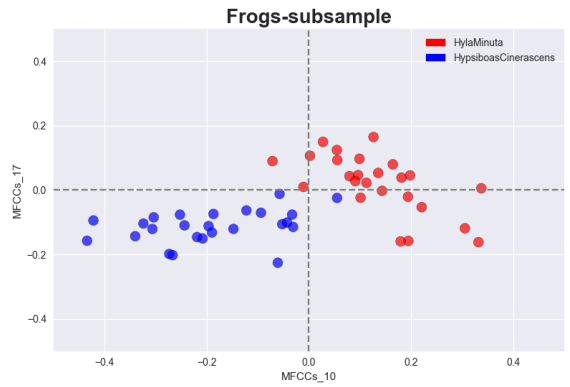
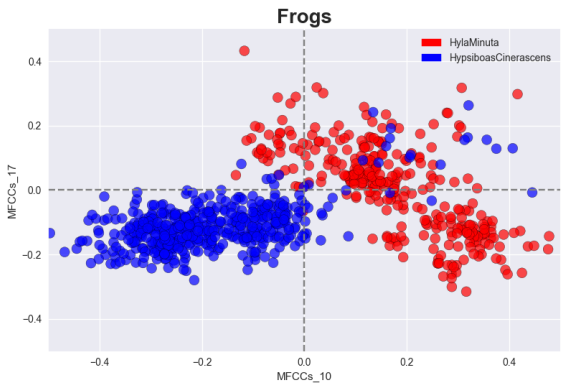
Q1)

Plots:

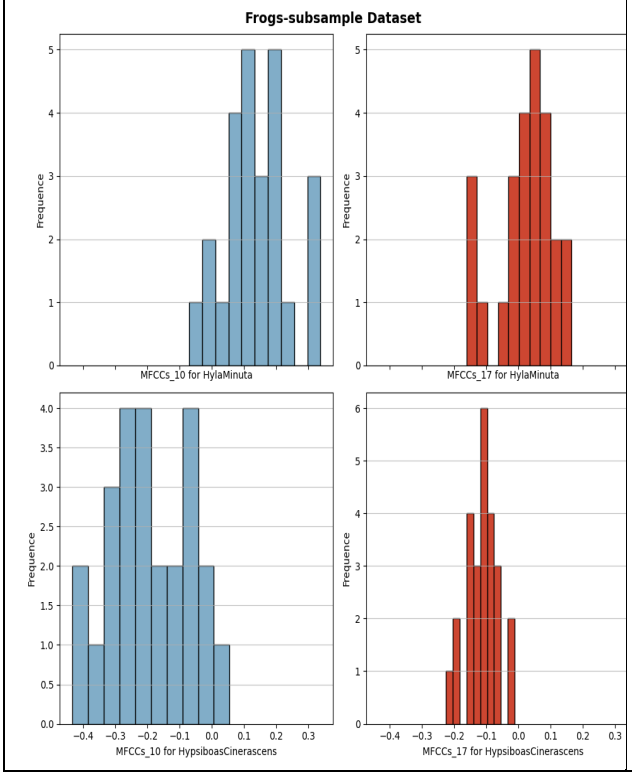
Scatter Plot:



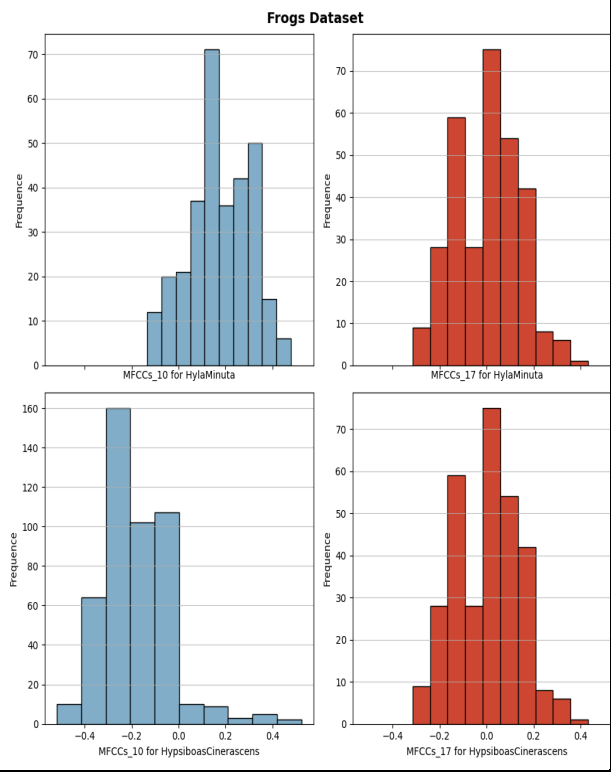
Scatter Plot:



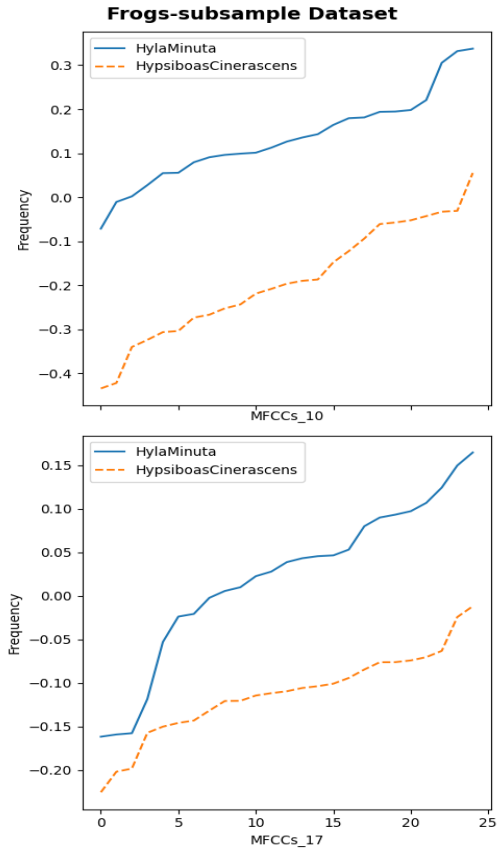
Histogram Plot:



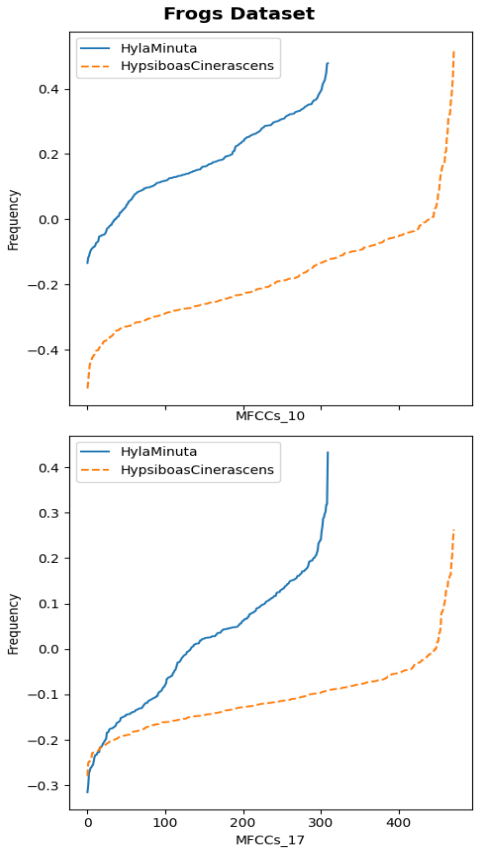
Histogram Plot:



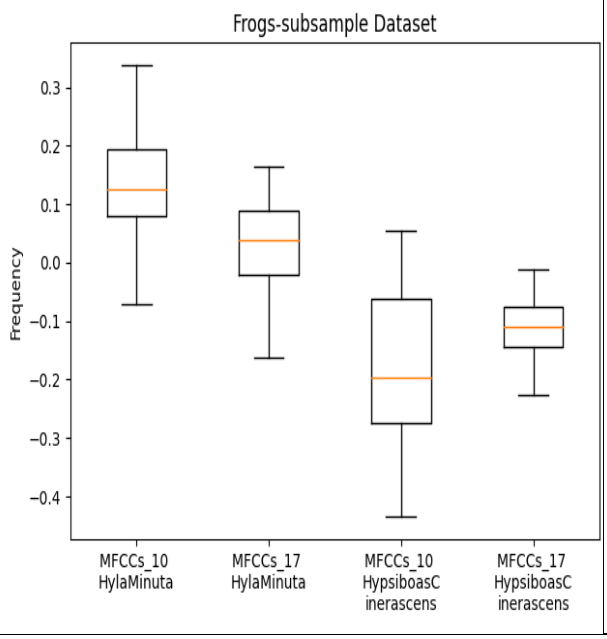
Line Graph:



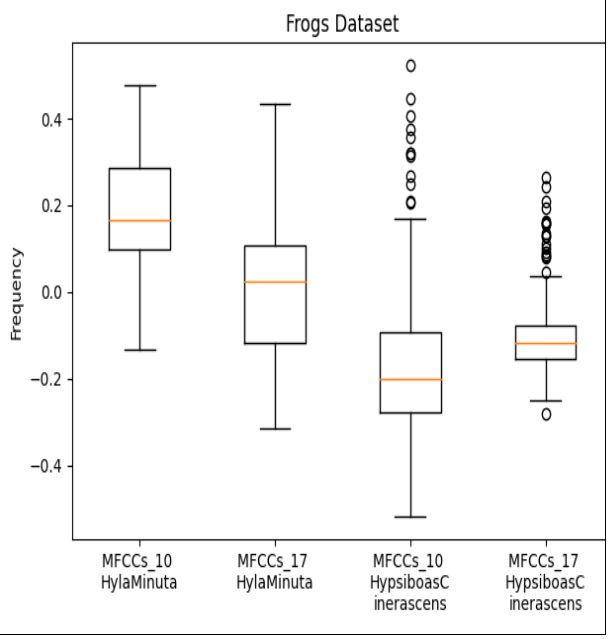
Line Graph:

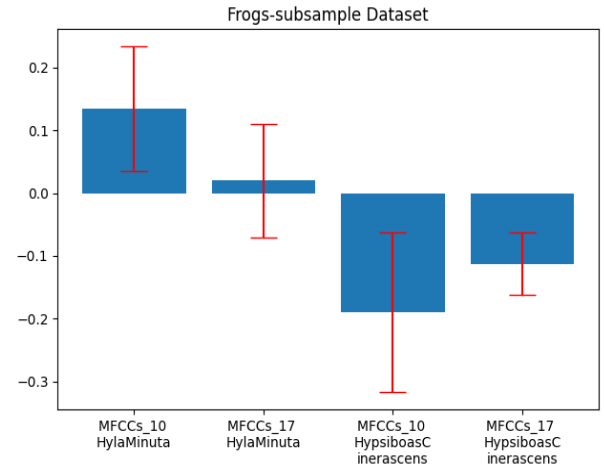
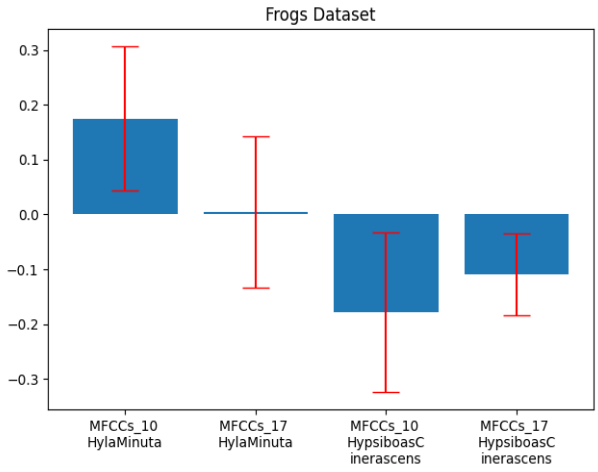


Box Plot:



Box Plot:



Bar Graph with Error Bar:	Bar Graph with Error Bar:
	

descriptive statistics:

Frogs-subsample Dataset	Frogs Dataset
MFCCs_10 mean= -0.02801130178	MFCCs_10 mean= -0.038321155240409216
MFCCs_17 mean= -0.046330259160000005	MFCCs_17 mean= -0.06451480131585678
co-variance of MFCCs_10= 0.040059573344532806	co-variance of MFCCs_10= 0.049606201135078756
co-variance of MFCCs_17= 0.009952778615942902	co-variance of MFCCs_17= 0.014038896296835126
standard deviation of MFCCs_10= 0.19813728038317813	standard deviation of MFCCs_10= 0.2225820435148893
standard deviation of MFCCs_17= 0.09876093885552142	standard deviation of MFCCs_17= 0.11841006605870075

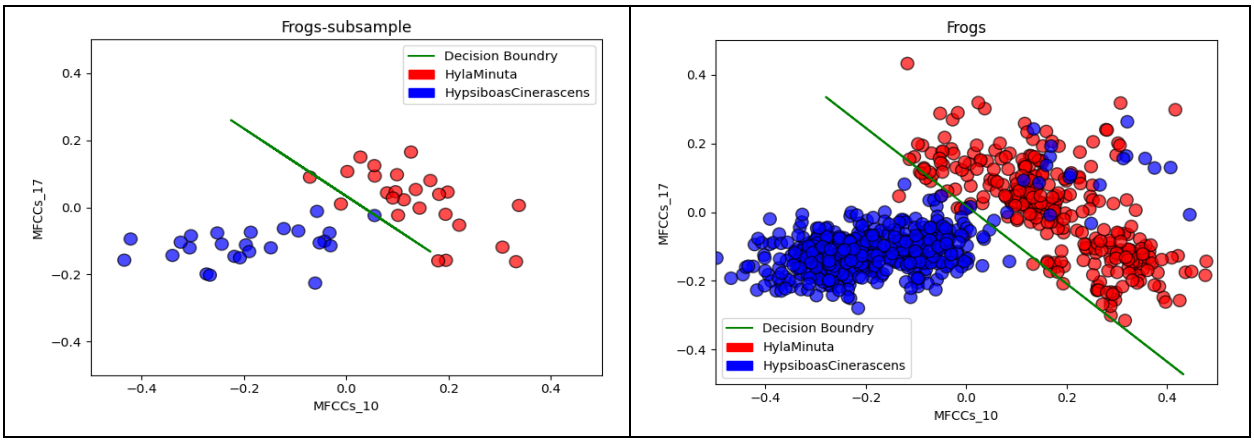
From the box plots it is visible that MFCCs_10 feature for Hyla Minuta class lies mostly in the positive range whereas for the Hypsiboas Cinerascens class lies mostly in the negative range with some exceptions (outliers) shows with circles in the plot. The mean of this feature lies approximately at -0.028 for the Frogs-subsample dataset whereas it is approximately -0.038 for the Frogs dataset. The deviation between two points in MFCCs_10 is about 0.198 for the Frogs-subsample dataset whereas it is 0.223 for the Frogs dataset. The line graph of the feature MFCCs_10 denotes a sigmoid like curvature for the Frogs data set whereas that of Frogs-subsample dataset shows a mild sigmoid like curve almost close to a linear equation. From the Histogram, we can learn that the most frequent occurring numbers in MFCCs_10 for class Hyla Minuta are in the range 0.1 to 0.2 whereas for class Hypsiboas Cinerascens are in the range -0.2 to -0.3.

The feature MFCCs_17 for Hyla Minuta does not lie mostly in the positive range but is evenly balanced between positive and negative ranges whereas the Hypsiboas Cinerascens class lies mostly in the negative ranges with some exceptions (outliers) shows with circles in the box plot. The mean of this feature lies approximately at -0.046 for the Frogs-subsample dataset whereas it is approximately -0.065 for the Frogs dataset. The deviation between two points in MFCCs_17 is about 0.1 for the Frogs-subsample dataset whereas it is 0.12 for the Frogs dataset. The line graph of the feature MFCCs_17 denotes a sigmoid like curvature for the Frogs data set whereas that of Frogs-subsample dataset shows a mild sigmoid like curve almost close to a linear equation. From the Histogram, we can learn that the most frequent occurring

numbers in MFCCs_17 for class Hyla Minuta are in the range 0.0 to 0.1 whereas for class Hypsiboas Cinerascens the most frequent range is not detectible.
While comparing the two classes Hyla Minuta (HM) and Hypsiboas Cinerascens (HC) respectively, it can be seen from the scatter plot that the HM class lies in the bottom left of the graph (with some outliers) indicating that they have comparatively lower values of MFCCs_10 and MFCCs_17 than the HC class. Also, it can be seen that HC lies in the top right of the graph (with some outliers) indicating that they have comparatively higher values of MFCCs_10 and MFCCs_17 than the HM class.

Q2)

Decision boundary for the binary classifier built using a single logistic regressor:



In order to obtain the decision boundary, we substitute the values of theta and x (input data) into the equation of line y = mx + c to get y (output data). When this line is plot along with the scatter plot as shown in the above figure, we get the decision boundary. This boundary separates the two classes HM and HC from each other. The boundary that I obtained has an accuracy of greater than 95% for both the datasets given. Which means that for the random testing data set created by the program q2.py, the predicted output is closely matched to what the actual output should be.
Since the Frogs-subsample dataset is small in quantity, it has very less outliers leading to a greater accuracy and a clearly separating decision boundary. But small data quantity also means less values to train the data leading to uncertainty about the correctness of the model for varied testing data sets. On the other hand, since the Frogs dataset is large, it covers a variety of possible cases which leads to a more stable accuracy even though it may not be a 100% but in overall cases will be better than the accuracy of the Frogs-subsample dataset.
The LogisticRegression class in the q2.py program has training and predicting methods which contain formulas to calculate the gradient descent and update the weights at every iterating through the training data to provide us with the model. When this model is provided with a testing dataset, the predicting function in the LogisticRegression class predicts the output values by using the model's values. When checked against the actual values, and calculating its percentage, we get the accuracy of our model. Lastly, by substituting the values of theta and the input data in the equation of the line, as mentioned above, in the first paragraph, we obtain our decision boundary.

Q3 Answers)

1. In creating a product, 85% are produced without defects. Of the products inspected, 10% of the good ones are seen as defective and not shipped, while only 5% of the defective products are approved and shipped. If a product is shipped, what is the probability that it has a defect?

Let Defective = D and Shipped = S
Given: P(~D) = 0.85; P(D) = 0.15; P(S|D) = 0.05; P(~S|~D) = 0.1; P(S|~D) = 0.9; P(~S|D) = 0.95
By using Bayes' theorem we get,
P(D|S)=(P(S|D)P(D))/(P(S|D)P(D)+P(S|~D)*P(~D)) = **0.0097**

2. Consider randomly generated bit strings of length four. Demonstrate whether or not the event of generating bit string with an even number of 1's is independent of the event producing bit strings that end in 1.

   Let A be the event of generating bit string with an even number of 1's and let B be the event producing bit strings that end in 1.

   Events A and B are independent if and only if → $P(A \cap B) = P(A) \cdot P(B)$

   All the possible strings of length 4 are as follows:

0	0	0	0
0	**0**	**0**	**1**
0	0	1	0
0	**0**	**1**	**1**
0	1	0	0
0	**1**	**0**	**1**
0	**1**	**1**	**0**
0	**1**	**1**	**1**
1	0	0	0
1	**0**	**0**	**1**
1	**0**	**1**	**0**
1	**0**	**1**	**1**
1	**1**	**0**	**0**
1	**1**	**0**	**1**
1	1	1	0
1	**1**	**1**	**1**

   As clearly seen 6 out of 16 strings are of event A.
   $P(A) = \frac{6}{16} = \frac{3}{8}$

   And 8 out of 16 are of event B.
   $P(B) = \frac{8}{16} = \frac{1}{2}$

   $P(A \cap B)$ (probability of event A and event B occurring) = $\frac{3}{16}$

   $P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{3}{8} = \frac{3}{16} = P(A \cap B)$
   Hence, events A and B are independent.

3. Let's flip a (fair) coin n times to generate a dataset, where we choose to represent the state of the coin, i.e., heads or tails, as the variable X = {0, 1}, where the first attribute value represents

"tails" and the other "heads". Suppose that the experiment's outcome yields the following state sequence:

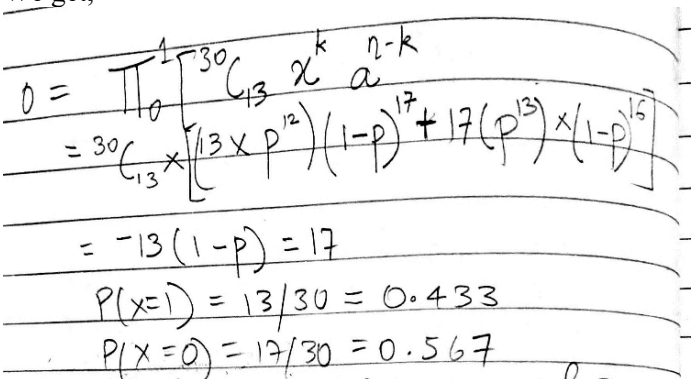
S = {1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0}

Estimate the probability that P(X = 1) from the data, using a maximum likelihood approach (Hint: the frequency approach). Also, what is the probability of getting tails under the sequence above, or P(X = 0)?

The derivative of P(X=1):

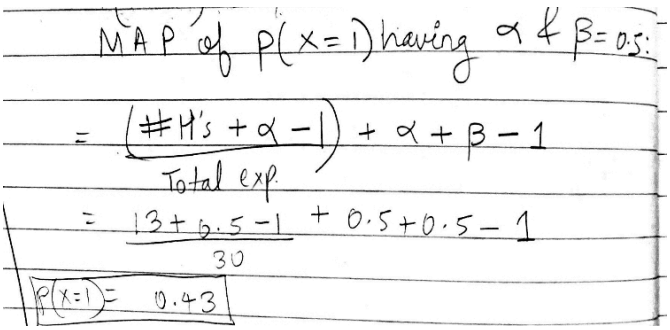
$\frac{d}{dp}L(p) = \prod_{k=0}^{n} {}^{n}_{k}C * x^k * a^{n-k}$

We have 13 heads out of a sample of 30-coin tosses.
Using notation pˆ for the MLE, set the above derivative to 0.
We get,



Probability that p(X = 1) = **0.433**
Probability that p(X = 0) = 1 – p(X=1) = 1- 0.433 = **0.567**

4.



Maximum a posteriori = **0.4**