CPSC 483, Project 2

Team Name: ‘AKMMR’

Team Members:

|  |  |  |  |
| --- | --- | --- | --- |
| Serial Number | Name | CWID | Email ID |
| 1 | Asit Kumar Dash | 893423350 | asitdash@csu.fullerton.edu |
| 2 | Kashyap Sodha | 893260984 | ksodha@csu.fullerton.edu |
| 3 | Melika Amini | 893423350 | melika\_am@csu.fullerton.edu |
| 4 | Milan Ghori | 893585703 | ghorimilan1994@csu.fullerton.edu |
| 5 | Rishabh Sharma | 893463695 | rishabh230795@csu.fullerton.edu |

Under guidance of Prof. Christopher Ryu

Machine Learning – CPSC483

Department of Computer Science

California State University, Fullerton

b) Brief description of the software and the tool

Description of Software and Tool:

The project was done in python 3 using the sci-kit learn library. The ide used is spyder. Spyder provides an IPython console which makes it easy for the developer to check the instances of the code at any point. We also used jupyter-notebook to research on some of our findings. Scikit learn provides the user with many tools to execute high end machine learning tasks. Scikit learn is developed by a large community of the developers and machine learning experts therefore new techniques are included in it in a very short span of time.

In order to help the users who have a hard time to find which algorithm should be used Andres Muller the developer of scikit learn made a very user friendly flow chart.

