CSCN8000 – Artificial Intelligence Algorithms and Mathematics

Assignment 2: Decision Trees

ID	Good Behaviour	Age<30	Drug Dependent	RECIDIVIST
1	False	True	False	True
2	False	False	False	False
3	False	True	False	True
4	True	False	False	False
5	True	False	True	True
6	True	False	False	False

A convicted criminal who reoffends after release is known as a recidivist. The table below lists a dataset that describes prisoners released on parole, and whether they reoffended within two years of release.

This dataset lists six instances where prisoners were granted parole. Each of these instances are described in terms of three binary descriptive features (GOOD BEHAVIOR, AGE < 30, DRUG DEPENDENT) and a binary target feature, RECIDIVIST. The GOOD BEHAVIOR feature has a value of true if the prisoner had not committed any infringements during incarceration, the AGE < 30 has a value of true if the prisoner was under 30 years of age when granted parole, and the DRUG DEPENDENT feature is true if the prisoner had a drug addiction at the time of parole. The target feature, RECIDIVIST, has a true value if the prisoner was arrested within two years of being released; otherwise it has a value of false.

a. ***Using this dataset, construct the decision tree that would be generated by the ID3 algorithm (as done in class), using entropy-based information gain. (15 Points)***

  b. ***You need to provide all the full steps and calculations of information gain done to reach the final decision tree. If you only provide the final decision tree you will receive 2 points for this question.***

Let's calculate the entropy $E(S)$ for the entire dataset

Therefore, $E(S) = -p_{true} \log(p_{true}) - p_{false} \log(p_{false})$

In **Recidivist** the number of $True$ & $False$ with there probabilities are $1/2$ respectively.

Therefore,

$$E(S) = -p_{true} \log(p_{true}) - p_{false} \log(p_{false})$$

$E(S) = [-\left(\frac{1}{2}\right)\log\left(\frac{1}{2}\right)] - [\left(\frac{1}{2}\right)\log\left(\frac{1}{2}\right)]$

$E(S) = -[0.5 * -1] - [0.5 * -1]$

$E(S) = 0.5 + 0.5$

$E(S) = 1$

Now, we will look at the entropy for each feature to look for maximum entropy for Information gain among the features

1. **Information Gain** for *Good Behaviour:*

   Good Behavior = True: (IDs 4, 5, 6) → 2 False, 1 True
   Good Behavior = False: (IDs 1, 2, 3) → 1 False, 2 True
   $IG(Good) = E(S) - E(S|Good)$
   $IG(Good) = 1 - [\left(\frac{3}{6}\right)E(S_{true}) + \left(\frac{3}{6}\right)E\left(S_{false}\right)]$
   $IG(Good) = 1 - \{0.5\left[-\left(\frac{1}{3}\right)\log\left(\frac{1}{3}\right) - \left(\frac{2}{3}\right)\log\left(\frac{2}{3}\right)\right] + 0.5[-\left(\frac{2}{3}\right)\log\left(\frac{2}{3}\right) - \left(\frac{1}{3}\right)\log\left(\frac{1}{3}\right)]\}$
   $IG(Good) = 1 - \{0.5[0.918] + 0.5[0.918]\}$
   $IG(Good) = 1 - 0.918$
   $IG(Good) = 0.082$

2. **Information Gain** for *AGE < 30:*

   Age < 30 = True: (IDs 1, 3) → 0 False, 2 True
   Therefore entropy = 0
   Age < 30 = False: (IDs 2, 4, 5, 6) → 2 False, 1 True
   Therefore,
   $E(S_{False}) = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right) = 0.918$

   Weighted entropy for *Age:*
   $E(S|Age) = \frac{2}{6}.0 + \frac{4}{6}.0.918 = 0.612$
   Therefore, Information Gain (Gain (Age < 30)):
   $IG(Age) = E(S) - E(S|Age)$
   $IG(Age) = 1 - 0.612$
   $IG(Age) = 0.388$

