Thema09_Report

Marcel Setz

13-11-2022

Contents

Introduction

Pancreatic ductal adenocarcinoma (PDAC) is a severe type of cancer with a poor prognosis. The current five-year survival rate is less than 10%. The main reason for this poor outcome is the lack of effective methods for early detection. In most cases, PDAC is not diagnosed until it has reached an advanced stage, at which point treatment options are limited and the chances of survival are greatly reduced.

The early detection of PDAC is crucial for improving patient outcomes, but current diagnostic methods are inadequate. Imaging techniques such as CT and MRI are not reliable for early detection and biomarkers for PDAC have been hard to identify. However, recent research suggests that urinary biomarkers may be useful for early detection of PDAC.

This report presents a machine learning approach for the early detection of PDAC using urinary biomarkers such as creatinine, LYVE1, REG1A, REG1B, and TFF1. A Java Wrapper is used to predict if a patient has PDAC, benign pancreatic conditions, or no PDAC. The data for this study was obtained from Kaggle and analyzed using RStudio Markdown. The goal of this report is to investigate the research question: "Is it possible to detect pancreatic cancer using values of the urinary biomarkers?" and demonstrate the potential usefulness of this machine learning approach for early detection of PDAC and to provide insight into the biomarkers that can be used for this purpose. By providing a way to detect PDAC early, this machine learning approach could help improve patient outcomes and increase the chances of survival.

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

The data for this study was obtained from Kaggle and analyzed using R in RStudio. To clean and process the data, an exploratory data analysis (EDA) was performed. A codebook was created to visualize the data, and the data was read and visualized using a table and a boxplot created with ggplot2 in R. Outliers identified in the boxplot were analyzed to determine if they could be removed or not. The correlation data was also obtained to investigate if the biomarkers correlated with each other in any way.

To determine the best machine learning algorithm, a quality metric that fit the data was selected. The false negatives must be as low as possible and therefore, the false negative rate (FNR) must be as low as possible for the machine learning steps. The algorithm was then evaluated using a 10-fold cross-validation for accurate results and a baseline accuracy was established using ZeroR.

The data was then processed in the Weka Experimenter using all machine learning algorithms. A table was created using the outputted data of all the quality metrics. Based on the results, a java wrapper was created, allowing the user to input a file with urinary biomarker values, and the program will predict the diagnosis.

Results

The correlation data obtained during the EDA process helped to investigate if the biomarkers correlated with each other in any way. By understanding the correlation between the biomarkers, it was possible to identify which biomarkers could be used as predictors of others.

By analyzing the correlation matrix, it was possible to identify if any of the biomarkers were highly correlated with others. For example, if creatinine and LYVE1 had a high positive correlation coefficient, it would indicate that when the creatinine levels were high, the LYVE1 levels were also high. This information can be used to identify which biomarkers could be used as predictors of others and which biomarkers are more informative.

Additionally, it also helped to identify if there is any multicollinearity in the data. Multicollinearity occurs when two or more independent variables in a regression analysis are highly correlated. This can lead to unstable and unreliable estimates of the regression coefficients and can also cause problems in interpreting

the results of the analysis. By identifying any multicollinearity, it was possible to address it and ensure that the machine learning algorithm was not affected by it.

Overall, the correlation data helped in identifying which biomarkers are more informative, which biomarkers could be used as predictors of others, and if there is any multicollinearity in the data. This information was used to select the best quality metric and machine learning algorithm.

