Midterm Q1

Name_____Course_____

Instructions

- Include enough details to show your reasoning. Clearly indicate your final answers.
- Notes, textbook, calculator, and computer are allowed; internet is only allowed for course materials.
- Each problem is 20 points. Total points = 40.

> **Important**
>
> This exam uses the starwars data set from the dplyr package and you must have the at least version 1.1.4 of the package to have accurate data.

1 Drug Testing *Do this by hand. Show the steps and results of your analysis.*

Patients in a clinical trial received either a placebo (0 mg) or different doses of a treatment. The treatment was a success if a patient improved after three days. Patient data was randomly split into training and testing sets and relabeled to be in sequence.

Train ID	A	B	C	D	E	F	G	Test ID	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

How well can the company predict success of the treatment given the dose? Use these training and testing sets to estimate the **error classification rate** for each classification method below.

(a) By just looking at the data, do you think KNN with $K = 3$ or $K = 5$ will provide better predictions? Why?

(b) KNN. Use your answer from (a) to choose $K = 3$ **or** $K = 5$. Split any ties in half.

You may use the following table to show your work. The test data is on the left. Use the remaining columns to identify the nearest neighbors from the training data, their responses (`Improvement`), and $\hat{Y}$ for the test observation. Then identify if the prediction is an error or not and calculate the overall prediction error rate.

Test ID	Dose	Improvement	Nearest Neighbors	Neighbor Responses	$\hat{Y}$	Error?
H	0	N				
I	0	N				
J	1	Y				
K	3	Y				
L	5	Y				
					Pred Error Rate:	

(c) Logistic regression. The training data was used in a logistic regression model which gave the following R output. Improvement was coded as `Success` with value 1. Assume equal loss for false positives and false negatives. You can use the table below to show your work.

```
glm(Success ~ Dose, family=binomial, data=training)
Coefficients:
  (Intercept)  Dose
         -1.0  2
```

Test ID	Dose	Improvement	Logit	$\hat{p}$	$\hat{Y}$	Error?
H	0	N				
I	0	N				
J	1	Y				
K	3	Y				
L	5	Y				
					Pred Error Rate:	

(d) Repeat question (c), assuming the penalty for a false negative (predicting failure when the treatment actually succeeded) is **1/3rd as high** as the penalty for a false positive (predicting success when the treatment failed). Note: this is equivalent to the Loss for a False Positive being 3 times as high as the loss for a False Negative. Calculate the new threshold and use the $\hat{p}$ from (c) to predict $\hat{Y}$ and calculate the error rate. You may use the space and table below to show your work.

Test	Dose	Improvement	$\hat{p}$	$\hat{Y}$	Error?
H	0	N			
I	0	N			
J	1	Y			
K	3	Y			
L	5	Y			

Test	Dose	Improvement	$\hat{p}$	$\hat{Y}$	Error?
				Pred Error Rate:	

(e) Linear discriminant analysis. Assume **equal prior probabilities of a success and a failure** of the treatment and **equal penalties for misclassification** (which allows you to simplify your analysis). You can use the table below to show your work.

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
$\bar{X}_Y$								$\hat{Y}$					
$\bar{X}_N$								Error?					
								Error Rate					

(f) **Summarize your analysis in the following table** and make a recommendation as to which method or methods (from, b, c, or e) you would recommend for this problem and why.

	Prediction Error Rate
(b) KNN	
(c) Logistic Regression	
(e) LDA	