**METHODOLOGY(APPROACH) SELECTION**

**Objective 1: Identify potential problem solving approaches (methods)**

There are many data mining algorithms including Naïve-Bayes, Decision Tree, Neural Network, Support Vector Machines etc. We decided to implement Naïve-Bayes at the beginning, but as Decision Tree algorithm is easier to visualize and simple to understand, right now we prefer to implement Decision Tree Algorithm method for modelling process.

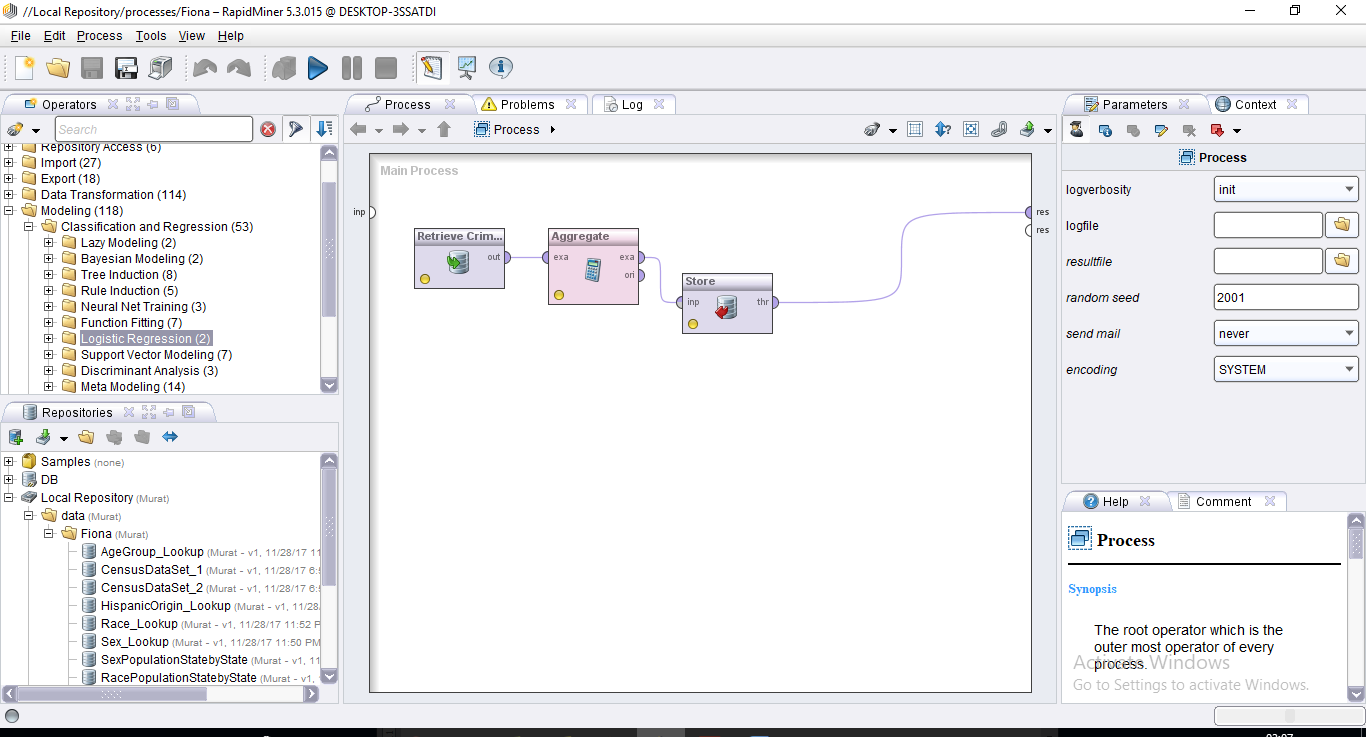
**Objective 2: Select software tools**

We will use Rapid Miner software tool for most of the phases of our project. Rapid Miner is a data mining process tool for implementing CRISP-DM methodology. So we decided to use this tool for our project.

