Few comments:

1 b) Scaling the data helps in centering the data around zero, which means that the mean of each feature becomes approximately zero. This can help the algorithm to converge quicker since it doesn't have to deal with large mean values which could slow down the optimization process.

The variance of the features has also been reduced, which is important when dealing with features with different scales. It means that no single feature will dominate the others in the logistic regression model.

1 c) l2 regularisation helps to prevent the case of overfitting. If we consider linearly separable data, gradient descent will cause the value of c (where w= cw\_ ) to grow out of bound. This leads to the sigmoid function tending towards the heavyside function. This results in there being no finite solution as the model would be infinitely steep at the inflection point.

The l2 regulariser helps circumvent this problem by ensuring that gradient descent will converge to a globally unqiue solution for any input data. Data scaling is a useful preprocessing step before applying the regularization because regularization methods like l2 are sensitive to the scale of the features. As discussed in the previous part, it ensures no indvidual feature dominates the process and speeds up convergence.

2 a) Is it worth transforming the data as well?

X\_train\_unscaled, X\_test\_unscaled, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.7, random\_state=42)scaler = StandardScaler()X\_train\_scaled = scaler.fit\_transform(X\_train\_unscaled)X\_test\_scaled = scaler.transform(X\_test\_unscaled)

Q6

The alpha parameter controls the strength of additive smoothing which can be used to handle the issue of zero probabilities. It is added to the frequency counts of features and class labels during model training to avoid zero probabilities when estimating probabilities in the training data.

This parameter is useful in particular when a discrete feature value is absent in the training set, which is a problem because if we assume the probability of that feature to be zero when an event happens, it's a form of overfitting since it might be present in test set or real life examples. However we must choose the alpha-value carefully as if it is too large it may lead to a case of underfitting.