Problem Statement

This dataset is about Human Heart Disease dataset. The objective is to predict

whether the patient has disease or not.

About 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.

Data Field description:

age

sex

chest pain type (4 values)

resting blood pressure

serum cholestoral in mg/dl

fasting blood sugar > 120 mg/dl

resting electrocardiographic results (values 0,1,2)

maximum heart rate achieved

exercise induced angina

oldpeak = ST depression induced by exercise relative to rest

the slope of the peak exercise ST segment

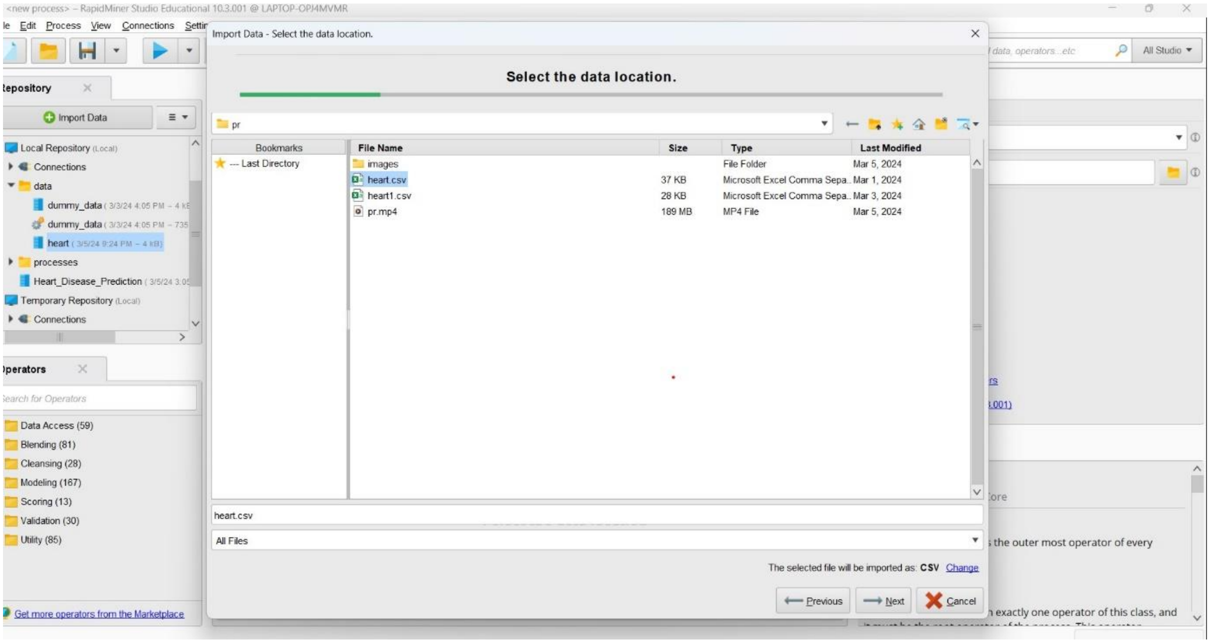
number of major vessels (0-3) colored by flourosopy

thal: 0 = normal; 1 = fixed defect; 2 = reversable defect

The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

Steps to convert dataset to prediction:

1. **Load data:** Import dataset from your computer to local repository.



Import Data - Select the data location.

Select the data location.

pr

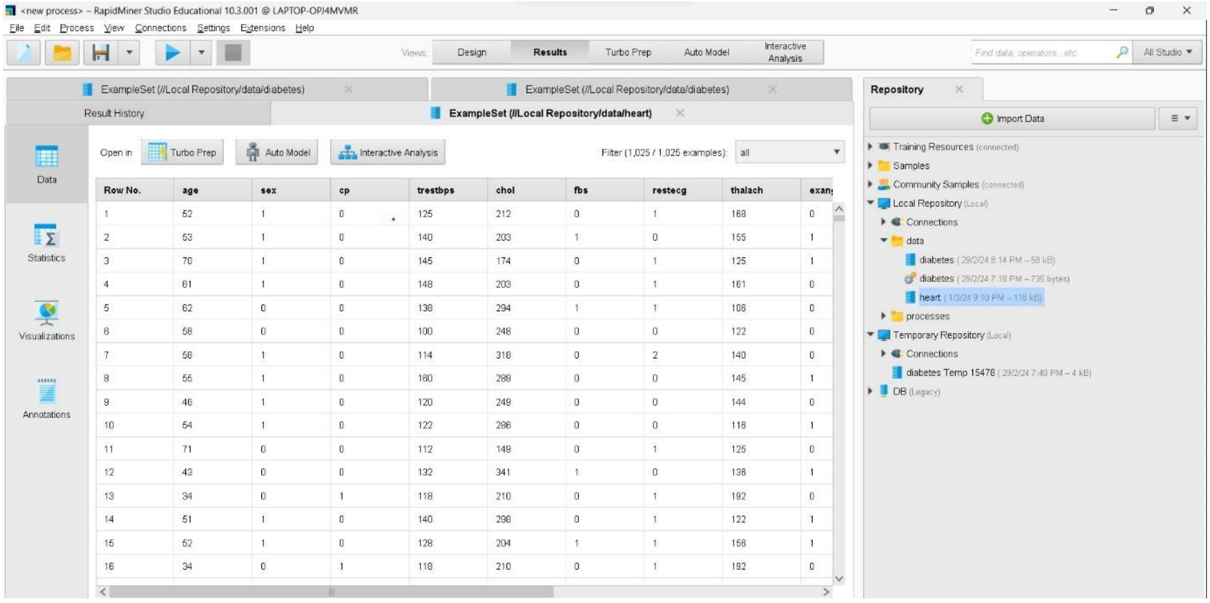
File Name	Size	Type	Last Modified
images		File Folder	Mar 5, 2024
heart.csv	37 KB	Microsoft Excel Comma Sepa..	Mar 1, 2024
heart1.csv	28 KB	Microsoft Excel Comma Sepa..	Mar 3, 2024
pr.mp4	189 MB	MP4 File	Mar 5, 2024

heart.csv

All Files

The selected file will be imported as: CSV Change

2. Exploring Dataset size:



ExampleSet (//Local Repository/data/heart)

Filter (1,025 / 1,025 examples): all

Row No.	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exan
1	52	1	0	125	212	0	1	168	0
2	53	1	0	140	203	1	0	155	1
3	70	1	0	145	174	0	1	125	1
4	61	1	0	148	203	0	1	161	0
5	62	0	0	138	294	1	1	106	0
6	58	0	0	100	248	0	0	122	0
7	58	1	0	114	318	0	2	140	0
8	55	1	0	160	289	0	0	145	1
9	46	1	0	120	249	0	0	144	0
10	54	1	0	122	286	0	0	116	1
11	71	0	0	112	149	0	1	125	0
12	43	0	0	132	341	1	0	136	1
13	34	0	1	118	210	0	1	192	0
14	51	1	0	140	298	0	1	122	1
15	52	1	0	128	204	1	1	158	1
16	34	0	1	118	210	0	1	192	0

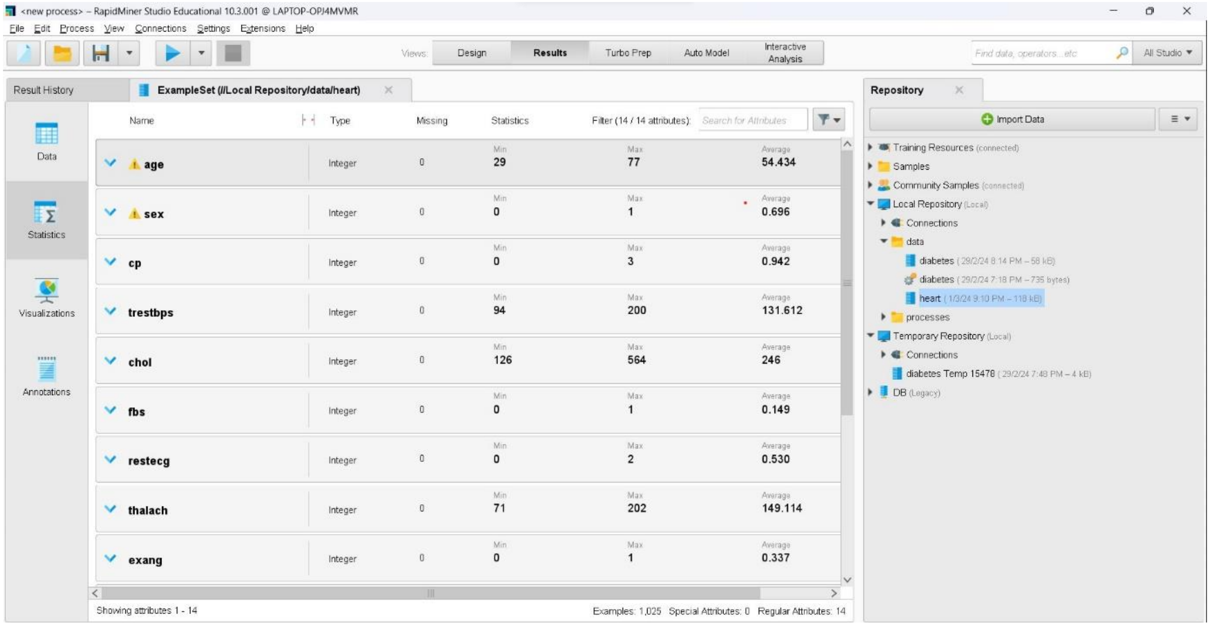
<new process> – RapidMiner Studio Educational 10.3.001 @ LAPTOP-OPJ4MVMR
File Edit Process View Connections Settings Extensions Help
Design
Results
Turbo Prep
Auto Model
Interactive Analysis
ExampleSet (//Local Repository/data/diabetes)
ExampleSet (//Local Repository/data/diabetes)
Result History
Open in
Data
Statistics
Visualizations
Annotations
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
diabetes
diabetes
heart
processes
Temporary Repository (Local)
Connections
diabetes Temp 15478
DB (Legacy)

