Group-2-Activity-Module-3

Question 1

Since the mentioned attributes are binary, the following deduction can be made.

1st round decision: there are $C_2^1 = 2$ decisions can be made. In total, there should be 2 possible situations;

2nd round decision: there are $C_2^1 = 2$ decisions can be made. In total, there should be $C_2^1 * C_2^1 = 2 * 2 = 2^2 = 4$ possible situations;

3rd round decision: there are $C_2^1 = 2$ decisions can be made. In total, there should be $C_2^1 * C_2^1 * C_2^1 = 2 * 2 * 2 = 2^3 = 8$ possible situations;

By parity of reasoning, when it reaches the nth round decision, there should be $(C_2^1)^n = 2^n$ possible situations.

Question 2

a. In this case, we have $Y = \{0, 1\}$, thus, $P(Y = 0) = \frac{1}{2}$, and $P(Y = 1) = \frac{1}{2}$.

Therefore, $Entropy(Y) = H(Y) = -\sum P(Y = y) * log_2\big(P(Y = y)\big) = -\frac{1}{2} * log_2\left(\frac{1}{2}\right) - \frac{1}{2} * log_2\left(\frac{1}{2}\right) = -\frac{1}{2} * (-1) - \frac{1}{2} * (-1) = 1$

b. To compare the information gained, the following calculation should be made.

Feature A

According to the dataset, $A = \{0, 1\}$, $P(A = 0) = \frac{1}{2}$, $P(A = 1) = \frac{1}{2}$.

$H(Y|A) = \sum P(A = a) * H(Y|A = a) = P(A = 0) * H(Y|A = 0) + P(A = 1) * H(Y|A = 1)$

When $A = 0$, $(A, Y) \in \{(0, 1), (0, 0)\}$, $P(0,1) = \frac{2}{3}$, $P(0,0) = \frac{1}{3}$;

When $A = 1$, $(A, Y) \in \{(1, 0), (1,1)\}$, $P(1,0) = \frac{1}{3}$, $P(1,1) = \frac{2}{3}$;

Thus, $H(Y|A) = \frac{1}{2} * \left(-\frac{2}{3} log_2\left(\frac{2}{3}\right) - \frac{1}{3} * log_2\left(\frac{1}{3}\right)\right) + \frac{1}{2} * \left(-\frac{1}{3} * log_2\left(\frac{1}{3}\right) - \frac{2}{3} log_2\left(\frac{2}{3}\right)\right) = -\frac{1}{3} * log_2\left(\frac{1}{3}\right) - \frac{2}{3} log_2\left(\frac{2}{3}\right) = 0.918$

Therefore, information gain $I(A; Y) = H(Y) - H(Y|A) = 1 - 0.918 = 0.082$.

Feature B

According to the dataset, $B = \{0,1\}$, $P(B = 0) = \frac{1}{2}$, $P(B = 1) = \frac{1}{2}$.

When $B = 0$, $(B, Y) \in \{(0,0), (0,1)\}$, $P(0,0) = \frac{1}{2}$, $P(0,1) = \frac{1}{2}$;

When $B = 1$, $(B, Y) \in \{(1,1), (1.0)\}$, $P(1,1) = \frac{1}{2}$, $P(1,0) = \frac{1}{2}$;

Thus, $H(Y|B) = \frac{1}{2} * \left(-\frac{1}{2} log_2\left(\frac{1}{2}\right) - \frac{1}{2} log_2\left(\frac{1}{2}\right)\right) + \frac{1}{2} * \left(-\frac{1}{2} log_2\left(\frac{1}{2}\right) - \frac{1}{2} log_2\left(\frac{1}{2}\right)\right) = \frac{1}{2} + \frac{1}{2} = 1$.

Therefore, information gain $I(B;Y) = H(Y) - H(Y|B) = 1 - 1 = 0$.

Feature C

According to the dataset, $C = \{0,1\}$, $P(C = 0) = \frac{1}{2}$, $P(C = 1) = \frac{1}{2}$.

