Statistical Machine Learning: Platelet Count and Strokes

By Emily Neuman, Marshall Roll, Kenny Nhan

Goals and Methods

1

Regression

OLS and LASSO modeling to **predict platelet levels**, which are related to stroke events

2

Classification

Logistic LASSO and a decision tree to **predict stroke events**

3

Clustering

Hierarchical clustering to **uncover patterns underlying stroke events**

01

Platelets

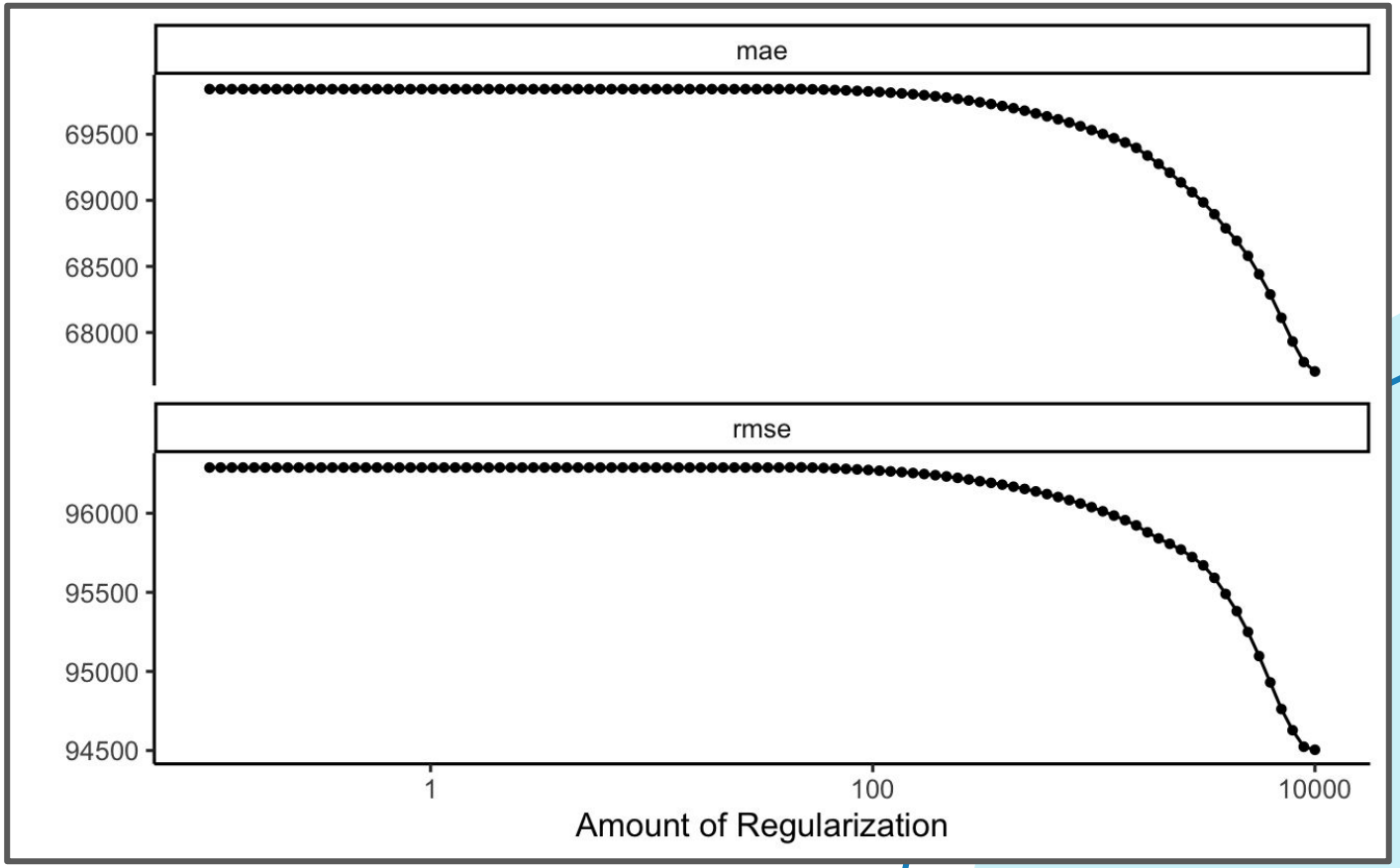
- Age
- Creatinine Phosphokinase
- Ejection Fraction
- **Platelets**
- Serum Creatinine
- Serum Sodium
- Diabetes
- High Blood Pressure
- Sex
- Smoking
- Anaemia

Regression

Research Question

Which of the available biological factors are most accurate in **predicting platelet levels**, which are related to stroke events?

Model Type	RMSE	Ease of Interpretability
OLS	96307	Low
LASSO	95305	High
GAM	95478	Low


mae
69500
69000
68500
68000
rmse
96000
95500
95000
94500
1
100
10000
Amount of Regularization

LASSO Takeaways

- Model shrinks all variables to zero except for **biological sex**
- All linear regression models **perform weakly**, but the LASSO is the most interpretable
- **More biological predictors** needed to determine platelet levels

