Mohita Chaudhary
20830560

SYDE-675 Assignment 2

Question 1:

If we select a node randomly from the Binary Decision Tree from nodes with Information Gain greater than 0 then,

a) **Each leaf of the tree contains at least one training data.**
b) **If we have n training data, what is the maximum number of leaves in the constructed decision tree. Compare the result with the state that we used Information Gain for selecting nodes.**

Solution:

a) If we choose any node randomly with IG>0, then definitely ID3 we will have at least one training data in each leaf. This is because ID3 perfectly classifies each training data, by perfectly we mean that on selecting an impure node it will split the node till it achieves a pure node(all the labels are the same for the training data). So, in a worst case scenario the leaf node will for sure have 1 data point.
b) The maximum number of leaves in the binary decision tree if we choose the nodes randomly would be **n.** Since, in the worst case scenario every time when a random node is chosen there is a possibility for the binary decision tree to be completely left skewed or completely right skewed and there would be a unique node for all the training data points and hence, for n training points we will obtain **n** leaves.

Question 2:

a) **To design a ID3 classifier and use it to classify the Tic Tac Toe Endgame and Wine Dataset. Report the mean accuracy, variance of accuracy and the Confusion matrix. Result using the Information Gain.**

Solution:

The Information Gain was used to build the decision tree recursively, to find the Information gain the following equation was used:

$$Information\ Gain = Entropy - Conditional\ Entropy$$

$$IG(X, Y) = H(X, Y) - H(X|Y)$$

here $H(X, Y)$ *is the entropy and* $H(X|Y)$ *is the conditional entropy*

Also, formula for entropy and conditional entropy are:

$$H(X) = -\sum_{x_i \varepsilon X} P(x_i)\ log_2 P(x_i)$$

$$H(X|Y) = -\sum_{i}\sum_{j} P(X = i,\ Y = j)\ log\ P(Y = j|\ X = i)$$

Mohita Chaudhary
20830560

For the Tic Tac Toe Endgame:

1. Mean Accuracy: **0.8417554985554163**
2. Variance of the mean: **0.0018504795714426481**
3. Confusion Matrix: **[[23 8]**
   [7 54]]

Following is the screenshot of the result:

```
(base) mohita@Mohitas-MacBook-Pro ~ % conda activate base
(base) mohita@Mohitas-MacBook-Pro ~ % /Users/mohita/anaconda3/bin/python /Users/mohita/Documents/trail.py
FOR TIC TAC TOE DATA
Accuracy with Information Gain for 1 th time is 0.8401535087719297
Accuracy with Information Gain for 2 th time is 0.8406633771929826
Accuracy with Information Gain for 3 th time is 0.8446747076023391
Accuracy with Information Gain for 4 th time is 0.8443366228070175
Accuracy with Information Gain for 5 th time is 0.8438947368421051
Accuracy with Information Gain for 6 th time is 0.8415241228070175
Accuracy with Information Gain for 7 th time is 0.8422305764411028
Accuracy with Information Gain for 8 th time is 0.8409251644736842
Accuracy with Information Gain for 9 th time is 0.8392190545808967
Accuracy with Information Gain for 10 th time is 0.8399331140350877
Accuracy with Information Gain for 10 iterations is 0.8417554985554163
The standard deviation for Information Gain for 10 iterations is 0.0018504795714426481
Confusion Matrix:
 [[23  8]
 [ 7 54]]
```

For the Wine data-set:

1. Mean Accuracy: **0.9406915395787945**
2. Variance of the mean: **0.004933564336759629**
3. Confusion Matrix: **[[2 0 0]**
   [2 7 1]
   [0 1 4]]

Following is the screenshot of the result:

```
FOR WINE DATA
Accuracy with Information gain for 1 th time  is  0.9375816993464052
Accuracy with Information gain for 2 th time  is  0.948529411764706
Accuracy with Information gain for 3 th time  is  0.9449891067538125
Accuracy with Information gain for 4 th time  is  0.9459967320261438
Accuracy with Information gain for 5 th time  is  0.944313725490196
Accuracy with Information gain for 6 th time  is  0.9432461873638344
Accuracy with Information gain for 7 th time  is  0.9359477124183005
Accuracy with Information gain for 8 th time  is  0.9361111111111112
Accuracy with Information gain for 9 th time  is  0.9350036310820624
Accuracy with Information gain for 10 th time  is  0.9351960784313725
Accuracy with Information Gain for 10 iterations is 0.9406915395787945
The standard deviation for Information Gain for 10 iterations is 0.004933564336759629
Confusion Matrix:
 [[2 0 0]
 [2 7 1]
 [0 1 4]]
```

Mohita Chaudhary
20830560

Analysis of the above results:

1. We observe that the accuracy which we attain 10 times for the 10 fold cross validation is almost very same every time for both the datasets. This is also verified from the standard deviation obtained for each dataset which is very less, showing that all the accuracies are more or less the same. This clearly shows that the training of both the datasets converges to some minima, thus global minima is obtained each time.
2. The mean accuracy for the wine dataset is more than the mean accuracy for the Tic tac Toe dataset that is clearly because we know that the number of training samples in Tic tac toe dataset are more than the Wine dataset that might have lead to overfitting in the decision tree model, whereas the training for the wine dataset must have been more generalised and thus more accurate.
3. From the confusion matrix we deduce that the True positives and true negatives are more than the False positives and False negatives also we get to know that both the datasets are a little biased. The samples for the positives are more than the samples for negatives.

b) Result using the Gain Ratio:

Here the gain ratio was used to find the attribute for the best split and the equation for the Gain Ratio is as follows:

$$Gain\ Ratio(S,A) = \frac{Gain\ (S,A)}{SplitInformation(S,A)}$$

$$\text{Split Information (S,A)= -} \sum_{v\varepsilon\ Values(A)} \frac{|S_v|}{|S|} log_2 \frac{|S_v|}{|S|}$$

For this S_v is the subset for which attribute A has the value v.

<u>For the Tic Tac Toe Endgame:</u>

1. Mean Accuracy: : **0.8355777686403508**
2. Variance of the mean: **0.007960186403850527**
3. Confusion Matrix: **[[21 3]**
   [10 60]]

Mohita Chaudhary
20830560

Following is the screenshot of the result:

```
Accuracy with Gain Ratio for 1 th time  is  0.8140679824561403
Accuracy with Gain Ratio for 2 th time  is  0.8302576754385965
Accuracy with Gain Ratio for 3 th time  is  0.8373501461988304
Accuracy with Gain Ratio for 4 th time  is  0.8365021929824561
Accuracy with Gain Ratio for 5 th time  is  0.836188596491228
Accuracy with Gain Ratio for 6 th time  is  0.8366648391812866
Accuracy with Gain Ratio for 7 th time  is  0.8395394736842106
Accuracy with Gain Ratio for 8 th time  is  0.8396175986842106
Accuracy with Gain Ratio for 9 th time  is  0.842108918128655
Accuracy with Gain Ratio for 10 th time  is  0.8434802631578949
Accuracy with Gain Ratio for 10 iterations is 0.8355777686403508
The standard deviation for Gain Ratio for 10 iterations is 0.007960186403850527
Confusion Matrix:
 [[21  3]
 [10 60]]
```

For the Wine data-set:

1. Mean Accuracy: **0.9214396332607115**
2. Variance of the mean: **0.003591598844860223**
3. Confusion Matrix: **[[3 0 0]**
   [1 7 0]
   [0 0 6]]

Following is the screenshot of the result:

```
Accuracy with Gain Ratio for 1 th time  is  0.9202614379084968
Accuracy with Gain Ratio for 2 th time  is  0.9263071895424837
Accuracy with Gain Ratio for 3 th time  is  0.928540305010893
Accuracy with Gain Ratio for 4 th time  is  0.9236928104575164
Accuracy with Gain Ratio for 5 th time  is  0.9162091503267973
Accuracy with Gain Ratio for 6 th time  is  0.9178649237472768
Accuracy with Gain Ratio for 7 th time  is  0.9199346405228758
Accuracy with Gain Ratio for 8 th time  is  0.9194035947712418
Accuracy with Gain Ratio for 9 th time  is  0.9201888162672475
Accuracy with Gain Ratio for 10 th time  is  0.9219934640522877
Accuracy with Gain Ratio for 10 iterations is 0.9214396332607115
The standard deviation for Gain Ratio for 10 iterations is 0.003591598844860223
Confusion Matrix:
 [[3 0 0]
 [1 7 0]
 [0 0 6]]
```

Analysis of the above results:

1. The mean accuracy obtained from the Gain Ratio is quite similar to the mean accuracy obtained using the Information Gain for both the datasets.
2. The mean accuracy for the tic tac toe dataset is almost the same $\simeq 83\%$ whereas the mean accuracies for wine dataset is different for Gain Ratio and Information Gain.

Mohita Chaudhary
20830560

3. In the case when two features have the same entropy but different set of values for that feature then the Information Gain algorithm cannot decide which one is the best feature to select so a random feature is picked on the other hand the Gain ratio Algorithm selects the feature with the less set of values. Thus, the accuracy for the tic tac toe is almost same for both the algorithms since the set of values for all features is same on the other hand for the wine data the possible values for every feature is very different thus we observe more difference in the mean accuracies obtained using both the algorithms for wine dataset.

Question 3:

A) Analysis the effect of attribute noise:

a) Plots:

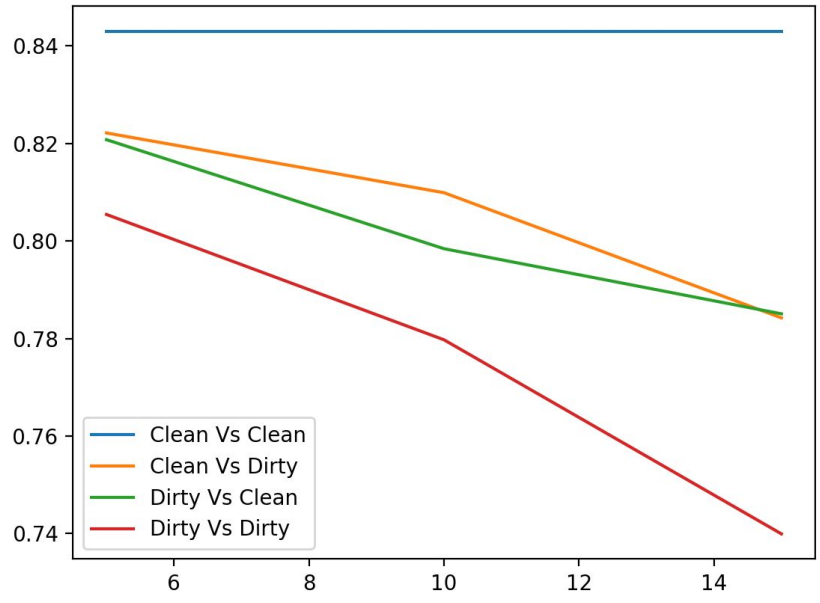
Tic Tac Toe Endgame

Following is the screenshot of the result for the Tic Tac Toe Endgame :

```
(base) mohita@Mohitas-MacBook-Pro ~ % /Users/mohita/anaconda3/bin/python /Users/mohita/Documents/trial22.py

 For Clean Train and Clean Test

The mean acciuracy is  0.8430197368421053 The standard deviation is 0.010939985042849251

 For Dirty Train and Clean Test

For noise level  5 % the mean accuracy is 0.8207850877192984 and the Standard Deviation is 0.005919454787653369

 For Dirty Train and Clean Test

For noise level  10 % the mean accuracy is 0.7984243421052631 and the Standard Deviation is 0.009852457910172967

 For Dirty Train and Clean Test

For noise level  15 % the mean accuracy is 0.7850679824561404 and the Standard Deviation is 0.012033834920856723

 For Clean Train and Dirty Test

For noise level  5 % the mean accuracy is 0.8221644736842105 and the Standard Deviation is 0.008700888590116383

 For Clean Train and Dirty Test

For noise level  10 % the mean accuracy is 0.8099111842105262 and the Standard Deviation is 0.009468900384631553

 For Clean Train and Dirty Test

For noise level  15 % the mean accuracy is 0.7842368421052631 and the Standard Deviation is 0.013612678889426221

 For Dirty Train and Dirty Test
For noise level  5 % the mean accuracy is 0.8054484649122807 and the Standard Deviation is 0.012920749892610573

 For Dirty Train and Dirty Test
For noise level  10 % the mean accuracy is 0.7797412280701754 and the Standard Deviation is 0.01410278714111996

 For Dirty Train and Dirty Test
For noise level  15 % the mean accuracy is 0.7399276315789474 and the Standard Deviation is 0.009082056474589515
```

Plot Tic Tac Toe Endgame (noise level vs accuracy)

Mohita Chaudhary
20830560



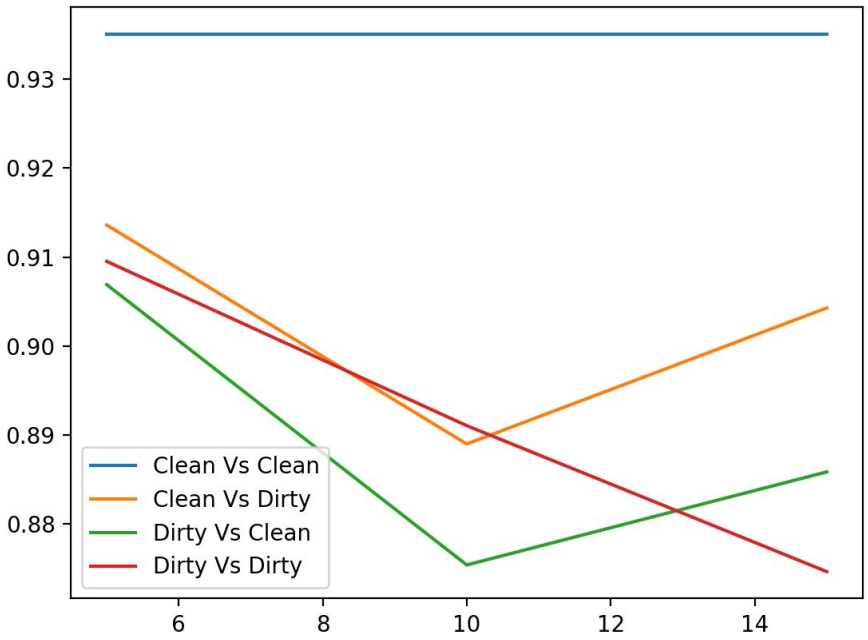
1. **Wine Dataset (noise level vs accuracy)**

   Following is the screenshot of the result for the Wine dataset :

```
mohita@Mohitas-MacBook-Pro ~ % source /Users/mohita/anaconda3/bin/activate
(base) mohita@Mohitas-MacBook-Pro ~ % conda activate base
(base) mohita@Mohitas-MacBook-Pro ~ % /Users/mohita/anaconda3/bin/python "/Users/mohita/Documents/Question1 copy.py"
/Users/mohita/Documents/Question1 copy.py:44: DeprecationWarning: object of type <class 'numpy.float64'> cannot be safely interpreted as
  pot_th=np.linspace(data_x[:,i].min(),data_x[:,i].max(),num=(data_x[:,i].max()-data_x[:,i].min()+0.5)/0.01)

 For Clean Train and Clean Test
The mean accuracy is 0.9350653594771241 and The Standard Deviation is 0.00819610188569678

 For Dirty Train and Clean Test
For noise level  5 % the mean accuracy is 0.9068954248366013 and the Standard Deviation is 0.016884043144245648

 For Dirty Train and Clean Test
For noise level  10 % the mean accuracy is 0.8753921568627451 and the Standard Deviation is 0.024904469535508292

 For Dirty Train and Clean Test
For noise level  15 % the mean accuracy is 0.8858496732026143 and the Standard Deviation is 0.011979743179937915

 For Clean Train and Dirty Test
For noise level  5 % the mean accuracy is 0.91359477124183 and the Standard Deviation is 0.017051935603879867

 For Clean Train and Dirty Test
For noise level  10 % the mean accuracy is 0.8889869281045751 and the Standard Deviation is 0.022511121889862096

 For Clean Train and Dirty Test
For noise level  15 % the mean accuracy is 0.9042810457516339 and the Standard Deviation is 0.02335055935785582

 For Dirty Train and Dirty Test
For noise level  5 % the mean accuracy is 0.9095098039215687 and the Standard Deviation is 0.024030365293714204

 For Dirty Train and Dirty Test
For noise level  10 % the mean accuracy is 0.8910457516339868 and the Standard Deviation is 0.022243453909635197

 For Dirty Train and Dirty Test
For noise level  15 % the mean accuracy is 0.874640522875817 and the Standard Deviation is 0.010535047562059835
```

Mohita Chaudhary
20830560

Plot Wine dataset (noise level vs accuracy)



b) Analysis of the plots:

1. For the Tic tac toe dataset the Dirty train vs Dirty test data gives the worst accuracy whereas the accuracies for CvsD and DvsC are quite similar, with CvsD still performing overall better than DvsC.
2. For the wine dataset the DvsD accuracy decreases with increasing the attribute noise we observe a clear negative slope. We observe DvsC giving the worst accuracy. We also observe CvsD performing better than both CvsC and DvsC.
3. The best accuracy is observed for the CvsC for both the datasets.
4. For the wine dataset we observe that for CvsD and DvsC first the accuracy decreases till 10% of noise but then it starts increasing after 10%.

B) Analysis the effect of class-label noise:

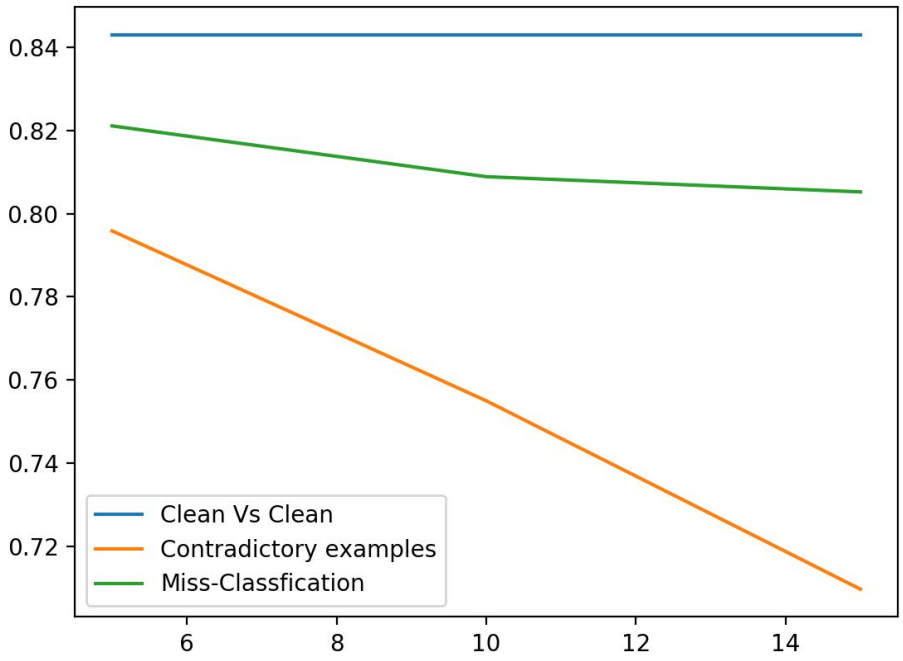
For the Tic Tac Toe endgame:

Following is the screenshot for the result obtained:

Mohita Chaudhary
20830560

```
 For Contradictory examples
For noise level  5 % the mean accuracy is 0.7958311403508771 and the Standard Deviation is 0.012347916584390616

 For Contradictory examples
For noise level  10 % the mean accuracy is 0.7549561403508772 and the Standard Deviation is 0.017061533875161503

 For Contradictory examples
For noise level  15 % the mean accuracy is 0.7096381578947368 and the Standard Deviation is 0.01198702545420535

 For the misclassified labels
For noise level  5 % the mean accuracy is 0.8210910087719296 and the Standard Deviation is 0.004897876089032644

 For the misclassified labels
For noise level  10 % the mean accuracy is 0.8088750000000001 and the Standard Deviation is 0.012041927257737436

 For the misclassified labels
For noise level  15 % the mean accuracy is 0.8052160087719298 and the Standard Deviation is 0.007936469937259265
(base) mohita@Mohitas-MacBook-Pro ~ %
```

