MACHINE LEARNING AND PATTERN RECOGNITION

MODEL CODE: B9DA109
LECTURER: Satya Prakash
STUDENT: Peter Ibeabuchi
STUDENT ID: 20007349

November 2023

INTRODUCTION

This assessment looks into the application of decision tree entropy and information gain to make predictions. By calculating entropy and information gain, we can identify the most informative attributes in a dataset and construct a decision tree that effectively partitions the data based on these attributes.

THE DATASET

The dataset shows a table of five features, Age, Hair size, Browneyes, Sex and Won. The class we are interested in predicting is "Won", thus, this is our decision column. This class has a binary value of "Yes" or "No", indicating whether or not the person won the competition.

Age	Hair_Size	Brown_Eye	Sex	Won
youth	long	no	male	no
youth	long	no	female	no
middle_age	long	no	male	yes
senior	medium	no	male	yes
senior	short	yes	male	yes
senior	short	yes	female	no
middle_age	short	yes	female	yes
youth	medium	no	male	no
youth	short	yes	male	yes
senior	medium	yes	male	yes
youth	medium	yes	female	yes
middle_age	medium	no	female	yes
middle_age	long	yes	male	yes
senior	medium	no	female	no

DETERMINING FEATURE IMPORTANCE

The aim here is to determine what attribute classes in the dataset are more important than others in determining whether or not a person won the fashion show. To achieve this, we first find the entropy of the decision column "Won". Entropy of j is given as.

$$Entropy(t) = -\sum p(fj|t) \log p(j|t)$$

WON = Yes- 9, NO – 5

E(WON) =E(9,5)

= [- (9/14) *log2 (9/14) + (5/14) *log2(5/14)] = 0.94

CALCULATING THE AVERAGE WEIGHTED ENTROPY (AvgE)

In this session we will calculate the average weighted average of each independent column.

For the Age column, E(Won ,AGE)

	WON FASHION COMPETITION			
AGE		YES	NO	TOTAL
	YOUTH	2	3	5
	MIDDLE AGE	4	0	4
	SENIOR	3	2	4
				14

Calculating the average weighted entropy

E(Won, Age) = (5/14)*E(2,3) + (4/14)*E(4,0) + (5/14)*E(3,2)

= (5/14)(-[(2/5)log2(2/5) + (3/5)log2(3/5)]) + (4/14)(-[(4/4)log2(4/4) + (0/4)log2(0/4)]) +

(5/14)(-[(3/5)log2(3/5) + (2/5)log2(2/5)])

= (5/14) * 0.97 + (4/14) * 0 + (5/14) * 0.97 = **0.693**

For the Hair-size column, E(Won, Hair_size)

	WON FASHION COMPETITION			
HAIR SIZE		YES	NO	TOTAL
	LONG	2	2	4
	MEDIUM	4	2	6
	SHORT	3	1	4
				14

Calculating the average weighted entropy

E(Won, Hair_Size) = (4/14)*E(2,2) + (6/14)*E(4,2) + (4/14)*E(3,1)
= (4/14)(-[(2/4)log2(2/4) + (2/4)log2(2/4)]) + (6/14)(-[(4/6)log2(4/6) + (2/6)log2(2/6)]) + (4/14)(-[(3/4)log2(3/4) + (1/4)log2(1/4)])
= (4/14) * 1 + (6/14) * 0.918 + (4/14) * 0.811 = **0.910**

For the Brown-Eyes column, E(Won, Brown_Eyes)

	WON FASHION COMPETITION			
BROWN EYES		YES	NO	TOTAL
	YES	6	1	7
	NO	3	4	7
				14

Calculating the average weighted entropy

E(Won, Brown_Eyes) = (7/14)*E(6,1) + (7/14)*E(3,4)
(7/14)(-[(6/7)log2(6/7) + (1/7)log2(1/7)]) + (7/14)(-[(3/7)log2(3/7) + (4/7)log2(4/7)])
=(7/14) * 0.591 + (7/14) * 0.985 = **0.788**

For the Sex column, E(Won ,Sex)

	WON FASHION COMPETITION			
SEX		YES	NO	TOTAL
	MALE	6	2	8
	FEMALE	3	3	6
				14

Calculating the average weighted entropy

E(Won, Sex) = (8/14)*E(6,2) + (6/14)*E(3,3)
(8/14)(-[(6/8)log2(6/8) + (2/8)log2(2/8)]) + (6/14)(-[(3/6)log2(3/6) + (3/6)log2(3/6)])
=(8/14) * 0.811 + (6/14) * 1 = **0.892**

CALCULATING INFORMATION GAIN

The information Gain represents how much information a feature provides for the target variable. It is represented as thus:

$$Information\ Gain = \ Entropy_{\ parent} - Entropy_{\ children}$$

In our analysis therefore, the information gain is the entropy of the decision column, minus the entropy of each weighted average of the attribute column. The column with the highest information gain is the most important feature.

