Thema09_Report

Marcel Setz

13-11-2022

Contents

Introduction

Pancreatic ductal adenocarcinoma (PDAC) is one of the deadiest cancers. The chances of survival are increased when diagnosed in an early stage. However, PDAC shows symptoms when it already spread throughout the body. Most of the time, it's too late by then. There may be a way to detect PCAD in an early stage with a simple urine test, with the use of the following biomarkers: creatinine (Urinary biomarker of kidney function) LYVE1 (Urinary levels of Lymphatic vessel endothelial hyaluronan receptor 1, a protein that may play a role in tumor metastasis), REG1A and REG1B (Urinary levels of a protein that may be associated with pancreas regeneration.), and TFF1 (Urinary levels of Trefoil Factor 1, which may be related to regeneration and repair of the urinary tract)

the attributes in the data interesting for this research are the biomarker values mentioned before. There is also an attribute called diagnosis, in which the diagnosis of the sample is stated, where 1 means no PDAC, 2 means benign hepatobiliary disease (non cancerous, non harmful pancreatic condition), and 3 means that the sample has PDAC.

Material & Methods

Materials

The data used was obtained from kaggle (https://www.kaggle.com/datasets/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer).

The project was developed using github: github for analysis: https://github.com/marcelsetz/Thema09_2223 github for Weka API: https://github.com/marcelsetz/Thema09_JavaWrapper

Table with software used:

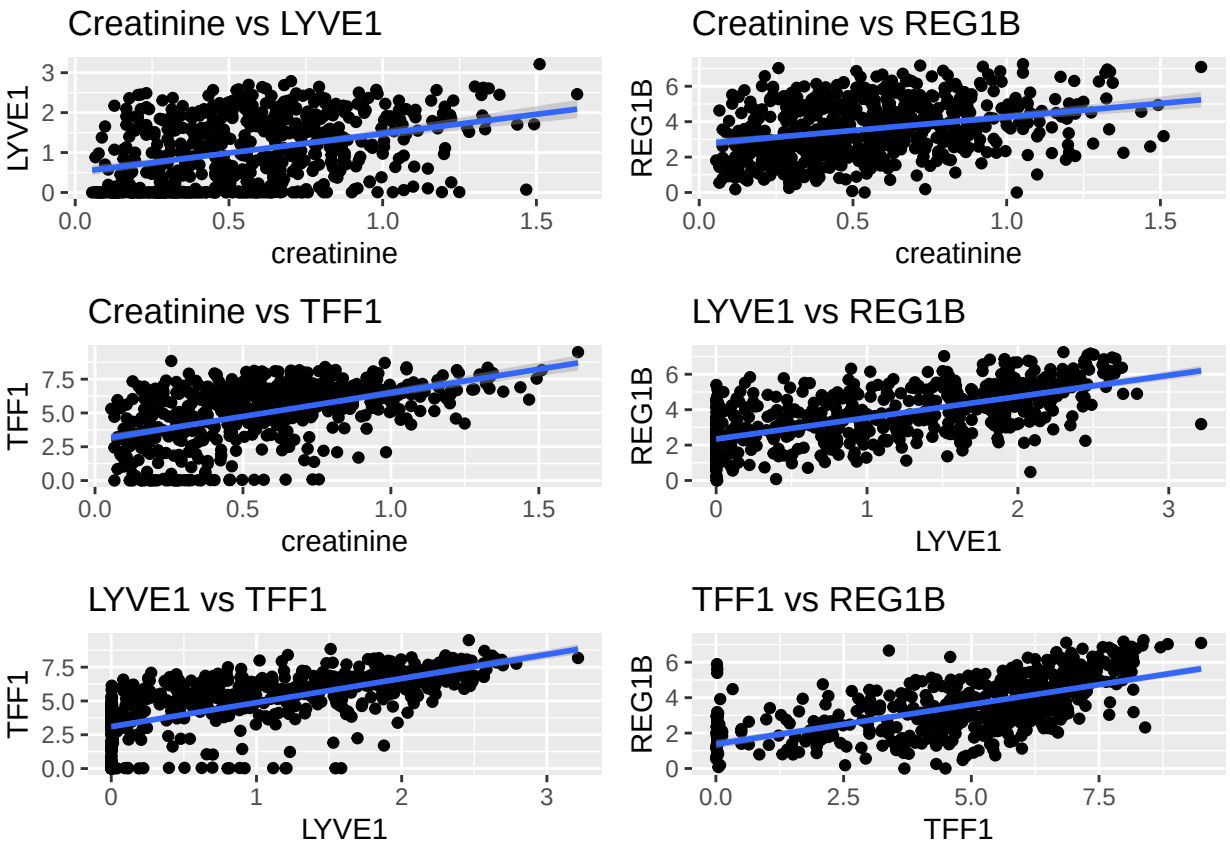
Software	Version
R	4.0.4
RStudio	4.0.3
RMarkdown	2.11
Java	17
JDK (Java)	11
SDK (Java)	17
Gradle	7.1
Weka	3.8.5