From above figure there are 1025 examples/rows/records and 14 columns/attributes.

Out of 14 columns, 13 columns are predictors and remaining column is target.

3. Data cleanup:

a. In this dataset there is no missing values.


<new process> – RapidMiner Studio Educational 10.3.001 @ LAPTOP-OPJ4MVMR
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Result History
ExampleSet (//Local Repository/data/heart)
Data
Statistics
Visualizations
Annotations
Filter (14 / 14 attributes):
Search for Attributes
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
diabetes (29/2/24 8:14 PM – 58 kB)
diabetes (29/2/24 7:18 PM – 735 bytes)
heart (1/3/24 9:10 PM – 118 kB)
processes
Temporary Repository (Local)
Connections
diabetes Temp 15478 (29/2/24 7:48 PM – 4 kB)
DB (Legacy)

Name	Type	Missing	Statistics		
age	Integer	0	Min 29	Max 77	Average 54.434
sex	Integer	0	Min 0	Max 1	Average 0.696
cp	Integer	0	Min 0	Max 3	Average 0.942
trestbps	Integer	0	Min 94	Max 200	Average 131.612
chol	Integer	0	Min 126	Max 564	Average 246
fbs	Integer	0	Min 0	Max 1	Average 0.149
restecg	Integer	0	Min 0	Max 2	Average 0.530
thalach	Integer	0	Min 71	Max 202	Average 149.114
exang	Integer	0	Min 0	Max 1	Average 0.337

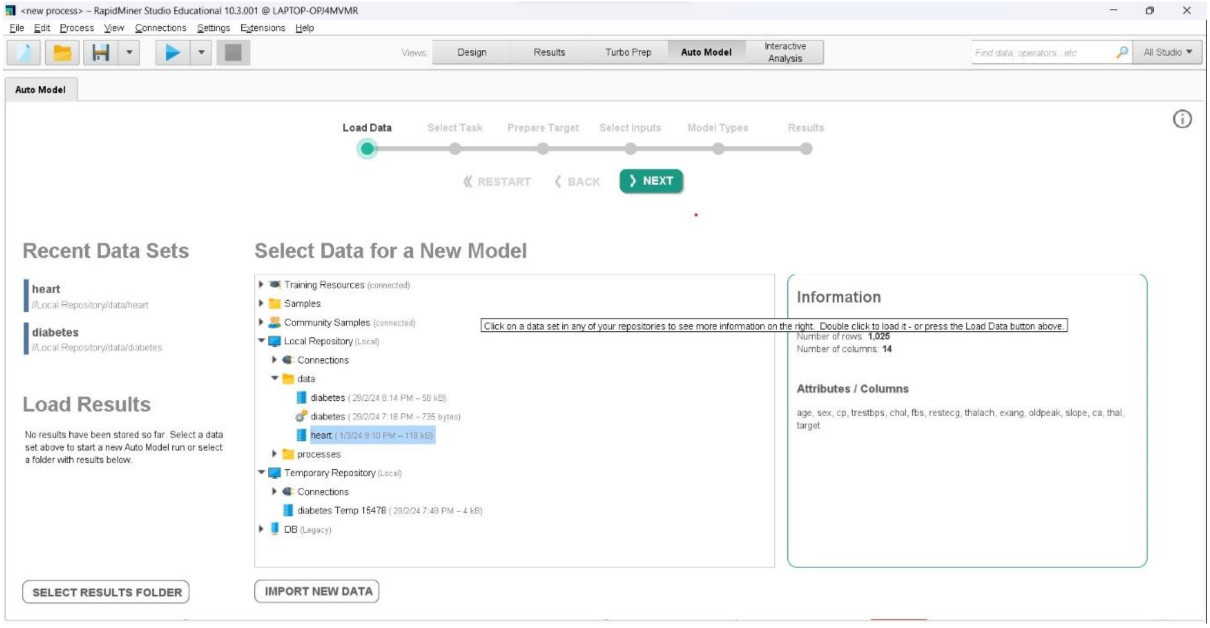
Showing attributes 1 - 14

Examples: 1,025 Special Attributes: 0 Regular Attributes: 14

b. Drop attributes if required.

We can drop attributes if required. In this dataset there is no need of dropping the attribute.

c. Target selection:


<new process> – RapidMiner Studio Educational 10.3.001 @ LAPTOP-OPJ4MVMR
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Auto Model
Load Data Select Task Prepare Target Select Inputs Model Types Results
RESTART BACK NEXT
SELECT RESULTS FOLDER
IMPORT NEW DATA

Recent Data Sets

heart
//Local Repository/data/heart

diabetes
//Local Repository/data/diabetes

Load Results

No results have been stored so far. Select a data set above to start a new Auto Model run or select a folder with results below.

Select Data for a New Model

- Training Resources (connected)
- Samples
- Community Samples (connected)
- Local Repository (Local)
  - Connections
  - data
    - diabetes (29/2/24 8:14 PM – 58 kB)
    - diabetes (29/2/24 7:18 PM – 735 bytes)
    - heart (1/3/24 9:10 PM – 118 kB)
  - processes
- Temporary Repository (Local)
  - Connections
  - diabetes Temp 15478 (29/2/24 7:48 PM – 4 kB)
- DB (Legacy)

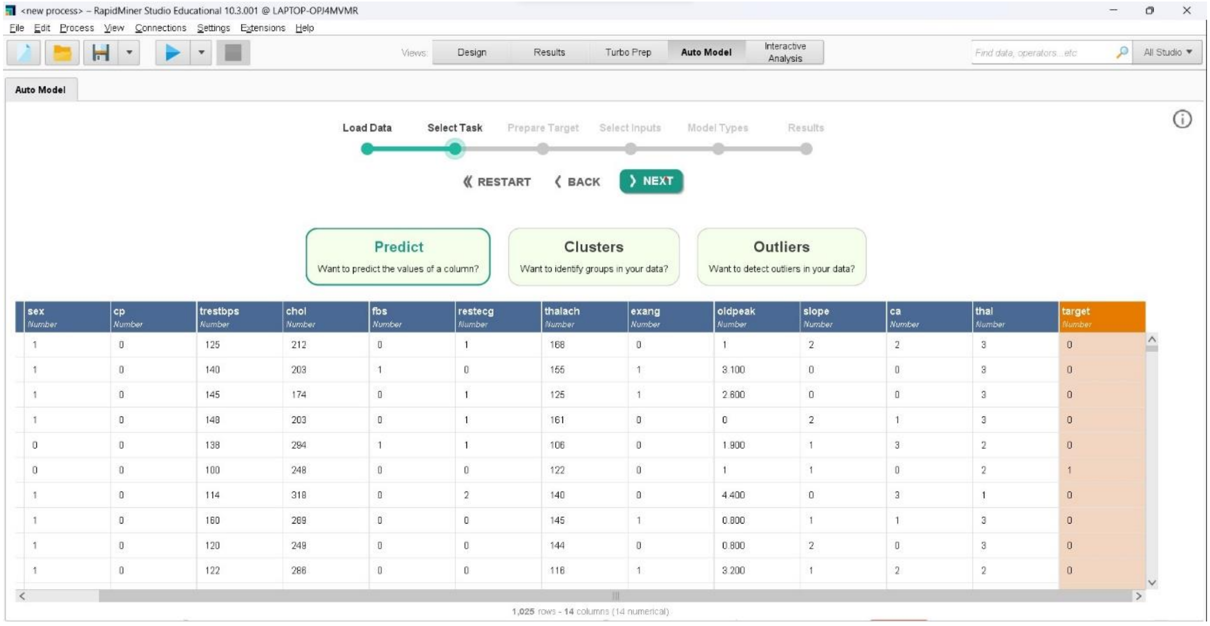
Click on a data set in any of your repositories to see more information on the right. Double click to load it - or press the Load Data button above.

