Report

1.1 There is no informative feature in the data set since individual features' data has a close distribution of data range amongst the two classes.

1.2 No, the data does not contain a learnable pattern as there is no possible function that can be used to map the input to output with a good enough accuracy in this case. This is because the normal distributions with outputs 1 and 0 have an almost entirely overlapping range and were assigned an output of 1 or 0 randomly.

1.3 In terms of accuracy we could expect 55.5% accuracy as the average accuracy obtained after running the ML algorithm several times with random seeds gives the average accuracy of 0.56 in step 6.2.

2.1

Number of neighbors (k)	Accuracy obtained step 4.2	Accuracy obtained step 6.2
1	1	0.56
3	0.72	0.48
5	0.76	0.56
7	0.72	0.6
9	0.64	0.44

2.1 The calculated Accuracy of the prediction with 1 neighbor where 100% of data is used as training data is equal to 1.0 and the highest obtained. This is because when the data looks for the closest neighbor of the test in the training data, it will find the same data with 0 distance, hence the accuracy would be 100%.

2.2 The drop in accuracy in step 6.2 is because now we have two sets of training and test data, so when testing the data, the same data which was used for training won't be used for testing therefore the accuracy will decrease. When I run step 6.2 1000 times, the mean accuracy would be reported as:
K=1: 0.76 Accuracy
K= 3: 0.56 accuracy
K= 5 neighbors 0.56 accuracy
K= 7 neighbors: 0.48 accuracy
K=9 neighbors: 0.36 accuracy

2.3 The best performing model would be the model with K=5. Although the accuracy of the model with K=1 is high, it would not be able to classify very well as it would over-fit the given data (high variance). The models with K=7 and K= 9 on the other hand would suffer from high bias, hence can't generalize very well. Also, out of all the other models K=3 and K=5 is shown to have the highest accuracy. Since, k=5 is shown to empirically yield test errors that suffer neither from high bias nor from very high variance we chose it as the best performing model. That being said from the training/test error vs k-values graph attached, it is deduced since the proposed

model over fits the data even with a relative high accuracy amongst other models(different k neighbours) it will have a high generalization error rate.

2.4 As k grows in a KNN bias increases and variance decreases so the models with the lowest bias and highest variance will be with k=1 on both step 4.2 and step 6.2.