Bronco ID: |0|1|3|4|8|4|6|7|9|

Last Name: Francisco

First Name: Serrano

1. A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E (Mitchell, 1997). Explain this definition of a machine learning system informing in your answer how E, T, P correlate with each component of the image below.

   In each iteration, we study the problem and apply what we've learned to develop the next version of our model. The system is exposed to experience E where we are training the machine learning algorithm, informing it on the task T. During this stage, we use a specific set of data known as the training data to teach the model how to perform the task. To evaluate our solution, the model, we introduce it to new, unseen data—known as validation or test data. This is where the performance measure PP comes into play, quantifying how well the model has learned from E. If P shows that the model's performance is good enough, we launch/deploy. Otherwise, we might need to analyze potential errors in the current iteration before revisiting E by collecting more data or different types of data to improve the model's performance, to make P better. E, T, and P are linked, creating a feedback loop that allows for continuous improvement.

2. Some authors present a KDD/Data Mining pipeline process with only 3 main phases instead of those 6 shown in the image below (see the dashed arrows). Name those 3 main phases and explain their corresponding relevance to building knowledge.

Those three phases would be preprocessing, machine learning, and postprocessing.

The preprocessing phase encompasses the selection, processing, and transformation phases in the 6-step version. In this step we would be identifying potential sources of data we can use, whether that be a database, data warehouse, or some file from places like Kaggle. Once we determine the data we will be using, we need to analyze the structure/format of the data to convert it into something that is useable/preferrable for our algorithms. This might involve normalizing or standardizing data, converting from nominal categorical to ordinal, assigning identifiers to the record to track it, or it might mean converting continuous features into discrete representations(discretization). We need to account for missing values, make decisions on what features we are going to be using from the source— meaning we need to select our features and decide if some kind of simplification process like feature aggregation or dimension reduction is necessary.

The machine learning phase is going to cover to main parts, one is going to be training the actual algorithm on a dataset and another where we are going to validate the model using data separate from the training dataset to see how well it performs. This is where we are going to compute the performance measure.

Post-processing phase is where we interpret the results of the machine learning phase, preparing what you learned, and analyzing the results.

3. Machine learning algorithms face multiple challenges while analyzing data such as scalability, distribution, sparsity, resolution, class imbalance, noise, outliers, missing values, and duplicated data. For each image below, name and explain what the corresponding challenge is from this list (you do not need to explain how to solve the challenge).
    1. I believe the challenge is the data distribution, the data distribution that the machine algorithm was trained on is not how the input data's distribution is. The model might perform well on training data but poorly on other kinds of data.
    2. Outliers is the challenge; they can skew the data such that it makes some algorithms like clustering and regression models perform poorly since they depend on the distribution of the data.
    3. Missing values is the challenge in this diagram, there is a lot of values that are missing inside of most of the feature columns, but especially so in the last two. The challenge with this is that most algorithms assume that the data being passed to it is complete, so as developers we need to handle the missing values prior to passing it so that we get that fine grained control over how we treat the data and any assumptions we make in handling it.
    4. Noise because of how random the interruption of the data is, to me it looks like all the data points are there, but it looks like some is missing because of how random the disbursement is relative to how tightly packed the data points were prior to that. This is a challenge because it can make it harder for the model to understand the overall pattern.
    5. Here the challenge is going to be sparse data, because of the larger number of feature columns it is going to be hard to train the model on this dataset since we would need a larger number of records to compensate for it. It becomes harder for the model to generalize from high-dimensional feature space to the desired output. If the data is dissimilar to the training data, it will perform poorly.
4. Analyze the dataset below and answer the proposed questions:
    1. What is the most likely task that data scientists are trying to accomplish?
        1. They are trying to predict whether we should recommend contact lenses based on the four features, age, spectacle prescription, astigmatism, and tear production rate.
    2. In general, what is a feature, and how would you exemplify it with this data?
        1. A feature is one way we are observing the record, it is a property or characteristic of it. For example, Astigmatism is going to be one of the features we are using to describe each instance/record in the table. Each record/instance, which in this case is representing some person, has a corresponding feature-value for astigmatism which denotes whether that person (the record) has astigmatism or not.
    3. In general, what is a feature value, and how would you exemplify it with this data?
        1. A feature-value is going to be the value some record has for a given feature. Back to the astigmatism example, there are two possible feature-

        values : yes or no. Each record (or person in this instance) is being described along the lines of a set of features, astigmatism is one of these and each one will have a feature-value for it to denote whether it has it.

