Pancreatic Ductal AdenoCarcinoma

Joshua Tolhuis

11/12/2021

Name: Joshua Tolhuis
Studentnumber: 390799
Study: Bio-informatica
Class: BFV3
Date: 05/10/2021
Teachers: Michiel Noback

Abstract

Introduction

In this paper, the main goal is to research the effect of urinary biomarker on patients with Pancreatic Ductal AdenoCarcinoma (PDAC). It is further used to predict the stages of cancer and using the least amount of variables possible. This is done, to make a patients live easier and testing on patients more efficient.

Material & Methods

This paper has made use of several programs such as, Weka, Rstudio and Java to further this research. The Java was used to make a tool which uses Weka to predict the stages of PDAC. The Rstudio was used to create a research paper and make an Exploratory Data Analysis (EDA).
Results
The result of this study were that it is possible to leave out the factors, age and the sex. The most important variables were the biomarkers used, the creatinine and the blood plasma.

Conclusion

The conclusion gotten is that creatinine, blood plasma and the biomarkers are very useful for prediction for prediction of cancer stages. It is also very possible to predict the stages to an accuracy of 75 percent.

Table of Contents

Contents

Abbreviations

- Pancreatic Ductal AdenoCarcinoma (PDAC)
- Exploratory Data Analysis (EDA)
- lymphatic vessel endothelial hyaluronan receptor 1 (LYVE1)
- Regenerating family member 1 beta (REG1B)
- Regenerating family member 1 alpha(REG1A)
- Trefoil factor 1(TFF1)
- Receiver Operating Characteristic (ROC)

Introduction

This report is about visualization of the data gained by the "Urinary biomarkers for pancreatic cancer" data set [1]. The paper referenced on this website uses biomarkers in order to check if somebody has pancreatic cancer. The main reason for this paper was to test if their newly improved biomarker (REG1B), was better than their other older biomarker (REG1A). Which they concluded to be true.

In this report the data is visualized using the programming language R. The goal of this report is to discuss and conclude the results of the Exploratory Data Analysis (EDA).

In the EDA the main thesis was, "What's the minimal amount of data combinations to predict a patient has pancreatic ductal adenocarcinoma?". In order to answer this, the first step was to visualize the data using R, and later on WEKA will be used in order to try and find the least amount of data values.

Material & Methods

Materials

The data obtained was gainen from kaggle[1] The project was maintained via github:
github for report, report
github for Weka Api, Weka Api

For this project mainly programming software and libraries were used.
Also Weka was used in order to obtain an ROC curve.
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

And here's a library table:

Library	Version
ggplot2 (R)	3.3.5
tidyr (R)	1.1.3
gridExtra (R)	2.3
dplyr (R)	1.0.7
plotly (R)	4.9.4.1
cluster (R)	2.1.1
ggfortify (R)	0.4.3

Methods

This software was used as followed: The R, Rstudio and RMarkdown, were used to visualize the data. Create several plots correlating different variables and looking at those results. It was also used to create a report. Java, Gradle and Java's SDK and JDK were used to create an tool which uses a Weka algorithm to predict cancer stages from a specific file. Java also used R to correct the file inserted into via the command line.

Project Proposal

In This project I also had to decide which minor to follow in my upcoming semester. My choice for minor was “Application design”.
My choice for this minor was a difficult choice considering I liked both minors and found both very interesting.
The main reason I chose Application design in the end was due to the projects concerning the 2 minors.
I found myself enjoying the web design project a lot more than the data mining project.
I also did my research on both minors and decided that even if data mining & machine learning is a very popular topic, it doesn’t tackle how to handle clients like Application design does.
This in my opinion is a very important skill as communication between scientists, marketing and bio bioinformatics is key.
The main reason this is key is to have knowledge on if certain aspects are possible to code as to not over or under sell my job in the future.
This was confirmed by one of the senior students I spoke to when they presented their project on the minor.
This is the main reason why I want to follow the Application design minor.

