CPSC392 FINAL PROJECT

# HOSPITAL MORTALITY

ARIEL KUO, MCKINLEY PIEPER, JONATHAN VERGONIO

# DATASET: In Hospital Mortality Prediction

**from Kaggle.com**

Our motivation: We wanted to work with data that had meaning and could impact everyday life.

It Is essential to understand what features can lead to hospital mortality and what other diseases could possibly hold more weight in predictions.

# VARIABLES

**What are the predictors? What are we predicting?**

Age, Gender, BMI, Hypertensive, Atrialfibrillation, CHD with no MI, Diabetes, Deficiencyanemias, Depression, Hyperlipemia, Renal failure, COPD, Heart Rate, Systolic blood pressure, Diastolic blood pressure, Respiratory rate, Temperature, SP O2, Urine output, Hematocrit, RBC, MCH, MCHC, MCV, RDW, Leucocyte, Platelets, Neutrophils, Basophils, Lymphocyte, INR, NT-proBNP, Creatine kinase, Creatinine, Urea nitrogen, Glucose, Blood potassium, Blood sodium, Blood calcium,Chloride, Anion gap, Magnesium ion, PH, Bicarbonate, Lactic acid, PCO2, EF

# DATA CLEANING & STANDARDIZING

- Check NULL Values
- Drop them
- Z-scale to make sure all variables are on the same scale

# Question 1

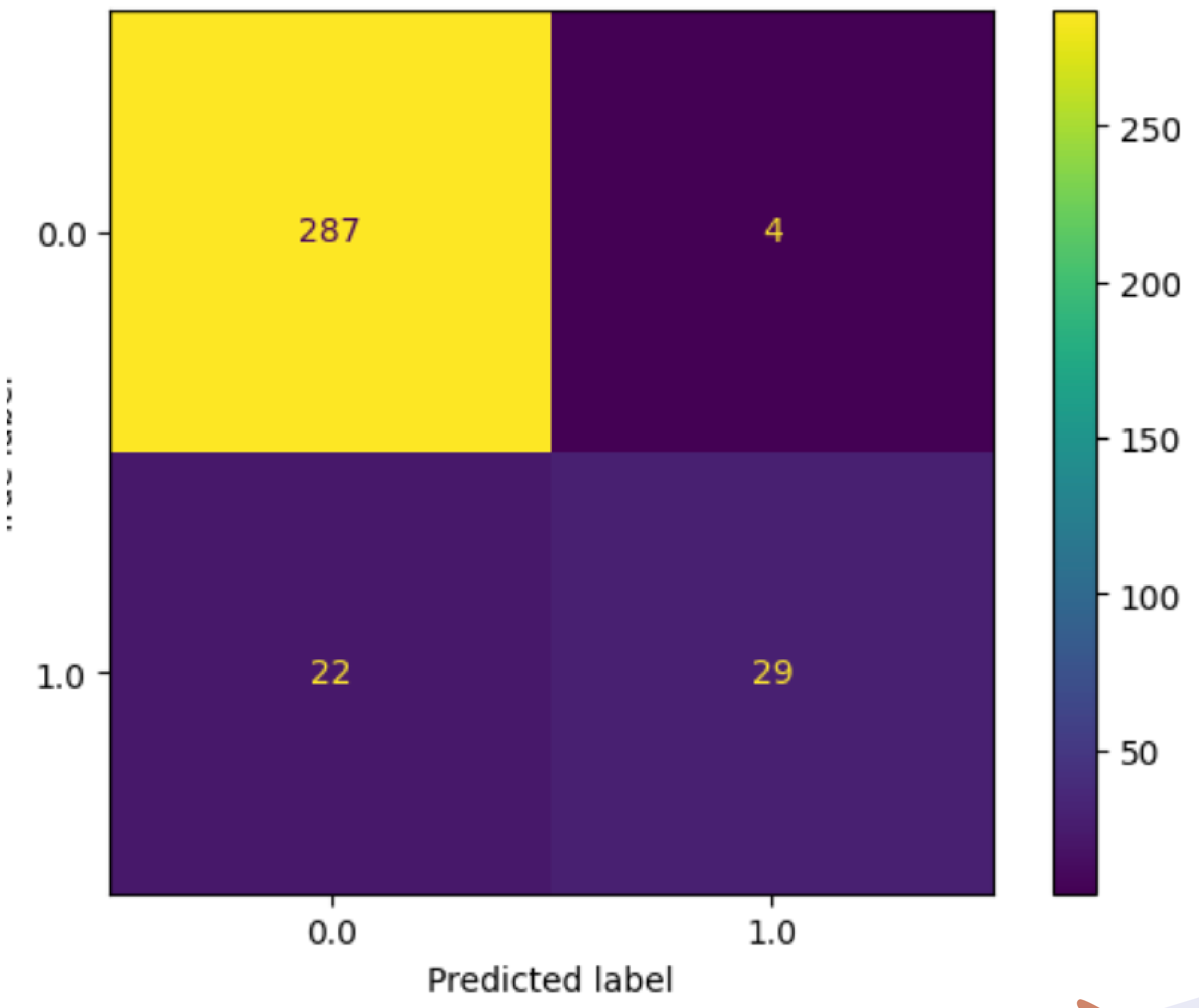
**What are the most common health conditions that lead to in-hospital mortality?**

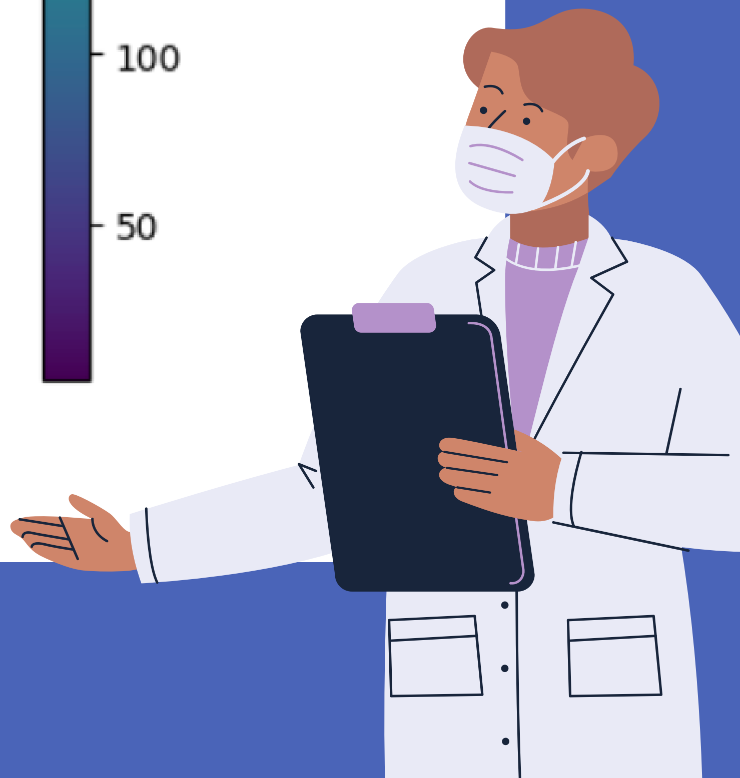
```
# metrics
print("Accuracy: ", accuracy_score(y_test, predictedVals))
print("F1 Score: ", f1_score(y_test, predictedVals))
print("Recall: ", recall_score(y_test, predictedVals))
print("Precision: ", precision_score(y_test, predictedVals))
```

