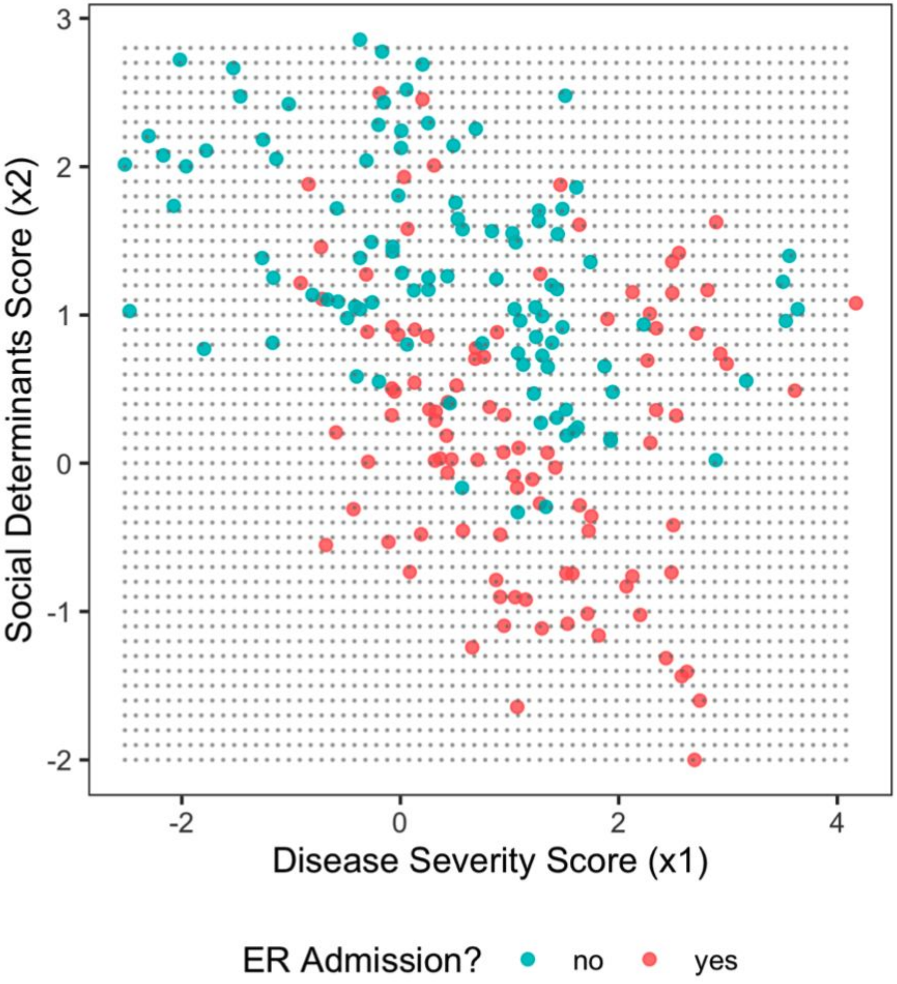
Chapter 2: The Basics of Classification

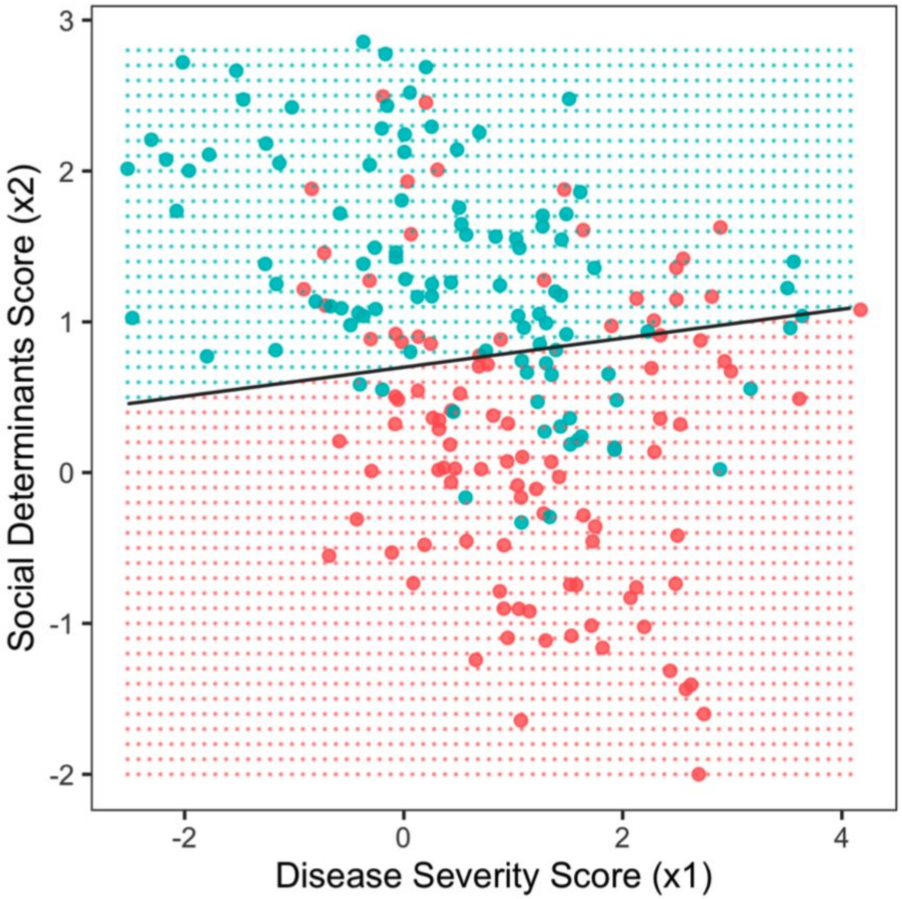
Modern Clinical Data Science
Chapter Guides
Bethany Percha, Instructor

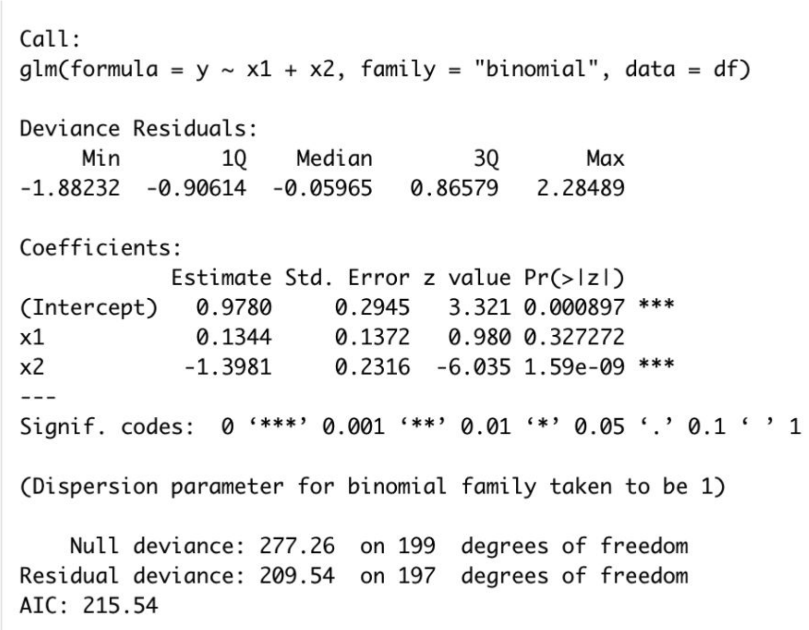
Icahn School of Medicine at Mount Sinai

How to Use this Guide

- Read the corresponding notes chapter first
- Try to answer the discussion questions on your own
- Listen to the chapter guide (should be 15 min, max) while following along in the notes