Methods

First the data was analyzed, cleaned and processed with the use of R in Rstudio. Please refer to the log to see the detailed information of this process.
Then using gradle and Java, a wrapper was created which wraps the whole classifying in a single program. To see the source code or to run the program, refer to the Weka API github seen above, the Readme.md file explains exactly how to use the program.

Results

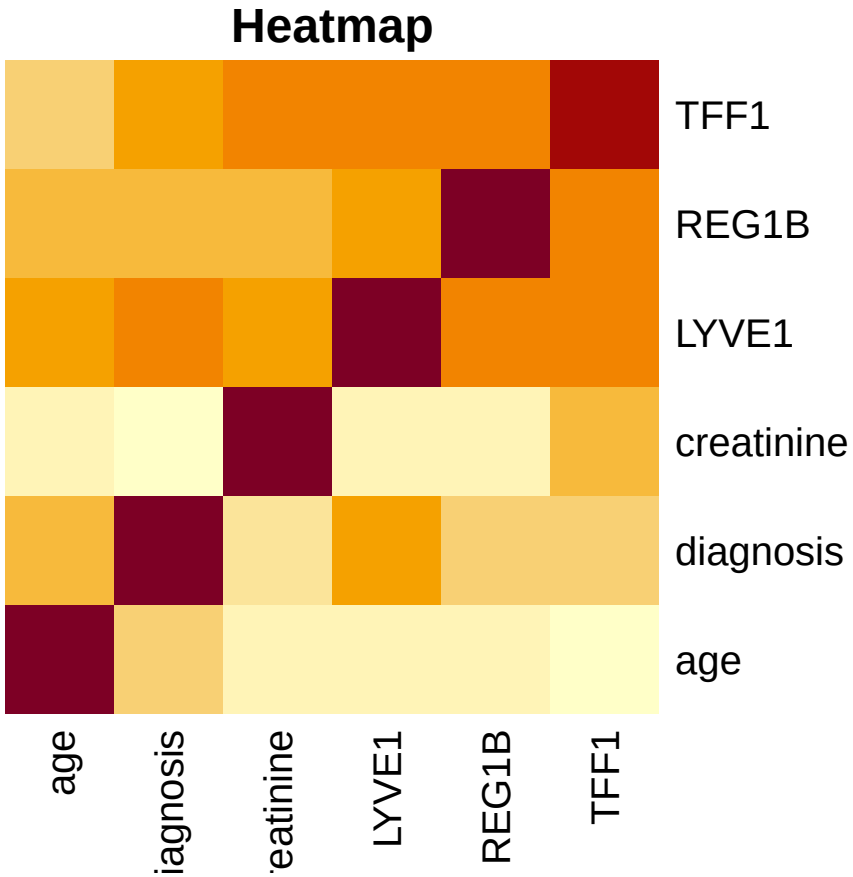
Here's a comparison of every biomarker to see if they correlate with each other in any way. REG1A was excluded because REG1B was an improvement of this bio marker so it wasn't deemed necessary.



