Ai 305 Assignment

Question 1:

a) IQ = 110

GPA = 4.0

Regression model:

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 (X_1 \cdot X_2) + \beta_5 (X_1 \cdot X_3)$$

X_1 = GPA = 4.0 $\quad$ X_3 = level (1 university, 0, high school)

X_2 = IQ = 110 $\quad$ $X_4 = X_1 \cdot X_2 = 4.0 \times 110$

$X_5 = X_1 \cdot X_3 = 4.0 \times$ level

The coefficients:-

$\beta_0 = 50 \quad \beta_1 = 20 \quad \beta_2 = 0.07 \quad \beta_3 = 35$

$\beta_4 = 0.01 \quad \beta_5 = -10$

$\therefore$ for ($X_3 = 1$)

$$\hat{y} = 50 + 20(4) + 0.07(110) + 35(1) + 0.01(4 \times 110) + (-10)(4 \times 1)$$

$$= 50 + 80 + 7.7 + 35 + 4.4 - 40 =$$

$$= 137.1$$

$\therefore$ The predicted salary is 137.1 $\Leftrightarrow$ 137,100

B)

$\beta_2 = 0.07$: This is a small effect, which means the IQ does not have a strong impact on starting salary.

$\beta_4 = 0.01$: This is also a small effect stating that the combined effect of GPA and IQ is not a strong determinant of salary.

Ci : false

we look at the coefficient $\beta_3 = 35$, which represent the effects of being a University student (compared to high school student). and since $\beta_3 = 35$ is positive, this implies that for a fixed value of IQ and GPA, University student earn 35 more, than high school student.

Cii : True

if you compare the predicted salary for University student ($X_3 = 1$) and high school graduate ($X_3 = 0$)

* for university student, salary equation includes

$$\beta_5 (X_1 \cdot X_3 = -10 \cdot X$$

* for high school ($X_3 = 0$) so the term $\beta_5 (X_1 \cdot X_3)$ drops out.

∴ high school student could potentially earn more if GPA is very high because the negative interaction between GPA & Uni status become dominant.

Question 2:

a) Logistic regression model is given by:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

where:

p = probability of getting an (A) (response variable)

X_1 = Number of hours studied

X_2 = Undergrad GpA

Coefficients = $\beta_0 = -6$ $\beta_1 = 0.05$ $\beta_2 = 1$

∴ if $X_1 = 40$ hrs, $X_2 = 3.5$ GPA then,

$\log\left(\frac{p}{1-p}\right) = -6 + 0.05(40) + 1(3.5) =$

=

∴ the equation became:

$$\log\left(\frac{p}{1-p}\right) = -0.5$$

$$p = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}$$

Substitue $\beta_0 + \beta_1 x_1 + \beta_2 x_2 = -0.5$

$$p = \frac{e^{-0.5}}{1 + e^{-0.5}}$$

∴ $e^{-0.5} \approx 0.6065$

∴ $p = \frac{0.6065}{1 + 0.6065} = \frac{0.6065}{1.6065} \approx 0.377$

∴ the estimated prob. that student who study 40 hrs with gpa 3.5 to get an A = 37.7%

B)

to find X_1 when the probability of getting A is 50% = 0.5

$$\log\text{-odd} = \log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$\therefore$

$$P = 0.5 \quad \therefore \log\left(\frac{0.5}{1-0.5}\right) = \log(1) = 0$$

$\therefore$ the equation becomes:

$$0 = -6 + 0.05X_1 + 1(3.5)$$

$$0 = -6 + 0.05X_1 + 3.5$$

$$0 = -2.5 + 0.05X_1$$

$$2.5 = 0.05X_1$$

$$\frac{2.5}{0.05} = \frac{0.05X_1}{0.05} \Rightarrow X_1 = \frac{2.5}{0.05} = 50$$

$\therefore$ Student will need to study for 50 hours to have 50% chance of getting A in the class.

Question 3:

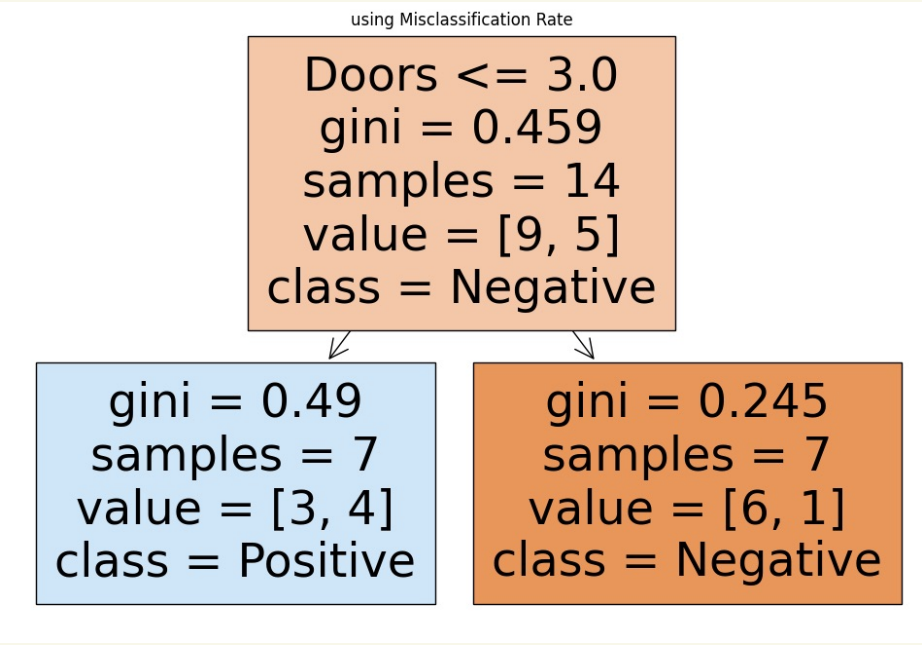
P_+ = proportion of positive class
P_- = proportion of Negative class

Total example = 14
positive class = 6
Negative class = 8

a: mis calculation rate.

Miscalculation rate = $1 - \max(P_+, P_-)$

$\therefore$ miscalculation rate (S) = $1 - \text{Max}\left(\frac{6}{14}, \frac{8}{14}\right)$



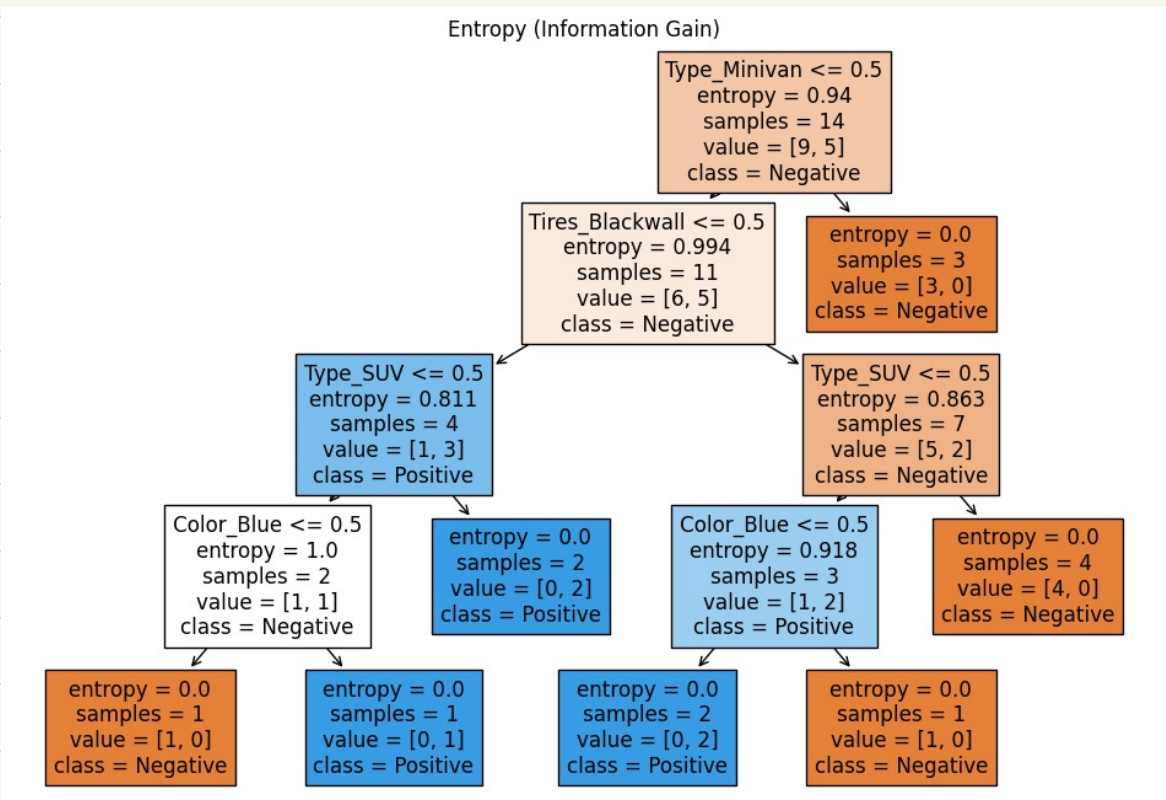
bc entropy & Information gain

$$\text{Entropy}(S) = -p_+ \log_2(p_+) - p_- \log_2(p_-)$$

$$IG(S,A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

where: S = dataset, S_v = subset of the dataset S corresponding to value A.