Information

Number of rows: 1,025
Number of columns: 14

Attributes / Columns

age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, target


<new process> – RapidMiner Studio Educational 10.3.001 @ LAPTOP-OPJ4MVMR
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Auto Model
Load Data Select Task Prepare Target Select Inputs Model Types Results
RESTART BACK NEXT

Predict
Want to predict the values of a column?

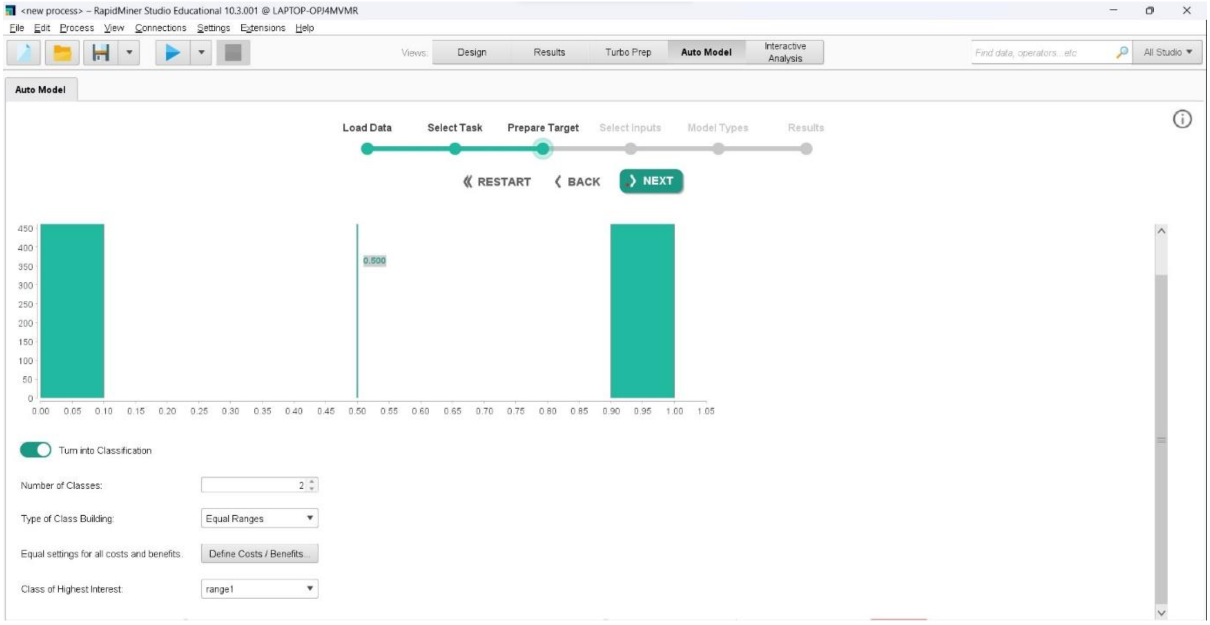
Clusters
Want to identify groups in your data?

Outliers
Want to detect outliers in your data?

sex Number	cp Number	trestbps Number	chol Number	fbs Number	restecg Number	thalach Number	exang Number	oldpeak Number	slope Number	ca Number	thal Number	target Number
1	0	125	212	0	1	168	0	1	2	2	3	0
1	0	140	203	1	0	155	1	3.100	0	0	3	0
1	0	145	174	0	1	125	1	2.600	0	0	3	0
1	0	148	203	0	1	161	0	0	2	1	3	0
0	0	138	294	1	1	106	0	1.900	1	3	2	0
0	0	100	248	0	0	122	0	1	1	0	2	1
1	0	114	318	0	2	140	0	4.400	0	3	1	0
1	0	160	289	0	0	145	1	0.800	1	1	3	0
1	0	120	249	0	0	144	0	0.800	2	0	3	0
1	0	122	286	0	0	116	1	3.200	1	2	2	0

1,025 rows - 14 columns (14 numerical)

d. Prepare target:



Turn into Classification

Number of Classes: 2

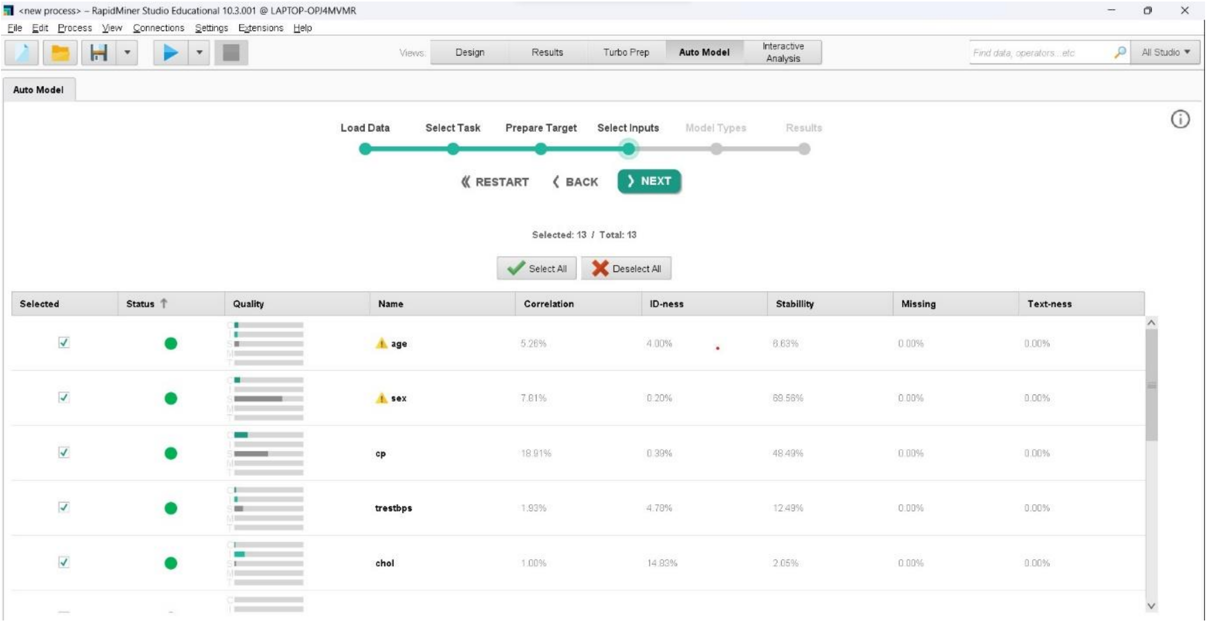
Type of Class Building: Equal Ranges

Equal settings for all costs and benefits: Define Costs / Benefits...

Class of Highest Interest: range1

Here target is to predict the target attribute.

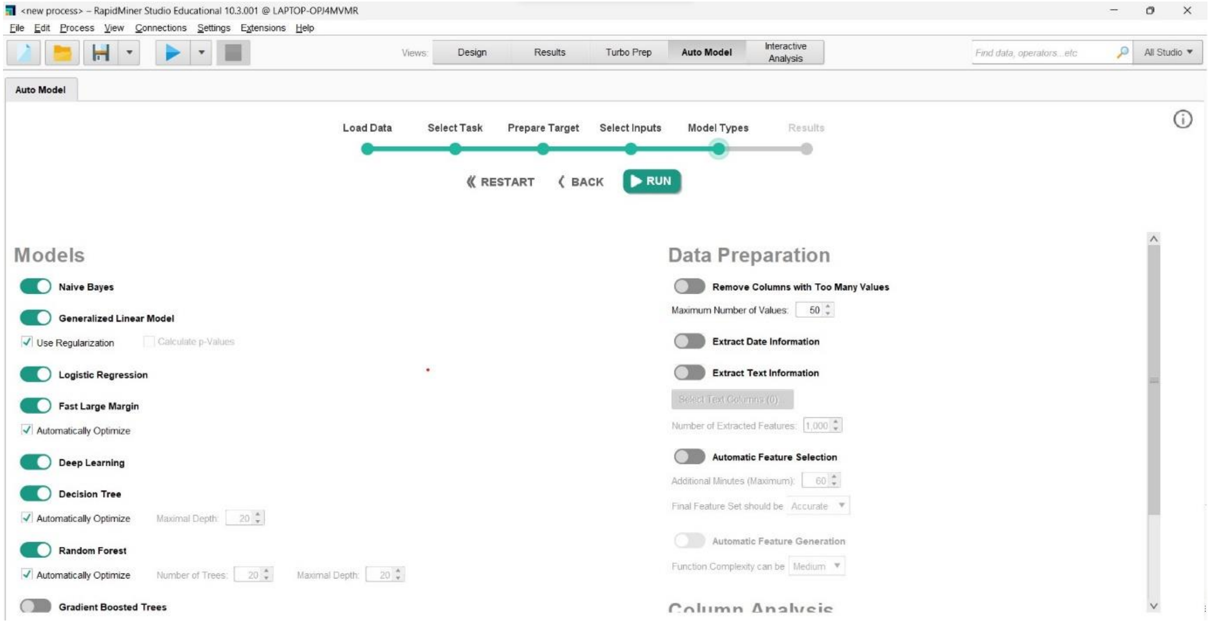
e. Select inputs:


<new process> – RapidMiner Studio Educational 10.3.001 @ LAPTOP-OPJ4MVMR
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Auto Model
Load Data
Select Task
Prepare Target
Select Inputs
Model Types
Results
RESTART
BACK
NEXT
Select All
Deselect All

Selected: 13 / Total: 13

Selected	Status	Quality	Name	Correlation	ID-ness	Stabillity	Missing	Text-ness
☑	●		age	5.26%	4.00%	6.63%	0.00%	0.00%
☑	●		sex	7.81%	0.20%	69.56%	0.00%	0.00%
☑	●		cp	18.91%	0.39%	48.49%	0.00%	0.00%
☑	●		trestbps	1.93%	4.78%	12.49%	0.00%	0.00%
☑	●		chol	1.00%	14.83%	2.05%	0.00%	0.00%

f. Select the model:



<new process> – RapidMiner Studio Educational 10.3.001 @ LAPTOP-OPJ4MVMR
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Auto Model
Load Data
Select Task
Prepare Target
Select Inputs
Model Types
Results
RESTART
BACK
RUN

Models

● Naive Bayes

● Generalized Linear Model

☑ Use Regularization ☐ Calculate p-Values

● Logistic Regression

● Fast Large Margin

☑ Automatically Optimize

● Deep Learning

● Decision Tree

☑ Automatically Optimize Maximal Depth: 20

● Random Forest

☑ Automatically Optimize Number of Trees: 20 Maximal Depth: 20

○ Gradient Boosted Trees

Data Preparation

○ Remove Columns with Too Many Values

Maximum Number of Values: 50

○ Extract Date Information

○ Extract Text Information

Select Text Columns (0)

