**MODEL BUILDING**

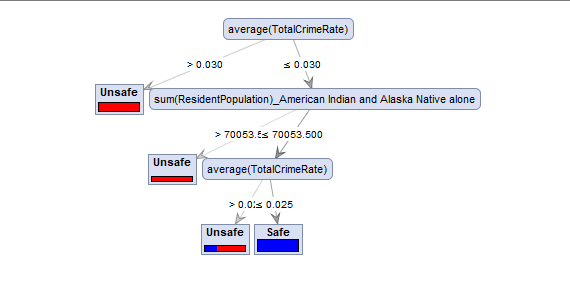
**Objective 1: Identify model structures**

In Data Mining Models, there is a common structure for resolving the problem. We followed these general steps;

* First we have created the mining structure and include the columns of data that might be required for the model analysis. We explained the data transformation we already made in the “Methodology Selection” document.
* Then we selected the algorithm as “Decision Tree” as we discussed it the previous document.
* We will choose the columns from the structure to use in the model according to our needs. We will determine the input and output columns in the final data set.
* We can set parameters to fine-tune the system.
* Finally, we’ll populate the model with data by processing the structure.

**Objective 2: Run and evaluate the models**

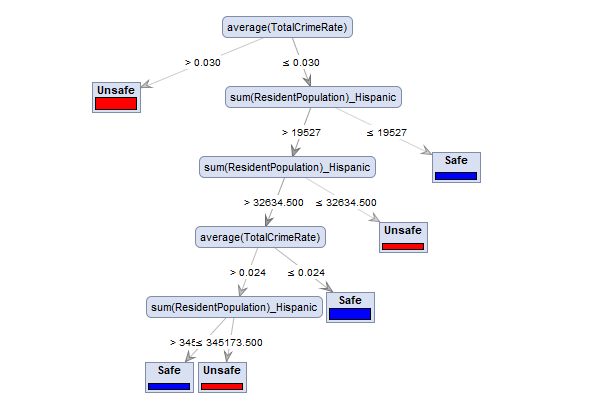
1. We run the model by selecting the all columns which are Per Capita Personal Income Average by State”, Total Crime Rate by State”, Sum of Resident Population of Age Groups, Sum of Resident Population of Race Groups, Sum of Population of Hispanic Origin, Sum of Resident Population of Male and Female separately. By running the model we get Figure-1.



**Figure-1:** The decision tree output of selecting all columns.

According to this output, we say that if the crime rate of a state which was below the rate of 0.03 then that state is being considered as unsafe. If not, the attribute of Sum of Resident Population of American Indian and Alaska Native alone is considered. If the Sum of Resident Population of American Indian and Alaska Native alone is below the number of 70.053, that state is being considered as unsafe. If not, the Crime rate is considered again. If the Crime rate of a state is below 0.025, it is considered as unsafe. If not, it is safe.

1. We run the model by selecting the columns which are Total Crime Rate by State, Sum of Population of Hispanic Origin. By running the model we get Figure-2.