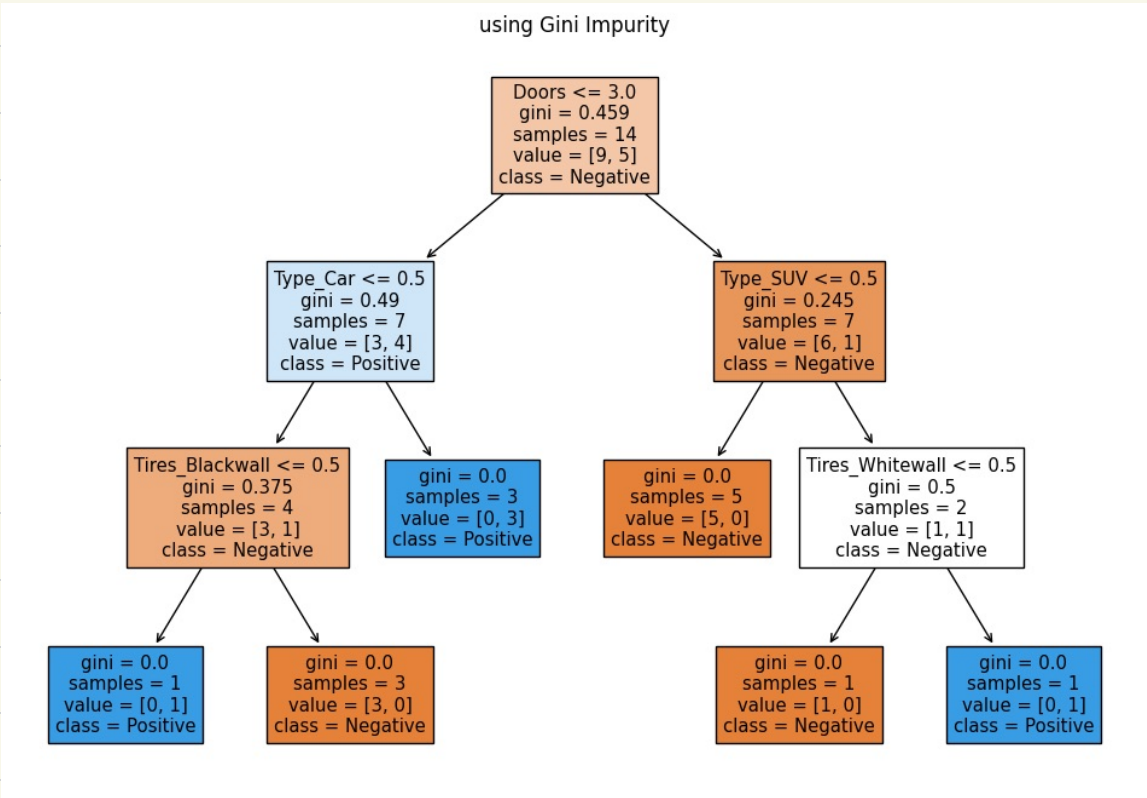
$$\therefore \text{Entropy}(S) = -\left(\frac{6}{14}\log_2\frac{6}{14} + \frac{8}{14}\log_2\frac{8}{14}\right)$$



C : Gini :

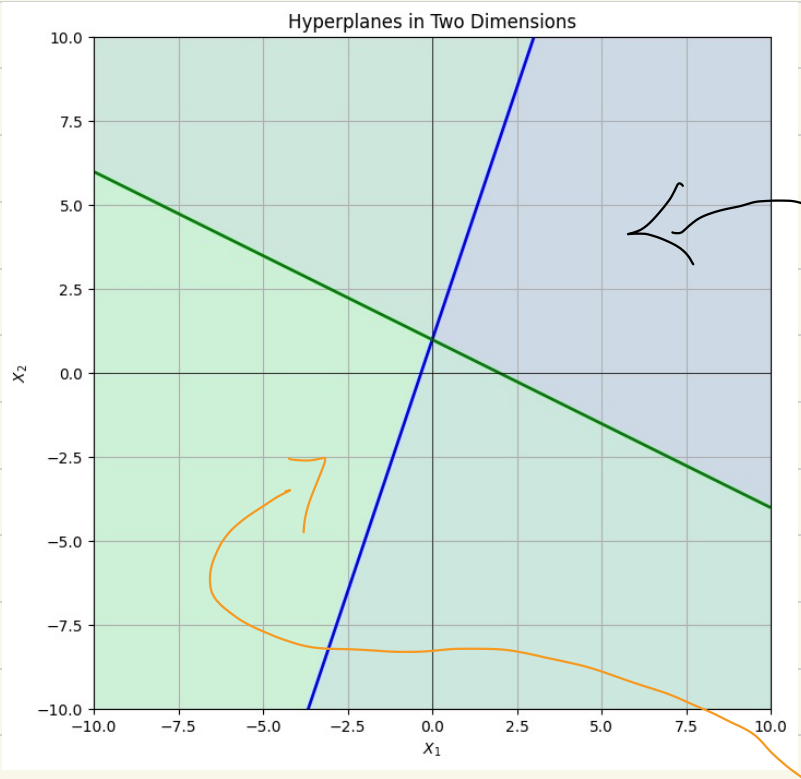
$$gini(S) = 1 - (P_+^2 + P_-^2)$$

$$\therefore \quad gini(S) = 1 - \left(\left(\frac{6}{14}\right)^2 + \left(\frac{8}{14}\right)^2\right)$$