Results

The EDA gave some useful insights regarding the obtained data set. The first step was to see all the data by making a small table containing the data to test if it all loaded in properly and if any data was missing, here's a small preview of this table.

plasma_CA19_9	creatinine	LYVE1	REG1B	TFF1	REG1A
11.7	1.83222	0.8932192	52.94884	654.2822	1262.000
NA	0.97266	2.0375850	94.46703	209.4882	228.407
7.0	0.78039	0.1455889	102.36600	461.1410	NA
8.0	0.70122	0.0028049	60.57900	142.9500	NA
9.0	0.21489	0.0008596	65.54000	41.0880	NA
NA	0.84825	0.0033930	62.12600	59.7930	NA

As can be seen in this small table, A few problems arise, the stage and benign_sample_diagnosis. appear to be empty, but after closer inspection of the data it just means that these people aren't yet in a stage and also do not have a benign sample because they are the control group and don't have Pancreatic Ductal AdenoCarcinoma(PDAC). After checking the code book, All the data was accounted for and no cleaning needed to be done.

The first thing to look at was just a general view of the bio marker and how it corresponded to different stages of PDAC. The bio marker chosen was REG1B.

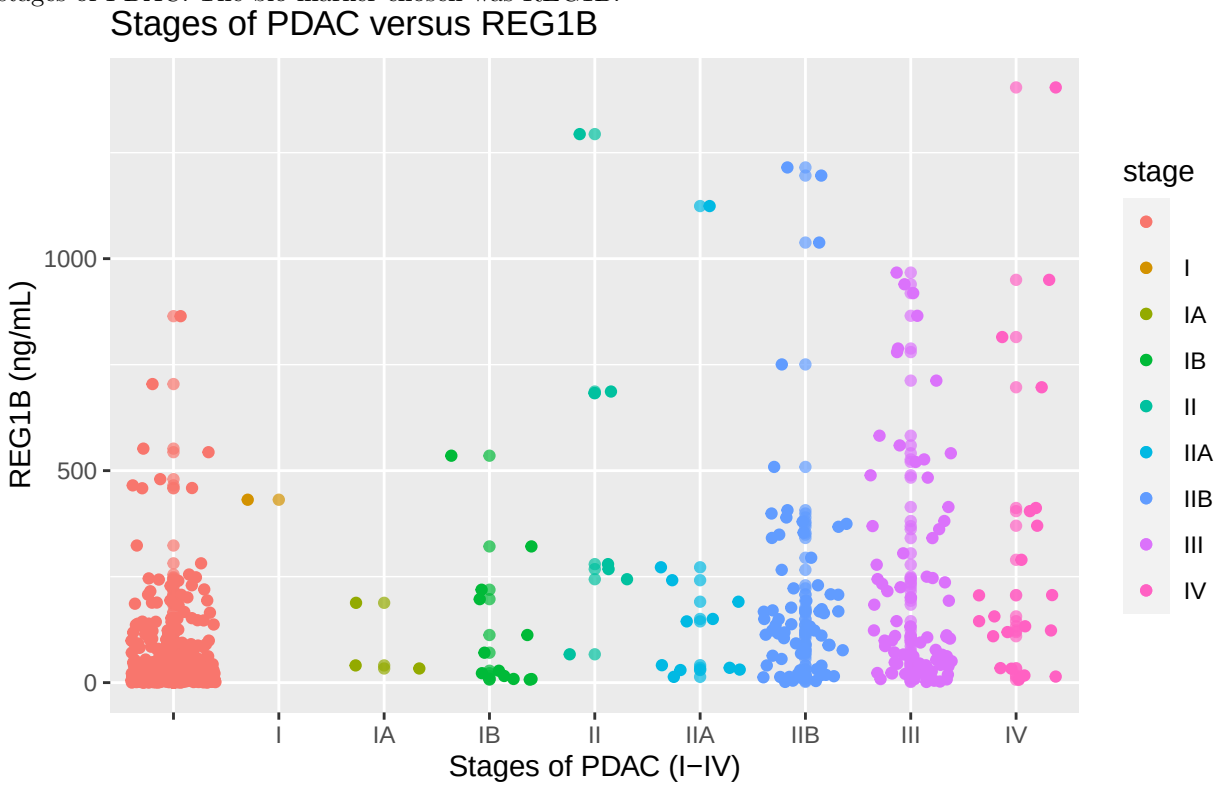


figure 1, in this figure cancer stage is plotted against REG1B value in a patients urine. Each color shows a different stage, and the red points are the 'null' stage or control group.