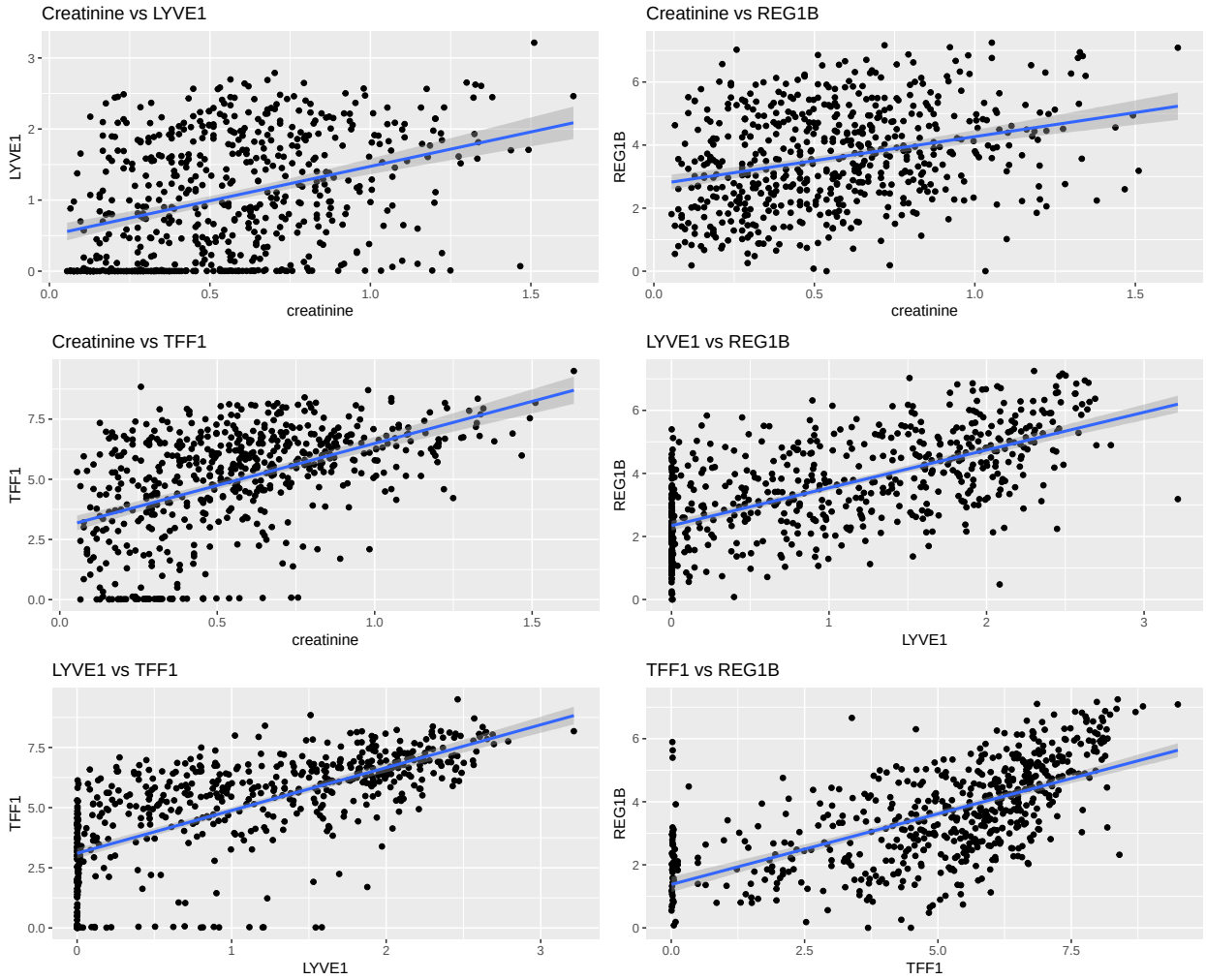
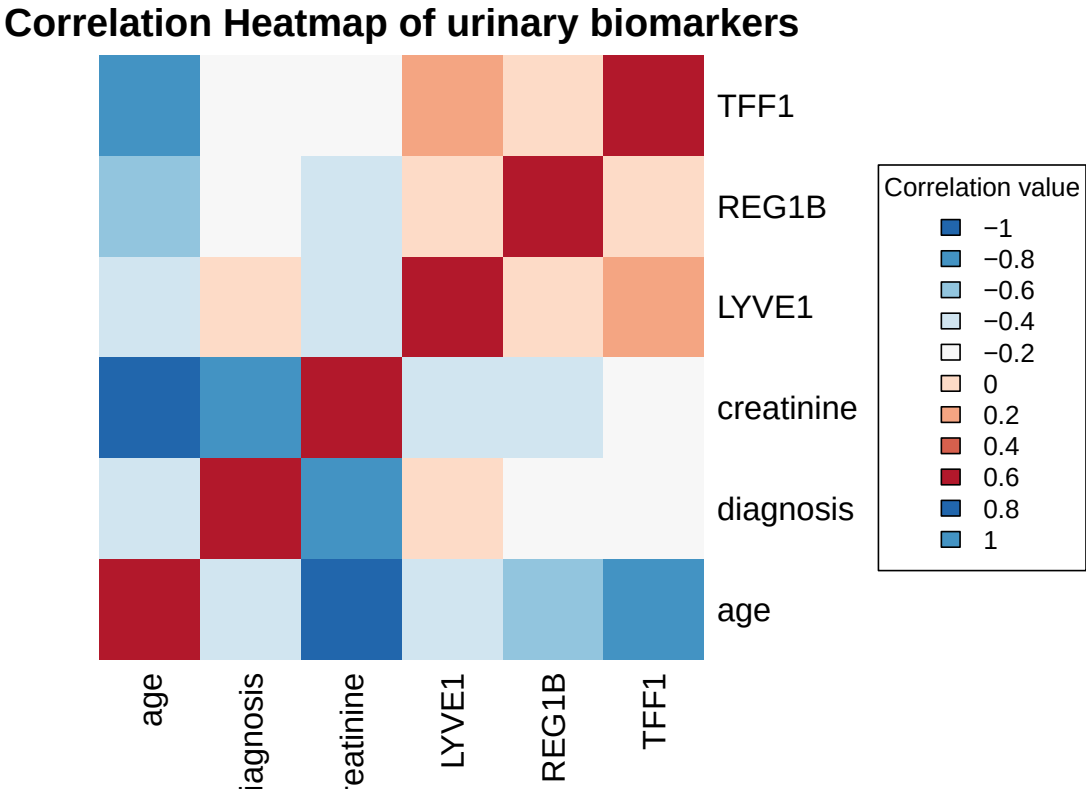