Classification

Research Question

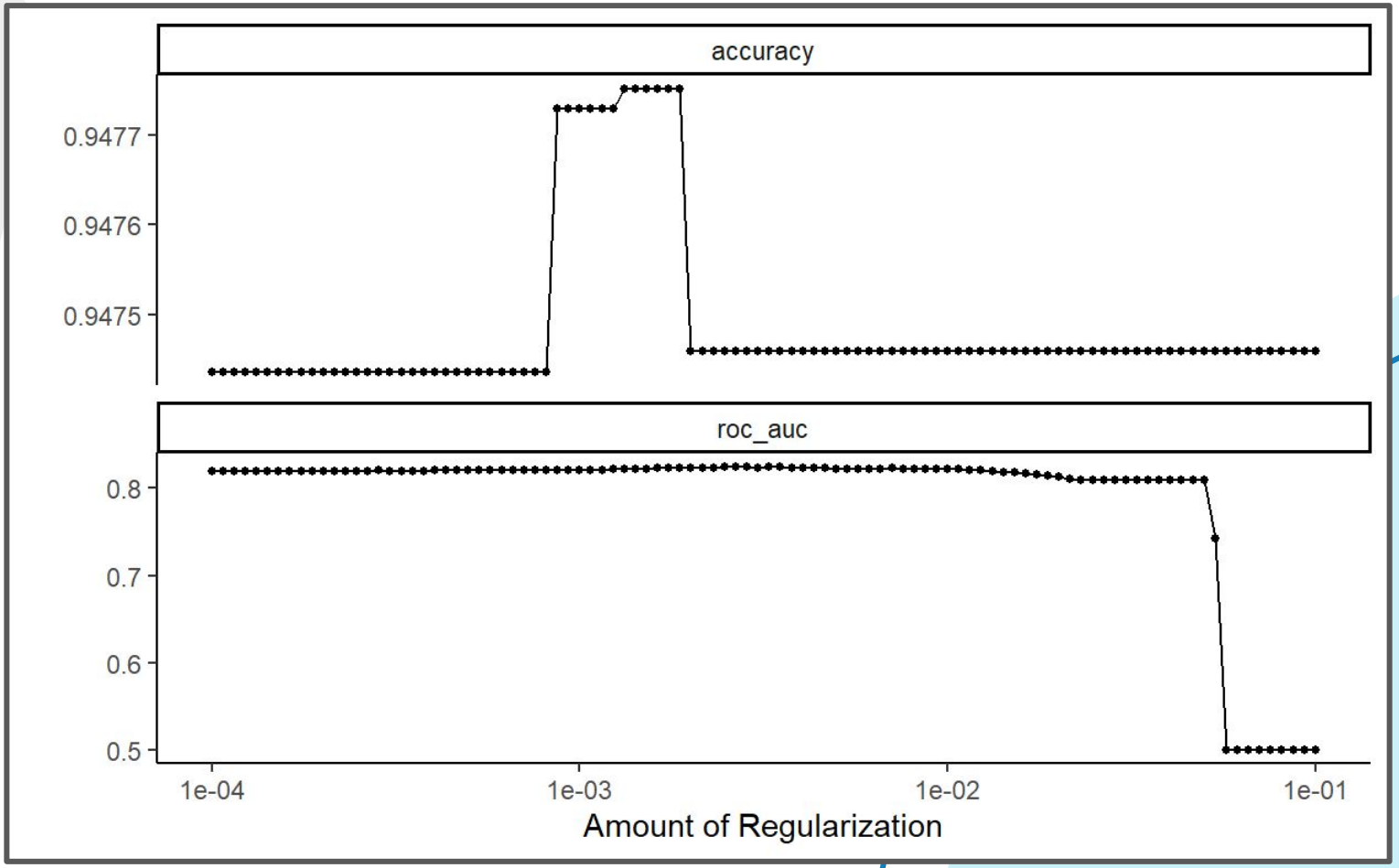
Can we predict whether or not a patient will have a stroke based on certain characteristics?

