Pancreatic Cancer Prediction Report

Jorick Baron

2022-11-10

Contents

Introduction

Pancreatic cancer is a difficult to diagnose cancer and when a patient is diagnosed it's often too late with only 4% of patients remaining alive after 5 years after diagnosis [1]. Because of this an early diagnosis is of vital importance but, "There is no single diagnostic test that can tell you if you have pancreatic cancer. Definitive diagnosis requires a series of imaging scans, blood tests and biopsy—and those tests are typically only done only if you have symptoms." [2]. These methods are invasive and difficult to justify as a periodical screening tool, therefore an alternative method of diagnosing or predicting pancreatic cancer, that could also function as a screening tool, is worth researching. Using the data of 590 urine samples together with some patient data could it be possible to have a machine learning model that is capable to create a viable screening method for pancreatic cancer?

Research question

How accurate can a model be trained to detect the difficult to diagnose pancreatic cancer utilising a patient's urine sample?

Material & Methods

Data source

The data for this research project is from here, this data has been sampled from 590 patients and has been provided by: Barts Pancreas Tissue Bank, London, UK; Spanish National Cancer Research Centre, Madrid, Spain; Liverpool University, UK and University College London, UK.

Data cleaning

The data downloaded from kaggle needed to be cleaned to make it viable for this research purpose.

For this purpose the data has been explored using the R programming language and displayed in R markdown (R version 4.0.4) supported by the following packages: dplyr [3], tiydr [4], ggplot2 [5], gridExtra [6], knitr [7], kableExtra [8], e1071 [9] and foreign [10].

The codebook below is to display and explain the data that was downloaded from kaggle.

Table 1: Codebook to explain each variable in the downloaded data.

Name	Fullname	Description	Type	Unit
sample_id	Sample ID	Unique string identifying each subject	string	NA
patient_cohort	Patient's Cohort	Cohort 1, previously used samples; Cohort 2, newly added samples	string	NA
sample_origin	Sample Origin	BPTB: Barts Pancreas Tissue Bank; ESP: Spanish National Cancer Research Centre; LIV: Liverpool University;UCL: University College	string	NA
age	Age	Age in years	int	years
sex	Sex	M = male, F = female	char	NA
diagnosis	Diagnosis (1=Control, 2=Benign, 3=PDAC)	1 = control, 2 = benign hepatobiliary disease; 3 = Pancreatic ductal adenocarcinoma, i.e. pancreatic cancer	int	NA
stage	Stage	For those with pancratic cancer, what stage was it? One of IA, IB, IIA, IIIB, III, IV	string	NA
benign_sample_diagnosis	Benign Samples Diagnosis	For those with a benign, non-cancerous diagnosis, what was the diagnosis?	string	NA
plasma_CA19_9	Plasma CA19-9 U/ml	Blood plasma levels of CA 19–9 monoclonal antibody that is often elevated in patients with pancreatic cancer.	float	plasma units/milliliter
creatinine	Creatinine mg/ml	Urinary biomarker of kidney function	float	mg/ml
LYVE1	LYVE1 ng/ml	Urinary levels of Lymphatic vessel endothelial hyaluronan receptor 1, a protein that may play a role in tumor metastasis	float	ng/ml
REG1B	REG1B ng/ml	Urinary levels of a protein that may be associated with pancreas regeneration.	float	ng/ml
TFF1	TFF1 ng/ml	Urinary levels of Trefoil Factor 1, which may be related to regeneration and repair of the urinary tract	float	ng/ml
REG1A	REG1A ng/ml	Urinary levels of a protein that may be associated with pancreas regeneration.	float	ng/ml

The project is focused on detecting cancer therefore the diagnosis column has been dropped and changed into the has_cancer column, with the values 0 for a control sample and 1 for a pancreatic cancer sample.
After this the dataset was trimmed to remove data that has either no realistic impact on if a patient has pancreatic cancer like sample origin, data that negativily impacts machine learning e.g. sample ids, data that is out of the scope of the project e.g. blood plasma data or data that we can not know of a fresh sample like stage.
Eventually ending up with our data consisting of patient age and sex and the following 5 bio-markers: creatinine, LYVE1, REG1B, TFF1, REG1A. The values of the bio-markers have also been log2 transformed to reduce the skewness in the data.

Table 2: The skewness in the untransformed data.

