Data 558: Statistical Machine Learning

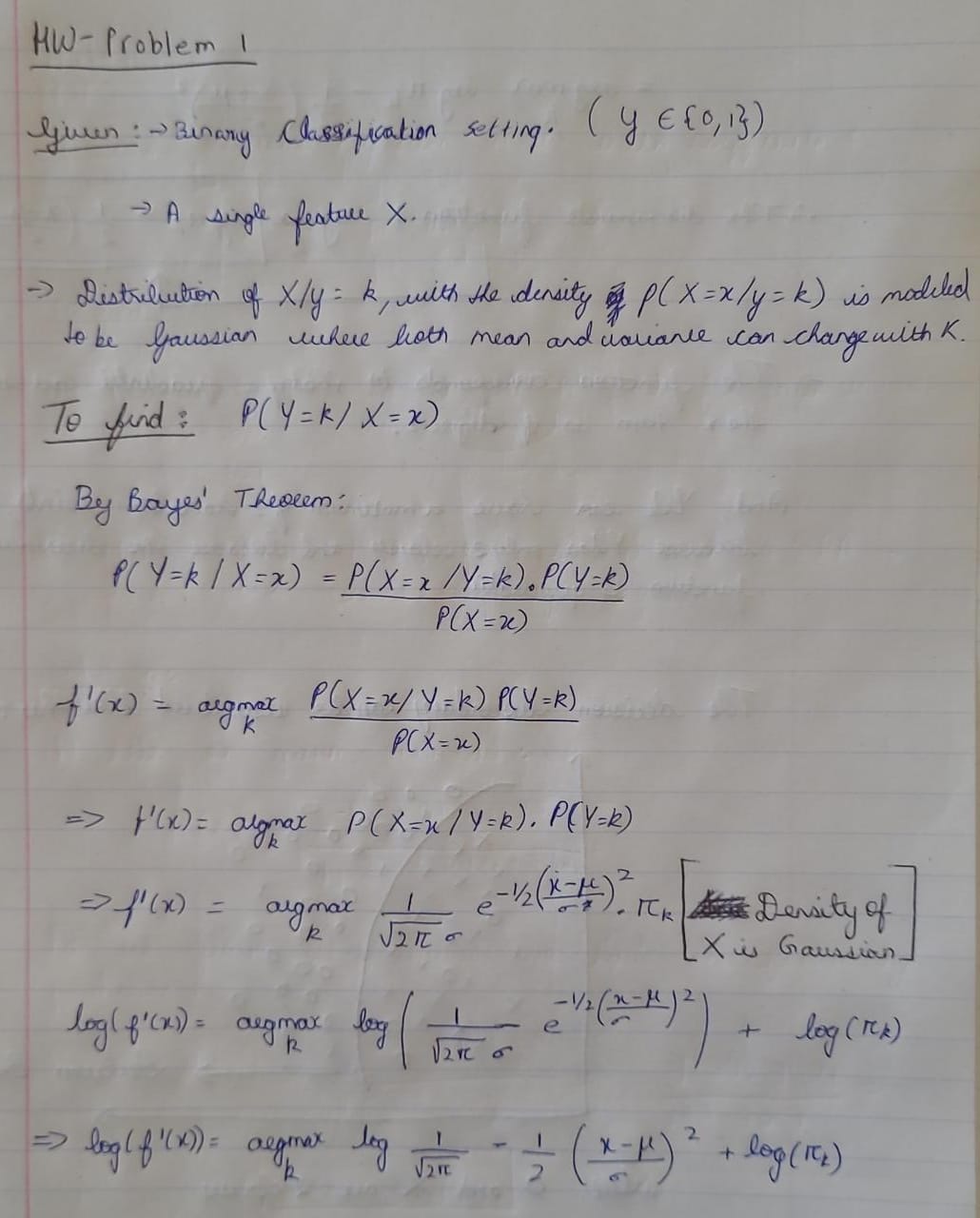
Spring 2023

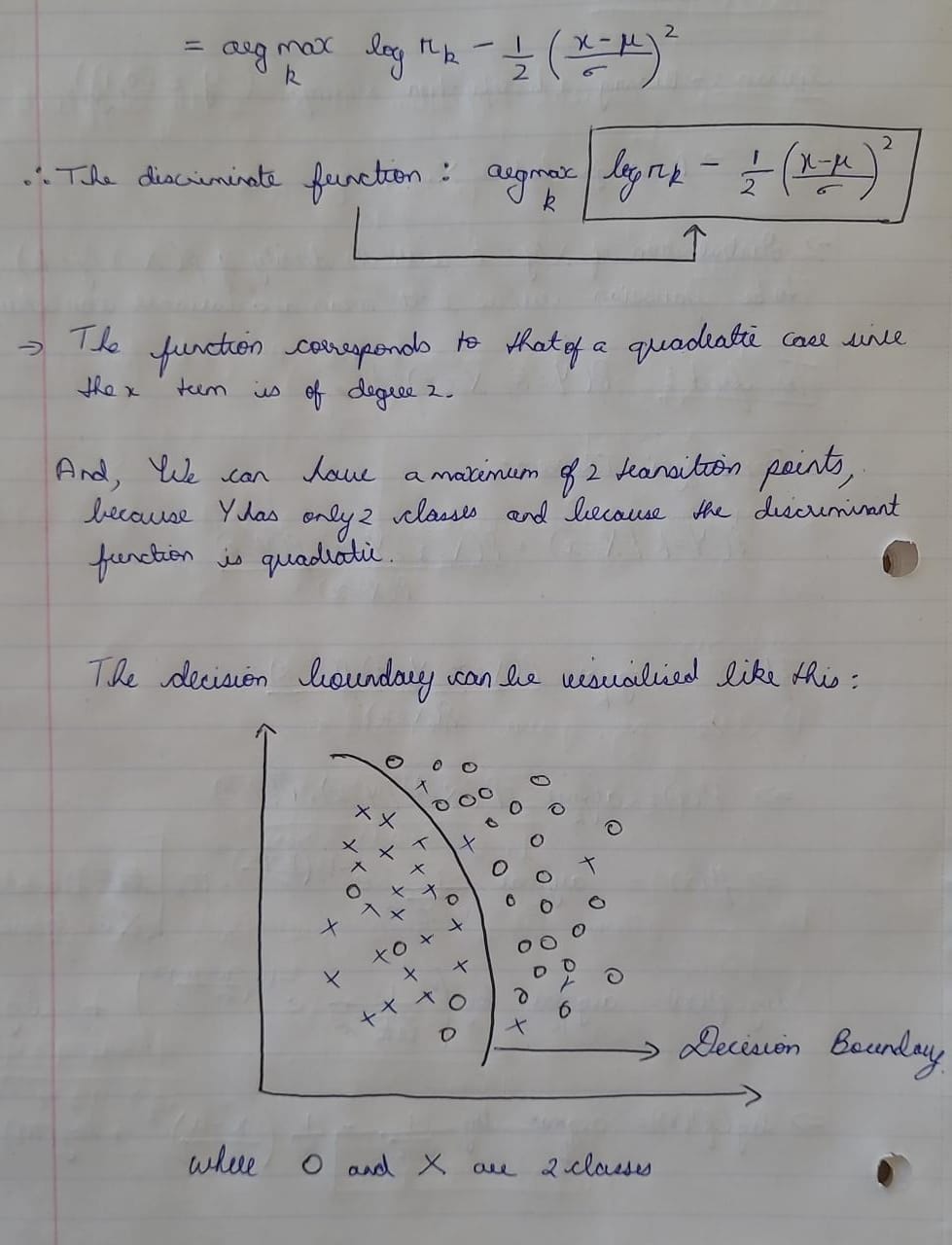
Homework - 2

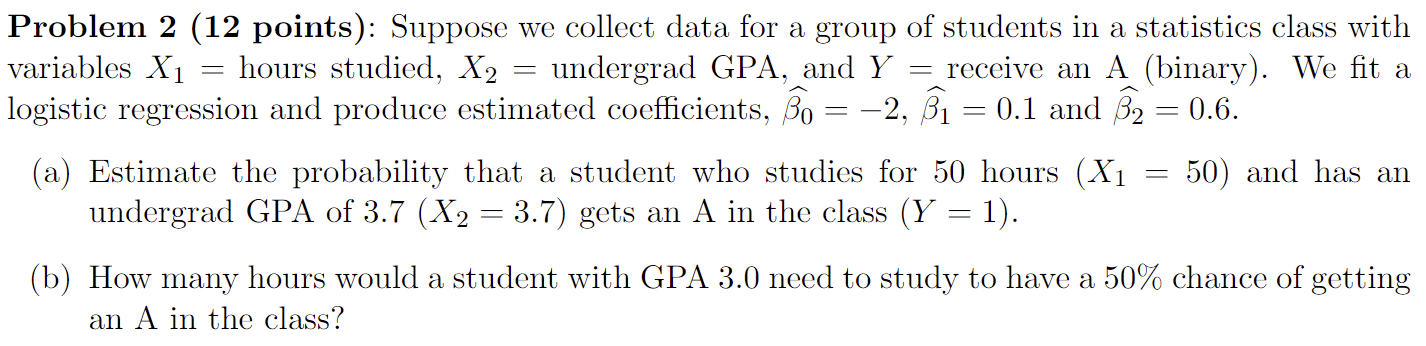
Arjun Sharma

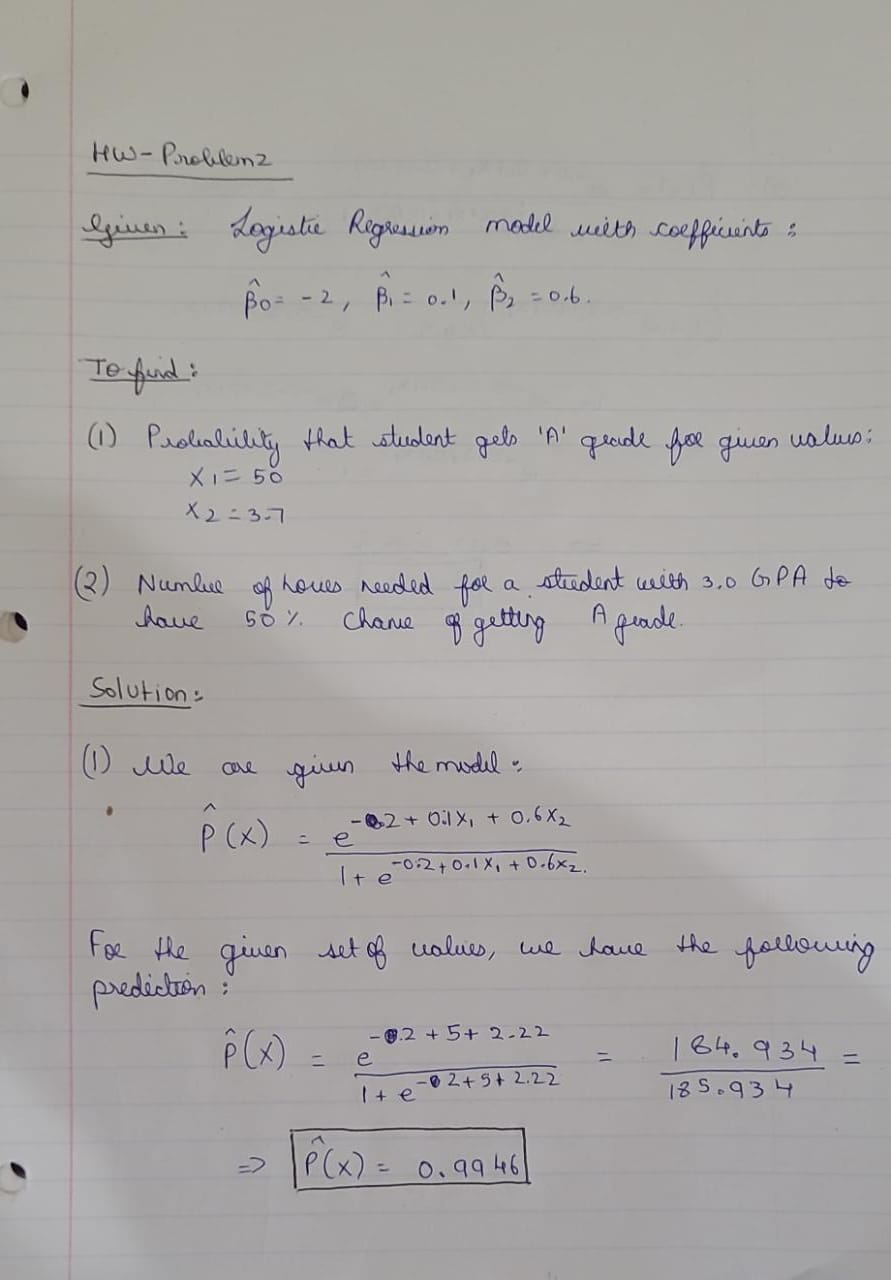
**Conceptual Questions:**

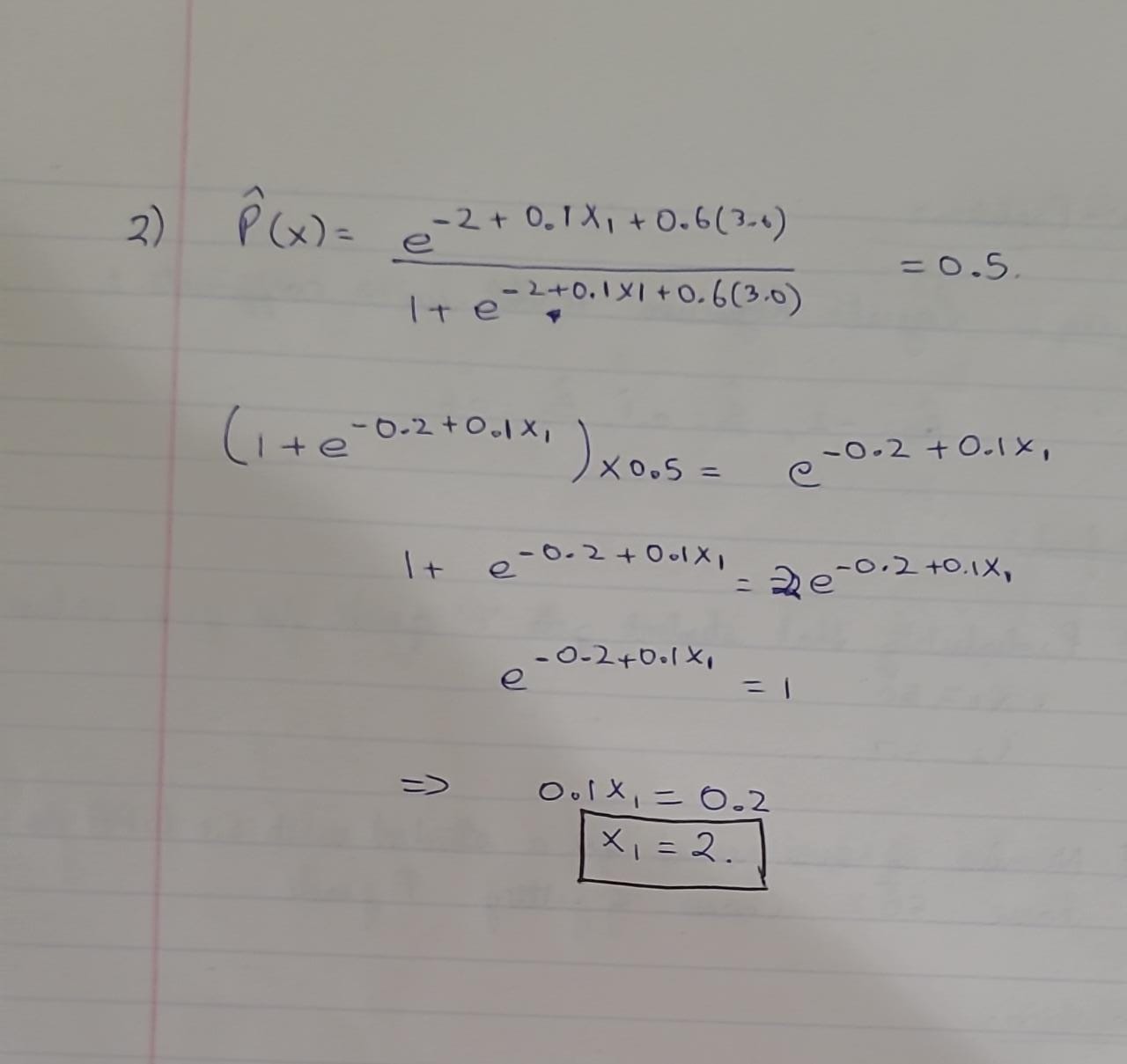


