Phelan-Mcdermid Syndrome Project

CPSC 4300 - Applied Data Science, Spring 2023

Sumanth Pandiri
Models & Writing
Sumanth Pandiri

CJ Boni
Models & Writing
Cameron Boni

Yash Patel
Models & Writing
Yash Patel

ABSTRACT

In this paper, we identify significant differences between PMS (Phelan-McDermid syndrome) patients and control cell lines, identify candidate pathways that are particularly affected in individuals with PMS, and identify subgroups within the PMS population based on their metabolic profiles. The data for this project was generated using the Biolog Phenotype Mammalian Microarrays: this technology measures the production of energy (NADH) in the presence of different metabolic compounds. GitHub Link

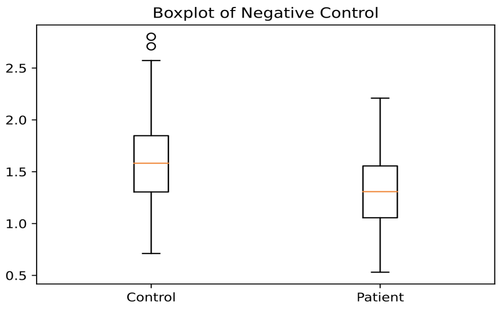
1. INTRODUCTION

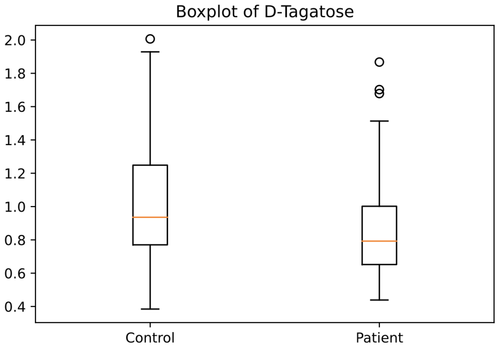
The goal of our project is to use the endpoint absorbance of metabolic pathways in human cells to determine the presence of Phelan-McDermid Syndrome (PMS) in a subject. PMS is a rare heterogeneous genetic condition caused by genes 22 and/or 23, with symptoms such as developmental and speech delays, behavioral problems, and a weakened or no ability to feel pain or sweat. The cells used in the study were derived from the participant's blood, and the endpoint absorbance of the cells was observed using the Biolog Phenotype Mammalian Microarrays. The dataset measures endpoint absorbance across 768 different wells, which tested various metabolic sources and effectors. Not all of these wells were unique, but we were still provided with a very wide variety. The data set provided us with 768 well readings for 50 control subjects and 48 patient subjects, giving us 75,264 individual observations. The control data samples were collected on various dates from June 21st, 2016, to September 12, 2017, and the patient data samples were collected from October 24th, 2016, to December 30th, 2019. The endpoint absorbance data were measured using the color reading values of Redox Dye Mixes, which uses a tetrazolium dye that can be reduced to purple formazan.

2. SUMMARY OF EDA

We were provided with normalized versions of the raw control and patient data, which we used for this project. We imported the data from the Excel sheets into Pandas data frames and cleaned the data. This process included steps such as removing blank or unnecessary columns and rows, such as the dates or well numbers.

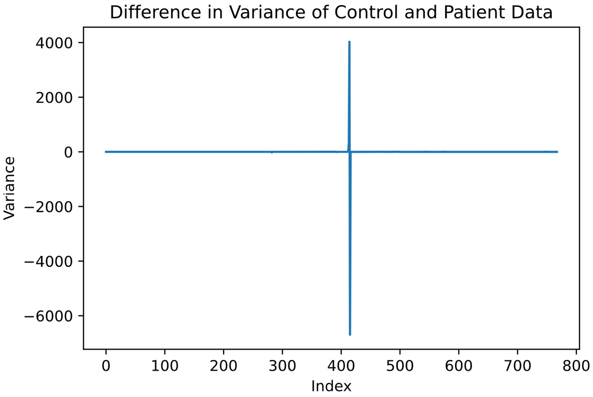
We then performed a preliminary visualization of our data to check for any obvious patterns. We first converted the data to NumPy arrays. Using MatPlotLib, we made side-by-side boxplot figures comparing the control and patient data for each well. Here are 2 sample boxplots for PM-M1, with the y-axis being a measurement of the dye reading:

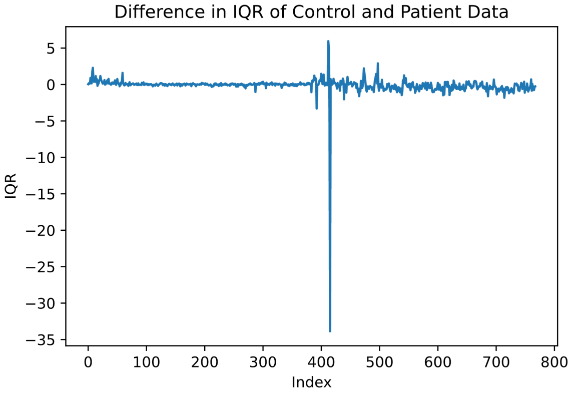


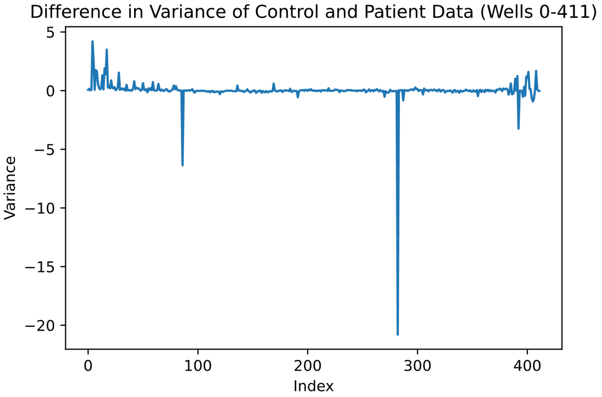


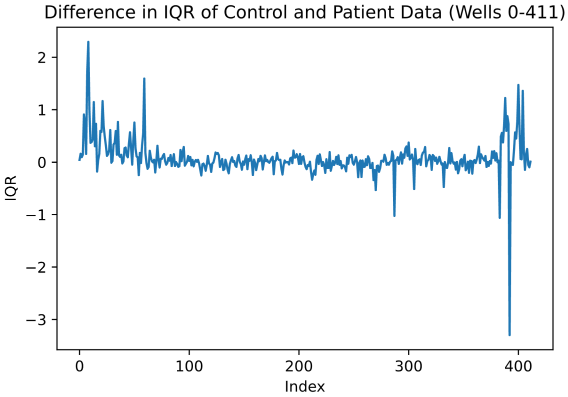
From the 768 boxplots, we noticed that the control medians tended to be higher than the patient medians. We also noticed that there were a significant amount of outliers throughout the graphs.

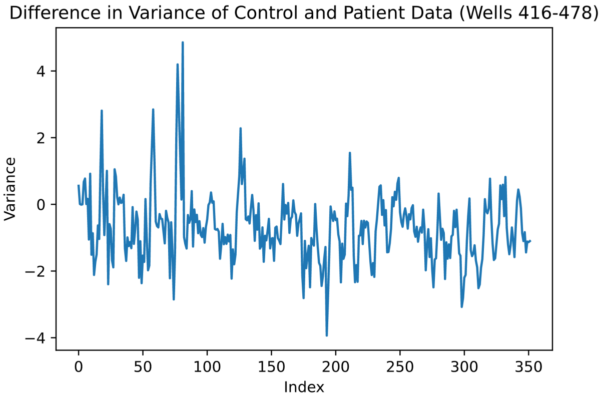
We decided to perform further statistical analysis. We decided to look at the median difference between the control and patient medians in each well, as well as the median difference between the control and patient means across each well. We also plotted the differences in patient and control medians and means, as well as the differences of IQR and variance between patient and control groups per well. Upon examination of these visualizations, we noticed an outlier from wells 412-415, which all represented Manganese Chloride.