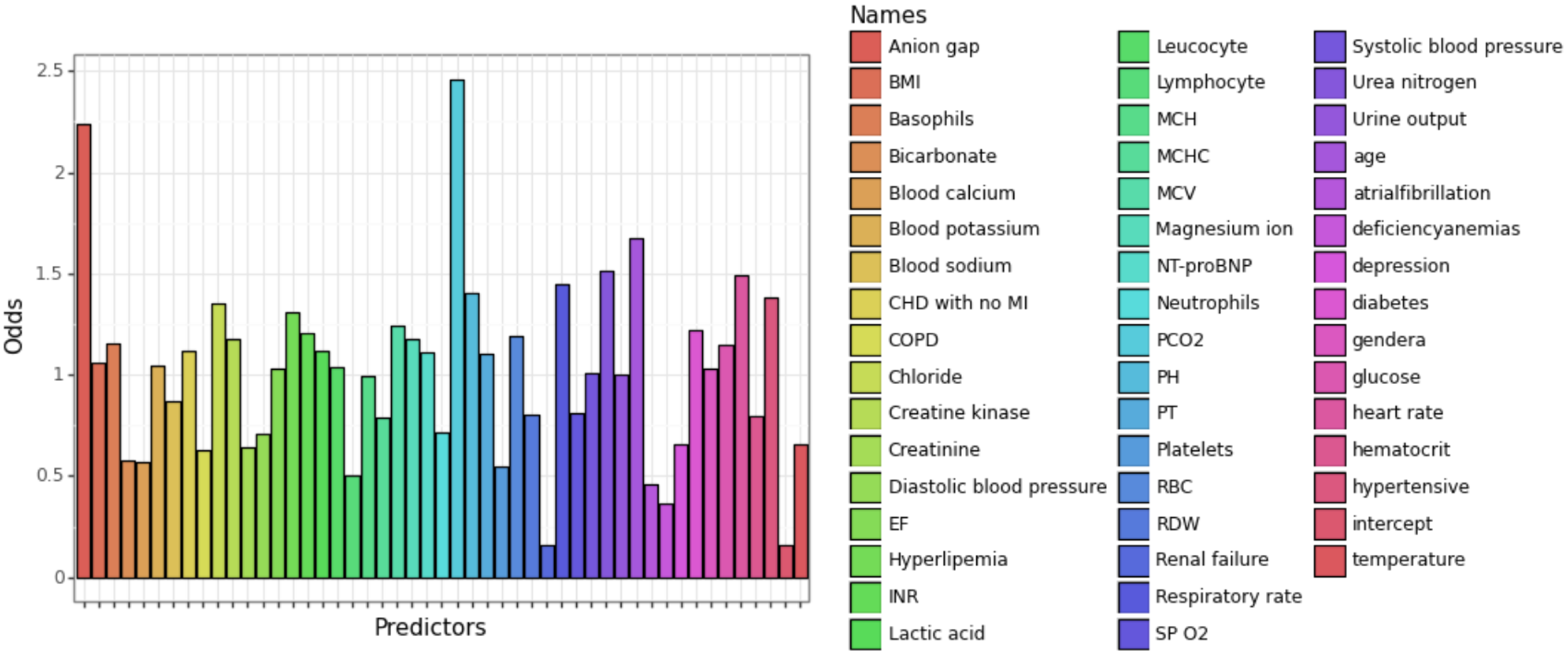
```
Accuracy:  0.8953488372093024
F1 Score:  0.6086956521739131
Recall:  0.5
Precision:  0.7777777777777778
```

```
# metrics
print("Accuracy: ", accuracy_score(y_train, myLogit.predict(X_train)))
print("F1 Score: ", f1_score(y_train, myLogit.predict(X_train)))
print("Recall: ", recall_score(y_train, myLogit.predict(X_train)))
print("Precision: ", precision_score(y_train, myLogit.predict(X_train)
```

```
Accuracy:  0.9239766081871345
F1 Score:  0.6904761904761905
Recall:  0.5686274509803921
Precision:  0.8787878787878788
```