Figure 2 is an boxplot containing all 3 bio markers. REG1A was excluded because REG1B was an improvement of this bio marker so it wasn't deemed necessary.

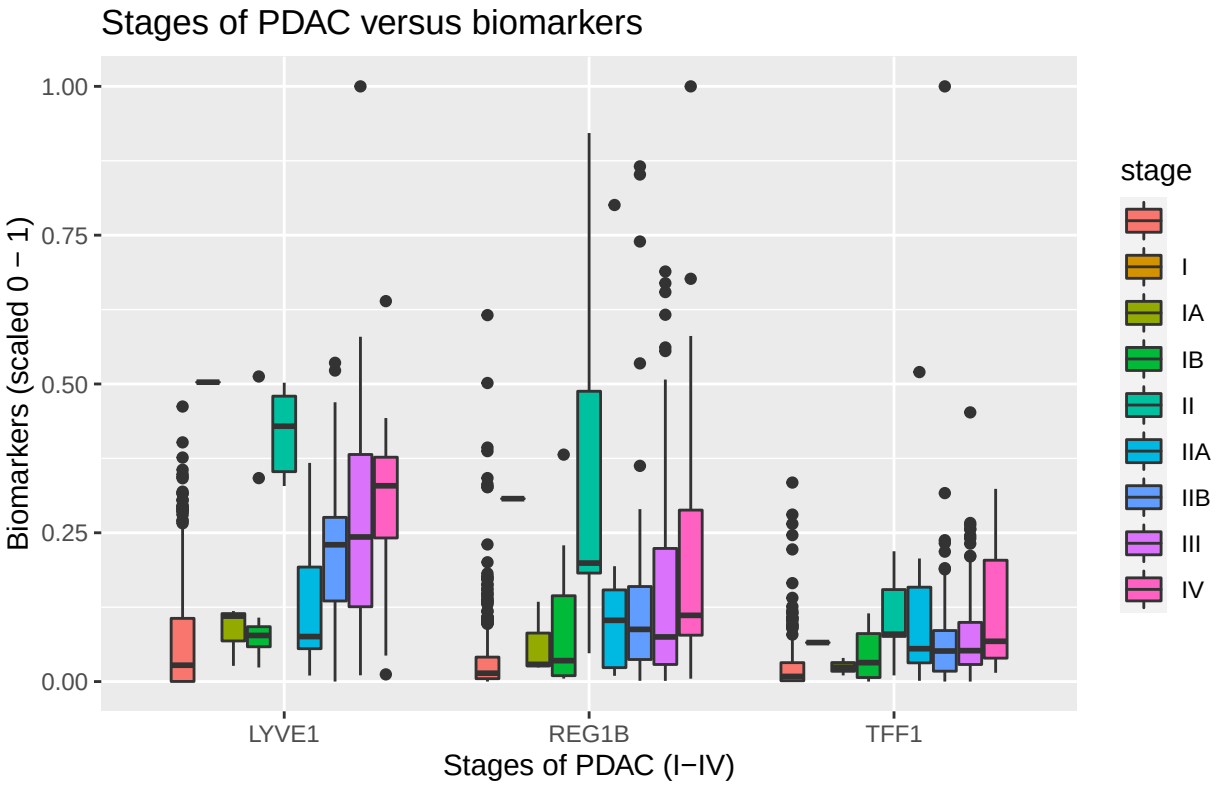
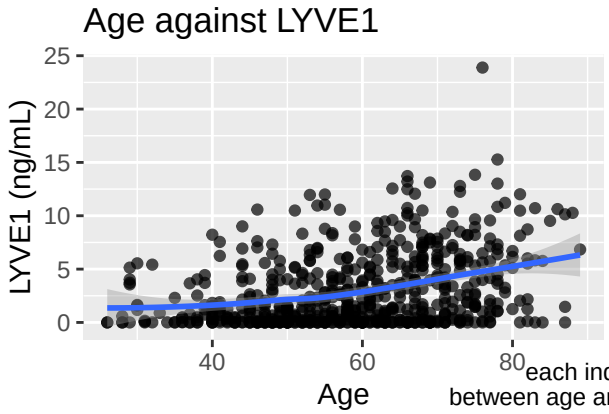