Number of Extracted Features: 1,000

○ Automatic Feature Selection

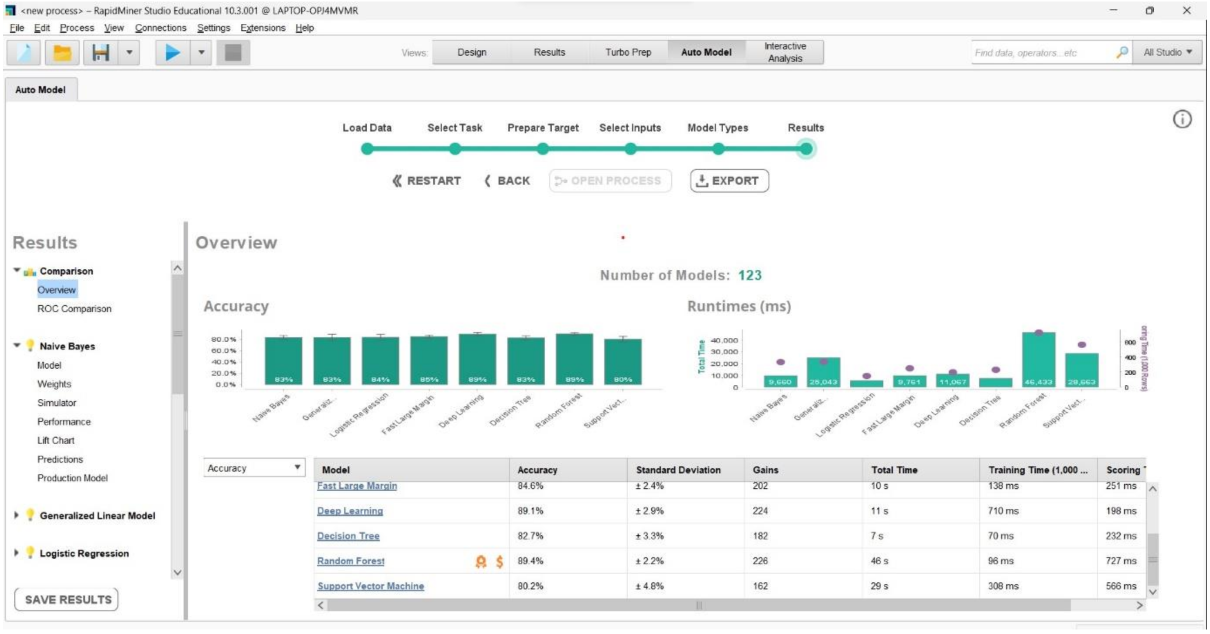
Additional Minutes (Maximum): 60

Final Feature Set should be Accurate

○ Automatic Feature Generation

Function Complexity can be Medium

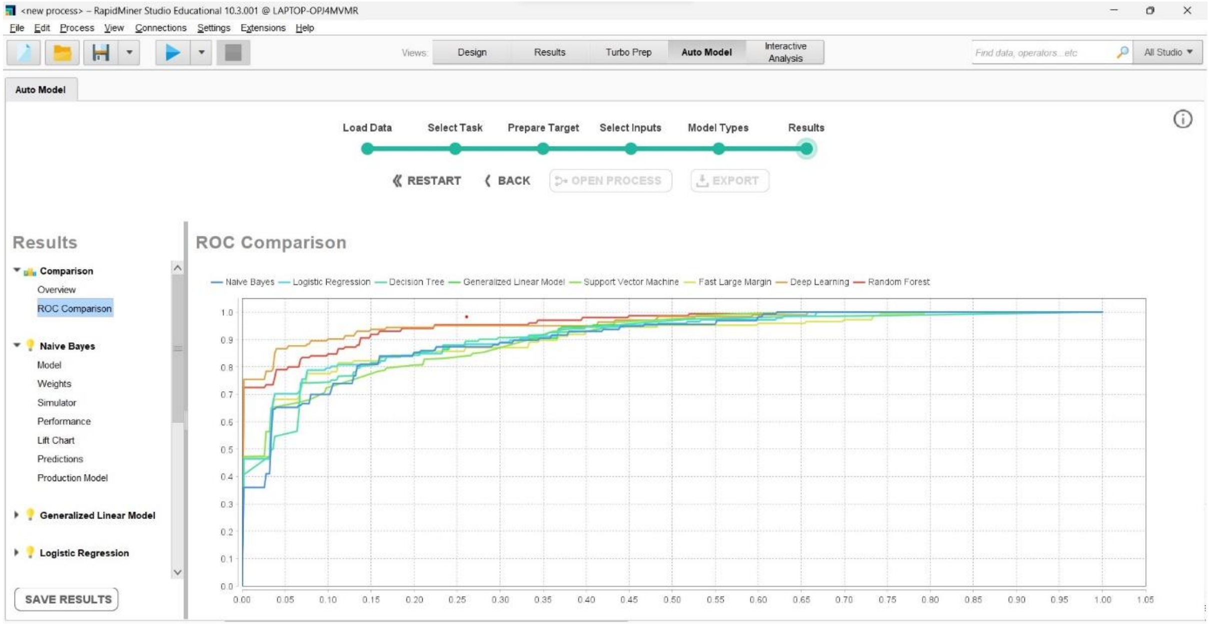
Column Analysis



Number of Models: 123

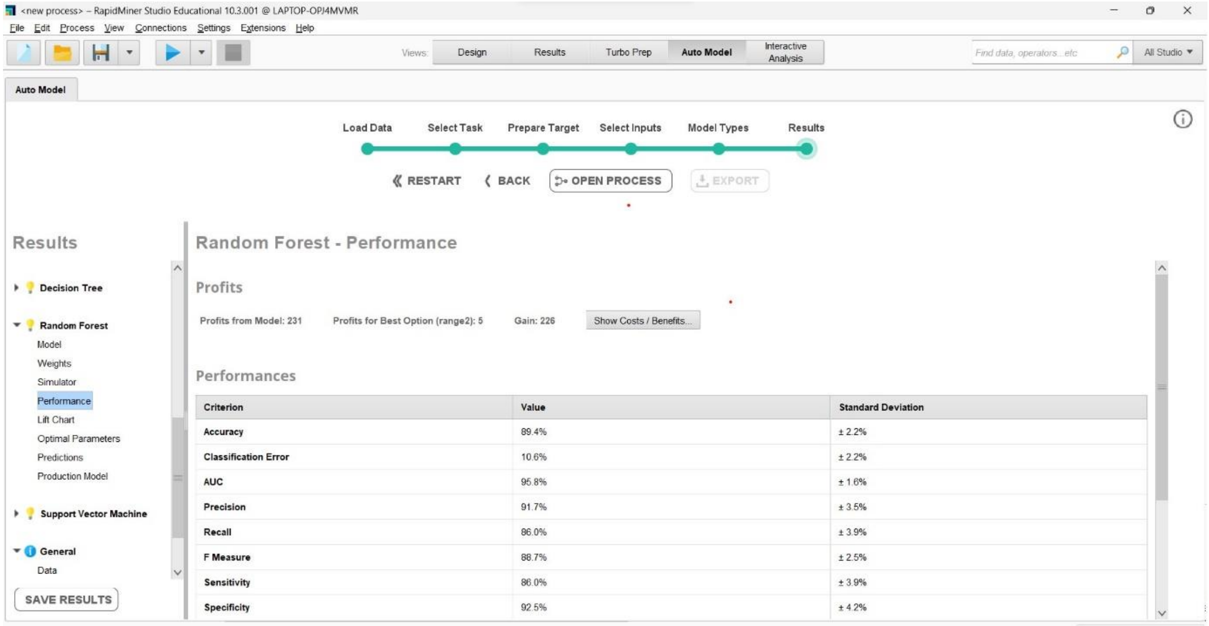
Accuracy ▾

Model	Accuracy	Standard Deviation	Gains	Total Time	Training Time (1,000 ...	Scoring
Fast Large Margin	84.6%	± 2.4%	202	10 s	138 ms	251 ms
Deep Learning	89.1%	± 2.9%	224	11 s	710 ms	198 ms
Decision Tree	82.7%	± 3.3%	182	7 s	70 ms	232 ms
Random Forest $	89.4%	± 2.2%	226	46 s	96 ms	727 ms
Support Vector Machine	80.2%	± 4.8%	162	29 s	308 ms	566 ms



Out of all models Random Forest has given best performance and best gain.

4. Baseline Model:



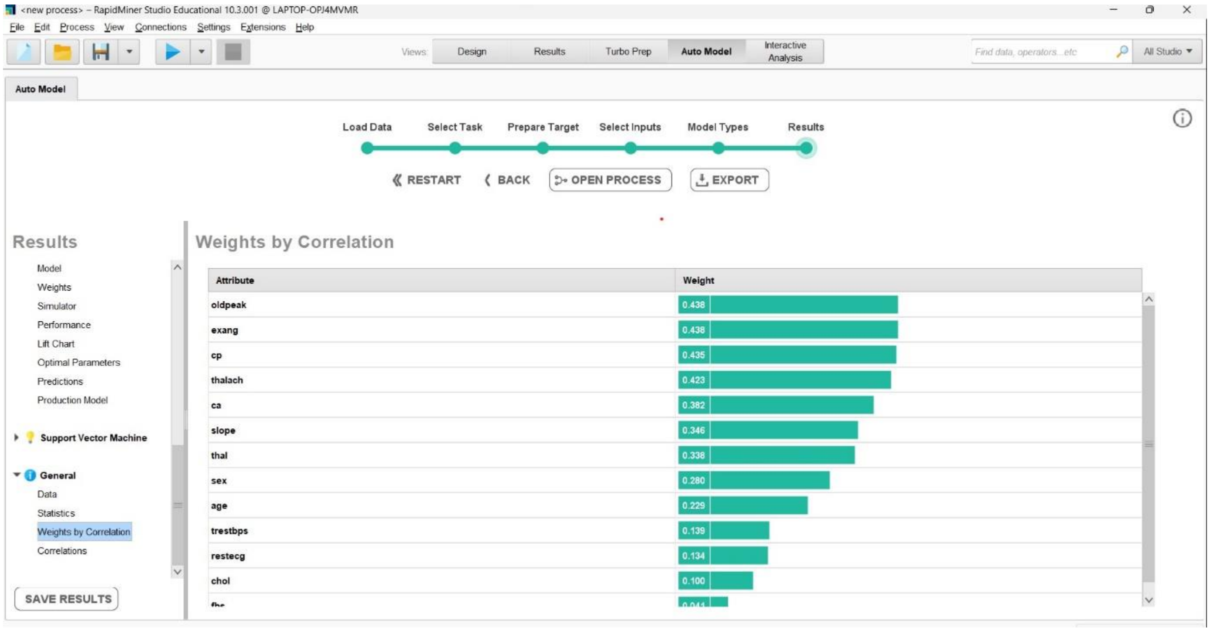
Random Forest - Performance

Profits

Profits from Model: 231   Profits for Best Option (range2): 5   Gain: 226

Performances

Criterion	Value	Standard Deviation
Accuracy	89.4%	± 2.2%
Classification Error	10.6%	± 2.2%
AUC	95.8%	± 1.6%
Precision	91.7%	± 3.5%
Recall	86.0%	± 3.9%
F Measure	88.7%	± 2.5%
Sensitivity	86.0%	± 3.9%
Specificity	92.5%	± 4.2%

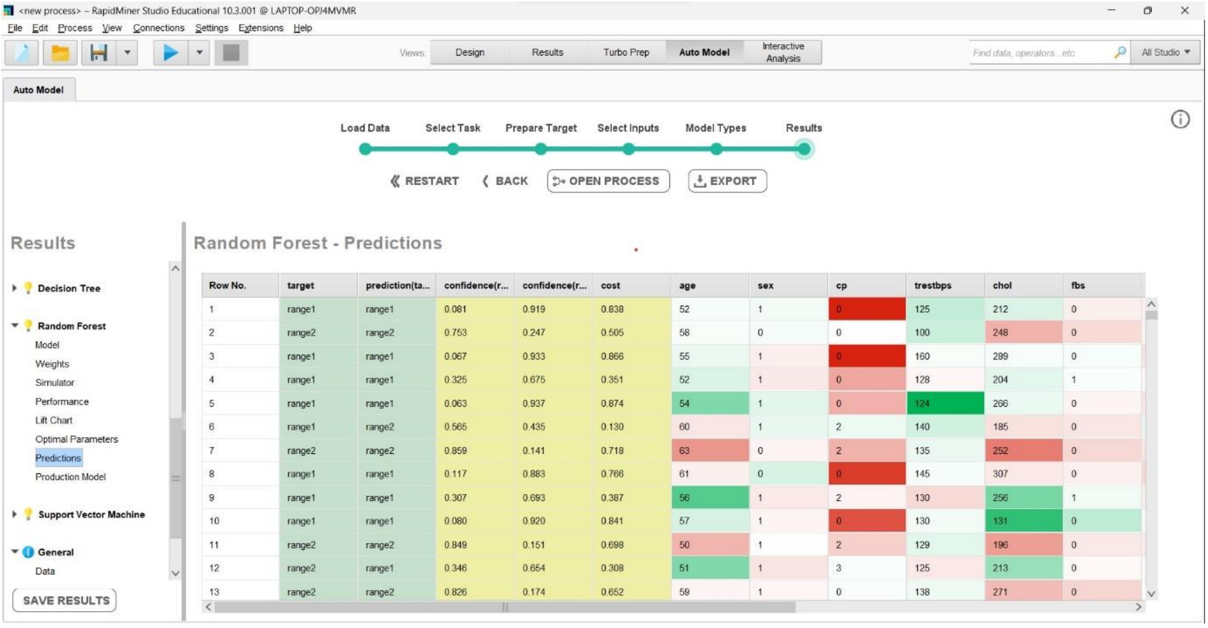


Weights by Correlation

Attribute	Weight
oldpeak	0.438
exang	0.438
cp	0.435
thalach	0.423
ca	0.382
slope	0.346
thal	0.338
sex	0.280
age	0.229
trestbps	0.139
restecg	0.134
chol	0.100
fbs	0.041

