MDL Assignment-4
2018101075

1. Dataset

(2018101075 % 10 + 1) = 6

2. Flipped Entries

a) (5 + 7) % 12 + 1 = 1

b) 7 % 12 + 1 = 8

3. Dataset After Flipping

Sleep Pattern f[0]	Junk Food Consumption f[1]	Exercise daily f[2]	Healthy
Irregular Sleep	High	Yes	Yes
Irregular Sleep	Normal	No	No
Irregular Sleep	Low	Yes	Yes
Irregular Sleep	Low	No	Yes
Good Sleep	High	Yes	Yes
Good Sleep	Normal	No	No
Good Sleep	Low	Yes	Yes
Good Sleep	High	No	Yes
Long Sleep	High	No	No
Long Sleep	Normal	Yes	Yes
Long Sleep	Low	Yes	Yes
Long Sleep	Normal	No	Yes

Step 1: Calculating the Entropy of the Dataset

Entropy of a boolean random variable that is true with probability q is:

$B(q) = - q \log_2(q) - (1 - q) \log_2(1 - q)$

There are p = 9 Yes and n = 3 No in the dataset

Entropy of the dataset $= B(p / (p + n)) = B(6 / 9)$

$= - 9 / 12 * \log_2(9 / 12) - 3 / 12 * \log_2(3 / 12)$

$= 0.9852$

Step 2: Calculating the Importance value of each attribute

Information Gain:

$$Gain(A) = B(\frac{p}{p+n}) - Remainder(A)$$

Where Remainder(A) is :

$$Remainder(A) = \sum_{k=1}^{d} \frac{p_k+n_k}{p+n} B(\frac{p_k}{p_k+n_k})$$

Where p_k is the positive entries and n_k is the negative entries for the E_k value of attribute A.

More the Information Gain of the Attribute, More is the Importance of that Attribute.

1. Sleep Pattern

	Yes (p_k)	No (n_k)
Irregular Sleep	3	1
Good Sleep	3	1
Long Sleep	3	1

Remainder(Sleep Pattern) = 4/12 * B(3/4) + 4/12 * B(3/4) + 4/12 * B(3/4)
= 0.4635
Gain(Sleep Pattern) = B(9/12) - Remainder(Sleep Pattern)
= 0.5216

2. Junk Food Consumption

	Yes (p_k)	No (n_k)
High	3	1
Normal	2	2
Low	4	0

Remainder(Junk Food Consumption)
= 4/12 * B(3/4) + 4/12 * B(2/4) + 4/12 * B(4/4)
= 0.3450
Gain(Junk Food Consumption)
= B(9/12) - Remainder(Junk Food Consumption)
= 0.6402

3. Exercise daily

	Yes (p_k)	No (n_k)
Yes	6	0
No	3	3

Remainder(Exercise daily) = 6/12 * B(6/6) + 6/12 * B(3/6)
= 0.8571
Gain(Exercise daily) = B(9/12) - Remainder(Exercise daily)
= 0.1280

Maximum Gain Value is of Junk Food Consumption. Junk Food Consumption will be the root node. Now the dataset is divided into 3 parts: High, Normal and Low Junk Food Consumption.

Step 3: Forming the subtree for Junk Food Consumption: High

There are p = 3 Yes and n = 1 No for this dataset

1. Sleep Pattern

	Yes (p_k)	No (n_k)
Irregular Sleep	1	0
Good Sleep	2	0
Long Sleep	0	1

Remainder(Sleep Pattern) = 1/4 * B(1/1) + 2/4 * B(2/2) + 1/4 * B(0/1)
= 0.0
Gain(Sleep Pattern) = B(3/4) - Remainder(Sleep Pattern)
= 0.81

2. Exercise daily

	Yes (p_k)	No (n_k)
Yes	2	0
No	1	1

Remainder(Exercise daily) = 2/4 * B(2/2) + 2/4 * B(1/2)
= 0.5
Gain(Exercise daily) = B(3/4) - Remainder(Exercise daily)
= 0.3112

Maximum Gain Value is of Sleep Pattern. Sleep Pattern will be the child of Junk Food Consumption with edge value High.