    4. In general, what is dimensionality, and how would you exemplify it with this data?
        1. Dimensionality is going to be the number of features you have describing an instance, in this case it would be 4 of them since an object is described by their age, spectacle prescription, astigmatism, and tear production rate.
    5. In general, what is an instance, and how would you exemplify it with this data?
        1. An instance is going to be one of the rows in the table that are described by the feature columns, in this case there are more than one. The first instance is the first row, they are young, have a myope prescription, they do not have astigmatism, and they reduced their tear production rate.
    6. In general, what is a class, and how would you exemplify it with this data?
        1. Assuming that the last column is indeed a target variable, then it's possible values serve as an example of the possible classes. It would mean that we are dealing with a binary classification, either recommend lenses or not. The two possible classes are yes and no.

5. Identify and explain what kind of machine learning (supervised, unsupervised, semi-supervised, reinforcement) system should be used for each scenario below including in your answer information about data labels. Hint: check the images to figure out which data sample is labelled.
    1. The data is labeled so we can use some kind of supervised machine learning in this scenario.
    2. Unsupervised learning since in image b they clustered and drew a circle around the two groups of data points, the data is not labeled.
    3. Some of the data points are labeled and others are not, so we can use semi-supervised learning.
6. Explain the tasks addressed by each classifier below. K and C must be present on your answer.

The binary classifier is going to be one where K=2, so there are two classes that we can classify the input data as. For binary classifiers C=1, so that means that any input value can only be mapped to one class (any item can only belong to one class, not more than one). The binary classifier is tasks with taking in input and predicting which of the two it belongs to.

Multiclass classifier is going to be one where K>2 and C=1, so you must place into at least 3 or more classes, but any input can still only belong to one class. This one is a little more complicated than a binary classifier since it is no longer simply a yes or no question.

On the other hand, a multi-label classifier where K>2 and C≥0 is going to be where you need at least 3 or more classes that you place into, but any given input can belong to any of the classes, none of the classes, or multiple classes.

7. Regarding the training data shown in question 4:

1. [20 points] Derive the decision tree produced by the standard ID3 algorithm. Show your calculations for entropy and information gain for all splits. Plot your final tree at the end

A1: ⑦ Derive the decision tree using ID3 algo

features: age, spectacle, perscription, tear prod rate
target var: reccommend lens

$P(\text{"yes"}) = 4/10 = 2/5$; $P(\text{"no"}) = 6/10 = 3/5$

$H(S) = -\left(\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{3}{5}\log_2\left(\frac{2}{5}\right)\right)$

$= -(-.97) = .9709$

root: → for age: age: {young, presbyopic, prepresbyopic}

for age == young :

young && no : $\frac{2}{4} = \frac{1}{2}$

young && yes : $\frac{2}{4} = \frac{1}{2}$

$H_{young} = -\left[\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right]$

$= -[-1] = 1$

for age == presbyopic

presbyopic && no : $\frac{2}{3}$

presbyopic && yes : $\frac{1}{3}$

$H_{presbyopic} = -\left[\frac{2}{3}\log_2\left(\frac{2}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right]$

$= 0.9183$

for age == prepresbyopic

prepresbyopic && no : $\frac{2}{3}$

prepresbyopic && yes : $\frac{1}{3}$

$H_{prepresbyopic} = 0.9183$ (same as last calculation)

$I.G.(S, Age) = .9709 - \left(\frac{4}{10}\cdot 1 + \frac{3}{10} * .9183 + \frac{3}{10} * .9183\right)$

$= 0.01942$

Shape (S)

for spectacle : spectacle = {myope, hypermetrope}

for spectacle == myope.