Variable	Skewness	Interpretation
age	-0.2157312	Fairly symmetrical
creatinine	1.4589654	Greatly positively skewed
LYVE1	1.3869334	Greatly positively skewed
REG1B	3.3169925	Greatly positively skewed
TFF1	5.1321035	Greatly positively skewed
REG1A	4.4254038	Greatly positively skewed

The log transformed data contained less outliers and a better distributed dataset. But this data was not fully normally distributed but no issues relating to this uneven distribution occurred. A full in depth EDA detailing the process step by step can be found at at this github repository.

Results

Skew

First the important task of exploring the data for potential mistakes or outliers has been carried out. To easily catch outliers a boxplot visualisation is applied, this boxplot will also be coloured based on the has_cancer boolean value to also show the difference in data between those with pancreatic cancer and without.

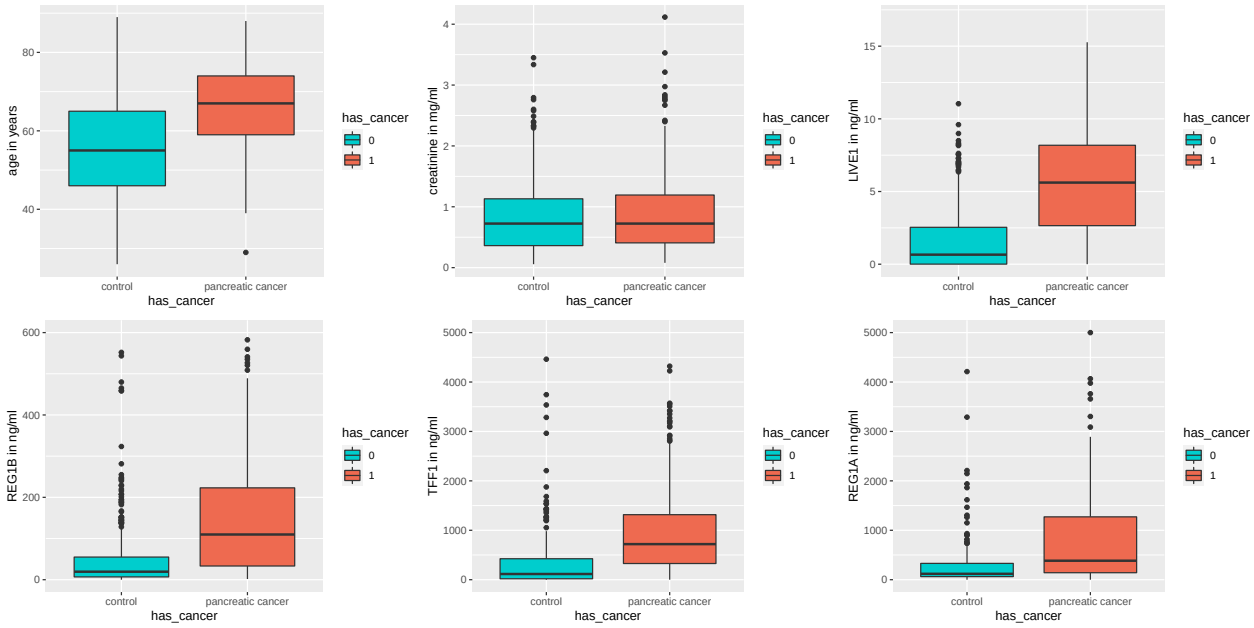

Figure 1: Boxplots of age and the 5 bio-marker data (some figures has been croped for readablity).

In the above boxplots it is visible the data contained many outliers however with how the outlyng data is located above the average at seemingly exponentially higher rates the question that these outliers are no outliers at all and that the data is simply in need of a restructuring to better display the full picture can be asked. This is further corroborated with the data of the skewness from table 2. And after applying the log2 transformation the following boxplots where produced.

The data after the transformation has visibly improved showing less outliers and not needing to be cropped to be visualised properly are notable improvements to the data quality.

