Aim: Experiment to explore Rapid Miner and implement classification models like Decision Tree and Naive Bayes etc.

To do:

1. Preprocess data. Split data into train and test set
2. Build Classification model using Rapid miner on training data (use decision tree and naive Bayes method)
3. Calculate metrics based on test data
4. Compare the models based on metrics and find which model is best suited for the dataset.

Link of Dataset:

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease.
https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

Theory:

Rapidminer is a comprehensive data science platform with visual workflow design and full automation. It means that we don't have to do the coding for data mining tasks. Rapidminer is one of the most popular data science tools.

This is the graphical user interface of the blank process in rapidminer. It has the repository that holds our dataset. We can import our own datasets. It also offers many public datasets that we can try. We can also work with a database connection.

Below the repository window, it has an operator. The operators include everything we need to build a data mining process, such as data access, data cleansing, modeling, validation, and scoring.

Activity Selection

This is the first interface that will pop up when we launch the rapidminer application. The blank process is to build from scratch. It works by dragging-and-dropping operators to the process field manually. This is the menu that you want to choose if you are at the intermediate level with this application.

Turbo Prep is for dataset preparation only. It includes transform, cleaning, and combines datasets. Auto Model will bring us the wizard to perform data mining tasks. Just like installing an application on Windows. Next-Next and Finish.
It also has a lot of templates for us to start with. We will choose the Auto Model for this study case.

Importing Dataset

Here we can choose the dataset that we will use. We can import our own dataset or select from the available dataset provided by rapidminer. Import-New-Data button below the select data list is to import our own dataset.
To use the available dataset from Rapidminer, click on the samples folder, then expand the data folder, and let's choose the Titanic dataset for our study case and click the green Next button.
Notice in the progress bar there are only six easy steps to perform a data mining task with rapidminer.

Data Mining Method Selection

The details from a chosen dataset will be displayed. The titanic is a dataset to predict whether the passenger will survive on the titanic ship from the available input parameters. There are eleven input (x) parameters and one label (y) from this dataset.
There are three actions that we can choose for our dataset. Predict, clusters and outliers. The outliers button will help us detect outliers in our data. Clusters will help us detect common groups in our data. Predict will classify the data from the given input parameter.
Here we can observe our dataset's input parameter. We can see that the Titanic dataset consists of both categorical and numerical data. The target label is in categorical, yes, or no format.
Select the predict button to do classification, select the Survived column as a label or classification target, and click on the Next button.

Data Balance

After selecting the data mining method and selecting the target column, we will be served with the data balance in a chart. See that the No data is more than the Yes data. This condition is quite common in reality. The ratio is around 60:40 which is still acceptable.
We need to start to worry when the ratio is higher than 70:30. A highly imbalanced class will lead to imbalanced prediction. The classification is usually predicted to the majority class.

Input Selection

In this section, we can exclude the columns from the input parameters. The default is that all columns are included. Rapidminer will recommend which columns should be included or excluded.

Notice that the first three rows are excluded by default. This happens because the status is red. The red status will automatically be excluded by rapidminer though we still can include it. You can hover the red circle in the status column to see the details.

Classification is defined as the process of recognition, understanding, and grouping of objects and ideas into preset categories a.k.a "sub-populations." With the help of these pre-categorized training datasets, classification in machine learning programs leverage a wide range of algorithms to classify future datasets into respective and relevant categories.

Based on training data, the Classification algorithm is a Supervised Learning technique used to categorize new observations. In classification, a program uses the dataset or observations provided to learn how to categorize new observations into various classes or groups. For instance, 0 or 1, red or blue, yes or no, spam or not spam, etc. Targets, labels, or categories can all be used to describe classes. The Classification algorithm uses labeled input data because it is a supervised learning technique and comprises input and output information. A discrete output function (y) is transferred to an input variable in the classification process (x).

In simple words, classification is a type of pattern recognition in which classification algorithms are performed on training data to discover the same pattern in new data sets.

Types of Classification Algorithms

You can apply many different classification methods based on the dataset you are working with. It is so because the study of classification in statistics is extensive. The top five machine learning algorithms are listed below.

1. Logistic Regression

It is a supervised learning classification technique that forecasts the likelihood of a target variable. There will only be a choice between two classes. Data can be coded as either one or yes, representing success, or as 0 or no, representing failure. The dependent variable can be predicted most effectively using logistic regression. When the forecast is categorical, such as true or false, yes or no, or a 0 or 1, you can use it. A logistic regression technique can be used to determine whether or not an email is spam.

2. Naive Bayes
Naive Bayes determines whether a data point falls into a particular category. It can be used to classify phrases or words in text analysis as either falling within a predetermined classification or not.

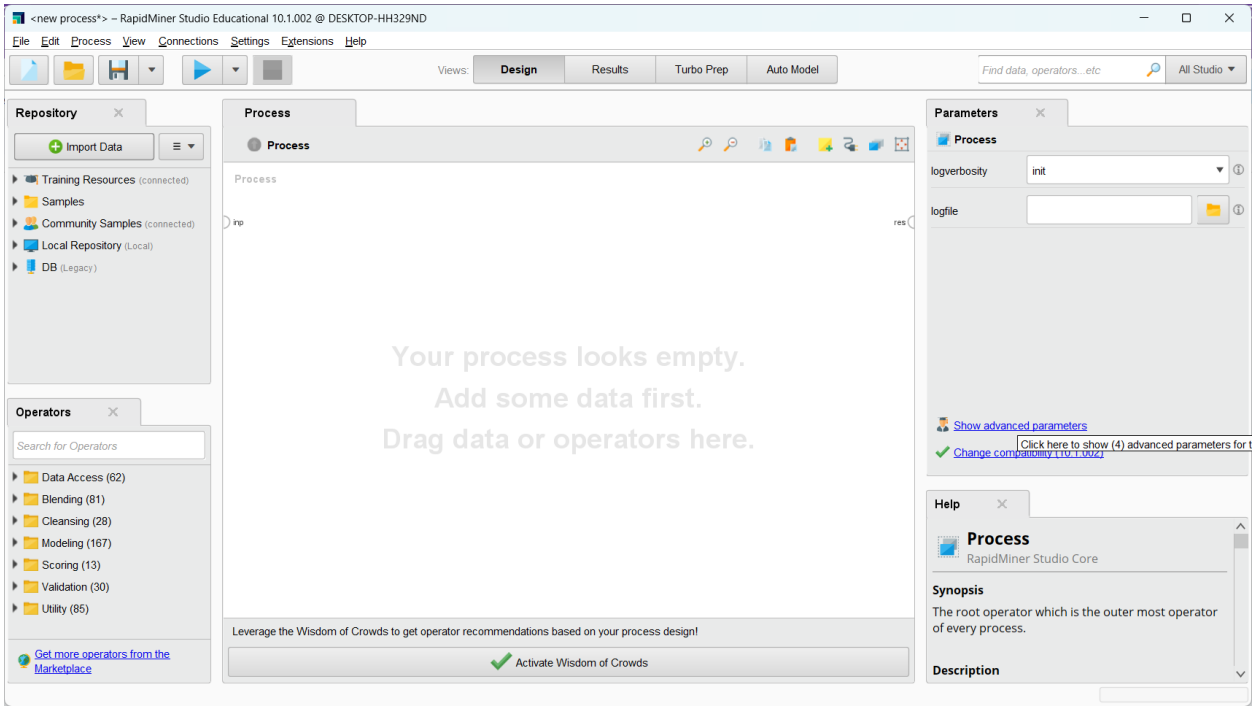
3. K-Nearest Neighbors
It calculates the likelihood that a data point will join the groups based on which group the data points closest to it are a part of. When using k-NN for classification, you determine how to classify the data according to its nearest neighbor.

