1 A

Daniel Trhan

I Think K=3 will provide a better prediction. In the training data there are 7 observations. I Think taking 5 Ks would make the model too bias towards the majority class.

B

K=3

| Test ID | Dose | Improvement | Nearest Neighbors | Neighbor Responses | $\hat{Y}$ | Error? |
|---|---|---|---|---|---|---|
| H | 0 | N | A, B, C | N, Y, N | N | correct |
| I | 0 | N | A, B, C | N, Y, N | N | correct |
| J | 1 | Y | B, A, C | Y, N, N | N | incorrect |
| K | 3 | Y | D, C, E | N, N, Y | N | incorrect |
| L | 5 | Y | E, F, G | Y, Y, Y | Y | correct |
| | | | | | Pred Error Rate: | $\frac{2}{5} = 40\%$ |

E & F are both Y so I'll just take E

C

$$\text{logit} = \frac{1}{1+e^{-(-1+2\cdot\text{dose})}} \qquad \text{Threshold} = 0.5$$

| Test ID | Dose | Improvement | Logit | $\hat{p}$ | $\hat{Y}$ | Error? |
|---|---|---|---|---|---|---|
| H | 0 | N | $-1+2\cdot0$ | $\frac{1}{1+e} \approx 0.269$ | N | correct |
| I | 0 | N | 1 | 0.269 | N | correct |
| J | 1 | Y | -1 | 0.731 | Y | correct |
| K | 3 | Y | -5 | 0.993 | Y | correct |
| L | 5 | Y | -9 | 0.999 | Y | correct |
| | | | | | Pred Error Rate: | 0% |

(D) $\text{loss}(FP) = 3\,\text{loss}(FN)$

$$\text{new threshold} = \frac{1}{1+\frac{\text{loss}(FP)}{\text{loss}(FN)}} = \frac{1}{1+3} = 0.25$$

| Test | Dose | Improvement | $\hat{p}$ | $\hat{Y}$ | Error? |
|---|---|---|---|---|---|
| H | 0 | N | 0.269 | Y | incorrect |
| I | 0 | N | 0.269 | Y | incorrect |
| J | 1 | Y | 0.731 | Y | correct |
| K | 3 | Y | 0.993 | Y | correct |
| L | 5 | Y | 0.999 | Y | correct |

Pred error rate: $\frac{2}{5}$ 40%

(E)

| Train | A | B | C | D | E | F | G | Test | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dose | 0 | 1 | 2 | 3 | 4 | 4 | 6 | | 0 | 0 | 1 | 3 | 5 |
| Imp | N | Y | N | N | Y | Y | Y | | N | N | Y | Y | Y |
| $\bar{X}_Y$ | (1+4+4+6)/4 = 3.75 | | | | | | | $\hat{Y}$ | N | N | N | Y | Y |
| $\bar{X}_N$ | (0+2+3)/3 = 1.67 | | | | | | | Error? | ✓ | ✓ | ✗ | ✓ | ✓ |
| | | | | | | | | Error Rate | 1/5 = 20% | | | | |

$\frac{3.75+1.67}{2} = 2.71$ ← Threshold

(F)

| | Prediction Error Rate |
|---|---|
| (b) KNN | 40% |
| (c) Logistic Regression | 0% & 40% |
| (e) LDA | 20% |

I would recommend using Model C, the Logistic regression because it has the lowest error rate.