Question 1

a) Prove by contradiction
Assume that there exists one leaf which contains no training data for one chosen attribute, and because this is a binary decision tree, all of the training data of that chosen attribute will in another attribute value path. According to the equation of IG, $I(X,Y) = H(Y) - H(Y|X) = H(Y) - P(attribute = only\ on\ attribute\ value) * H(Y|only\ on\ attribute\ value)$, where $P(attribute = only\ on\ attribute\ value) = 1$, and all of training data under that attribute value path; therefore, $H(Y|only\ on\ attribute\ value)$ = H(Y), which mean IG = 0.

Thus, contradicting our assumption that choose a node according to IG > 0.

b) The worst case for both randomly choosing the IG>0 and choosing best IG, is same, which is that to perfectly classify all the samples, when there is too much noise in the training data. The maximum number for both methods is n.

Question 2

a)

Tic-tac-toe dataset mean and variance of accuracy based on info gain to choose attribute

```
accuarcy tic tac toe ig:  0.8558947368421053
variance tic tac toe ig:  0.0011274238227146813
```

Tic-tac-toe dataset confusion matrix based on info gain to choose attribute

```
[[29  4 33]
 [ 3 59 62]
 [32 63 95]]
```

Wine dataset mean and variance of accuracy based on info gain to choose attribute

```
accuarcy wine ig:  0.9352941176470588
variance wine ig:  0.003480968858131488
```

Wine dataset confusion matrix based on info gain to choose attribute

```
[[ 4  0  0  4]
 [ 0  7  0  7]
 [ 0  0  6  6]
 [ 4  7  6 17]]
```

From the mean accuracy results, we can get that the Tic-tac-toe has a lower mean accuracy compared with the result of Wine dataset. The main reason I think is that for each feature in Tic-tac-toe dataset has three feature value when the feature value for each feature in Wine dataset is 2 (I use binary split the continuous feature) , therefore, we have more times choosing best feature using information gain in Tic-tac-toe dataset, which could lead to a local optimization result. We need to focus on the confusion matrix that, only predicted positive while the actual is negative, misprediction happens, this could be caused by this kind of local optimization.

I believe that another reason for a high error rate could be that the test sample has some feature value could not be trained in the train dataset. To solve this, we need more data to for training, however, this could lead to overfit.

For the variance result, the Wine dataset has a higher variance. The reason is that the test data for the Wine dataset is too small, only 17 samples. So if the trained decision tree predict a wrong result on test data, it has a large impacts on the mean accuracy; therefore, the Wine dataset has a higher variance

For confusion matrix, I choose the best prediction in 100 tests, according to 10-times-10-fold cross validation, which clearly reflect the mean accuracy and variance of error rate results in the experiments. Even in the best result confusion matrix, the error rate of Tac-tic-toe dataset cannot be zero, and in the confusion matrix for Wine dataset, the best result has zero error rate. And I can guess that the worst mean accuracy for the Wine dataset is lower than that of Tic-tac-toe dataset, because the Wine dataset has a higher variance of mean accuracy.

b)

Tic-tac-toe dataset mean and variance of accuracy based on gain ratio to choose attribute

```
accuarcy tic tac toe gain ratio:  0.8614736842105264
variance tic tac toe gain ratio:  0.0007964542936288088
```

Tic-tac-toe dataset confusion matrix based on gain ratio to choose attribute

```
[[29  1 30]
 [ 5 60 65]
 [34 61 95]]
```

Wine dataset mean and variance of accuracy based on gain ratio to choose attribute

```
accuarcy wine gain ratio:  0.9335294117647059
variance wine gain ratio:  0.004647058823529412
```

Wine dataset confusion matrix based on gain ratio to choose attribute

```
[[ 0  0  0  0]
 [ 0 10  0 10]
 [ 0  0  7  7]
 [ 0 10  7 17]]
```

The results for Gain ratio are similar to that of the information gain. The reason is that, the gain ratio addresses the problem that the feature has more feature value, which lead to a higher chance to be chosen as a best feature as node. However, in this question, both dataset feature value are the same for each feature in both dataset, i.e. in Wine dataset, the number of feature value for each feature is 2, while in Tic-tac-toe, the number of feature value for each feature is 3. Therefore, using gain ratio method cannot improve a lot.

Question 3

A)

a)

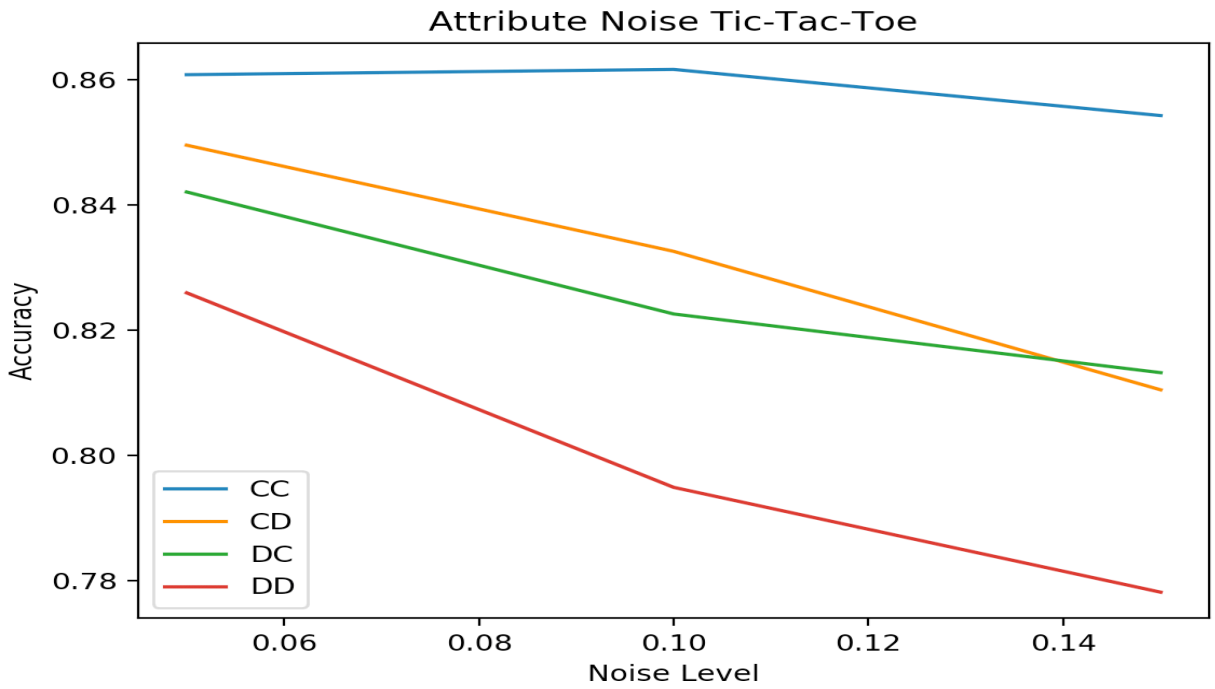

Figure 1

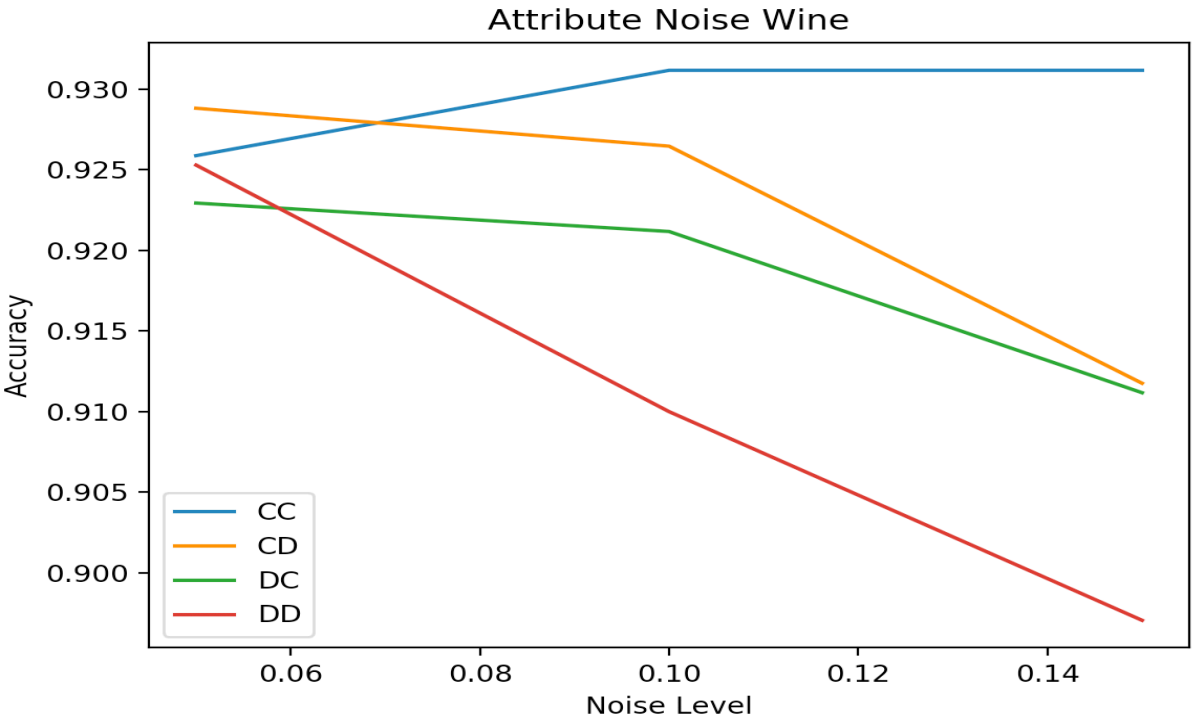

Figure 2

b)
From Figure1, we can see that in these four situations, the CC has the most stable result with the increase of noise level. And the mean accuracy for the other three situations decreases steadily with the rise of noise level. Besides, we can see that the DD has the lowest mean accuracy, which means that when we dirty both train and test dataset, we get the worst result. For DC and CD, they have a similar trend and test result, therefor we might get that we no matter we dirty the train dataset or test dataset, it has a similar impact on our result.

From Figure 2, we can see that when the noise level is 0.05, it does not have a big impact on the test result. For DC, CD and DD, they all have a decreasing trend; however, when the noise level increases from 0.1 to 0.15, the mean accuracy for CD and DC decreases more rapidly; meanwhile, the error rate for DD is the highest. The error rate for CC, is both most stable and lowest, with the increase of noise level.

Compared Figure 1 and Figure 2, we can also get that the attribute noise has more impact on the nominal cases, because for the numerical samples, the subtle noise could not change the classification for most of the attribute value.

B)

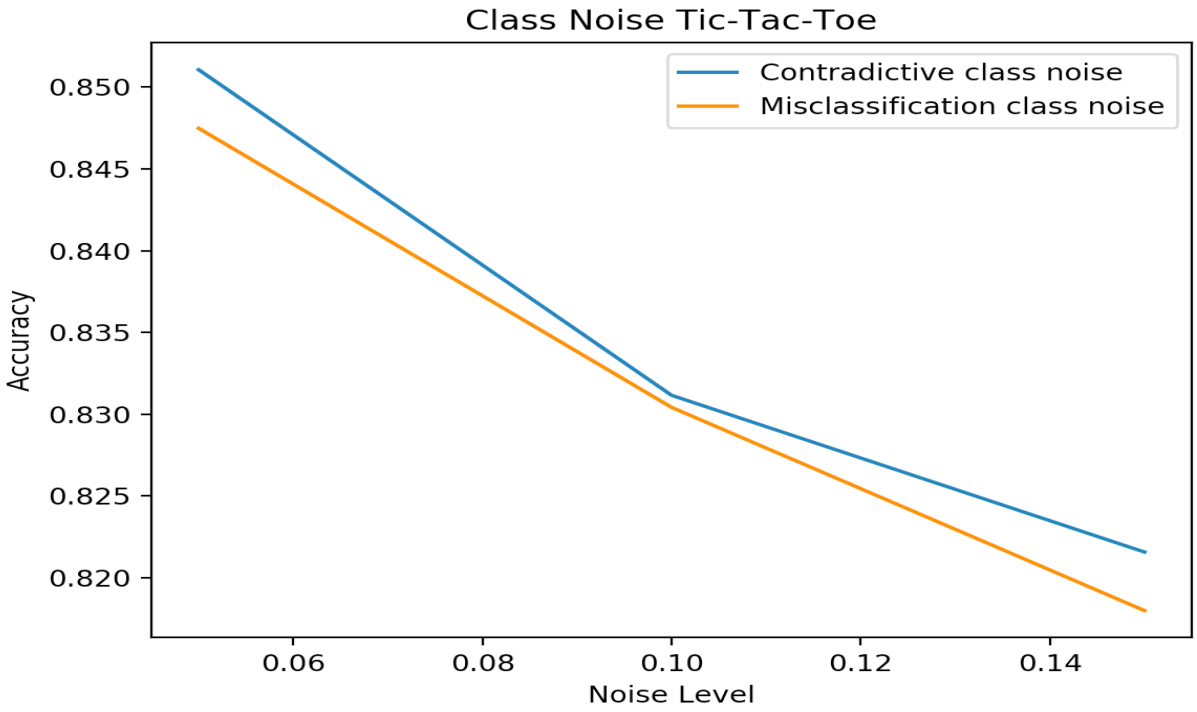

Figure 3

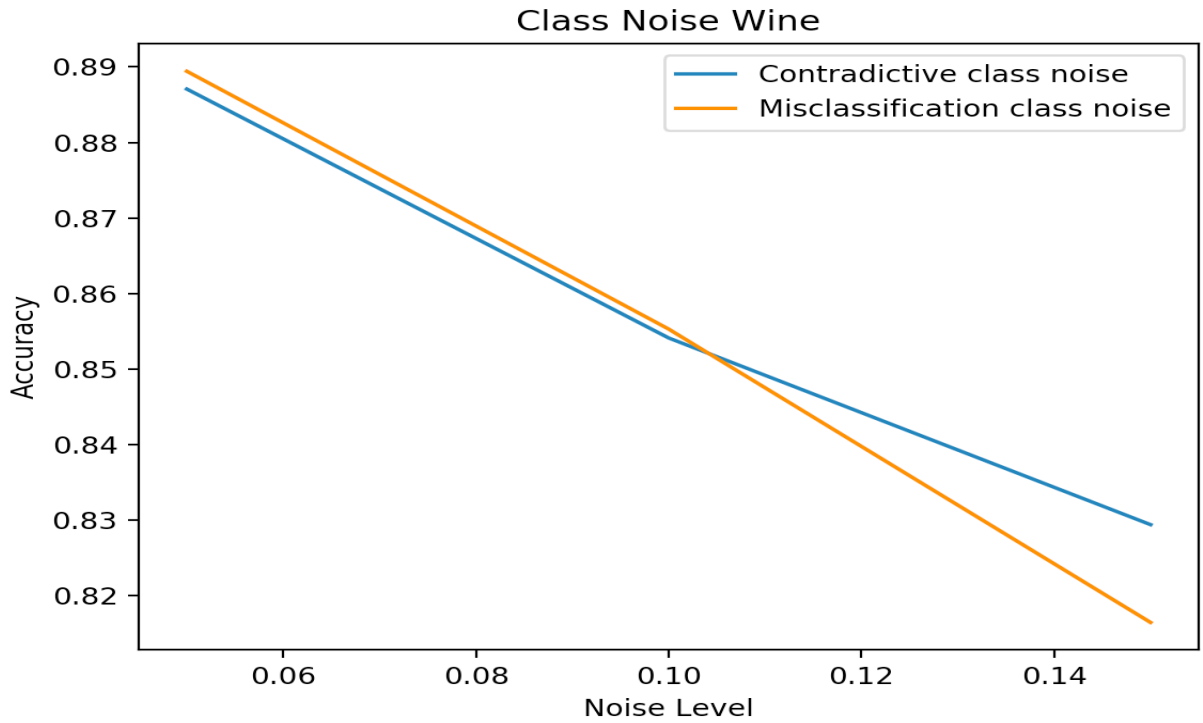

Figure 4

c)
From Figure 3 and Figure 4, we can get that the accuracy for both situations (contradictive class noise and misclassification class noise) decreases steadily with the rise of noise level rate.

When there are some class noises in train dataset, the decision tree will be constructed bigger to completely classify all the samples according to other attributes, which should not be used in the conflict free decision tree; therefore, the built decision tree will have a worse performance on test dataset, because of this kind of overfitting.

d)

From Figure 2 and Figure 4, which are results of wine dataset, we can clearly see that the class noise has more harmful impact on results.

This is because the class noise will have 100% percent to construct a bigger decision tree to lead to overfitting problem, reducing the performance of our decision tree; however, for numerical cases, the attribute noises have a large chance not to change the decision tree built.

From Figure 1 and Figure 3, which are results of tic tac toe dataset, we can not clearly see that the class noise has more harmful impact on results, because the DD situation has the worst test result compared with class noise. The class noise has a similar result as the DC and CD.

For DD case in tic tac toe dataset, the dirty training set has an impact on the building the decision tree, which has a similar result as class noise; Furthermore, the dirty test set also has a similar result as class noise; therefore, the DD case clearly has a worse performance compared with class noise in tic tac toe dataset.