**Assignment #2: Decision Tree**

|  |
| --- |
| **Submission Instructions**   * Submit the following files on Blackboard:  1. The completed answer sheet provided on the last two pages (for Q1-Q10 in Part 1 and Q11-Q13 in Part 2).  * If you do not follow the instructions, your assignment will be counted late. |

**Part 1. Decision Tree Analysis in Python**

**Before you start**

For this assignment, you’ll be working with the **BankLoan.csv** file and the **Lab2.py** script (which we used in Lab #2). The BankLoan.csv file has data about 550 customers that received personal loans from a bank. The president of the bank wants to predict how likely a future customer is to pay back their loan so she can make better loan approval decisions.

The data file contains the following variables:

|  |  |
| --- | --- |
| **Variable Name** | **Variable Description** |
| **ID** | Customer identification number |
| **age** | The age of the customer, in years |
| **Male** | The gender of the customer |
| **income** | Customer’s yearly income in dollars |
| **married** | Whether the customer is married |
| **children** | How many children the customer has |
| **car** | Whether the customer has a car |
| **save\_act** | Whether the customer has ever had a savings account with SchuffBank! |
| **current\_act** | Whether the customer has an active account with SchuffBank! |
| **mortgage** | Whether the customer has a mortgage |
| **payback** | Whether the customer paid back their loan (0 = no, 1 = yes)  **NOTE: payback** is the outcome variable we are interested in here. It describes a categorical event (0 = no, 1 = yes). |

**Guidelines:**

1. You’ll need to modify the script with the following information to perform the analysis:

* Set the input filename to the bank’s dataset (i.e., BankLoan.csv).
* Set the training partition (using TRAINING\_PART) to 50% of the data set.
* Set the minimum split (using MINIMUMSPLIT) to 5.
* Set the maximum\_depth to 5.
* Make sure the outcome column setting is correct for your dataset.
* You will need to modify the model to reflect the dataset. This requires editing the Lab2.py script. Make sure you choose the correct outcome variable and exclude the variables that are inappropriate for the analysis. (HINT: ID is irrelevant to the analysis.)

1. Once you finish modifying the script, you can run the script.
2. **Based on your script output, answer Questions 1-6 in the answer sheet at the end of this document:**
3. **Now change the maximum\_depth to 10** **and re-run the script. Using the new tree, answer Questions 7-11 in the answer sheet at the end of this document.**

**Part 2. Compute and Evaluate Decision Trees**

Consider the following based on a different dataset than what you have done so far in this assignment.

**Question 12.** (write your answer in the answer sheet)

Suppose we run the decision tree algorithm and get a decision tree (called it Tree #1): compute the correct classification rate based on the following confusion matrix (*Compute it by hand. No need to use R/RStudio*):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Predicted outcome:** | |  |
|  |  | 1 | 0 |  |
| **Observed outcome:** | 1 | 700 | 270 |  |
| 0 | 130 | 900 |  |

Table 1. Confusion Matrix (Tree #1)

**Question 13.** (write your answer in the answer sheet)

Suppose we re-run the decision tree algorithm and get another decision tree (called it Tree #2): compute the correct classification rate based on the following confusion matrix (Compute it *by hand. No need to use R/RStudio*):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Predicted outcome:** | |  |
|  |  | 1 | 0 |  |
| **Observed outcome:** | 1 | 550 | 150 |  |
| 0 | 250 | 1050 |  |

Table 2. Confusion Matrix (Tree #2)

**Question 14.** (write your answer in the answer sheet)

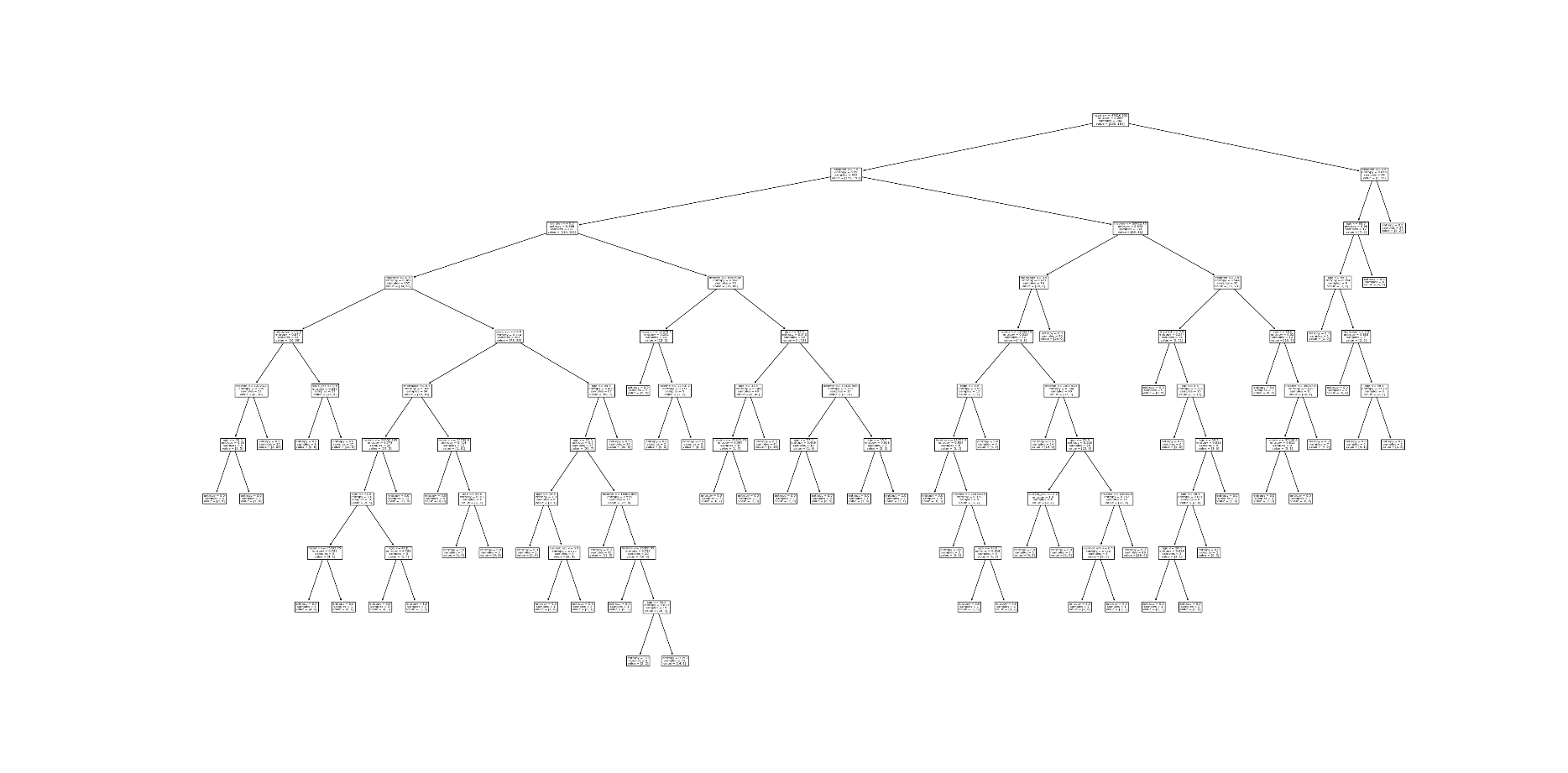
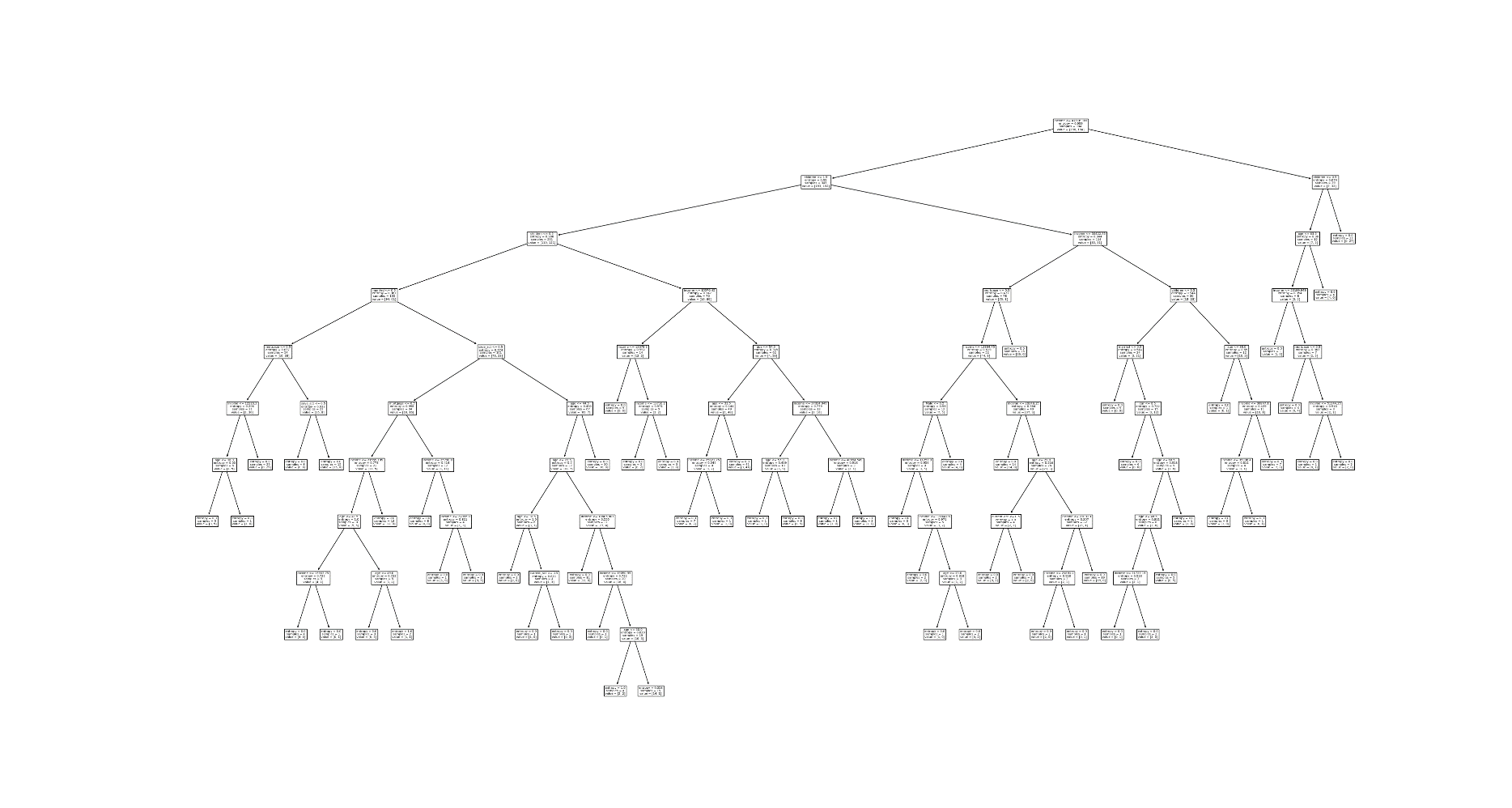
Which decision tree (Tree #1 versus Tree #2) has higher classification accuracy?

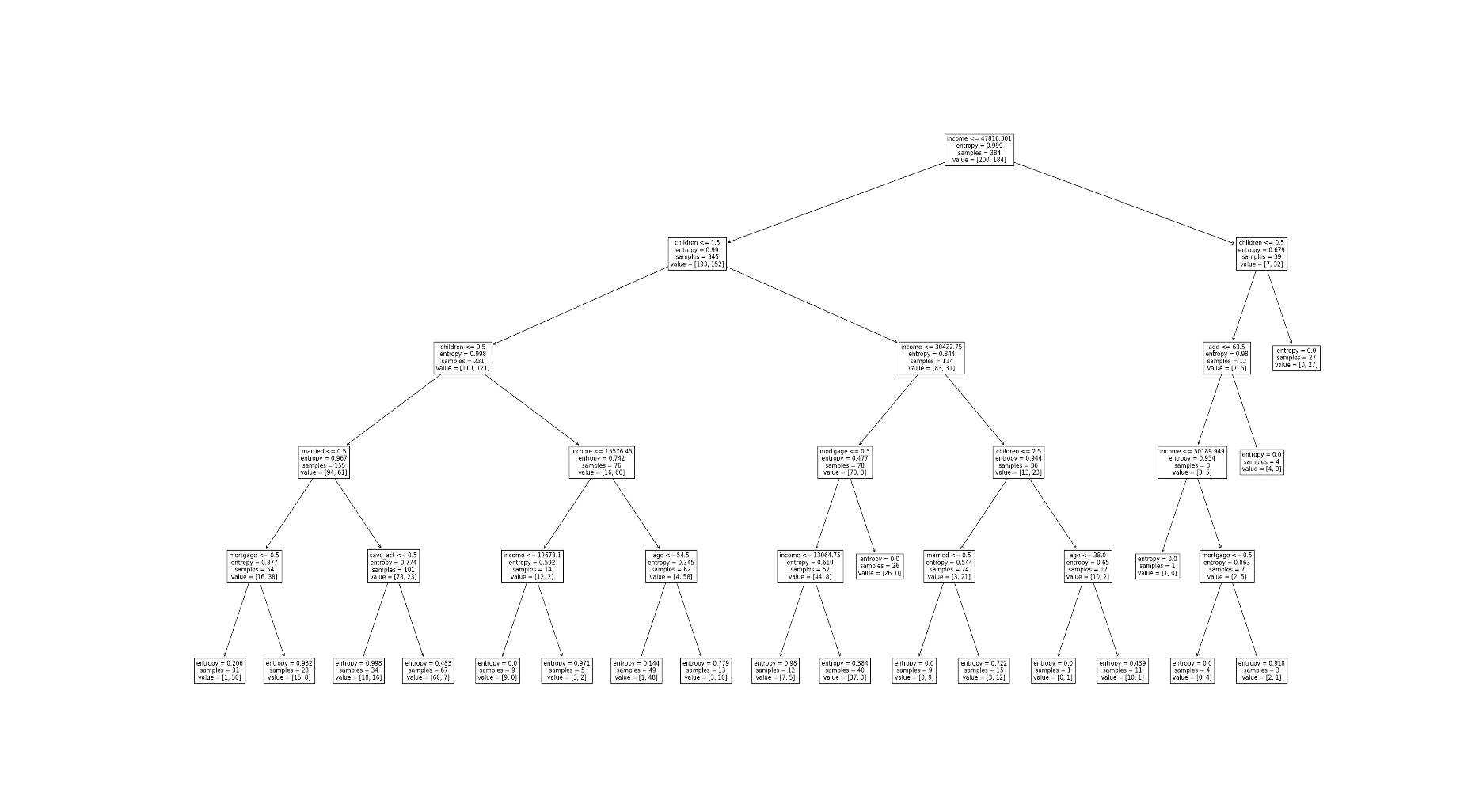
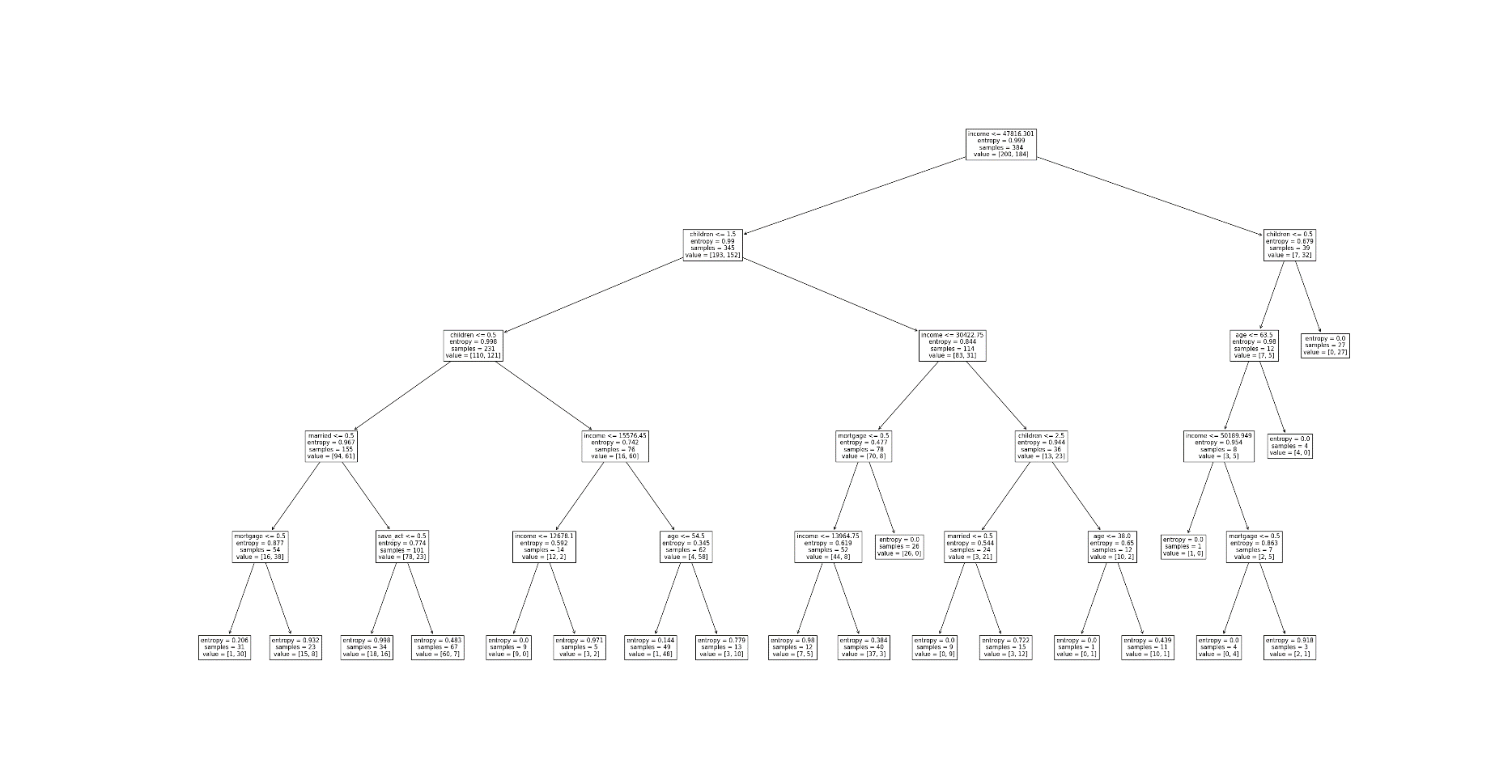
**Answer Sheet on the Next Two Pages……**

Answer Sheet for Assignment: Decision Trees in Python

*Fill in the answer sheet below.*

|  |  |  |
| --- | --- | --- |
|  | **Question** | **Answer** |
| **Part 1. Decision Tree in Python**  **(maximum\_depth = 5)** | | |
| 1 | How often will this tree make a correct prediction (including decimals)? Provide your answer for both the training set and the validation set. | Accuracy of training set: 0.8671875  Accuracy of validation set: 0.8545454545454545 |
| 2 | How likely is a customer to pay back their loan if he/she has one child and makes $35,000 per year? | 0.925 |
| 3 | How likely is a customer to pay back their loan if he/she makes $50,000 per year and has no children? | 0.416 |
| 4 | How likely is a customer to pay back their loan if he/she makes $83,000 per year and has two children? | 1.000 |
| 5 | Describe the profile of the least likely customer to successfully repay their loan. | People who have an income of less than $12678 and have more than one child |
| 6 | Describe the profile of the most likely customer to successfully repay their loan. | People who earn more than $47816 and have at least one child |
| **(maximum\_depth = 10)** | | |
| 7 | How often will this new tree make a correct prediction (including decimals)? Provide your answer for both the training set and the validation set. | Accuracy of training dataset(criterion=entropy, max\_depth=4, min\_samples\_split=3): 0.9869791666666666  Accuracy of validation dataset(criterion=entropy, max\_depth=4, min\_samples\_split=3): 0.8606060606060606  Accuracy of training dataset (criterion=entropy, max\_depth=6, min\_samples\_split=3): 0.9869791666666666  Accuracy of validation dataset (criterion=entropy, max\_depth=6, min\_samples\_split=3): 0.8666666666666667 |
| 8 | Is this model better or worse than the first model at predicting who will repay their loan in the training and testing dataset? | Based on the accuracy scores alone, the first model with a maximum depth of 10 is better than the second model with a maximum depth of 5 at predicting who will repay their loan, both on the training dataset (0.987 vs 0.867) and validation dataset (0.867 vs 0.855). However, it's important to keep in mind that accuracy alone may not be the best metric to evaluate the performance of a model, especially if the dataset is imbalanced or has other issues. Additionally, it's possible that the first model with a maximum depth of 10 may be overfitting the training data, so it's important to evaluate the model using other metrics and techniques to confirm its performance. |
| 9 | How likely is a customer to pay back their loan if he/she has one child and makes $35,000 per year? | 0.935 |
| 10 | How likely is a customer to pay back their loan if he/she makes $50,000 per year and is married but has no children? | 0.416 |
| 11 | How likely is a customer at the age of 30 to pay back his/her loan if he/she makes $33,000 per year and has two children? | 0.875 |
| **Part 2 Compute and Evaluate Decision Trees** | | |
| 12 | What is the correct classification rate for Confusion Matrix (Tree #1)? | (270+130)/2000=0.21-0.2=0.813 |
| 13 | What is the correct classification rate for Confusion Matrix (Tree #2)? | (250+150)/2000=0.21-0.2=0.814 |
| 14 | Which decision tree (Confusion Matrix (Tree #1) versus Confusion Matrix (Tree #2)) has higher classification accuracy? | The classification accuracy for both Confusion Matrix (Tree #1) and Confusion Matrix (Tree #2) is the same, which is 0.8 or 80%. Therefore, both decision trees have the same classification accuracy. |

**Maximum depth=10**

**Max capacity 5 tree plot**