1)

mean: $\begin{vmatrix} 2 \\ 3 \end{vmatrix}$ covariant: $\begin{vmatrix} 2 & 2 \\ 2 & 3 \end{vmatrix}$

characteristic polynomial: $\begin{vmatrix} 2-\lambda & 2 \\ 2 & 3-\lambda \end{vmatrix} = (2-\lambda)(3-\lambda) - (2)(2) = \lambda^2 - 5\lambda + 2$

$\lambda_1, \lambda_2 = \frac{5 \pm \sqrt{(-5)^2 - 4(1)(2)}}{2(1)}$

$\lambda_1 = \frac{5-\sqrt{17}}{2} = 0.438 \qquad \lambda_2 = \frac{5+\sqrt{17}}{2} = 4.56$

$A - \lambda_1 I = \begin{vmatrix} \frac{\sqrt{17}-1}{2} & 2 \\ 2 & \frac{\sqrt{17}+1}{2} \end{vmatrix} \qquad A - \lambda_2 I = \begin{vmatrix} \frac{-\sqrt{17}-1}{2} & 2 \\ 2 & \frac{-\sqrt{17}+1}{2} \end{vmatrix}$

$v_1 = \begin{vmatrix} \frac{-\sqrt{17}-1}{4} \\ 1 \end{vmatrix} = \begin{vmatrix} -1.281 \\ 1 \end{vmatrix} \qquad v_2 = \begin{vmatrix} \frac{\sqrt{17}-1}{4} \\ 1 \end{vmatrix} = \begin{vmatrix} 0.781 \\ 1 \end{vmatrix}$

2)

The variance explained by the first component $4.5616/(4.5616+0.4384) = 0.9123$.

3)

The variance explained by the second component is $0.4384/(4.5616+0.4384)=0.0877$.

4)

Code is included in the Jupyter notebook.

Question 1

What is the difference between the value of the coefficients of L1 versus L2? Which one do you think is better?

The difference is the coefficients for L1 has many zeros. The coefficients are for L2 are all nonzero. Computationally, the L1 is probably better as having the zeros saves time. The L2 which keeps all the coefficients non zero and may have better accuracy at the expense of longer computational times.

Question 2

What do you observe from the plot above?

As C increases (decrease in regularization), we see the weight coefficient get more spread out. And we are more able to see which weights are more influential. As C decreases (increase in regularization), we see all the coefficients go to zero.

Question 3

Form the plot above, can you make the decision what is the best number of feature to keep for this dataset?

From the plot above, we see after 5 features, the accuracy does not seem to increase. Therefore, we should keep 5 features.

Question 4

What feature and how many features do you want to keep based on the feature importance of decision tree model?

From the decision tree model, we see there's a big dip in variable importance after Hue. Therefore, we should keep 6 features.

Question 5

Compare LDA to PCA algorithm. Which one works better for this problem?

The LDA algorithm works better because it has better separation between classes than PCA algorithm from looking at the graph.