CSCN8000 – Artificial Intelligence Algorithms and Mathematics

Assignment 2: Decision Trees

Student ID: 8883828

Name: Parthasarathy Rajendiran

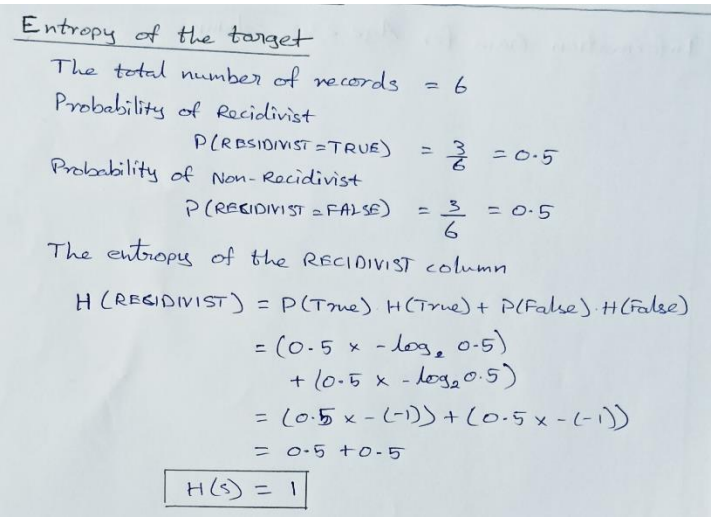
Dataset

The given dataset contains 6 observations, three predictors (Good Behaviour, Age<30, and Drug Dependent), and a target (RECIDIVIST)

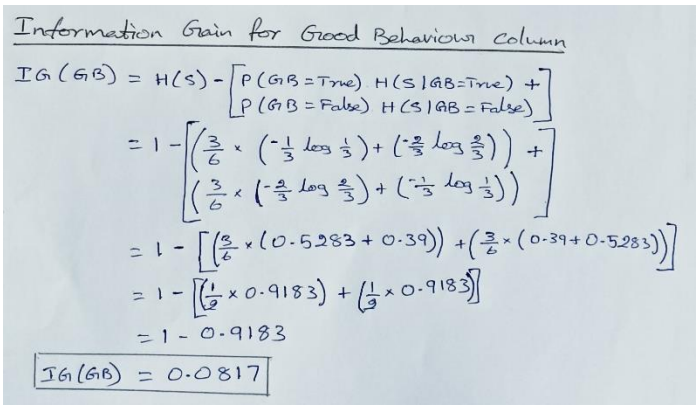
ID	Good Behaviour	Age<30	Drug Dependent	RECIDIVIST
1	FALSE	TRUE	FALSE	TRUE
2	FALSE	FALSE	FALSE	FALSE
3	FALSE	TRUE	FALSE	TRUE
4	TRUE	FALSE	FALSE	FALSE
5	TRUE	FALSE	TRUE	TRUE
6	TRUE	FALSE	FALSE	FALSE

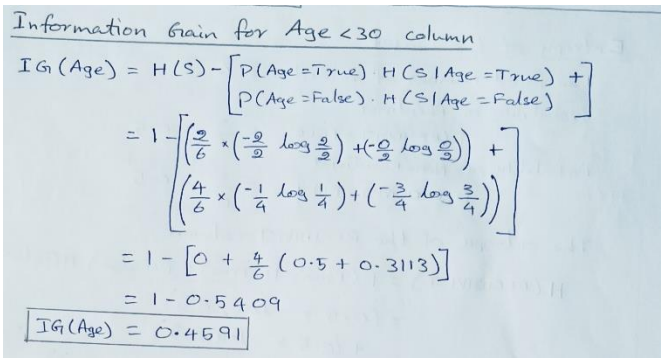
Part a.

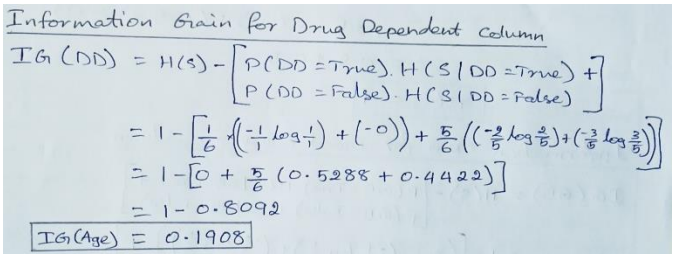
Firstly, the entropy of the target variable should be calculated.



With the use of this entropy, the Information Gain for each column should be calculated.







Root Node Selection

The information gain from the “Age<30” column is higher than that of other columns. So, Age<30 can be used as the root node.

