

# DM Assignment 06

23.2.2016

Adithya Abraham Philip

1PI13CS008

## Overview

The assignment involved experimenting with the C4.5 and Gini decision trees available in Weka, on a provided bank dataset, to predict whether a person was likely to purchase a PEP.

## Questions

1. Difference between the decision trees.

**C4.5 Decision Tree** may result in an **n-ary** tree, where  $n$  is the highest cardinality of an attribute, among all the attributes used for decision making. Also, an attribute is branched on only once.

A **Gini Tree** on the other hand, ensures that it is a **binary** tree. Also, an attribute may be branched on multiple times in the tree.

2. Why is there a difference?

**Gini Tree** ensures a **binary** tree by splitting the values of an attribute into two sets, where each branch represents one of the sets. Consequently, since a branch may contain multiple nominal values for a given attribute, it may be split on again further down the tree.

On the other hand, a **C4.5 Decision Tree** is **n-ary** as it uses all possible values of an attribute as possible branches when evaluating it. As a result, it does not make sense to split on an attribute more than once, as all the tuples that travel down that branch have only one value for that attribute. (The exception to this rule is numerical attributes that have a range of values, so there is no factor *preventing* the repeated use of numerical attributes. However **C4.5** is implemented to not split more than once on any attribute, numerical or otherwise.).

3. Take one example you create on your own and explain how each decision tree will be used to predict the class for your example.

The decision trees are included in the appendix. This section shows merely the steps of traversal.

E.g.

Age	Sex	Region	Income	Married	Children	Car	save_act	current_act	mortgage
21	MAL E	INNER_CITY	11000	YES	2	NO	NO	NO	NO

#### C4.5 Classifier: (attributes shown in order of traversal)

children = 0 && married = NO && mortgage = NO => YES (classified for PEP)

#### Gini Classifier: (attributes shown in order of traversal)

children = (3)|(0)|(2) && married != YES && mortgage != YES

&& children != (3)|(2)|(1) && income < 16479.6 => YES (classified for PEP)

## Final Observations

1. Accuracy seemed to lean towards the **C4.5** decision tree over several different test sets, but the each test set was too small for a conclusive observation.
2. The **Gini Tree** implementation in **Weka**, when re-computing an attribute further down the tree, seems to include values that are **impossible** for that attribute at that stage. E.g. (from traversal in previous section) event after children = (3) | (0) | (2) is followed, a node in the subtree branches for children != (3) | (2) | (1). The (1) here is impossible to arrive at in this subtree and can be **excluded** from the tree altogether. If this were done, **efficiency** would **improve** as the Gini index would have to be calculated only for the split between combinations of (3), (2), (0) instead of (3), (2), (1) and (0). This would be more pronounced for attributes with higher cardinalities.

## Appendix

### C4.5 Decision Tree:

children = 0

| married = NO

| | mortgage = NO: YES (45.0/3.0)

| | mortgage = YES

| | | save\_act = NO: YES (12.0)  
| | | save\_act = YES: NO (22.0)  
| married = YES  
| | save\_act = NO  
| | | mortgage = NO  
| | | | income <= 21506.2  
| | | | | age <= 41: NO (11.0/1.0)  
| | | | | age > 41: YES (5.0/1.0)  
| | | | income > 21506.2: NO (20.0)  
| | | mortgage = YES: YES (25.0/3.0)  
| | save\_act = YES: NO (118.0/12.0)

children = 1

| income <= 15538.8  
| | age <= 41: NO (22.0/2.0)  
| | age > 41: YES (2.0)  
| income > 15538.8: YES (109.0/5.0)

children = 2

| income <= 30189.4: NO (81.0/8.0)  
| income > 30189.4: YES (50.0/5.0)

children = 3

| income <= 44288.3: NO (60.0/5.0)  
| income > 44288.3: YES (8.0)