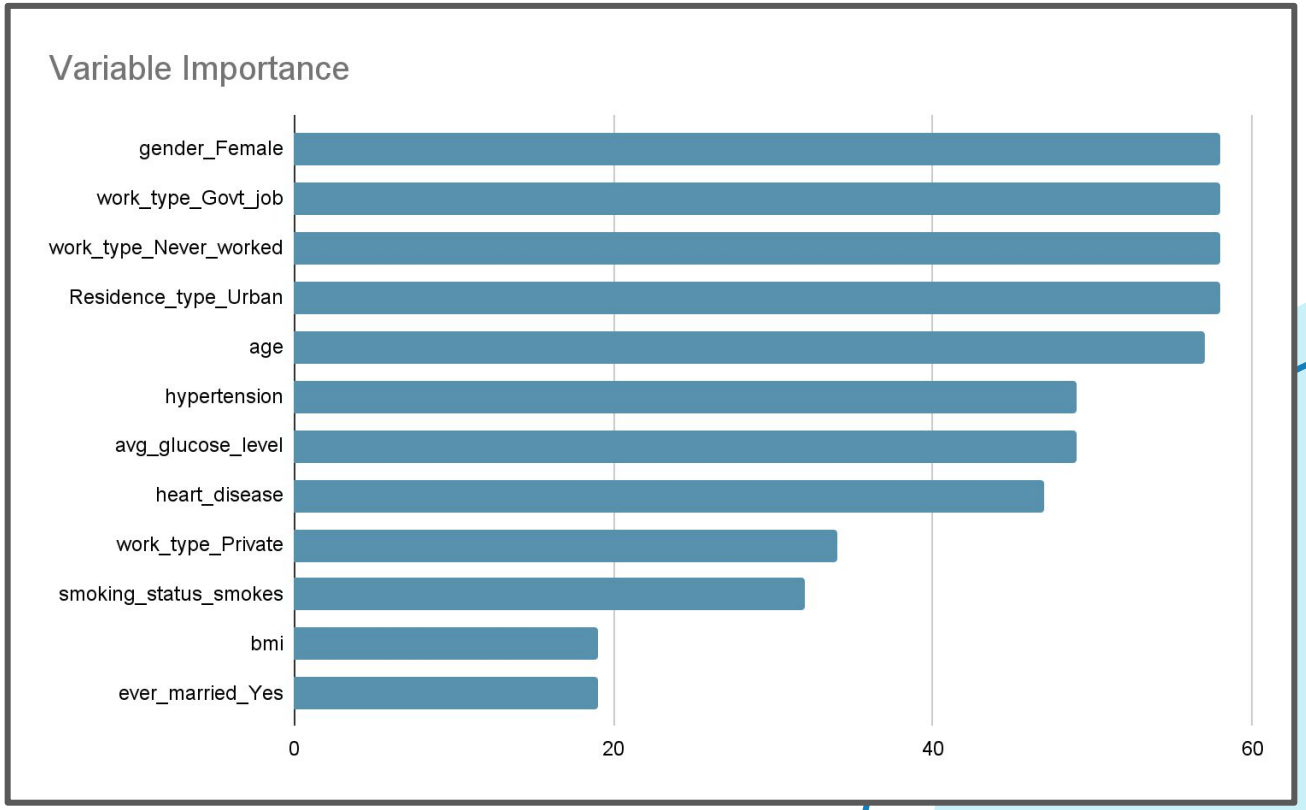
02

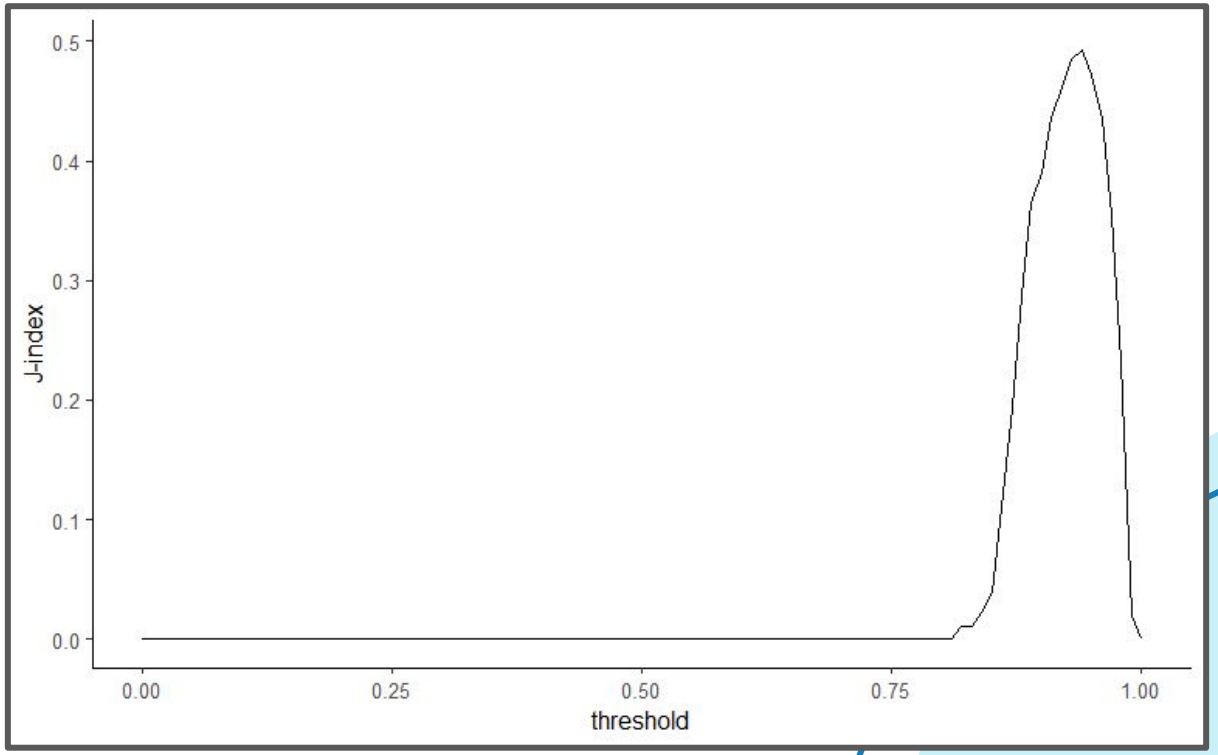
Stroke

- Age
- Average Glucose Level
- BMI
- Gender
- Hypertension
- Heart Disease
- Ever Married
- Work Type
