2a)

The machine learning model predicts 90 as having X but the data shows only 10 actually have X and the rest are false positives. The information we know is that the victims shows no symptoms and it affects a small proportion of the population. Due to these limitations being a property if the problem domain it is infeasible to obtain more information and a class imbalance occurs. The minority class(victims) is harder to predict because there are too few examples of this class and makes it harder for the model to learn the characteristics of this class. This means to train a model to predict if someone has this disease the data set has too little information to distinguish between healthy and sick. A performance metric we can use to is classification accuracy. Here the classification accuracy is (No of predictions/Total no of predictions made) 10/90 is 11.11%.

Another method to measure performance metric we can use the confusion matrix. There are 10 true positives, 10 true negatives, 80 false positives and 0 false negatives. The accuracy is (TP+TN)/TOTAL=20/100=0.2, the precision (TP/(TP+FP)) is 10/90=0.1111, the recall is (TP/(TP+FN)) is 10/10=1 and the F1 score (2*precision*recall)/(Precision*recall) is (2*0.111*1)(1+0.111)= 0.1998. Using the F1 score of a classifier dealing with imbalanced data we can say that the higher the F1 score the better the model.

2b)

This is not the right approach as the features added could have very little to no relation to the desired output , provide no information gain and it might take the algorithm much longer to process the unnecessary features.. Ex- In the titanic model if we use the Name feature to model if a passenger survived or not, it could give unexpected results to different test cases. Then there is the introduction of variance error.This leads to the algorithm being highly sensitive to high degrees of variation in your training data, which can lead the model to overfit the data that is the model memorizes the data rather than generalize it. Adding more data without adding features in this case, can be helpful as increasing the data provided increases the accuracy and can help generalize the model but the data also needs to be free of outliers.