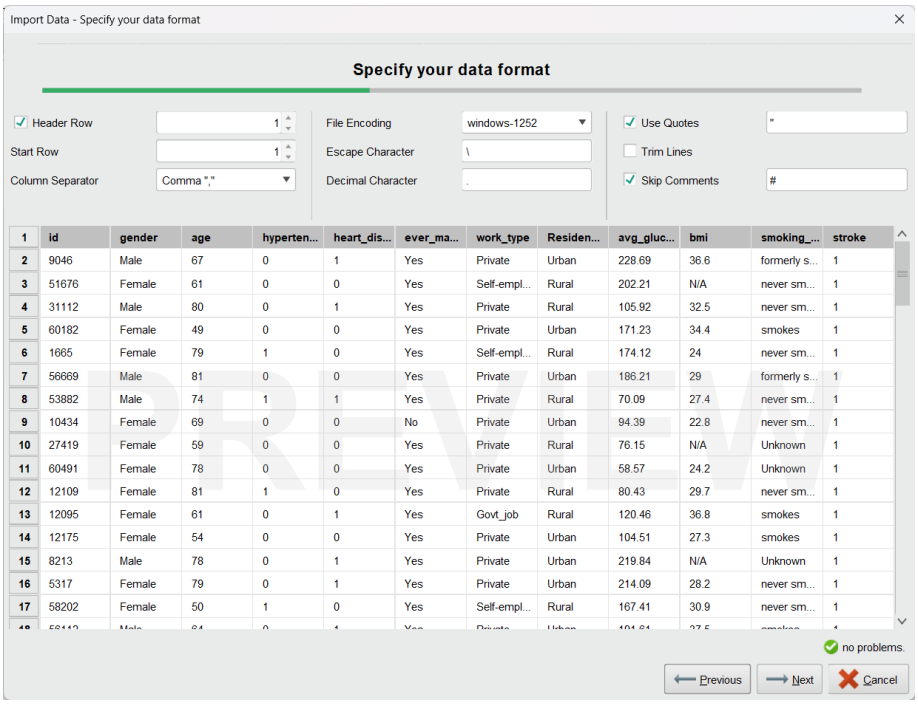
4. Decision Tree
A decision tree is an example of supervised learning. Although it can solve regression and classification problems, it excels in classification problems. Similar to a flow chart, it divides data points into two similar groups at a time, starting with the "tree trunk" and moving through the "branches" and "leaves" until the categories are more closely related to one another.

5. Random Forest Algorithm
The random forest algorithm is an extension of the Decision Tree algorithm where you first create a number of decision trees using training data and then fit your new data into one of the created ‘trees’ as a ‘random forest’. It averages the data to connect it to the nearest tree data based on the data scale. These models are great for improving the decision tree’s problem of forcing data points unnecessarily within a category.

6. Support Vector Machine
Support Vector Machine is a popular supervised machine learning technique for classification and regression problems. It goes beyond X/Y prediction by using algorithms to classify and train the data according to polarity.

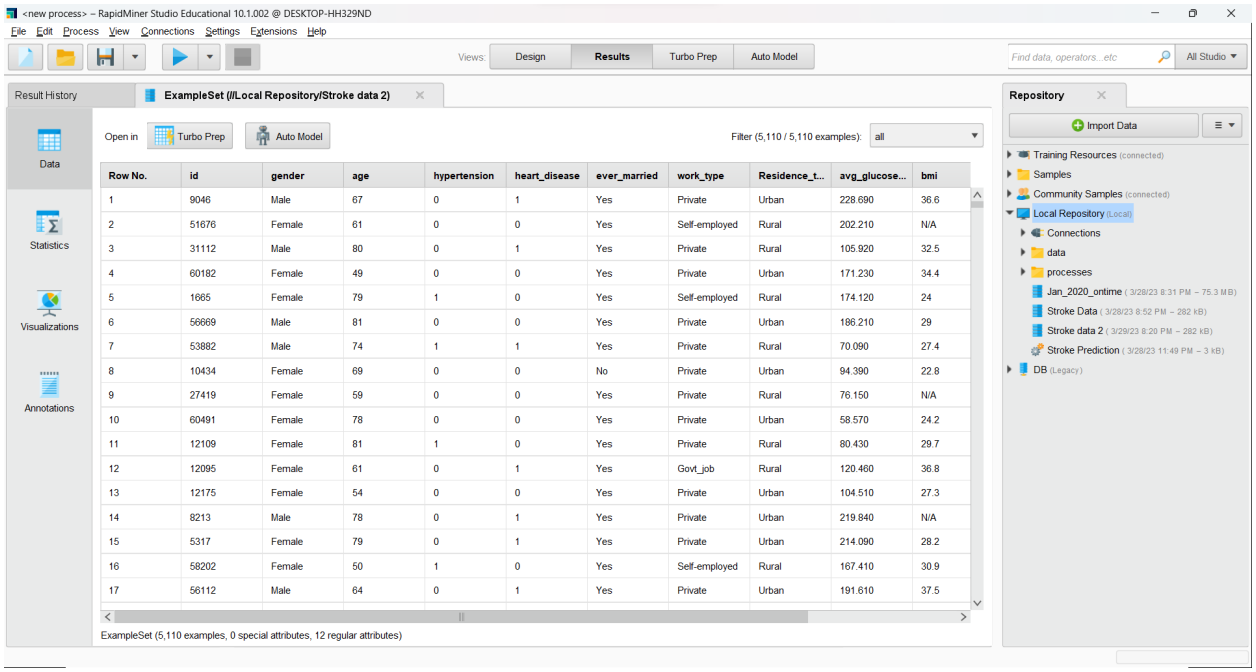
Implementation of Decision tree in Rapid Minier
Step 1 : Import the dataset from your local directory



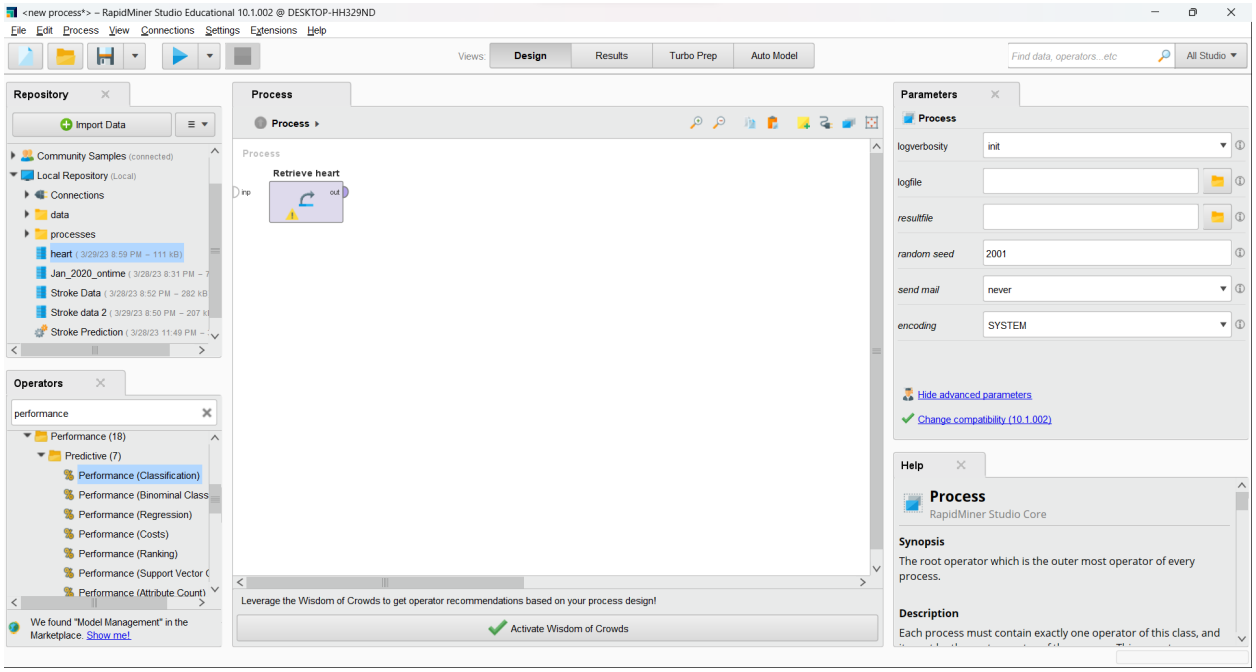


	id	gender	age	hyperten...	heart_dis...	ever_ma...	work_type	Residen...	avg_gluc...	bmi	smoking_...	stroke
2	9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly s...	1
3	51676	Female	61	0	0	Yes	Self-empl...	Rural	202.21	N/A	never sm...	1
4	31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never sm...	1
5	60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
6	1665	Female	79	1	0	Yes	Self-empl...	Rural	174.12	24	never sm...	1
7	56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly s...	1
8	53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never sm...	1
9	10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never sm...	1
10	27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
11	60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12	12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never sm...	1
13	12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
14	12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
15	8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1
16	5317	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never sm...	1
17	58202	Female	50	1	0	Yes	Self-empl...	Rural	167.41	30.9	never sm...	1