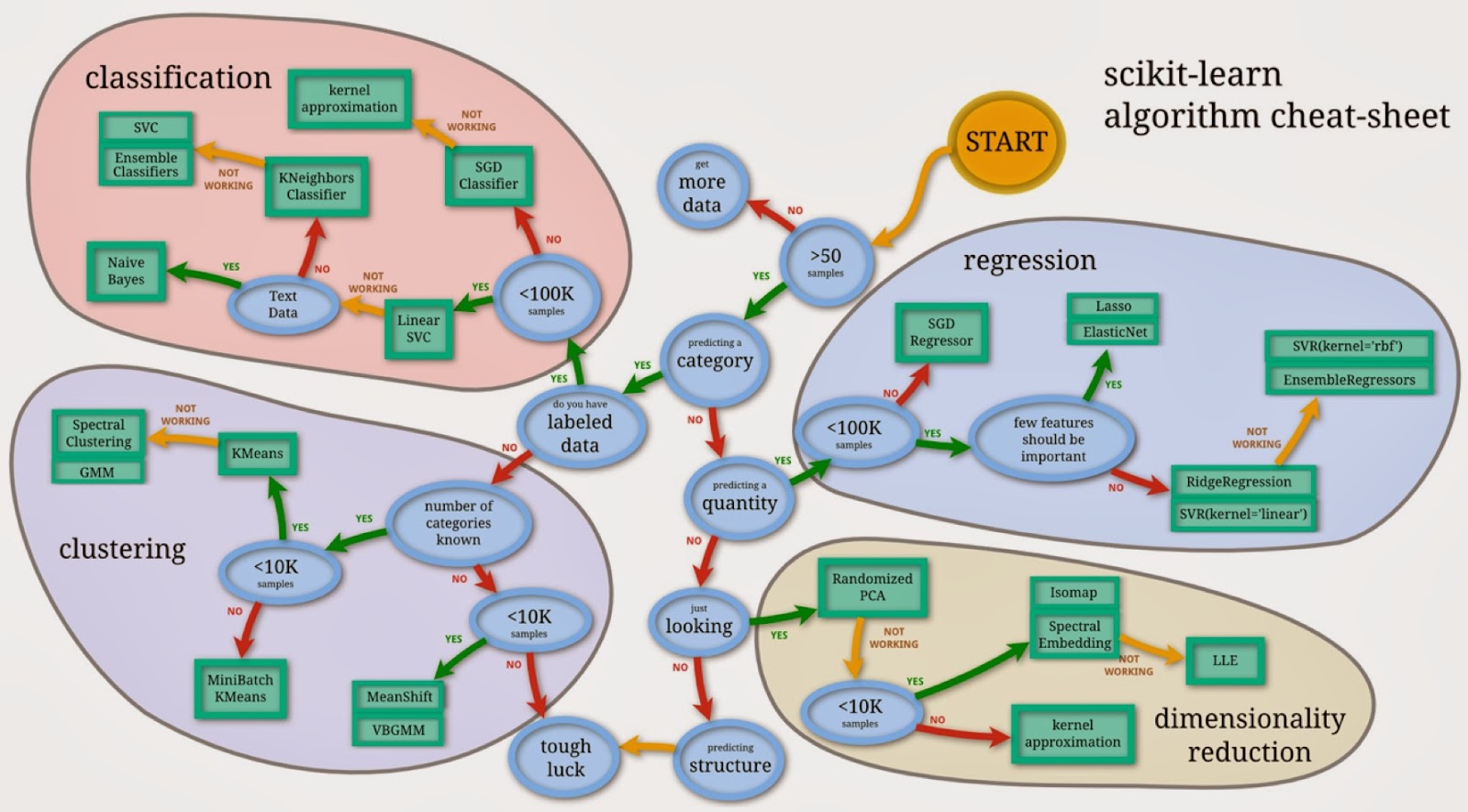


Image Source: <https://www.oreilly.com/ideas/six-reasons-why-i-recommend-scikit-learn>

C) i: Conduct classification using each of those methods specified above and come up with the best classifier from each method

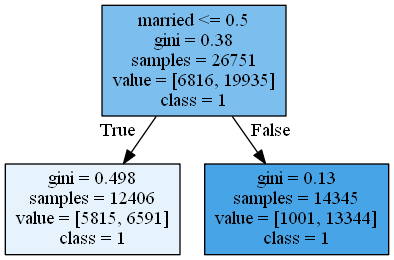
Answer: For classification we first start with the decision tree based classification method. Decision trees is a non-parametric supervised learning method that is used for classification. The goal is to **create a model that predicts the value of a target variable by rules that will be inferred with the help for the data.**

Decision trees work according to the following way:

1. Place the best attribute of the dataset at the **root** of the tree.
2. Split the training set into **subsets**. Subsets should be made in such a way that each subset contains data with the same value for an attribute.
3. Repeat step 1 and step 2 on each subset until you find **leaf nodes** in all the branches of the tree.

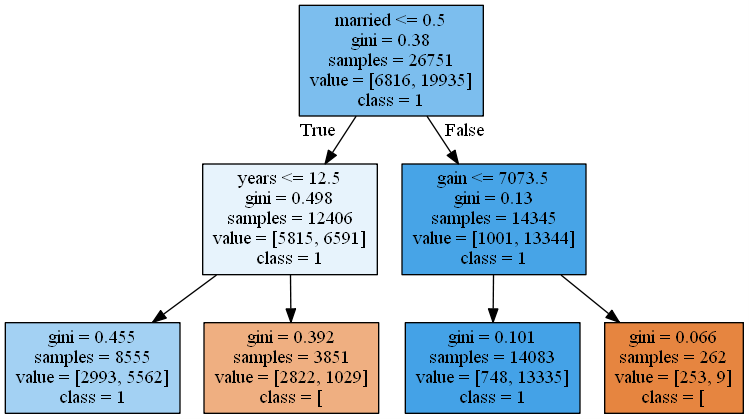
We conduct the decision tree classification and use different level of tree depth to predict the data. This give us different decision tree models that can be used to find the parameter of the dataset.

Decision tree with depth 1.