Note:- I used python code to generate the decision tree. Using decision tree classifier. The python code is provided in a seperate file.

Question 4:



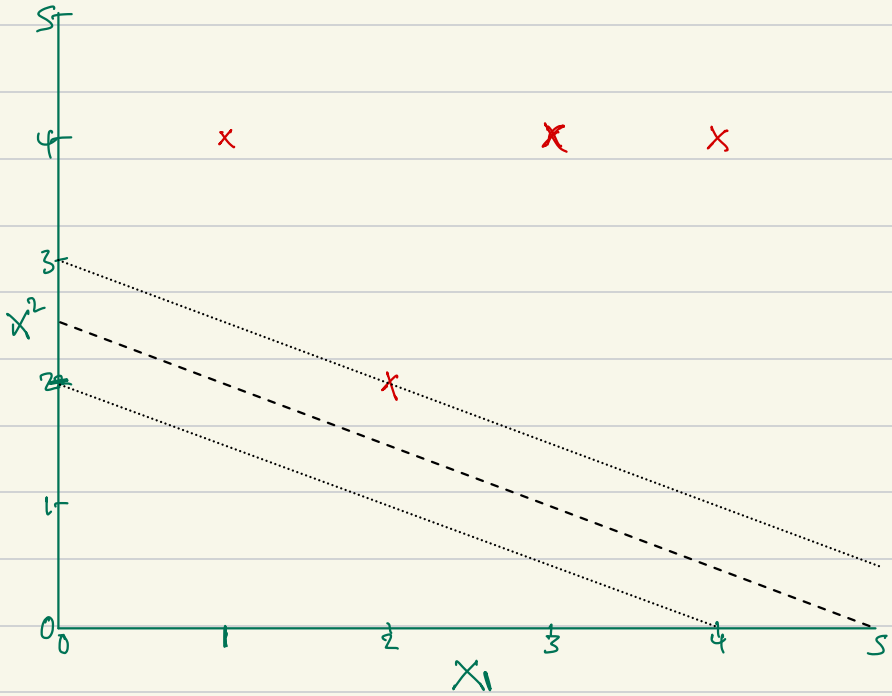
Blue: represent the hyperplane : $1 + 3X_1 - X_2 = 0$

* $1 + 3X_1 - X_2 > 0$
* $1 + 3X_1 - X_2 < 0$

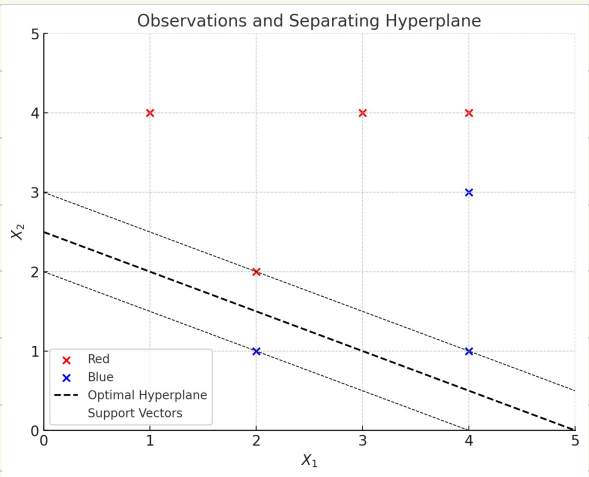
Green: represent the hyper-plane : $-2 + X_1 + 2X_2 = 0$

* $-2 + X_1 + 2X_2 > 0$
* $-2 + X_1 + 2X_2 < 0$

Question 5:



* red
* Blue
---- optimal hyperplane
support vectors.



Observation Summary:

Blue class : point (2, 1), (4, 3), (4, 1)
Red class : point (3, 4), (2, 2), (4, 4)

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0$$

let assume : $x_2 = -0.5 x_1 + 2.5$

$$\therefore \quad 0.5 x_1 + x_2 - 2.5 = 0$$

Coefficient :

$\beta_0 = -2.5$

$\beta_1 = 0.5$

$\beta_2 = 1$

$\therefore$ The final equation of the optimal hyperplane is:

$$0.5 x_1 + x_2 - 2.5 = 0$$