Risk Prediction of Pancreatic Cancer Using Urine Biomarkers

Jenny Harston
harstonj@uw.edu
Computing and Software Systems
University of Washington Bothell
Bothell, WA, USA

ABSTRACT

Pancreatic ductal adenocarcinoma (PDAC) is a deadly cancer that is mostly diagnosed late [1]. Earlier detection of this cancer would help increase the survival rate of patients. Using a three-biomarker urine panel along with creatinine levels, age, and sex, a prediction model was created to identify patients with PDAC. Multiple machine learning algorithms were modeled and their results were compared to determine which model had the best performance. Support vector classification had the highest performance score and further research should be conducted to fine tune the model and increase prediction accuracy.

CCS CONCEPTS

• **Computing methodologies → Classification and regression trees**.

KEYWORDS

pancreatic cancer, classification, machine learning

1 INTRODUCTION

Pancreatic ductal adenocarcinoma (PDAC) is an aggressive cancer with an average 5-year survival rate of less than 10% [4]. One of the major challenges to effectively fighting this disease is the ability to detect it at early stages [3]. It is seldom detected early because symptoms of this disease typically do not appear until it has progressed and metastasized. The stage at the time of diagnosis greatly influences the response to treatments, which are not sufficiently effective for patients with late stages of PDAC [4]. If caught early, the odds of surviving are much higher.

Few models have been designed for risk prediction of PDAC. Having a prediction model that uses a non-invasive urinary biomarker panel would aid in earlier detection of pancreatic cancer [1]. There are two studies that have inspired the goal of this research paper.

The first study developed PancRisk [1], "A urine biomarker-based risk score for stratified screening of pancreatic cancer patients." The sample data they collected included healthy individuals and patients diagnosed with PDAC. The data provided measurements for three urine biomarkers (LYVE1, REG1B, and TFF1), creatinine, and age, to train multiple machine learning algorithms and compared their performance. They found that none of the algorithms significantly outperformed the others and used a logistic regression model to create the PancRISK score.

The second study by Silvana Debernardi and colleagues, [2] continued the work of this first study, with the goal to establish the accuracy of an improved panel, using the PancRISK score, and increased the number of samples in the dataset. The new dataset included samples for benign hepatobiliary diseases.

The goal of this research is to build off of the two studies previously conducted in the hopes of improving prediction accuracy. I will use the dataset from the previously conducted studies to identify patients with pancreatic cancer using the three urine biomarkers, creatinine, age, along with the data attribute of sex, to create a risk prediction model.

2 METHOD

The task of this research is to train a machine learning model to identify patients with pancreatic cancer using the data attributes: age, sex, creatinine, LYVE1, REG1B, and TFF1. Figure 1 shows an overview of the system used for training and evaluating a prediction model.

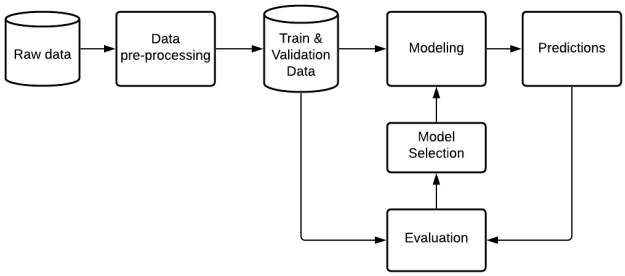

Figure 1: Workflow of Proposed Machine Learning Framework.

2.1 Dataset

The dataset used for this work was acquired from the website Kaggle and is a verbatim copy of the raw data provided in the paper by Debernardi et al. [2], with column names changed for easier importing and use. A description of the data is as follows, with sample distributions for diagnosis and sex shown in Tables 1 and 2:

- Total number of samples: 590
- age: Age in years
- sex: M = male, F = female
- creatinine: mg/ml Urinary biomarker of kidney function
- LYVE1: ng/ml Urinary levels of Lymphatic vessel endothelial hyaluronan receptor 1, a protein that may play a role in tumor metastasis
- REG1B: ng/ml Urinary levels of a protein that may be associated with pancreas regeneration.
- TFF1: ng/ml Urinary levels of Trefoil Factor 1, which may be related to regeneration and repair of the urinary tract

- diagnosis: 1 = control (no pancreatic disease), 2 = benign hepatobiliary disease, 3 = Pancreatic ductal adenocarcinoma (PDAC).

Table 1: Diagnosis sample distribution

Diagnosis	Samples (n)
Control	183
Benign	208
PDAC	199

Table 2: Sex sample distribution

Sex	Samples (n)
Male	291
Female	299

2.2 Data Pre-processing

To prepare the dataset for modeling, the data csv file was first loaded into python as a pandas dataframe. The dataset contained unwanted variables, so the next step was to remove the unwanted attributes from the set. To get a better visual understanding of the data, a pair plot was produced to see the distribution of the individual attributes and relationships between pairs of attributes, as shown in Figure 2. To use the attribute of sex in the training, the male and female variables needed to be encoded to binary. In order to compare the different diagnoses, the diagnosis data needed to be separated into individual datasets. To prepare the diagnosis data for comparison, the separated diagnosis sets were combined into comparison sets: Control vs. PDAC, and Benign vs. PDAC. After combining, the data in each set was shuffled and standardized.

