CS 4210
Assignment 1
Brandon Tang

1)
The main purpose of using machine learning is to make computers learn from data without being explicitly programmed. The definition given contrasts with the diagram shown because maintaining and reanalyzing problems becomes very complex and hard to maintain. This is particularly bad when it comes to data with higher dimensionality, where detecting patterns and updating the list of rules is very difficult.

2)
The 3 main phases are preprocessing, machine learning, and postprocessing. Preprocessing ensures that raw data is properly prepared for analysis by handling null values, noise, and inconsistencies. Preprocessing is vital because poor quality data will lead to poor quality results. Machine learning is used to extract meaningful patterns, trends, and relationships from the data. The choice of machine learning techniques used will determine the quality and usefulness of the knowledge gained. Postprocessing involves filtering patterns, visualization, and interpretations. The discovered patterns are assessed for their usefulness, novelty, and accuracy. Irrelevant or misleading patterns are filtered out, and the results are visualized or structured in a way that enables decision-making. The final knowledge must be interpretable and actionable for stakeholders.

3a)
The challenge in the image is data distribution. The graph is showing the input density for training and testing data. Since the graph is displaying the density distributions between training and testing, it could lead to faulty results when running a model.

b)
Outliers are the challenge. There is a single point that is completely far off from the rest of the data inputs. Outliers are a challenge because they are threats to misleading results.

c)
Missing values is the challenge in this scenario. Many of the cells either have irrelevant characters, random numbers, or are empty. This significantly degrades the quality of the data.

d)
The random points floating in the mid area describe noise. Noise is the unpredictable fluctuations that make it difficult to identify patterns and relationships.

e)
The dataset shown is considered to be sparse. Sparsity is a machine learning challenge where a dataset or matrix contains many zero values.

4a)
Data scientists are most likely going to predict whether a patient needs contact lenses using features in the dataset as determining factors.

b)
A feature would be considered a column label in the dataset. For example, “Age”, “Spectacle Prescription”, and “Astigmatism” would be dataset features.

c)
Feature value would be an individual value in one of the cells in the dataset. Examples of this would be “Reduced”, “Myope”, and “Yes”.

d) Dimensionality refers to the number of features in a dataset. Since the dataset has 5 features, it’s considered to have low dimensionality.

e) An instance is a row entry in a dataset. An example of this in the dataset would be the first row entry at the top of the dataset.

f) A class is a collection of data that are related to each other in some way. The dataset can demonstrate a class by having features that are similar to each other that can be used to predict a result in machine learning.

5a)
Supervised learning would be best used in the scenario. In this scenario, the main task is to classify and label the X’s and O’s. The input and output are fully labeled so we know the desired result is achieved.

b)

Semi-supervised learning would be used in this scenario because a small portion of labels in each class are classified incorrectly. We can use the labeled data to train a model and then leverage the large pool of unlabeled data to further improve the model's ability to be better.

c)
Unsupervised learning would be needed in this scenario. Assuming the data in the image is unlabeled, we would need to utilize unsupervised learning to find out if there are any patterns in the data to then classify them into clusters.

6)
The task for the binary classifier is to properly address problems where there are exactly two mutually exclusive outcomes. An example of this could be spam or not spam.

The task for the multiclass classifier handles problems where inputs need to be classified into exactly one of several classes.

The task for the multilabel classifier is to label data that could belong to multiple classes. This could involve tagging an image with the labels, “car” “vehicle” “transportation”.

7a)
Yes = 4
No = 6
Total = 10
Initial Entropy = $- \{ [(4/10) \log_2 (4/10)] + [(6/10) \log_2 (6/10)] \} = 0.971$

Attribute: Age
Values(Age) = young, presbyopic, prepresbyopic
S_{Young} = [2+, 2-]
$S_{Presbyopic}$ = [1+, 2-]
$S_{Prepresbyopic}$ = [1+, 2-]

Entropy(S_{Young}) = 1
Entropy($S_{Presbyopic}$) = 0.918
Entropy($S_{Prepresbyopic}$) = 0.918
Gain(S, Age) = Entropy(S) - $\sum$ ($|S_v| / |S|$) Entropy(S_v)
Gain(S, Age) = Entropy(S) - 4/10 Entropy(S_{Young}) - 3/10 Entropy($S_{Presbyopic}$) - 3/10 Entropy($S_{Prepresbyopic}$)

= 0.971 - (0.4)(1) - (0.3)(0.918) - (0.3)(0.918) = 0.0202

Attribute: Spectacle Prescription
Values(SP) = myope, hypermetrope
S_{Myope} = [4+, 4-]
$S_{Hypermetrope}$ = [0+, 2-]

Entropy(S_{Myope}) = 1
Entropy($S_{Hypermetrope}$) = 0

Gain(S, SP) = 0.971 - (0.8)(1) - (0.2)(0) = 0.171

Attribute: Astigmatism
Values(Astigmatism) = yes, no
S_{Yes} = [3+, 1-]
S_{No} = [1+, 5-]

Entropy(S_{yes}) = - { [(3/4) $\log_2$ (3/4)] + [(1/4) $\log_2$ (1/4)] } = 0.811
Entropy(S_{no}) = - { [(1/6) $\log_2$ (1/6)] + [(5/6) $\log_2$ (5/6)] } = 0.650

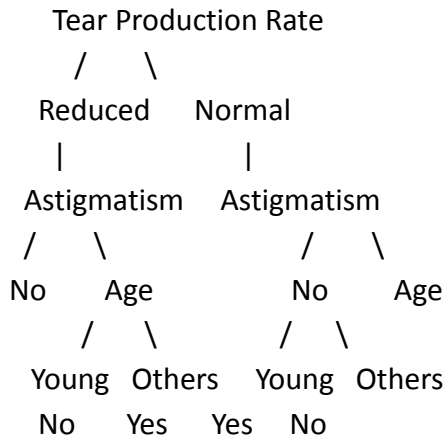
Gain(S, Astigmatism) = 0.971 - (0.4)(0.811) - (0.6)(0.650) = 0.257

Attribute: Tear Production Rate
Values(TPR) = normal, reduced
S_{Normal} = [3+, 1-]
$S_{Reduced}$ = [1+, 5-]

Entropy(S_{normal}) = - { [(3/4) $\log_2$ (3/4)] + [(1/4) $\log_2$ (1/4)] } = 0.811
Entropy($S_{reduced}$) = - { [(1/6) $\log_2$ (1/6)] + [(5/6) $\log_2$ (5/6)] } = 0.650

Gain(S, TPR) = 0.971 - (0.4)(0.811) - (0.6)(0.650) = 0.257

Gain(S, Age) = 0.0202
Gain(S, SP) = 0.171
Gain(S, Astigmatism) = 0.257
Gain(S, TPR) = 0.257



b)
https://github.com/BrandonTang95/cs4210.git
(inside assignment1 folder)

c)
The decision tree might be different due to reasons such as:
#1: tie breaking rules in information gain selection
#2: slight floating point differences
#3: automatic optimizations built in sklearn
#4: different strategies for resolving identical entropy values
Both trees should be very similar though.