Project 18 - Outline

Chau Zhong Belkis, Damergi Yassine, Gusinac Alem, Kappelmeyer Maurice, Tudelska Karolina Zofia

2022-04-06

Data Link

https://www.kaggle.com/datasets/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer

https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003489

Dataset Background

This dataset is about pancreatic cancer. It contains information about levels of four biomarkers (creatinine, LYVE1, REG1B, and TFF1) found in urine from individuals with the following diagnoses:

- Healthy controls
- Patients with non-cancerous pancreatic conditions, like chronic pancreatitis
- Patients with pancreatic ductal adenocarcinoma

In total 590 samples are included. They are described by following values: ID, cohort, sample origin, age, sex, stage of cancer, non-cancer diagnosis.

Aim of the Project

The original study was about early-diagnosis of pancreatic cancer based on the mentioned biomarkers. In this project, we would like to predict the pancreatic diagnosis of a patient, pancreas cancer, non-cancerous pancreas or healthy pancreas, or even a specific stage of cancer. It could be possible to predict the diagnosis outcome by the level of biomarkers.

Project Outline

1. Tidying and cleaning the dataset, including treatment of missing values and stratification into subgroups.
2. Characterization of the subject cohort by e.g. histograms for their age, sex and sample origin.
3. Clustering of samples and evaluate if it matches their diagnosis.
4. Modelling – creating a linear model attempting to predict clinical outcome from biomarker levels