3. **Information Gain** for *Drug Dependent:*

   Drug Dependent = True: (IDs 5) → 0 False, 1 True
   Similarly, entropy = 0
   Drug Dependent = False: (IDs 1, 2, 3, 4, 6) → 2 False, 2 True
   $E(S_{False}) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4} = 1$
   Weighted entropy for Drug:
   $E(S|Drug) = \frac{1}{6}.0 + \frac{5}{6}.1 = 0.833$

Information Gain (Gain (Drug Dependent)):

$IG(Drug) = E(S) - E(S|Drug) = 1 - 0.833 = 0.167$

Since, the *Age < 30* has highest information gain then the 1st split would be between *Age < 30* for which, while the prediction is *True* its homogenous but when its *False* ,we need to choose between *Good Behaviour & Drug Dependent.*

ID	Good Behaviour	Age<30	Drug Dependent	RECIDIVIST
2	False	False	False	False
4	True	False	False	False
5	True	False	True	True
6	True	False	False	False

Above is the dataset now only consisting *Age < 30 = False*, we will choose between *Good Behaviour & Drug Dependent*

In **Recidivist** the number of $True$ & $False$ probabilities are $\frac{1}{4}$ & $\frac{3}{4}$ respectively

Therefore,

$$E(S) = -p_{true}\log(p_{true}) - p_{false}\log(p_{false})$$

$$E(S) = -(\frac{1}{4})\log(\frac{1}{4}) - (\frac{3}{4})\log(\frac{3}{4})$$

$$E(S) = -0.25(-0.6021) - 0.75(-0.1250)$$

$$E(S) = 0.1501 + 0.09375$$

$$E(S) = 0.24385$$

Now, we will look at the entropy for each feature to look for maximum entropy for Information gain among the features

1. **Information Gain** for *Good Behaviour:*

   Good Behavior = True: (IDs 4, 5, 6) → 2 False, 1 True
   Good Behavior = False: (IDs 2) → 1 False, 0 True

   $$IG(Good) = E(S) - E(S|Good)$$
   $$IG(Good) = 0.24385 - [\left(\frac{3}{4}\right)E(S_{true}) + \left(\frac{1}{4}\right)E(S_{false})]$$
   $$IG(Good) = 0.24385 - \{0.75\left[-\left(\frac{1}{3}\right)\log\left(\frac{1}{3}\right) - \left(\frac{2}{3}\right)\log\left(\frac{2}{3}\right)\right] + 0.25[-\left(\frac{0}{1}\right)\log\left(\frac{0}{1}\right) - (1)\log(1)]\}$$
   $$IG(Good) = 0.24385 - \{0.74[0.918] + 0.25[0]\}$$
   $$IG(Good) = 0.24385 - 0.6781$$
   $$IG(Good) = -0.43425$$

