**CS677 TERM PROJECT**

Name: Allegra Marsiglio

BU ID: U24621582

Semester: SPRING ‘23

Section: A1

Project Title: *Urinary Biomarkers for Pancreatic Cancer*

**Introduction**

Pancreatic cancer has an exceptionally high mortality rate. Indeed, just 9% of individuals with pancreatic cancer survive more than 5 years following their initial diagnosis. Additionally, the early stages of this type of cancer are asymptomatic, making a diagnosis in the early stages challenging. According to the National Cancer Institute, there would be 64.000 new cases of pancreatic cancer in 2023, with an anticipated 50.000 deaths—or roughly 78% of new cases. This statistic makes it quite evident how important it is today to find solutions for the early detection of this disease.

**Dataset**

In order to investigate the above-described matter, I will use a dataset used for research that was published in Plos Medicine by Silvana Debernardi and colleagues. The dataset contains data on 590 patients that are divided into three groups:

1. Healthy patients
2. Patients with non-cancerous pancreatic conditions
3. Patients with pancreatic ductal adenocarcinoma (a specific type of pancreatic cancer)

The information that was collected for these patients include age, sex, and a series of biomarkers from the urine. In particular, the biomarkers that I will focus on in my study are:

* Creatinine: a 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.
* REGB1: urinary levels of a protein that may be associated with pancreas regeneration.
* TFF1: urinary levels of Trefoil Factor 1, which may be related to regeneration and repair of the urinary tract.

Since there were no missing data for the four features I extracted, I won't go into great detail on the data cleaning I did on this dataset. But before diving into the research procedure, allow me to give you a few statistics:

* Number of patients by diagnosis:
  + healthy status: 183
  + non-cancerous pancreatic condition: 208
  + pancreatic cancer: 199
* Age:
  + Mean: 59
  + Standard deviation: 13
* Patients by sex:
  + Female: 299
  + Male: 291
* Creatinine:
  + Mean: 0.855
  + Standard deviation: 0.639
* LYVE1:
  + Mean: 3.064
  + Standard deviation: 3.439
* REGB1:
  + Mean: 111.8
  + Standard deviation: 196.3
* TFF1:
  + Mean: 597.9
  + Standard deviation: 1010.5

The "healthy" and "non-cancerous condition" patients were compiled into a single "cancer negative" class as the last step of my data preprocessing, which I will contrast with the "cancer positive" class. My research does aim to create a model that can predict if a person will acquire pancreatic cancer.

**Features analysis**

According to the correlation matrix of the chosen features, there were some features with relatively strong correlations. As a matter of fact, the absolute correlation between LYVE1 and REG1B was 0.54; LYVE1 and TFF1 was 0.58; and REG1B and TFF1 was 0.69. Due to the high correlation, I decided to first check for multicollinearity before continuing with the analysis.

The Variance Inflation Factor (VIF) quantifies the degree to which an independent variable's behavior (variance) is affected or inflated by its interaction with or connection with other independent variables. Generally speaking, variables have no correlation when the VIF is equal to 1, moderate correlation when it is between 1 and 5, and strong correlation when it is greater than 5. The results of the VIF test on the features under consideration is the following:

Text

Description automatically generated

Since the VIF for all of the features was higher than 1, I could notice some level of multicollinearity. Therefore, in my analysis, I decided to apply models that are generally less affected by multicollinearity.

**Modelling**

Following the previous analysis, the classifiers I chose to model my dataset with are Decision Tree, Random Forest, Linear SVM, Gaussian SVM, Polynomial SVM with degree 3, and Logistic Regression.

While Decision Tree, Random Forest, and Support Vector Machine operate well even with features collinearity, Logistic Regression is often more influenced by it, especially if one of the goals of modeling Logistic Regression is to estimate the importance of features. However, I chose to include it in my modelling because it is a widely used classifier for predicting diseases.

The way Logistic Regression is affected by multicollinearity can be mitigated by the use of Principal Component Analysis (PCA), which is a dimensionality reduction technique that is frequently used to reduce the dimensionality of big data sets by condensing a large collection of variables into a smaller set that retains the majority of the large set's information. In applying it to the dataset under considerations, I could observe that 5 features were enough to explain more than 90% of the variability in the model. Therefore, I used PCA to reduce the number of components from 6 to 5.

**Models’ Comparison and Results**

The following table summarizes the performance indicators of each model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classifier | **Accuracy** | **Sensitivity (TPR)** | **Specificity (TNR)** | **AUC score** |
| **Decision Tree** | 0.77 | 0.59 | 0.86 | 0.72 |
| **Random Forest** | 0.82 | **0.75** | 0.82 | **0.79** |
| **SVM Linear** | 0.80 | 0.56 | 0.92 | 0.74 |
| **SVM Gaussian** | **0.83** | 0.59 | 0.93 | 0.76 |
| **SVM Polynomial** | 0.74 | 0.37 | **0.96** | 0.67 |
| **Logistic Regression** | 0.79 | 0.62 | 0.88 | 0.75 |
| **Logistic Regression + PCA** | 0.81 | 0.63 | 0.89 | 0.76 |

As you can notice, the Gaussian SVM model produced the highest accuracy, Random Forest produced the most sensitivity, Polynomial SVM produced the highest specificity, and Random Forest produced the highest AUC score. Overall, Random Forest was the most accurate model for identifying whether a patient had cancer or not.

**Other Observations and Conclusions**

Since my dataset was very small, a way to improve my research would be using cross-validation to train the models. Additionally, in my research I am not considering the differences in biomarkers depending on the age and sex of the patients, which might have its relevance. A possible solution would be trying to perform the analysis on different age groups separately, and see in results change depending on whether the patient is male or female.

Overall, Random Forest resulted in a quite good level of accuracy, which is approximately 82%. It is also good that Random Forest is the model with the highest Sensitivity, since disease prediction models have usually higher interest in correctly predicting positive cases. Finally, Random Forest also produced the highest AUC scores, which assesses the model's performance at distinguishing between the positive and negative classes and, therefore, takes into account both Sensitivity and Specificity.

**References**:

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

<https://towardsdatascience.com/how-do-you-apply-pca-to-logistic-regression-to-remove-multicollinearity-10b7f8e89f9b#:~:text=PCA%20(Principal%20Component%20Analysis)%20takes,effectively%20eliminate%20multicollinearity%20between%20features>.

<https://towardsdatascience.com/7-techniques-to-handle-multicollinearity-that-every-data-scientist-should-know-ffa03ba5d29>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2849171/>

<https://www.geeksforgeeks.org/principal-component-analysis-with-python/>

<https://www.geeksforgeeks.org/test-of-multicollinearity/>

<https://statisticalhorizons.com/multicollinearity/>

[https://www.kaggle.com/code/dansbecker/cross-validation#](https://www.kaggle.com/code/dansbecker/cross-validation)

<https://seer.cancer.gov/statfacts/html/pancreas.html>