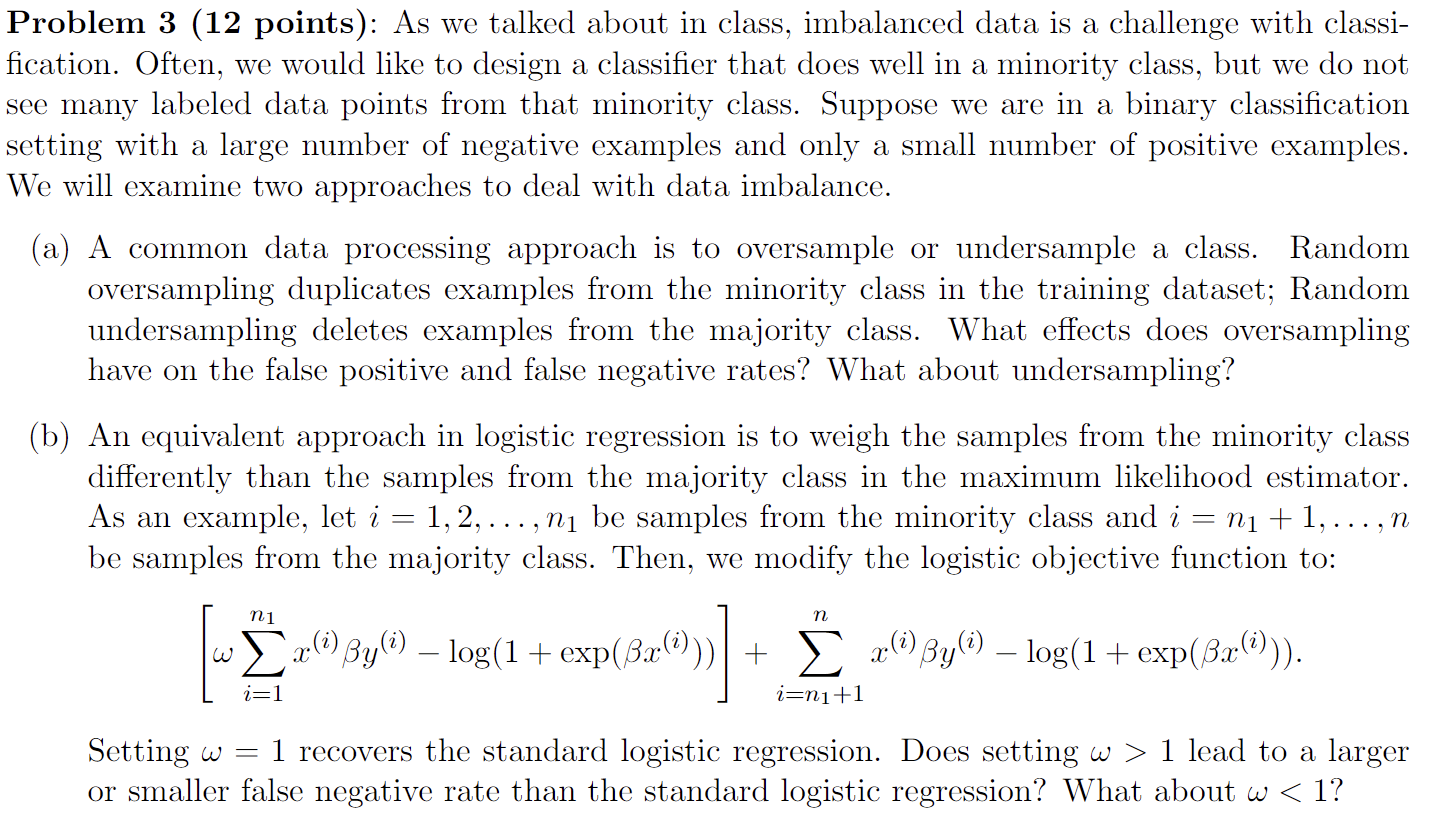












(a)

Oversampling the minority class leads to duplication of the positive class labelled datapoints. Now, because the frequency of positive class datapoints increases, the model becomes likely to predict the positive class more often. This can