Now, let us split the dataset using the “Age<30” column.

Age<30 = TRUE

Count 2

ID	Good Behaviour	Age<30	Drug Dependent	RECIDIVIST
1	FALSE	TRUE	FALSE	TRUE
3	FALSE	TRUE	FALSE	TRUE

Good B...	Age<30	Drug D...
FALSE	FALSE	FALSE
TRUE	TRUE	TRUE

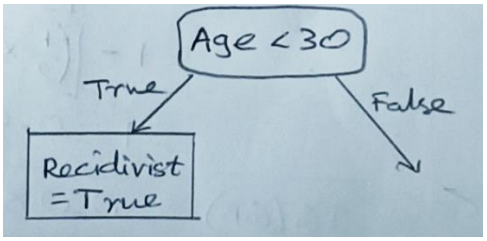
Age<30 = FALSE

Count 4

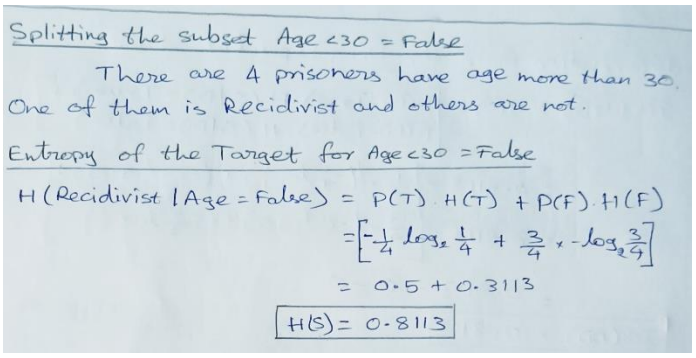
ID	Good Behaviour	Age<30	Drug Dependent	RECIDIVIST
2	FALSE	FALSE	FALSE	FALSE
4	TRUE	FALSE	FALSE	FALSE
5	TRUE	FALSE	TRUE	TRUE
6	TRUE	FALSE	FALSE	FALSE

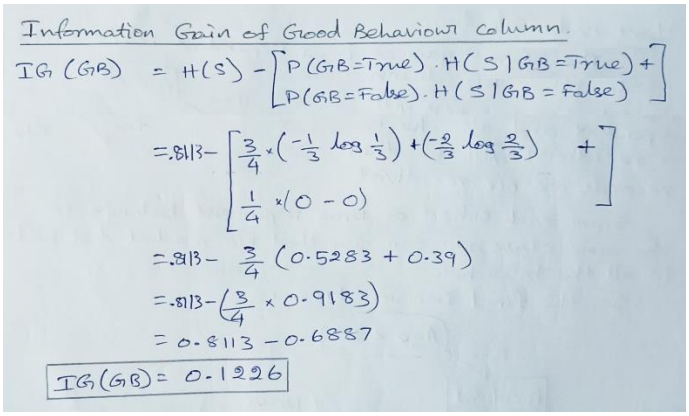
Good B...	Age<30	Drug D...
FALSE	FALSE	FALSE
TRUE	TRUE	TRUE

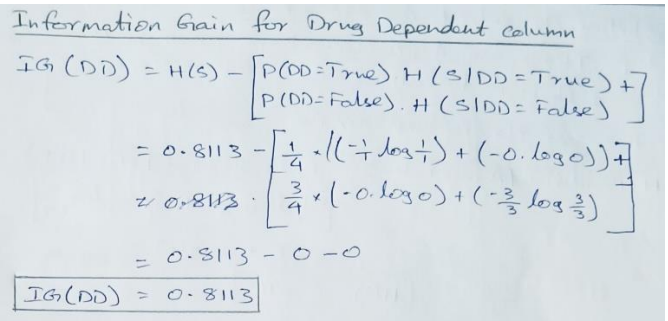
For the Age < 30 = TRUE subset, all the records have Recidivist = TRUE. Since this subset contains only one class, this branch reached the leaf node.



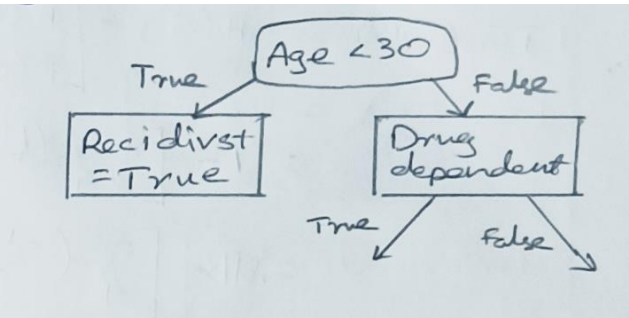
The above steps must be repeated for the subset Age<30 = FALSE







Since the information gain of the “Drug Dependent” column is higher, that can be considered for the next split.