Figure 6: Scatter plot showing the correlation between creatinine, LYVE1, REG1A, REG1B, and TFF1 biomarkers in urinary samples of pancreatic cancer patients

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.



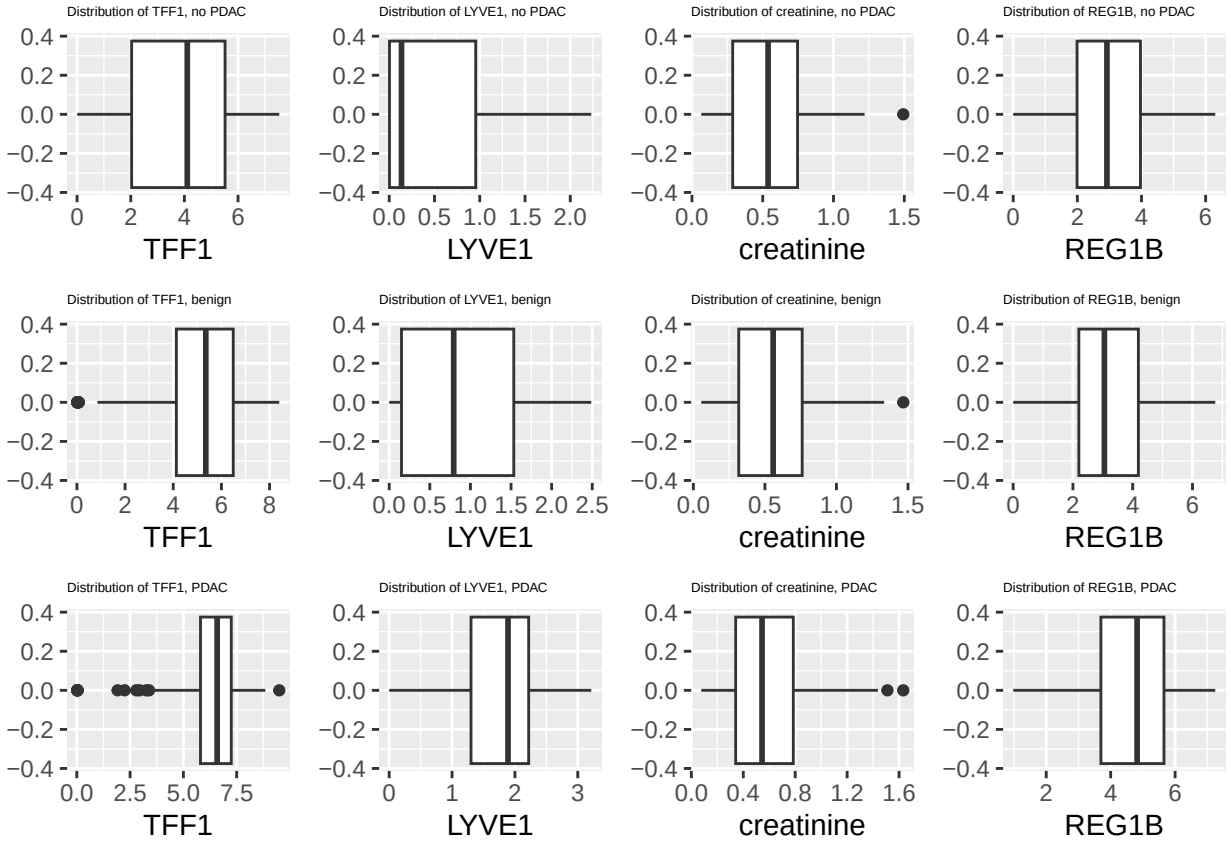
The heatmap created above represents the correlation between the creatinine, LYVE1, REG1A, REG1B, and TFF1 biomarkers in urinary samples of pancreatic cancer patients. The correlation values range from -1 to 1, with -1 indicating a strong negative correlation, 0 indicating no correlation, and 1 indicating a strong positive correlation.