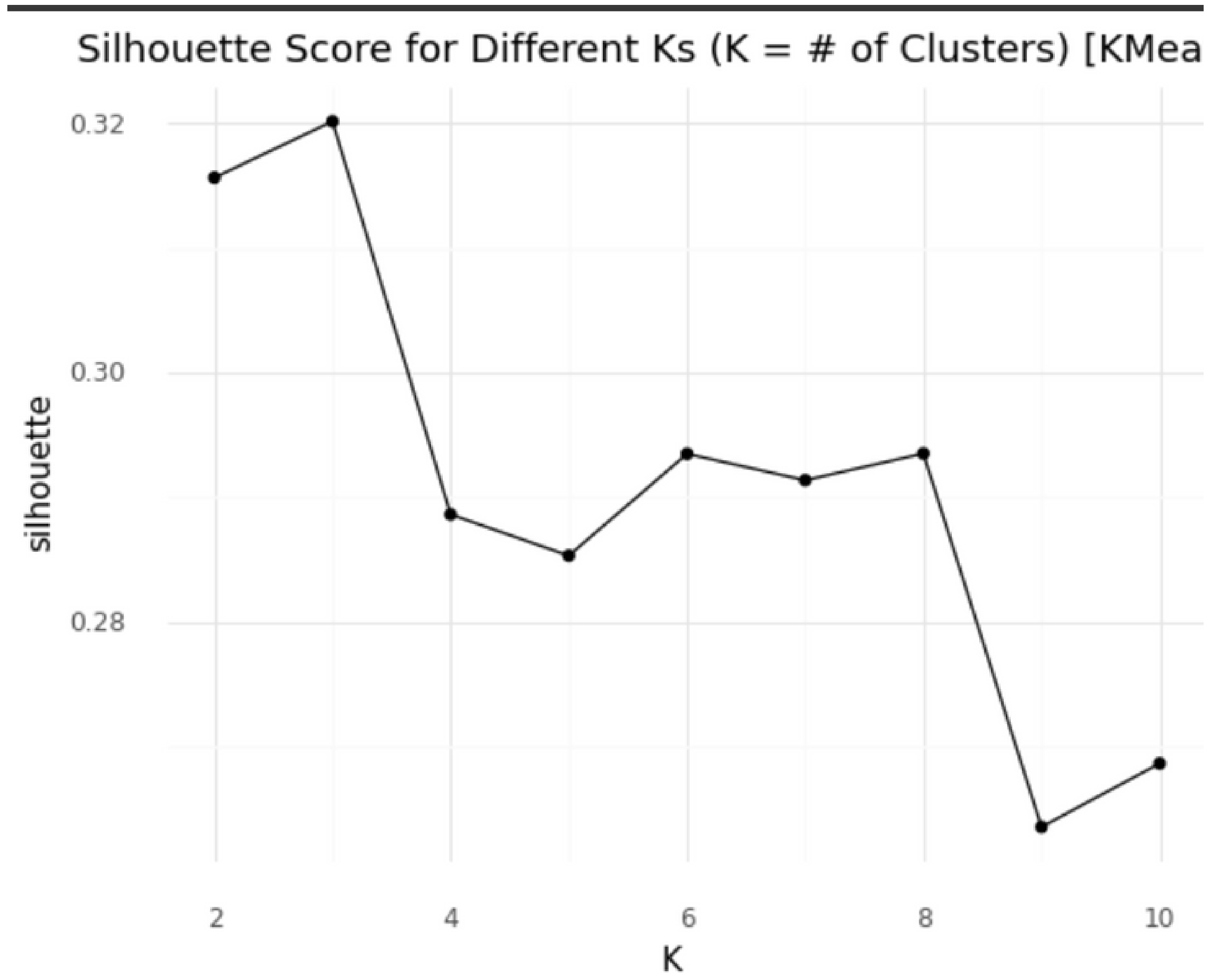


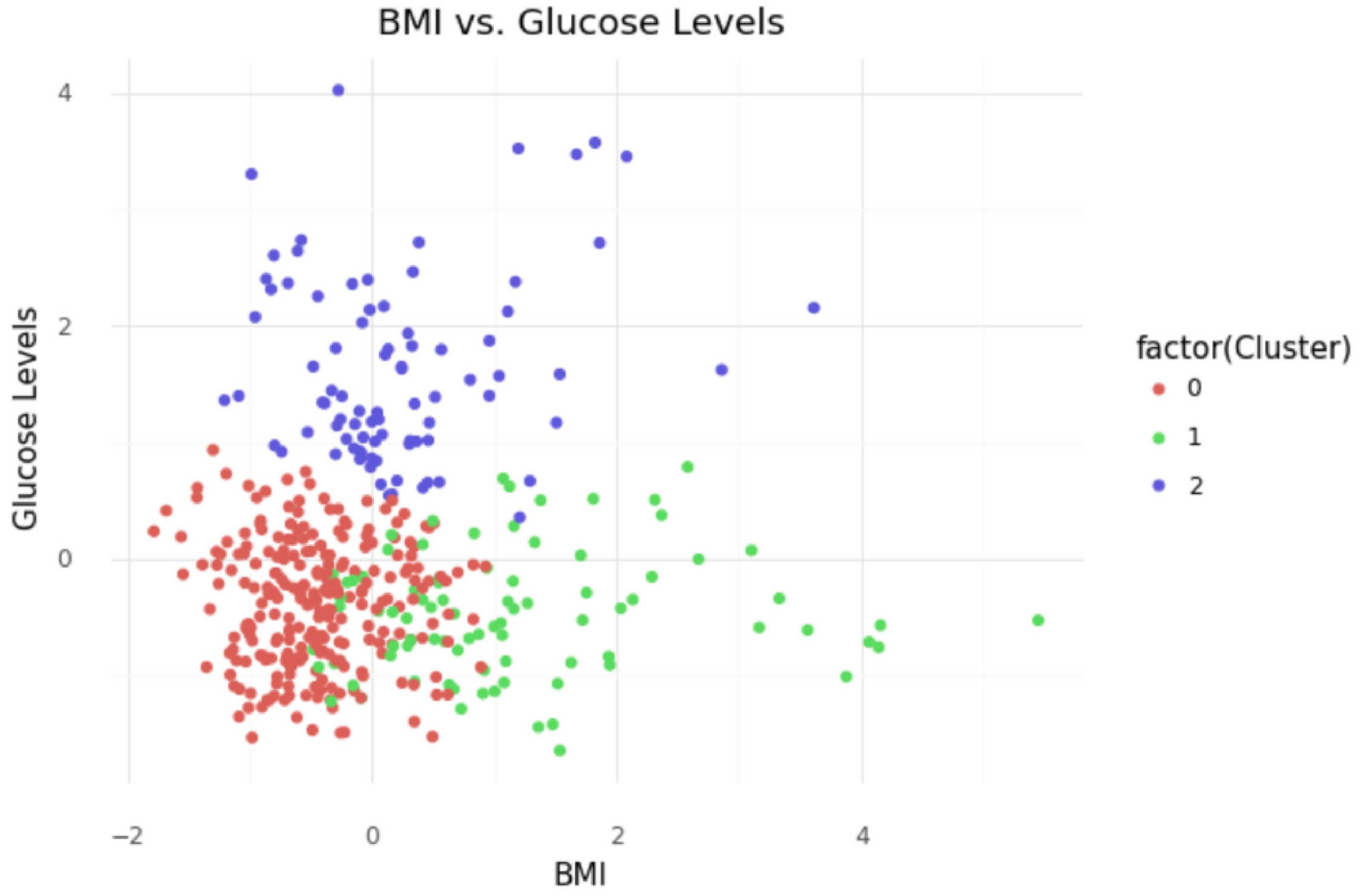




Names
Odds
Predictors
2.5
2
1.5
1
0.5
0
Anion gap
BMI
Basophils
Bicarbonate
Blood calcium
Blood potassium
Blood sodium
CHD with no MI
COPD
Chloride
Creatine kinase
Creatinine
Diastolic blood pressure
EF
Hyperlipemia
INR
Lactic acid
Leucocyte
Lymphocyte
MCH
MCHC
MCV
Magnesium ion
NT-proBNP
Neutrophils
PCO2
PH
PT
Platelets
RBC
RDW
Renal failure
Respiratory rate
SP O2
Systolic blood pressure
Urea nitrogen
Urine output
age
atrialfibrillation
deficiencyanemias
depression
diabetes
gendera
glucose
heart rate
hematocrit
hypertensive
intercept
temperature

# Question 2

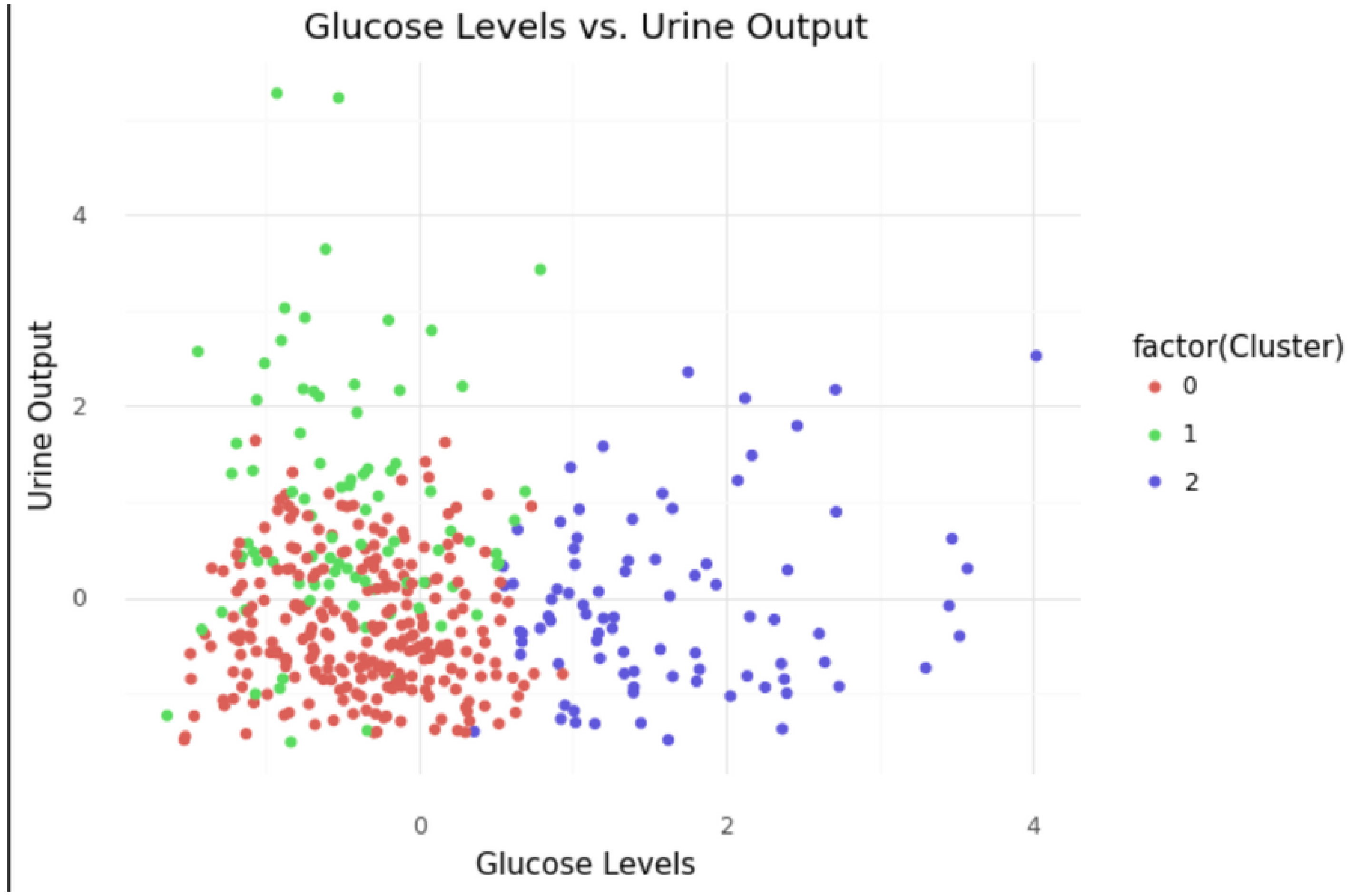
**When considering features BMI, Glucose, and Urine Output, what clusters may emerge and how can we characterize those clusters?**