Figure 2, this figure shows that there is also a trend line in the other 2 bio markers.
s are represented in their respective color and the red color represents the 'null' stage or control group

Please note that the data of figure 2 was scaled using a simple min max scale, Figure 2 shows that there is indeed a trend in bio markers vs stage.

Next up was finding out if age affected any of these bio markers, this was done by creating a simple scatter plot and a trend line of the three bio markers.



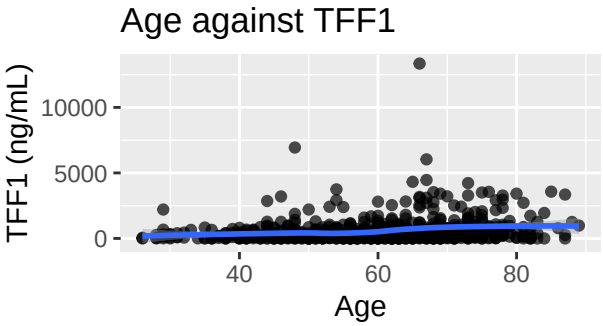
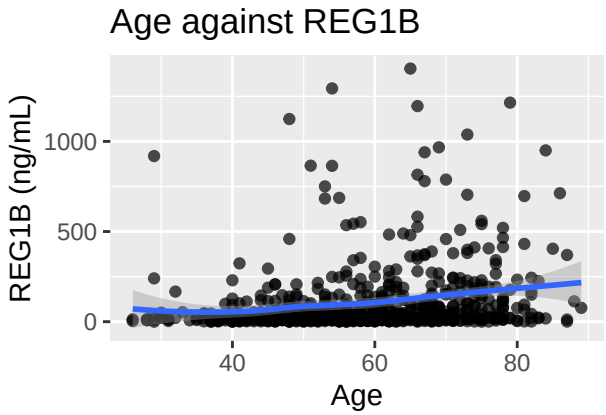


figure 3, this is an representation of age against each indivudal biomarker (except REG1A).In this plot a correlation between age and biomarker LYVE1 in contrast to the other 2 bio markers.



Interestingly it appears all three have an upward trend regarding the age of the patient tested. This might affect the algorithm positively, but it may also mean that the biomarkers have less of an impact than thought.

In order to further study correlations between numeric variables A heatmap was created to look at possible effect of the bio markers.

```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(selection)` instead of `selection` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```
## # A tibble: 7 x 8
##   var1             age plasma_CA19_9 creatinine LYVE1  TFF1 REG1B  REG1A
##   <chr>          <dbl>         <dbl>      <dbl> <dbl> <dbl> <dbl>  <dbl>
## 1 age            1             0.139     -0.119  0.393 0.302 0.241 0.104
## 2 plasma_CA19_9  0.139         1         -0.0952 0.198 0.148 0.182 0.123
## 3 creatinine    -0.119        -0.0952     1      0.308 0.299 0.159 0.0689
## 4 LYVE1          0.393         0.198      0.308  1     0.611 0.506 0.214
## 5 TFF1           0.302         0.148      0.299  0.611 1     0.627 0.340
## 6 REG1B          0.241         0.182      0.159  0.506 0.627 1     0.484
## 7 REG1A          0.104         0.123      0.0689 0.214 0.340 0.484 1
```

