CSE438 Assignment-1
Name: Joyanta ID: 19141016

Types of Data

BMI → Numerical (Interval)
Smoking → Categorical (Nomial)
Sex → Categorial (Nomial)
GenHealth → Categorical (Nomial)
SleepTime → Categorical
SkinCancer → Categorical (Nomial)

Calculating Entropy

Selecting 100 rows from the dataset,

BMI (making data intervals):

$x < 18.5 \longrightarrow 2$

$18.5 < x < 25 \rightarrow 32$

$25 < x < 30 \rightarrow 36$

$30 < x < 35 \rightarrow 22$

$x > 35 \rightarrow 8$.

$$\therefore \text{Entropy}(BMI) = -\frac{2}{100}\log_2\left(\frac{2}{100}\right) - \frac{32}{100}\log_2\left(\frac{32}{100}\right) - \frac{36}{100}\log_2\left(\frac{36}{100}\right) - \frac{22}{100}\log_2\left(\frac{22}{100}\right) - \frac{8}{100}\log_2\left(\frac{8}{100}\right)$$

$= 1.941688$. Ans.

Smoking:

0 — 51

1 — 49

$\therefore$ Entropy (Smoking) $= -\frac{51}{100}\log_2\left(\frac{51}{100}\right) - \frac{49}{100}\log_2\left(\frac{49}{100}\right)$

$= 0.9997$. Ans.

Sex:

0 - 70

1 — 30

$\therefore$ Entropy (Sex) $= -\frac{70}{100}\log_2\left(\frac{70}{100}\right) - \frac{30}{100}\log_2\left(\frac{30}{100}\right)$

$= 0.88132$

GenHealth:

0 - 9

1 — 18

2 - 31

3 — 8

4 — 34

$\therefore$ Entropy (GenHealth)

$= -\frac{9}{100}\log_2\left(\frac{9}{100}\right) - \frac{18}{100}\log_2\left(\frac{18}{100}\right)$

$- \frac{31}{100}\log_2\left(\frac{31}{100}\right) - \frac{8}{100}\log_2\left(\frac{8}{100}\right)$

$- \frac{34}{100}\log_2\left(\frac{34}{100}\right)$

$= 2.102347$ Ans

SleepTime:

4 → 3
5 → 11
6 → 17
7 → 21
8 → 35
9 → 5
10 → 5
12 → 2
15 → 1

∴ Entropy (SleepTime)

$$= -\frac{3}{100}\log_2\left(\frac{3}{100}\right) - \frac{11}{100}\log_2\left(\frac{11}{100}\right) - \frac{17}{100}\log_2\left(\frac{17}{100}\right) - \frac{21}{100}\log_2\left(\frac{21}{100}\right) - \frac{35}{100}\log_2\left(\frac{35}{100}\right) - \frac{5}{100}\log_2\left(\frac{5}{100}\right) - \frac{5}{100}\log_2\left(\frac{5}{100}\right) - \frac{2}{100}\log_2\left(\frac{2}{100}\right) - \frac{1}{100}\log_2\left(\frac{1}{100}\right)$$

$$= 2.551087.$$ Ans.