IG(AGE) = 0.94 - 0.693 = 0.247

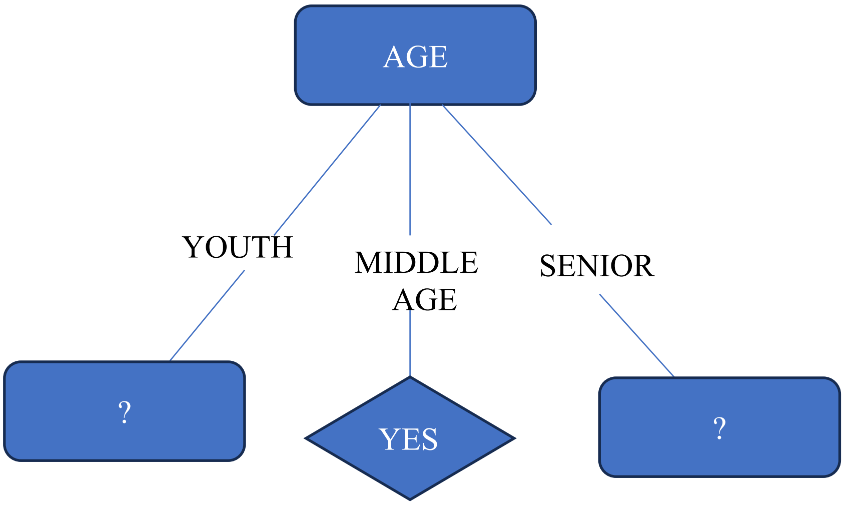
IG(HAIR_SIZE) = 0.94 - 0.910 =0.03

IG(BROWN_EYE) = 0.94 - 0.788 = 0.152

IG(SEX) = 0.94 - 0.892 = 0.048

Since age has the highest information gain, it is our most important feature, and thus our root node.

.



Now that we have the root node, we need to find the next node under “Youth”. Here is what the table look like:

Age	Hair_Size	Brown_Eye	Sex	Won
youth	long	no	male	no
youth	long	no	female	no
youth	medium	no	male	no
youth	short	yes	male	yes
youth	medium	yes	female	yes

Calculate the Entropy of Youth;
E(Youth) = E(3,2)
= E(Youth) = (-(3/5)log2(3/5)-(2/5)log2(2/5)) = **0.971**

Next, we calculate the entropy for each column just as before.

Average weighted entropy Hair_size E(Youth, Hair_Size)

	WON FASHION COMPETITION			
HAIR SIZE		YES	NO	TOTAL
	LONG	0	2	2
	MEDIUM	1	1	2
	SHORT	1	0	1
				5

E(Youth, Hair_Size) = (2/5)*E(0,2) + (2/5)*E(1,1) + (1/5)*E(1,0) = 2/5 = **0.4**

Average weighted entropy Brown Eyes E (Youth, Brown Eyes)

	WON FASHION COMPETITION			
BROWN EYES		YES	NO	TOTAL
	YES	0	2	2
	NO	3	0	3
				5

E(Youth, Brown Eye) = (2/5)*E(0,2) + (3/5)*E(3,0) = 0

Average weighted entropy Sex E(Youth, Sex)

	WON FASHION COMPETITION			
SEX		YES	NO	TOTAL
	MALE	1	2	3
	FEMALE	1	1	2
				5

E(Youth, Sex) = (3/5)*E(3,2) + (2/5)*E(1,1)
= (3/5) [-(1/3)log2(1/3)+(2/3)log2(2/3)] + (2/5)[-(1/2)log2(1/2)+(1/2)log2(1/2)]
= 0.95

INFORMATION GAIN

The information gain in this next phase is as follows:

IG (YOUTH, HAIR_SIZE) = 0.971-0.4 = 0.571
IG (YOUTH, BROWN EYE) = 0.971-0 = 0.971
IG (YOUTH, SEX) = 0.971-0.95 = 0.021

Thus, our next node is Brown eyes, since it has the highest information gain in the phase

Similarly, we will follow similar steps to find the information gain using “Senior”. Here is what the table looks like:

Age	Hair_Size	Brown_Eye	Sex	Won
senior	medium	no	male	yes
senior	short	yes	male	yes
senior	short	yes	female	no
senior	medium	yes	male	yes
senior	medium	no	female	no

Calculate the Entropy of Senior
E(Senior) = E (3,2)

E(Senior) = (-(2/5)log(2/5)-(3/5)log(3/5)) = **0.971**

Calculating the Average Weighted Entropy in E(Senior, Hair_size)

	WON FASHION COMPETITION			
HAIR SIZE		YES	NO	TOTAL
	MEDIUM	2	1	3
	SHORT	1	1	2
				5

E(Senior, Hair_Size) = (3/5)*E(2,1) + (2/5)*E(1,1)
= (3/5)*(-[(2/3)log2(2/3) + (1/3)log2(1/3)]) + (2/5)*(-[(1/2)log2(1/2) + (1/2)log2(1/2)])
(3/5) * 0.918 + (2/5) * 1 = **0.9508**