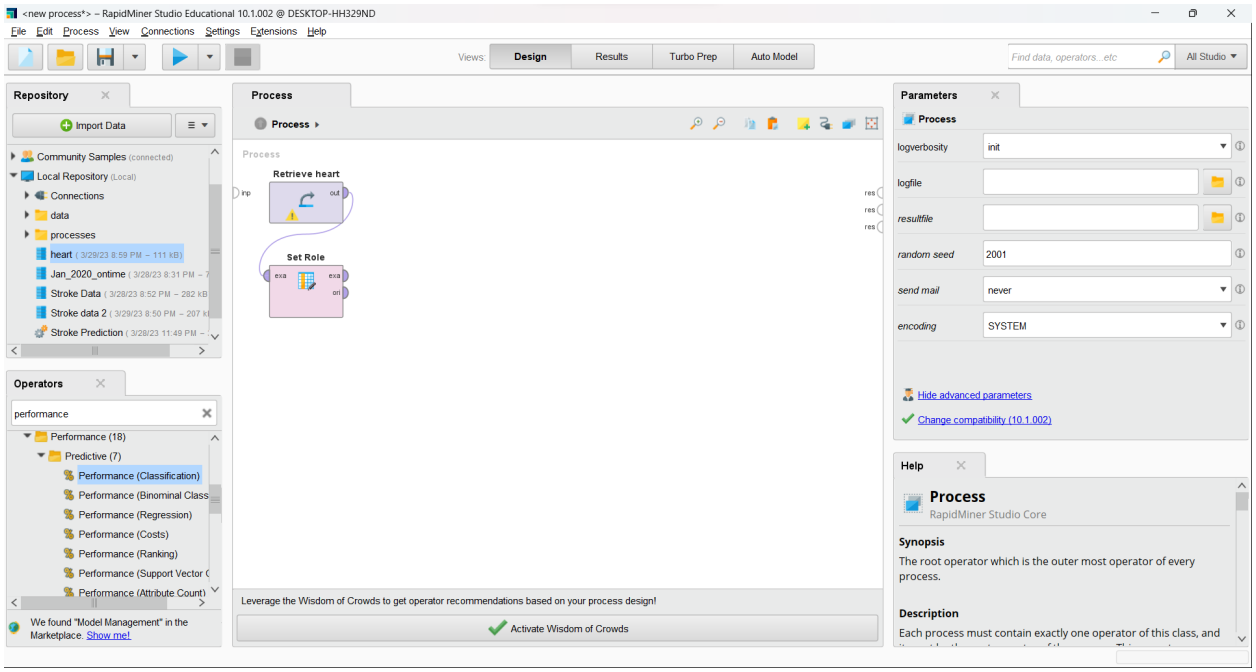
PREVIEW



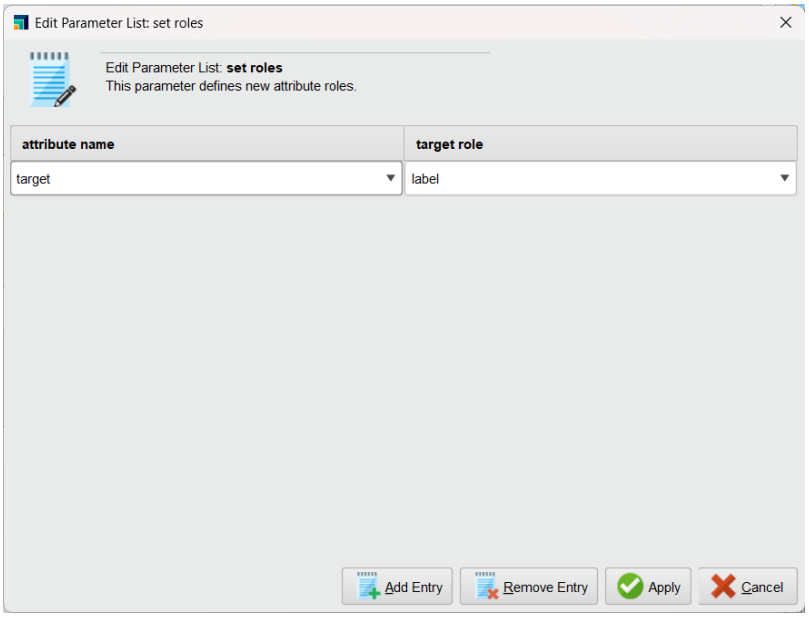
Step 2:Drag and drop the desired data set



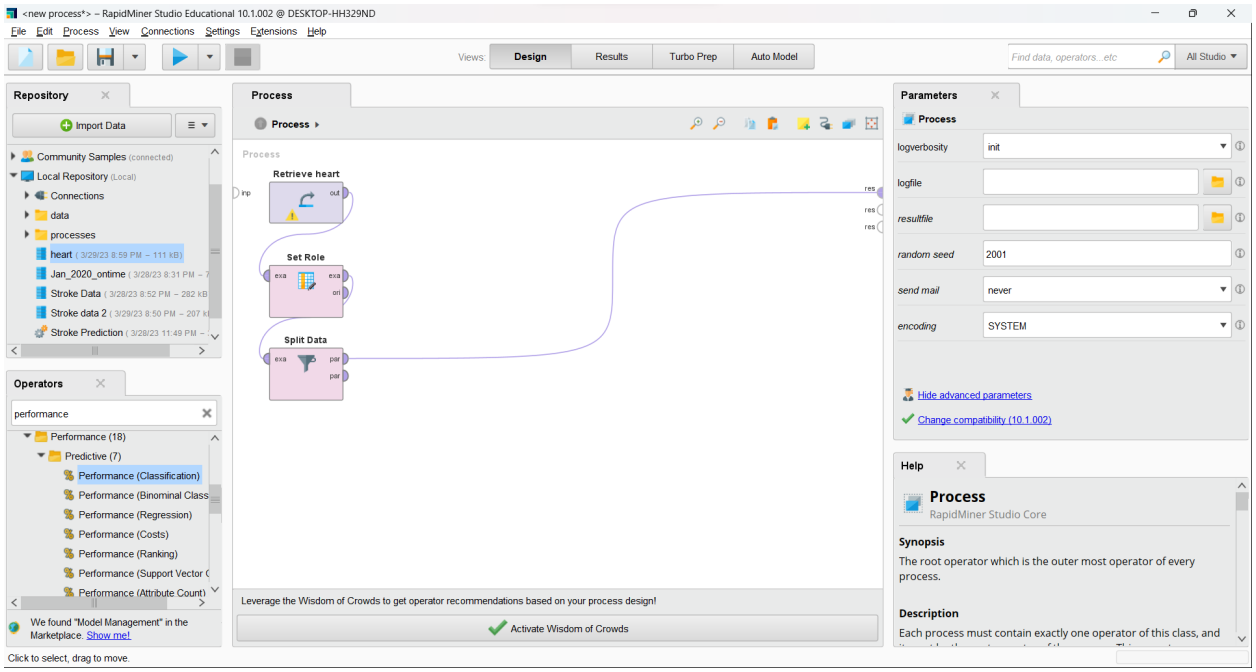
Step 3:Now search Set role and then drag and drop it



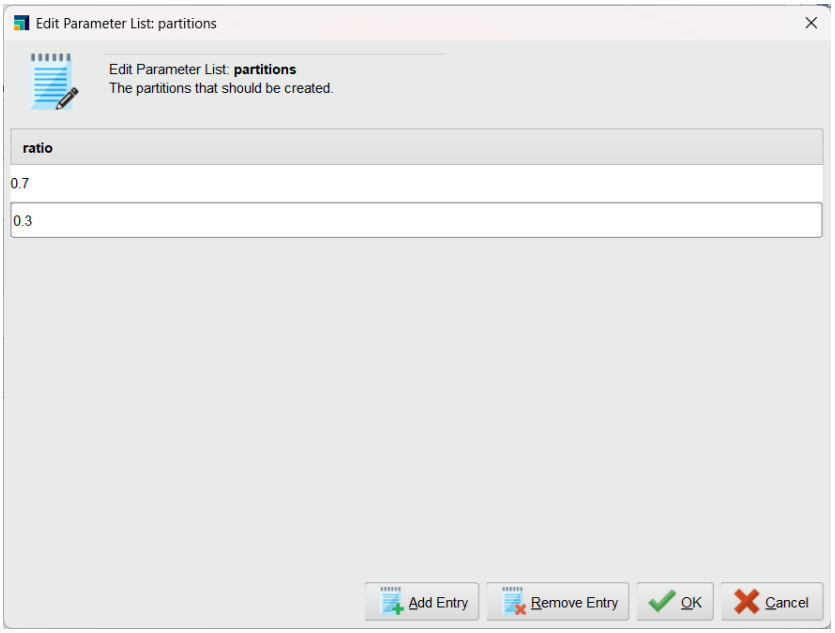
And after than set the role as target variable