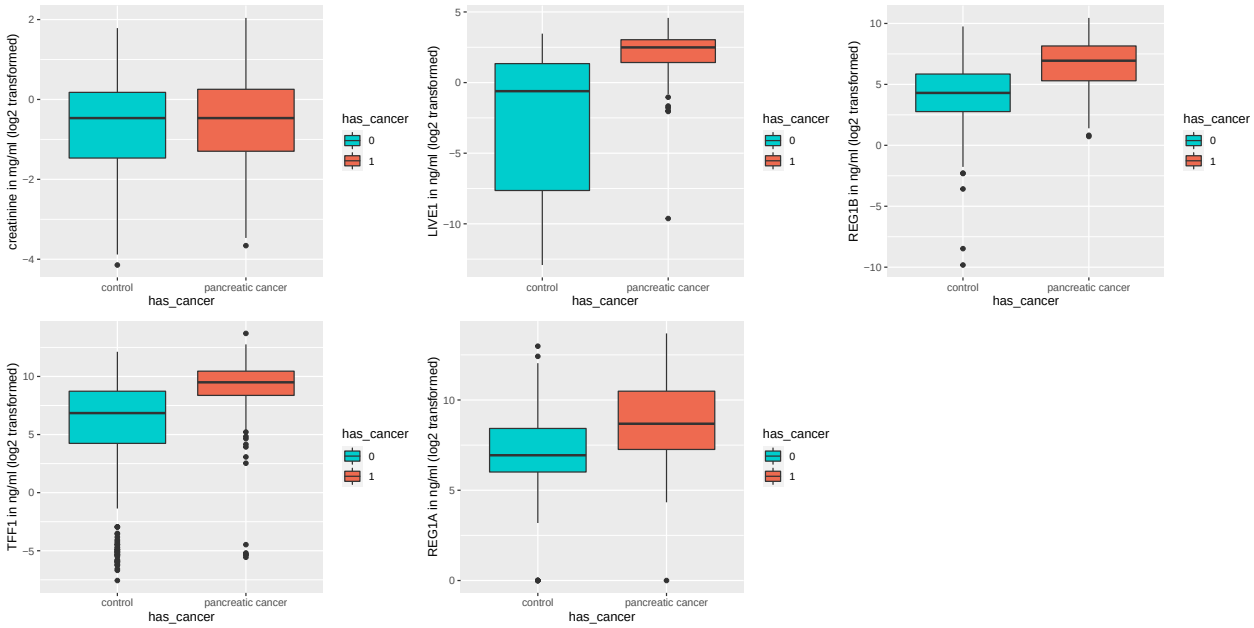

Figure 2: Boxplots of the 5 bio-markers after log2 transformation.

Distribution

However looking at if the data is normally distributed is something that is considered especially good practice after data transformation, shows the following.

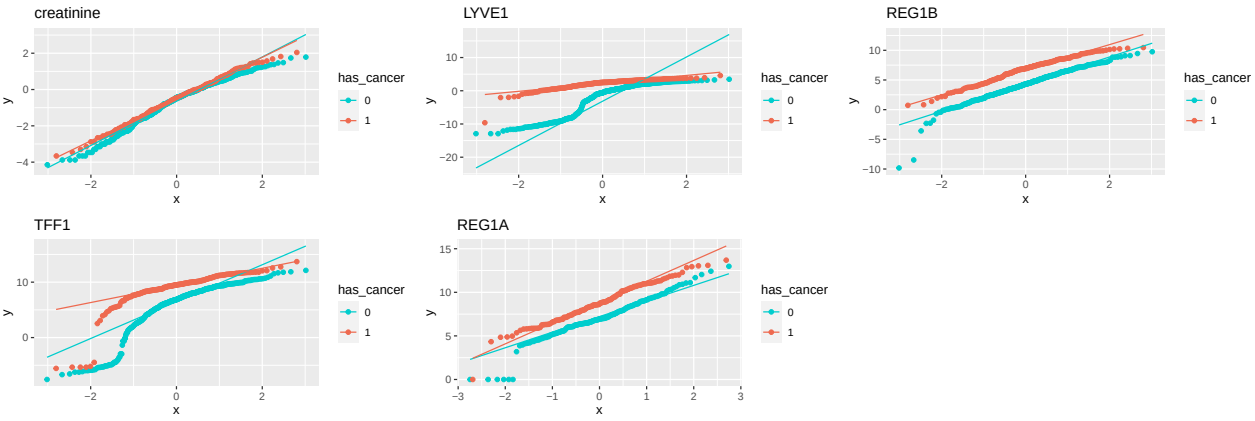

Figure 3: qqplots to show distribution, if the data is normally distributed the dots should follow the line.