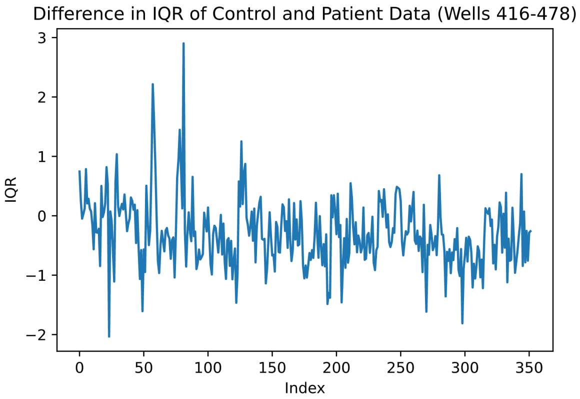

Difference in Variance of Control and Patient Data
Variance
4000
2000
0
−2000
−4000
−6000
0
100
200
300
400
500
600
700
800
Index

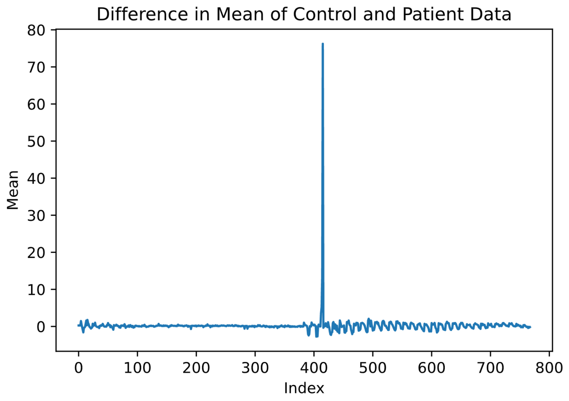

Difference in IQR of Control and Patient Data
IQR
5
0
−5
−10
−15
−20
−25
−30
−35
0
100
200
300
400
500
600
700
800
Index

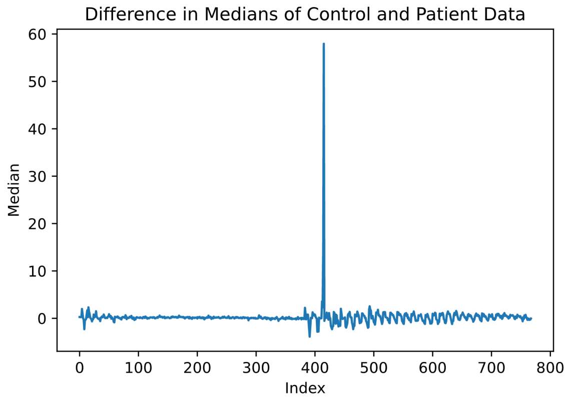

Difference in Variance of Control and Patient Data (Wells 0-411)
Variance
5
0
−5
−10
−15
−20
0
100
200
300
400
Index

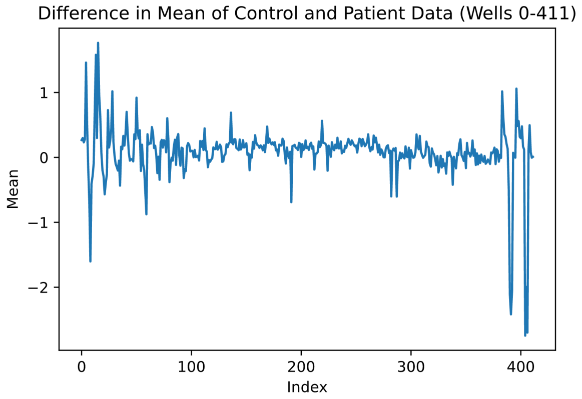

Difference in IQR of Control and Patient Data (Wells 0-411)
IQR
2
1
0
−1
−2
−3
0
100
200
300
400
Index