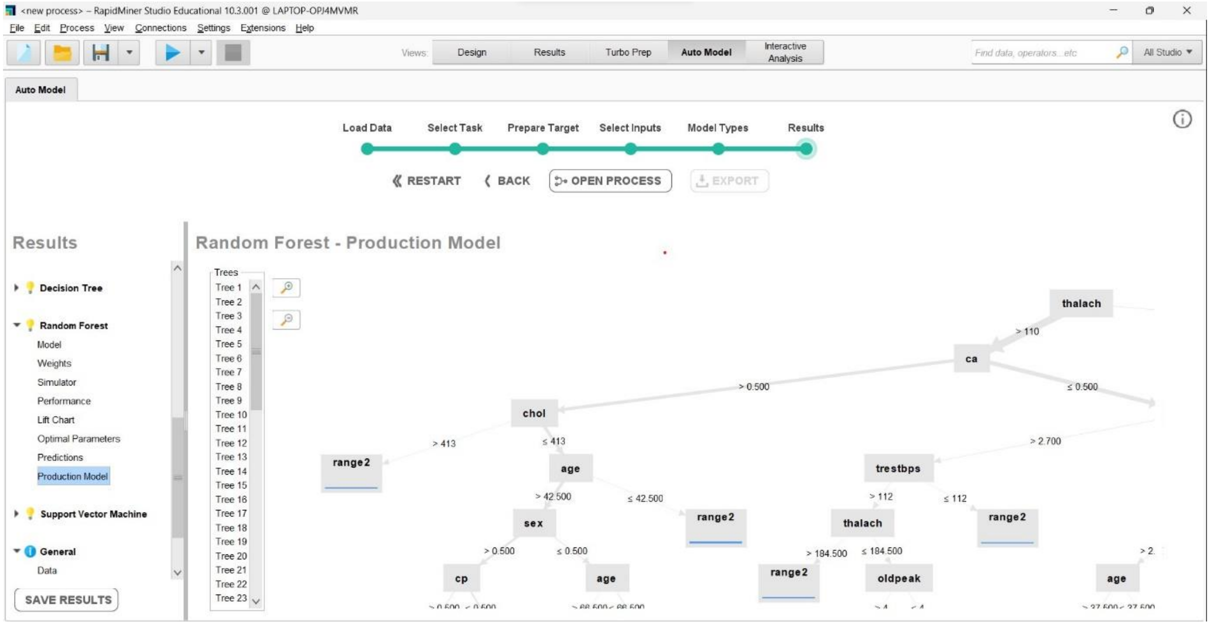
Model is more correlated on the oldpeak column and it has given almost equal preference to remaining attributes to predict outcome.

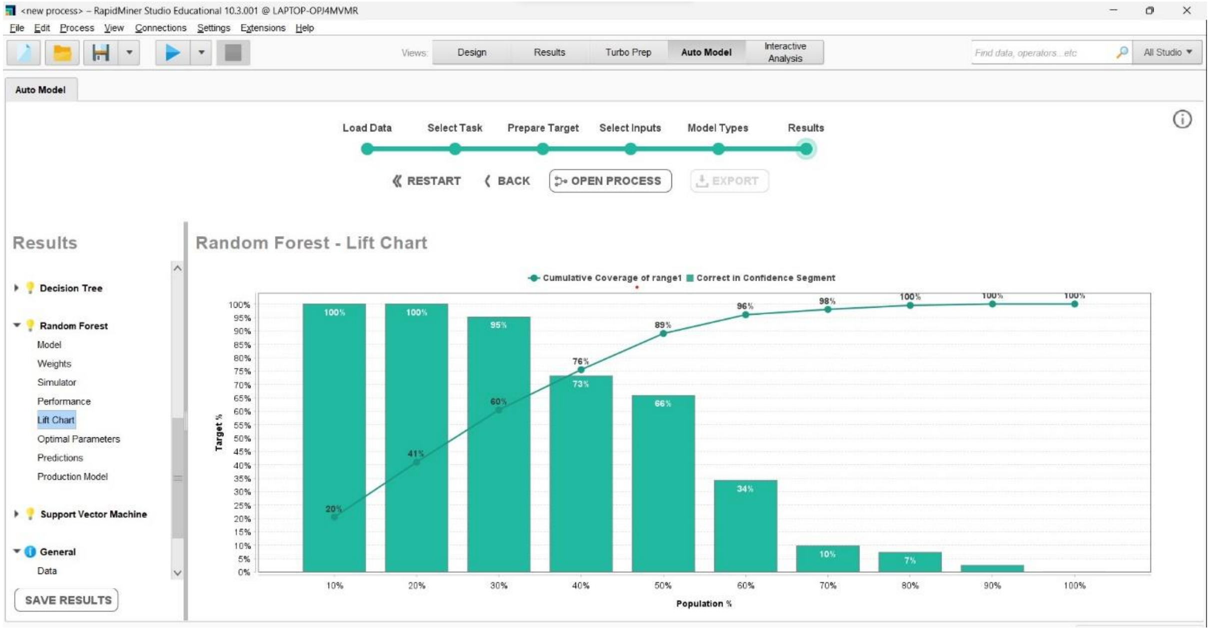
Out of all models Random Forest has given max accuracy about 89.4% with classification error of 10.6%.

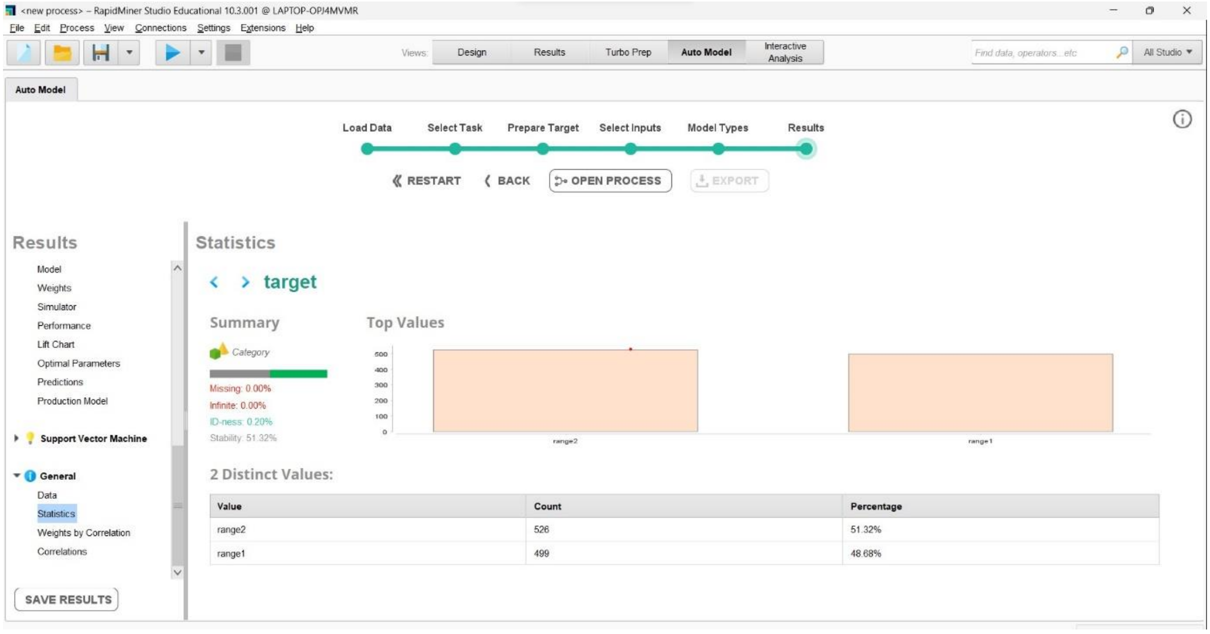


Random Forest - Predictions

Row No.	target	prediction(ta...	confidence(r...	confidence(r...	cost	age	sex	cp	trestbps	chol	fbs
1	range1	range1	0.081	0.919	0.838	52	1	0	125	212	0
2	range2	range2	0.753	0.247	0.505	58	0	0	100	248	0
3	range1	range1	0.067	0.933	0.866	55	1	0	160	289	0
4	range1	range1	0.325	0.675	0.351	52	1	0	128	204	1
5	range1	range1	0.063	0.937	0.874	54	1	0	124	266	0
6	range1	range2	0.565	0.435	0.130	60	1	2	140	185	0
7	range2	range2	0.859	0.141	0.718	63	0	2	135	252	0
8	range1	range1	0.117	0.883	0.766	61	0	0	145	307	0
9	range1	range1	0.307	0.693	0.387	56	1	2	130	256	1
10	range1	range1	0.080	0.920	0.841	57	1	0	130	131	0
11	range2	range2	0.849	0.151	0.698	50	1	2	129	196	0
12	range2	range1	0.346	0.654	0.308	51	1	3	125	213	0
13	range2	range2	0.826	0.174	0.652	59	1	0	138	271	0


