CS 4210 – Assignment #1
Maximum Points: 100 pts.

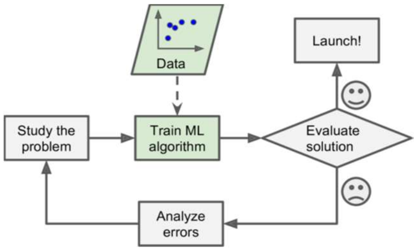
Bronco ID: 015412449

Last Name: Langgeng

First Name: Leonardo

Note 1: Your submission header must have the format as shown in the above-enclosed rounded rectangle.
Note 2: Homework is to be done individually. You may discuss the homework problems with your fellow students, but you are NOT allowed to copy – either in part or in whole – anyone else's answers.
Note 3: Your deliverable should be a .pdf file submitted through Gradescope until the deadline. Do not forget to assign a page to each of your answers when making a submission. In addition, source code (.py files) should be added to an online repository (e.g., github) to be downloaded and executed later.
Note 4: All submitted materials must be legible. Figures/diagrams must have good quality.
Note 5: Please use and check the Canvas discussion for further instructions, questions, answers, and hints. The bold words/sentences provide information for a complete or accurate answer.

1. [6 points] 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.



Experience are correlated with the data being fed to the ML algorithm. The more data being fed to the ML for training, the more experience the machine will learn from, and assuming there is no problem with the input data (bias, noise, etc), the machine will become more accurate.

Task are correlated with the objective of the ML. The machine will try to predict the output from a given input, based on the experience from training.

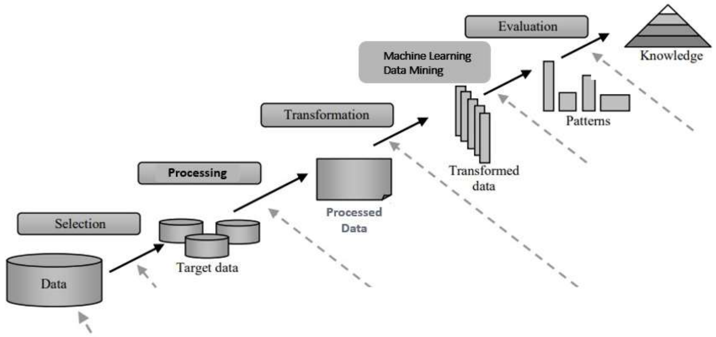
Performance are the accuracy of the machine's prediction. It either could be a % of error prediction.

2. [6 points] Some authors present a machine learning/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 build knowledge.

Preprocessing: preparing the dataset to be processed. This phase cleans up the dataset and making sure that the data are as optimal as possible thus making the ML process to be accurate with.

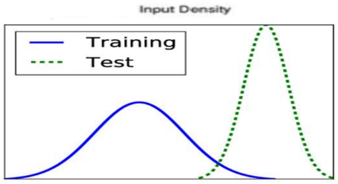
Machine learning: process where the machine will analyze the data using certain algorithm and outputs certain patterns to be analyzed later.

Postprocessing: Analyzing and interpret the information produced by ML

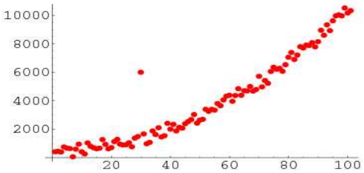


3. [15 points – 3 points each] Machine learning algorithms face multiple challenges while analyzing data such as scalability, distribution, sparsity, resolution, class imbalance, noise, outliers, missing values, and duplicated data. For **each** image below, **name** and **explain** what the corresponding challenge is from this list (you do not need to explain how to solve the challenge).

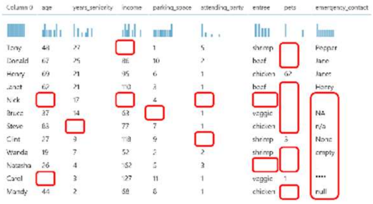
a.



b.

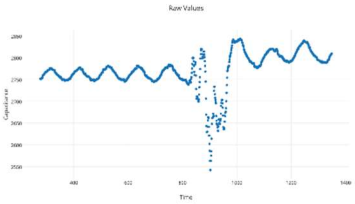


c.

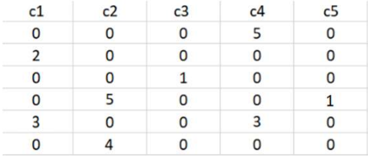


Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick		17		4				
Bruce	37	14	63		1	veggie		NA
Steve	63		77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	152	5	3			
Carol		3	127	11	1	veggie	1	••••
Mandy	44	2	58	8	1	chicken		null

d.



e.



c1	c2	c3	c4	c5
0	0	0	5	0
2	0	0	0	0
0	0	1	0	0
0	5	0	0	1
3	0	0	3	0
0	4	0	0	0

A. Biased training/test sets. If the training and testing sets are not balanced/equal, the prediction from it will be biased toward the training set, not the expected result from test set.

