# Parallel and Distributed Computing Assignment 2

Krista Miller

# The dataset:

According to the Mayo Clinic, pancreatic cancer has a survival rate of about 7 percent. One primary reason that the survival rate is so low is because it is often diagnosed very late in the disease process. Several researchers are trying to isolate a diagnostic test that could identify pancreatic cancer, including exploring biomarkers present in urine samples. The dataset used for this project originates from an open-access paper published December 10th, 2020, titled “A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case control study” (Debernardi, S., O’Brien, H., Algahmdi, A.S, et. al), and is available on Kaggle.

# Machine learning problem:

The goal of this machine learning problem is to accurately differentiate between 3 labeled classes of patients in the dataset: Healthy (coded as ‘1’), non-cancerous pancreas condition (coded as ‘2’), and pancreatic cancer (coded as ‘3’).

The key features for this dataset include four biomarkers:

1. Creatinine (protein that is used as an indicator of kidney function)
2. LYVE1 (protein that may play a role in metastatic tumors)
3. REG1B (protein that may be associated with pancreas regeneration)
4. TFF1 (may be related to regeneration and repair of the urinary tract).

Other features that are available in this dataset:

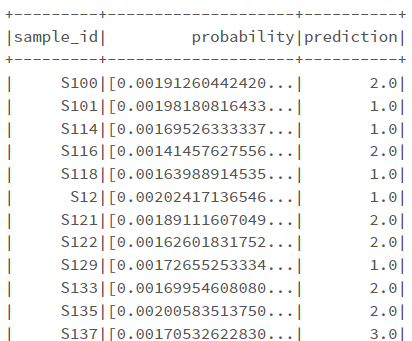
1. Age
2. Sex

Multinomial logistic regression was used as an attempt to solve this problem. First, two columns were dropped from the dataset (‘benign\_sample\_diagnosis’ and ‘stage’) due to the high percentage of null values. The file was split into a 70% training (417 patients) and 30% test (173 patients) set using the randomSplit function.

Next, to build the machine learning pipeline, the Bucketizer function was used to split the ages into five bins. The StringIndexer function transformed the ‘M’ and ‘F’ labels into numbers. The remaining features (creatinine, LYVE1, REG1B, TFF1) were left as float numbers. All of the transformed variables were used as inputs in the VectorAssember function with the output column labeled ‘features’.

The logistic regression model was built using ‘features’ for the featuresCol, and diagnosis as the labelCol and the following hyperparameters: maxIter = 8, regParam = 0.02, elasticNetParam = 0.07.

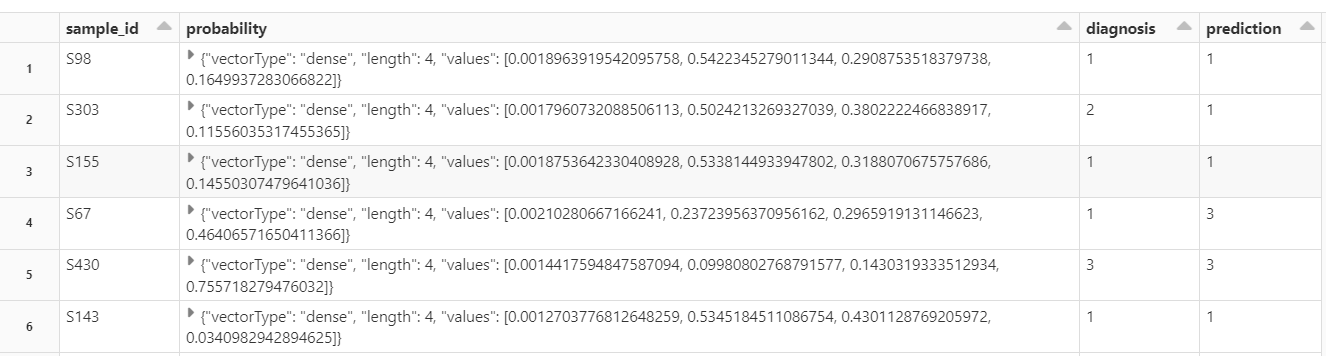
A pipeline was built using the stages outlined above to fit the training data. Once the pipeline model was built, the testing data was used. Below is a sample of the prediction classification on test set patients:



The accuracy for the model using the MulticlassClassificationEvaluator was 62.35%. For future experiments, I would consider scaling the feature variables and using a technique to optimize the hyperparameters.

# Streaming the data:

The pancreas test dataset was read into the databricks environment with a defined schema. The data was then repartitioned and written into 10 csv files in order to simulate streaming data. For the sourceStream, the maxFilesPerTrigger was set to 1. Finally, the pipeline that was fit to the training set is used to predict incoming data, seen below:



Given the poor accuracy of the model, we can see that there are 2 erroneous predictions from the actual diagnosis in the above sample of streaming predictions.

# Data issues:

This dataset is fairly clean, although a couple additional potential feature columns had to be dropped because of the high percentage of null values. Because the accuracy of the model was not great, it would be interesting to repeat this problem by grouping the healthy and non-cancerous pancreatic condition samples together and use a binary logistic regression model to see if the evaluation metrics performed better than this model.

# References:

<https://www.kaggle.com/datasets/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer>