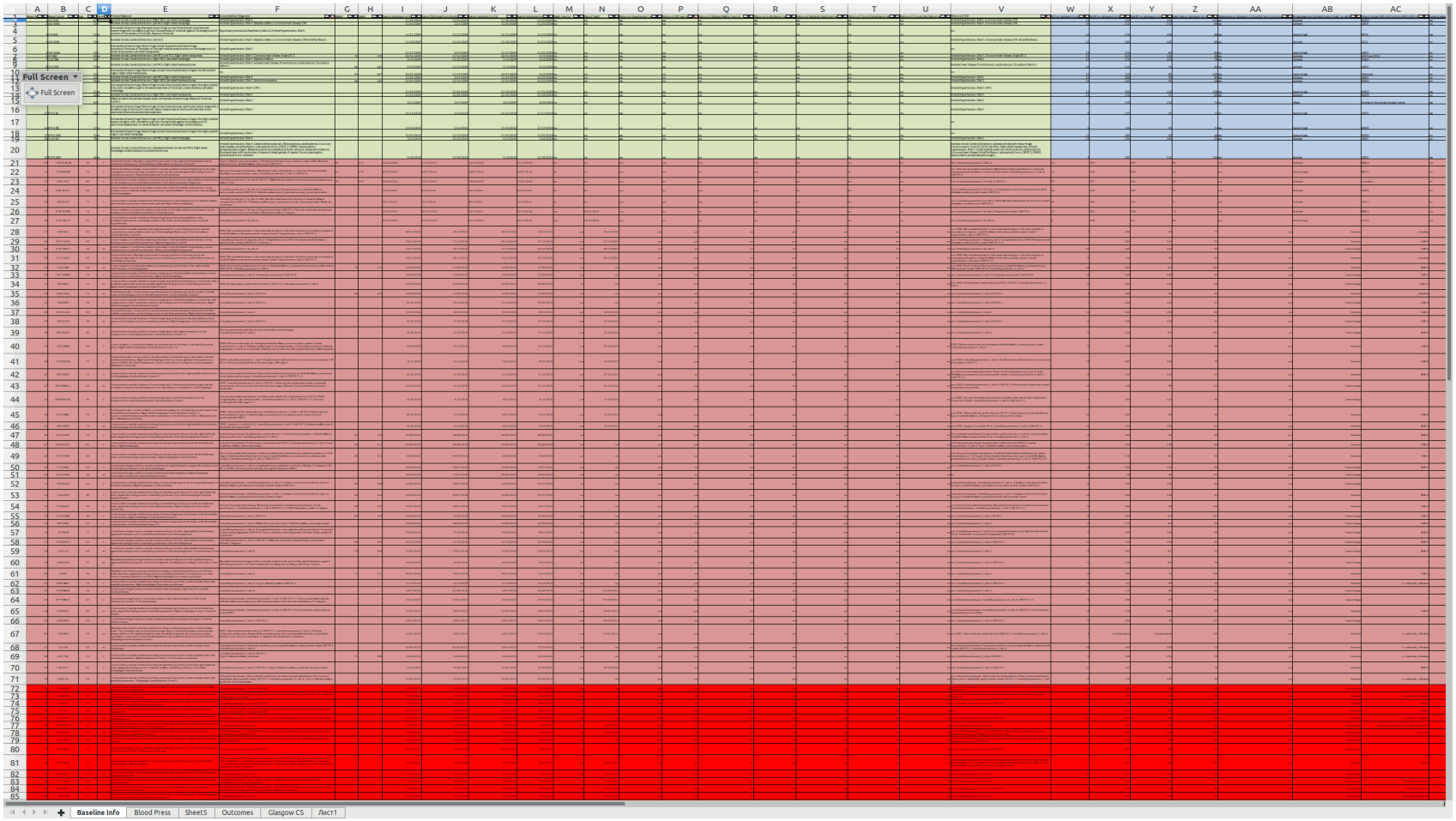
In other words:

1. What are the major predictors of negative outcomes in stroke patients?
2. How well can we predict negative outcomes in stroke patients with classification techniques?
3. When does the algorithm fail and why? Can we find reasons when we look at wrongly classified subjects?

The actual data

- There is a lot of data regarding the individual patients
- The data is organized in an excel sheet.. Unfortunately the data is not perfectly organized:
  1. some features are missing for of the some subjects
  2. sometimes data is numeric, sometimes in textform, sometimes discrete, sometimes continuous
  3. some data is static (age, weight, etc), while other data may be dynamic (day 1, day 2, etc.)
- Also, please keep in mind that the data may be imbalanced.

The actual data



- The data is available in the corresponding drive folder, please check and download.

Tasks in the project

- Examine the data and try to understand the format / what the features mean. We can probably schedule a meeting with Prof. Viderman, if we have gathered a list of questions regarding the data.
- Create a plan of which features you would like to include for the analysis and why.
- Transform the data into an appropriate format, i.e. matrix with $D \times N$ or similar
- perform cross-validation of the data with various methods and find out how well they work **and try to get idea why**

Tasks in the project

In this project the focus is on multiple things:

1. make sense of the data (probably lots of medical terminology that you need to look up and digest)
2. think of how to create features, i.e. transform text data to numerical, z-transforming features, etc
3. think about how to deal with missing data
4. application a number of ML classification techniques
5. interpretation the results

Good Luck!