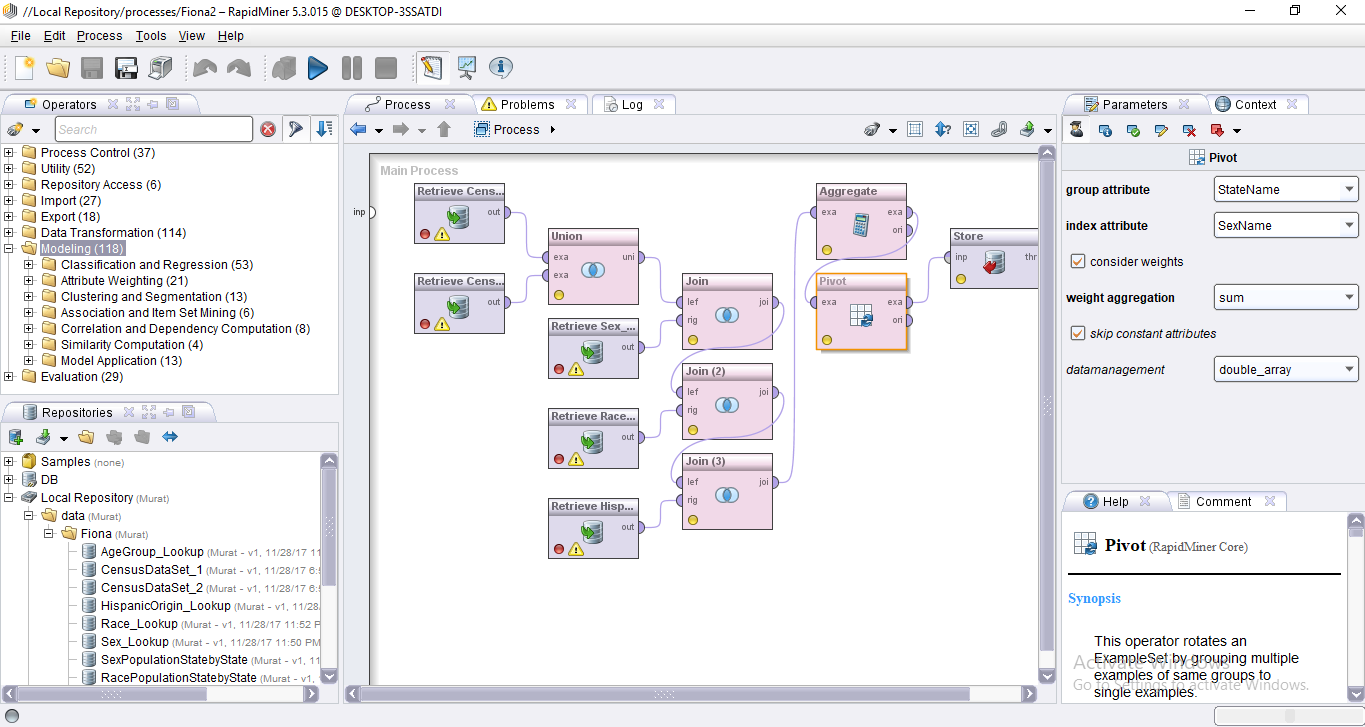
At the beginning, we have three main data sets which are about Crime, Personal Income and Census properties. When we start to analyze and use the data in Rapid Miner, we saw that we need to have some supplementary data sets with them. So we created lookup and SafeStates data sets afterwards. Lookup data set includes some numbers that Personal Income data set has which we explained in the Data document previously. We make the data sensible by using lookup data set. SafeStates data set has the labels of top 10**1** and bottom 10**2** safest states in US. By these data sets we labeled our data before passing into the Decision Tree Algorithm.

As seen in Figure-1 we aggregated the Crime Data by grouping by “State” column and calculated the average of total crime rates.

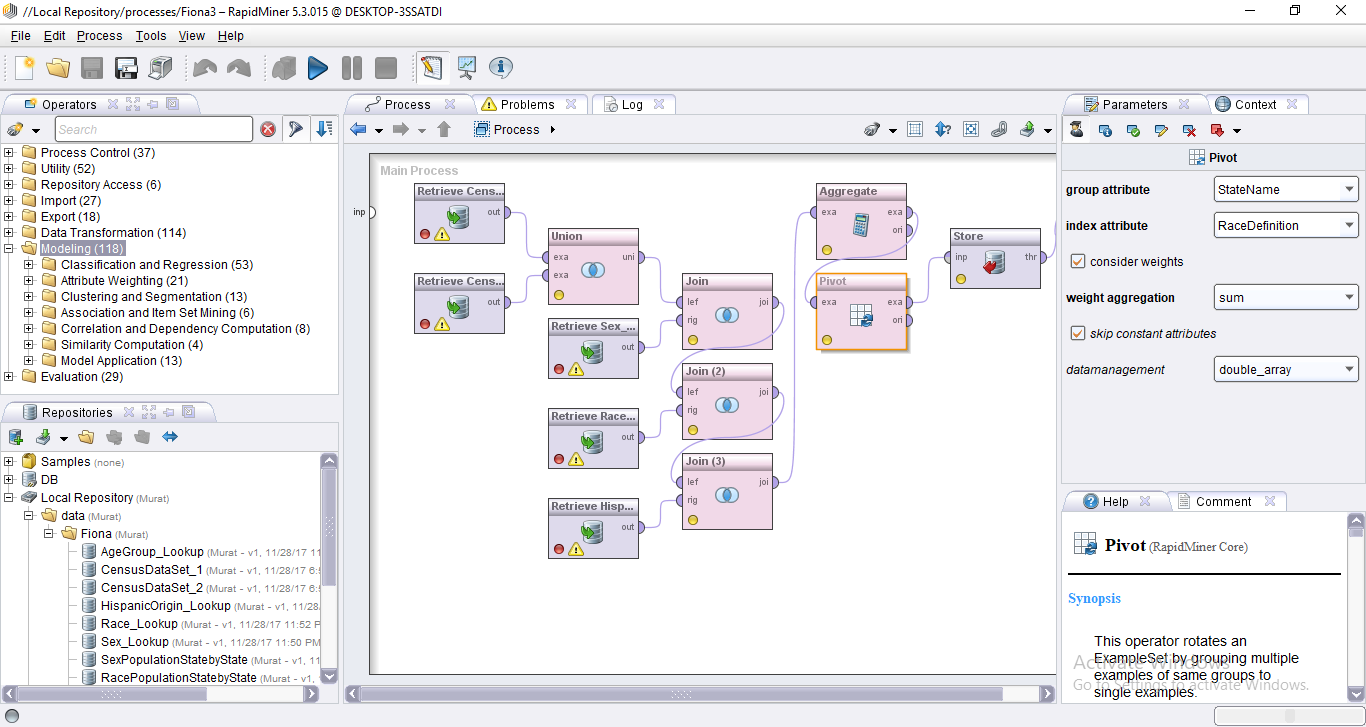
In Figure-2,3,4,5, we played with the Census data. As we have two separate csv files imported to the Rapid Miner repository called “CensusDataset\_1” and “CensusDataSet\_2”, we combined these two data sets by using “Union” operator. Then, we joined the Census data set with the lookup tables of Sex, Origin, Race and Age Group to make them sensible and readable. After aggregating the data by grouping by “State” column and calculating by sum function of the number of sex, origin, race and age groups, we pivoted the data set and have the final data asset as including State column and the number of people including male, female, Hispanic origin, white race etc. accordingly.