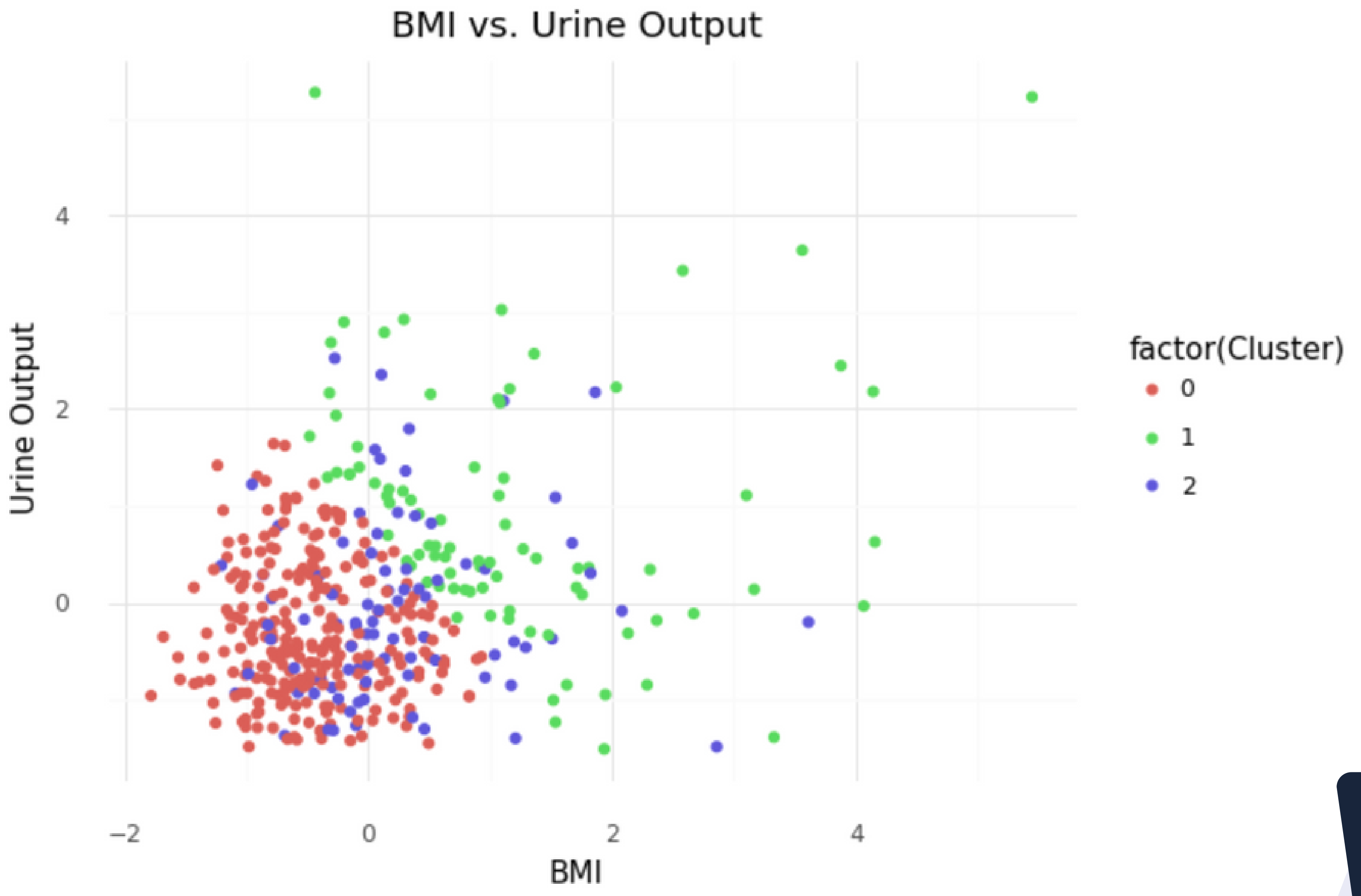


# Question 2

**When considering features BMI, Glucose, and Urine Output, what clusters may emerge and how can we characterize those clusters?**




# Question 3

**Are there any differences in the predictive performance of our model across different subgroups of patients, such as patients who are diagnosed with depression, or comorbidity status (renal failure, diabetes, hypertensive)?**

## Accuracy Scores:

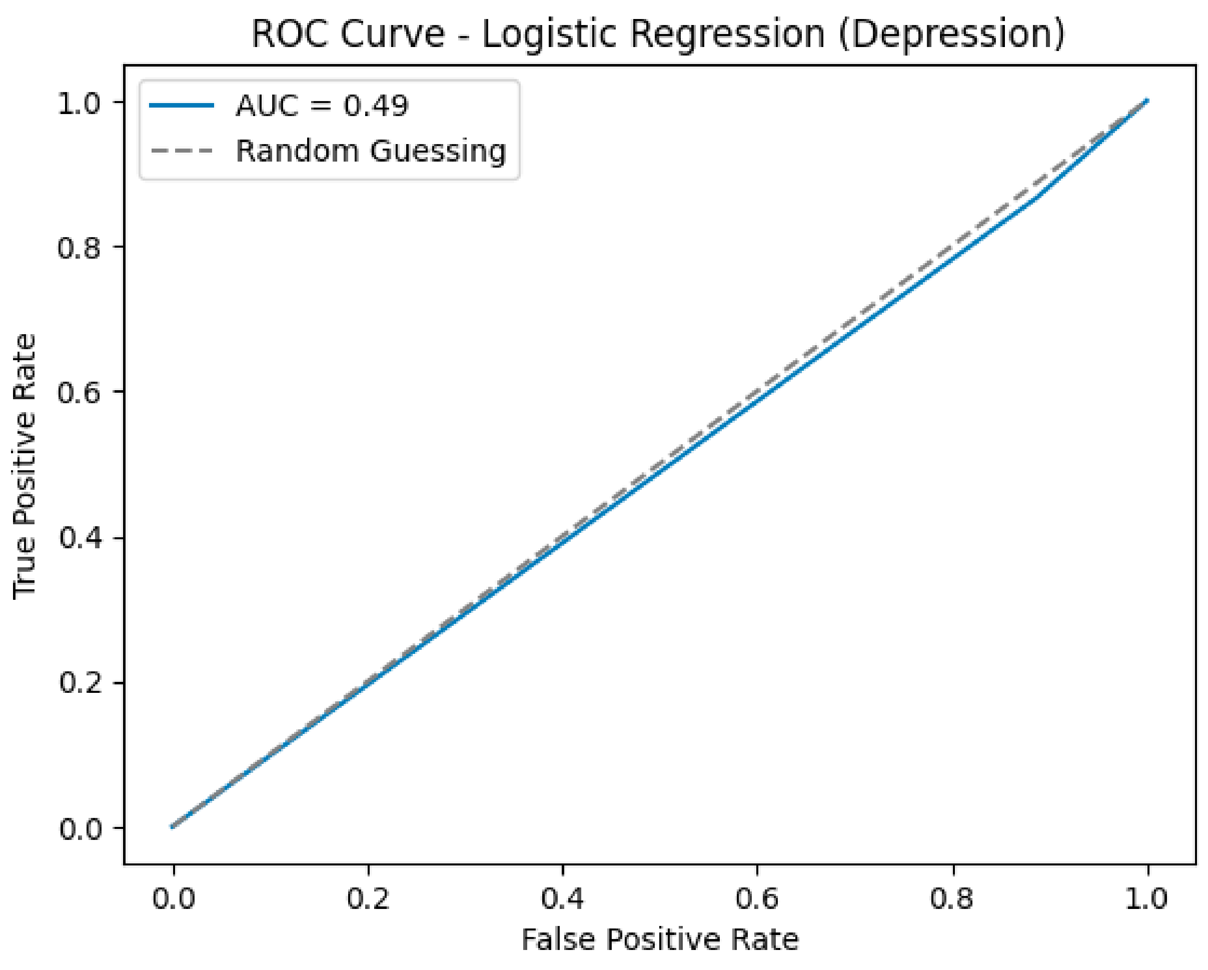
**Accuracy (Depression): 0.8255**
**Accuracy (Renal Failure): 0.8605**
**Accuracy (Diabetes): 0.8604**
**Accuracy (Hypertensive): 0.8372**
**Accuracy (All): 0.8139**

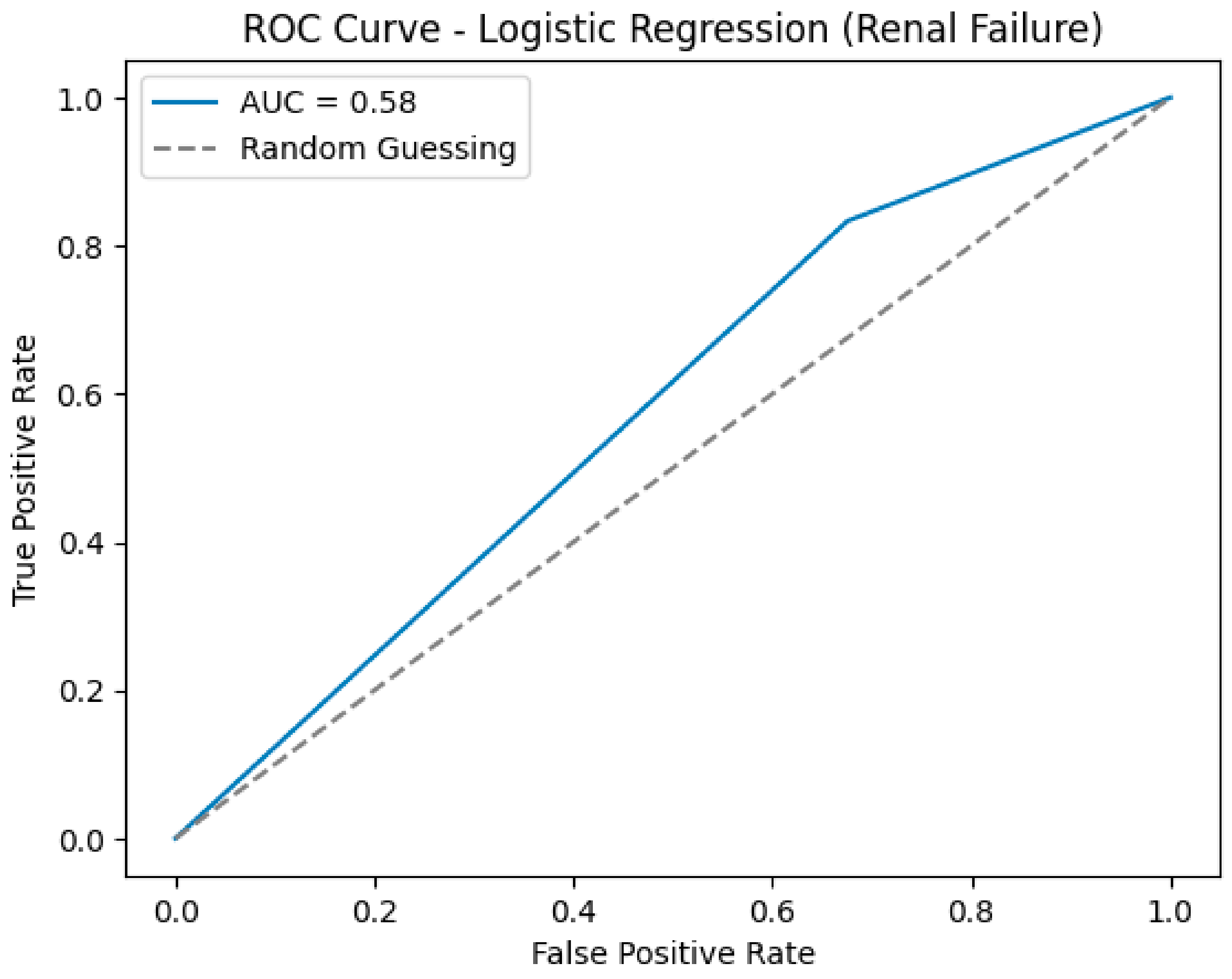
## Odds Coefficients:

| | Coef | Names | Odds |
|---|---|---|---|
| 0 | -1.232983 | depression | 0.291422 |
| 1 | -0.741782 | Renal failure | 0.476264 |
| 2 | -0.005079 | diabetes | 0.994934 |
| 3 | -0.004830 | hypertensive | 0.995182 |
| 4 | -1.444756 | intercept | 0.235804 |

