11/6/2021

*Nathan A. Horak*

# ASSIGNMENT 1T: DECISION TREES

Please use this template in creating your response. Retain the gray text and MS Word headings, and supply your responses where indicated. Your materials, in black 12-point Times New Roman, should not exceed 5 pages excluding this gray text, the references and your figures. You may add appendices, which should be referred to in the body of the paper, and which will be read on an as-needed basis. Note the evaluation criteria below, and leave plenty of time for editing so that your paper responds to them and you obtain the most favorable grade. Responses considered “good” should go beyond the minimum of what’s requested.

Go [here](https://colab.research.google.com/drive/1DGZdokU6DOXCtDjYjGU_yev0_6KZOP2o?usp=sharing), and copy to your Google drive.

## 1. SUMMARY DESCRIPTION

In one or two sentences, describe a decision tree application that you will implement by modifying the code you copied.

Using a decision tree machine learning algorithm, I would like to predict the education status of an individual by their date of birth, income, number of kids at home, and amount spent yearly on wine from a customer personality dataset. The tree will predict based on inputs whether an individual has a Basic, Graduate, Masters, or PhD education level.

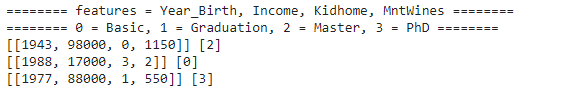
## 2. DATA SOURCE

Explain the source of your data (e.g., collected by hand; Kaggle). Point to a URL so we can see the data.

The data was obtained via Kaggle. The URL is: https://www.kaggle.com/imakash3011/customer-personality-analysis

## 3. EXAMPLES

Create three different input sets to the application and show their outputs.



The following inputs return outputs of Master, Basic, and PhD as predicted education levels for inputs 1, 2, and 3, respectively.

## 4. SOURCE

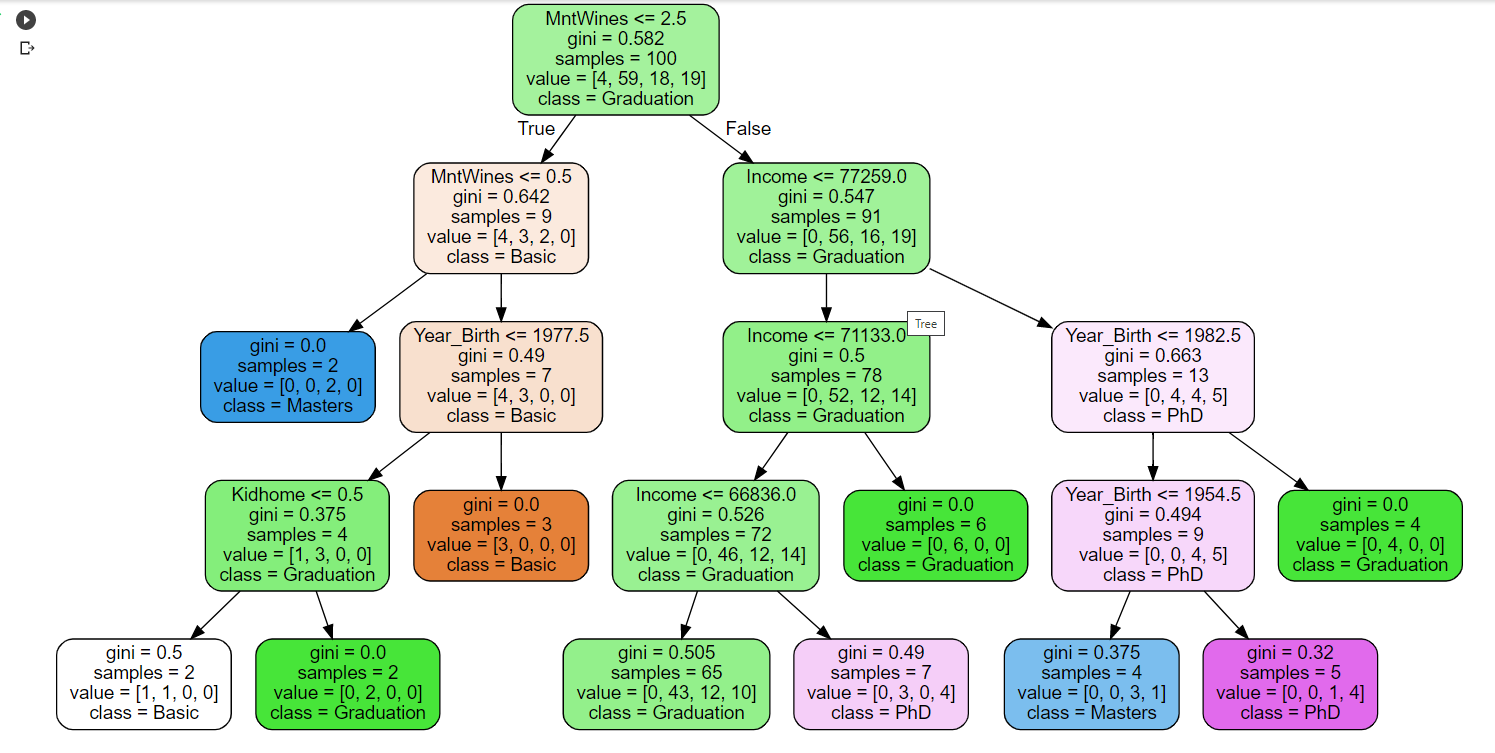
Please supply the URL of your (shared) Colab code.