- Residence Type
- Smoking Status
- Stroke

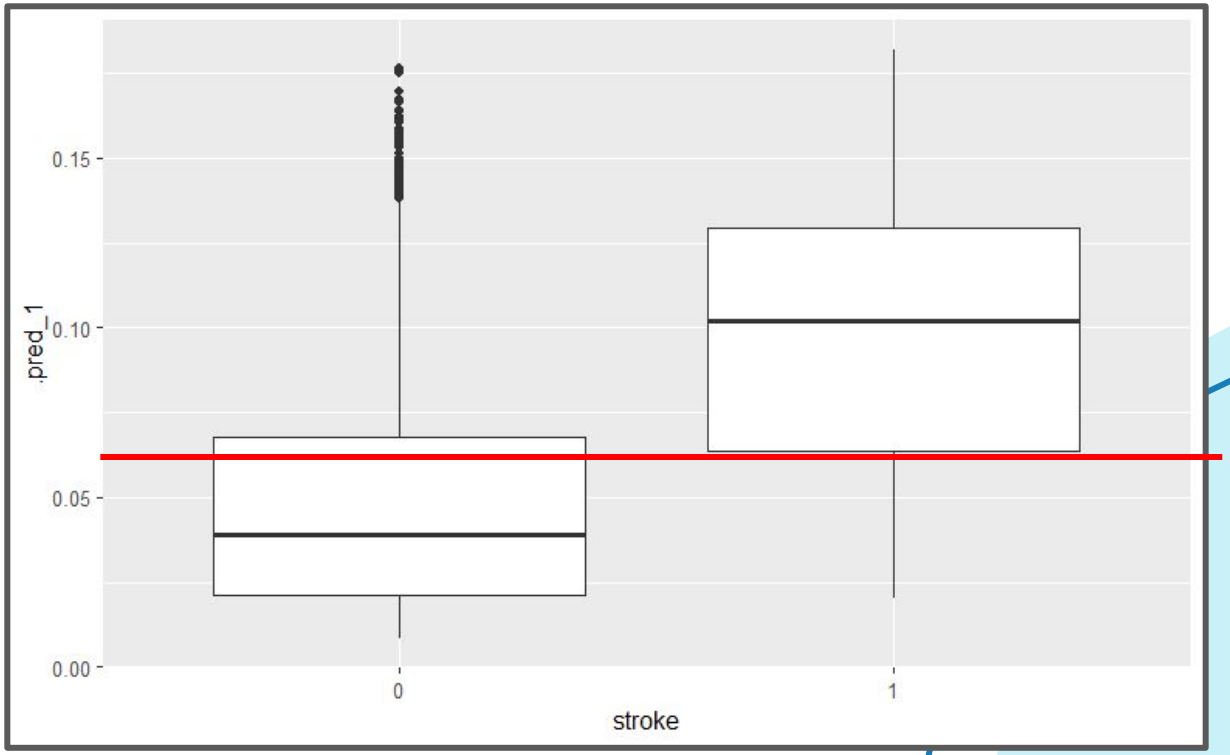
Logistic LASSO


accuracy
0.9477
0.9476
0.9475
roc_auc
0.8
0.7
0.6
0.5
1e-04
1e-03
1e-02
1e-01
Amount of Regularization