2.3 Modeling

After pre-processing the data, the next step was to split out training and validation sets, with a ratio of 80% training/20% validation. The test harness was set up using Stratified K-fold cross-validation, using n = 10 splits. Multiple machine learning models were built using the scikit-learn package. The test data set was then split and trained on the models. After training, predictions were made using the validation set to test the trained models, and then evaluated.

Models:

- Logistic Regression (LR)
- Linear Discrimination Analysis (LDA)
- K-Nearest Neighbors (KNN)
- Decision Tree (CART)
- Gaussian Naïve Bayes (NB)
- Support Vector Classification (SVC)

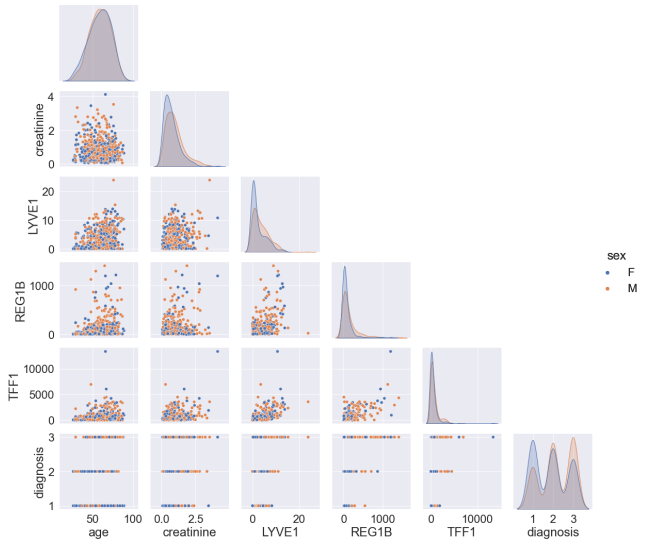

Figure 2: Plotting of pairwise relationships in the dataset.

3 EXPERIMENT AND RESULT

After training, model predictions were evaluated using the following metrics: accuracy score, confusion matrix, precision, recall, f-1 score, and ROC curve with AUC score. Comparing these metrics helps determine the best model to select for creating a prediction model. The best overall model performance for both comparison sets appears to be support vector classification, with one of the top scores in both accuracy and ROC-AUC.

3.1 Control vs. PDAC

Model performance for Control vs. PDAC is shown in Table 3. The models with the highest accuracy scores were K-nearest neighbors with 89.3%, logistic regression with 89.3%, and support vector classification with 88%. The ROC curve and AUC score for Control vs. PDAC is shown in Figure 3. This was computed by scikit-learn using an estimator. Here we can see that the best performance was accomplished with SVC at 95%, followed by logistic regression at 94%.

3.2 Benign vs. PDAC

Model performance for Benign vs. PDAC is shown in Table 4. The model with the highest accuracy score was support vector classification with 85%. The ROC curve and AUC score for Benign vs. PDAC is shown in Figure 4. This was also computed by scikit-learn using an estimator. Here we can see that four of the models tied for best performance at 90%: logistic regression, linear discriminant analysis, gaussian Naïve Bayes, and SVC. Predicting benign pancreatic disease vs. pancreatic cancer was not as accurate as predicting no pancreatic disease, which should lead to future investigation of this prediction.

4 CONCLUSION AND DISCUSSION

Creating an accurate prediction model for early detection of pancreatic ductal adenocarcinoma could help save many lives. The results

Table 3: Model Performance for Control vs. PDAC

		Precision		Recall		F-1		
Model	Accuracy	Control	PDAC	Control	PDAC	Control	PDAC	AUC
LR	0.893	0.94	0.85	0.84	0.95	0.89	0.90	0.94
LDA	0.827	0.82	0.83	0.84	0.81	0.83	0.82	0.92
KNN	0.893	0.88	0.91	0.92	0.86	0.90	0.89	0.90
CART	0.840	0.84	0.84	0.84	0.84	0.84	0.84	0.84
NB	0.853	0.83	0.88	0.89	0.81	0.86	0.85	0.93
SVC	0.880	0.89	0.87	0.87	0.89	0.88	0.88	0.95

Table 4: Model Performance for Benign vs. PDAC

		Precision		Recall		F-1		
Model	Accuracy	Benign	PDAC	Benign	PDAC	Benign	PDAC	AUC
LR	0.825	0.81	0.84	0.85	0.80	0.83	0.82	0.90
LDA	0.825	0.81	0.84	0.85	0.80	0.83	0.82	0.90
KNN	0.800	0.80	0.80	0.80	0.80	0.80	0.80	0.85
CART	0.750	0.73	0.78	0.80	0.70	0.76	0.74	0.75
NB	0.788	0.73	0.87	0.90	0.68	0.81	0.76	0.90
SVC	0.850	0.89	0.82	0.80	0.90	0.84	0.86	0.90