https://colab.research.google.com/drive/1Zg-jmyW9LZ45MFiJlTjWaHAbtJ34GZxb?usp=sharing

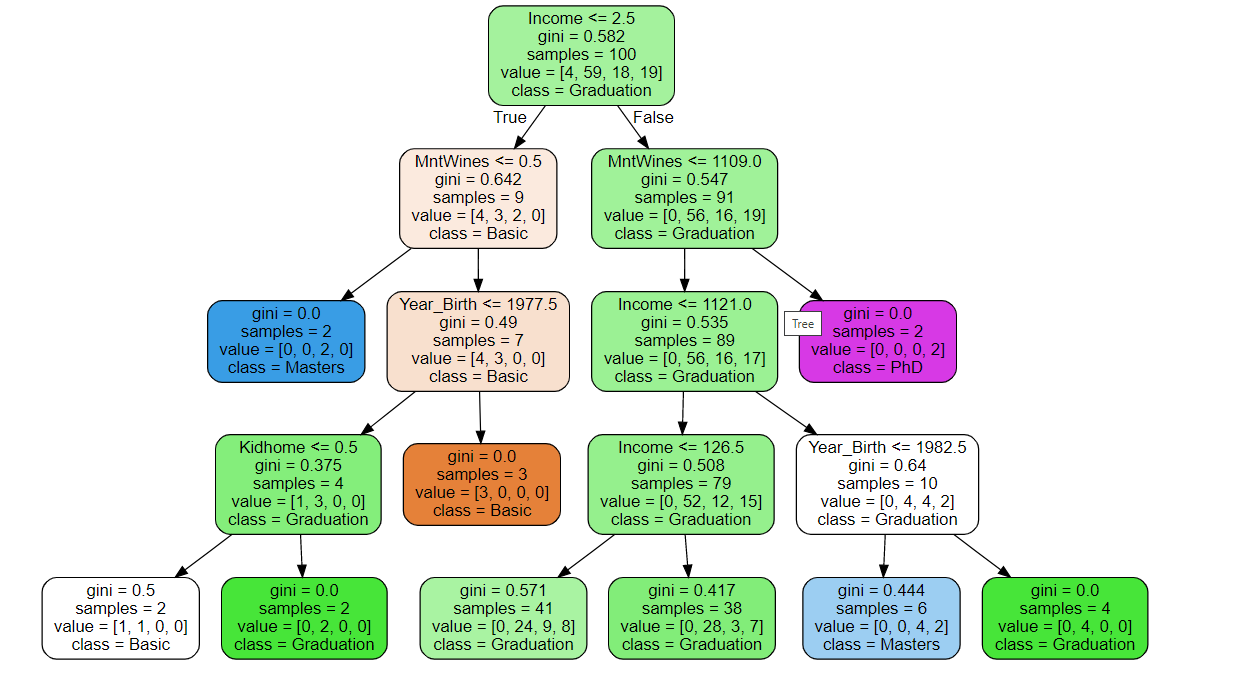
## 5. DATA ALTERATION FOR CHANGED RESULTS

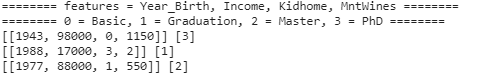
Alter the data so that it changes in the outputs for the above inputs. Explain why. Limit: 2 normal paragraphs (remember that you can use appendices for reference material).

When altering the data for this problem, I wanted to alter the data across the entire dataset rather than cherry pick individual points. A decision tree functions based on the effectiveness of thresholds that the algorithm finds and sets as conditional markers. These conditional statement sorters are crucial to the functioning of a decision tree. So, I decided to try to alter the data in a way relevant to these markers. See original tree, clf\_tree, below:



There exist a total of 78 data points that have an income value of less than 77259 and after lots of experimentation, including counting hefty wine bills as an additional kid amongst other things, I decided to use this threshold value, 77259, in my data alteration. I subtracted 77259 from all income values in this column and then I changed all values less than or equal to 0 in the income column to be equal to the value of the MntWines column entry. In essence, what I did was treat every individual with less than 77259 income from the original dataset as having their income be equal to the amount that they spend yearly on wine[3]. See the updated tree, clf\_tree2, below:



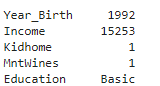


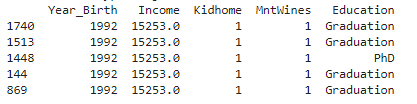
Given the updated tree with alterations the following inputs return outputs of PhD, Graduation, and Master as predicted education levels for inputs 1, 2, and 3, respectively.

## 6. INCONSISTENT DATA

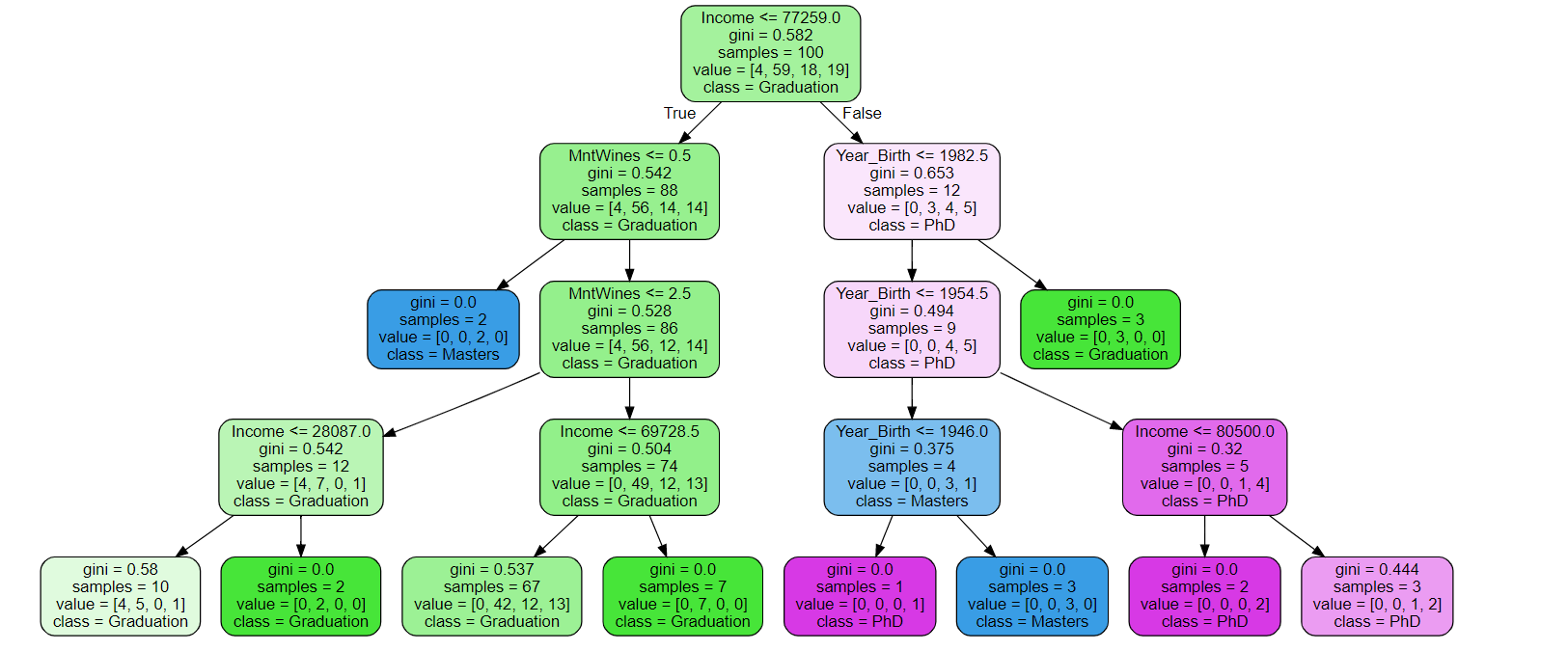
Alter the data so that it contains inconsistencies, and show changes in the I/O of the examples in part 3. Explain. Limit: 2 normal paragraphs (remember that you can use appendices for reference material).

Per the lecture slides, inconsistent data refers to two sets of data that have matching values among feature variables (input) but differing values among the target variable (output). To show the effect that inconsistent data can have on decision trees and specific input/expected output, I wanted to introduce consistencies that would have a noticeable impact on expected results. Examining the original tree, the classification ‘Basic’ stuck out to me given that there are only 4 in the dataset. Furthermore, the left side of the tree that contains the only ‘Basic’ values has many less data points than the right side. Introducing inconsistencies related to this classification would be certain to alter the tree and subsequent classification- especially for input 2 from which the original tree predicts a ‘Basic’ education level. I found the following data at row 14, the first ‘Basic’ from the dataset:

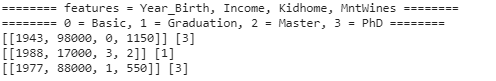


From here, I altered the data of the features for rows 19, 39, 59, 79, and 99 so that the feature data matched the feature data from row 14, while the ‘Education’ labels were different.

Given that row still exists, there are now 6 rows with this same feature data. Four of them predict a ‘Graduation’ education level, one predicts a ‘PhD’ level, and one predicts a ‘Basic’ level. See below for clf\_tree3 and the new predicted output:



There are a few things to note here. First, given the parameters set and the given dataset, this decision tree never predicts a ‘Basic’ education level. In fact, looking at the lower leftmost tree node, all input data of a true ‘Basic’ education level would predict a ‘Graduation’ education level. Secondly, the tree branches by weight have swapped places. The heaviest tree branch is now on the left (with the ‘Basic’ education level data points) while the lighter branch (without any ‘Basic education level data points) is on the right. Third, this decision tree has more incorrectly predicted data points within the tree nodes. The tree with inconsistencies has 31 incorrectly predicted points while the original tree has 28 incorrect predictions. Lastly, as seen below, the added inconsistencies not only changed the predicted output for the input array that originally predicted a ‘Basic’ education level (input 2), but also for input 1 that predicted a ‘Master’ education level.



Given the updated tree with inconsistencies the following inputs return outputs of PhD, Graduation, and PhD as predicted education levels for inputs 1, 2, and 3, respectively.

## 7. BENEFITS

In at most a page and a half (of 12-point text), explain the pros and cons of decision trees *applied to the application you have chosen*. You may build on the work of others but (1) show clearly that you understand this work and (2) observe all plagiarism rules scrupulously, including clear citations.

There were a few pros utilizing the decision tree algorithm for this application. The tree was simple to set up and it was easy to view the methodology and rules behind how it made its classifications. I could easily see how the decision tree was classifying any input by examining the tree and its associated nodes. This aspect was very helpful when I was experimenting with input arrays and their associated outputs, along with changes to any data alterations or introduction of data inconsistencies. Per geeksforgeeks.org[1], this is a commonly cited strength of the decision tree algorithm. Another strength of the decision tree algorithm, again cited by geeksforgeeks.org[1], is that decision trees are capable of handling both continuous and categorical variables. The four feature variables I used were all continuous variables (including date of birth, which can when represented as a date function as a categorical variable) while the target output was a categorical variable- specifically, education level. I could have (although I chose not to) use a categorical variable as an input variable. Indeed, I began my analysis using ‘Marital\_Status’ as a feature but decided against using it. Lastly, per geeksforgeeks.org[1], a strength of the decision tree algorithm which I fully utilized was that a decision tree provides a clear indication which fields are most important for prediction or classification. By examining the conditional statements within the tree nodes, it is easy to see what variables are most important to an array of data being classified. When considering data alterations, I knew I wanted to alter ‘Income’ and I came to this conclusion by examining the decision tree. ‘Income’ is the feature variable linked to the tree nodes at several important locations on the tree. It is present at both upper levels of the tree as well as within nodes containing a large sample of data points. Additionally, when considering the introduction of data inconsistencies, I knew I wanted to introduce inconsistency towards predicting the ‘Basic’ class. I did this noting how few data points were predicted to have the ‘Basic’ education level.

