Harsh Sanjay Agrawal
CWID - 10475285

Homework 3

Q1. Explain what is the bias-variance trade-off? Describe few techniques to reduce bias and variance respectively.

Bias is defined as the inability of the machine learning algorithm to accurately find the relationship of the data, or the mean of the given data. For example, if a linear regression model is applied to data which is not linearly related, the prediction will be assumed to be linear, resulting in a high bias and consequently under fitting.

On the contrary, variance is the over fitting of a dataset. If the model learns the training dataset very closely, it fails to generalize well enough in the real world or testing data. As the model becomes more and more complex, it tends to have a higher variance and a lower bias.

The concept of the bias of a model decreasing as the variance rises is called bias-variance tradeoff and this needs to be used to find the minimum training and testing error of the model.

Example:

Low-variance ML algorithms: Linear Regression, Logistic Regression, Linear Discriminant Analysis.

High-variance ML algorithms: Decision Trees, k-NN, and Support Vector Machines.

The main method to reduce bias in a machine learning model is random sampling in data selection. Constant monitoring using callbacks is often required to ensure that the AI doesn't develop any bias during the training process. Increasing features and employing feature engineering are other ways.

The main method to reduce variance in a machine learning model is to reduce noise in the training data. The other way is to introduce randomness during the learning process. For example, random splits while working on random forests, random initialization of weights in neural networks and shuffling of training data in stochastic gradient descent. Increasing the training data size and reducing the number of features can also be extremely helpful.

Harsh Sanjay Agrawal
CWID - 10475285

Q2. Assume the following confusion matrix of a classifier. Please compute its

1) precision
2) recall
3) F1-score.

Precision = TP / (TP + FP) = 50 / (50 + 40) = **0.556**

Recall = TP / (TP / FN) = 50 / (50 + 30) = **0.625**

F1 Score = 2 * (Precision * Recall) / (Precision + Recall) = 2 * (0.625 * 0.556) / (0.625 + 0.556)

F1 Score = 0.695 / 1.181 = 0.588484335 = **0.588**

Harsh Sanjay Agrawal
CWID - 10475285

Question 3: Build a decision tree using the following training instances (using information gain approach):

Q3) Entropy of playing tennis is given as

$$\frac{-6}{10}\log\frac{6}{10} - \frac{4}{10}\log\frac{4}{10} = 0.97$$

Let T be event of playing tennis

$$H(T) = 0.97$$

Now for calculating gain outlook (O)

$H(T/O)$ = , for sunny (4) $P(+) = 1/4$, $P(-) = 3/4$
for overcast (2) $P(+) = 2/2$, $P(-) = 0$
for rain (4) $P(+) = \frac{3}{4}$ $P(-) = 1/4$

$$\therefore H(T/O) = \frac{4}{10} \times 0.811 + \frac{2}{10} \times 0 + \frac{4}{10} \times 0.811$$

$$= 0.648$$

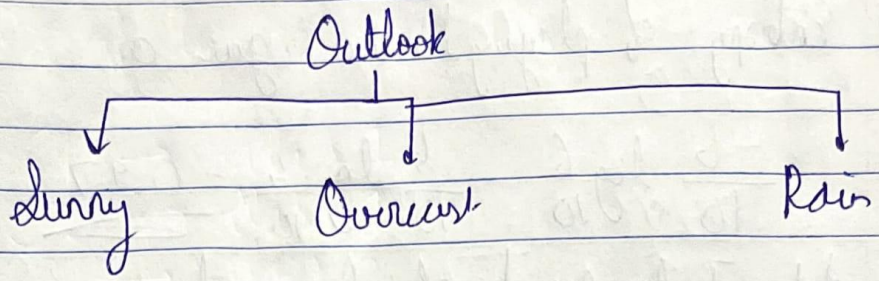
$$\therefore \text{Gain} = 0.97 - 0.648 = 0.321$$

Similarly we calculate for Temp (Te), Humidity (H), Wind (W)

$$\therefore G(T/Te) = 0.096$$
$$G(T/H) = 0.124$$
$$G(T/W) = 0.195$$

∴ Step 1 is calculated,

Harsh Sanjay Agrawal
CWID - 10475285



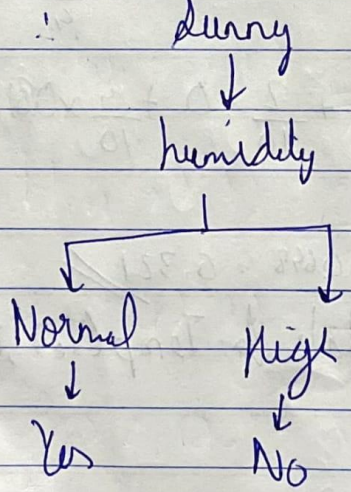
WKT overast ⟹ Yes

(i) ∴ Calculating for Sunny,

Looking at table, we see that,

for sunny, Gain(Humidity) = 0

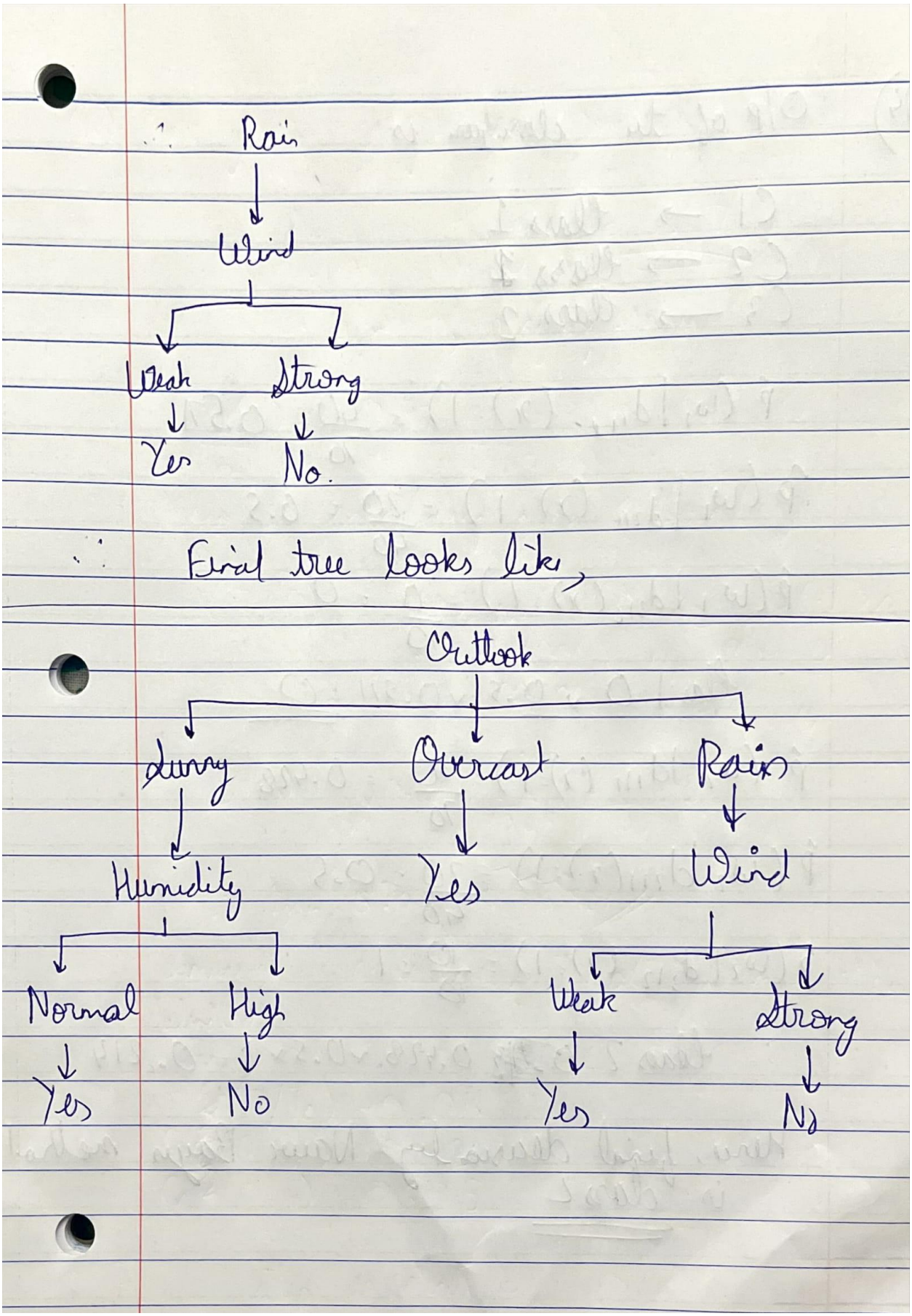
∴



(ii) ∴ Calculating for raing)

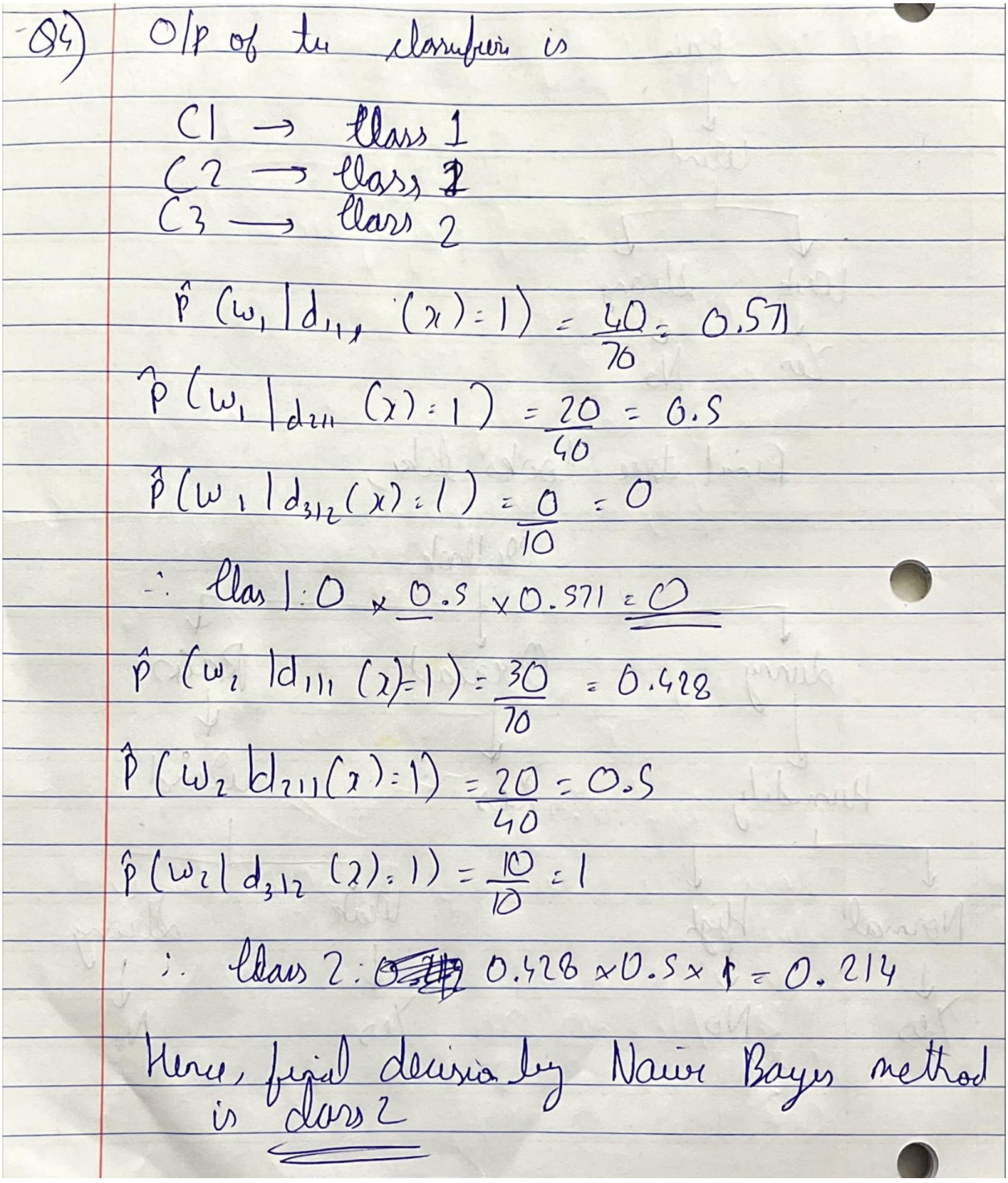
Gain (wind) = 0

Harsh Sanjay Agrawal
CWID - 10475285



Harsh Sanjay Agrawal
CWID - 10475285

Question 4. The naïve Bayes method is an ensemble method as we learned in Module 5. Assuming we have 3 classifiers, and their predicted results are given in the table 1. The confusion matrix of each classifier is given in table 2. Please give the final decision using the Naïve Bayes method:



Q4) O/p of the classifiers is

C1 → Class 1
C2 → Class 2
C3 → Class 2

$\hat{P}(\omega_1 | d_{1,1}(x)=1) = \frac{40}{70} = 0.571$

$\hat{P}(\omega_1 | d_{2,1}(x)=1) = \frac{20}{40} = 0.5$

$\hat{P}(\omega_1 | d_{3,2}(x)=1) = \frac{0}{10} = 0$

∴ Class 1: $0 \times 0.5 \times 0.571 = 0$

$\hat{P}(\omega_2 | d_{1,1}(x)=1) = \frac{30}{70} = 0.428$

$\hat{P}(\omega_2 | d_{2,1}(x)=1) = \frac{20}{40} = 0.5$

$\hat{P}(\omega_2 | d_{3,2}(x)=1) = \frac{10}{10} = 1$

∴ Class 2: $0.428 \times 0.5 \times 1 = 0.214$

Hence, final decision by Naive Bayes method is class 2