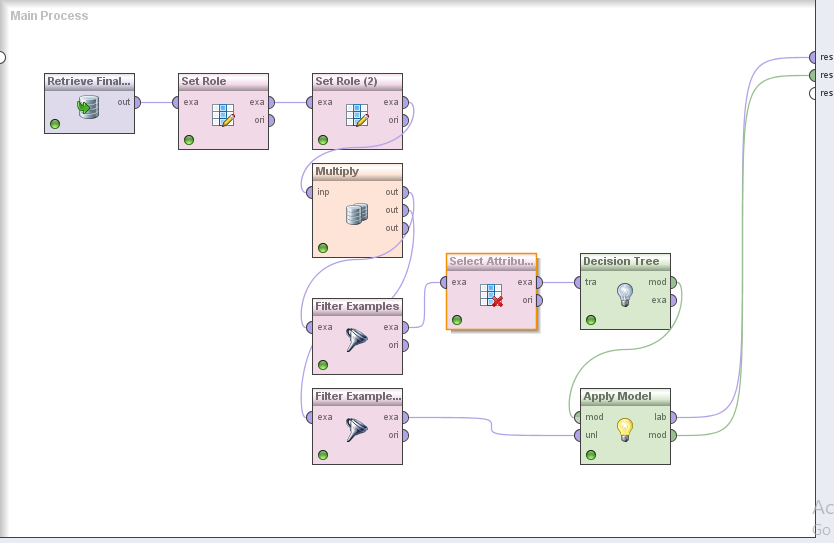


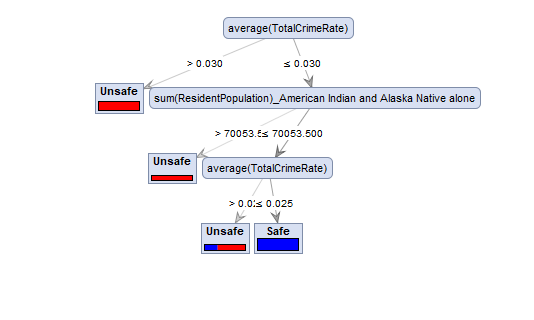
**Figure-2:** The decision tree output of selecting Crime Rate and Hispanic Origin columns.

According to this output, we say that if the crime rate of a state which was below the rate of 0.03 then that state is being considered as unsafe. If not, the attribute of Sum of Resident Population of Hispanic Origin is considered. If the Sum of Resident Population of Hispanic Origin is below the number of 19.527, that state is being considered as safe. If not, again the Sum of Resident Population of Hispanic Origin is considered again. If the Sum of Resident Population of Hispanic Origin of a state is below 32.634, it is considered as unsafe. If not, it is going on as seen in the tree.

**Objective 3: Calibrate models and data**

Now, it is time to apply our model to the test data set. The missing values of SAFE column will be determined according to the decision tree model we created. The **Figure-3** show the model we prepared in Rapid Miner. We multiplied the data set into two, one of them consists of non-missing values of SAFE column that are labeled and the other one is missing values of SAFE column that will be predicted according to the model.



**Figure-3:** The process of applying Decision Tree model to our test data set.

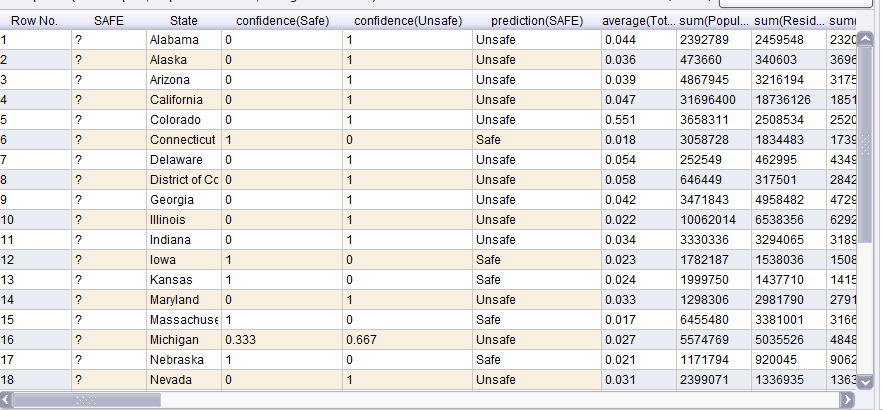
**Figure-4:** The output of Decision Tree model according to our training data set.

**Figure-4** shows the model created according to the labeled data set. According to our labels which are 10 safe states and 10 unsafe states, the model learns by Decision Tree Algorithm that if the Total Crime Rate is below 0.03 then it is unsafe. If not, the resident population of American Indian and Alaska native alone is considered. If the resident population of American Indian and Alaska native alone is below 70053 than it is unsafe. If not, if Total Crime Rate is above 0.025 then it is unsafe, all the other possibilities are safe.

**Objective 4: Document and communicate findings (incl. assumptions, limitations, constraints)**

As our findings are accurate if the data sets reflect the real world. As we projected the years of 2010-2013 only, this problem may be investigated by the comprehensive data sets. By doing this, more sound solutions will, in no doubt, come out. Actually, we showed in our project that this big problem of a society as Crime Findings can be solved or decreased in the future by analyzing the historic data. We can get learned lessons from the historic data. But the limitations will always occur that a society cannot be defined by the numbers %100. There are also some intangible factors affecting the problem. Anyway, it is always good to analyze data and take learned lessons.

**Figure-5** shows the prediction of being SAFE or UNSAFE of the states according to the factors of crime rate, personal income, and census properties like sex, age groups, Hispanic origin and race.



**Figure-5:** The output of Decision Tree model according to our test data set.