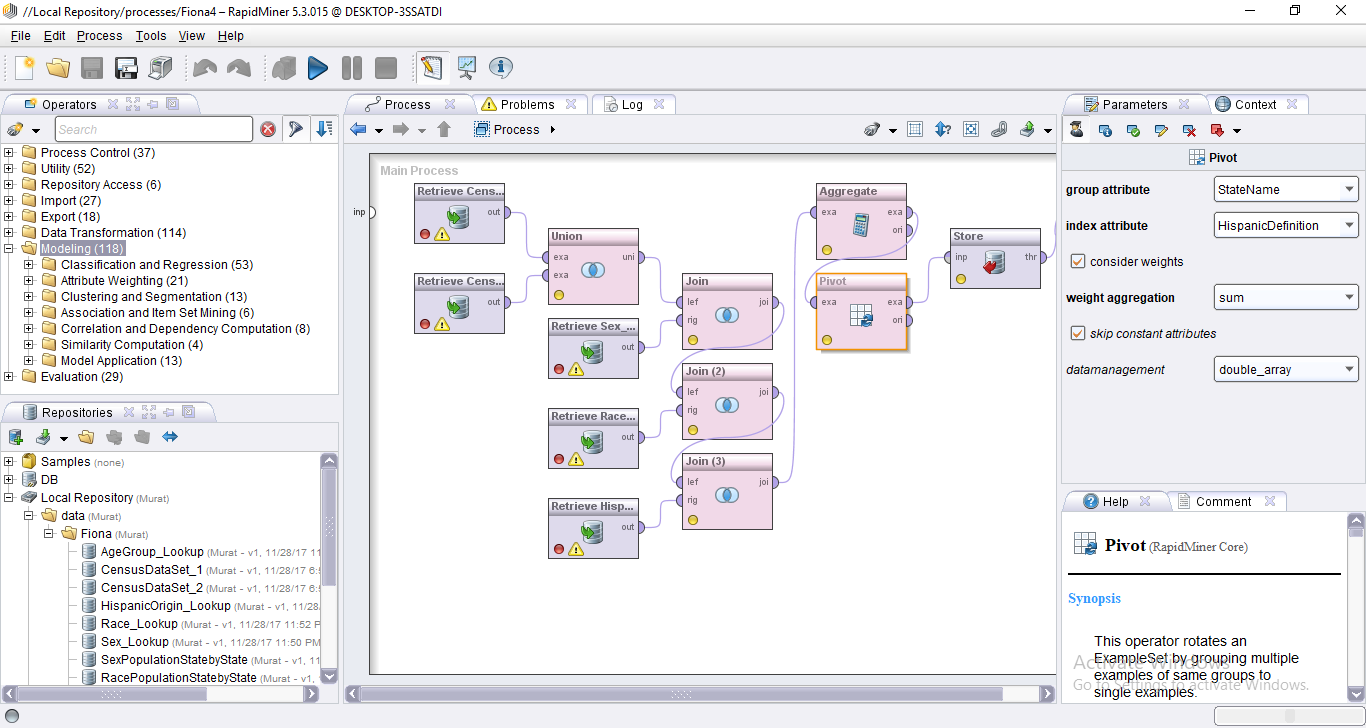
**Figure-1:** The Process of aggregating Crime Data Set in Rapid Miner.