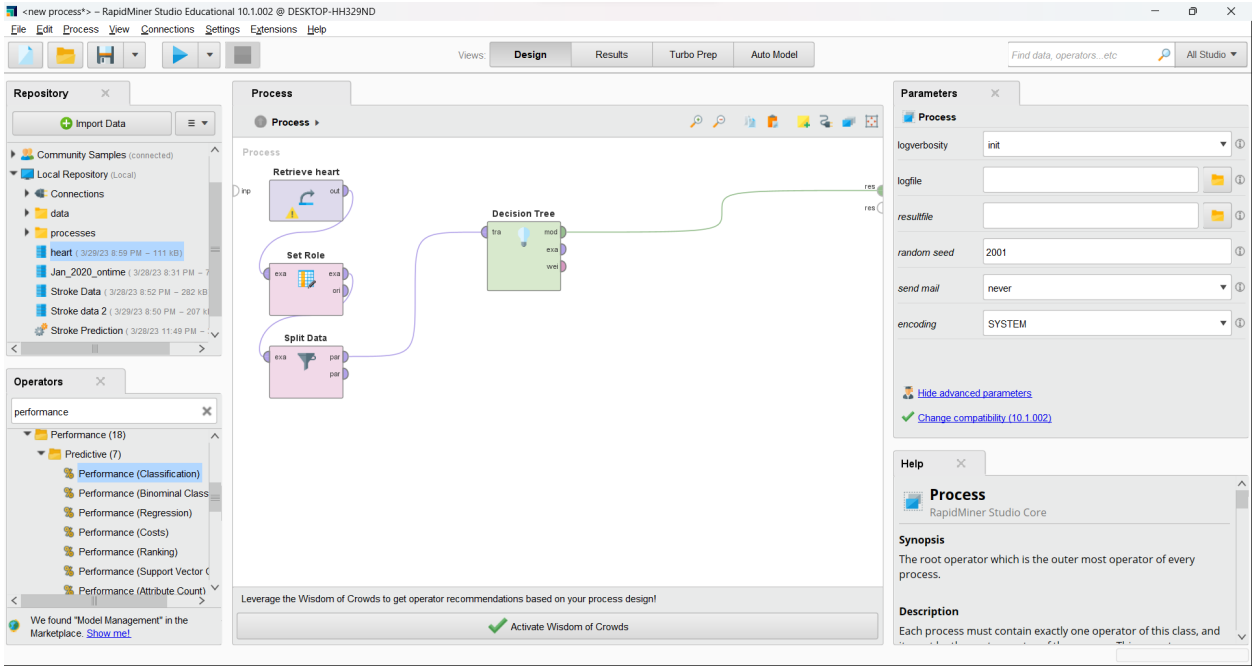
Step 4: As now we want to split the data we will use the split operator



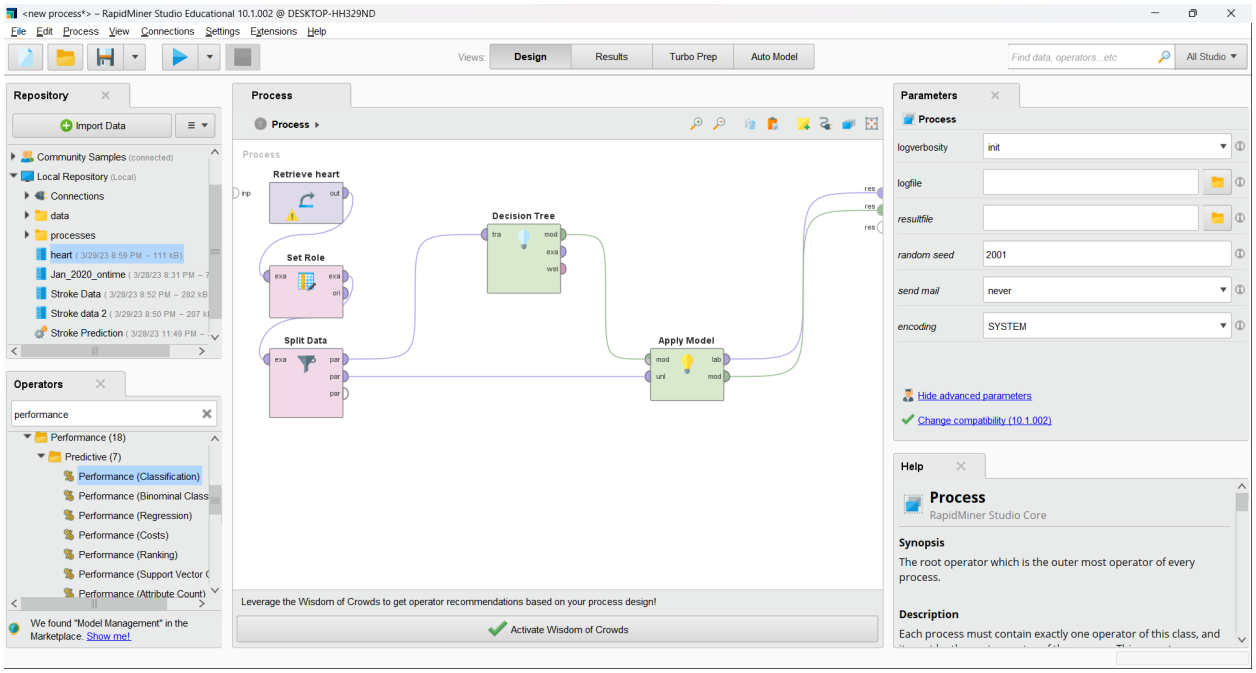
Step 5:Splitting the data in the ratio of 70 as to 30%