Variable Importance
gender_Female
work_type_Govt_job
work_type_Never_worked
Residence_type_Urban
age
hypertension
avg_glucose_level
heart_disease
work_type_Private
smoking_status_smokes
bmi
ever_married_Yes
0
20
40
60

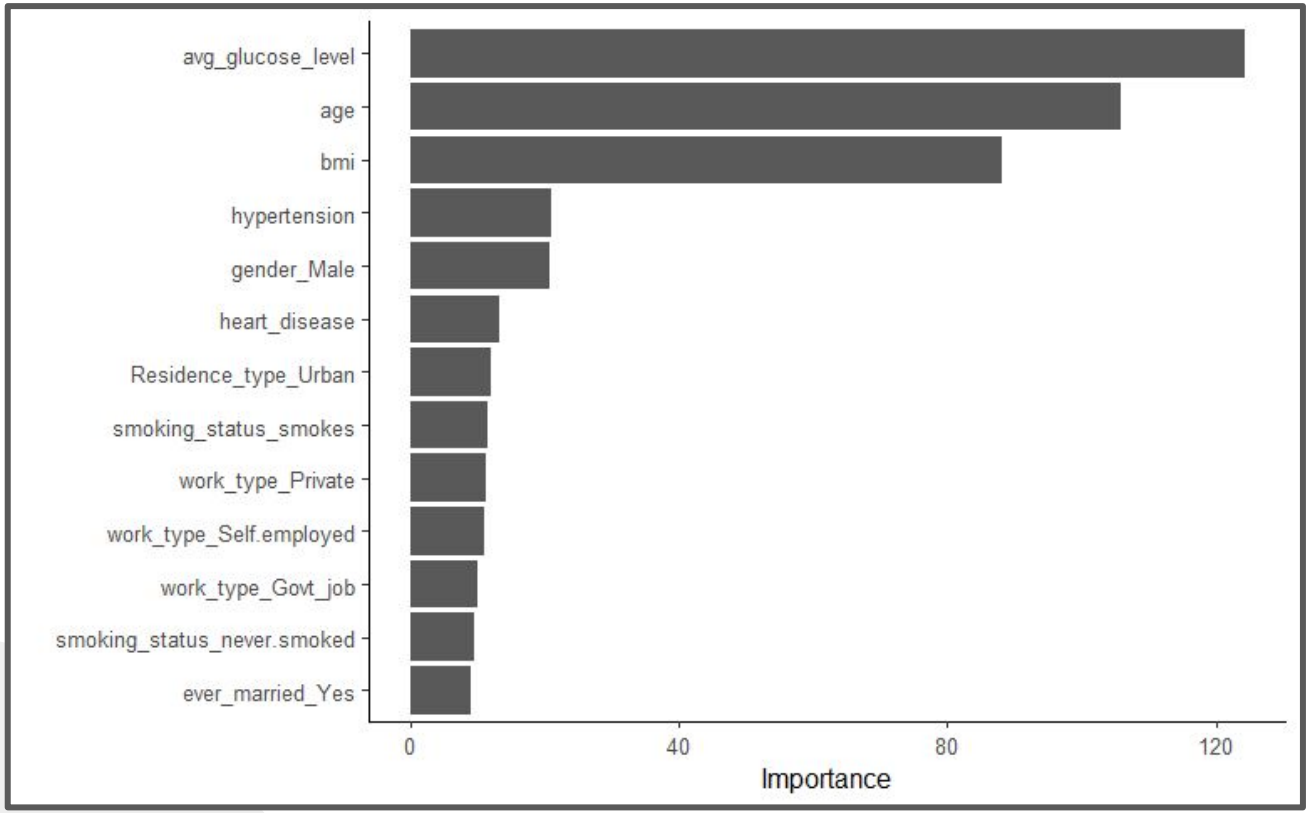


Accuracy	0.70
Sensitivity	0.78
Specificity	0.70



Threshold: 0.94

Decision Tree

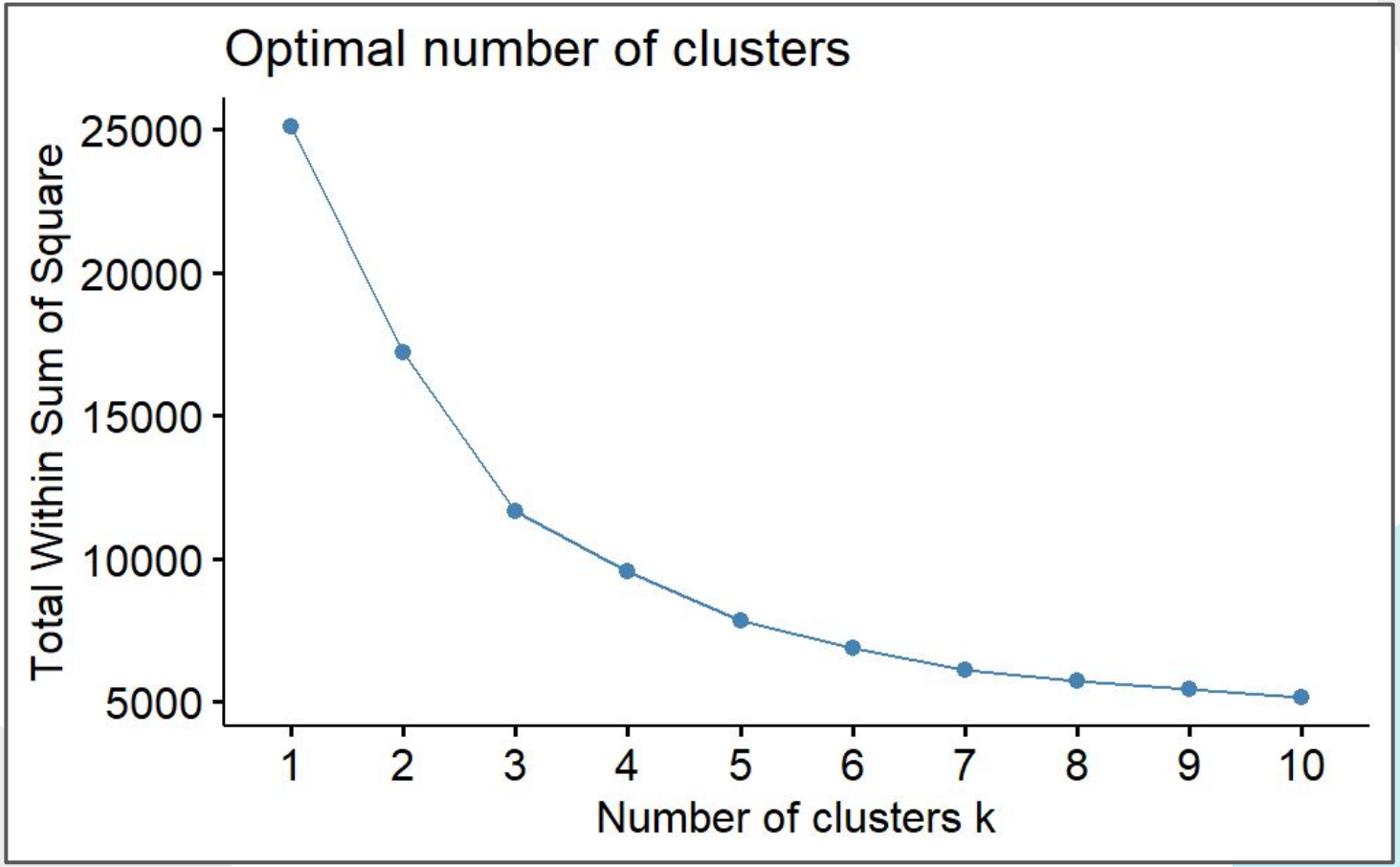

avg_glucose_level
age
bmi
hypertension
gender_Male
heart_disease
Residence_type_Urban
smoking_status_smokes
work_type_Private
work_type_Self.employed
work_type_Govt_job
smoking_status_never.smoked
ever_married_Yes
0
40
80
120
Importance