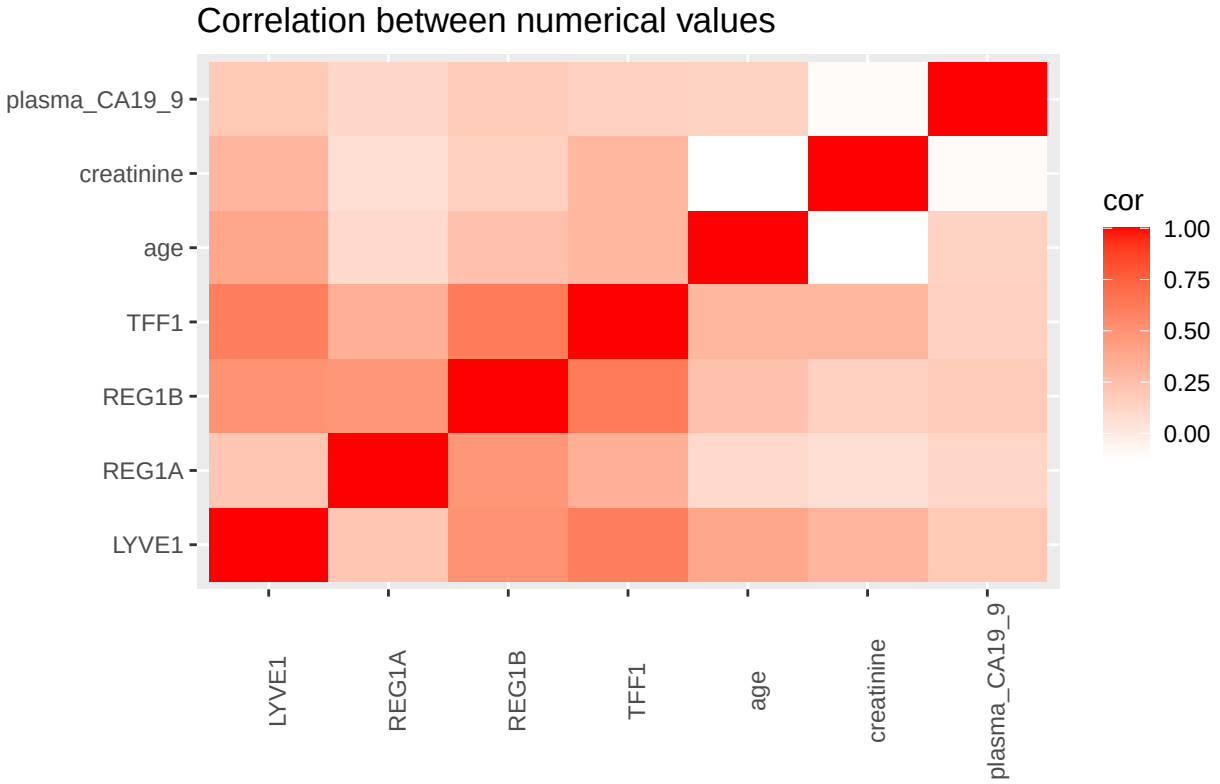


figure 4, this figure shows an heatmap of the numeric data in the data set.
In this figure a clear correlation between severel variables can be seen

This plot showcased another correlation between LYVE1 and creatinine, this can be useful in order to further the algorithm.

Weka results

A result of the Logistic function used to create an ROC curve to look at the sensitivity and specificity of the algorithm

