Question 4:

4.1.(d) Implemented feature extraction and generated training and testing sets

4.2.(b) If class proportions aren't maintained across the folds, the training set can be a poor representation of the underlying distribution. By stratifying the folds, the training set becomes a better sample of the dataset and the test set accuracy can also be improved.

4.2.(d)

C	Accuracy	F1-score	AUROC
0.001	0.7089419539640778	0.8296828227419593	0.8105494821634063
0.01	0.7107437557658796	0.8305628004640422	0.8110783467587265
0.1	0.8060326761654195	0.875472682955829	0.8575527426160339
1.0	0.8146271113085273	0.8748648327495685	**0.8712327387802071**
10	**0.8181827370986664**	**0.876562152886752**	0.8695790180283852
100	**0.8181827370986664**	**0.876562152886752**	0.8695790180283852
Best *C*	**10/100**	**10/100**	**1.0**

The performance for accuracy and F1-score increase gradually until they reach a plateau at C = 10, where the performance metric stays the same. For AUROC, the performance metric increases until C = 1.0. It then decreases as C increases to 10, where it plateaus similar to the other metrics.

For the next few questions, I have assumed C = 10 as the best hyperparameter for the first two metrics.

4.3.(c)

C	Metric	Performance on test set
10	Accuracy	**0.7428571428571429**
10	F1-score	**0.43749999999999994**
1.0	AUROC	**0.7405247813411079**

The accuracy and AUROC measures were pretty similar with a value of ~0.74. However, the F1-score reduced drastically from the training to the test set.