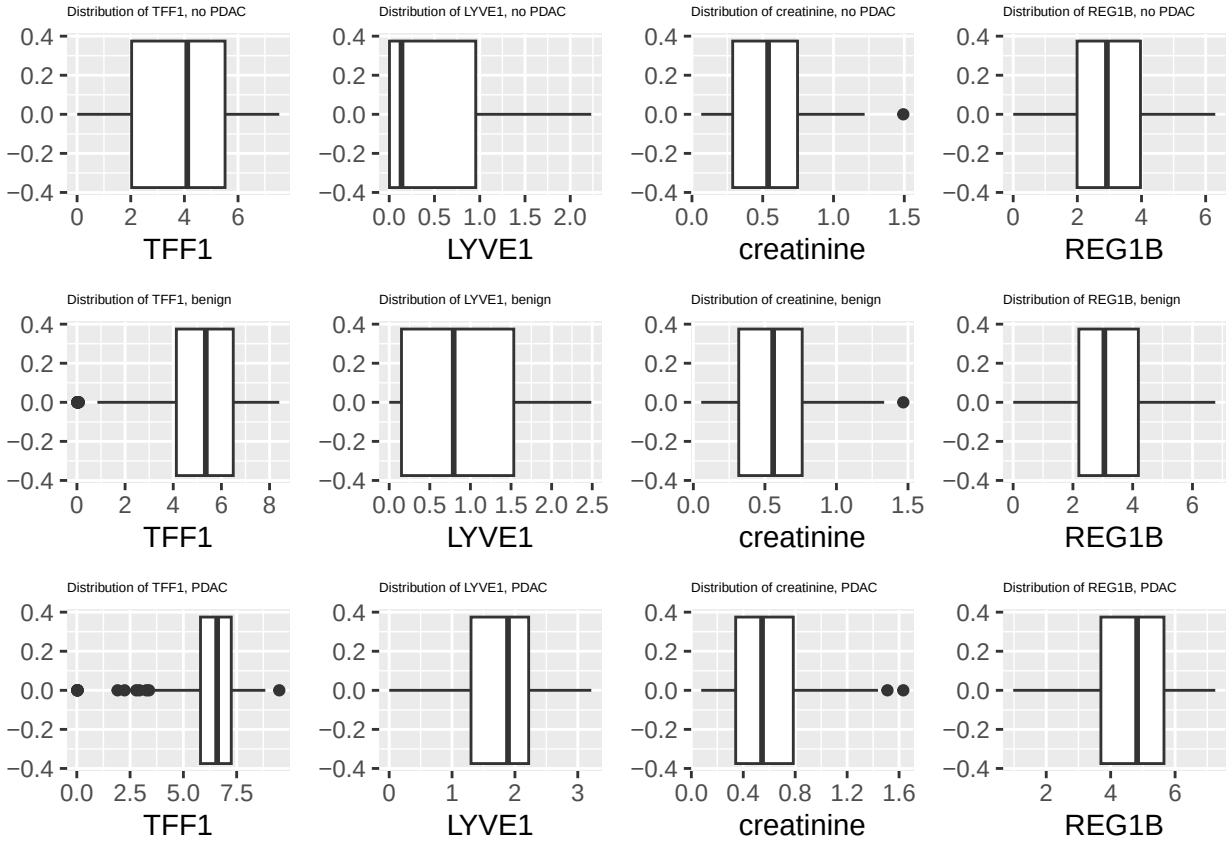
There seems to be a slight correlation in a few of these. This is important for the classification stage, because these biomarkers do tell something about the value of the others and also the possible diagnosis. Here's the heatmap to further analyze the correlation.

Table 2: correlation matrix

	age	diagnosis	creatinine	LYVE1	REG1B	TFF1
age	1	0.3083	-0.05705	0.3179	0.277	0.1707
diagnosis	0.3083	1	0.06435	0.5775	0.4355	0.4749
creatinine	-0.05705	0.06435	1	0.3559	0.3001	0.4856
LYVE1	0.3179	0.5775	0.3559	1	0.642	0.6743
REG1B	0.277	0.4355	0.3001	0.642	1	0.634
TFF1	0.1707	0.4749	0.4856	0.6743	0.634	1



It is visible here that the TFF1 biomarker correlates the most with every other biomarker. In the correlation matrix, it is possible to see that the highest correlation is between the biomarkers TFF1 and LYVE1. However in this matrix there are some outliers that may be important. It is in about 67% of the samples the case that the higher the value of TFF1, the higher the value of LYVE1.



The values in these biomarkers show quite the difference here, With the TFF1 as well as the LYVe1 biomarker, the higher the value, the more likely it is that this sample has PDAC.

Conclusion & Discussion

Using our research and data analysis it is safe to assume that some biomarkers may be important for the diagnosis of PDAC. Especially the biomarkers TFF1 and LYVE1 show a significant correlation with each other and it was visible that the higher those values, the higher the chance of PDAC is. The effect of non-cancerous pancreatic conditions and PDAC on TFF1 and LYVE1 is the best visible, the rest of the biomarkers don't show a significant effect on pancreatic conditions. Ofcourse this is extremely reliable, due to all the outliers. But if you take a urine sample from a patient and both the TFF1 and LYVE1 biomarkers are a high value, then maybe it's wise to further investigate if the patient has PDAC or not.

Project proposal

This research can be improved in multiple ways. These possibilities are written for the minor Application Design.
The main issue here is that the Java wrapper is not very user-friendly. The program can now only be runned via the command line with a few options. For a non-programmer, it can be really hard to understand.

A web application implementing this program within a user-friendly environment can drastically improve the way this program is used.

The target audience would be people working in health care with large amounts of biomarker data, that needs to be diagnosed. The program can process a lot of patients very quickly.