<new process> – RapidMiner Studio Educational 10.3.001 @ LAPTOP-OPJ4MVMR
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Auto Model
Load Data
Select Task
Prepare Target
Select Inputs
Model Types
Results
RESTART
BACK
OPEN PROCESS
EXPORT
Results
Decision Tree
Random Forest
Model
Weights
Simulator
Performance
Lift Chart
Optimal Parameters
Predictions
Production Model
Support Vector Machine
General
Data
SAVE RESULTS
Random Forest - Production Model
Trees
Tree 1
Tree 2
Tree 3
Tree 4
Tree 5
Tree 6
Tree 7
Tree 8
Tree 9
Tree 10
Tree 11
Tree 12
Tree 13
Tree 14
Tree 15
Tree 16
Tree 17
Tree 18
Tree 19
Tree 20
Tree 21
Tree 22
Tree 23
thalach
> 110
ca
> 0.500
≤ 0.500
chol
> 413
≤ 413
range2
age
> 2.700
trestbps
> 42.500
≤ 42.500
sex
range2
> 112
≤ 112
thalach
range2
> 0.500
≤ 0.500
cp
age
> 184.500
≤ 184.500
range2
oldpeak
> 2
age


<new process> – RapidMiner Studio Educational 10.3.001 @ LAPTOP-OPJ4MVMR
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Auto Model
Load Data
Select Task
Prepare Target
Select Inputs
Model Types
Results
RESTART
BACK
OPEN PROCESS
EXPORT
Results
Decision Tree
Random Forest
Model
Weights
Simulator
Performance
Lift Chart
Optimal Parameters
Predictions
Production Model
Support Vector Machine
General
Data
SAVE RESULTS
Random Forest - Lift Chart
Cumulative Coverage of range1
Correct in Confidence Segment
Target %
100%
95%
90%
85%
80%
75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Population %
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
100%
100%
95%
73%
66%
34%
10%
7%
20%
41%
60%
76%
89%
96%
98%
100%
100%
100%



Statistics

target

Summary

Category

Missing: 0.00%

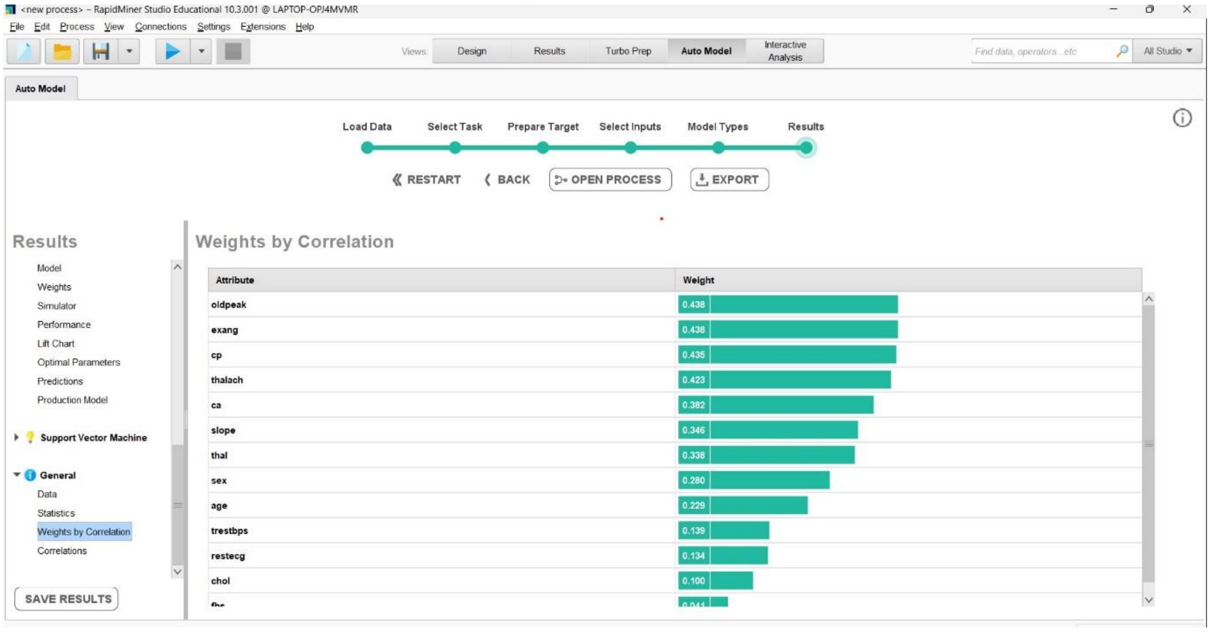
Infinite: 0.00%

ID-ness: 0.20%

Stability: 51.32%

2 Distinct Values:

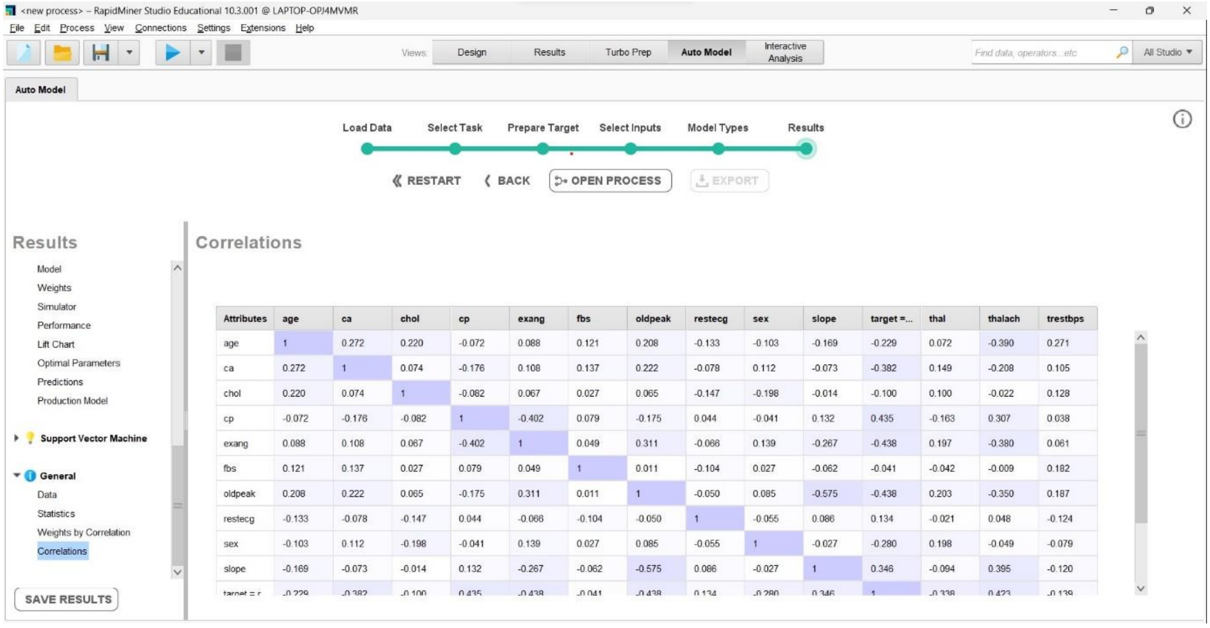
Value	Count	Percentage
range2	526	51.32%
range1	499	48.68%



Weights by Correlation

Attribute	Weight
oldpeak	0.438
exang	0.438
cp	0.435
thalach	0.423
ca	0.382
slope	0.346
thal	0.338
sex	0.280
age	0.229
trestbps	0.139
restecg	0.134
chol	0.100
fbs	0.041