# Question 3

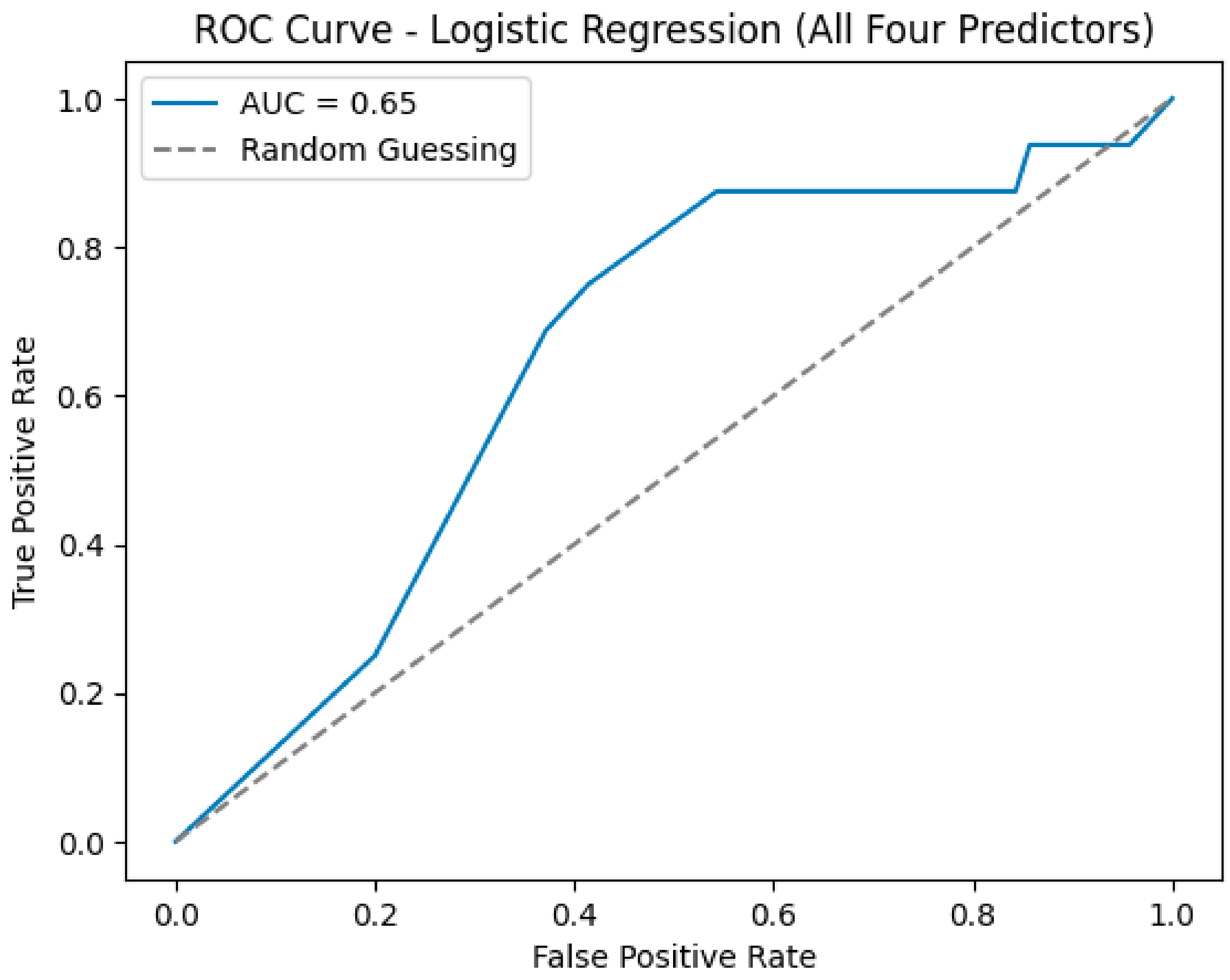
**Are there any differences in the predictive performance of our model across different subgroups of patients, such as patients who are diagnosed with depression, or comorbidity status (renal failure, diabetes, hypertensive)?**



# Question 3

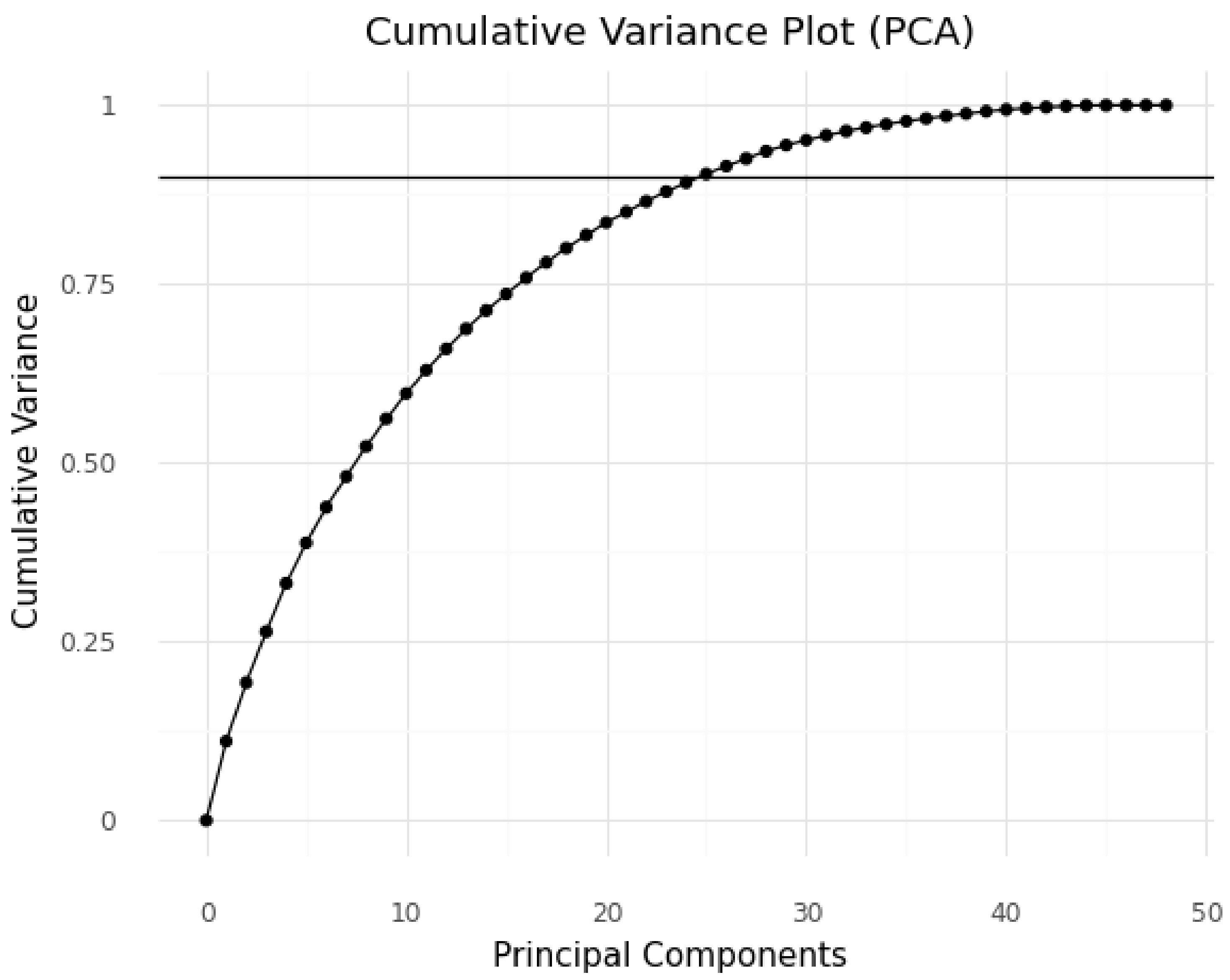
**Are there any differences in the predictive performance of our model across different subgroups of patients, such as patients who are diagnosed with depression, or comorbidity status (renal failure, diabetes, hypertensive)?**




# Question 4

**How does the mean absolute error differ between the train and test data when using Principle Component Analysis on all continuous variables, and retaining enough Principle Components to keep 90% of the variance, to predict hospital mortality with our model(s)?**

| | expl_var | pc | cum_var |
|---|---|---|---|
| 0 | 0.110861 | 1 | 0.110861 |
| 1 | 0.081857 | 2 | 0.192718 |
| 2 | 0.071281 | 3 | 0.263999 |
| 3 | 0.067548 | 4 | 0.331547 |
| 4 | 0.056565 | 5 | 0.388113 |




**Using 25 PCs for 90% variance**

# Question 4

How does the mean absolute error differ between the train and test data when using Principle Component Analysis on all continuous variables, and retaining enough Principle Components to keep 90% of the variance, to predict hospital mortality with our model(s)?