lead to a higher false positive rate. Conversely, the false negative rate decreases since the model is relatively more likely to predict the positive class compared to a model which has been trained on data without any sampling.

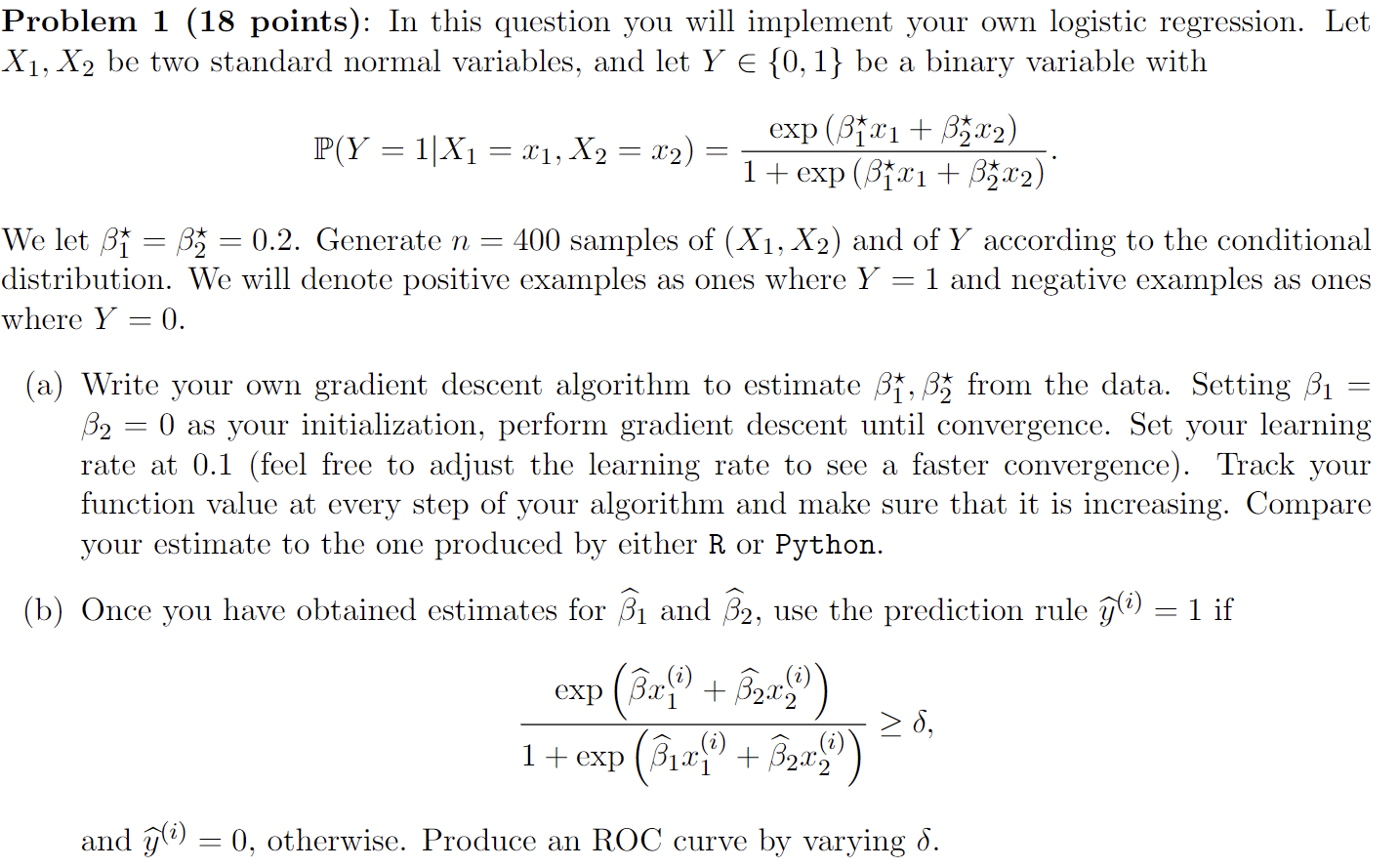
Under-sampling the majority class leads to some information loss about the relationship between the predictors and outcome. In addition, due to the relatively balanced composition of the classes in the data, the model would have a higher false negative rate. The increase in false negatives would also imply a decreased false positive rate.