Step 4: Forming the subtree of Junk Food Consumption: High → Sleep Pattern:

Sleep Pattern has 3 edges: Irregular Sleep, Good Sleep and Long Sleep. As only one attribute is left, we need to process Exercise daily for the child node for each edge.

1. Irregular Sleep → Exercise daily: Yes → Yes
   As we have only one classification at the end, classification Yes will be returned to the Irregular Sleep as the child node. Finally,
Junk Food Consumption: High → Sleep Pattern: Irregular Sleep → Yes

2. Good Sleep → Exercise daily: Yes → Yes
   Good Sleep → Exercise daily: No → Yes
   As we have only one classification at the leaf node, classification Yes will be returned to the Good Sleep as the child node. Finally,
Junk Food Consumption: High → Sleep Pattern: Good Sleep → Yes

3. Long Sleep → Exercise daily: No → No
   As we have only one classification at the leaf node, classification No will be returned to the Long Sleep as the child node. Finally,
Junk Food Consumption: High → Sleep Pattern: Long Sleep → No

Step 5: Forming the subtree for Junk Food Consumption: Normal

There are p = 2 Yes and n = 2 No for this dataset

1. Sleep Pattern

	Yes (p_k)	No (n_k)
Irregular Sleep	0	1
Good Sleep	0	1
Long Sleep	2	0

Remainder(Sleep Pattern) = 1/4 * B(0/1) + 1/4 * B(0/1) + 2/4 * B(2/2)
= 0.0
Gain(Sleep Pattern) = B(2/4) - Remainder(Sleep Pattern)
= 1.0

2. Exercise daily

	Yes (p_k)	No (n_k)
Yes	1	0
No	1	2

Remainder(Exercise daily) = 2/4 * B(2/2) + 2/4 * B(1/2)
= 0.6887
Gain(Exercise daily) = B(2/4) - Remainder(Exercise daily)
= 0.3112

Maximum Gain Value is of Sleep Pattern. Sleep Pattern will be the child of Junk Food Consumption with edge value Normal.

Step 6: Forming the subtree of Junk Food Consumption: Normal → Sleep Pattern:

1. Irregular Sleep → Exercise daily: No → No
   As we have only one classification at the end, classification No will be returned to the Irregular Sleep as the child node. Finally,

Junk Food Consumption: High → Sleep Pattern: Irregular Sleep → No

2. Good Sleep → Exercise daily: No → No

   As we have only one classification at the leaf node, classification No will be returned to the Good Sleep as the child node. Finally,
Junk Food Consumption: High → Sleep Pattern: Good Sleep → No

3. Long Sleep → Exercise daily: Yes → Yes
   Long Sleep → Exercise daily: No → Yes

   As we have only one classification at the leaf node, classification Yes will be returned to the Long Sleep as the child node. Finally,
Junk Food Consumption: High → Sleep Pattern: Long Sleep → Yes

Step 7: Forming the subtree for Junk Food Consumption: Low

There are p = 4 Yes and n = 0 No for this dataset

1. Sleep Pattern

	Yes (p_k)	No (n_k)
Irregular Sleep	2	0
Good Sleep	1	0
Long Sleep	1	0

Remainder(Sleep Pattern) = 2/4 * B(2/2) + 1/4 * B(1/1) + 1/4 * B(1/1)
= 0.0
Gain(Sleep Pattern) = B(4/4) - Remainder(Sleep Pattern)
= 0.0

