6.3) Problem Solving on Paper

A. Research the concept of "Information Gain Ratio" (which was discussed in lectures). In 100 words, discuss what is wrong with regular "Information Gain" and how does "Information Gain Ratio" partially remedy this. Explain with one example.

(total words: 100).

Answer: The issue with information gain is bias towards attributes with all distinct values, as information gain in that scenario is equivalent to the original total entropy of the data provided, thus always selecting that attribute as the first attribute to split tree, this could lead to model learning training data too well, leading to overfitting.

As in the example the distinct valued attribute is the Name, this has highest initial Information gain, which when divided by intrinsic information provided by the feature gives correct result.

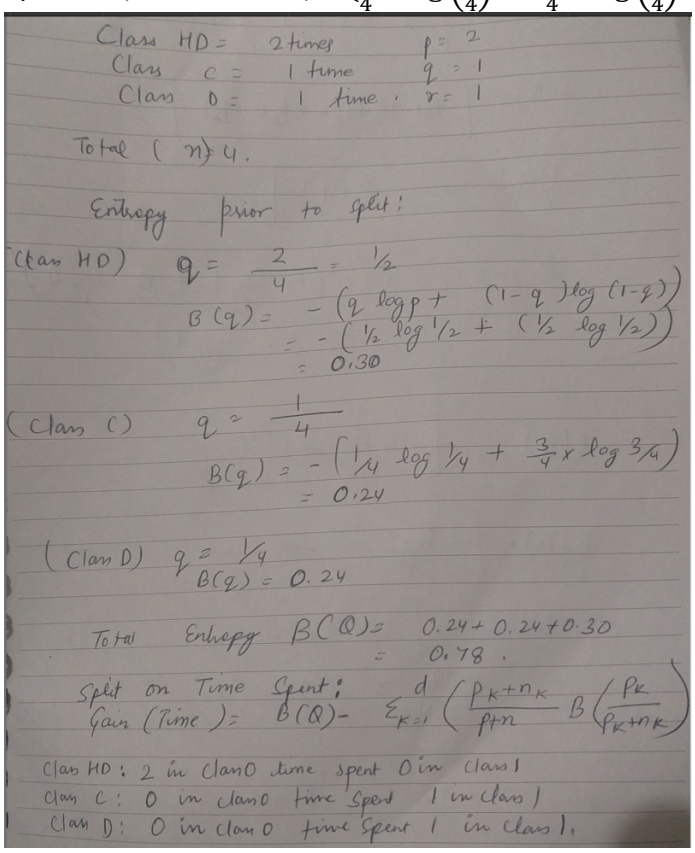
EXAMPLE:

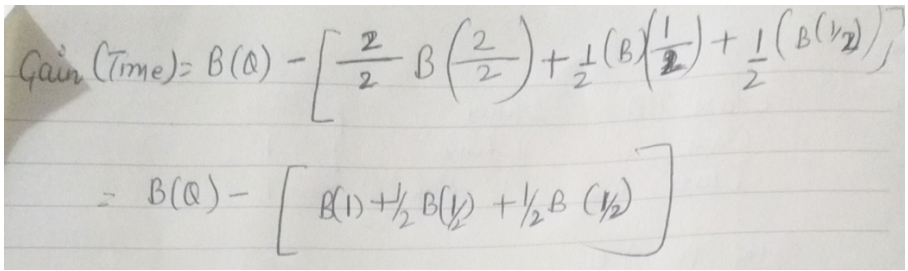
Name	Time Spent on Unit(in hrs)	Grade in Unit
Navdeep	120 (0 Class)	HD
Michael	110 (0 class)	HD
Abhishek	80 (1 class)	C
Gurpreet	90 (1 class)	D