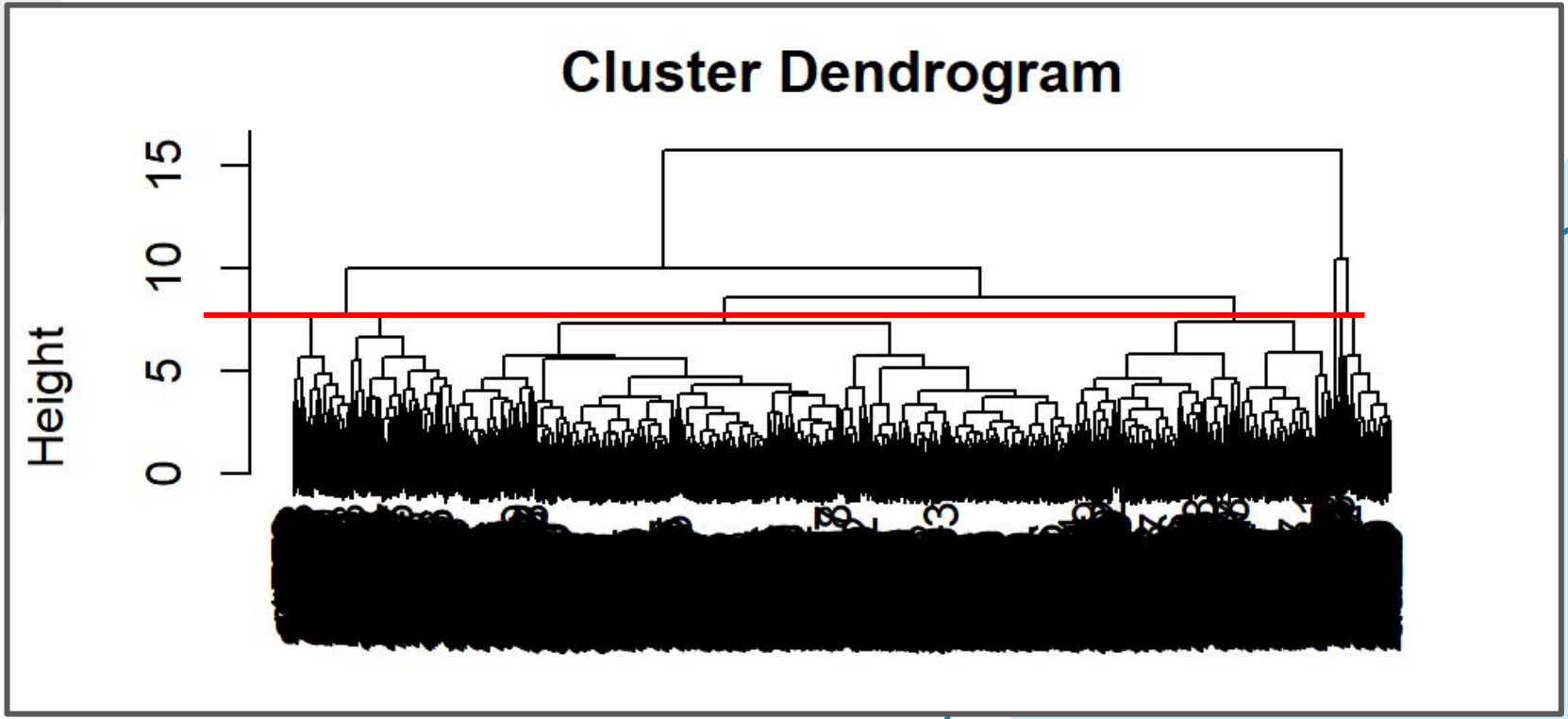
Classification Takeaways

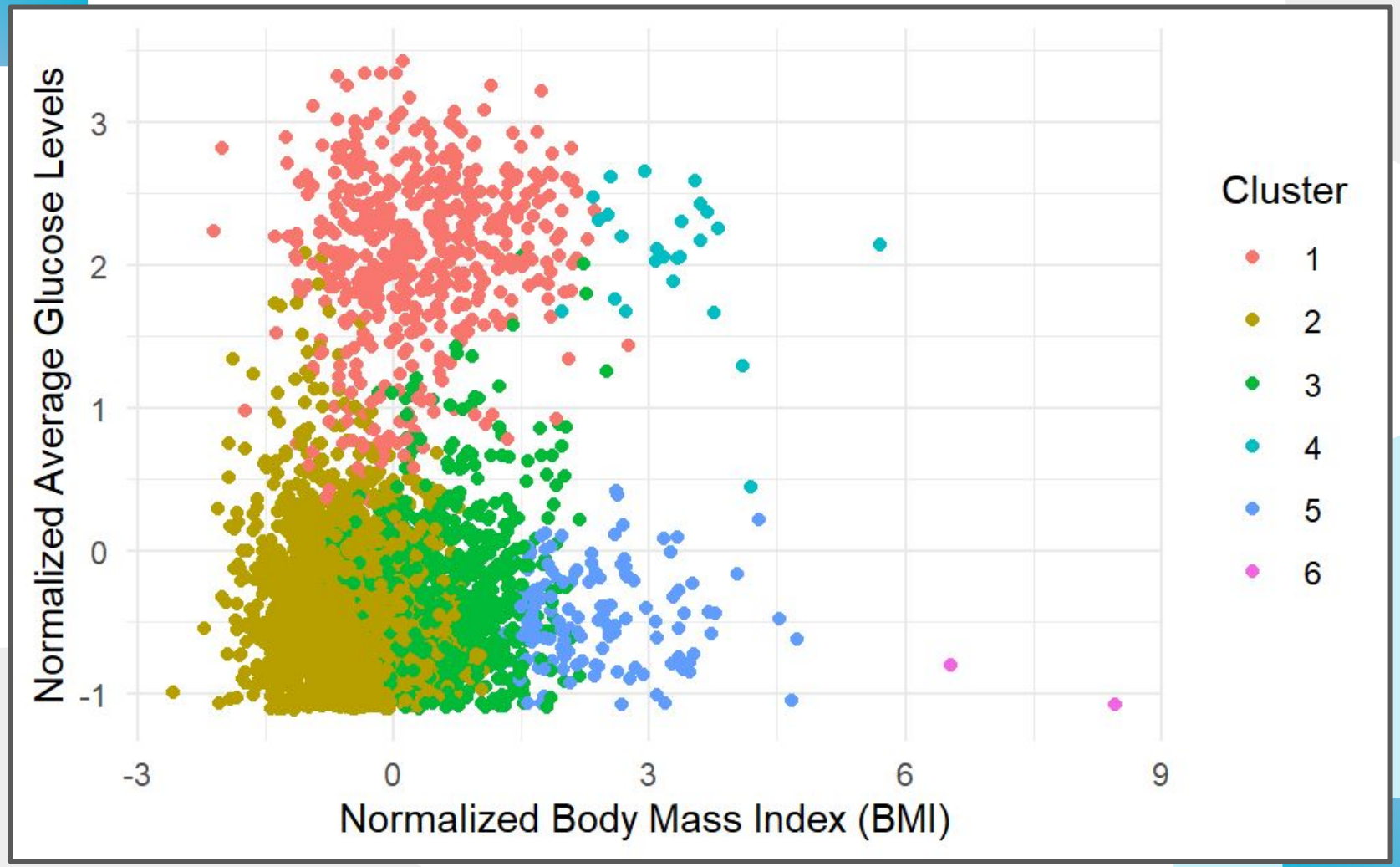
- **Logistic LASSO is a stronger model** because it is more interpretable and gives clearer metrics of variable importance
- Preferred thresholds with **higher specificity** while maintaining a reasonable sensitivity and accuracy
- Biological women, age, hypertension, working a government job, and unemployment are the strongest predictors