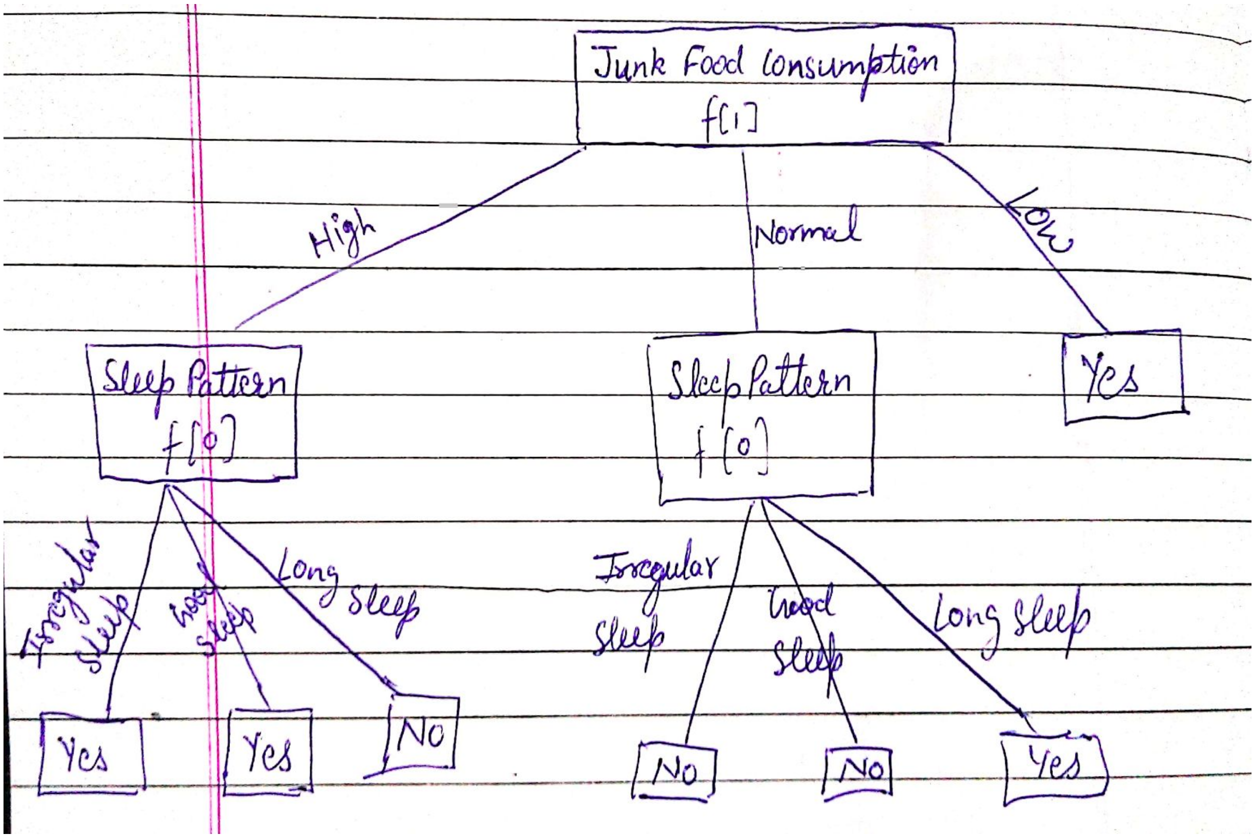
2. Exercise daily

	Yes (p_k)	No (n_k)
Yes	3	0
No	1	0

Remainder(Exercise daily) = 2/4 * B(3/3) + 2/4 * B(1/1)
= 0.0
Gain(Exercise daily) = B(4/4) - Remainder(Exercise daily)
= 0.0

Both have the same Gain Value, so we can choose any one of the two. Whichever we choose, we have only one classification further, so classification Yes will be returned to Low as the child node. Finally, Junk Food Consumption: Low → Yes

Decision Tree


Junk Food consumption
f[1]
High
Normal
Low
Sleep Pattern
f[0]
Sleep Pattern
f[0]
Yes
Irregular Sleep
Good Sleep
Long Sleep
Irregular Sleep
Good Sleep
Long Sleep
Yes
Yes
No
No
No
Yes