When $C = 0$, $(C,Y) \in \{(0,0),(0,1)\}$, $P(0,0) = \frac{1}{3}$, $P(0,1) = \frac{2}{3}$;

When $C = 1$, $(C,Y) \in \{(1,0),(1,1)\}$, $P(1,0) = \frac{2}{3}$, $P(1,1) = \frac{1}{3}$;

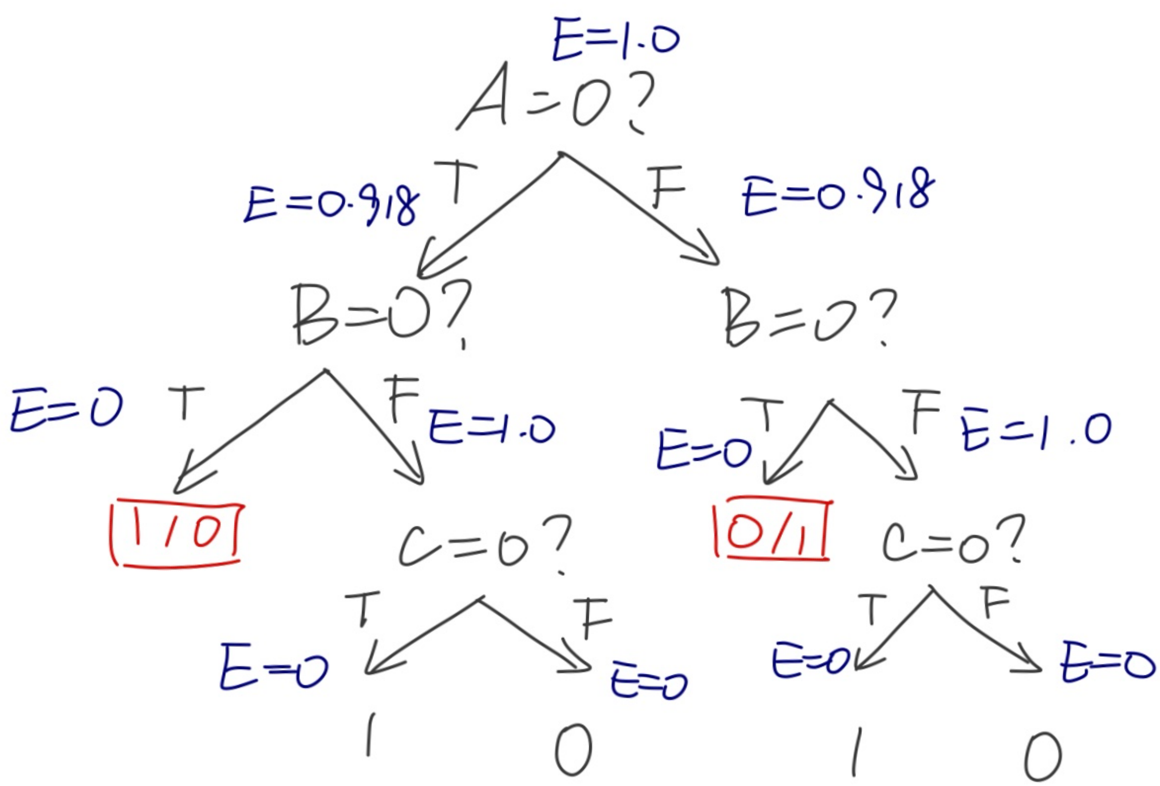
Thus, $H(Y|C) = \frac{1}{2} * \left(-\frac{1}{3} * log_2\left(\frac{1}{3}\right) - \frac{2}{3} log_2\left(\frac{2}{3}\right)\right) + \frac{1}{2} * \left(-\frac{2}{3} log_2\left(\frac{2}{3}\right) - \frac{1}{3} * log_2\left(\frac{1}{3}\right)\right) = -\frac{2}{3} log_2\left(\frac{2}{3}\right) - \frac{1}{3} * log_2\left(\frac{1}{3}\right) = 0.918$.

Therefore, information gain $I(C;Y) = H(Y) - H(Y|A) = 1 - 0.918 = 0.082$.

Conclusion

Since $I(A;Y) = I(C;Y) > I(B;Y)$, either feature A or C could be firstly split at.

c. Unique.