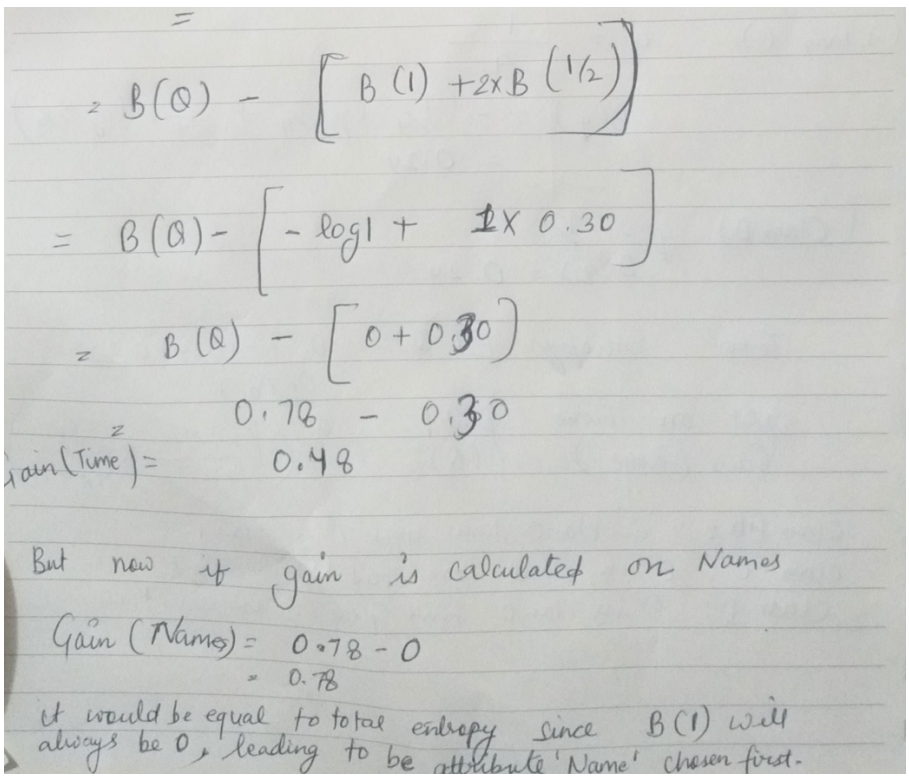
Split Info(Name)=($-(\frac{1}{4} * \log\left(\frac{1}{4}\right) + \frac{1}{4} * \log\left(\frac{1}{4}\right) + \frac{1}{4} * \log\left(\frac{1}{4}\right) + \frac{1}{4} * \log\left(\frac{1}{4}\right)$))= 0.60

Split Info(Time Spent On Unit)= $-(\frac{2}{4} * \log\left(\frac{2}{4}\right) + \frac{2}{4} * \log\left(\frac{2}{4}\right)$ =0.15

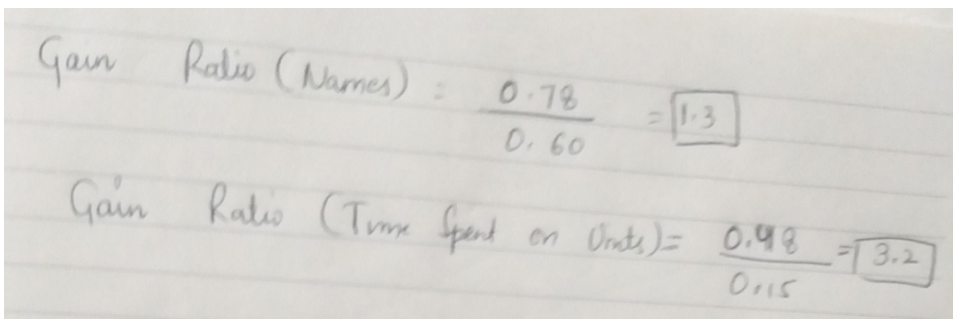
Split Info(Grade in Unit) =$-(\frac{2}{4} * \log\left(\frac{2}{4}\right) + \frac{1}{4} * \log\left(\frac{1}{4}\right) + \frac{1}{4} * \log\left(\frac{1}{4}\right)$)= 0.45







Since this is when the information gain fails, the ratio can be calculated by diving the gain found with the Split info of relevant attribute, this gives true and corrected values which can be used in building decision tree.



B. Research the concept of cross validation. In 100 words, discuss what is different between StratifiedKFold and ShuffleSplit. Explain with one example.
(100 words)
Answer:

Stratified K Fold	Shuffle Split
1)Each test should not overlap with previous split. 2) The data is first shuffled at the start and then split is done depending on the size, each time, testing and training dataset being distinct.	1)Tests could overlap with previous split. 2) The data is shuffled every time the model is tested and split is done depending on the size, each time, testing and training dataset, having the chance of overlap with last split.

Here is an example K fold shuffled once, shuffle split, every time data is shuffled.

```
KFold
TRAIN: [0 2 3 4 5 6 7 9] TEST: [1 8]
TRAIN: [0 1 2 3 5 7 8 9] TEST: [4 6]
TRAIN: [0 1 3 4 5 6 8 9] TEST: [2 7]
TRAIN: [1 2 3 4 6 7 8 9] TEST: [0 5]
TRAIN: [0 1 2 4 5 6 7 8] TEST: [3 9]
```

```
Shuffle Split
TRAIN: [8 4 1 0 6 5 7 2] TEST: [3 9]
TRAIN: [7 0 3 9 4 5 1 6] TEST: [8 2]
TRAIN: [1 2 5 6 4 8 9 0] TEST: [3 7]
TRAIN: [4 6 7 8 3 5 1 2] TEST: [9 0]
TRAIN: [7 2 6 5 4 3 0 9] TEST: [1 8]
```