3
2
1
0
-1
-2
Social Determinants Score (x2)
-2
0
2
4
Disease Severity Score (x1)
ER Admission?
● no
● yes

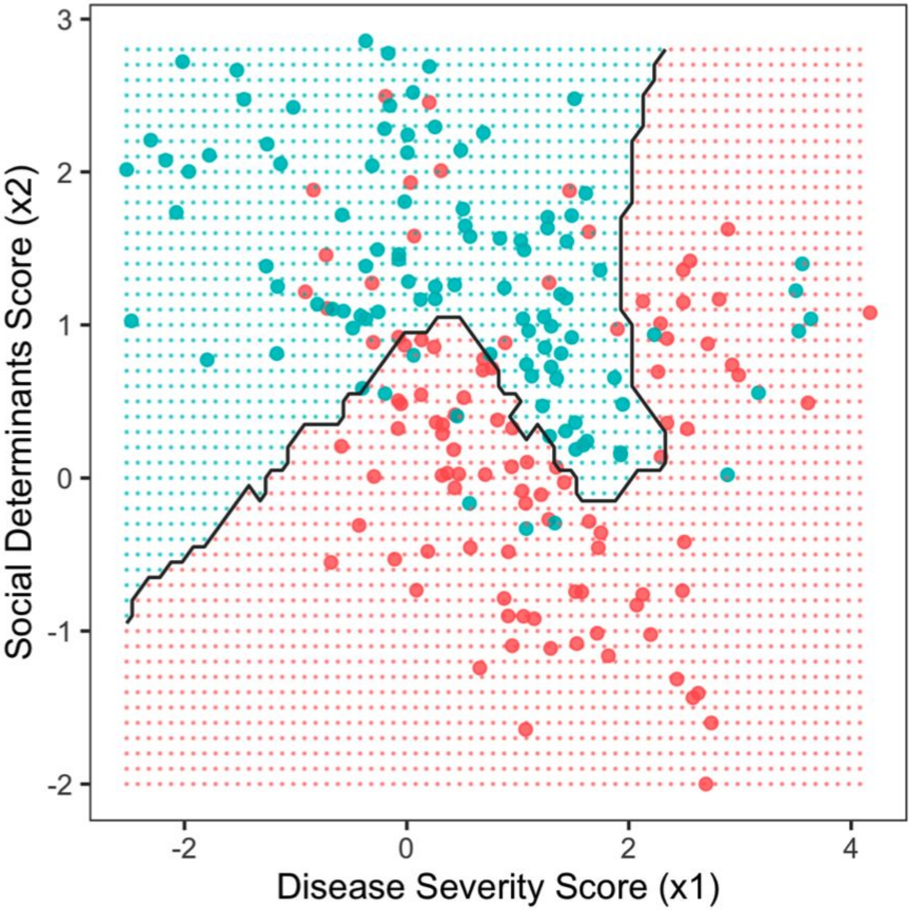


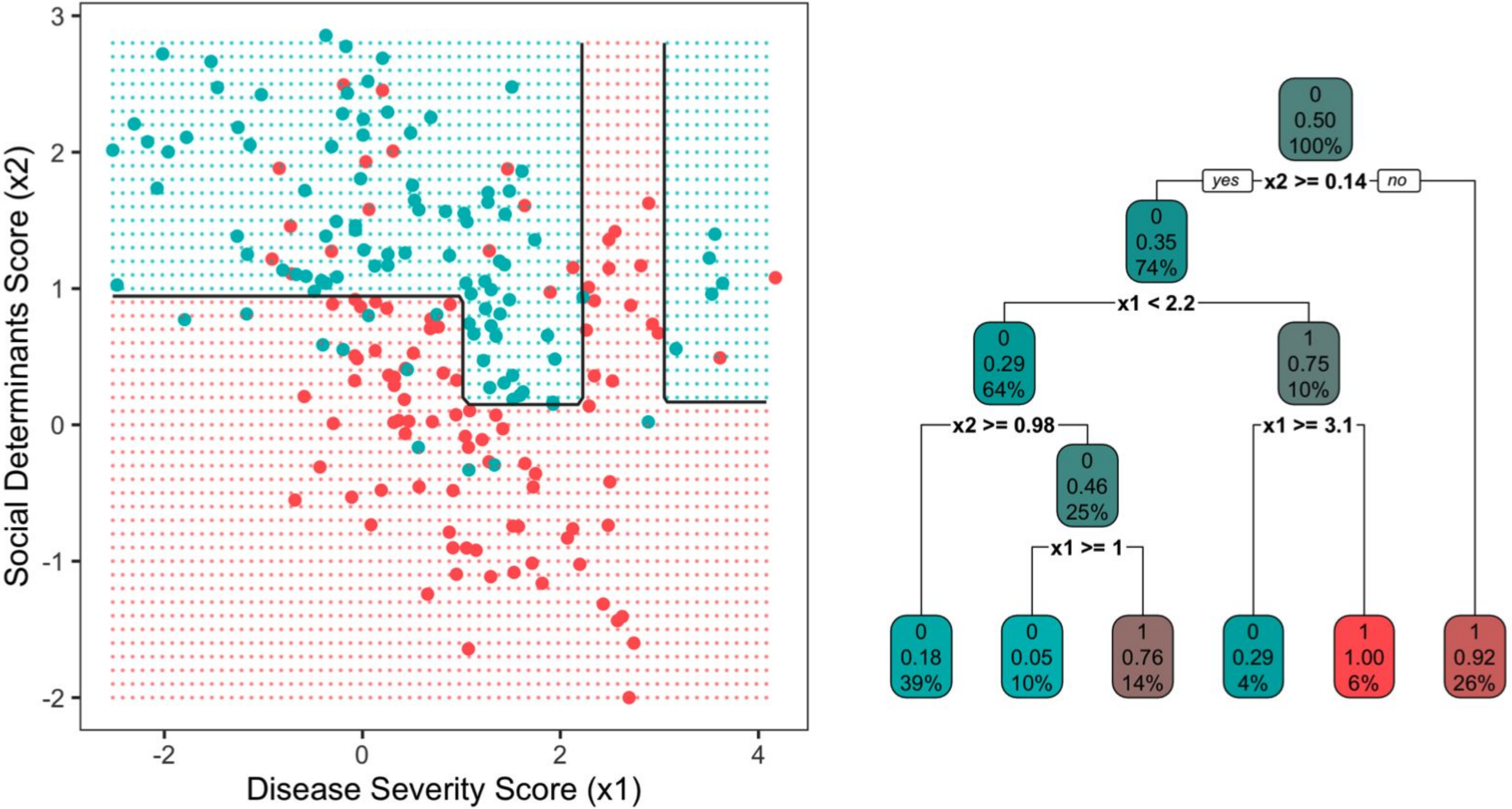


```
Call:
glm(formula = y ~ x1 + x2, family = "binomial", data = df)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.88232  -0.90614  -0.05965   0.86579   2.28489

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.9780     0.2945   3.321 0.000897 ***
x1            0.1344     0.1372   0.980 0.327272
x2           -1.3981     0.2316  -6.035 1.59e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 277.26  on 199  degrees of freedom
Residual deviance: 209.54  on 197  degrees of freedom
AIC: 215.54
```