The above qqplots are displaying that the data is not normally distributed. To aid in the interpretation of the qqplots the below table will display the results of a Shapiro-Wilk Normality Test.

Table 3: Results of the Shapiro-Wilk Normality Test and interpretation.

Variable	p-value	Interpretation
creatinine	1.2542643×10^{-6}	this data is not normally distributed
LYVE1	$1.7752934 \times 10^{-25}$	this data is not normally distributed

Variable	p-value	Interpretation
REG1B	$9.8879478 \times 10^{-10}$	this data is not normally distributed
TFF1	$1.0585334 \times 10^{-24}$	this data is not normally distributed
REG1A	5.5640181×10^{-6}	this data is not normally distributed

Due to every p-value being below the standard alpha value of 0.05 the null hypothesis that the data is normally distributed is rejected for each variable. The fact that the data is not normally distributed is not fatal for machine learning but any result needs to be held under extra scrutiny.

Correlations

Furthermore the data needed to be checked for correlations as highly correlated data can be counter-productive for machine learning. A good visualisation of correlations is a heatmap.

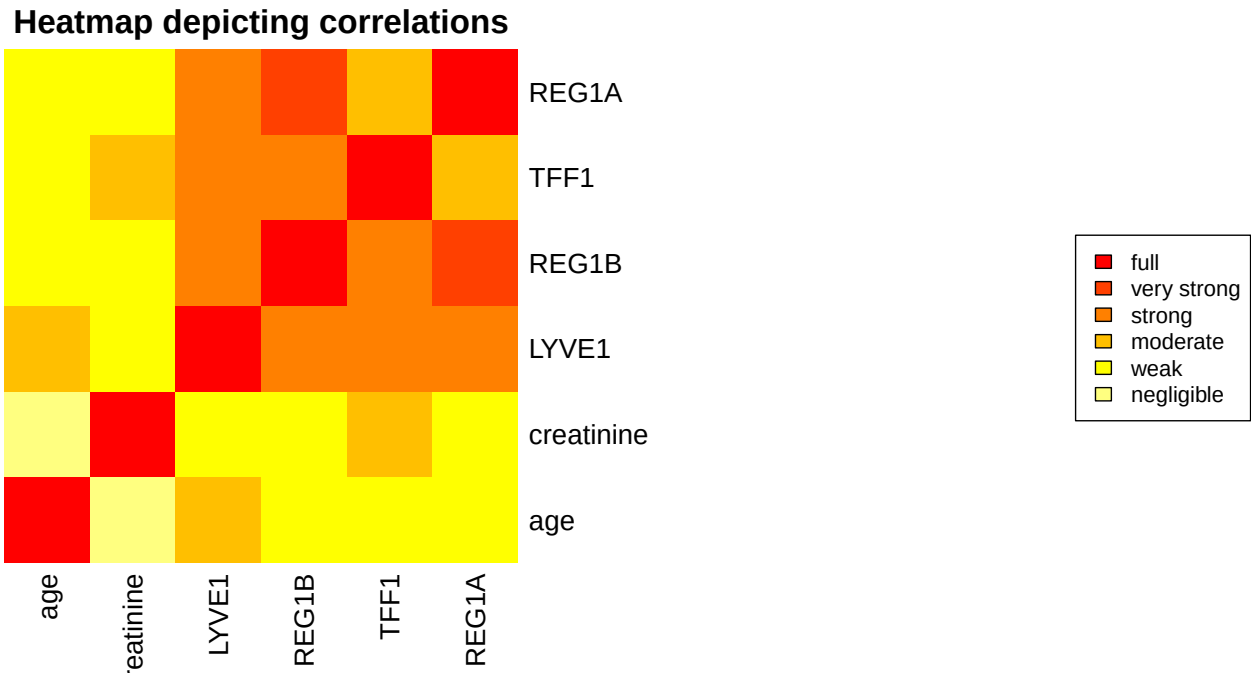

Figure 4: Correlations displayed in a heatmap

From this heatmap it is visible that REG1B and LYVE1 have correlations with other variables but with the highest correlation efficient between REG1B and REG1A (0.7641084) is very strongly positively related but otherwise no datapoints are too strongly correlated.

Finally the class label (has_cancer) has been tested against all the other data using a Welch Two Sample t-test to test if those datapoints are of influence to the class label.

Table 4: T-test results and interpretation.