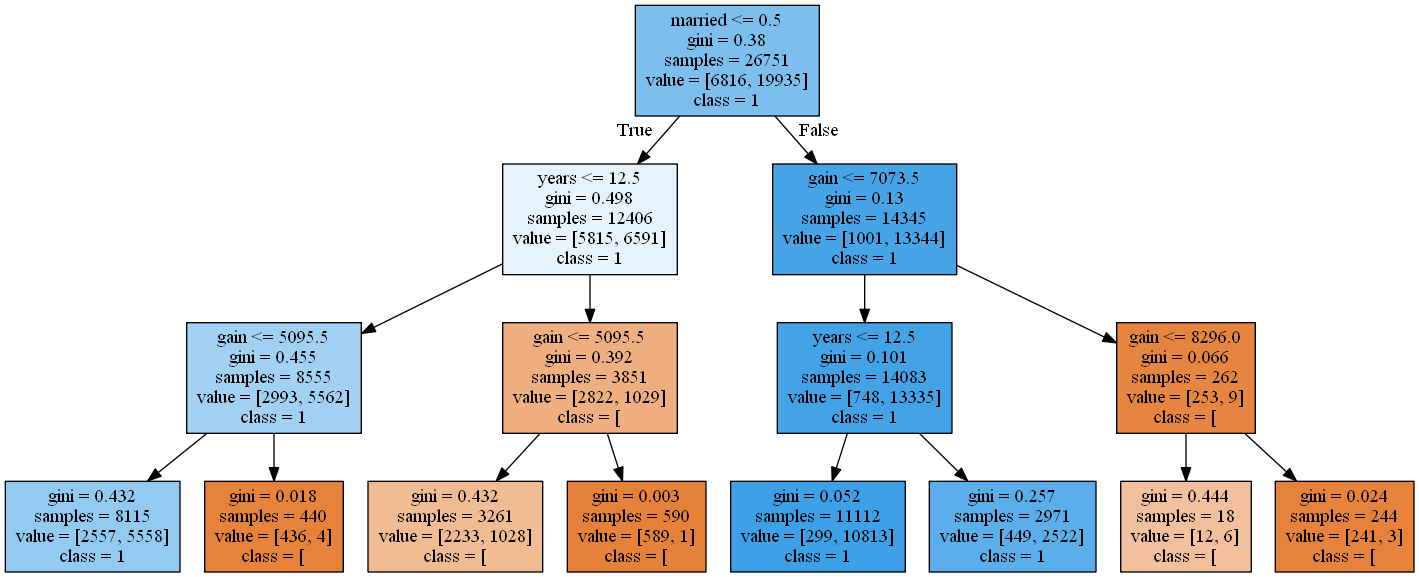
The above decision tree takes married as a classification status based on the **gini index** method and gives the above model. We can not use this method because both the datasets are classified to class 1. Class 1 means that the total income will be less that 50K and class 2 means that total income will be more than 50K. **Left branch is also the true condition and right branch is always the false condition.**

Then we increase the depth of the tree to 2.



New model gives us two leaf node that are the lowest level node with class 0 and two leaf node with class 1. Therefore at level 2 the person who is married-civ-spouse and whose years of education is more than 12.5 years can make more than 50k or else the person who has any other marital status but has a capital gain of more than 7073.5 can make more than 50K. At every node the parameter to decide the root value for that node is calculated using the gini index for all the samples at that node.

We keep on increasing our tree now to level 3 and we get the following.



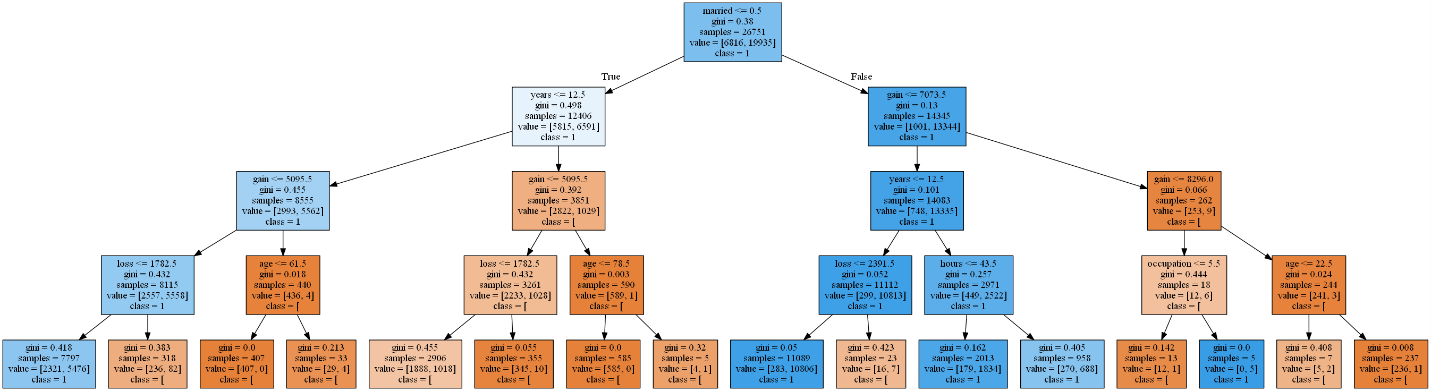
Same way as we following the above method we can find out the profile for people who will earn more than 50K.