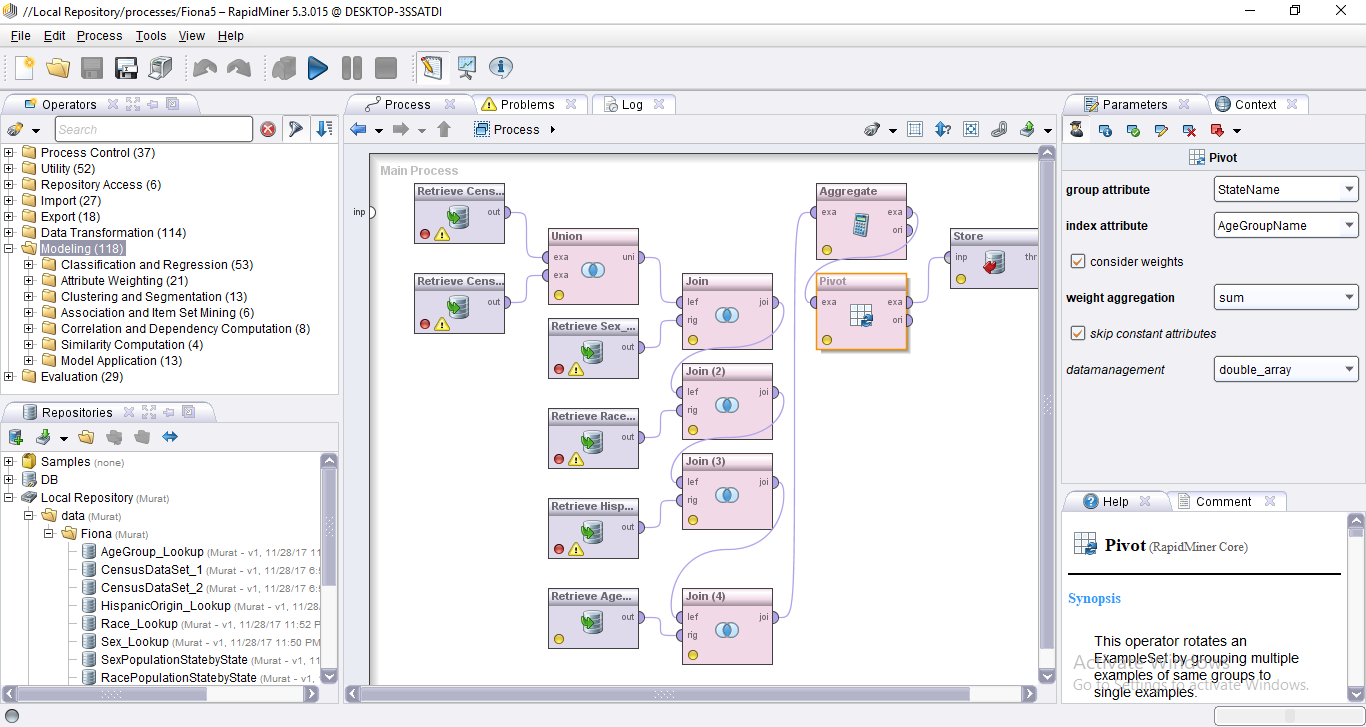
**Figure-2:** The Process of transforming Census Data Set in Rapid Miner (Sex Population).