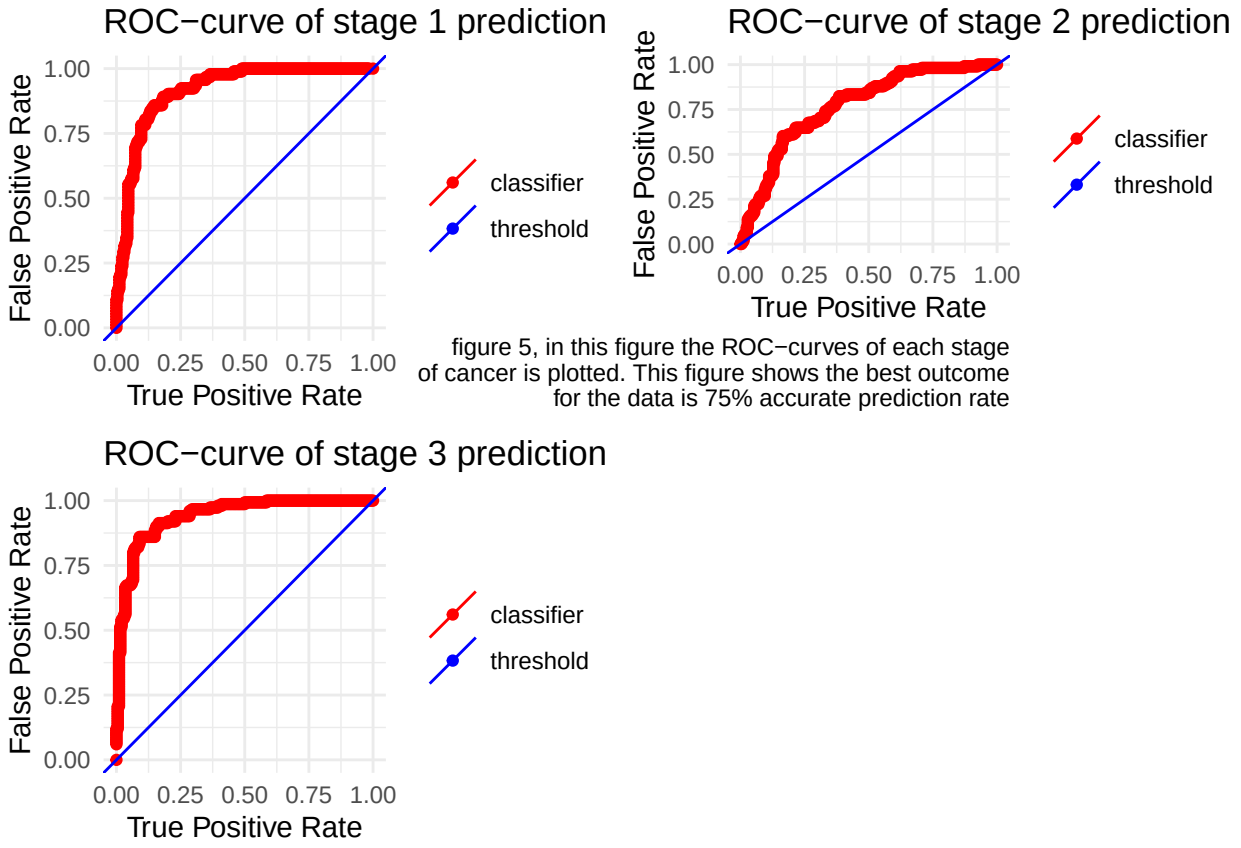


figure 5, in this figure the ROC−curves of each stage of cancer is plotted. This figure shows the best outcome for the data is 75% accurate prediction rate

In weka it was also found out that the maximum percentage to correctly guess the cancer stage was 75%.

Discussin & Conclusion

results

To answer the question, "What's the minimal amount of data combinations to predict a patient has pancreatic ductal adenocarcinoma?" The data succeeded in giving a better image of if this is possible. However It's still quite unclear if it's possible to decide if this data can be compromised and still gain good results if it is subjected to an algorithm. In the plots created, it shows clearly that there is a trend line in the stages and biomarkers. but as found in the later plots it shows that this might be due to the age. This makes the reliability of the first plots considerably lower, nonetheless it should still be possible to use all the data values combined to predict if a patient has PDAC.

discussion

Comparing the plots created against the official paper it shows that indeed there is a valuable correlation between the biomarkers and the stage. But what was surprising to see is that age also correlates strongly to one of the biomarkers as does creatinine. This does expose a few cracks in my thesis.

This research is weakened by the lack of further research into creatinine and blood plasma. These are valuable points and should have been created to get a better view if the future algorithm is helped or not by these values. This could easily be avoided by creating some small plots which show the correlation between the biomarkers and these bodily factors.

Perspective and conclusion:

The results of this report made a better insight if less factors could be used to predict if a patient has cancer or not, In the future however a careful eye must be casted over the bodily factors such as creatinine, blood plasma and age. If this research is successful it could mean that cancer can be predicted without need of a person's blood or personal data.

References

[1] https://www.kaggle.com/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer