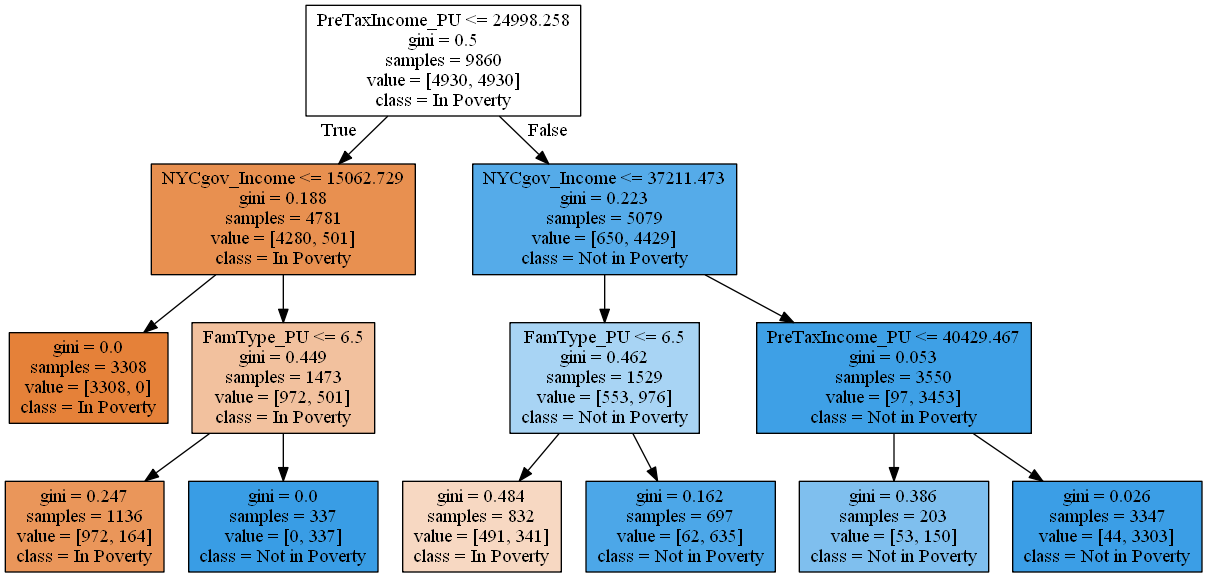
Decision Tree:

**Background**

Decision tree is a non-parametric and supervised algorism that go through all the available attributes to make the best prediction base on information gain. The information gain can tell us the importance of a certain attribute by calculating the impurity of the parent and child nodes. For our study, we use the Gini as our cost function because it would calculate the result faster, as we have a relatively big dataset 68644\*79. The function for the Gini is .

**Preprocessing**

Before we can produce our decision tree, we need to make sure our dataset has balanced target classes, right dimensionality and most representative observations. The preprocess techniques we implemented in our study are feature selection and instance selection. For the feature selection, The goal is to drop irrelevant or redundant features to reduce the dimensionality of the dataset. Some of the variables are: *Household Unit ID*, *Age Category, Poverty Gap, Tax Unit, etc.* We also decided to drop the variables that have high correlation with the total income, since income and poverty status shared the same characteristic (low income directly decides whether the person is in poverty or not). If we include total income, the decision tree would pick income as criteria in every nodes, which will underestimate the impact of other variables (as figure 1 shows below).

* ( Figure 1 shows the model will pick income 4 out of 6 times as the node)*

The instance selection will select the observations that are most representative for the study. Our dataset records 70000 individuals and are grouped by the household unit. Table 1 shows the first 7 observations of the dataset, the observations that have same SERIALNO indicates the individuals are belong in the same household. The Poverty Status of each individual is decided by the head of family (*Povunit\_Rel = 1*), if the head of family is in poverty then rest of the members are in in poverty too regardless of other features. In the case, we only kept the head of family as our unit of measurement.

 Another issue our dataset has is the imbalanced target classes. The original dataset is dominated by the non-poverty class, which is about 80% of the total observations. This would cause our training set biased toward the dominated class, which is non-poverty. To fix this problem,

we downsize the poverty class in the training set (EliteDataScience, 2017). We used *resample* function to randomly draw from the non-poverty class without replacement and keep drawing until it matches the number of the poverty class. *(table 1)*

The new training set now would have equal distribution between poverty and non-poverty.

**The Decision Tree**

I am using Sklearn for the model building, I set the max\_depth to 3 in order to avoid overfitting. The graph below is the decision tree. The accuracy for the training set is 83% and 79 % for the testing set.