**Logistic Regression Results:**

**Accuracy for Original Train Set: 0.9239**

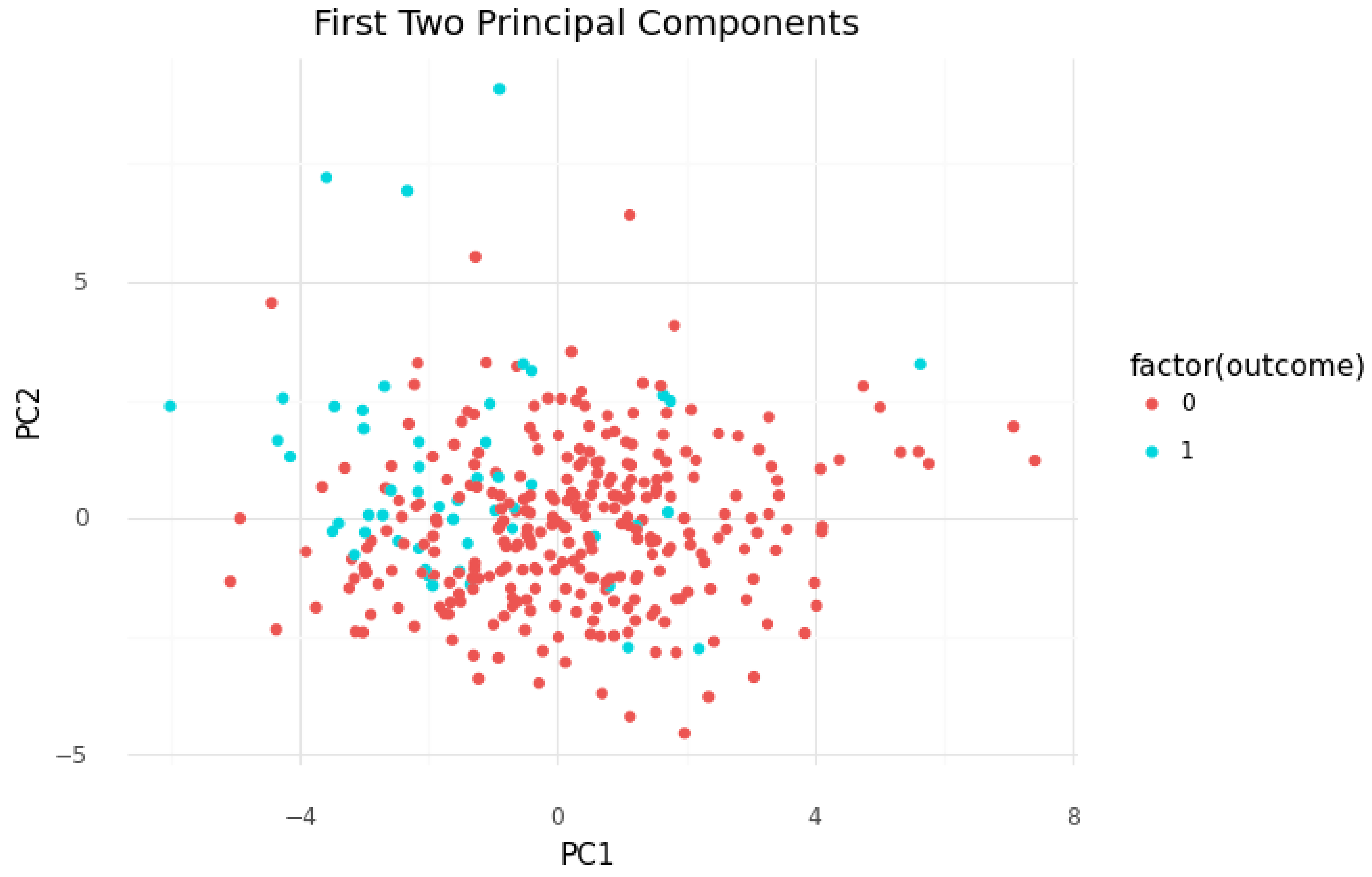
**Accuracy for Original Test Set: 0.8953**

**Accuracy for PCA Train Set: 0.8976**

**Accuracy for PCA Test Set: 0.8721**

# Question 4

**How does the mean absolute error differ between the train and test data when using Principle Component Analysis on all continuous variables, and retaining enough Principle Components to keep 90% of the variance, to predict hospital mortality with our model(s)?**




# Question 5

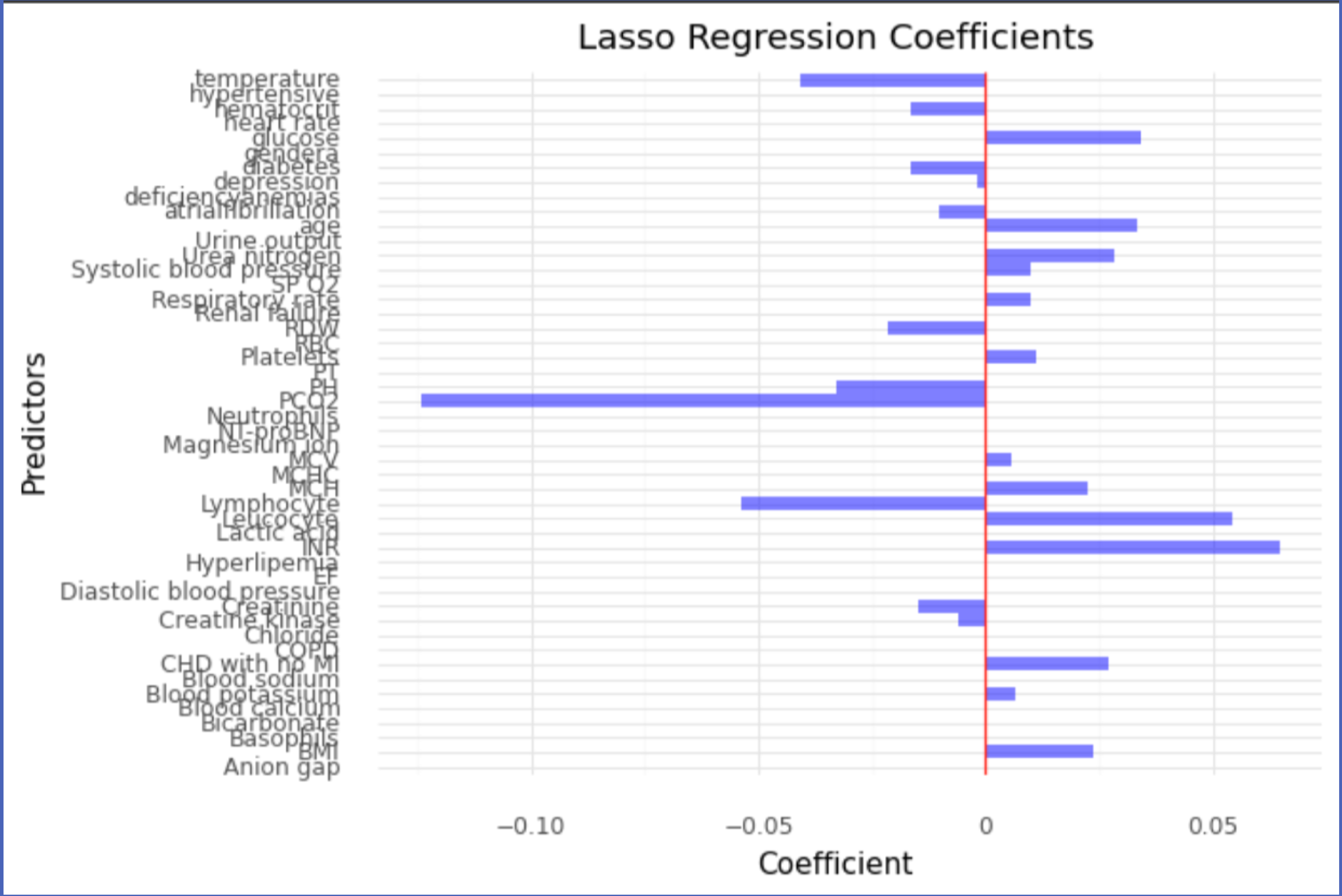
**Out of the 48 variables observed in the dataset (excluding ground and ID), how can we utilize the Regularization method LASSO to select which variables have the most impact in predicting hospital mortality?**