Out of 4 prisoners with Age < 30, only one is Drug dependent and also he is a recidivist. All the other records are non-recidivist.

Count 1

ID	Good Behaviour	Age<30	Drug Dependent	RECIDIVIST
5	TRUE	FALSE	TRUE	TRUE

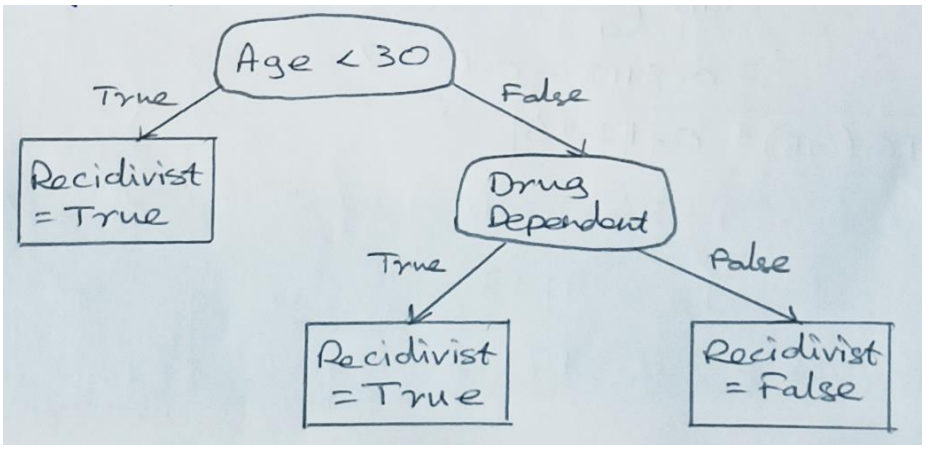
Good B...	Age<30	Drug D...
TRUE	FALSE	FALSE
FALSE	TRUE	TRUE

Count 3

ID	Good Behaviour	Age<30	Drug Dependent	RECIDIVIST
2	FALSE	FALSE	FALSE	FALSE
4	TRUE	FALSE	FALSE	FALSE
6	TRUE	FALSE	FALSE	FALSE

Good B...	Age<30	Drug D...
FALSE	FALSE	FALSE
TRUE	TRUE	TRUE

Since each subset of Drug Dependents belongs to the same class, we can say that we reached leaf nodes in all the branches. So, the final decision tree is.

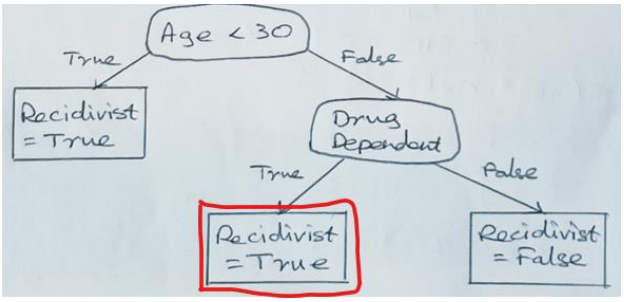


This decision tree is optimal and gives below insights,

- All the prisoners released on parole and youngsters (Age < 30) are RECIDIVIST.
- All the Drug Dependents are Recidivist = TRUE.
- The prisoners who are not Drug Dependents and older than 30 years are not RECIDIVIST.

Part b.

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

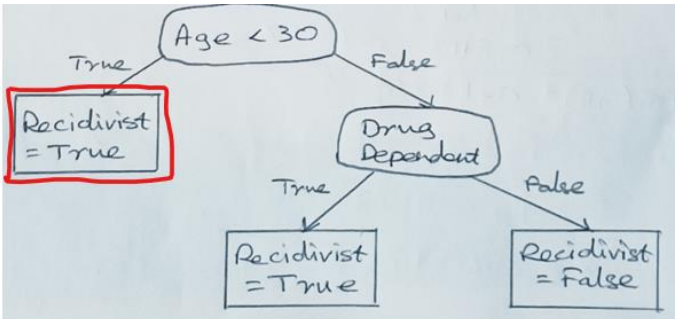


GOOD BEHAVIOR = false, AGE < 30 = false, DRUG DEPENDENT = true

Based on our decision tree, a prisoner with an age of more than 30 and a drug-dependent will be a Recidivist.

Part c.

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



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

Based on our decision tree, a prisoner with an age of less than 30 and not drug-dependent will be a Recidivist.