Calculating the Average Weighted Entropy in Sex, E(Senior, Sex)

	WON FASHION COMPETITION			
SEX		YES	NO	TOTAL
	MALE	3	0	3
	FEMALE	0	3	2
				5

E(Senior, Sex) = (3/5)*E(3,0) + (2/5)*E(0,2)
(3/5)*(-[(3/3)log2(3/3) + (0/3)log2(0/3)]) + (2/5)*(-[(0/2)log2(0/2) + (2/2)log2(2/2)]) (3/5) * 0 + (2/5) * 0 = **0**

INFORMATION GAIN

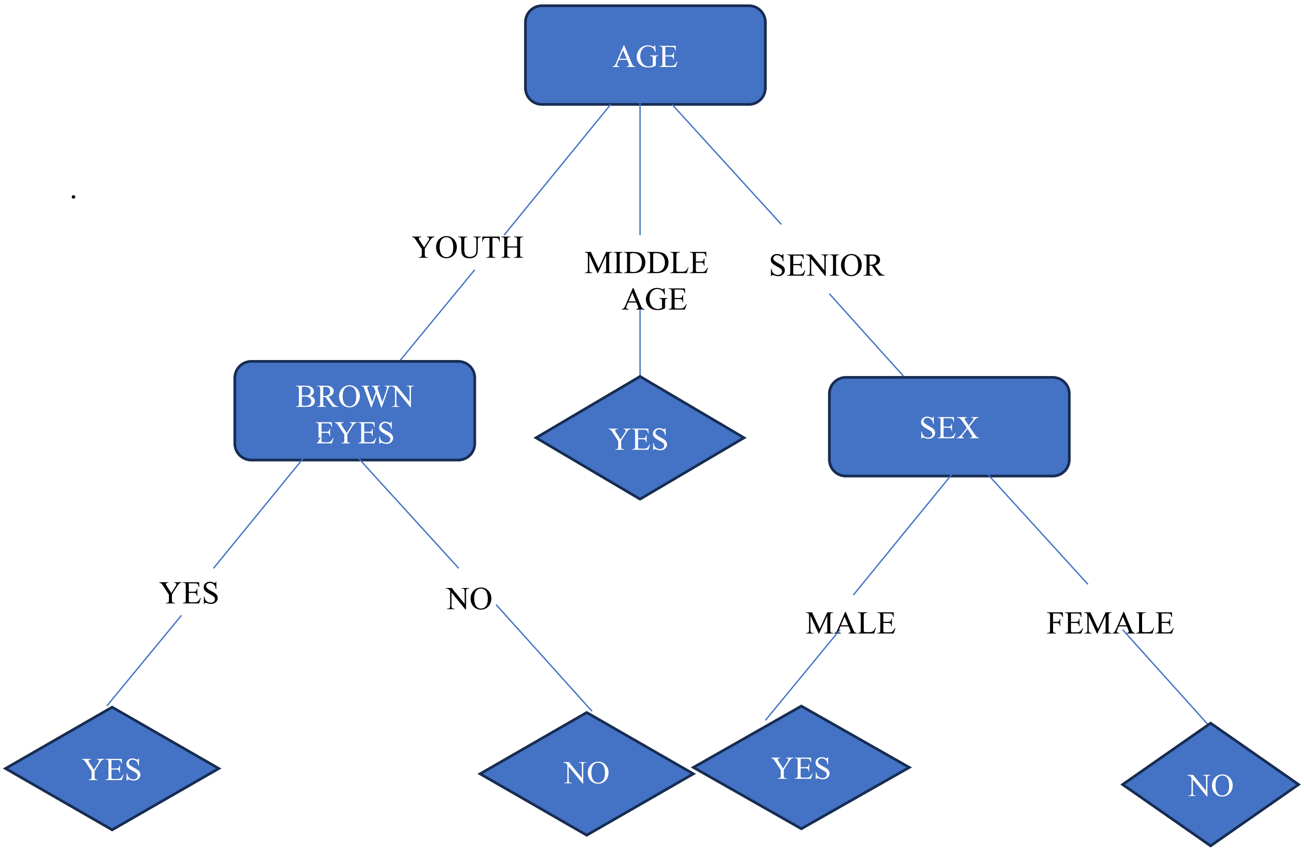
The information gain in this next phase is as follows:

IG (YOUTH, HAIR_SIZE) = 0.971-0.9508= 0.0202

IG (YOUTH, SEX) = 0.971- 0 = 0.971

Thus, the next node under Senior is sex. Our decision three looks like this

.



This is therefore our final tree as we no longer have attributes to consider.

Final Observation:

The Age column is the most important feature and the highest information gain.