Homework 5   
Answer questions 5 through 10 in the Exercises section from Chapter 8 of the Larose   
textbook.   
In your submission, you must:   
• Clearly answer each question in a Word document and submit the Word   
document to the Module 5 Homework assignment.   
• Write your answers completely and elaborate on your ideas as much as   
possible.

Table

Description automatically generated

1. **Construct a classification and regression tree to classify *salary* based on the other variables. Do as much as you can by hand, before turning to the software.**

**Ans.**  Classification and regression tree to classify salary based on the other variable.

As mentioned Salary can be categorized in -   
1. x< $35,000

2. $35,000 <=x < $45,000

3. $45,000 <=x<$55,000

4. x>$55,000

Please refer to the excel sheet used to calculate.

Diagram

Description automatically generated

1. **Construct a C4.5 decision tree to classify *salary* based on the other variables. Do as much as you can by hand, before turning to the software.**

**Ans.** C4.5 decision tree to classify *salary*  based on the other variable.

Graphical user interface, application, table, Excel

Description automatically generated

1. **Compare the two decision trees and discuss the benefits and drawbacks of each.**

**Ans.**  The two decision trees are CART and C4.5 decision trees which are developed based on different meausres for leaf node purity.

|  |  |
| --- | --- |
| **CART decision tree** | **C4.5 decision tree** |
| * CART recursively partitions data into two categories wit subset of similar values | * C4.5 decision tree categorise based branches created for each value set |
| * CART identifies possible splits based on the values it contains | * Depends on the values of categorical parameter |
| * For every root node/ decision node find the training records associated with decision node | * For every decision node there are values associated branches |
| * CART is restricted to binary splits | * C4.5 is not restricted to binary splits |
| * CART generates a binary tree, having greater level and deeper | * C4.5 generated a bushier tree or of variable shape, provides greater breadths |
| * For categorical attributes, Φ(s|t) be a measure of the “goodness” of a candidate split s at node t,   Φ(*s*|*t*) = 2*PLPR* ∑ *j*=1 |*P*(*j*|*tL*) − *P*(*j*|*tR*)| | * C4.5 creates a separate branch for each value of the categorical attribute based on values having different frequencies or association with different values |
| * It generates the error rate that is weighted avg rate for overall decision tree | * C4.5 method for measuring homogeneity is based on information gain and entropy reduction to select optimal split |
|  |  |

1. **Generate the full set of decision rules for the CART decision tree.**

**Ans.** Decision rules for CART decision tree—

* Trees produced are strictly binary having exactly two branched for each decision node
* It recursively partition the records in training dataset to subset of records with similar value for target attribute
* CART Algo grows tree by conducting for each decision node an exhaustive search for all available variables and all possible splitting values, selecting optimal split according to measure of “goodness” of the criteria based on formula –
* Φ(*s*|*t*) = 2*PLPR* ∑ *j*=1 |*P*(*j*|*tL*) − *P*(*j*|*tR*)|

Where *tL* = left child node of node *t* , *tR* = right child node of node *t* ,

*PL*= number of records at *tL* number of records in training set

*PR*= number of records at *tR*   
 number of records in training set

*P*(*j*|*tL*) = number of class *j* records at *tL* number of records at *t*

*P*(*j*|*tR*) = number of class *j* records at *tR* number of records at *t*

1. **Generate the full set of decision rules for the C4.5 decision tree.**

**Ans.**

* Target variable must be categorical not continuous
* Pre classified target variable must be included in training set
* **C4.5 decision tree is not restricted to binary splits, it could create a tree of more variable shape.**
* For categorical attribute, it produces separate branch of value of the categorical attribute.
* Some values may have low frequency or may naturally associated with other values
* Method of measuring homogeneity is by the concept of information gain or entropy reduction to select optimal split.

It is derived from the entropy of X which is defined as -   
H(X) = − ∑*pj* log2(*pj*)

1. **Compare the two sets of decision rules and discuss the benefits and drawbacks of each.**

**Ans.**

|  |  |
| --- | --- |
| **CART** | **C4.5** |
| **Benefits –**  creates a strictly Binary Decision Tree CART algo prunes the node and branch that would reduce the generalizability of classification esults.  Classification error rate is calculated for entire decision tree to be weighted avg of the individual leaf error rate | **Benefits –** Creates a bushier tree not restricted to binary  Measure of homogeneity is based on information gain or entropy reduction to select the optimal split  Entropy/Node homogeneity is calculated by information gain from the classification and optimal split. |
| **Drawbacks -** Restricted to either or of binary resultants Variables having More than one results are tricky to be classified by CART | **Drawbacks -** Trees get bushier as we keep on classifying further for a splitting the branch for each categorical attribute |
|  |  |