The color ramp used in the heatmap ranges from blue for negative correlation to red for positive correlation. As you move from blue to red, the correlation becomes stronger. The darker the color, the stronger the correlation. The diagonal line of the heatmap is where the variable is correlated to itself, this is always 1, and this is represented by a dark red color.

To interpret the heatmap, you should look for patterns and trends in the colors. For example, if you notice that a particular biomarker has strong positive correlations with several other biomarkers, it might indicate that this biomarker is a good predictor of pancreatic cancer.

On the other hand, if you notice a biomarker that has mostly weak or negative correlations with the other biomarkers, it might indicate that this biomarker is not a good predictor of pancreatic cancer. You should also look for clusters of biomarkers that are positively or negatively correlated, which could indicate that these biomarkers work together in some way.

It is also worth noting that this heatmap is only one part of the overall analysis, and that correlation does not imply causality. Therefore, it is important to consider other factors such as the p-values of the correlation coefficients and the biological plausibility of the correlations before making any conclusions.



Creating the distribution of every biomarker with every diagnosis allows for a visual representation of the distribution of biomarker values for each diagnosis. This can be useful for identifying patterns or trends in the data that may not be immediately apparent from a simple summary of the data. For example, in this case, the distribution plots reveal that higher LYVE1 and TFF1 values are associated with a higher likelihood of PDAC. This information can be used to inform further research and the development of diagnostic tools for pancreatic cancer. Additionally, these distribution plots can be used to inform the selection of cut-off values for biomarkers in diagnostic tests, as well as to identify potential biomarkers for use in combination tests for improved diagnostic accuracy.

Conclusion & Discussion

The results of this study indicate that the urinary biomarkers creatinine, LYVE1, REG1A, REG1B, and TFF1 have the potential to be used for early detection of pancreatic ductal adenocarcinoma (PDAC). The analysis of the distribution of these biomarkers in relation to the diagnosis of the samples showed that higher values of LYVE1 and TFF1 were associated with a higher likelihood of a sample being diagnosed with PDAC. Additionally, machine learning algorithms were applied to the data to determine the best method for predicting the diagnosis of a sample based on its biomarker values. The SMO algorithm was found to generate the least amount of false negatives, making it the best option for use in a diagnostic tool.

The correlation and heatmap plots further support the potential utility of these biomarkers in detecting PDAC. The correlation between the biomarkers and the diagnosis was found to be strongest for LYVE1 and TFF1, providing additional evidence for their usefulness in early detection.

It is important to note that this study is based on a limited dataset and further research is needed to confirm the findings. Additionally, the use of these biomarkers for diagnostic purposes would need to be validated

in larger, clinical studies. Nonetheless, the results of this study provide a promising starting point for the development of a simple, non-invasive test for early detection of PDAC.

In conclusion, the findings of this study suggest that the urinary biomarkers creatinine, LYVE1, REG1A, REG1B, and TFF1 have potential utility in the early detection of PDAC. Further research is needed to confirm these results and validate their use in a diagnostic tool, but the results of this study provide a promising starting point for the development of a simple, non-invasive test for PDAC.

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.