# Homework 4

Your Name: Srujan Shekar Shetty

Student ID: A20529733

**1. Build a decision Tree [50]. Note: solve this problem by manual calculations**

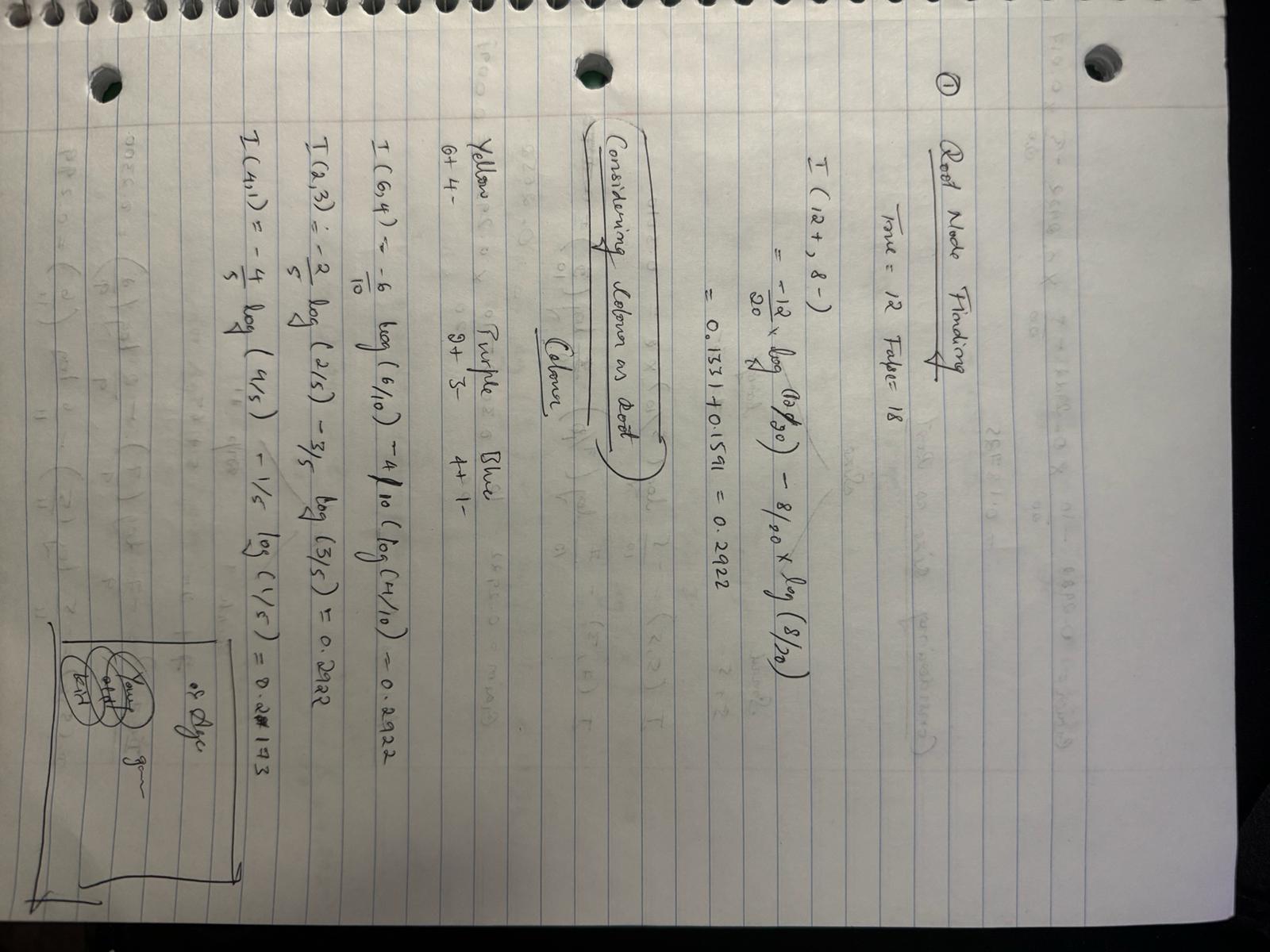
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Color** | **Size** | **Act** | **Age** | **Inflated** |
| YELLOW | SMALL | STRETCH | Young | T |
| YELLOW | SMALL | STRETCH | Old | T |
| YELLOW | SMALL | STRETCH | Old | T |
| YELLOW | SMALL | DIP | Kid | F |
| YELLOW | SMALL | DIP | Kid | F |
| YELLOW | LARGE | STRETCH | Old | T |
| YELLOW | LARGE | STRETCH | Old | T |
| YELLOW | LARGE | DIP | Young | F |
| YELLOW | LARGE | DIP | Young | T |
| YELLOW | LARGE | DIP | Young | F |
| PURPLE | SMALL | STRETCH | Young | F |
| PURPLE | SMALL | STRETCH | Old | T |
| PURPLE | SMALL | STRETCH | Old | F |
| PURPLE | SMALL | DIP | Kid | T |
| PURPLE | SMALL | DIP | Kid | F |
| Blue | LARGE | STRETCH | Kid | T |
| Blue | LARGE | DIP | Young | T |
| Blue | LARGE | DIP | Young | F |
| Blue | LARGE | DIP | Old | T |
| Blue | LARGE | DIP | Young | T |

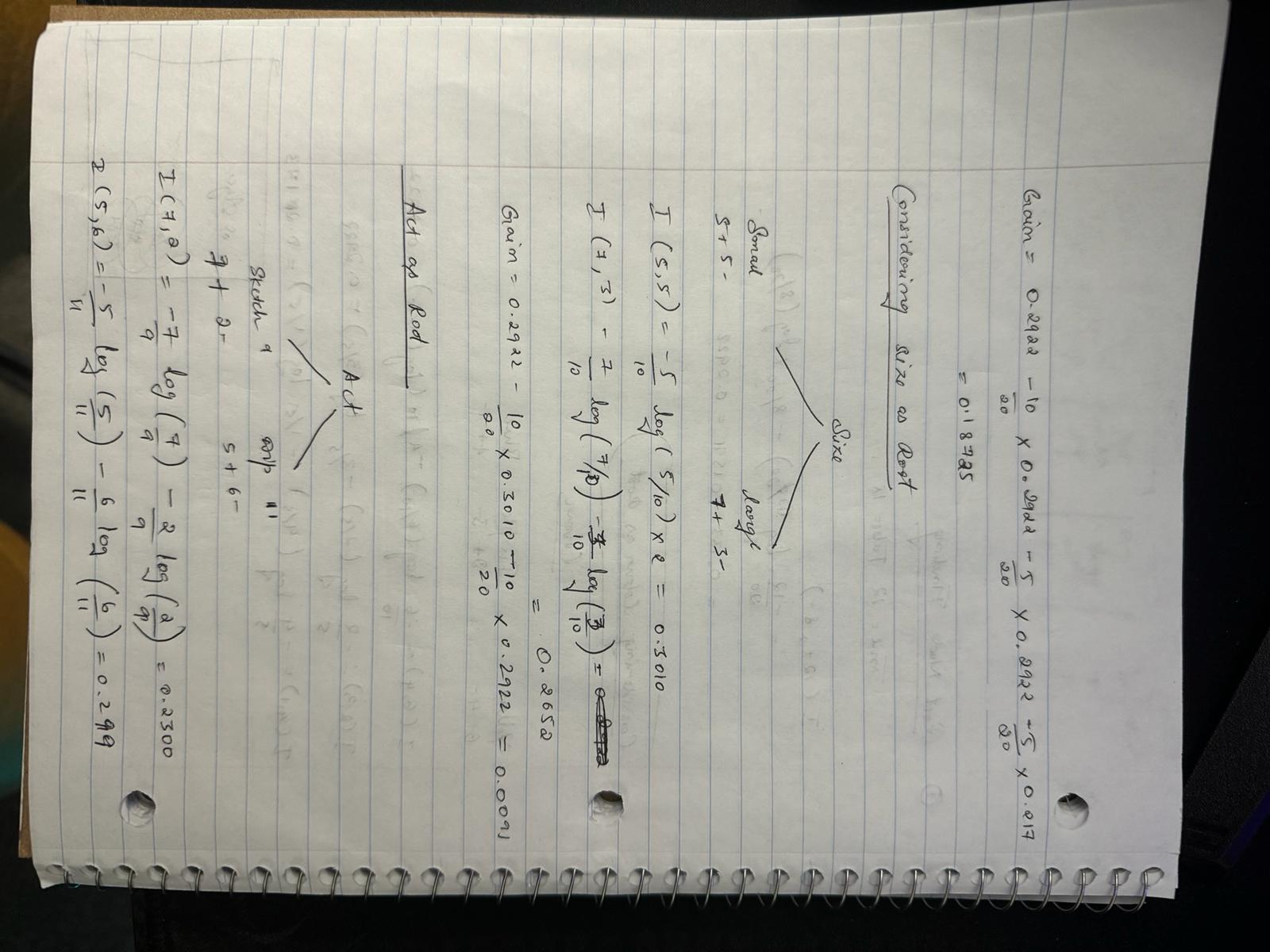
1). The data above is used to classify whether a balloon is inflated or not. [30] Note: use 10 as the log base

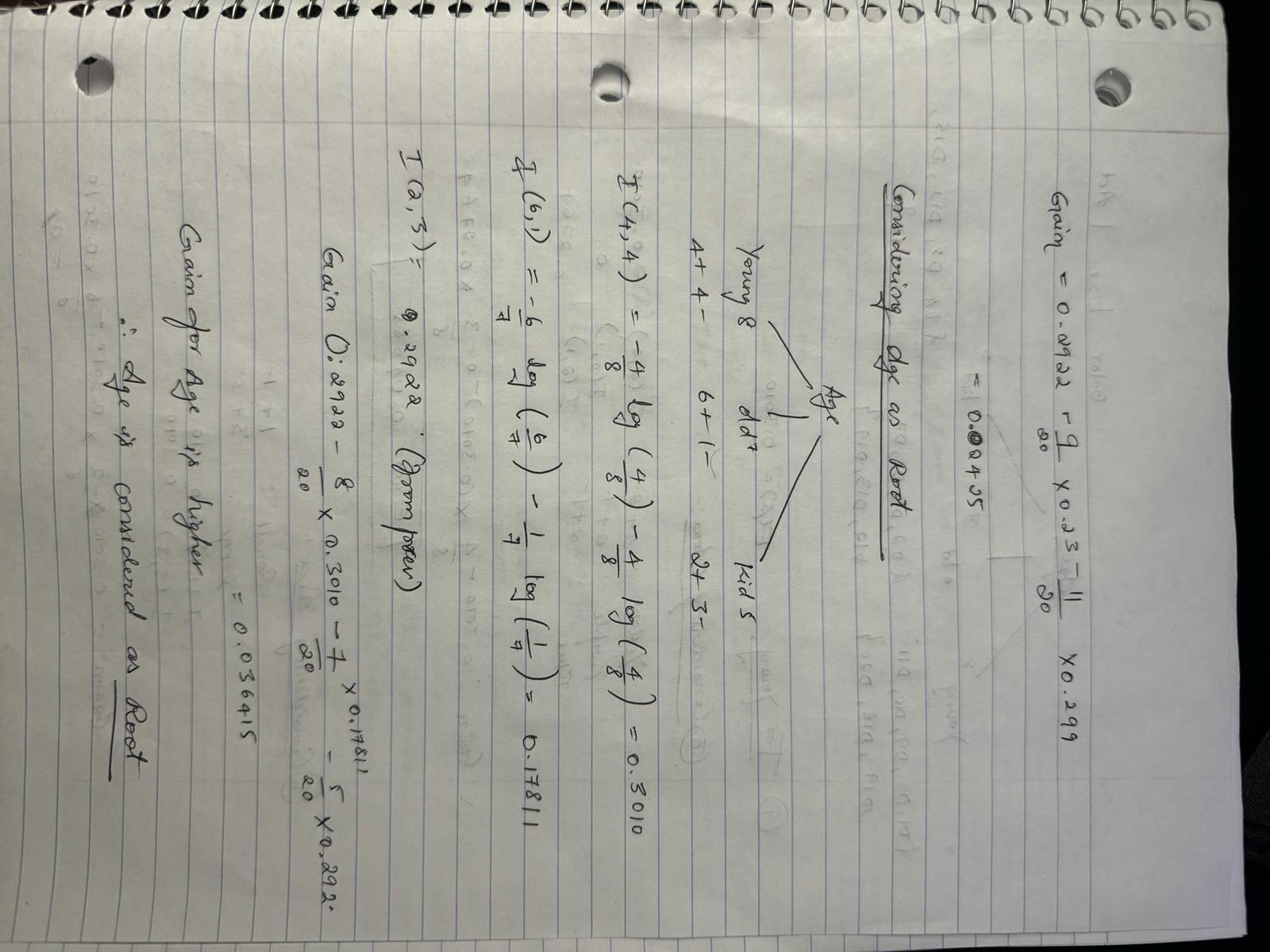
In this question, you need to build a decision tree, but only build the 1st (root node) and 2nd levels – in other words, you need to figure out the features to be filled in the first two levels, but you do not need to find the features to be filled in the 3rd level. On the 3rd level, you need to use leaf nodes as representations.

You should show the process and calculations about how to build the tree.

Note: if two variables have the same largest information gain, you should choose the one with less number of the values in the variable (i.e., fewer branches). For example, if you find that act and age have the same information gain value, and the value is the largest one, then you choose act instead of age since there are only two values in the variable act

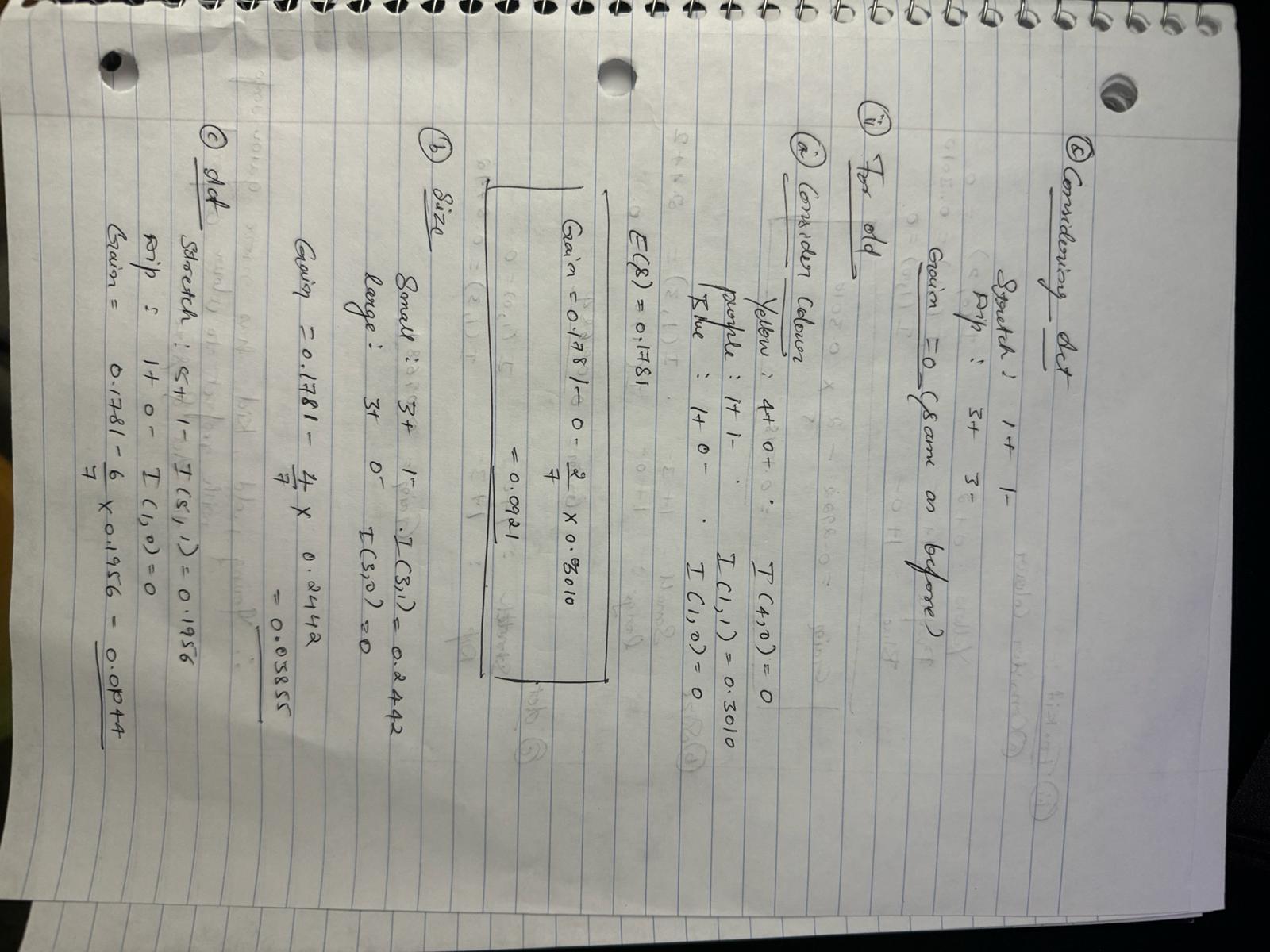


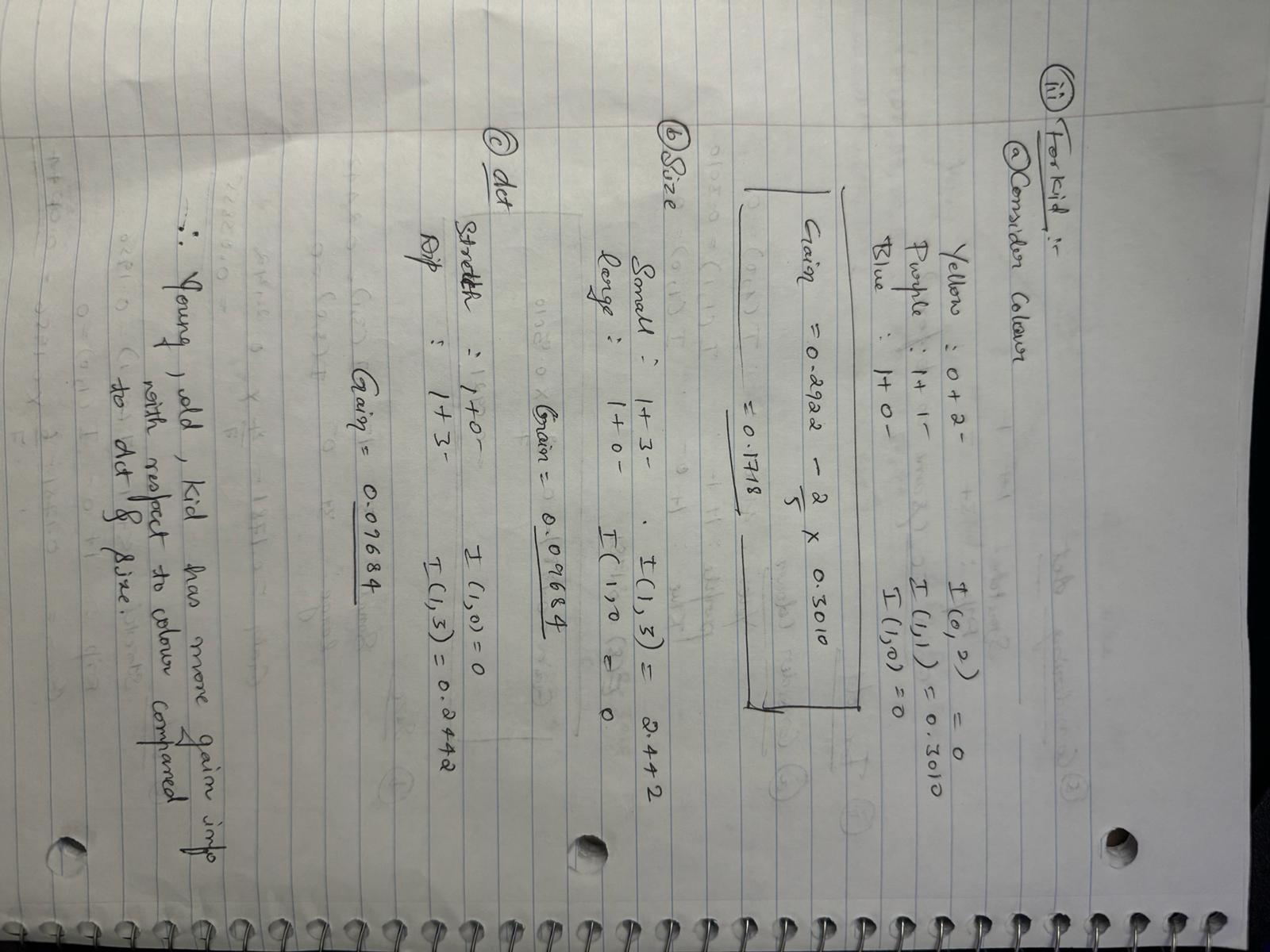


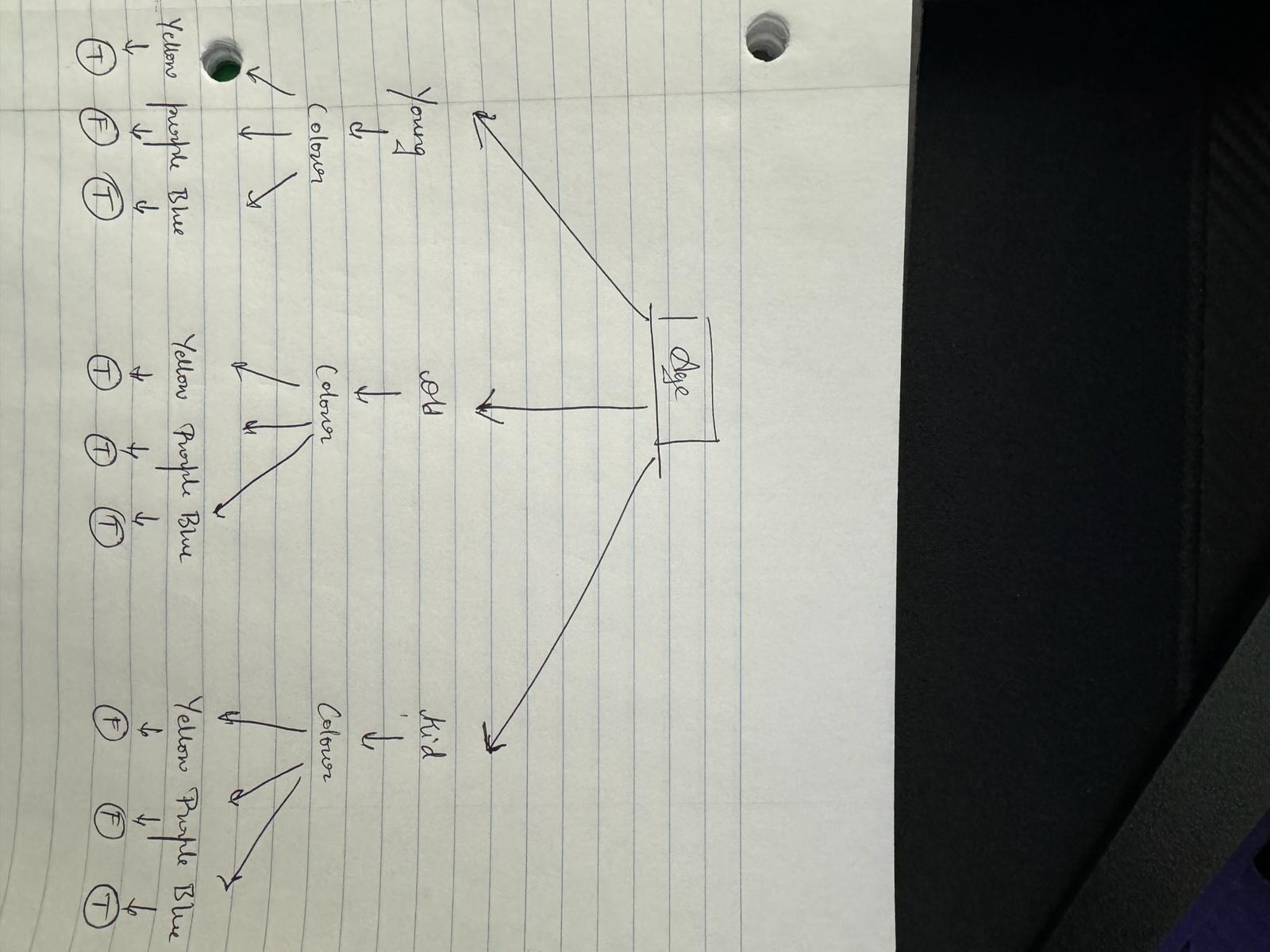


A notebook with writing on it

Description automatically generated



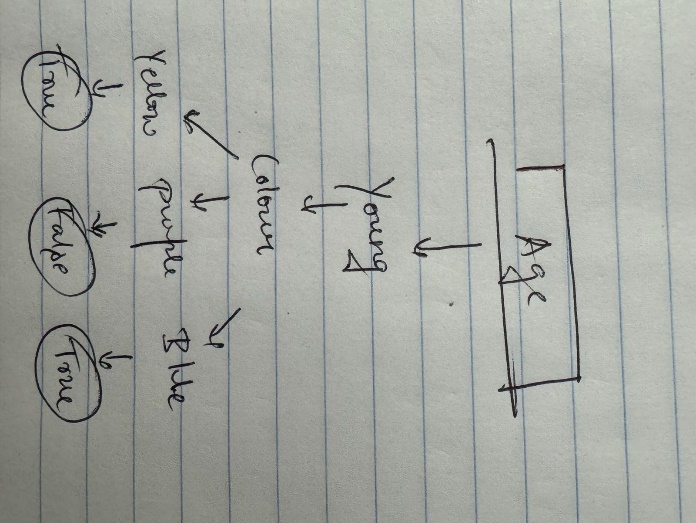




2). On the 3rd level of the tree, represent them by using leaf nodes (i.e., labels) on this level – you can simply use the majority of the label in each group as the predicted label. Given a new data “YELLOW, SMALL, DIP, YOUNG”, make your prediction [10]

A paper with lines drawn on it

Description automatically generated



From the above screenshot, we have inferred the conclusion (Note that: we have built a decision tree, but only the 1st (root node) and 2nd level while level 3 are leaves)

Given the case “YELLOW, SMALL, DIP, YOUNG”, we can say that majority is True based on the above logic.

From which we can conclude that the balloon may be inflated.

3). List at least three solutions to alleviate overfitting in decision tree [10]

Ans- 1. Using N -Cross-fold validations

2. Stop earlier: If we stop growing the tree earlier, we avoid overfitting the model to the training set

To avoid growing the tree we can use:

a. Classification metric (such as accuracy)

b. Minimum Description Length (MDL)

3. Post-prune: We can also allow overfitting and post-prune the tree later

Pruning increases classification accuracy by deleting branches and subtrees that are created because of noise.

**2. (50 points) Python practice for Decision Tree classification**

**Run decision tree techniques to find the best parameters and performance**

* Use Malware\_MultiClass.csv data by using 5-fold cross validation
* Build decision tree models

Note:

* You need to change different/multiple parameters to find the best model.
* You should try multiple models you have built, with the parameters you set
* Find the best model by using AUC score as the metric
* Finally, visualize the best tree model, Hint: you need to select a model from the 10-fold cross validation. See the example coding in the class.
* You can find data sets from “slide & data” on blackboard system

Submission

* The ipynb and saved html files