Bronco ID: | 0 | 1 | 4 | 5 | 5 | 6 | 3 | 7 | 3 |
Last Name: Leos
First Name: Hugo

Assignment 1

1. Starting with “study the problem”, this component of the ML system requires understanding of the task **T**. During the “train ML algorithm” component we use training data to give the algorithm experience **E**. The more data, the more experience the algorithm gets. At the “evaluate solution” component, we use performance measure **P** to decide whether the performance has reached the desired threshold. If so, then the ML system is launched. If it is not satisfactory, then we must “analyze errors” by adjusting the algorithm without changing **E**, **T**, or **P** to maintain consistency.

Bronco ID: | 0 | 1 | 4 | 5 | 5 | 6 | 3 | 7 | 3 |
Last Name: Leos
First Name: Hugo

2. The three main phases are:

    a. Preprocessing – is where ~70% of the time is spent since having the best possible data will give the best results/knowledge. If no preprocessing is done, then the results will be garbage.

    b. Machine Learning/Data Mining – is where the algorithm processes the data and can give out results which will be used to gain knowledge.

    c. Postprocessing – is where we use the results from the ML algorithm and interpret or infer knowledge from them.

Bronco ID: | 0 | 1 | 4 | 5 | 5 | 6 | 3 | 7 | 3 |
Last Name: Leos
First Name: Hugo

3.

a. This is the challenge of data distribution since the training data and test data are not evenly distributed.

b. This is the challenge of outliers since one of the data points is far from the rest.

c. This is the challenge of missing values because many instances do not have input for one or more features.

d. This is the challenge of noise since the data appears to have interference noted by the data points do not make sense.

e. This is the challenge of dimensionality/sparsity because the features' values range from 0-5 and there are 5 features per instance giving a total possible combination of instances $6^5 = 7776$. However, there are only 6 instances which will leave out a lot of possible combinations.

Bronco ID: | 0 | 1 | 4 | 5 | 5 | 6 | 3 | 7 | 3 |
Last Name: Leos
First Name: Hugo

4.

a. They most likely are trying to find which contact lenses should be recommended to a user based on age, spectacle prescription, astigmatism and tear production rate.

b. A feature is a characteristic of an object or instance. This can be seen in the age feature which characterizes whether that instance is young, presbyopic, or prepresbyopic person.

c. A feature value is the value that the feature has/can have. As seen in (b), the feature values are young, presbyopic, or prepresbyopic

d. Dimensionality is usually the number of features. Therefore, the data has a dimensionality of 4 because the four features are age, spectacle prescription, astigmatism, and tear production rate. While the "recommended lenses" is the label.

e. An instance or sample is a data entry that contains feature values for the features. Thus, this is an instance from the data:

Young	Myope	No	Reduced	No

f. A class is the label that the instance has or is given. For example, the instance from (e) has a class of "no".

Bronco ID: | 0 | 1 | 4 | 5 | 5 | 6 | 3 | 7 | 3 |
Last Name: Leos
First Name: Hugo

5.

a. This is supervised learning since the goal of supervised learning is to map the inputs and label them by learning from a labeled data set.

b. This is unsupervised learning since the goal of unsupervised learning is to group or cluster the data based on their features using an unlabeled data set.

c. This is semi-supervised learning since the goal of semi-supervised learning is to map the inputs and label them by learning from a partially labeled data set.

Bronco ID: | 0 | 1 | 4 | 5 | 5 | 6 | 3 | 7 | 3 |
Last Name: Leos
First Name: Hugo

6. The binary classifier addresses tasks that have two classes and only one label. Such as determining if an object is a bird since it is either a bird or it is not a bird. The Multiclass classifier is like binary classifier but has more than just two classes. For example, having apples, bananas, and oranges as classes, we determine if an object is either one of those three options thus only having one label. The multilabel classifier, on the other hand, has more than two classes and allows for more than just one label. This can be exemplified in an article about religion in politics for the United States. The article can then be labeled as being a religious article, political article, and an article about the United States.

Bronco ID: | 0 | 1 | 4 | 5 | 5 | 6 | 3 | 7 | 3 |
Last Name: Leos
First Name: Hugo

7.

a.

$\text{Entropy}(S) = -\ 4/10 \log_2(4/10) - 6/10 \log_2(6/10)$
$= -\ (-0.52877) - (-0.44217)$
$= 0.971$

$\text{Gain}(S, \text{Age}) = 0.971 - 4/10\ \text{Entropy}(S_{young}) - 3/10\ \text{Entropy}(S_{prebyopic}) - 3/10\ \text{Entropy}(S_{preprebyopic})$
$= 0.971 - 4/10\ [-\ 2/4 \log_2(2/4) - 2/4 \log_2(2/4)] - 3/10\ [-\ 1/3 \log_2(1/3) - 2/3 \log_2(2/3)] - 3/10\ [-\ 1/3 \log_2(1/3) - 2/3 \log_2(2/3)]$
$= 0.971 - 4/10\ [\ 1\ ] - 3/10\ [\ -\ (-0.52832) - (-\ 0.38997)] - 3/10\ [\ -\ (-0.52832) - (-\ 0.38997)]$
$= 0.971 - 0.4 - 0.27548 - 0.27548$
$= 0.02$