(b)

For ω > 1, the added weights to the minority class increases the model’s likelihood to predict in favor of the minority class. Consequently, the false negative rate would decrease since the model’s frequency of prediction to be negative decreases.

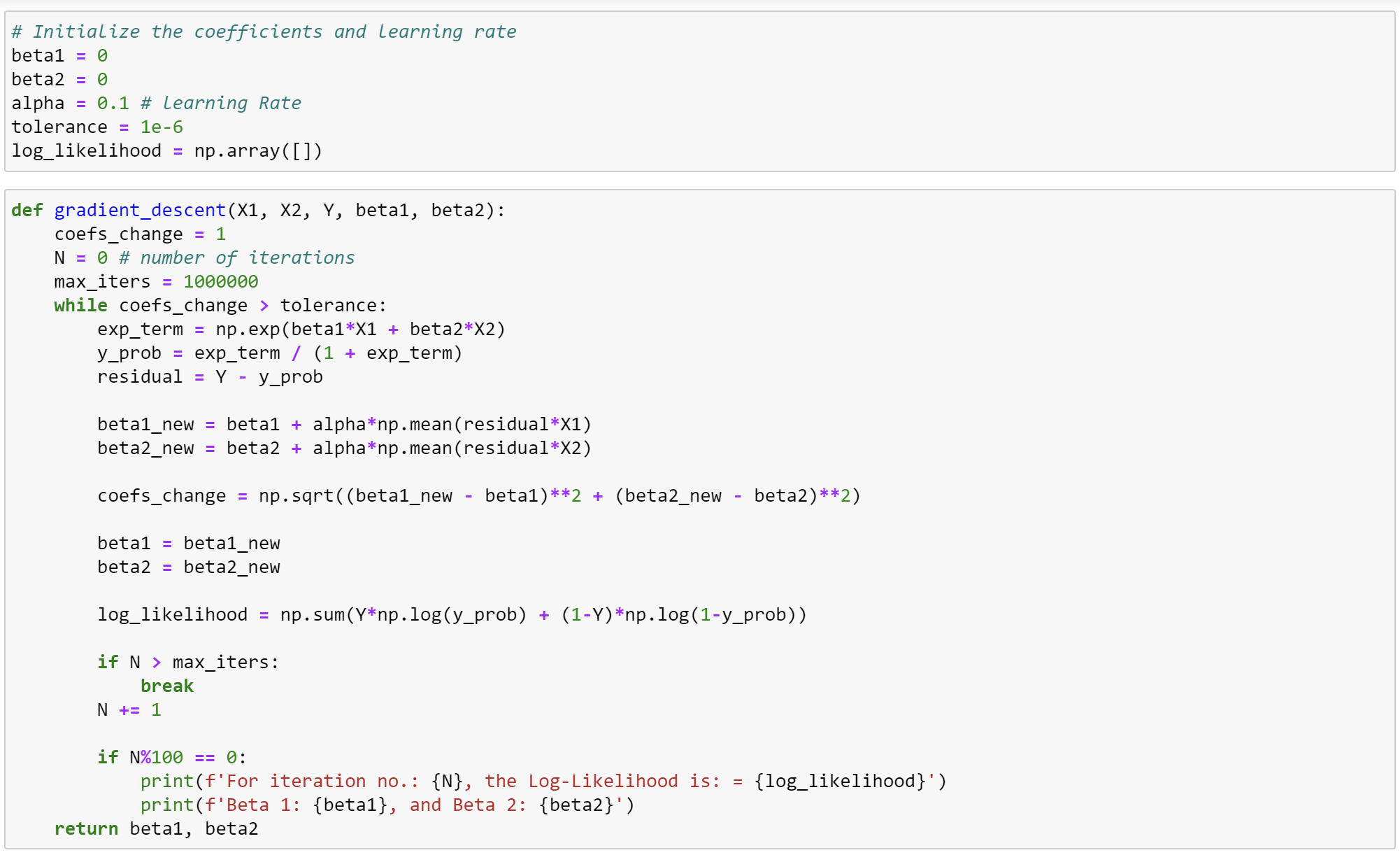
For ω < 1, the reduced weights to the minority class increases the model’s likelihood to predict in favor of the majority class. Thus, the false negative rate would increase since the model is more likely to predict an outcome as negative.