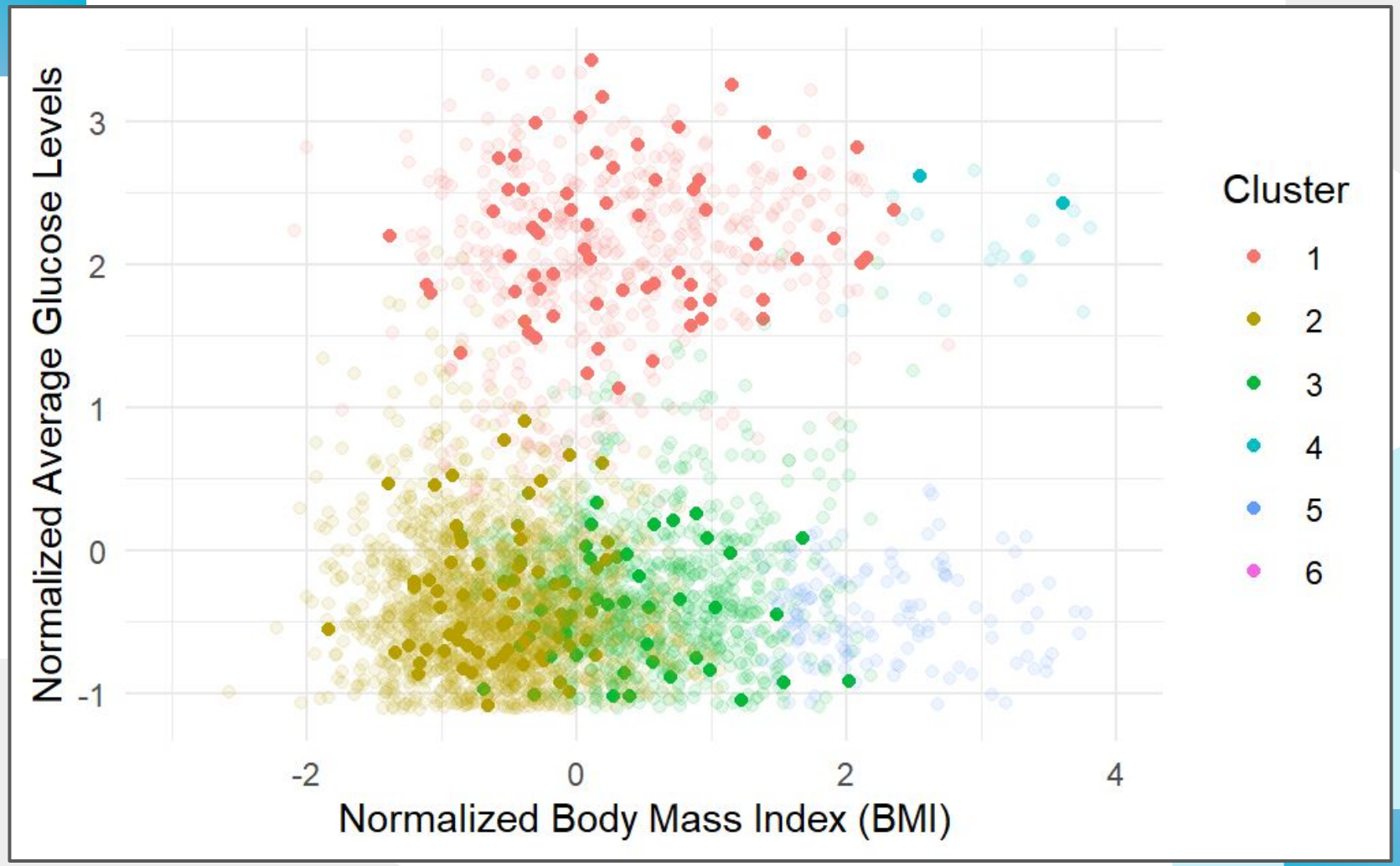
Clustering

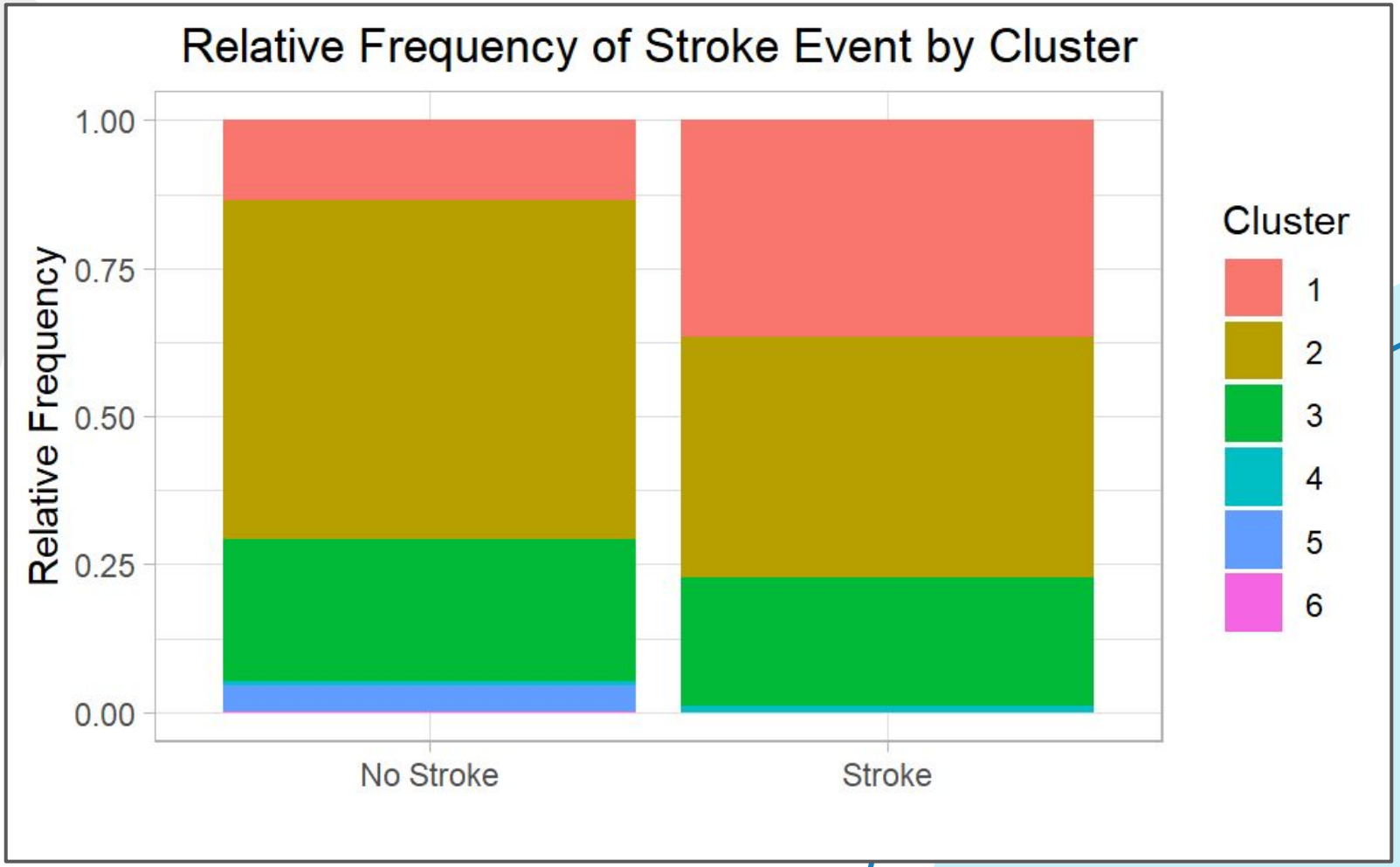
- Uncover **underlying patterns** determining whether a patient is at risk for a stroke
- Hierarchical clustering using complete linkage
- Created **6 clusters** cutting at a height of ~12.5