Now search for Decision tree operator and drag and drop it

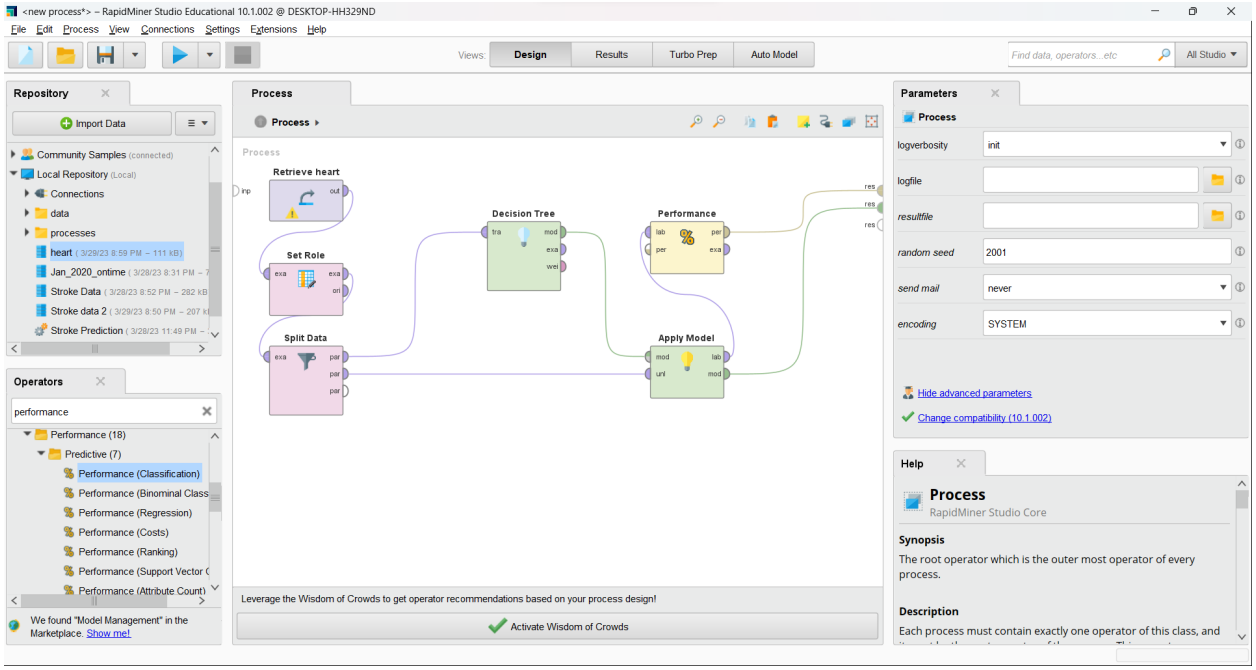


Now add apply model operator as shown below

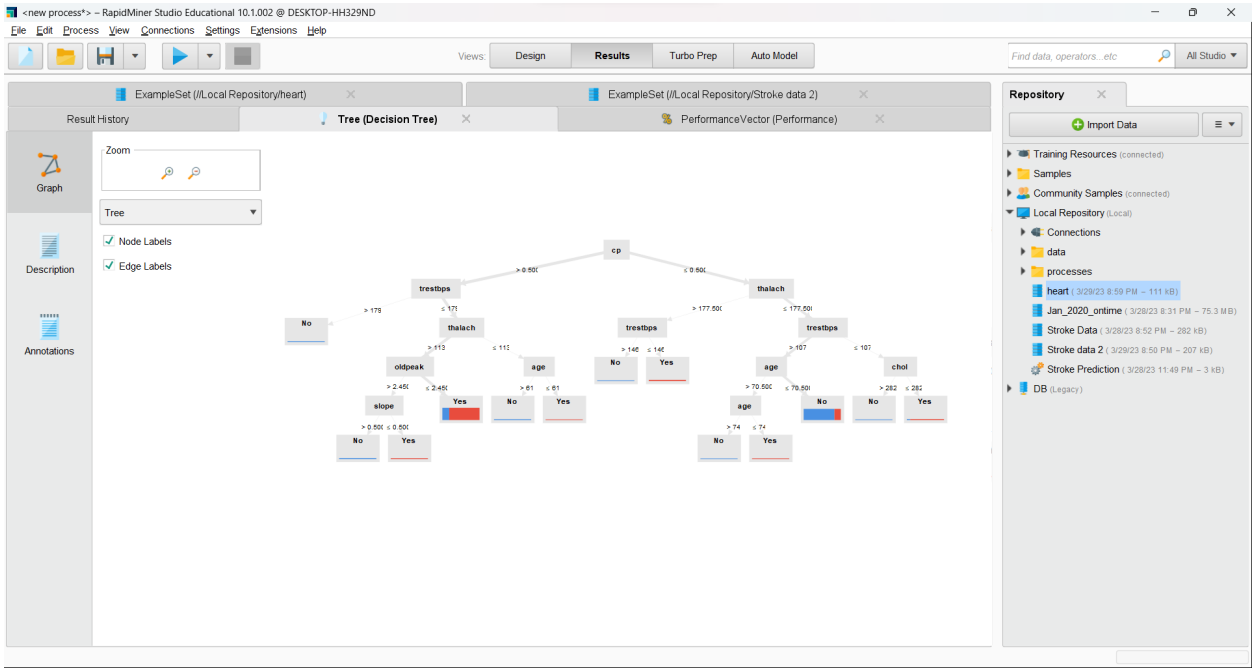


Step 6: Now add Performance measure operator and then connect it as shown below

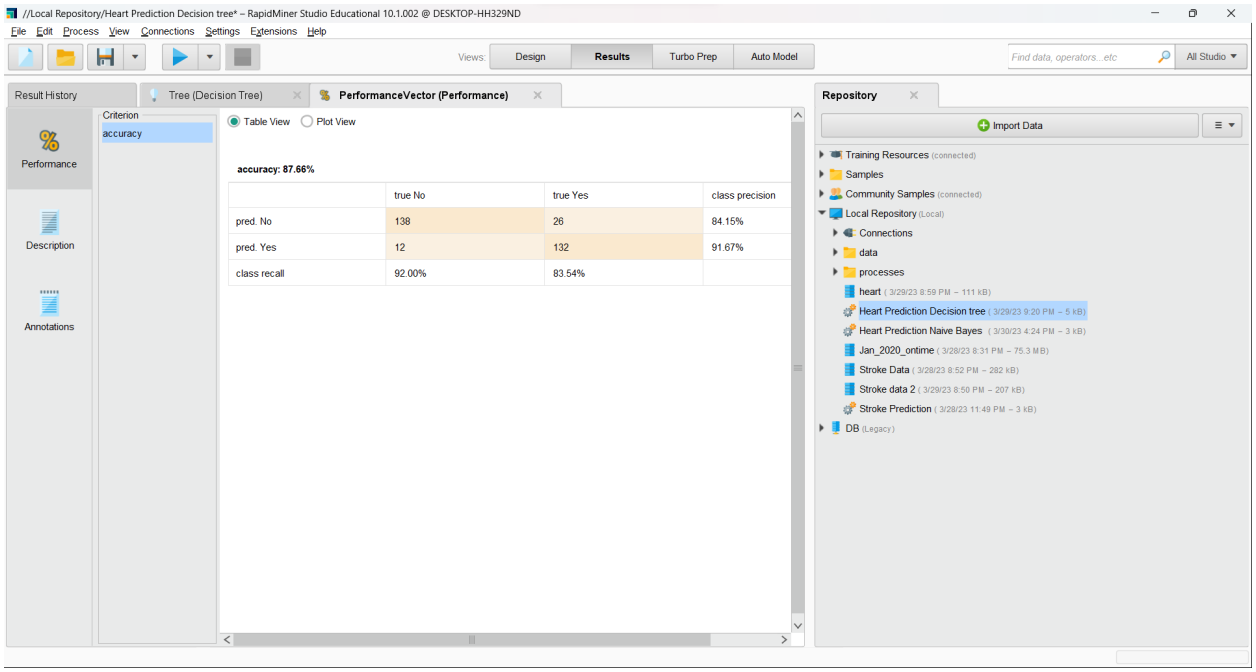
And also do the connection for model as shown below



After running the tree will be visible as below



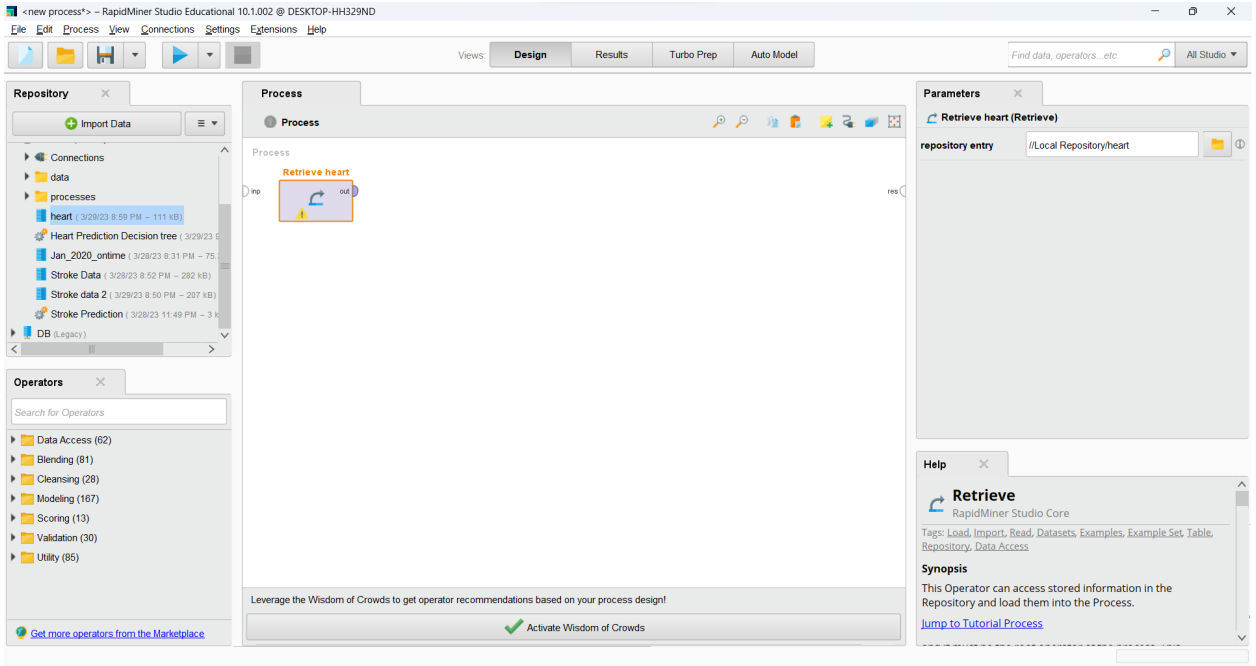
This is the decision tree formed by the given dataset