The decision tree algorithm is not perfect, and there was a major downside to using this algorithmic method for this application. Per geeksforgeeks.org[1], decision trees can be computationally expensive to train and the process of growing a decision tree is computationally expensive. A big reason why this is true is the data must be split at each node before the optimal split is found. Although keeping a smaller sample size and limiting max depth of the tree (addressed further below in COMMENTS) assisted in limiting excess computational burden, these were decisions that inherently led to a less accurate tree. Increasing the max tree depth parameter would have allowed for additional splits at the bottom of the tree. This would have reduced the number of incorrect classifications and thus lowered the Gini measure at these bottom nodes. It should be noted that the two other weaknesses of the decision tree algorithm were mitigated by the nature of the application. According to geeksforgeeks.org[1], two other main weaknesses of the decision tree algorithm are that decision trees are less able for estimation tasks with the goal of predicting a continuous variable and that decision trees are prone to errors in classification problems with many classes and relatively small number of training examples. By predicting a categorical data output rather than a continuous attribute, the first weakness was avoided. The second weakness was avoided by maintaining a relatively large sample size and by some data wrangling. As addressed below, a target variable class ‘2n Cycle’ was relabeled as ‘Master’ given that both designations refer to roughly the same amount of educational achievement. This label seems to be used in non-Great Britain Europe in lieu of a ‘Master’ designation[2].

## EVALUATION



## COMMENTS

While working on this problem, I decided to work with a sample size of 100 rather than the full dataset (size 2240). I did this for two reasons. First, I wanted the original tree to contain tree nodes with conditional statements for each feature variable and I wanted each target output predicted by at least one leaf node without too large of a tree. For visual purposes a smaller, more compact tree is preferred but I also wanted to not get carried away given the decision tree algorithm’s run-time. Second, I knew that I would be tweaking with alterations and inconsistencies and it is much easier to account for changes due to these when you are working with a smaller data set. Given that the original dataset (and indeed the sample dataset I sampled) contain a much smaller number of outputs for ‘Basic’ education level compared to the other three target outputs, I debated wrangling these data points out. However, I had already used data wrangling to label ‘2n Cycle’ education level as ‘Master’ education level and I didn’t want to spend too much time wrangling when I could instead be experimenting with various data alterations and introducing data inconsistencies.

## APPENDIX

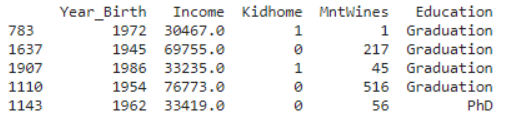
[1] “*Decision Tree*”. GeeksforGeeks. 22 June, 2021. <https://www.geeksforgeeks.org/decision-tree/>

[2] “*Structure of studies in Poland*”. National Agency for Academic Exchange. Accessed 9 November, 2021. <https://study.gov.pl/structure-studies-poland/>

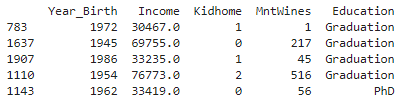
[3]

As mentioned above, I experimented with multiple ways of altering the data before settling on subtracting 77259 from each individual’s income and setting each individual with 0 or less income after this subtraction to have income equal to the amount they spend yearly on wine. One avenue I experimented with that I thought would work well but was not so impactful in practice was considering individuals with hefty wine expenses as having additional children. I decided to treat each 250 spent on wine yearly as an additional child in the home, up to a maximum of 4 kids for wine expenses greater than 1000. Here are the first five rows from the original dataset (top) and the altered dataset (bottom) with the above methodology applied. Note the difference with regards to the ‘Kidhome’ column for the fourth data entry, numbered 1110:

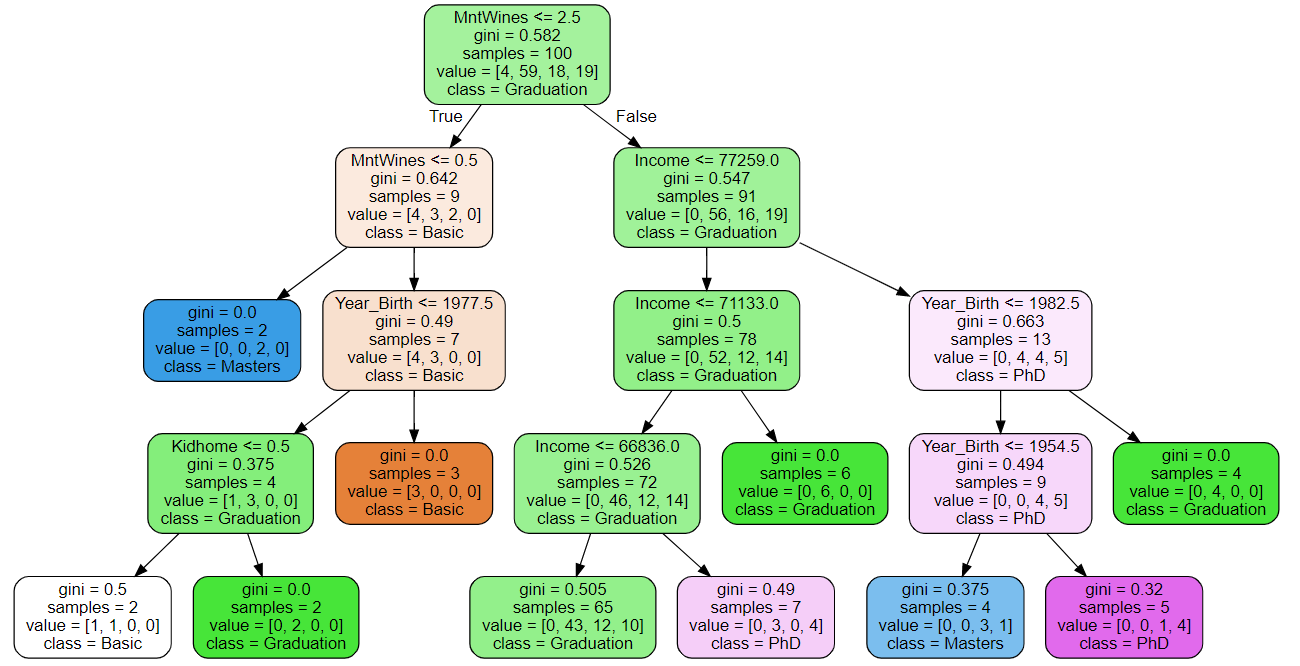
*Head of original data frame-*



*Head of appendix altered data frame-*



I had high hopes for this data alteration and I was let down when I noticed that not only the resulting predicted outputs were the same as for the original tree, but that the decision tree was identical to the original.



Upon examination of the tree, I realized why this data alteration changed nothing. There was only one node that made a decision based on the ‘Kidhome’ feature and it considered only 4 data points. Unless the changes I made to the ‘Kidhome’ column were drastic enough to make it more impactful than the ‘Year\_Birth’, ‘MntWines’, or ‘Income’ features as the decision tree algorithm performed its analysis of Gini coefficients, the structure of the tree would not change. Furthermore, the leaf nodes below the ‘Kidhome’ decision node would not change unless the data alterations caused a meaningful difference with regard to the statement of ‘Kidhome <= 0.5’. In this decision node, 1 ‘Basic’ Education level and 3 ‘Graduation’ Education level individuals would be divided into leaf nodes of 0 children at home or 1 or more children at home. With the data alteration, an individual’s ‘Kidhome’ value would increase if they spent heavily on wine. However, all four of these individuals were already placed on this side of the tree at the very top-most node by having a yearly wine expense of less than or equal to 2.5 (equivalent to less than 3). Given that the first threshold for an additional kid was set at 250 dollars spent yearly on wine, there is no way that any of these individuals could have gained an additional child at home and thus the tree is completely unchanged.