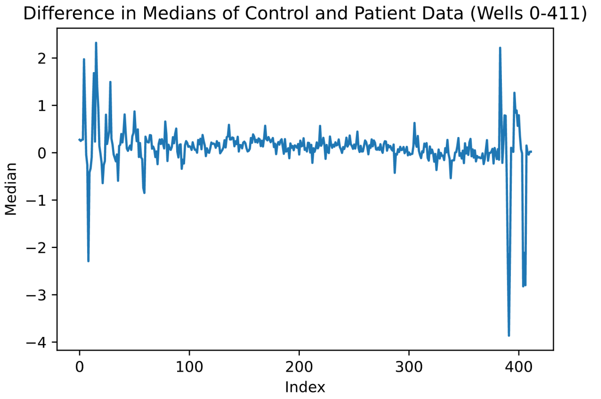

Difference in Variance of Control and Patient Data (Wells 416-478)
Variance
4
2
0
−2
−4
0
50
100
150
200
250
300
350
Index

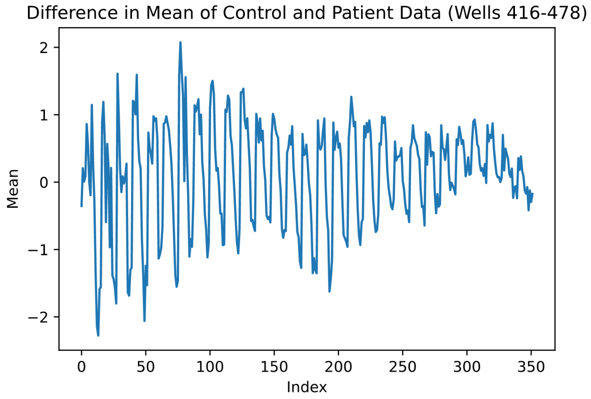

Difference in IQR of Control and Patient Data (Wells 416-478)
IQR
3
2
1
0
−1
−2
0
50
100
150
200
250
300
350
Index

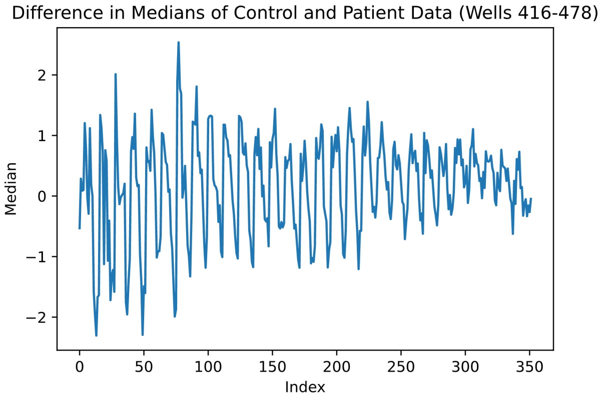

Difference in Mean of Control and Patient Data


Difference in Medians of Control and Patient Data


Difference in Mean of Control and Patient Data (Wells 0-411)


Difference in Medians of Control and Patient Data (Wells 0-411)


Difference in Mean of Control and Patient Data (Wells 416-478)


Difference in Medians of Control and Patient Data (Wells 416-478)

We compared the median differences in patient and control medians and means of the overall dataset, as opposed to the Manganese Chloride, excluded dataset, and these were the results:

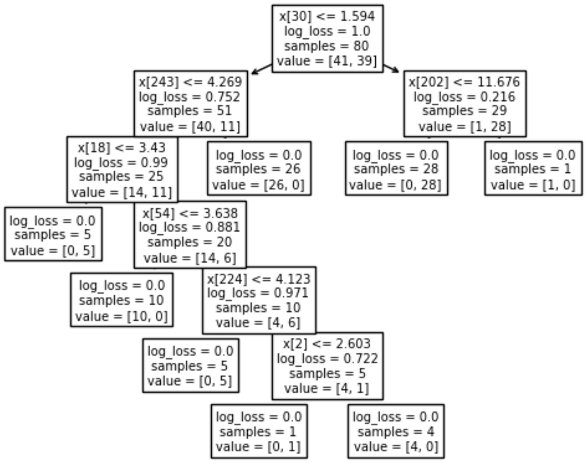
	Median Difference in Patient and Control Means Across Each Well:	Median Difference in Patient and Control Medians Across Each Well:
Including Manganese Chloride	0.24432858	0.2609712967

Excluding Manganese Chloride	0.24097688	0.258738528

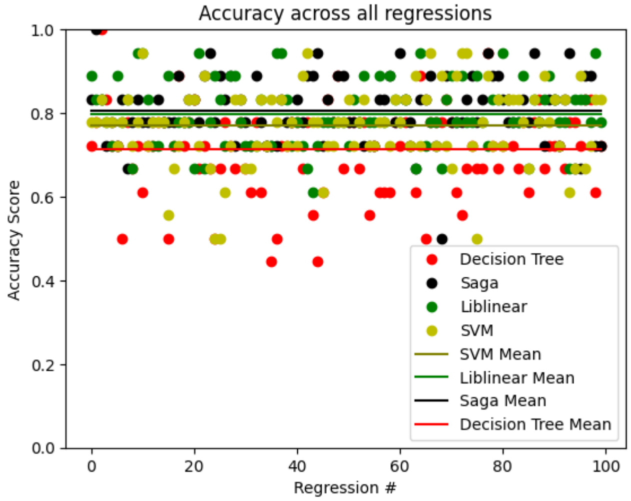
We noticed that the wells after Manganese Chloride have a different trend, with the difference between the patient and control seeming to be at greater magnitudes. Due to the size of our feature number, 768, we decided to remove wells that didn't show much difference between patients and controls so that we could prevent overfitting. We removed any pathways that had a difference in the control and patient medians between -0.5 and 0.5, leaving us with 254 wells. Unsurprisingly, 90.5% of the wells before Manganese Chloride were excluded, while only 40% of the wells after Manganese Chloride were excluded. As shown in the visualization above, the wells after Manganese Chloride tended to show greater differences between patients and controls. In doing this, we hoped to isolate the more prominent pathways that would be predictors for whether or not a person had Phelan-Mcdermid Syndrome. We made 2 sets of training data, one with Manganese Chloride, and one without, so we could compare the results.