Optimal number of clusters
Total Within Sum of Square
25000
20000
15000
10000
5000
1
2
3
4
5
6
7
8
9
10
Number of clusters k

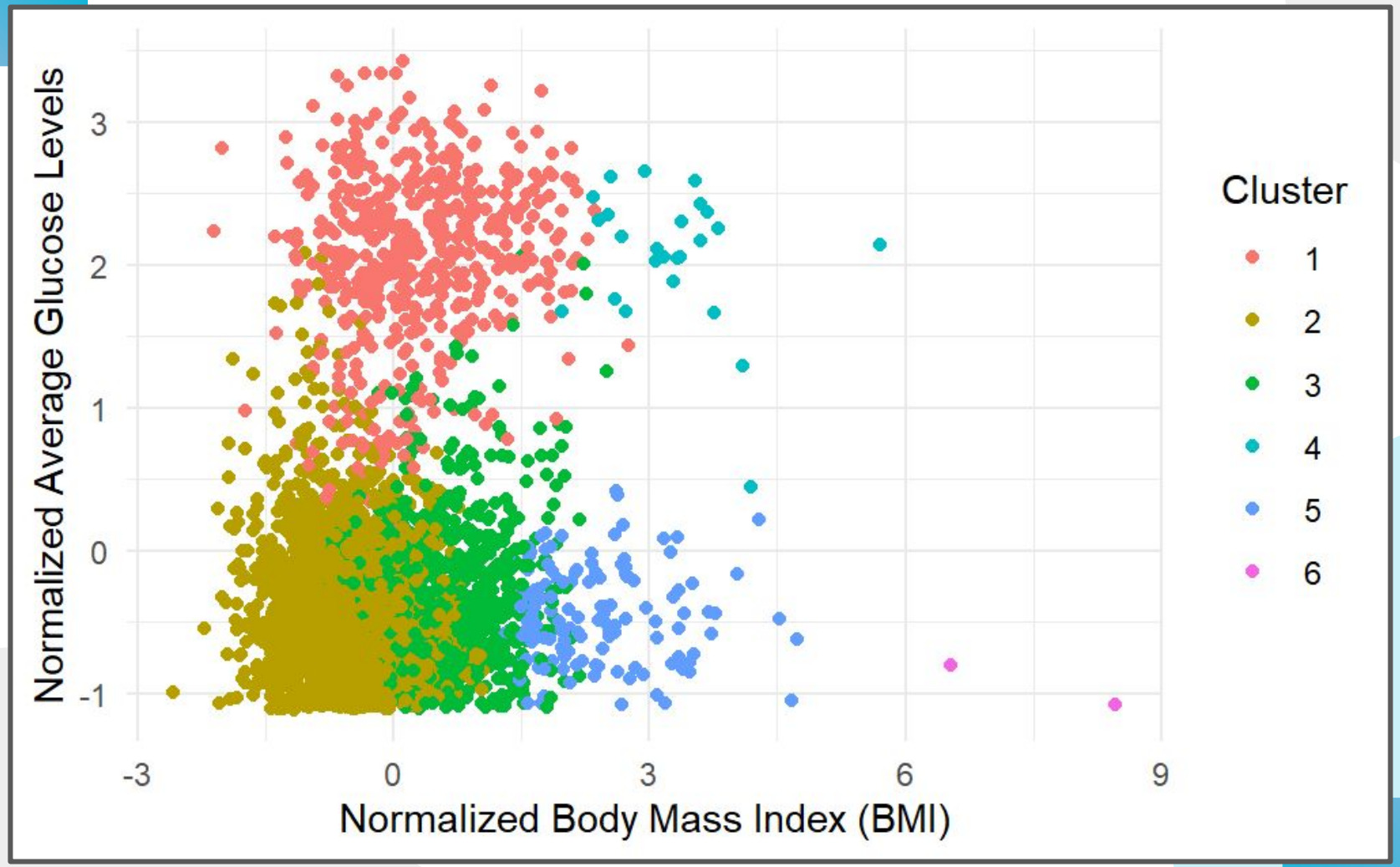

Cluster Dendrogram
Height
0
5
10
15


Normalized Average Glucose Levels
3
2
1
0
-1
-3
0
3
6
9
Normalized Body Mass Index (BMI)
Cluster
1
2
3
4
5
6


Normalized Average Glucose Levels
3
2
1
0
-1
-2
0
2
4
Normalized Body Mass Index (BMI)
Cluster
1
2
3
4
5
6


Relative Frequency of Stroke Event by Cluster
Relative Frequency
1.00
0.75
0.50
0.25
0.00
No Stroke
Stroke
Cluster
1
2
3
4
5
6

Cluster	Total Cases	Number of Strokes	Proportion of Strokes
1	508	66	**0.13**
4	25	2	**0.08**
3	816	39	**0.05**
2	1929	73	**0.04**
5	146	0	**0**
6	2	0	**0**


Normalized Average Glucose Levels
3
2
1
0
-1
-3
0
3
6
9
Normalized Body Mass Index (BMI)
Cluster
1
2
3
4
5
6

Clustering Takeaways

- Hierarchical clustering primarily groups cases by **average glucose level and BMI**
- Clusters with **higher average glucose levels tend to have higher instances of stroke**, whereas BMI does not tend to be as important in the dataset