$$0.9780 + 0.1344x_1 - 1.3981x_2 = 0$$

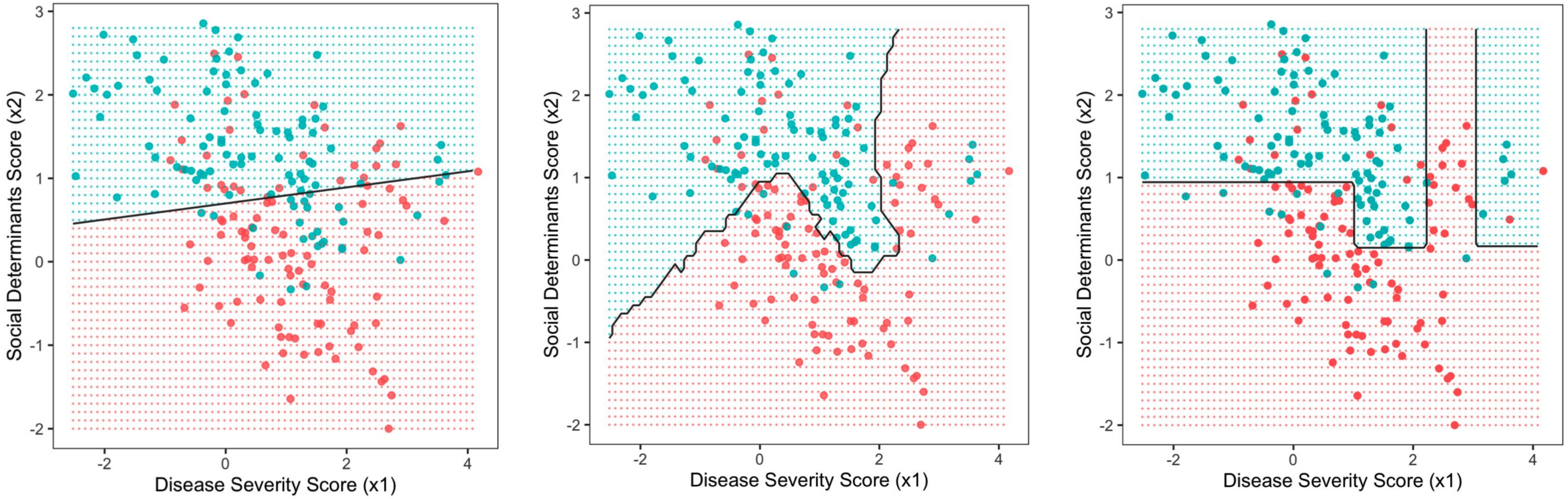
$$\implies x_2 = \frac{0.9780 + 0.1344x_1}{1.3981}$$


3
2
1
0
-1
-2
Social Determinants Score (x2)
-2
0
2
4
Disease Severity Score (x1)


Social Determinants Score (x2)
Disease Severity Score (x1)
3
2
1
0
-1
-2
-2
0
2
4
0
0.50
100%
yes
x2 >= 0.14
no
0
0.35
74%
x1 < 2.2
0
0.29
64%
1
0.75
10%
x2 >= 0.98
x1 >= 3.1
0
0.46
25%
x1 >= 1
0
0.18
39%
0
0.05
10%
1
0.76
14%
0
0.29
4%
1
1.00
6%
1
0.92
26%