At level 3 we get that the people who are married-civ-spouse, have more than 12.5 years of education will get more that 50K, People who are married-civ-spouse, have less than 12.5 years of education but has a capital gain more than 5095.5 will earn more than 50K and also people who have any other marital status but has a gain of more than 7073.5 will earn more than 50K.

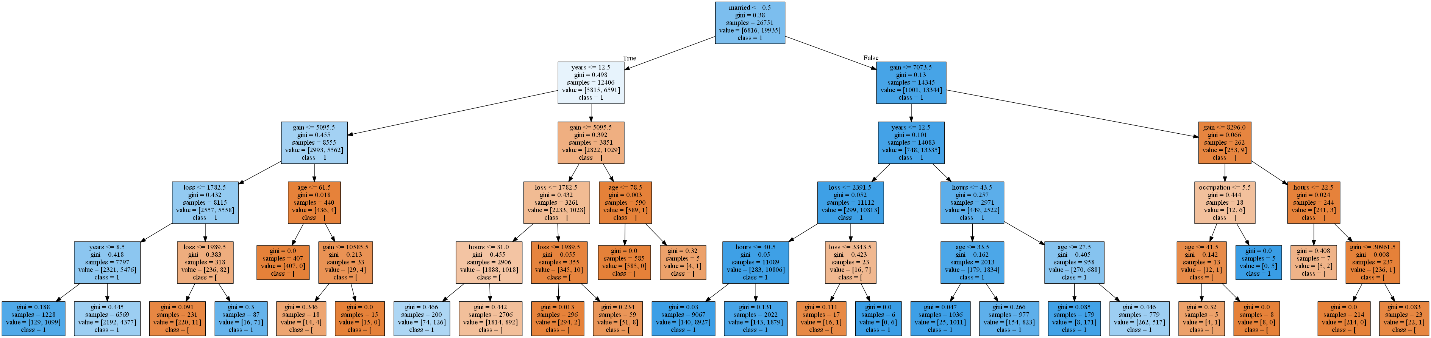
All the trees represent some model that can be used for classification and these models are the nodes represent the rules.

We made the trees to a depth of 6 and the following trees models were obtained.

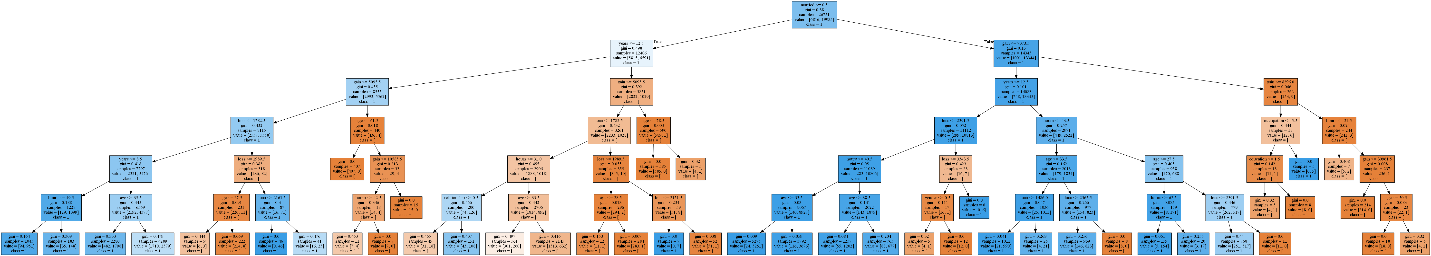
At the depth of 4:



At a depth of 5:



At a depth of 6:



We can keep on increasing the depth of our model to a specific tree depth that we want. However if we don’t pass any maximum depth value then the **algorithm keeps on computing the model till it doesn’t reach a leaf node that has an entropy of 0.**