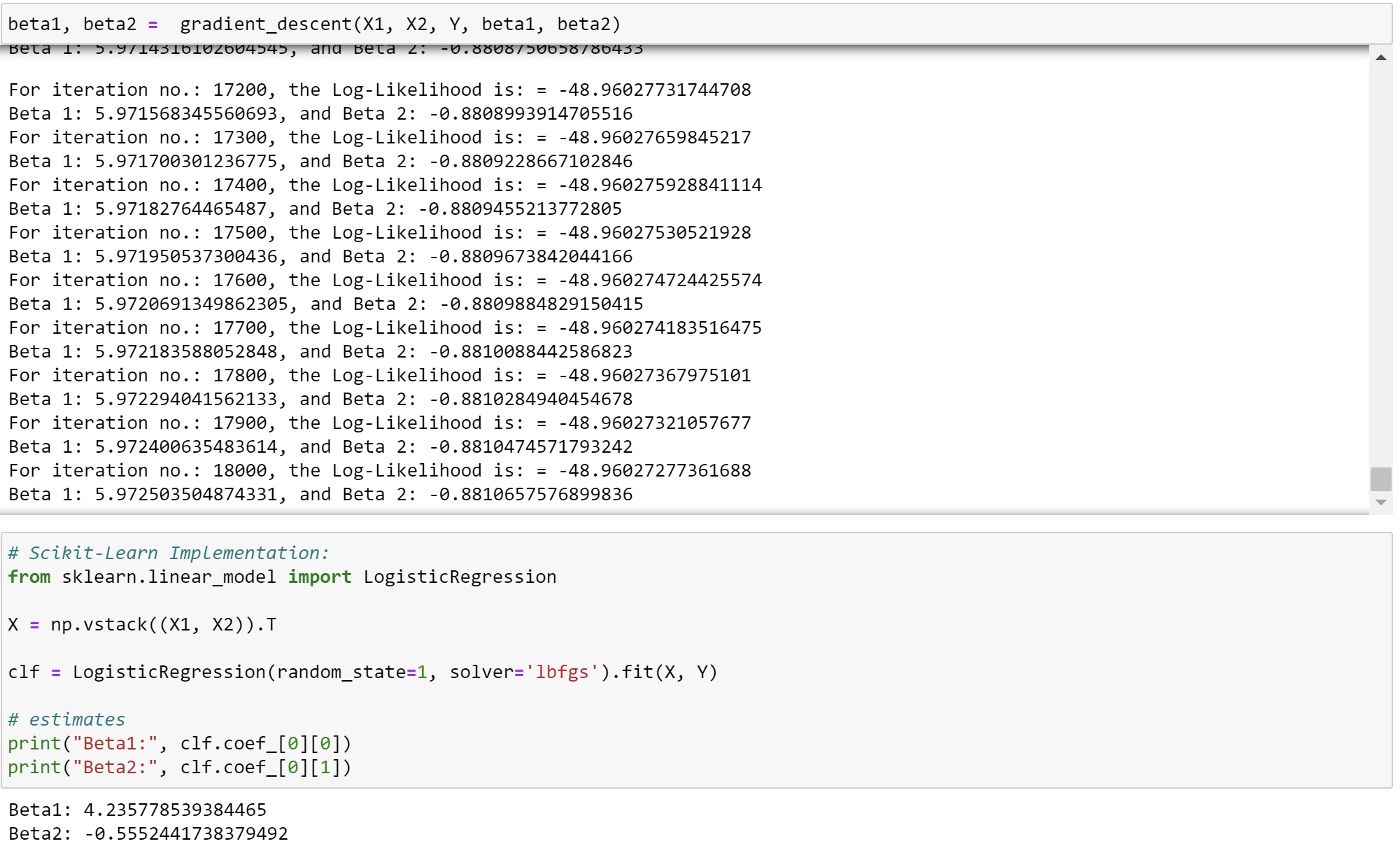
**Applied Questions:**



Answer:







The coefficients of my model are:

Beta1 = 5.972503

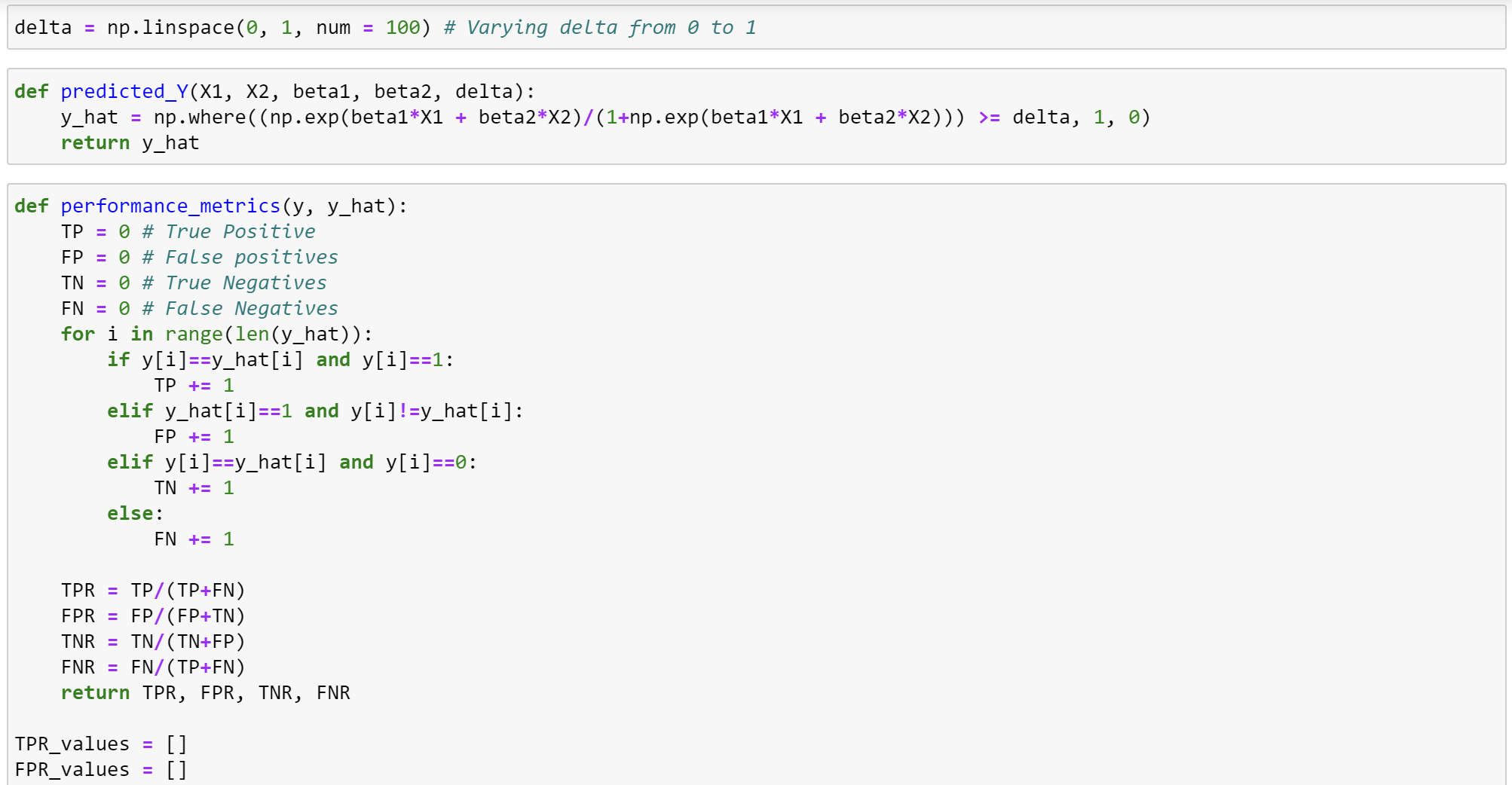
Beta2 = -0.881065

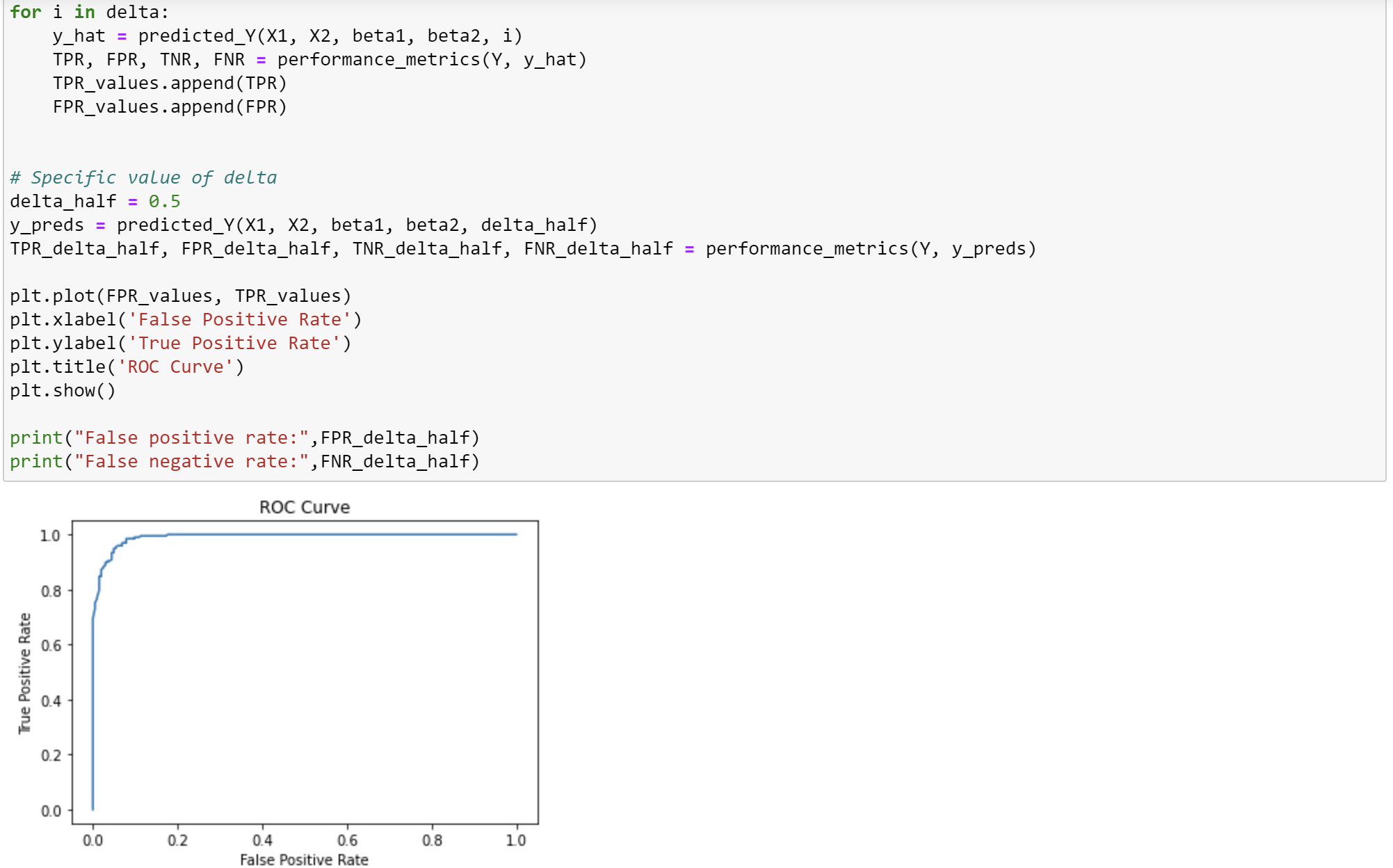
While the coefficients of the model built on scikit-learn are:

Beta1: 4.235778

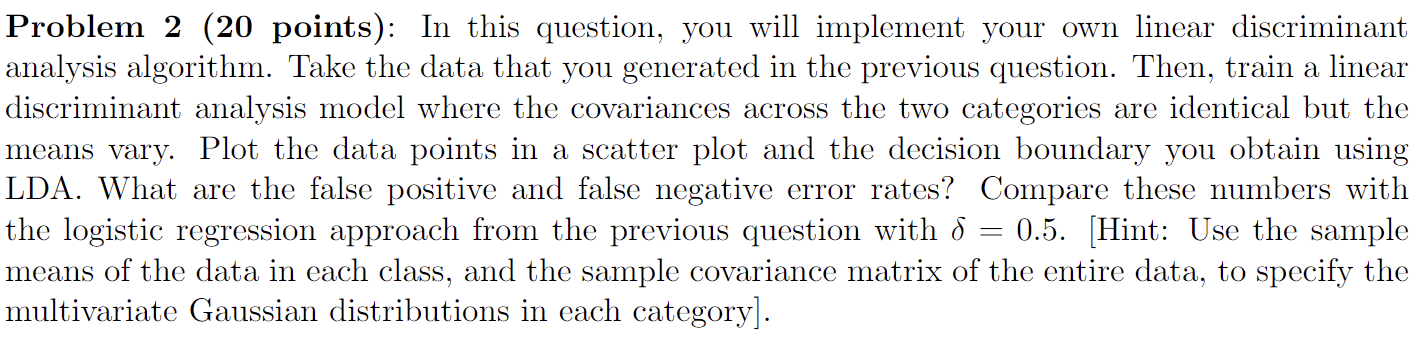
Beta2: -0.5552441

The coefficients of the models are similar, although they are not the same.

