Assignment 4

due to 27.4

1.Submission Instructions

1.1 Theoretical Part

a) The objective of this part is to practice the terms and concepts correlated to decision trees.

b) If needed, make sure to explain or present an example to any given answer to assure full understanding of your idea.

c) Submission: please submit a PDF version named "ex4-theoretical-part.pdf".

1.1 Practical Part

a) You are allowed to use the following Python packages: *numpy*, *glob*,*pandas*,*graphviz*.

b) Use python 3.7

c) Name your main code as ex4.py

Good Luck!

2. Questions

2.1 Theoretical Part

1. Decision Tree algorithm

dataset				
sample number	A	B	C	Y
1	F	F	F	F
2	F	F	F	T
3	F	F	F	T
4	F	F	F	F

1. Using the dataset above, we want to build a decision tree which classifies Y as True (T) or False (F) given the binary variables A, B, C. Draw the tree that would be learned by the greedy algorithm with zero training error. You do need to show computations.

2. Is this tree optimal (i.e.,does it get zero training error with minimal depth)? Explain in less than two sentences. If it is not optimal, draw the optimal tree as well.

dataset			
sample number	GPA	Studied	Passed
1	L	F	No
2	L	T	Yes
3	M	F	No
4	M	T	Yes
5	H	F	Yes
6	H	T	Yes

3. We will use the dataset below to learn a decision tree which predicts if people pass machine learning (Yes or No), based on their previous GPA (High, Medium, or Low) and whether or not they studied. On the following calculations use log on base of 2.

4. What is the entropy of H(Passed) ?

5. What is the entropy of H(Passed | GPA)?

6. What is the entropy of H(Passed | Studied)?

7. Draw the full decision tree, that would be learned for this dataset. You do not need to show any calculations.

2.2 Practical Part:

In this problem you are asked to implement a program that builds decision tree from categorical attributes and two-class classification tasks. The programming part requires building a tree from a training dataset and classifying unseen instances.

Our dataset comes from a risk loan dataset, which is being used to predict the risk quality of a loan application. Each instance is classified as good (class G) or bad(class B).

The tables of attributes and their possible values are shown in the table below:

A1: Checking status	x (no checking) n ($x < 0$, negative) b ($0 <= x < 200$, bad) g ($200 <= x < 200$, good)
A2: Saving status	n (no known savings) b ($x < 100$) m ($100 <= x < 500$) g ($500 <= x <= 1000$) w ($1000 <= x$)
A3: Credit history	a (all paid) c (critical/other existing credit) d (delayed previously) e (existing paid) n (no credits)
A4: Housing	r (rent) o (own) f (free)
A5: Job	h (high qualified job) s (skilled) n (unemployed) u (unskilled)
A6: Property magnitude	c (car) l (life insurance) r (real estate) n (no known property)
A7: Number of dependents	1, 2
A8: Number of existing credits	1, 2, 3, 4
A9: Own telephones or not	y(yes) ,n(no)
A10: Foreign workers or not	y(yes) ,n(no)

In each file, there will be a header that gives information about the dataset and example header. The example in the dataset is shown below. The header format includes several lines starting with // that provides some description and comments about the dataset. The line starting with %% will list all the class labels. Each line starting with ## will give the name of one attribute and all its possible values.

1. Implement a method called *decision_tree_build*, which builds the decision tree using the training data. You are allowed to use pandas library.

2. Use the graphviz library to plot the final tree. At each note write the computed Entropy and the Information-Gain. You should present the results with 5 decimal places. Submit your result as plot.png

3. Implement a method called *print_accuracy*. This method should evaluate and print the accuracy of your model.

4. Suggest three different improvements to the algorithm that could improve the result (add your answer to the theoretical part).