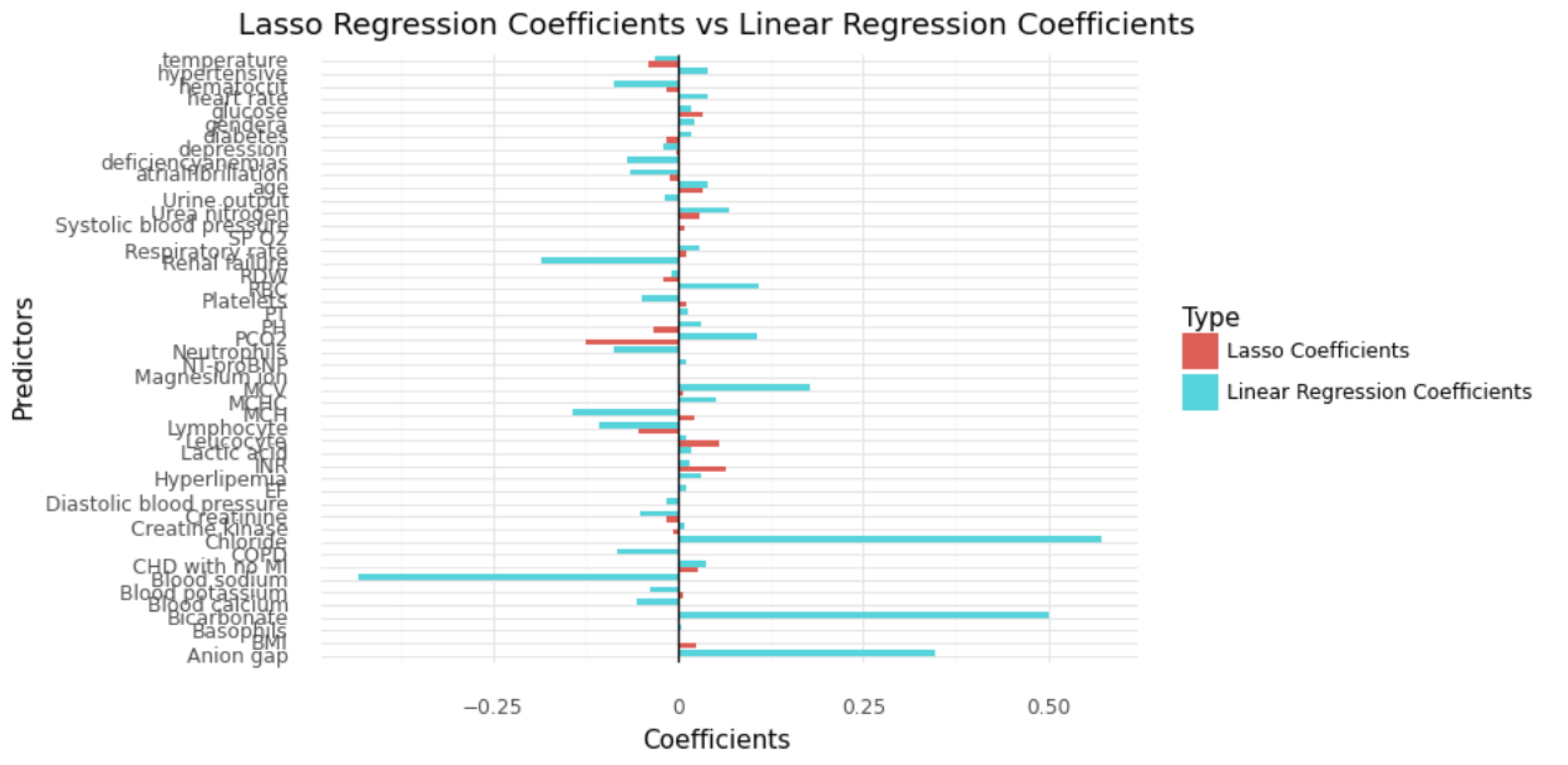
| | Coefficients | Predictors |
|---|---|---|
| 0 | 0.033263 | age |
| 1 | 0.000000 | gendera |
| 2 | 0.023522 | BMI |
| 3 | -0.000000 | hypertensive |
| 4 | -0.010378 | atrialfibrillation |
| 5 | 0.027015 | CHD with no MI |
| 6 | -0.016475 | diabetes |
| 7 | -0.000000 | deficiencyanemias |
| 8 | -0.002012 | depression |
| 9 | 0.000000 | Hyperlipemia |
| 10 | 0.000000 | Renal failure |
| 11 | 0.000000 | COPD |
| 12 | -0.000000 | heart rate |
| 13 | 0.009652 | Systolic blood pressure |
| 14 | 0.000000 | Diastolic blood pressure |
| 15 | 0.009747 | Respiratory rate |
| 16 | -0.041035 | temperature |
| 17 | -0.000000 | SP O2 |
| 18 | 0.000000 | Urine output |
| 19 | -0.016460 | hematocrit |
| 20 | 0.000000 | RBC |
| 21 | 0.022599 | MCH |
| 22 | 0.000355 | MCHC |
| 23 | 0.005712 | MCV |
| 24 | -0.021458 | RDW |
| 25 | 0.054432 | Leucocyte |
| 26 | 0.011131 | Platelets |
| 27 | 0.000000 | Neutrophils |
| 28 | 0.000000 | Basophils |
| 29 | -0.053990 | Lymphocyte |
| 30 | 0.000000 | PT |
| 31 | 0.064715 | INR |
| 32 | 0.000000 | NT-proBNP |
| 33 | -0.005930 | Creatine kinase |
| 34 | -0.014864 | Creatinine |
| 35 | 0.028233 | Urea nitrogen |
| 36 | 0.034051 | glucose |
| 37 | 0.006354 | Blood potassium |
| 38 | 0.000000 | Blood sodium |
| 39 | 0.000000 | Blood calcium |
| 40 | -0.000000 | Chloride |
| 41 | 0.000000 | Anion gap |
| 42 | 0.000000 | Magnesium ion |
| 43 | -0.033004 | PH |
| 44 | -0.000000 | Bicarbonate |
| 45 | 0.000000 | Lactic acid |
| 46 | -0.124233 | PCO2 |
| 47 | -0.000000 | EF |

```
PENALTY: 0.01
TRAIN: 0.3396888666530895
TEST : 0.3918317430757958
```

**Predictors that got shrunk to 0:**

- gender
- hypertensive
- Hyperlipemia
- Renal failure
- COPD
- heart rate
- Diastolic blood pressure
- SP O2
- Urine output
- RBC
- Neutrophils
- Basophils
- PT
- NT-proBNP
- Blood sodium
- Blood calcium
- Chloride
- Anion gap
- Magnesium ion
- Bicarbonate
- Lactic acid
- EF

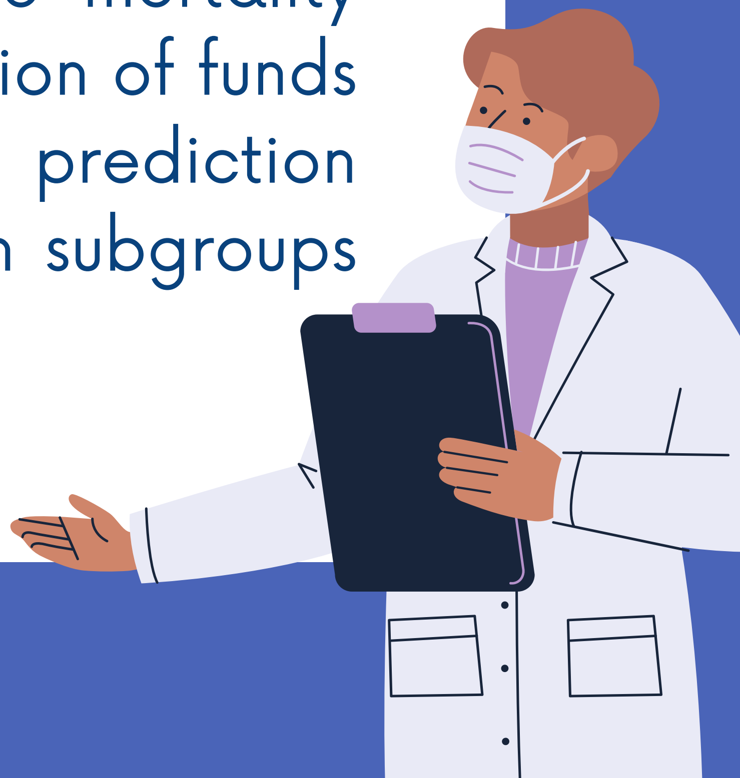

Lasso Regression Coefficients
temperature
hypertensive
hematocrit
heart rate
glucose
gendera
diabetes
depression
deficiencyanemias
atrialfibrillation
age
Urine output
Urea nitrogen
Systolic blood pressure
SP O2
Respiratory rate
Renal failure
RDW
RBC
Platelets
PT
PH
PCO2
Neutrophils
NT-proBNP
Magnesium ion
MCV
MCHC
MCH
Lymphocyte
Leucocyte
Lactic acid
INR
Hyperlipemia
EF
Diastolic blood pressure
Creatinine
Creatine kinase
Chloride
COPD
CHD with no MI
Blood sodium
Blood potassium
Blood calcium
Bicarbonate
Basophils
BMI
Anion gap
Predictors
−0.10
−0.05
0
0.05
Coefficient