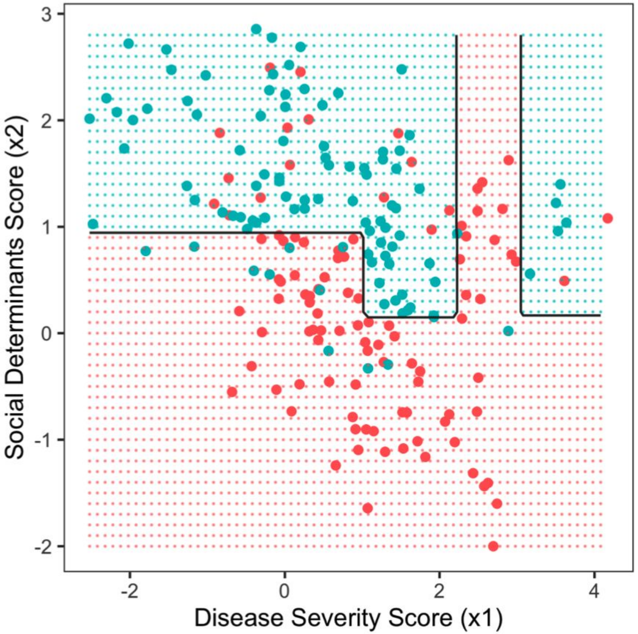
Question 2.1

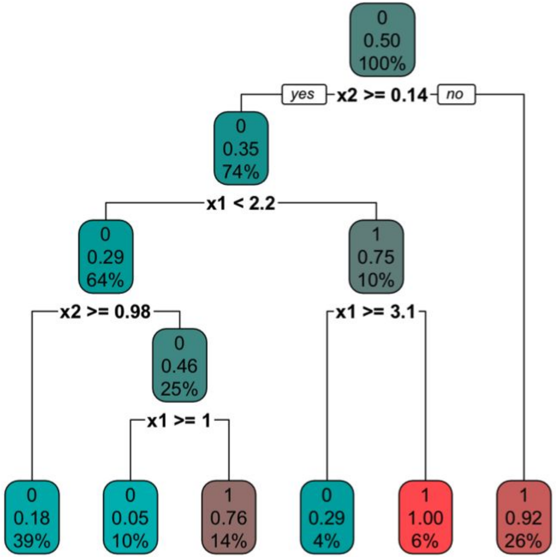
How can you tell, just by looking at these images, which feature (x_1 or x_2) impacts the outcome the most? Which one is it?

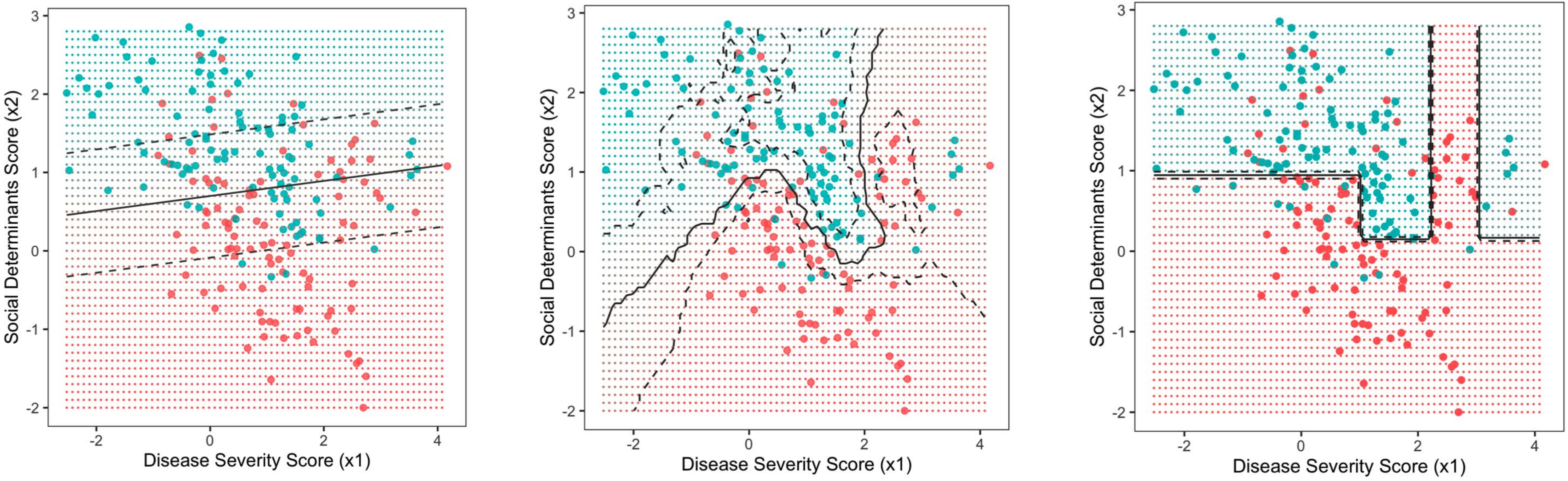


Question 2.2

There are six rectangular regions in the picture of the decision tree decision boundary. Each corresponds to one of the six leaves of the tree. Identify all six and which leaves they correspond to on the decision tree.