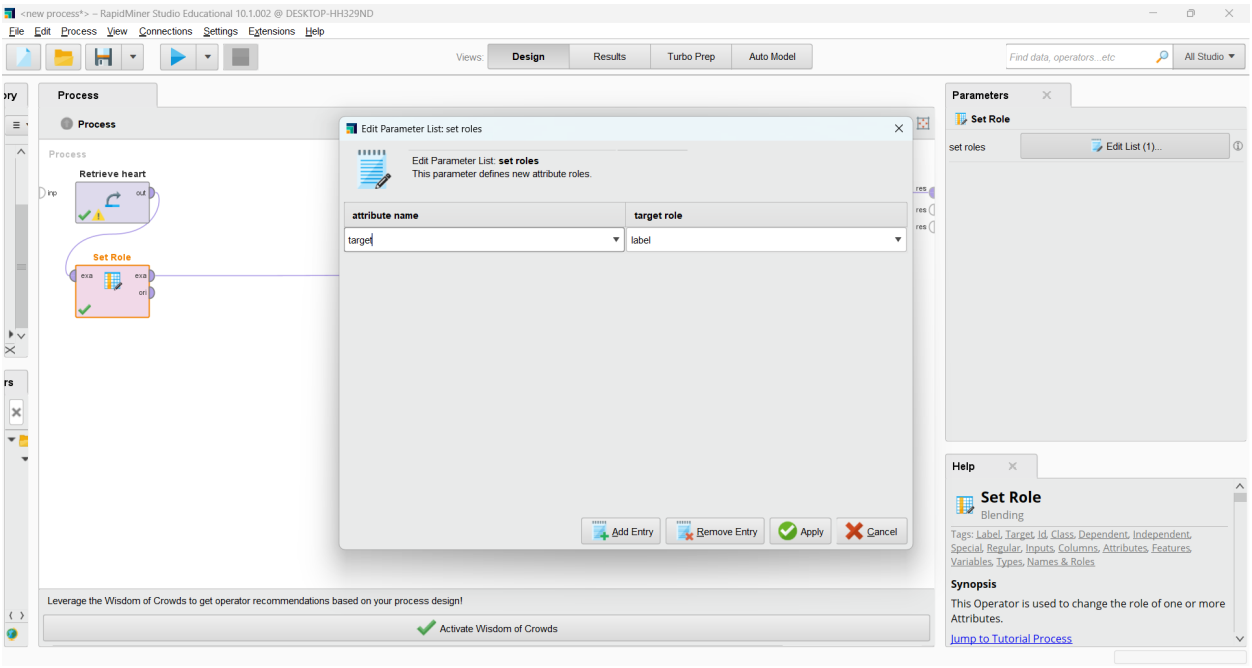
The overall **accuracy of Decision Tree is 87.66%**

Implementation of Naive Bayes in Rapid Minier

Step 1:Import the dataset using drag and drop

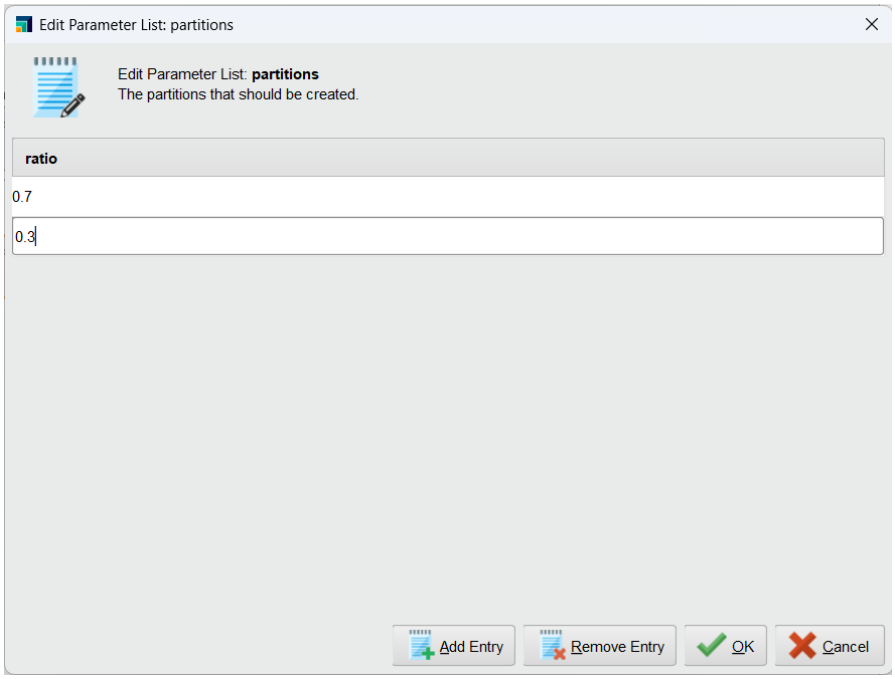


Here we set our target variable as target and target role as label

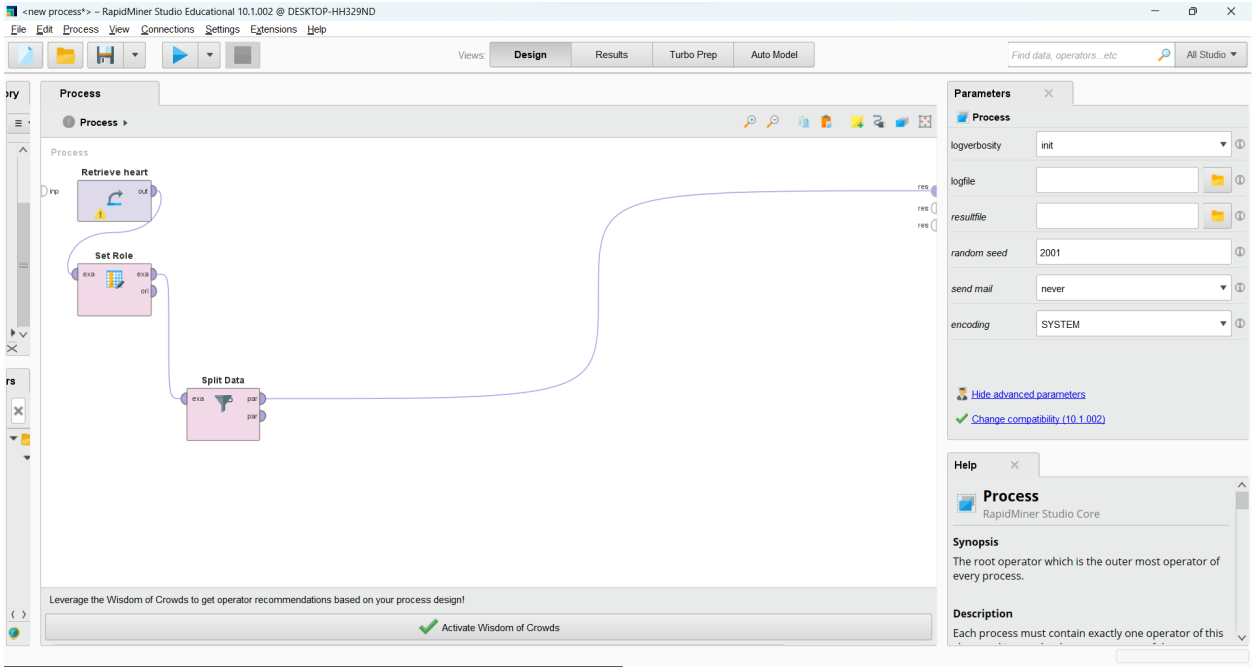


Step 2: Now we need to split the dataset into two parts that is Training and Testing dataset

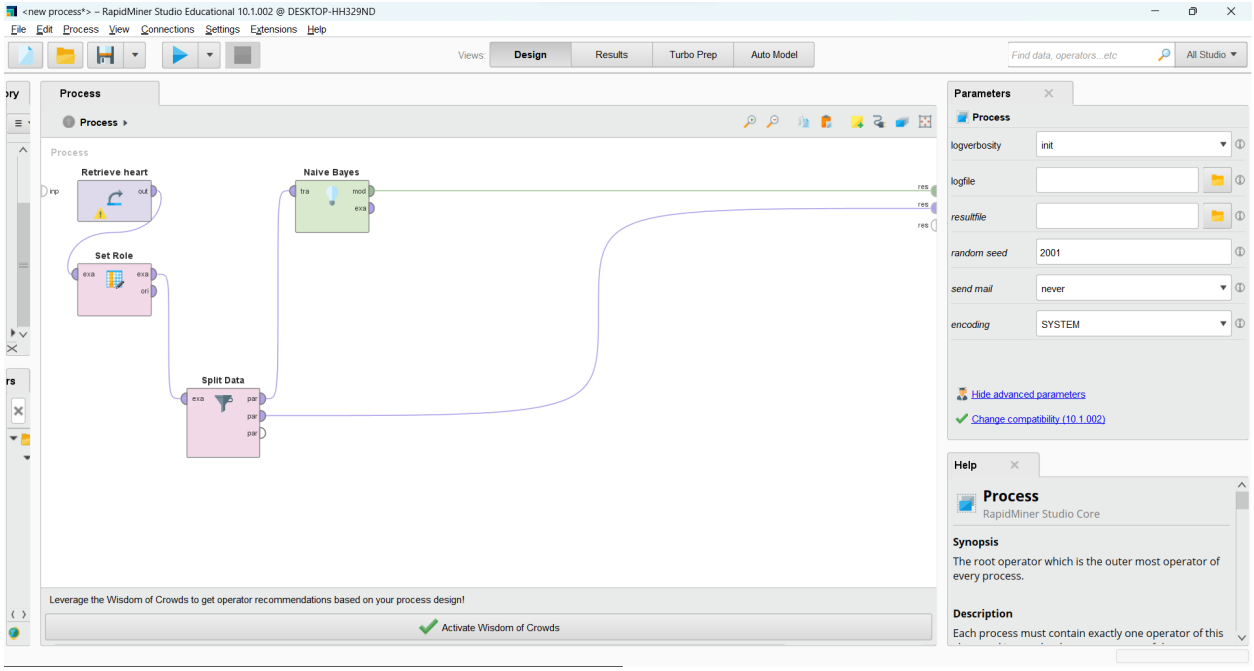
So in order to perform that we need to add split data operator and split data into 70% and 30%

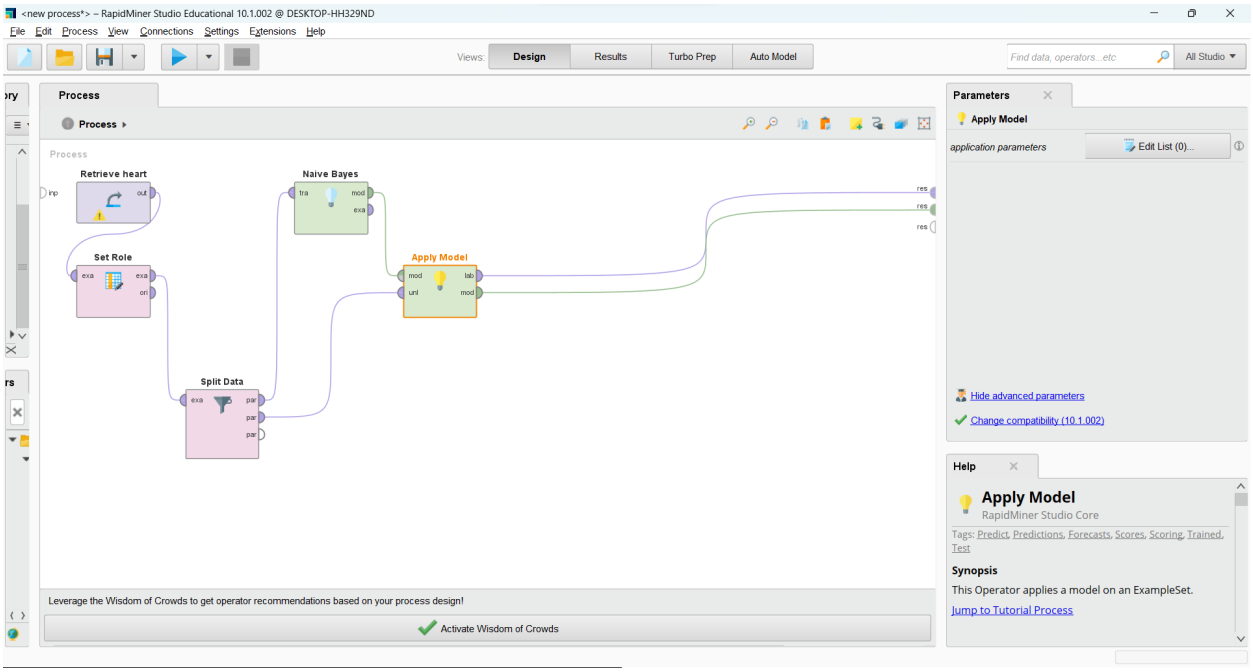


Step 3:Now for applying naive bayes on the dataset, we will drag and drop the Naive Bayes operator as shown below



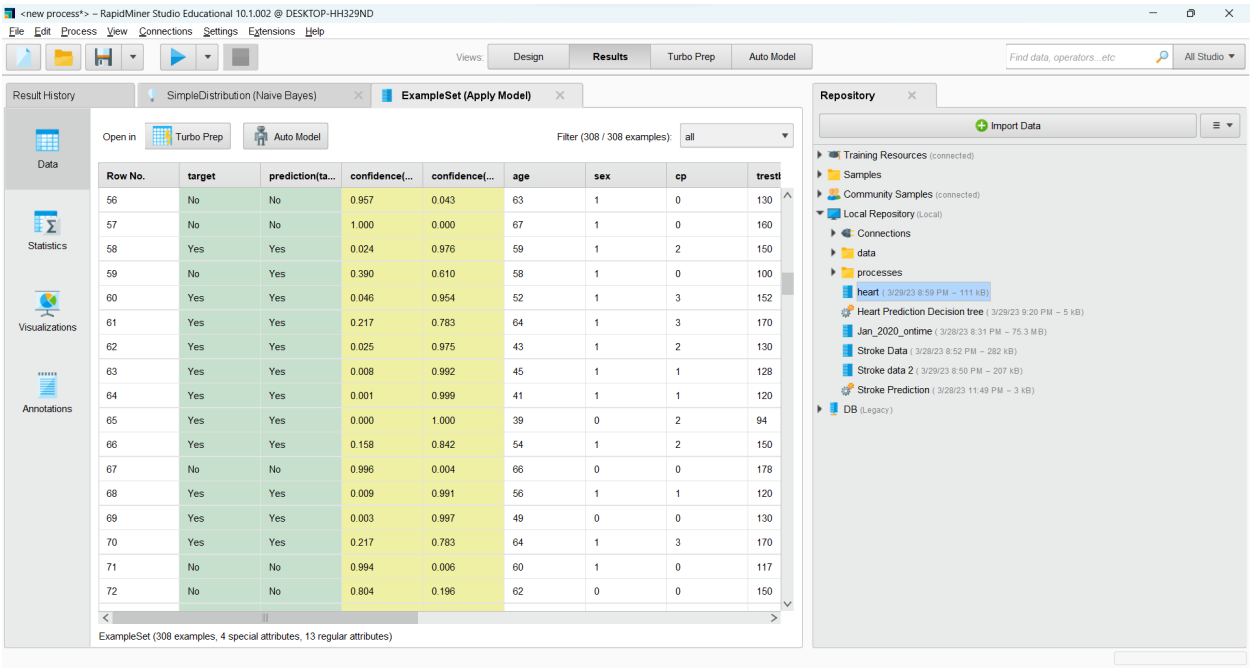
Now we will drag and drop Apply Model for the remaining 30% testing data



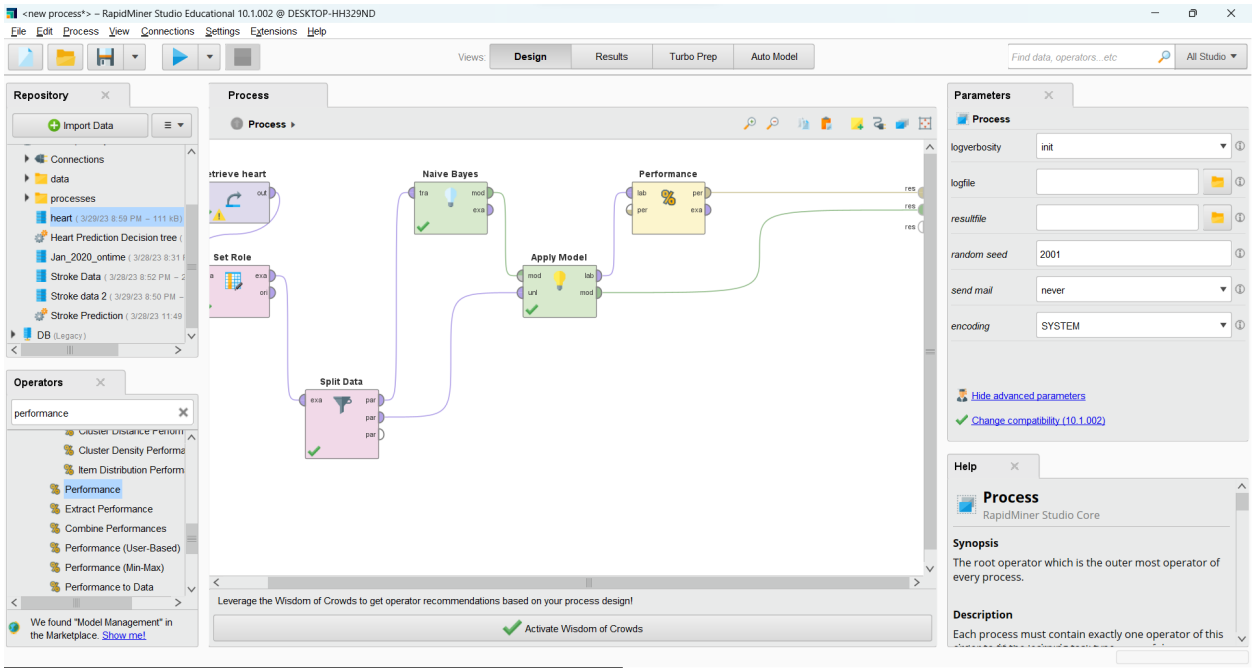


Step 4:After running this model we will get the following output:

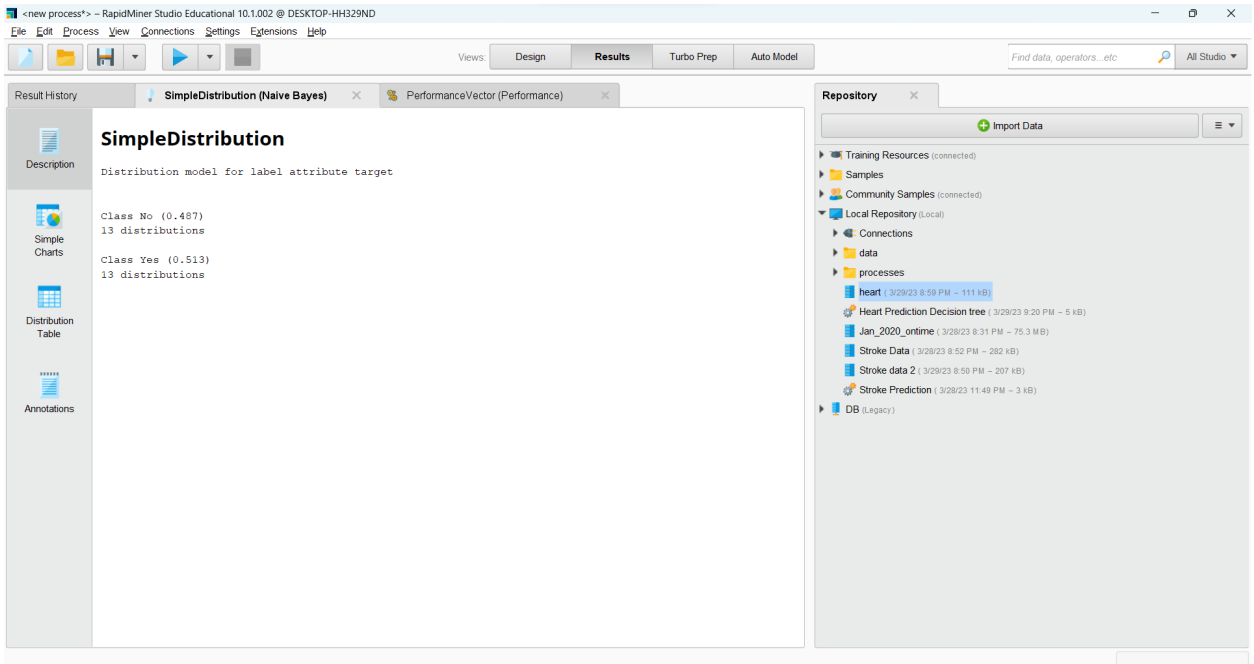
The following output shows the actual value and predicted value with confidence



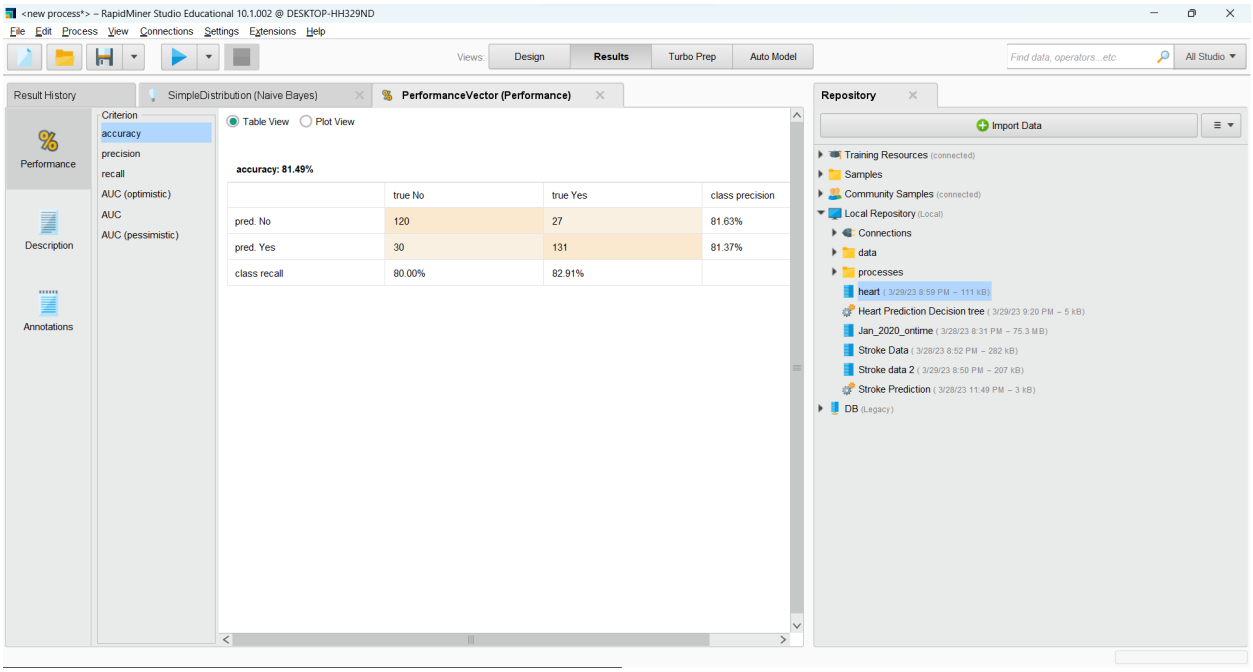
Now we want to apply performance operator, so we will drag and drop it



Step 5:After running this model we will get:
This will show the Simple Distribution of Datset

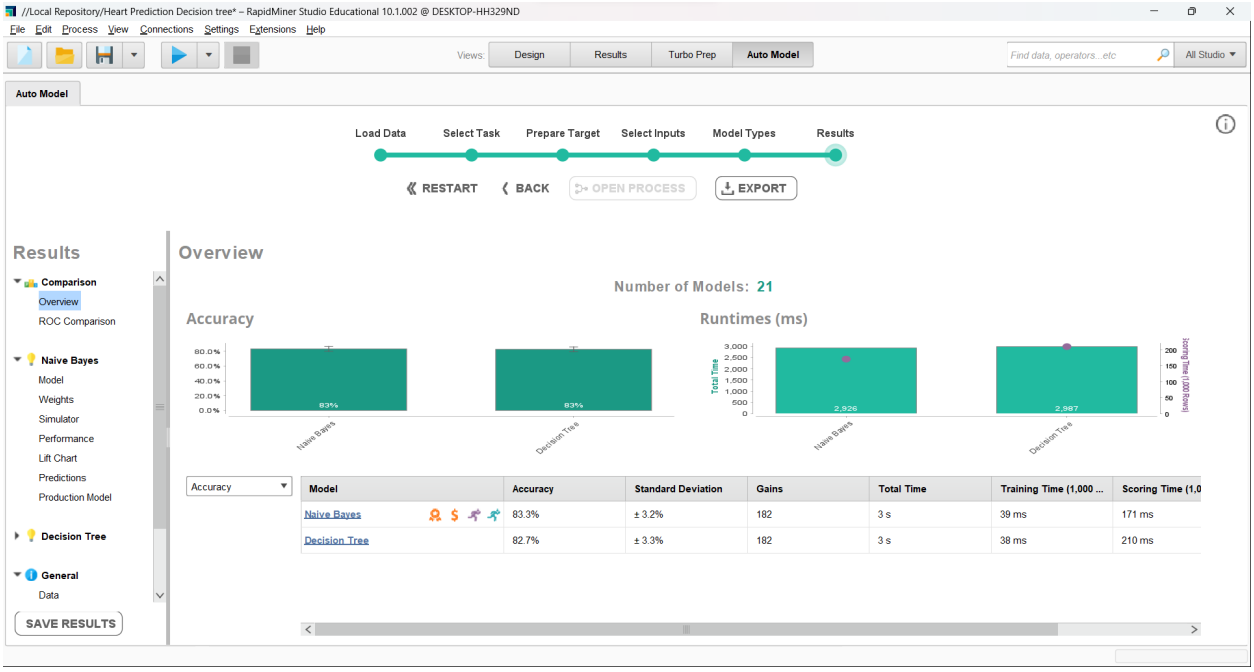


From the performance Vector we can see the accuracy as follows:



The overall **Accuracy of Naive Bayes is 81.49%**

From Given conditions we can clearly conclude that Decision tree has performed better than Naive Bayes. There is a huge difference between their accuracies.



This is the final Comparison of Decision Tree and Naive Bayes

Conclusion:

In this Experiment we saw two classification algorithms that are Decision Tree and Naive Bayes. We found their accuracies based on the provided dataset of heart disease prediction.

The performance of Decision Tree is better than Naive Bayes in RapidMiner. We have successfully implemented Experiment to explore Rapid Miner and implement classification models like Decision Tree and Naive Bayes etc.