2. **Information Gain** for *Drug Dependent:*

Drug Dependent = True: (IDs 5) → 0 False, 1 True

Similarly, entropy = 0

Drug Dependent = False: (IDs 2, 4, 6) → 3 False, 0 True

Again, entropy = 0

Therefore,

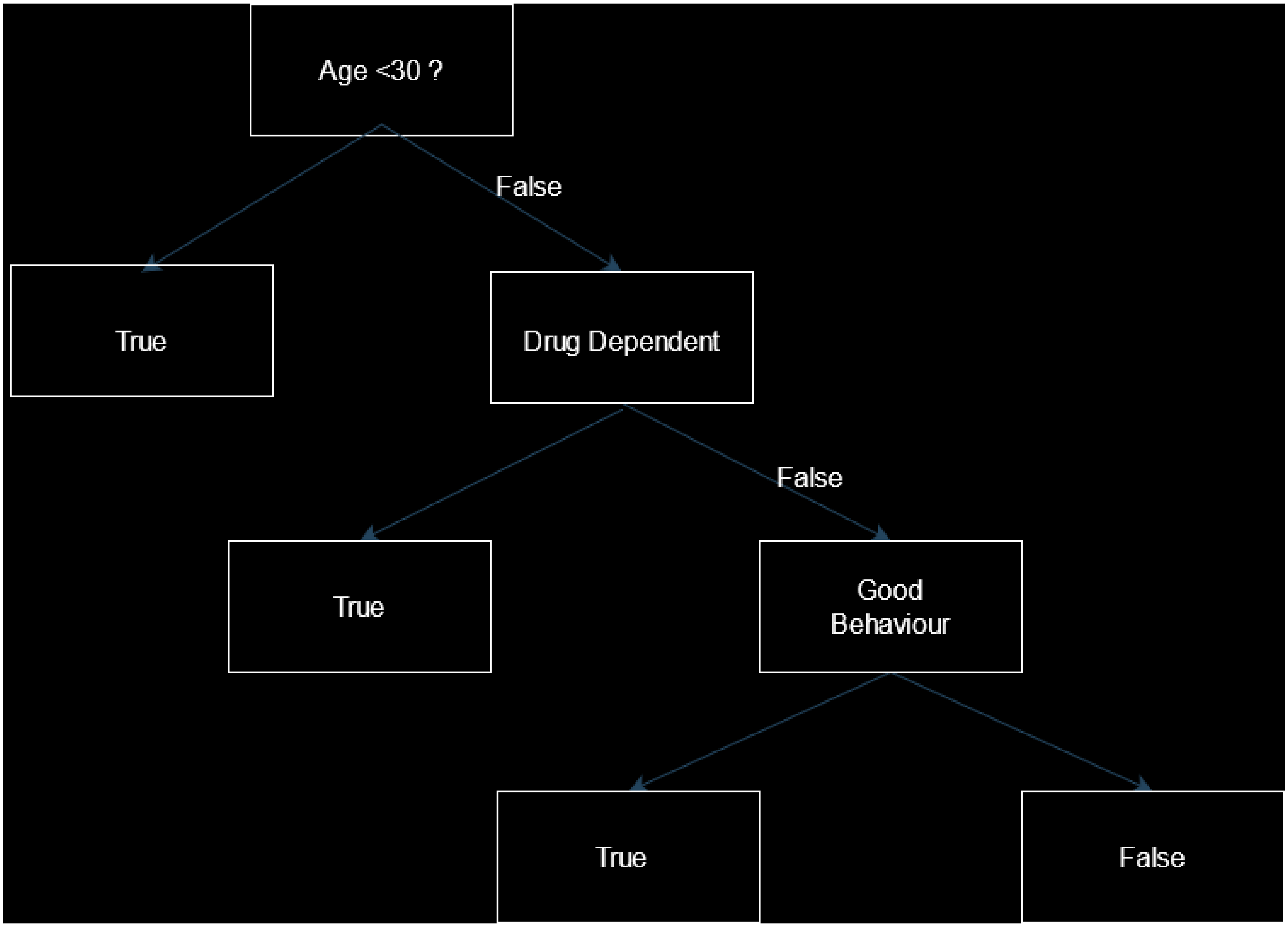
$$IG(Drug) = E(S) - E(S|Drug)$$

$$IG(Drug) = 0.24385 - [\frac{1}{4}E(S_{true}) + \frac{3}{4}E(S_{false})]$$

$$IG(Drug) = 0.24385 - \{\frac{1}{4}[-1log1 - (0)] + \frac{3}{4}[(0) - 1log1]\}$$

$$IG(Drug) = 0.24385$$

Since, *Drug Dependent* has highest information gain then the 2nd split would be from *Drug Dependent*



c. What prediction will the decision tree generated in part (a) of this question return for the following query? (5 Points)

GOOD BEHAVIOR = false, AGE < 30 = false, DRUG DEPENDENT = true

For the given query while traversing the decision tree:

1. Check *Age <30 = False, now* according to our decision tree if we check for false, we again check for *Drug Dependent*
2. Check *Drug Dependent* for *True* which in our case is the final prediction
   So, the final query returns True.

d. What prediction will the decision tree generated in part (a) of this question return for the following query? (5 Points)

GOOD BEHAVIOR = true, AGE < 30 = true, DRUG DEPENDENT = false

1. Check *Age <30 = True, now* according to our decision tree if we *True* then that is our final prediction because there is no further branch
   Therefore, the query returns True.