$\text{Gain}(S, \text{SP}) = 0.971 - 8/10\ \text{Entropy}(S_{myope}) - 2/10\ \text{Entropy}(S_{hypermetrope})$
$= 0.971 - 8/10\ [\ 1\ ] - 2/8\ [\ 0\ ]$
$= 0.971 - 0.8$
$= 0.171$

$\text{Gain}(S, \text{Astigmatism}) = 0.971 - 4/10\ \text{Entropy}(S_{yes}) - 6/10\ \text{Entropy}(S_{no})$
$= 0.971 - 4/10\ [-\ 3/4 \log_2(3/4) - 1/4 \log_2(1/4)] - 6/10\ [-\ 1/6 \log_2(1/6) - 5/6 \log_2(5/6)]$
$= 0.971 - 4/10\ [-\ (-0.31127) - (-0.5)] - 6/10\ [-\ (-0.43082) - (-0.21919)]$
$= 0.971 - 0.32451 - 0.39$
$= 0.256$

$\text{Gain}(S, \text{TPR}) = 0.971 - 6/10\ \text{Entropy}(S_{reduced}) - 4/10\ \text{Entropy}(S_{normal})$
$= 0.971 - 6/10\ [-\ 1/6 \log_2(1/6) - 5/6 \log_2(5/6)] - 4/10\ [-\ 3/4 \log_2(3/4) - 1/4 \log_2(1/4)]$
$= 0.971 - 6/10\ [-\ (-0.43082) - (-0.21919)] - 4/10\ [-\ (-0.31127) - (-0.5)]$
$= 0.971 - 0.39 - 0.32451$
$= 0.256$

$\text{Entropy}(S_{yes}) = -\ 3/4 \log_2(3/4) - 1/4 \log_2(1/4)$
$= -\ (-0.31127) - (-0.5)$
$= 0.81127$

$\text{Gain}(S_{yes}, \text{Age}) = 0.81127 - 2/4\ \text{Entropy}(S_{young}) - 1/4\ \text{Entropy}(S_{prebyopic}) - 1/4\ \text{Entropy}(S_{preprebyopic})$
$= 0.81127 - 2/4\ [-2/2 \log_2(2/2) - 0/2 \log_2(0/2)] - 1/4\ [-1/1 \log_2(1/1) - 0/1 \log_2(0/1)] - 1/4\ [-0/2 \log_2(0/2) - 1/2 \log_2(1/2)]$
$= 0.81127 - 2/4\ [\ 0\ ] - 1/4\ [\ 0\ ] - 1/4\ [\ 0\ ]$
$= 0.81127$

$\text{Gain}(S_{yes}, \text{SP}) = 0.81127 - 4/4\ \text{Entropy}(S_{myope}) - 0/4\ \text{Entropy}(S_{hypermetrope})$
$= 0.81127 - [-3/4 \log_2(3/4) - 1/4 \log_2(1/4)] - 0$
$= 0.81127 - 0.81127$
$= 0$

$\text{Gain}(S_{yes}, \text{TPR}) = 0.81127 - 2/4\ \text{Entropy}(S_{normal}) - 2/4\ \text{Entropy}(S_{reduced})$
$= 0.81127 - 2/4\ [-\ 2/2 \log_2(2/2) - 0/2 \log_2(0/2)] - 2/4\ [-\ 1/2 \log_2(1/2) - 1/2 \log_2(1/2)$
$= 0.81127 - 2/4\ [\ 0\ ] - 2/4\ [\ 1\ ]$
$= 0.31127$

Bronco ID: | 0 | 1 | 4 | 5 | 5 | 6 | 3 | 7 | 3 |
Last Name: Leos
First Name: Hugo

$$\begin{aligned}\text{Entropy}(S_{no}) &= -\,5/6\log_2(5/6) - 1/6\log_2(1/6)\\ &= -\,(-0.21919) - (-0.43082)\\ &= 0.65\end{aligned}$$

$$\begin{aligned}\text{Gain}(S_{no},\text{Age}) &= 0.65 - 2/6\,\text{Entropy}(S_{young}) - 2/6\,\text{Entropy}(S_{prebyopic}) - 2/6\,\text{Entropy}(S_{preprebyopic})\\ &= 0.65 - 2/6\,[-\,0/2\log_2(0/2) - 2/2\log_2(2/2)] - 2/6\,[-\,0/2\log_2(0/2) - 2/2\log_2(2/2)] - 2/6\,[-1/2\log_2(1/2) - 1/2\log_2(1/2)]\\ &= 0.65 - 2/6\,[\,0\,] - 2/6\,[\,0\,] - 2/6\,[\,1\,]\\ &= 0.31666\end{aligned}$$

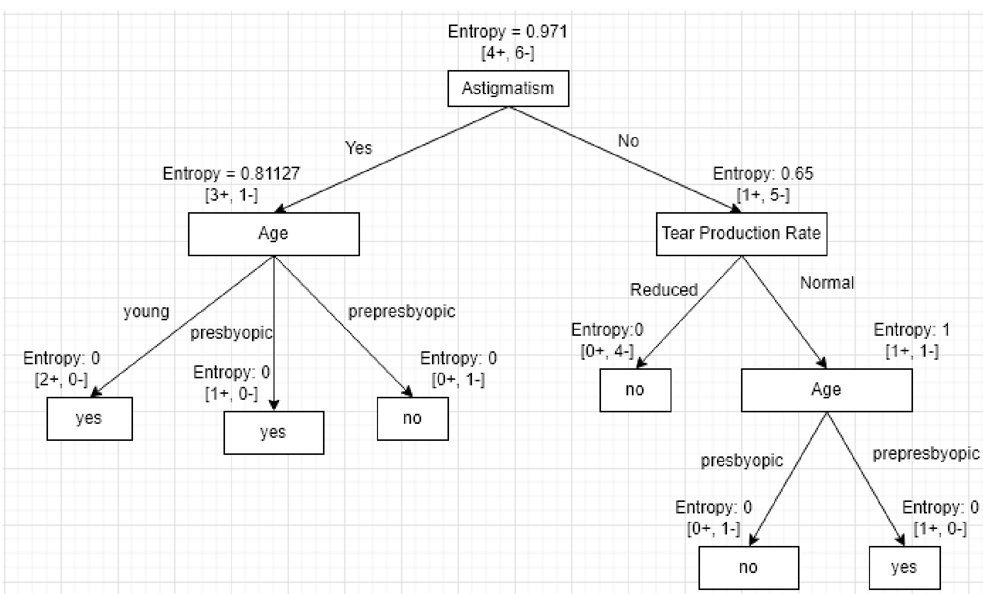
$$\begin{aligned}\text{Gain}(S_{no},\text{SP}) &= 0.65 - 4/6\,\text{Entropy}(S_{myope}) - 2/6\,\text{Entropy}(S_{hypermetrope})\\ &= 0.65 - 4/6\,[-\,1/4\log_2(1/4) - 3/4\log_2(3/4)] - 2/6\,[-\,0/2\log_2(0/2) - 2/2\log_2(2/2)]\\ &= 0.65 - 4/6\,[\,0.81127\,] - 2/6\,[\,0\,]\\ &= 0.1092\end{aligned}$$

$$\begin{aligned}\text{Gain}(S_{no},\text{TPR}) &= 0.65 - 2/6\,\text{Entropy}(S_{normal}) - 4/6\,\text{Entropy}(S_{reduced})\\ &= 0.65 - 2/6\,[-\,1/2\log_2(1/2) - 1/2\log_2(1/2)] - 4/6\,[-\,0/4\log_2(0/4) - 4/4\log_2(4/4)]\\ &= 0.65 - 2/6\,[\,1\,] - 4/6\,[\,0\,]\\ &= 0.31666\end{aligned}$$

$$\begin{aligned}\text{Entropy}(S_{normal}) &= -\,1/2\log_2(1/2) - 1/2\log_2(1/2)\\ &= 1\end{aligned}$$

$$\begin{aligned}\text{Gain}(S_{normal},\text{Age}) &= 1 - 0/2\,\text{Entropy}(S_{young}) - 1/2\,\text{Entropy}(S_{prebyopic}) - 1/2\,\text{Entropy}(S_{preprebyopic})\\ &= 1 - 0 - 1/2\,[-\,0/1\log_2(0/1) - 1/1\log_2(1/1)] - 1/2\,[-1/1\log_2(1/1) - 0/1\log_2(0/1)]\\ &= 1 - 1/2\,[\,0\,] - 1/2\,[\,0\,]\\ &= 1\end{aligned}$$

$$\begin{aligned}\text{Gain}(S_{normal},\text{SP}) &= 1 - 2/2\,\text{Entropy}(S_{myope}) - 0/2\,\text{Entropy}(S_{hypermetrope})\\ &= 1 - 2/2\,[-\,1/2\log_2(1/2) - 1/2\log_2(1/2)] - 0\\ &= 1 - 2/2\,[\,1\,]\\ &= 0\end{aligned}$$



Bronco ID: | 0 | 1 | 4 | 5 | 5 | 6 | 3 | 7 | 3 |
Last Name: Leos
First Name: Hugo

b. Github link to py file

c. A possible reason that there are differences is because there is never just one decision tree as a solution. Especially, when the root could be more than one since the gain was the same for two features. Also, the method of determining the root or next feature can affect the outcome of the decision tree.