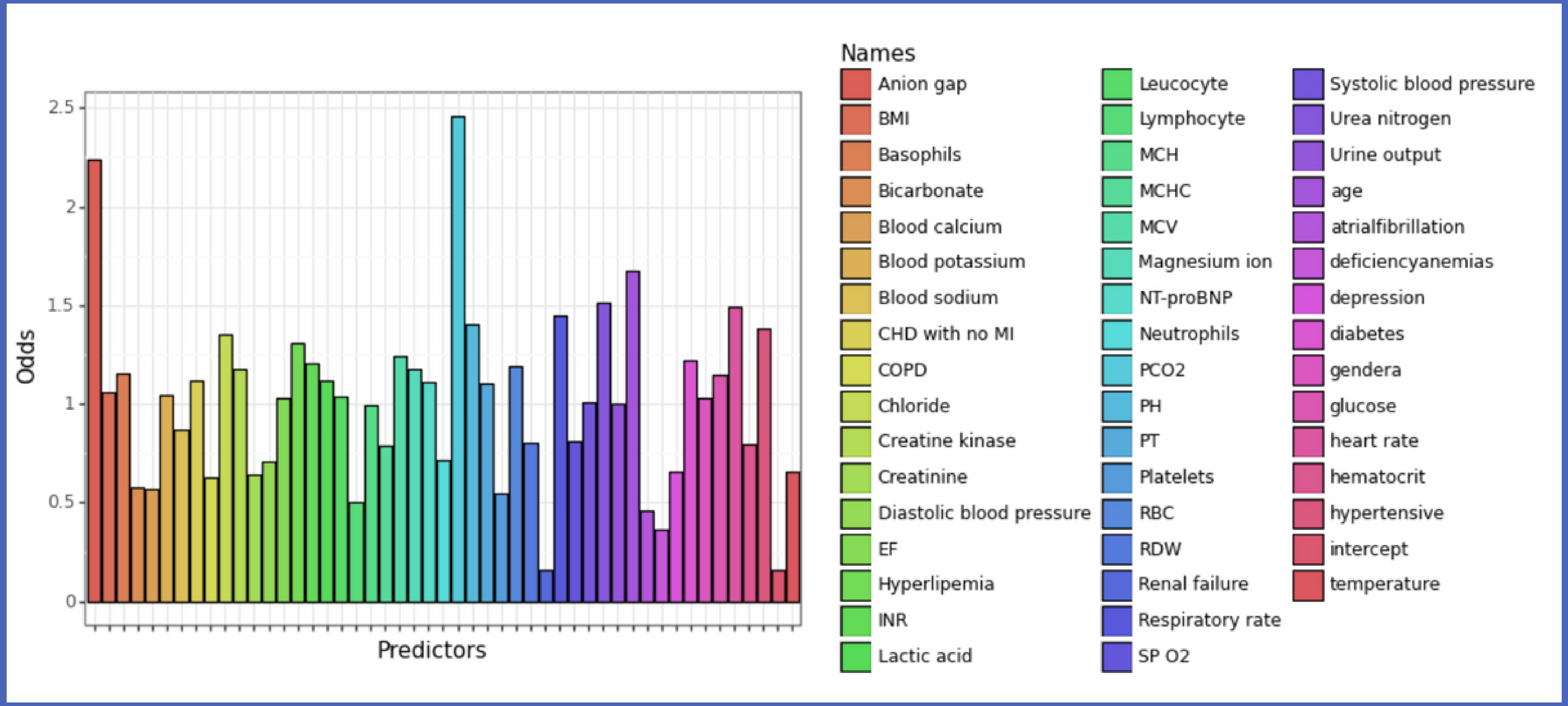

Lasso Regression Coefficients vs Linear Regression Coefficients
temperature
hypertensive
hematocrit
heart rate
glucose
gendera
diabetes
depression
deficiencyanemias
atrialfibrillation
age
Urine output
Urea nitrogen
Systolic blood pressure
SP O2
Respiratory rate
Renal failure
RDW
RBC
Platelets
PT
PH
PCO2
Neutrophils
NT-proBNP
Magnesium ion
MCV
MCHC
MCH
Lymphocyte
Leucocyte
Lactic acid
INR
Hyperlipemia
EF
Diastolic blood pressure
Creatinine
Creatine kinase
Chloride
COPD
CHD with no MI
Blood sodium
Blood potassium
Blood calcium
Bicarbonate
Basophils
BMI
Anion gap
Predictors
−0.25
0
0.25
0.50
Coefficients
Type
Lasso Coefficients
Linear Regression Coefficients

# Question 6

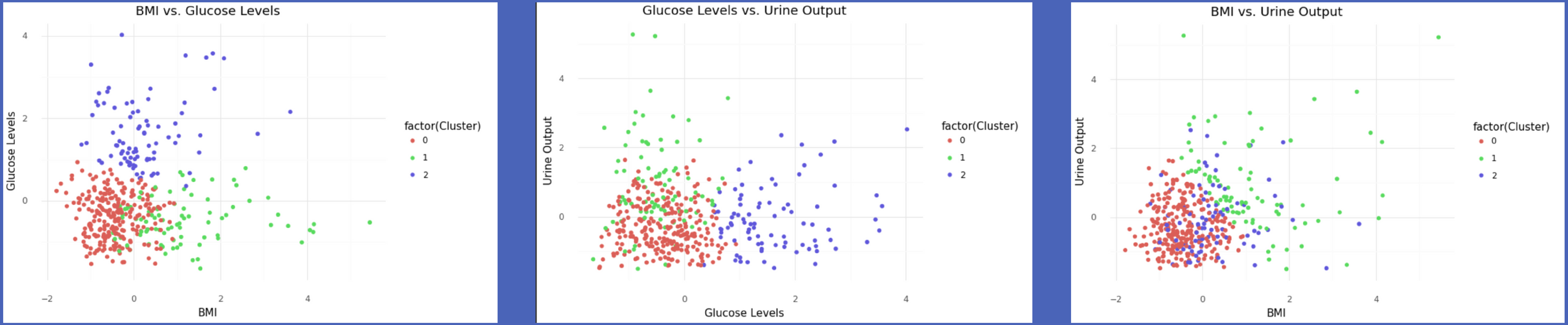
**How can healthcare providers and policymakers use the results of this study to improve healthcare outcomes and reduce costs?**

A model that could predict hospital mortality could be greatly beneficial for multiple parties, including healthcare providers, policymakers, and patients. First and foremost, with the prediction model, healthcare providers could easily identify patients that are at higher risk. This could allow the providers to pay extra attention to these high-risk patients and provide interventions and treatment at an early stage. Additionally, the ability to identify patients could also help healthcare providers to reduce costs and allocate resources, such as doctor/nurse staffing, medicine, and hospital equipment more efficiently. On the other hand, for policymakers, with the mortality prediction, policymakers could also utilize the results to make decisions on the distribution of funds to hospitals and healthcare providers. Moreover, policymakers could also apply the prediction results to conduct research and push for the demand to develop medicine to certain subgroups facing various medical conditions or diseases.






For example, we could use the logistic regression model from our first response to monitor the patients. According to the coefficient graph from our first response, we can see that feature "**PC O2**" is one of the health conditions that increase the odds of mortality. With that said, healthcare providers could tag those patients with these health conditions as the model indicates that they are at higher risk. In addition, with the predicted data, hospitals could delegate staffing and provide interventions/treatment more efficiently.




Another example, healthcare providers or policymakers could use the clustering model from our second response to study and develop pharmaceuticals for particular subgroups within the patients. Based on the clustering graphs from our second response, we can characterize the patients and classify them into different groups. With the classification, providers and policymakers can dive into patient groups, research their conditions for further purposes, and distribute resources and funds.