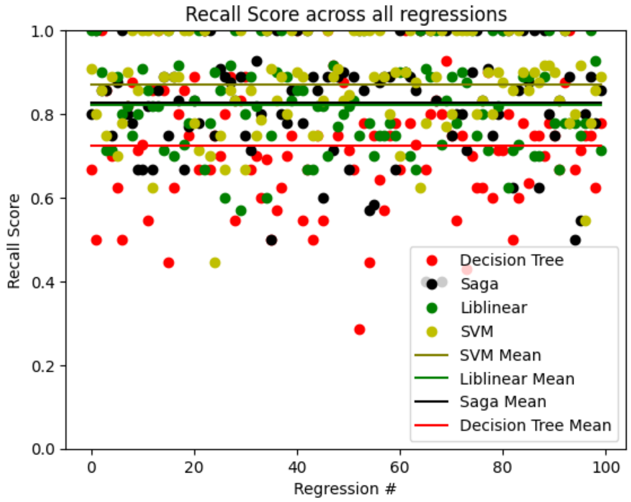
3. SUMMARY OF MACHINE LEARNING MODELS

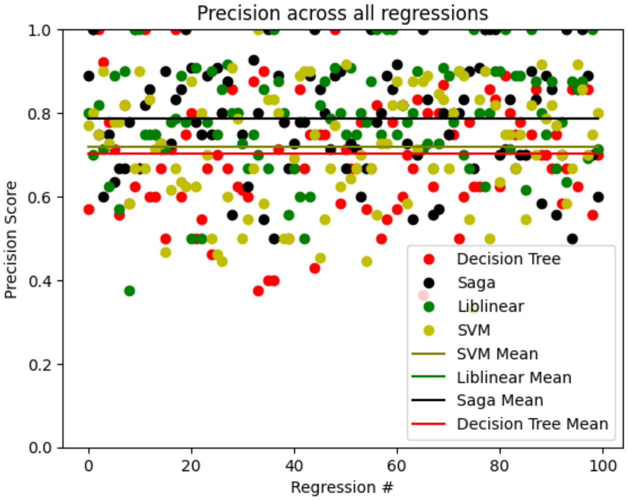
The goal of our machine learning models is to predict whether or not a person has Phelan-McDermid Syndrome via an analysis of their metabolic profile. Since our output is binary classification, we chose models that met those requirements. These four models are as follows: logistic regression using the "liblinear" solver, logistic regression using the "saga" solver, a decision tree classifier, and a support vector machine. We trained and tested those four regression/machine learning models on our data and then analyzed the results of each model. An example of how the decision tree classifier made its decisions is as follows:

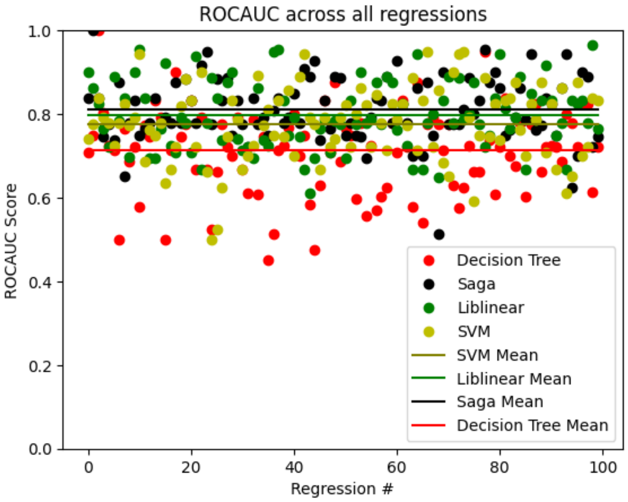


We compared all 4 models using accuracy, recall, precision, and area under the curve scores. We also compared the log loss for the regression models. Here are the plots used to visualize the results of the models trained on data excluding Manganese Chloride:

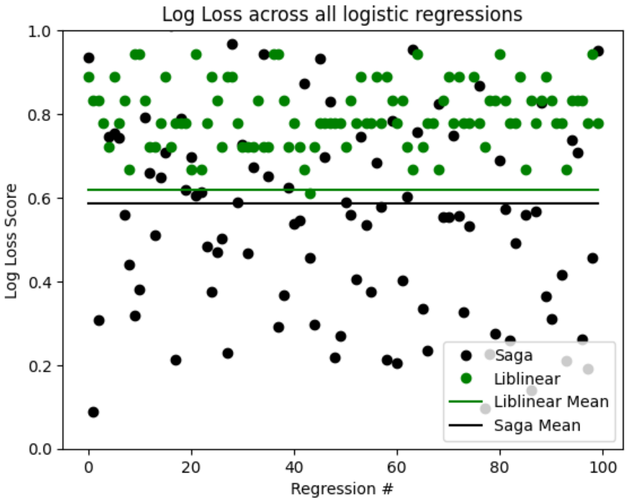








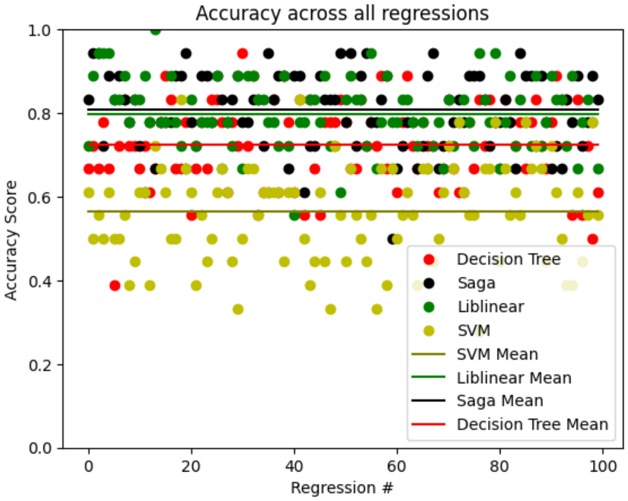
Since the logistic regression models are the only ones that used log loss as a factor, they were compared as follows:

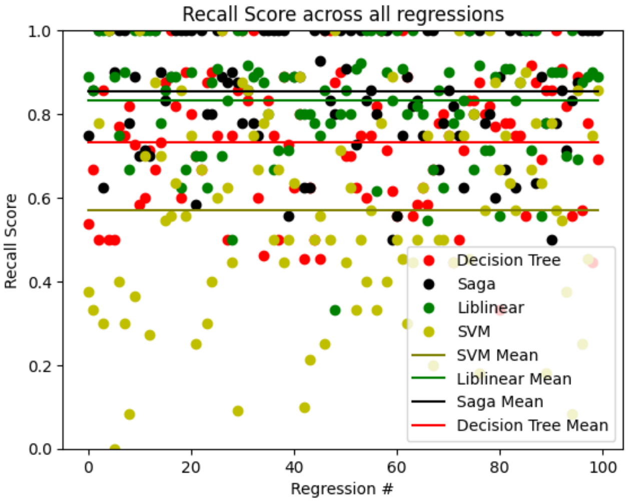


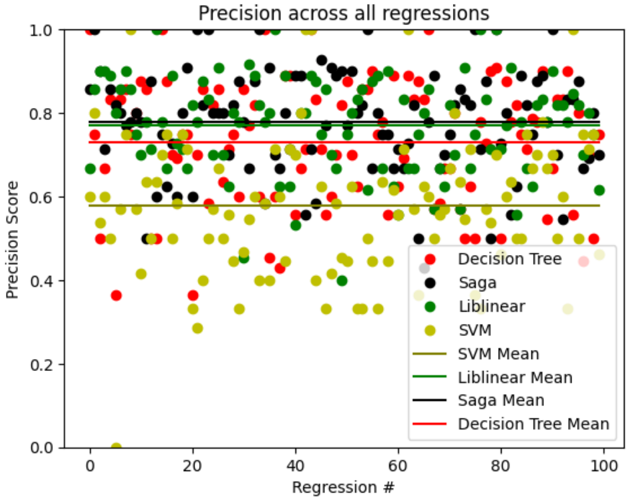
The actual means of each of the values (without Manganese Chloride) are as follows:

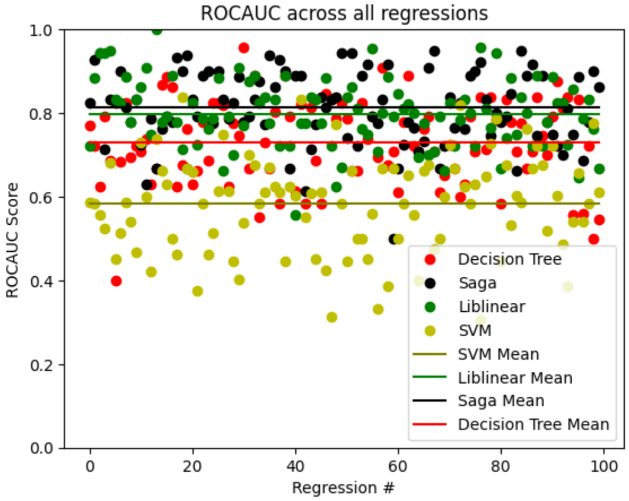
	Liblinear	Saga	Dec. Tree	SVM
Accuracy	0.797777	0.8061111	0.712222	0.5661111
Log Loss	0.620135	0.586015		
Recall	0.821345	0.826359	0.725044	0.569188
Precision	0.787778	0.787074	0.702016	0.598836
AUC	0.796241	0.810348	0.712283	0.595356