B. Outliers. Depending on algorithm used, it may throw off the model produced by the machine as outliers do not line up with the rest of the data points. Outliers are real data point, thus it should still be considered/included in the dataset.

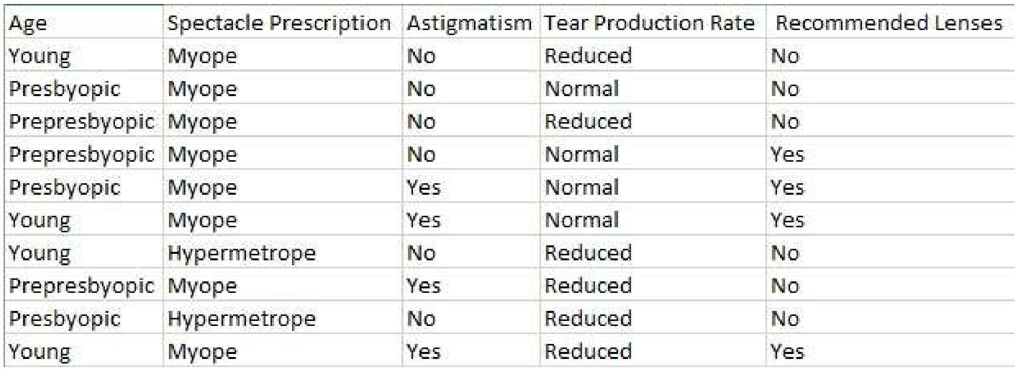
C. Missing values. It makes the machine harder to plot the data points with missing values, and in some/most cases we need to replace the missing values.

D. Noise. It puts a lot of inaccurate data points and thus will throw off machines prediction.

E. Sparse data. If the data is sparse, it is very hard for the machine to find all of the possible combinations of the features to build the model.

4. [18 points – 3 points each] Analyze the dataset below and answer the proposed questions:

The Contact Lens Data



Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
Young	Myope	No	Reduced	No
Presbyopic	Myope	No	Normal	No
Prepresbyopic	Myope	No	Reduced	No
Prepresbyopic	Myope	No	Normal	Yes
Presbyopic	Myope	Yes	Normal	Yes
Young	Myope	Yes	Normal	Yes
Young	Hypermetrope	No	Reduced	No
Prepresbyopic	Myope	Yes	Reduced	No
Presbyopic	Hypermetrope	No	Reduced	No
Young	Myope	Yes	Reduced	Yes

a. What is the most likely task that data scientists are trying to accomplish?
To predict whether or not we recommend lenses for certain people/patient based on certain information

b. **In general**, what is a feature and how would you **exemplify** it with **this data**?
Feature are the known trait from each data point. In this case, it is age, spectacle prescription, astigmatism, and tear rate

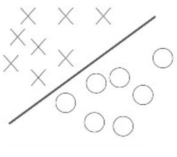
c. **In general**, what is a feature value and how would you **exemplify** it with **this data**?
The value of each feature of each data point. Age could be young, presbyopic, and prepresbyopic

d. **In general**, what is dimensionality and how would you **exemplify** it with **this data**?
The number of features from a dataset. The table have 4 datasets as we using recommended lenses as class/result

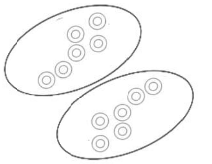
e. **In general**, what is an instance and how would you **exemplify** it with **this data**?
Instance are the data sample. The table have 10 instance (10 people/patient)

f. **In general**, what is a class and how would you **exemplify** it with **this data**?
Class is the expected result from each instance. In this case, it is whether or not we recommend lenses to the instance/patient

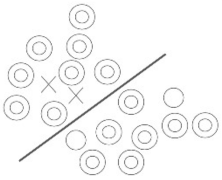
5. [9 points] Identify and explain what **kind of machine learning** (supervised, unsupervised, semi-supervised, reinforcement) **system** should be used for each scenario below including in your answer information about **data labels**. Hint: check the images to figure out which data sample is labelled.



a)



b)



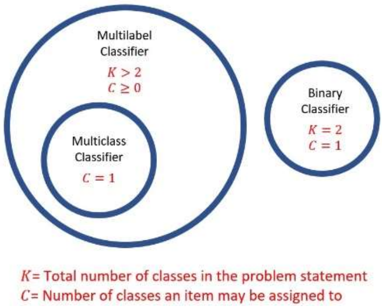
c)

A. Supervised learning. It is a classification tasks, and we try to predict the cutoff line / the divider between 2 or more class.

B. Unsupervised learning. The image trying to find interesting patterns or grouping in the data points.

C. Semi supervised learning. It is trying to map the data of a lot of unlabeled data points with a bit of labeled points.

6. [9 points] Explain the **tasks** addressed by each classifier below.



K= Total number of classes in the problem statement
C= Number of classes an item may be assigned to

Binary classifier guessing single output from 2 possible class (binary output; yes or no)

Multilabel classifier allows multiple class output from multiple possible class (result could be a percentage of each possible class)

Multiclass classifier produce 1 output class from multiple possible class (there is n# options, chose 1)

7. [37 points] Regarding the training data shown in question 4:

   a. [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.
   Excel attachment

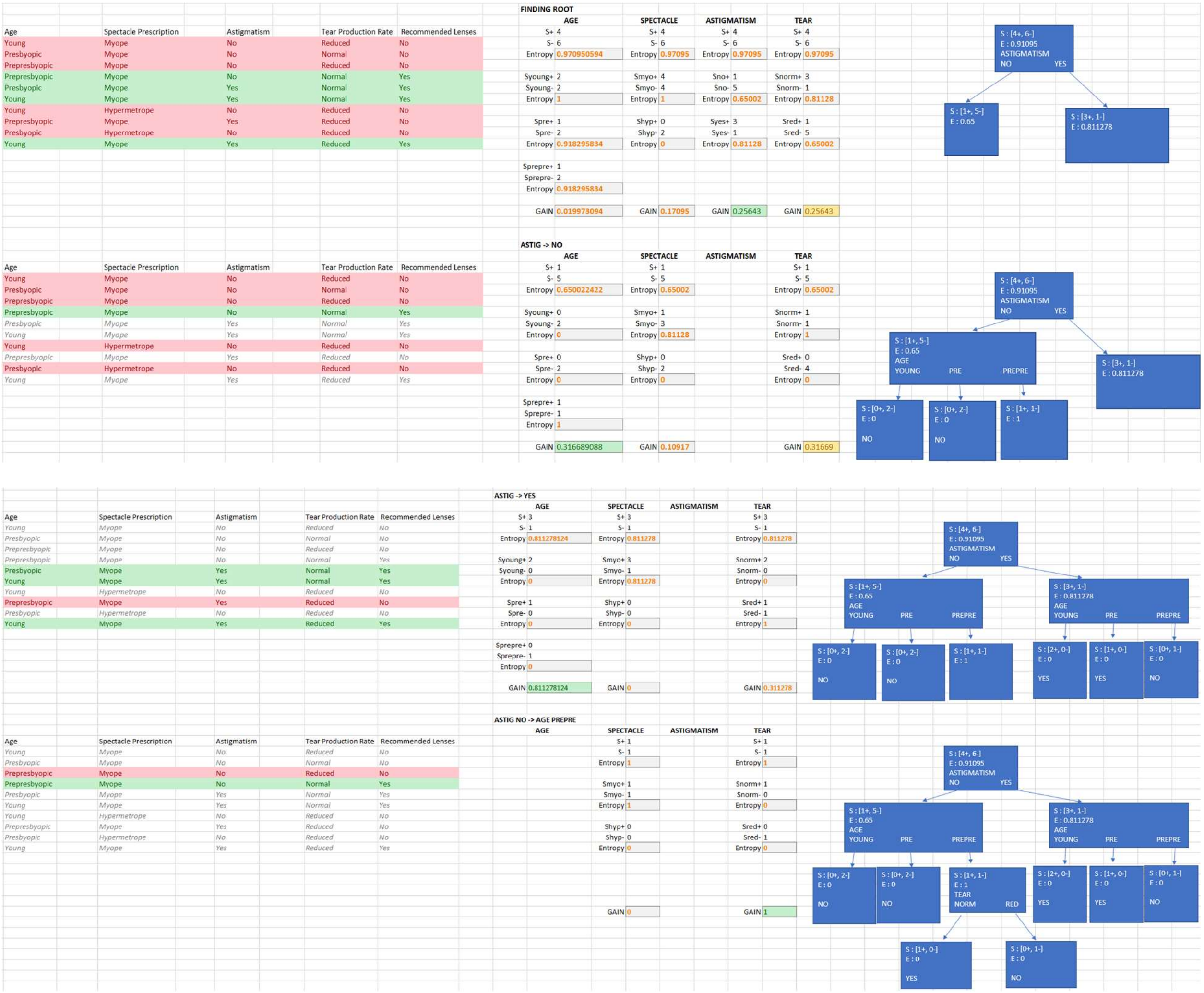
   b. [15 points] Complete the given python program (decision_tree.py) that will read the file contact_lens.csv and output a decision tree. Add the link to the online repository as the answer to this question.
   https://github.com/leolanggeng/contact_lens_ML

   c. [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.

While finding the root, astigmatism and tear have the same gain score, and also while finding the leaf of astigmatism -> age and tear also have the same gain. From this alone, there are at least 3 other possible trees, 1 the tree that I made and 2 alternate path to create the tree. This shows as the first time I ran the code I produced a different tree, but by repeating the code I do get the same tree that I made. There are also the case of the binary tree, the code cannot handle multi branch options so the 3 possibilities of age are combined.

Important Note: Answers to all questions should be written clearly, concisely, and unmistakably delineated. You may resubmit multiple times until the deadline (the last submission will be considered).

NO LATE ASSIGNMENTS WILL BE ACCEPTED. ALWAYS SUBMIT WHATEVER YOU HAVE COMPLETED FOR PARTIAL CREDIT BEFORE THE DEADLINE!



Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
Young	Myope	No	Reduced	No
Presbyopic	Myope	No	Normal	No
Prepresbyopic	Myope	No	Reduced	No
Prepresbyopic	Myope	No	Normal	Yes
Presbyopic	Myope	Yes	Normal	Yes
Young	Myope	Yes	Normal	Yes
Young	Hypermetrope	No	Reduced	No
Prepresbyopic	Myope	Yes	Reduced	No
Presbyopic	Hypermetrope	No	Reduced	No
Young	Myope	Yes	Reduced	Yes

FINDING ROOT

AGE		SPECTACLE		ASTIGMATISM		TEAR	
S+	4	S+	4	S+	4	S+	4
S-	6	S-	6	S-	6	S-	6
Entropy	0.970950594	Entropy	0.97095	Entropy	0.97095	Entropy	0.97095
Syoung+	2	Smyo+	4	Sno+	1	Snorm+	3
Syoung-	2	Smyo-	4	Sno-	5	Snorm-	1
Entropy	1	Entropy	1	Entropy	0.65002	Entropy	0.81128
Spre+	1	Shyp+	0	Syes+	3	Sred+	1
Spre-	2	Shyp-	2	Syes-	1	Sred-	5
Entropy	0.918295834	Entropy	0	Entropy	0.81128	Entropy	0.65002
Sprepre+	1						
Sprepre-	2						
Entropy	0.918295834						
GAIN	0.019973094	GAIN	0.17095	GAIN	0.25643	GAIN	0.25643

Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
Young	Myope	No	Reduced	No
Presbyopic	Myope	No	Normal	No
Prepresbyopic	Myope	No	Reduced	No
Prepresbyopic	Myope	No	Normal	Yes
Presbyopic	*Myope*	*Yes*	*Normal*	*Yes*
Young	*Myope*	*Yes*	*Normal*	*Yes*
Young	Hypermetrope	No	Reduced	No
Prepresbyopic	*Myope*	*Yes*	*Reduced*	*No*
Presbyopic	Hypermetrope	No	Reduced	No
Young	*Myope*	*Yes*	*Reduced*	*Yes*

ASTIG -> NO

AGE		SPECTACLE		ASTIGMATISM		TEAR	
S+	1	S+	1			S+	1
S-	5	S-	5			S-	5
Entropy	0.650022422	Entropy	0.65002			Entropy	0.65002
Syoung+	0	Smyo+	1			Snorm+	1
Syoung-	2	Smyo-	3			Snorm-	1
Entropy	0	Entropy	0.81128			Entropy	1
Spre+	0	Shyp+	0			Sred+	0
Spre-	2	Shyp-	2			Sred-	4
Entropy	0	Entropy	0			Entropy	0
Sprepre+	1						
Sprepre-	1						
Entropy	1						
GAIN	0.316689088	GAIN	0.10917			GAIN	0.31669

Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
Young	*Myope*	*No*	*Reduced*	*No*
Presbyopic	*Myope*	*No*	*Normal*	*No*
Prepresbyopic	*Myope*	*No*	*Reduced*	*No*
Prepresbyopic	*Myope*	*No*	*Normal*	*Yes*
Presbyopic	Myope	Yes	Normal	Yes
Young	Myope	Yes	Normal	Yes
Young	*Hypermetrope*	*No*	*Reduced*	*No*
Prepresbyopic	Myope	Yes	Reduced	No
Presbyopic	*Hypermetrope*	*No*	*Reduced*	*No*
Young	Myope	Yes	Reduced	Yes

ASTIG -> YES

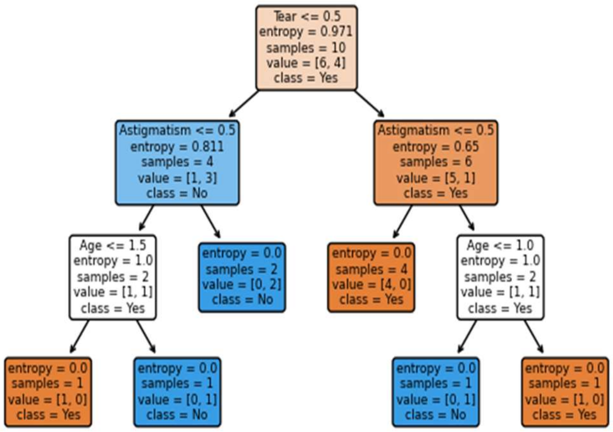
AGE		SPECTACLE		ASTIGMATISM		TEAR	
S+	3	S+	3			S+	3
S-	1	S-	1			S-	1
Entropy	0.811278124	Entropy	0.811278			Entropy	0.811278
Syoung+	2	Smyo+	3			Snorm+	2
Syoung-	0	Smyo-	1			Snorm-	0
Entropy	0	Entropy	0.811278			Entropy	0
Spre+	1	Shyp+	0			Sred+	1
Spre-	0	Shyp-	0			Sred-	1
Entropy	0	Entropy	0			Entropy	1
Sprepre+	0						
Sprepre-	1						
Entropy	0						
GAIN	0.811278124	GAIN	0			GAIN	0.311278

Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
Young	*Myope*	*No*	*Reduced*	*No*
Presbyopic	*Myope*	*No*	*Normal*	*No*
Prepresbyopic	Myope	No	Reduced	No
Prepresbyopic	Myope	No	Normal	Yes
Presbyopic	*Myope*	*Yes*	*Normal*	*Yes*
Young	*Myope*	*Yes*	*Normal*	*Yes*
Young	*Hypermetrope*	*No*	*Reduced*	*No*
Prepresbyopic	*Myope*	*Yes*	*Reduced*	*No*
Presbyopic	*Hypermetrope*	*No*	*Reduced*	*No*
Young	*Myope*	*Yes*	*Reduced*	*Yes*

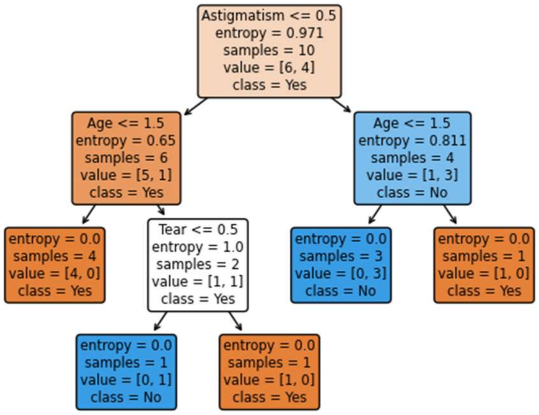
ASTIG NO -> AGE PREPRE

AGE		SPECTACLE		ASTIGMATISM		TEAR	
		S+	1			S+	1
		S-	1			S-	1
		Entropy	1			Entropy	1
		Smyo+	1			Snorm+	1
		Smyo-	1			Snorm-	0
		Entropy	1			Entropy	0
		Shyp+	0			Sred+	0
		Shyp-	0			Sred-	1
		Entropy	0			Entropy	0
		GAIN	0			GAIN	1

First run of the code



After re-running multiple times to find the same tree as mine



Astigmatism <= 0.5
entropy = 0.971
samples = 10
value = [6, 4]
class = Yes
Age <= 1.5
entropy = 0.65
samples = 6
value = [5, 1]
class = Yes
Age <= 1.5
entropy = 0.811
samples = 4
value = [1, 3]
class = No
entropy = 0.0
samples = 4
value = [4, 0]
class = Yes
Tear <= 0.5
entropy = 1.0
samples = 2
value = [1, 1]
class = Yes
entropy = 0.0
samples = 3
value = [0, 3]
class = No
entropy = 0.0
samples = 1
value = [1, 0]
class = Yes
entropy = 0.0
samples = 1
value = [0, 1]
class = No
entropy = 0.0
samples = 1
value = [1, 0]
class = Yes