$\text{myope \&\& no}: \frac{4}{8} = \frac{1}{2}$

$\text{myope \&\& yes}: \frac{4}{8} = \frac{1}{2}$

$H_{myope} = 1$ (same # across both)

for spectacle == hypermetrope

$\text{hypermetrope \&\& no}: \frac{2}{2}$

$\text{hypermetrope \&\& yes}: \frac{0}{2}$

$$H_{hypermetrope} = -\left[\frac{2}{2}\log_2\left(\frac{2}{2}\right) + \overbrace{\frac{0}{2}\log_2\left(\frac{0}{2}\right)}^{0}\right]$$

$$= 0$$

$$IG(S, \text{spectacle}) = .9709 - \left(\frac{8}{10} \cdot 1 + \frac{2}{10} \cdot 0\right)$$

$$= .1709$$

for astigmatism = {yes, no}

for astigmatism == yes

$\text{yes \&\& no}: \frac{1}{4}$

$\text{yes \&\& yes}: \frac{3}{4}$

$$H_{yes} = -\left[\frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{3}{4}\log_2\left(\frac{3}{4}\right)\right] = 0.8112$$

for astigmatism == no

$\text{no \&\& no}: \frac{5}{6}$

$\text{no \&\& yes}: \frac{1}{6}$

$$H_{no} = -\left[\frac{5}{6}\log_2\left(\frac{5}{6}\right) + \frac{1}{6}\log_2\left(\frac{1}{6}\right)\right]$$

persunption $= 0.6500$

$$IG(S, \text{astigmatism}) = .9709 - \left(\frac{4}{10} \cdot .8112 + \frac{6}{10} \cdot .6500\right)$$

$$= 0.25592$$

for tear prod rate: {reduced, normal}

for tear rate == reduced

reduced && no: $\frac{5}{6}$

reduced && yes: $\frac{1}{6}$

$H_{reduced} = -[\frac{5}{6}\log_2(\frac{5}{6}) + \frac{5}{6}\log_2(\frac{1}{6})] = 0.6500$

for tear rate == normal

normal && no: $1/4$

normal && yes: $3/4$

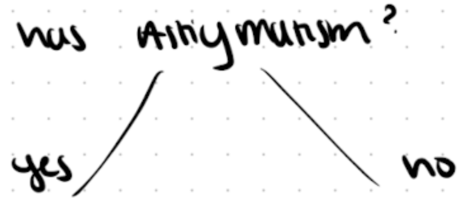
$H_{normal} = -[\frac{1}{4}\log_2(\frac{1}{4}) + \frac{3}{4}\log_2(\frac{3}{4})]$

$= .8112$

$IG(S, \text{tear rate}) = .9709 - (\frac{6}{10} \cdot 0.6500 + \frac{4}{10} * .8112)$

$= 0.25592$

since tear rate and astigmatism have same I.G., then I will randomly pick one (Astigmatism)



Astigmatism == yes $(\frac{4}{10})$

now choose betwen remaining 3:

$$\text{entropy} = -\left[\frac{3}{4}\log\left(\frac{3}{4}\right) + \frac{1}{4}\log\left(\frac{1}{4}\right)\right]$$

$$= 0.8112$$

for age:

for age == young:

young && no: $\frac{0}{2}$

young && yes: $\frac{2}{2}$

$$H_{young} = -\left[\frac{0}{2}\overbrace{\log_2\left(\frac{0}{2}\right)}^{0} + 1\overbrace{\log_2(1)}^{0}\right]$$

$$= 0$$

for age == presbyopic

presbyopic && no: $\frac{0}{1}$

presbyopic && yes: $\frac{1}{1}$

$$H_{presbyopic} = -\left[0\overbrace{\log_2(0)}^{0} + 1\overbrace{\log_2(1)}^{0}\right]$$

$$= 0$$

for age == prepresbyopic

prepresbyopic && no: $\frac{1}{1}$

prepresbyopic && yes: $\frac{0}{1}$

$$H_{prepresbyopic} = -\left[0\overbrace{\log_2(0)}^{0} + 1\overbrace{\log_2(1)}^{0}\right]$$

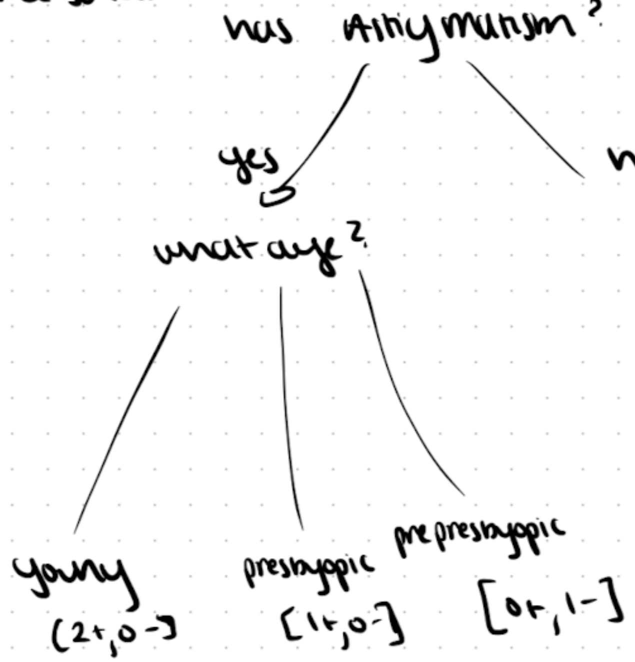
$$= 0$$

$$IG(S_{yes}, age) = 0.8112 - \left(\frac{2}{4}*0 + \frac{1}{4}*0 + \frac{1}{4}*0\right)$$

$$= 0.8112$$

since $IG(S_1, age) = H(S_1)$ then age is guaranteed to be the best feature (or it would be tied so we can pick any with that value)

tree so far:



Astigmatism == no:

$\frac{6}{10}$ of them are astig == no

for age:

$H(S_{no}) = -\left[\frac{5}{6}\log_2\left(\frac{5}{6}\right) + \frac{1}{6}\log_2\left(\frac{1}{6}\right)\right]$

$= 0.6500$

for age == young:

young && no: $\frac{2}{2}$

young && yes: $\frac{0}{2}$

$H_{young} = -\left[1\log_2(1) + 0\overbrace{\log_2(0)}^{0}\right]$

$= 0$

for age == presbyopic

presbyopic && no: $\frac{2}{2}$

presbyopic && yes: $\frac{0}{2}$

$H_{presbyopic} = -\left[1\log_2(1) + 0\log_2(0)\right]$

$= 0$

for age == prepresbyopic

prepresbyopic && no: $\frac{1}{2}$

prepresbyopic && yes: $\frac{1}{2}$

$H_{prepresbyopic} = -\left[\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right]$

$= -(-1) = 1$

$I.G.(S_{no}, age) = 0.6500 - \left(\frac{2}{6} * 0 + \frac{2}{6} * 0 + \frac{2}{6} * 1\right)$

$= 0.6500 - 0.3\overline{3} = 0.3167$

for spectacle : spectacle = {myope, hypermetrope}

spectacle == myope:

myope && no : $\frac{3}{4}$

myope && yes: $\frac{1}{4}$

$H_{myope} = -\left[\frac{3}{4}\log_2\left(\frac{3}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right]$

$= 0.8112$

spectacle == hypermetrope

hypermetrope && no : $\frac{2}{2}$

hypermetrope && yes: $\frac{0}{1}$

$H_{hypermetrope} = -[\overbrace{0\log_2(0)}^{0} + \overbrace{1\log_2(1)}^{0}]$

$= 0$

$IG(S, spectacle) = 0.6500 - \left(\frac{2}{6} * 0 + \frac{4}{6} * 0.8112\right)$

$= 0.1092$

for tear prod rate: {reduced, normal}

for tear rate == reduced

reduced && no: $\frac{4}{4}$

reduced && yes: $\frac{0}{4}$

$H_{reduced} = -[\overbrace{1\log_2(1)}^{0} + \overbrace{0\log_2(0)}^{0}] = 0$

for tear rate == normal

normal && no: $\frac{1}{2}$

normal && yes: $\frac{1}{2}$

$H_{normal} = -\left[\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right]$

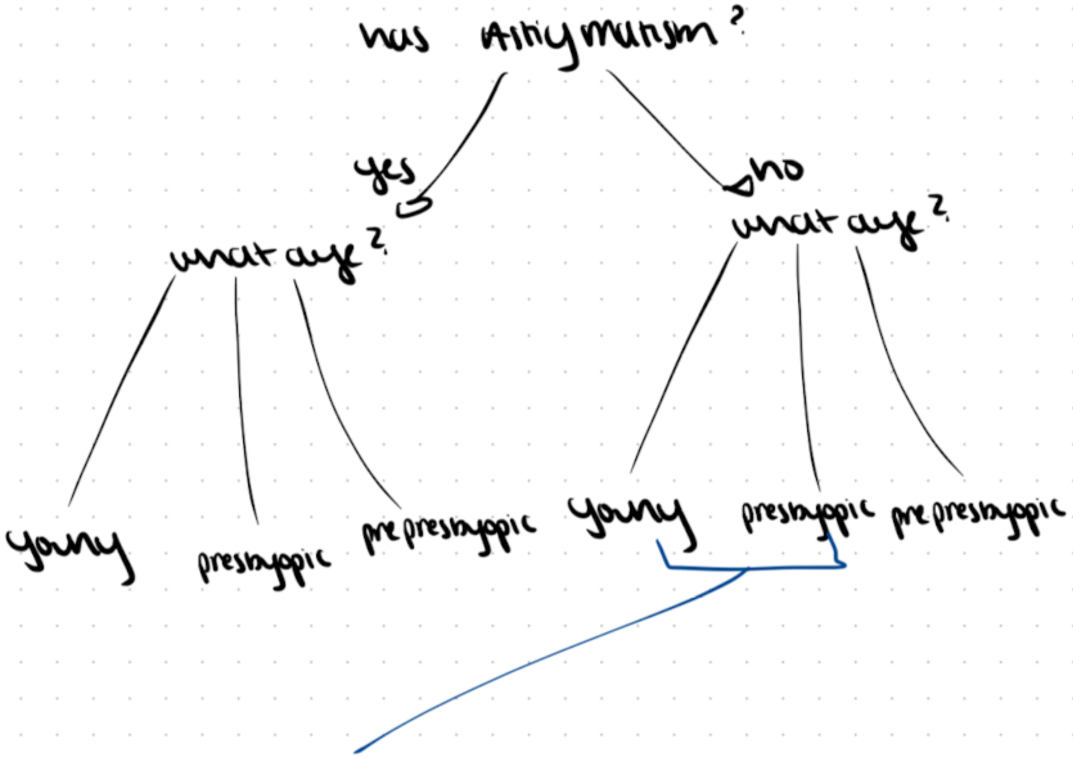
$= 1$

$IG(S, tear\ rate) = 0.6500 - \left(\frac{4}{6} * 0 + \frac{2}{6} * 1\right)$

$= 0.3167$

since $IG(S_{no}, \text{tear rate}) = IG(S_{no}, \text{age})$ we can pick any at random, I will select age

the tree now:



Since young and presbyopic both contain datapoints of a single class, they are considered pure enough. we now will do prepresbyopic

$$-[\tfrac{1}{2}\log_2(\tfrac{1}{2}) + \tfrac{1}{2}\log_2(\tfrac{1}{2})]$$

$S_{no,prepresbyopic} : \frac{2}{10} : H(S) = 1$

tear rate:

for tear rate == reduced

reduced && no : $\frac{1}{1}$

reduced && yes : $\frac{0}{1}$

$H_{reduced} = -[1 \log_2(1) + 0 \log_2(0)] = 0$

for tear rate == normal

normal && no : $\frac{0}{1}$

normal && yes : $\frac{1}{1}$

$H_{normal} = -[0 \log_2(0) + 1 \log_2(1)] = 0$

$IG(S, \text{tear rate}) = 1 - (\frac{1}{2} * 0 + \frac{1}{2} * 0)$

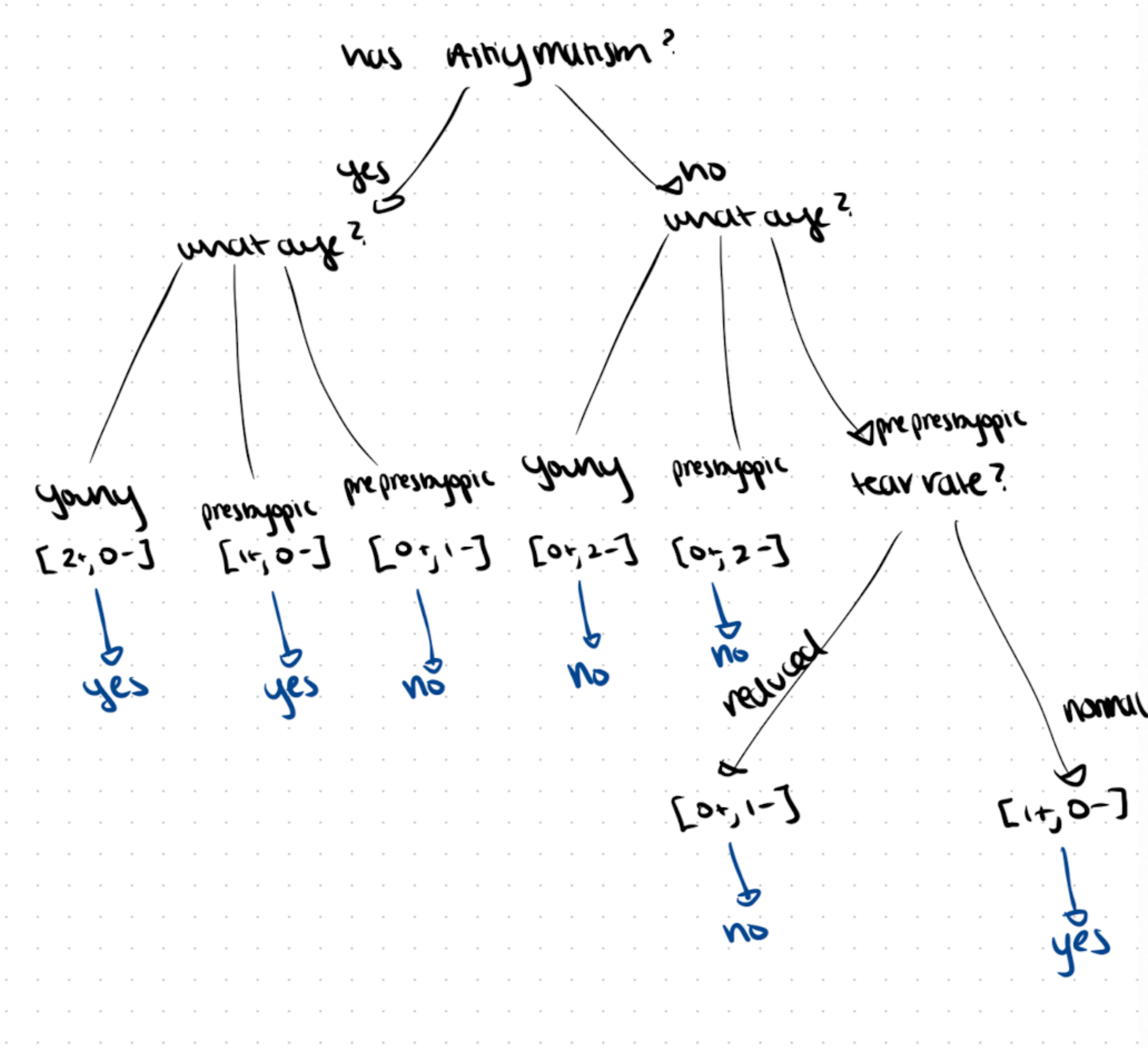
$= 1$

since $IG(S_{no,prepresbyopic}, \text{tear-rate}) = H(S_{no,prepresbyopic})$

then tear-rate has the largest possible information gain, if there is another feature with this then we can pick randomly. I select tear-rate.

Since tear-rate$_{reduced}$ and tear-rate$_{normal}$ both contain instances of one-kind, we can terminate here.

Here is my final tree:



2. [15 points] Complete the given python program (decision_tree.py) that will read the file contact_lens.csv and output a decision tree. Add the link to the online repository as the answer to this question.

https://github.com/franserr99/cs4210/tree/main/a1

note: that is going to be the directory containing all the files for this assignment.

3. [2 points] The tree you got in part b) should be the same one you got in part a), but there are probably some differences. Try to explain why.

Since the scikit library is dealing with what it things is a continuous input, instead of discrete categorical data, so when it is deciding on splits it will not know that each feature-value is mutually exclusive so it will/might make unnecessary splits based on intervals to lump together the subset falling within the internal. My graph did not have a split that they made, they created a more complicated tree because of the way we preprocessed the data.