Question 3

Before using the decision tree classifier, we checked the basic info of this dataset to determine the targets and features in this dataset. From the data types shown in the figure below, column "M" is likely to be the target, and the float columns are likely to be the features.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 568 entries, 0 to 567
Data columns (total 32 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   842302    568 non-null    int64
 1   M         568 non-null    object
 2   17.99     568 non-null    float64
 3   10.38     568 non-null    float64
 4   122.8     568 non-null    float64
 5   1001      568 non-null    float64
 6   0.1184    568 non-null    float64
 7   0.2776    568 non-null    float64
 8   0.3001    568 non-null    float64
 9   0.1471    568 non-null    float64
 10  0.2419    568 non-null    float64
 11  0.07871   568 non-null    float64
 12  1.095     568 non-null    float64
 13  0.9053    568 non-null    float64
 14  8.589     568 non-null    float64
 15  153.4     568 non-null    float64
 16  0.006399  568 non-null    float64
 17  0.04904   568 non-null    float64
 18  0.05373   568 non-null    float64
 19  0.01587   568 non-null    float64
 20  0.03003   568 non-null    float64
 21  0.006193  568 non-null    float64
 22  25.38     568 non-null    float64
 23  17.33     568 non-null    float64
 24  184.6     568 non-null    float64
 25  2019      568 non-null    float64
 26  0.1622    568 non-null    float64
 27  0.6656    568 non-null    float64
 28  0.7119    568 non-null    float64
 29  0.2654    568 non-null    float64
 30  0.4601    568 non-null    float64
 31  0.1189    568 non-null    float64
dtypes: float64(30), int64(1), object(1)
memory usage: 142.1+ KB
```

To verify our deductions, we also checked the dataset description, *wdbc.names*, from the UCI website, and its content supported our conjecture.

```
Results:

	- predicting field 2, diagnosis: B = benign, M = malignant
	- sets are linearly separable using all 30 input features
	- best predictive accuracy obtained using one separating plane
		in the 3-D space of Worst Area, Worst Smoothness and
		Mean Texture.  Estimated accuracy 97.5% using repeated
		10-fold crossvalidations.  Classifier has correctly
		diagnosed 176 consecutive new patients as of November
		1995.
```

According to the description given in the first figure, it can be observed that the dataset used in this case is a rather small dataset for the machine learning study topic. We repeated the required processes for 10 times and found that the metric values gained from this lab exercise are dynamic and highly related to the split training and testing set. As a result, we randomly selected one result set from the above trials.

For the first decision tree DT1, we used the default settings of DecisionTreeClassifier in Sklearn library. We got the following testing results, which are the average metric values of 20 times trails.

	Accuracy		**Precision**	**Recall**
	Train	**Test**		
DT1	1.00	0.93	0.95	0.94

For the second decision tree, we designed a repetitive parameter-tuning method to find the best match of the decision tree classifier. We increased the maximum depth of the decision tree from 1 to 20, and each decision tree was tested by the process used in DT1's testing. Then we got the following results.

max_depth	**train_accuracy**	**test_accuracy**	**precision**	**recall**
1	0.93	0.9	0.9	0.94
2	0.93	0.9	0.94	0.9
3	0.97	0.93	0.93	0.96
4	0.98	0.94	0.95	0.95
5	0.99	0.93	0.94	0.94
6	1	0.93	0.96	0.93
7	1	0.93	0.95	0.94
8	1	0.92	0.94	0.94
9	1	0.93	0.94	0.95
10	1	0.93	0.94	0.95
11	1	0.93	0.95	0.94

max_depth	train_accuracy	test_accuracy	precision	recall
12	1	0.94	0.95	0.95
13	1	0.93	0.95	0.94
14	1	0.92	0.94	0.94
15	1	0.93	0.95	0.94
16	1	0.94	0.96	0.94
17	1	0.93	0.95	0.94
18	1	0.93	0.94	0.95
19	1	0.93	0.95	0.94
20	1	0.92	0.94	0.94

According to the result set above, when the max depth of the decision tree $d \in [0,6]$, the accuracy of the training phrase is increasing. When $d = 6$, it met the upper boundary and has no improvement during the increase of the max depth. We found the first parameter match that made DT2 perform better than DT1 is $d = 12$. Therefore, we can supplement the result set for decision tree performance evaluation as the final results.

	Accuracy		**Precision**	**Recall**
	Train	**Test**		
DT1	1.00	0.93	0.95	0.94
DT2	1.00	0.94	0.95	0.95