Plot for accuracy Vs L% for Contradictory, Misclassified and Clean labels:

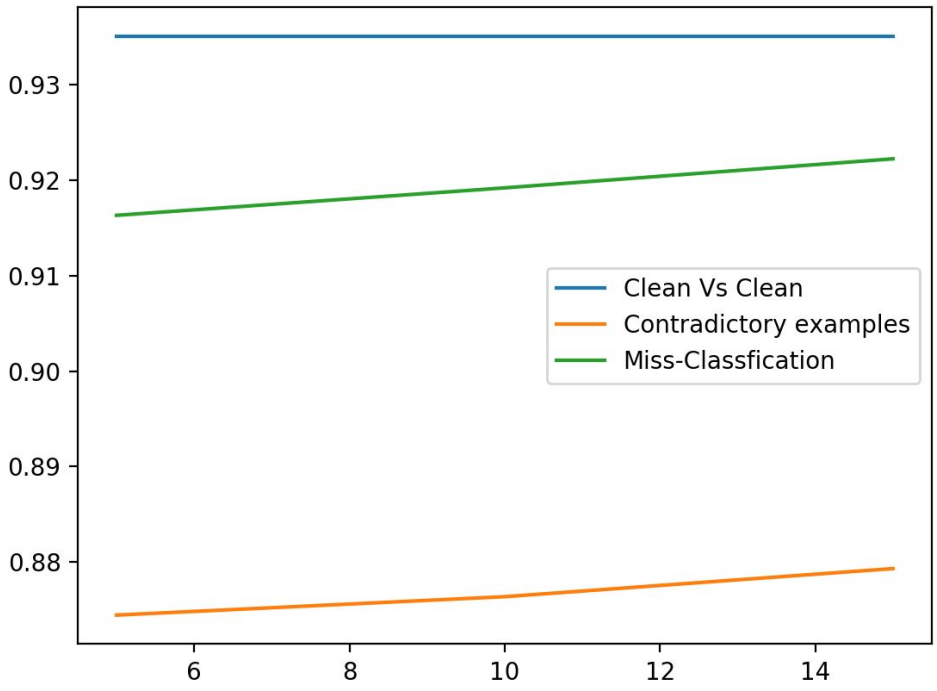


For the Wine Dataset:
Following is the screenshot for the result obtained:

Mohita Chaudhary
20830560

```
For Contradictory examples
For noise level  5 % the mean accuracy is 0.8744444444444444 and the Standard Deviation is 0.014865366144111233
For Contradictory examples
For noise level  10 % the mean accuracy is 0.8763725490196078 and the Standard Deviation is 0.02284033195487295
For Contradictory examples
For noise level  15 % the mean accuracy is 0.879313725490196 and the Standard Deviation is 0.013096423067031307

 For the misclassified Labels
For noise level  5 % the mean accuracy is 0.9163071895424837 and the Standard Deviation is 0.012725349204129455

 For the misclassified Labels
For noise level  10 % the mean accuracy is 0.9191830065359478 and the Standard Deviation is 0.020389145286251326

 For the misclassified Labels
For noise level  15 % the mean accuracy is 0.9222222222222223 and the Standard Deviation is 0.014128838459335715
```

Plot for accuracy Vs L% for Contradictory, Misclassified and Clean labels:



Observation from the above plots:
For both the datasets we observe that Clean vs Clean gives the best accuracies followed by the contradictory examples and then the miss classification examples.

Question: Effect of noise on ID3 when there are no pruning variables ?

Answer: When there are no pruning variables that means that there is no restriction in the depth of the decision tree. When we increase the noise in the dataset, this increases the

randomness in the dataset, which makes our decision tree deeper and leads to overfitting, thus giving poor accuracy as deduced from the plots above. Thus, without any pruning variables the depth of the decision tree increases.

Question: Which is more harmful attribute noise or class noise ?

Answer: The class noise is much more harmful than the attribute noise. In the above graphs which we obtained we can clearly see that the class noise gives a poorer accuracy than the attribute noise. This is clearly because the classes(labels) are used to find the entropy and when we add the contradictory examples or misclassified examples then we are unable to pick the best node(attribute) for the split thus leading to a poorer accuracy.