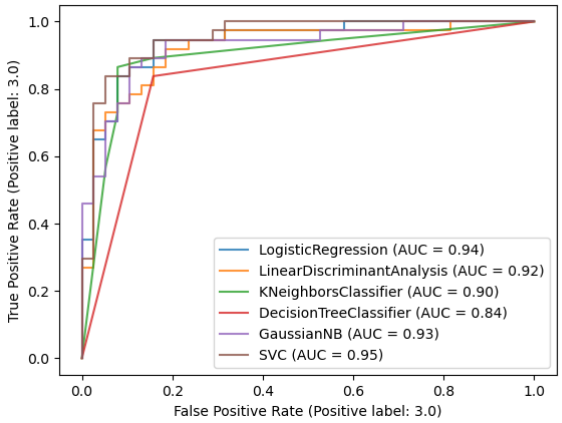

Figure 3: ROC Curve for Control vs. PDAC

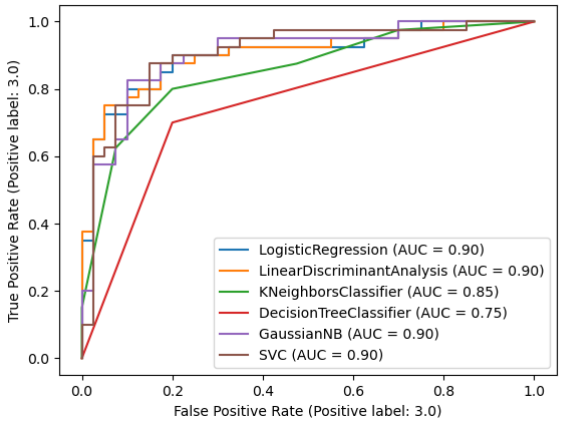

Figure 4: ROC Curve for Benign vs. PDAC

from the previous studies on this data and the results from my work were similar, but different testing methods were used. The original paper [2] was able to achieve an AUC of 0.936 on the validation set, where my work resulted in an AUC of 0.94 on the validation set using an estimator, but calculations using the actual predictions resulted in a slightly lower AUC. In the paper, the training/validation sets were split in a 1:1 ratio, and my work was split 80%/20%. The paper used bootstrap cross-validation and my work used stratified K-folds. The paper also excluded the data attribute of sex in their calculations.

Further adjustments to these selections could result in higher prediction accuracy and should continue to be evaluated. Accuracy may be improved if a 1:1 ratio of each diagnosis was selected before combining for modeling. Another way to try and improve accuracy could be to compare Control & Benign vs. PDAC. Perhaps in the future, other biomarkers will be discovered that can be added to the panel used in this dataset to provide an even better prediction model.

REFERENCES

[1] Oleg Blyuss, Alexey Zaikin, Valeriia Cherepanova, Daniel Munblit, Elena M Kiseleva, Olga M Prytomanova, Stephen W Duffy, and Tatjana Crnogorac-Jurcevic.

2020. Development of PancRISK, a urine biomarker-based risk score for stratified screening of pancreatic cancer patients. *Br. J. Cancer* 122, 5 (March 2020), 692–696. https://doi.org/10.1038/s41416-019-0694-0

[2] Silvana Debernardi, Harrison O'Brien, Asma S Algahmdi, Nuria Malats, Grant D Stewart, Marija Plješa-Ercegovac, Eithne Costello, William Greenhalf, Amina Saad, Rhiannon Roberts, Alexander Ney, Stephen P Pereira, Hemant M Kocher, Stephen Duffy, Oleg Blyuss, and Tatjana Crnogorac-Jurcevic. 2020. A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case-control study. *PLoS Med.* 17, 12 (Dec. 2020), e1003489. https://doi.org/10.1371/journal.pmed.1003489

[3] Michael Orth, Philipp Metzger, Sabine Gerum, Julia Mayerle, Günter Schneider, Claus Belka, Maximilian Schnurr, and Kirsten Lauber. 2019. Pancreatic ductal adenocarcinoma: biological hallmarks, current status, and future perspectives of combined modality treatment approaches. *Radiat. Oncol.* 14, 1 (Aug. 2019), 141. https://doi.org/10.1186/s13014-019-1345-6

[4] Panagiotis Sarantis, Evangelos Koustas, Adriana Papadimitropoulou, Athanasios G Papavassiliou, and Michalis V Karamouzis. 2020. Pancreatic ductal adenocarcinoma: Treatment hurdles, tumor microenvironment and immunotherapy. *World J. Gastrointest. Oncol.* 12, 2 (Feb. 2020), 173–181. https://doi.org/10.4251/wjgo.v12.i2.173