**Gini Decision Tree:**

children=(3)|(0)|(2)

| married=(YES)  
| | income < 45266.15  
| | | save\_act=(YES)  
| | | | children=(0)|(3)|(1): NO(114.0/13.0)  
| | | | children!=(0)|(3)|(1)  
| | | | | income < 29646.25

| | | | | | age < 29.0  
| | | | | | | age < 26.5  
| | | | | | | | income < 11408.35  
| | | | | | | | | sex=(MALE): NO(2.0/0.0)  
| | | | | | | | | sex!=(MALE): YES(2.0/1.0)  
| | | | | | | | | income >= 11408.35: NO(7.0/0.0)  
| | | | | | | age >= 26.5  
| | | | | | | | current\_act=(YES): NO(2.0/0.0)  
| | | | | | | | | current\_act!=(YES): YES(3.0/0.0)  
| | | | | | | age >= 29.0: NO(20.0/0.0)  
| | | | | | income >= 29646.25  
| | | | | | | region=(INNER\_CITY): NO(3.0/1.0)  
| | | | | | | region!=(INNER\_CITY): YES(6.0/0.0)  
| | | | save\_act!=(YES)  
| | | | mortgage=(NO)  
| | | | | children=(3)|(0)|(1)  
| | | | | | income < 21559.2  
| | | | | | | age < 41.5  
| | | | | | | | income < 15933.75: NO(10.0/0.0)  
| | | | | | | | | income >= 15933.75: NO(4.0/1.0)  
| | | | | | | | age >= 41.5  
| | | | | | | | | age < 49.0: YES(3.0/0.0)  
| | | | | | | | | age >= 49.0: YES(1.0/1.0)  
| | | | | | | | income >= 21559.2: NO(23.0/0.0)  
| | | | | | | children!=(3)|(0)|(1)  
| | | | | | | income < 30212.3  
| | | | | | | | income < 13700.75  
| | | | | | | | | age < 28.5: NO(2.0/0.0)  
| | | | | | | | | age >= 28.5: YES(2.0/0.0)  
| | | | | | | | | income >= 13700.75: NO(6.0/0.0)  
| | | | | | | | | income >= 30212.3: YES(6.0/0.0)

| | | | mortgage!=(NO)  
| | | | | children=(2)|(3): NO(10.0/0.0)  
| | | | | children!=(2)|(3)  
| | | | | | region=(TOWN): NO(2.0/1.0)  
| | | | | | region!=(TOWN): YES(21.0/1.0)  
| | income >= 45266.15  
| | | children=(0)|(1): NO(17.0/2.0)  
| | | children!=(0)|(1): YES(17.0/0.0)  
| married!=(YES)  
| | mortgage=(YES)  
| | | save\_act=(YES)  
| | | | children=(0)|(3)|(1): NO(24.0/0.0)  
| | | | children!=(0)|(3)|(1)  
| | | | | income < 30499.7: NO(7.0/0.0)  
| | | | | income >= 30499.7  
| | | | | | age < 60.5: YES(3.0/0.0)  
| | | | | | age >= 60.5: YES(1.0/1.0)  
| | | save\_act!=(YES)  
| | | | children=(2)|(3): NO(6.0/0.0)  
| | | | children!=(2)|(3): YES(12.0/0.0)  
| | mortgage!=(YES)  
| | | children=(3)|(2)|(1)  
| | | | income < 30340.85  
| | | | | region=(INNER\_CITY)|(RURAL)|(TOWN)  
| | | | | | age < 49.5: NO(24.0/0.0)  
| | | | | | age >= 49.5: YES(1.0/1.0)  
| | | | | region!=(INNER\_CITY)|(RURAL)|(TOWN)  
| | | | | | age < 41.5: YES(2.0/0.0)  
| | | | | | age >= 41.5: NO(3.0/0.0)  
| | | | income >= 30340.85  
| | | | | children=(3)



| | | | | | income < 43235.05: NO(4.0/0.0)  
| | | | | | income >= 43235.05: YES(3.0/0.0)  
| | | | | children!=(3): YES(16.0/0.0)  
| | | children!=(3)|(2)|(1)  
| | | | income < 16479.6: YES(9.0/2.0)  
| | | | income >= 16479.6: YES(33.0/1.0)  
children!=(3)|(0)|(2)  
| income < 15576.45  
| | age < 41.5  
| | | income < 12694.1: NO(14.0/0.0)  
| | | income >= 12694.1  
| | | | income < 13437.85: YES(2.0/0.0)  
| | | | income >= 13437.85: NO(6.0/0.0)  
| | age >= 41.5: YES(2.0/0.0)  
| income >= 15576.45: YES(104.0/5.0)