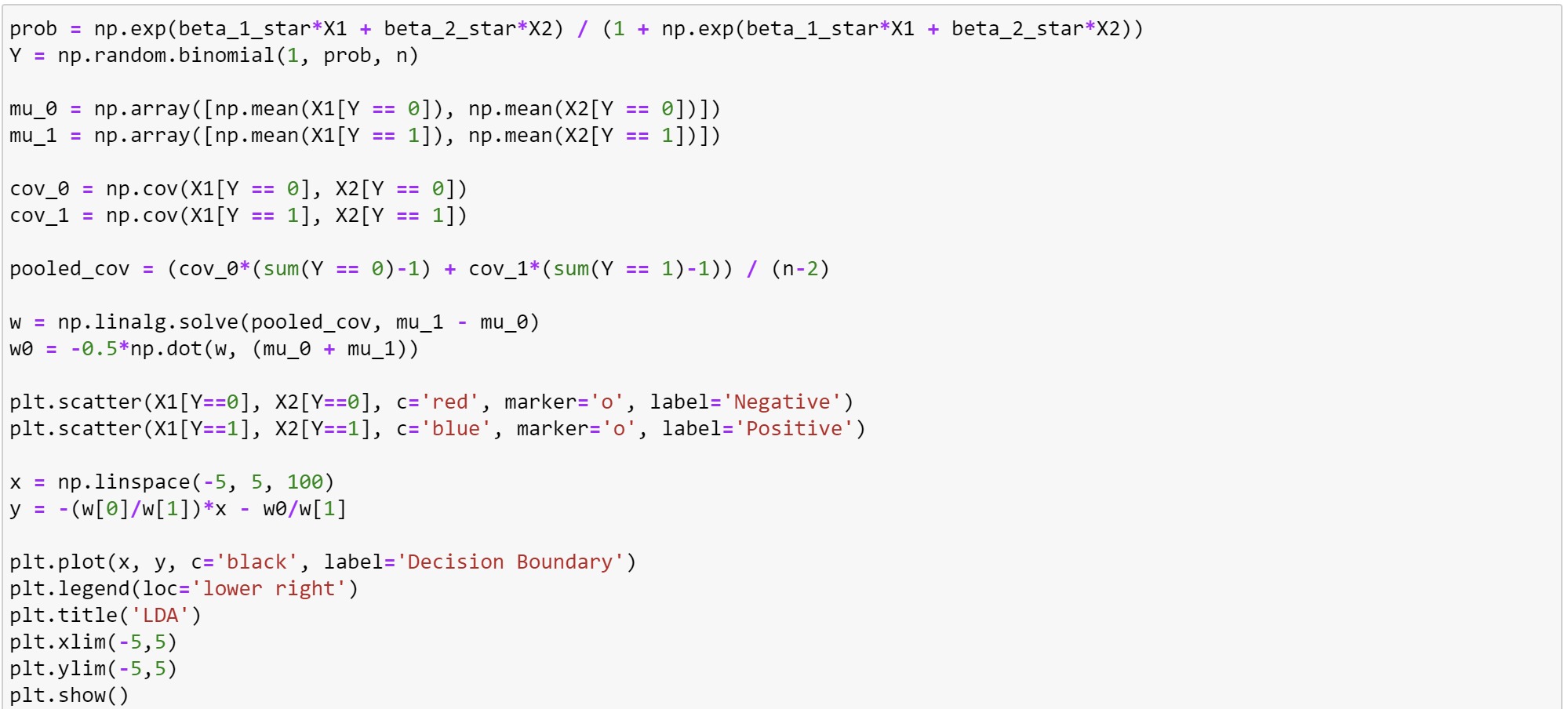


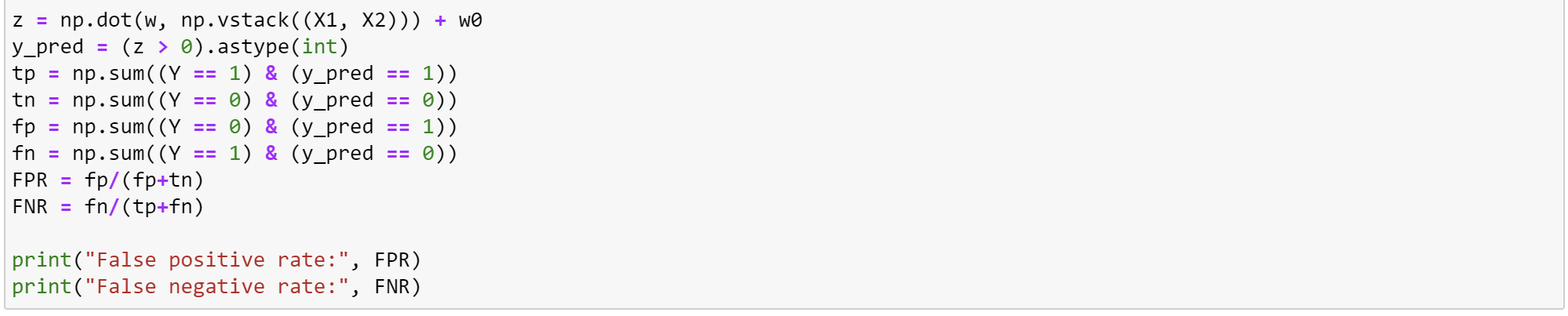






Note: Y is the data that was generated in Problem 1.





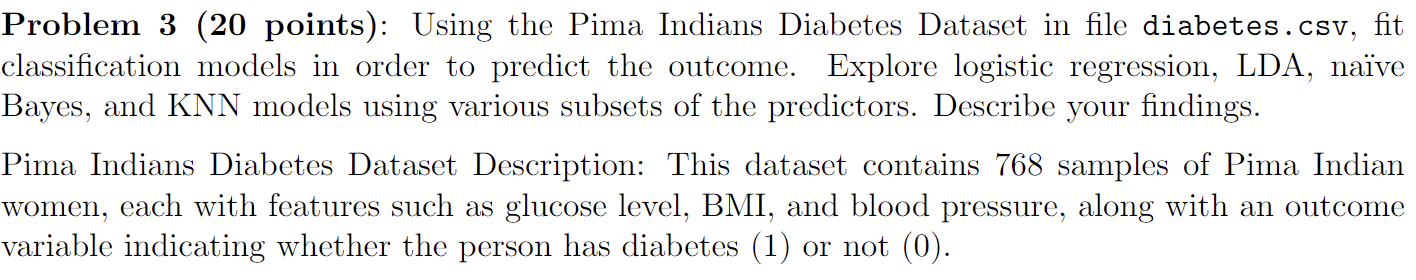


Above are the False Positive and False Negative Rates obtained using LDA.

Below are the False Positive and False Negative Rates obtained using Logistic Regression:



The FPR and FNR are quite similar for both Logistic Regression and LDA, however it is observed that the Logistic Regression performs marginally better.



Models were developed with the following sets of predictors:

1. 'Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', and 'Age'

2. 'Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', and 'Insulin'

3. 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', and 'Age'

Overall, the Logistic Regression Model performs best for this problem.