**Figure-3:** The Process of transforming Census Data Set in Rapid Miner (Race Population).

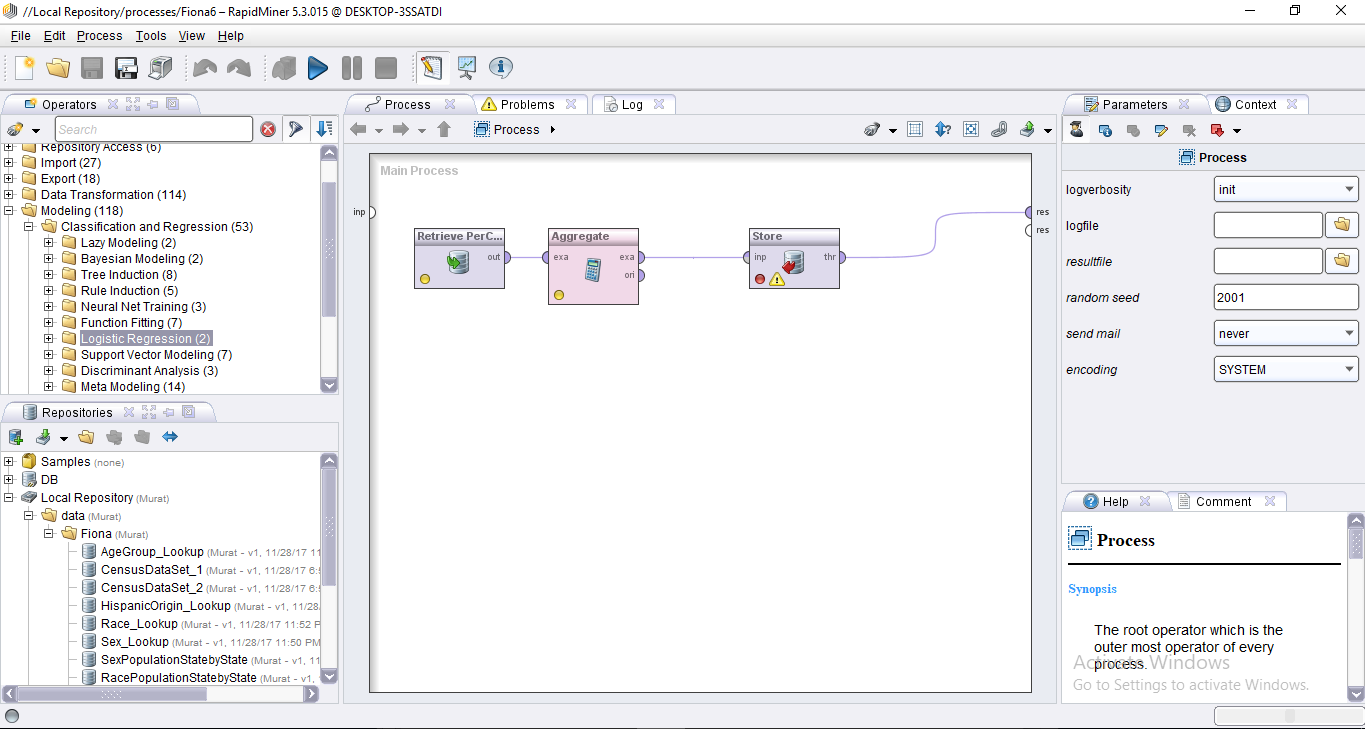


**Figure-4:** The Process of transforming Census Data Set in Rapid Miner (Hispanic Population).



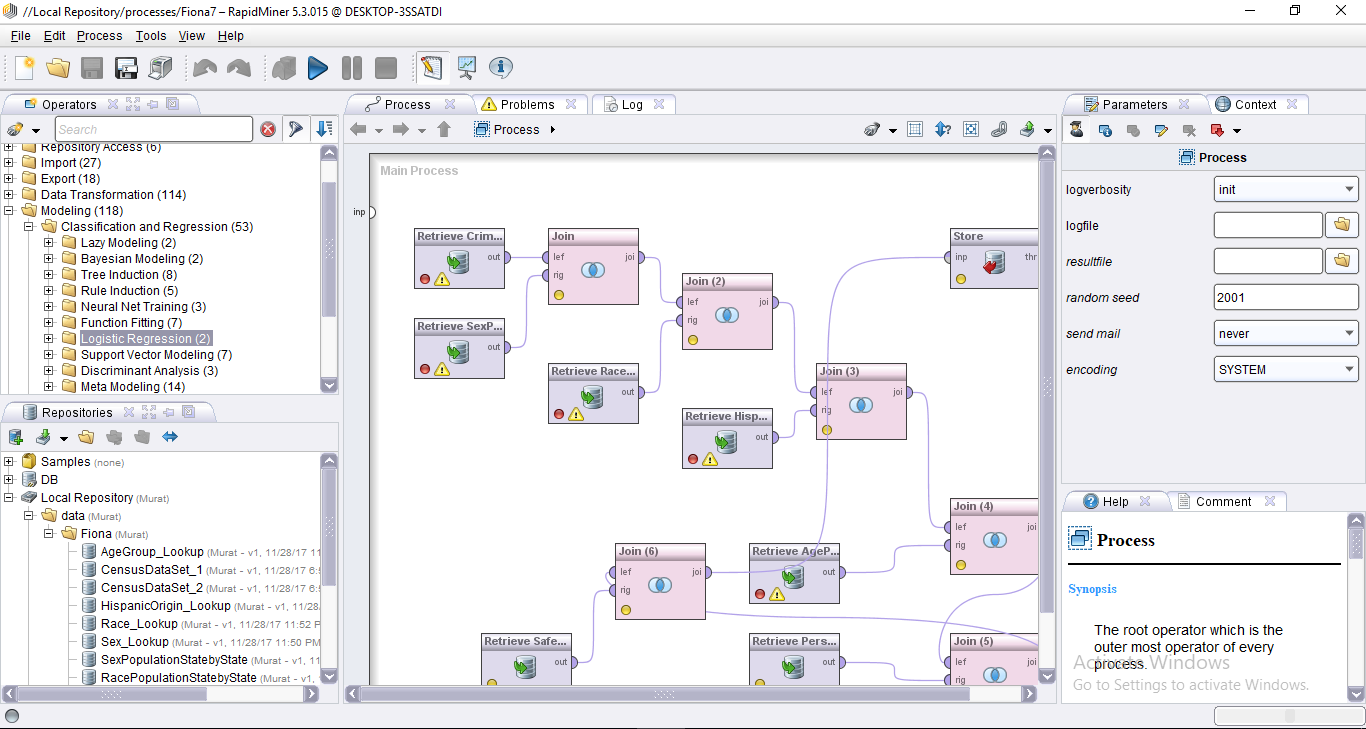
**Figure-5:** The Process of transforming Census Data Set in Rapid Miner (Age Group Population).

As seen Figure-6, we organized the Personal Income data set. We aggregated the data by grouping by “State” column by calculating the per capita Personal Income.



**Figure-6:** The Process of transforming Personal Income Data Set in Rapid Miner.

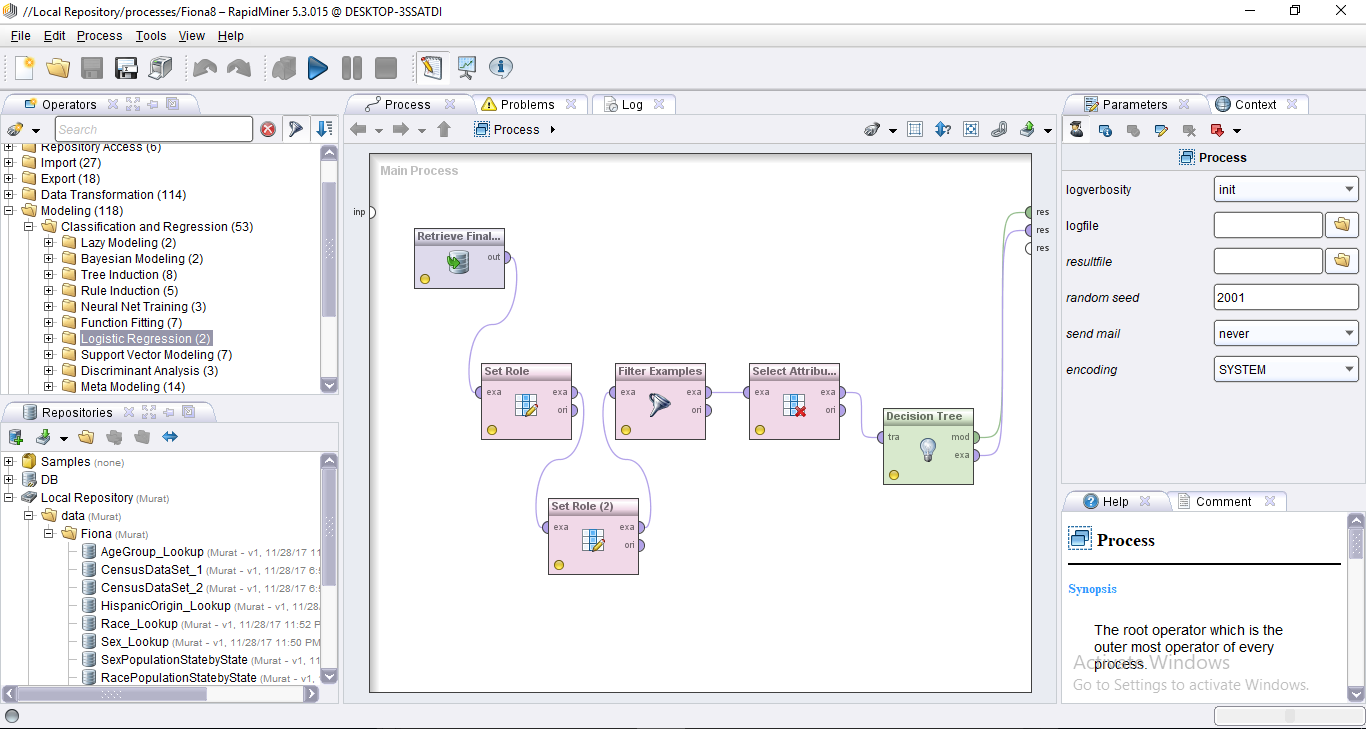
In Figure-7, we created a process of setting up the final data set before passing it to the data mining model algorithms. We combined all the stored data sets by State by using join operators.



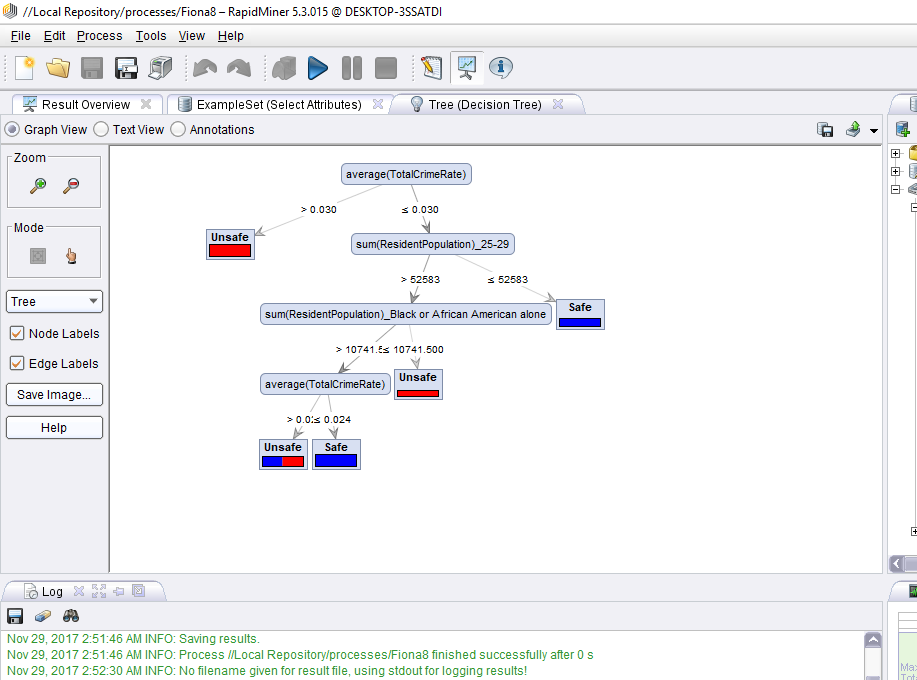
**Figure-7:** The Process of combining all data sets and making an output of final data set.

**Objective 3: Test approaches (methods)**

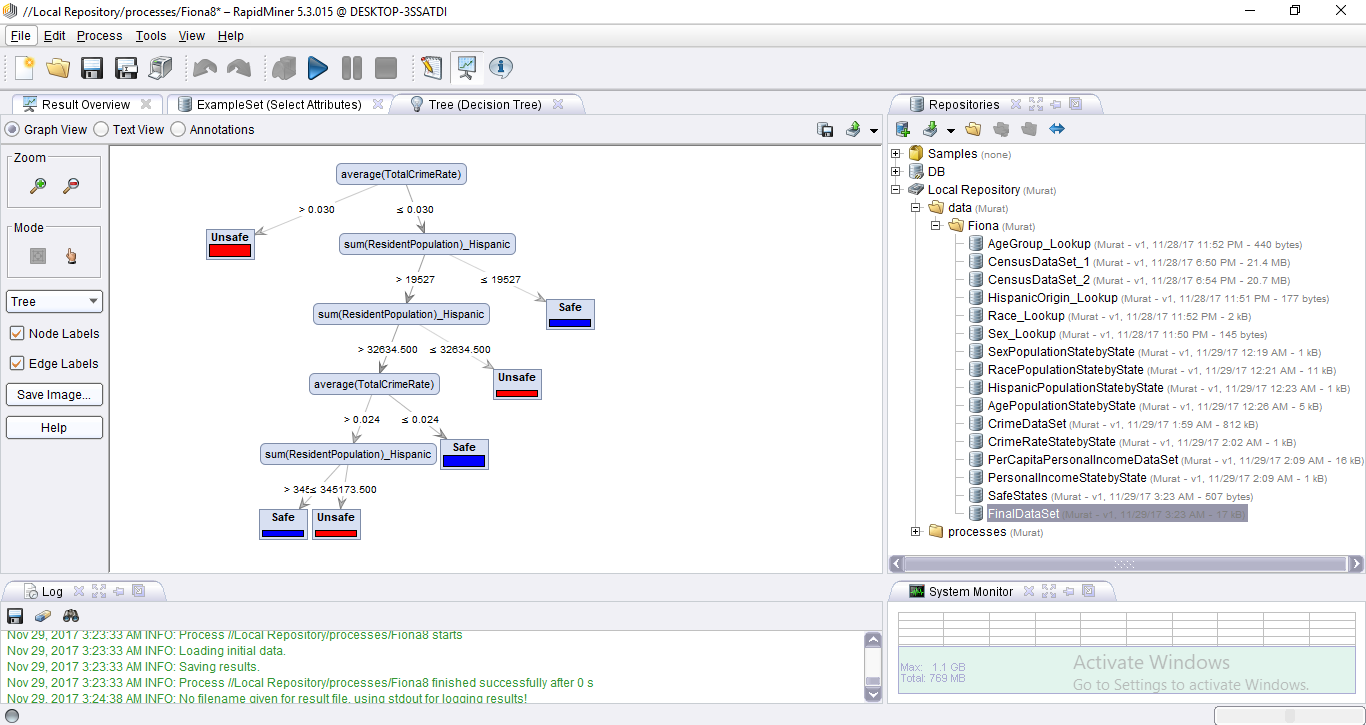
First we tried our final data set in the Decision Tree model. We get promising results as seen in Figure 9,10. In Figure 8, we identified the label as Safe column by Set Role operator. We filter data as non-missing label and select attributes to which we tried to classify. In Figure 9 we chose the attributes of Total Crime Rate, Population of Ages of 25-29, Black or African American Alone. In Figure 10 we chose the attributes of Total Crime Rate and Population of Hispanic origin people. So by iterations we can see the factors of the safety according to the states in the US. This model is easy to understand and so simple to visualize.



**Figure-8:** The Process of testing Decision Tree model.

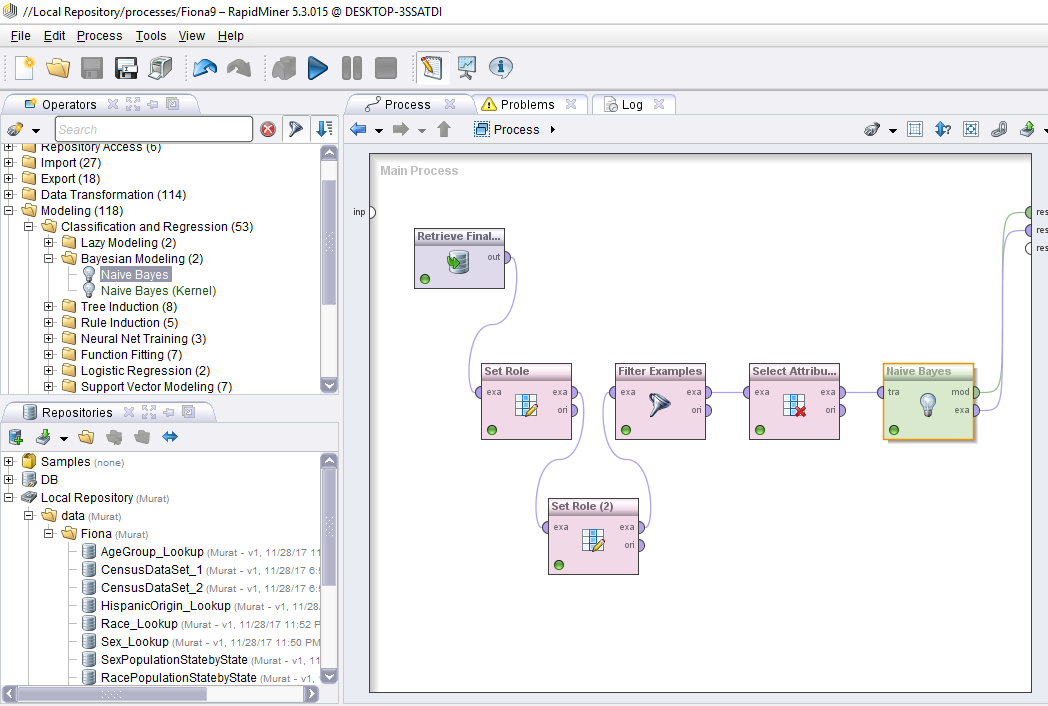


**Figure-9:** The output of Decision Tree algorithm of the attributes of Crime, Age and Race.

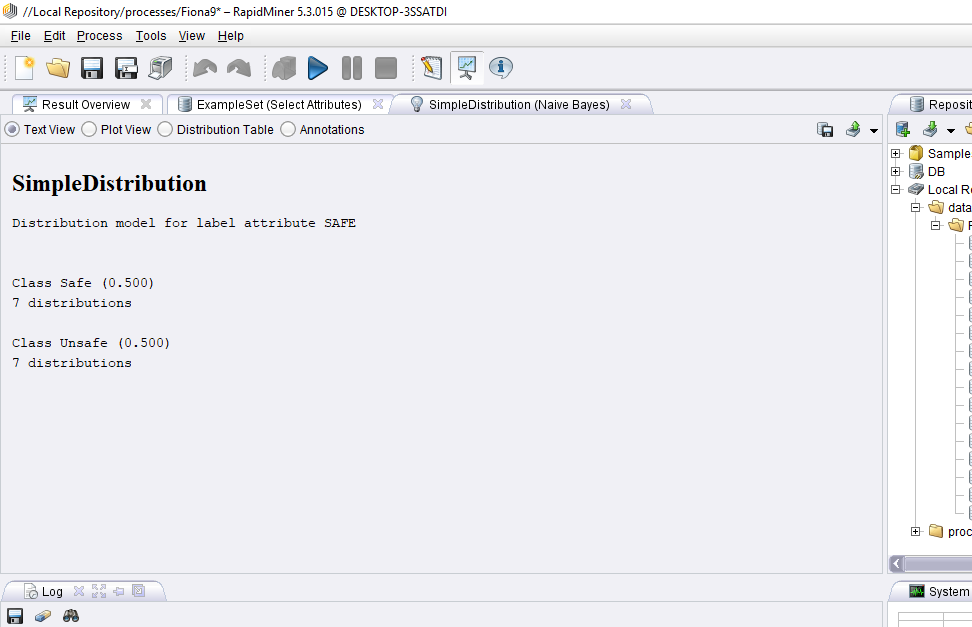


**Figure-10:** The output of Decision Tree algorithm of the attributes of Crime and Hispanic Origin.

Afterwards we tried our final data set in the Naïve Bayes model. We get the results as seen in Figure 12,13. In Figure 11, we identified the label as Safe column by Set Role operator. We filter data as non-missing label and select attributes to which we tried to classify. In Figure 12 we chose the attributes of Total Crime Rate, Population of Ages of 25-29, Black or African American Alone. This model gives the probability of being safe to the State as fifty-fifty.



**Figure-11:** The Process of testing Naïve Bayes model.

**Figure-12:** The output of Naïve Bayes algorithm of the attributes of Crime, Age and Race.

**Objective 4: Select approaches (methods)**

As a conclusion we preferred to use **Decision Tree algorithm method**. The reasons of selecting it are simplicity, easy to understand and evaluating the results of this model will be more sensible and proper.

**References**

**1** [**http://journalstar.com/lifestyles/the-safest-and-least-safe-states-in-america/article\_1be81e00-3b74-52d3-b5b8-3b39b9eae4bd.html**](http://journalstar.com/lifestyles/the-safest-and-least-safe-states-in-america/article_1be81e00-3b74-52d3-b5b8-3b39b9eae4bd.html)

**2** [**https://www.youtube.com/watch?v=6SwU-4UsSOU**](https://www.youtube.com/watch?v=6SwU-4UsSOU)