Variable	p-value	Significant
Age	$1.2428989 \times 10^{-24}$	yes
Creatinine	0.1543498	no
LYVE1	$3.0097865 \times 10^{-54}$	yes
REG1B	$2.4404512 \times 10^{-33}$	yes

Variable	p-value	Significant
TFF1	$2.6006468 \times 10^{-23}$	yes
REG1A	2.687216×10^{-11}	yes

Machine learning phase

After having explored and transformed the data has been written to a .arff file to be ready for use in the weka machine learning workbench [11]. After this the data has been used to try and train many machine learning models using different algorithms, the algorithms were then evaluated on performance in 3 key aspects.

Table 5: Preformance of diferent algorithms

X	ZeroR	OneR..B40	J48..M35	IBk..K19	NaiveBayes	RandomForest	SMO	SimpleLogistic
Percent_correct	66.27	78.03	78.27	79.44	74.95	80.73	82.20	83.19
True_negative_rate	0.00	0.59	0.59	0.61	0.85	0.67	0.71	0.73
Area_under_ROC	0.50	0.73	0.80	0.86	0.85	0.88	0.79	0.89

The 3 key aspects as seen in the above table are percent correct since the model is expected to predict accurately, the true negative rate since it is unwanted for patients to pass undetected and area under the ROC because this is a value that gives insight in the performance of a model to distinguish between the classes.
The SimpleLogistic algorithm scores highest in 2 of those metrics and second best in the true negative rate metric. Because of this the SimpleLogistic model has been chosen to be applied into the program to be used to predict pancreatic cancer.

In this repository the java program to predict pancreatic cancer can be found along with a guide on how to use it in the repository's README.md. For a quick overview the program is run through the command line and expects a arff file as input, and will print out for each sample it's prediction on if it is from a patient with pancreatic cancer or not.

Discussion & Conclusion

Discussion

Despite the fact that the data was not normally distributed the model seems to have handled this without any issues. Furthermore the data did only consist out of 590 samples, this seems to have been enough to train the model, but it is possible that a larger sample pool may have lead to a more accurate model.

Conclusion

With the machine learning results as seen in table 5 the answer to the research question is 83.19% which is quite accurate however not accurate enough to be a screening tool.
But perhaps as stated in the discussion more accuracy can be won by using a bigger sample size.

Future research

The model could be used for another project. The minor Application Design is such a project. The classifier should be accessible for doctors or lab workers without a deep understanding of computers. That is why a desktop application could be made for use in hospitals, clinics or labratories to aid the prediction of pancreatic cancer. The only thing necessary is a urine sample something a patient can easily give to a doctor, who can send it to the lab to extract the data necessary for the model. For privacy the application should be a desktop application that can operate on a local network within the hospital/clinic to make outside attacks more difficult and the data more secure/private.

References

1. Vincent, A., Herman, J., Schulick, R., Hruban, R.H., and Goggins, M. (2011). Pancreatic cancer. Lancet *378*, 607–620.
2. Pancreatic cancer screening (2021). Pancreatic Cancer Screening | Johns Hopkins Medicine.
3. Wickham, H., François, R., Henry, L., and Müller, K. (2022). Dplyr: A grammar of data manipulation.
4. Wickham, H., and Girlich, M. (2022). Tidyr: Tidy messy data.
5. Wickham, H. (2016). ggplot2: Elegant graphics for data analysis (Springer-Verlag New York).
6. Baptiste Auguie, A.A. (2022). Miscellaneous functions for "grid" graphics.
7. Xie, Y. (2022). Knitr: A general-purpose package for dynamic report generation in r.
8. Hao Zhu, T.T., Thomas Travison (2022). Construct complex table with 'kable' and pipe syntax.
9. David Meyer, K.H., Evgenia Dimitriadou (2022). Misc functions of the department of statistics, probability theory group (formerly: E1071), TU wien.
10. R Core Team, V.J.C., Roger Bivand (2022). Read data stored by 'minitab', 's', 'SAS', 'SPSS', 'stata', 'systat', 'weka', 'dBase', ...
11. Frank, E., Hall, M.A., Holmes, G., Kirkby, R., Pfahringer, B., and Witten, I.H. (2005). Weka: A machine learning workbench for data mining. In Data mining and knowledge discovery handbook: A complete guide for practitioners and researchers, O. Maimon and L. Rokach, eds. (Berlin: Springer), pp. 1305–1314.