With Manganese Chloride included, graphs appear as follows:

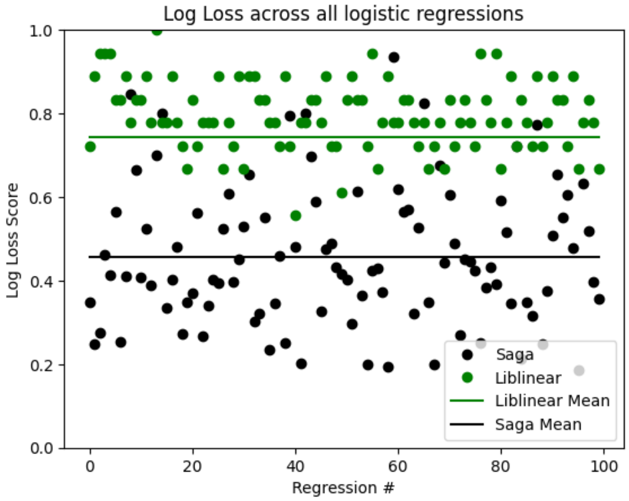








With Manganese Chloride included, the log loss plot is as follows:



The actual means of each of these values (with Manganese Chloride) are as follows:

	Liblinear	Saga	Dec. Tree	SVM
Accuracy	0.7961111	0.807222	0.723888	0.577222
Log Loss	0.742042	0.456110		
Recall	0.831942	0.853019	0.731921	0.545809
Precision	0.769503	0.777166	0.729575	0.620558
AUC	0.797089	0.812504	0.729909	0.601910

As seen by all of the plots and the tables of the means of each metric, the two logistic regression models have the best metrics. They have consistently higher accuracy, recall, precision, and AUC scores. This is present in both cases, with and without a seemingly outlier value of Manganese Chloride. However, when Manganese Chloride is included in the predictions the models make, the average loss of the logistic regression using the saga solver drops dramatically. Without Manganese Chloride, the logistic regressions with both solvers offer very similar results, with the saga solver having just a slightly lower average loss. Thus, logistic regression with the saga solver is deemed the most appropriate model for predicting whether or not a person has Phelan-McDermid Syndrome based upon their metabolic profile.

Outside of the logistic regression models, the decision tree classifier boasts significantly better metrics than the support vector machines and thus is better for predicting the presence of Phelan-McDermid Syndrome. The support vector machine approach has the worst metrics of any of the models.

4. SUMMARY AND CONCLUSION

Our primary goal for this project was to see if there was a significant difference between the control and patient cell lines with the data we retrieved using methods discussed in class. After conducting research, we noticed that the differences between logistic models were not as noticeable or only at a minuscule amount. If experts in this particular field analyzed our work, they would gain the most information from exploring the binary classification we did using logistic regression. With the classification, it identifies whether a patient has PMS or not at a decent accuracy. Manganese Chloride affected the distributions of the data significantly, as seen in the EDA visualizations, and excluding it increased the log loss. With more time with our project, we would like to see which other metabolic pathways affect PMS and if it's more of an outlier that only sometimes can affect a patient or if it is a constant factor in determining if someone has PMS. We also would have wanted to make a more systematic form when it came to our feature selection.