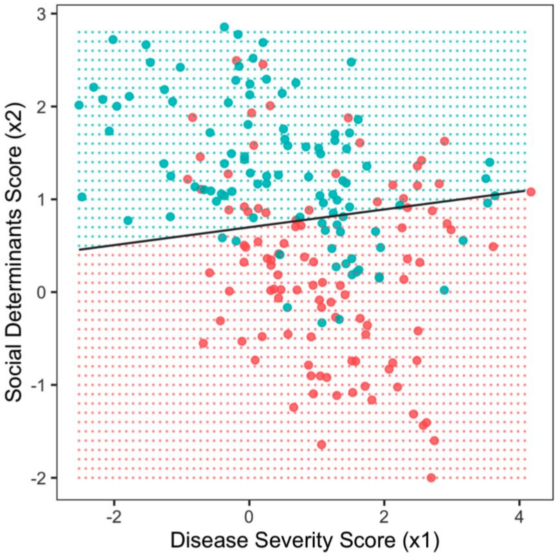


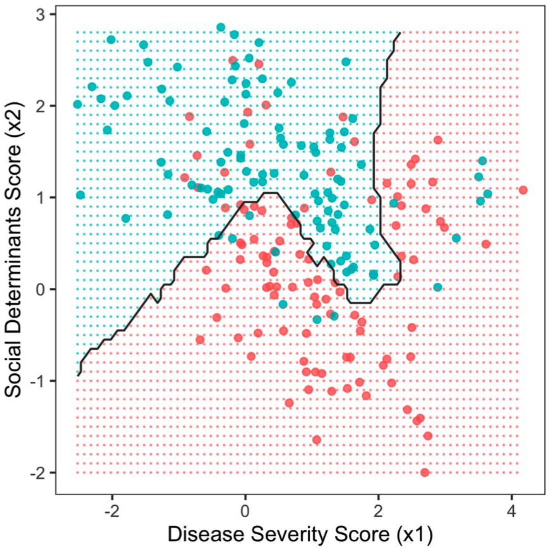

Social Determinants Score (x2)
3
2
1
0
-1
-2
-2
0
2
4
Disease Severity Score (x1)
Social Determinants Score (x2)
3
2
1
0
-1
-2
-2
0
2
4
Disease Severity Score (x1)
Social Determinants Score (x2)
3
2
1
0
-1
-2
-2
0
2
4
Disease Severity Score (x1)

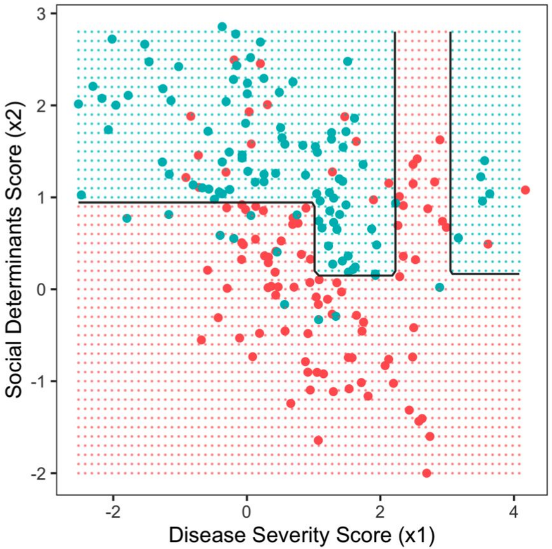
Question 2.3

What are the advantages and disadvantages of each algorithm?

1. Logistic regression?
2. KNN ($K = 15$)?
3. Decision tree?







Question 2.4

What makes a good classification algorithm? Consider issues of accuracy, generalizability, and speed (both to train the algorithm and to use it to make predictions on new samples).