5. Correlation matrix:

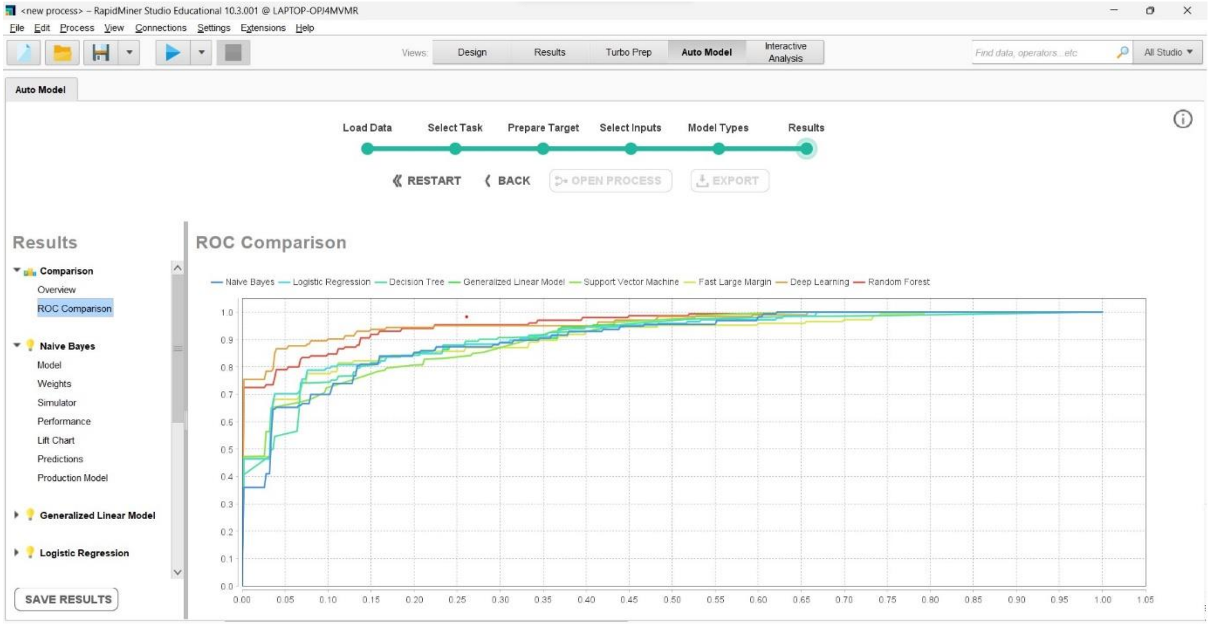


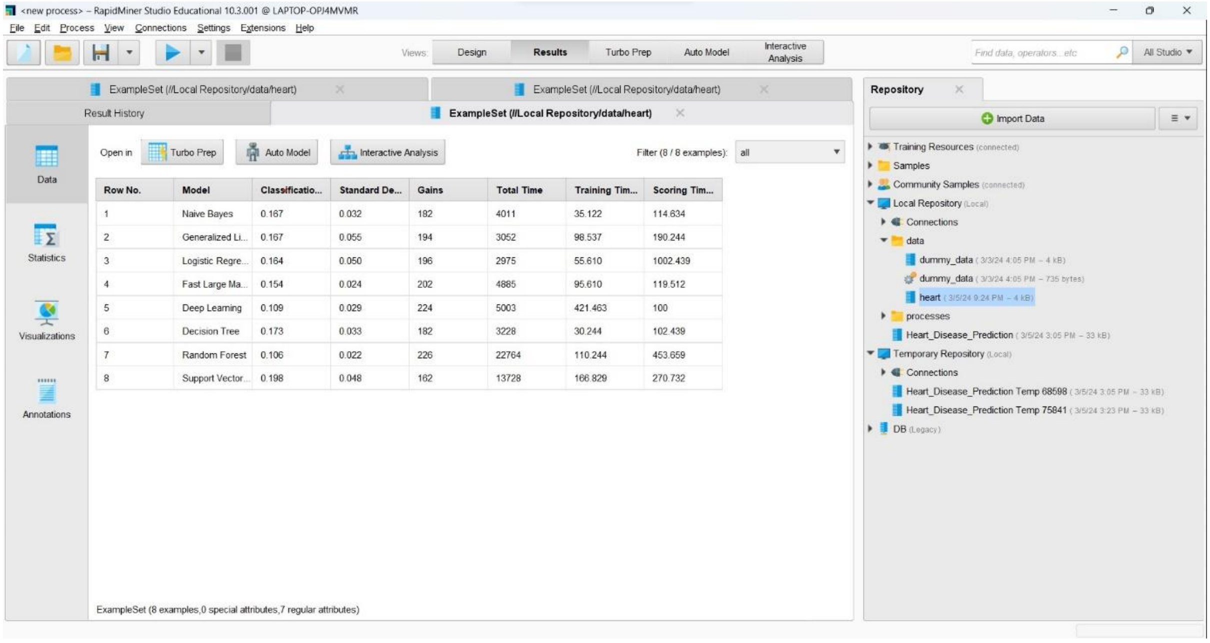
<new process> – RapidMiner Studio Educational 10.3.001 @ LAPTOP-OPJ4MVMR
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Auto Model
Load Data Select Task Prepare Target Select Inputs Model Types Results
RESTART BACK OPEN PROCESS EXPORT
Results
Model
Weights
Simulator
Performance
Lift Chart
Optimal Parameters
Predictions
Production Model
Support Vector Machine
General
Data
Statistics
Weights by Correlation
Correlations
SAVE RESULTS

Correlations

Attributes	age	ca	chol	cp	exang	fbs	oldpeak	restecg	sex	slope	target =...	thal	thalach	trestbps
age	1	0.272	0.220	-0.072	0.088	0.121	0.208	-0.133	-0.103	-0.169	-0.229	0.072	-0.390	0.271
ca	0.272	1	0.074	-0.176	0.108	0.137	0.222	-0.078	0.112	-0.073	-0.382	0.149	-0.208	0.105
chol	0.220	0.074	1	-0.082	0.067	0.027	0.065	-0.147	-0.198	-0.014	-0.100	0.100	-0.022	0.128
cp	-0.072	-0.176	-0.082	1	-0.402	0.079	-0.175	0.044	-0.041	0.132	0.435	-0.163	0.307	0.038
exang	0.088	0.108	0.067	-0.402	1	0.049	0.311	-0.066	0.139	-0.267	-0.438	0.197	-0.380	0.061
fbs	0.121	0.137	0.027	0.079	0.049	1	0.011	-0.104	0.027	-0.062	-0.041	-0.042	-0.009	0.182
oldpeak	0.208	0.222	0.065	-0.175	0.311	0.011	1	-0.050	0.085	-0.575	-0.438	0.203	-0.350	0.187
restecg	-0.133	-0.078	-0.147	0.044	-0.066	-0.104	-0.050	1	-0.055	0.086	0.134	-0.021	0.048	-0.124
sex	-0.103	0.112	-0.198	-0.041	0.139	0.027	0.085	-0.055	1	-0.027	-0.280	0.198	-0.049	-0.079
slope	-0.169	-0.073	-0.014	0.132	-0.267	-0.062	-0.575	0.086	-0.027	1	0.346	-0.094	0.395	-0.120
target = r	-0.229	-0.382	-0.100	0.435	-0.438	-0.041	-0.438	0.134	-0.280	0.346	1	-0.338	0.423	-0.139

Output for random input:





<new process> – RapidMiner Studio Educational 10.3.001 @ LAPTOP-OPJ4MVMR
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
ExampleSet (//Local Repository/data/heart)
ExampleSet (//Local Repository/data/heart)
Result History
ExampleSet (//Local Repository/data/heart)
Data
Statistics
Visualizations
Annotations
Open in Turbo Prep Auto Model Interactive Analysis
Filter (8 / 8 examples): all
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
dummy_data (3/3/24 4:05 PM – 4 kB)
dummy_data (3/3/24 4:05 PM – 735 bytes)
heart (3/5/24 9:24 PM – 4 kB)
processes
Heart_Disease_Prediction (3/5/24 3:05 PM – 33 kB)
Temporary Repository (Local)
Connections
Heart_Disease_Prediction Temp 68598 (3/5/24 3:05 PM – 33 kB)
Heart_Disease_Prediction Temp 75841 (3/5/24 3:23 PM – 33 kB)
DB (Legacy)

Row No.	Model	Classificatio...	Standard De...	Gains	Total Time	Training Tim...	Scoring Tim...
1	Naive Bayes	0.167	0.032	182	4011	35.122	114.634
2	Generalized Li...	0.167	0.055	194	3052	98.537	190.244
3	Logistic Regre...	0.164	0.050	196	2975	55.610	1002.439
4	Fast Large Ma...	0.154	0.024	202	4885	95.610	119.512
5	Deep Learning	0.109	0.029	224	5003	421.463	100
6	Decision Tree	0.173	0.033	182	3228	30.244	102.439
7	Random Forest	0.106	0.022	226	22764	110.244	453.659
8	Support Vector...	0.198	0.048	162	13728	166.829	270.732

ExampleSet (8 examples,0 special attributes,7 regular attributes)

Conclusion:

Out of all the classifiers tested Random Forest has gave better results in predicting outcome

The conclusion for the Human Heart Disease dataset depends on the specific analysis and modeling performed on it. However, here are some potential conclusions that could be drawn based on the information provided:

Predictive Modeling: Researchers could build predictive models using machine learning algorithms to predict the presence or absence of heart disease based on the provided attributes. These models could be evaluated based on their accuracy, sensitivity, specificity, and other performance metrics.

Feature Selection: Since the dataset contains 76 attributes, but most experiments focus on a subset of 14 attributes, researchers could perform feature selection to identify the most relevant attributes for predicting heart disease. This can help in simplifying models and improving their interpretability.

Risk Factors Identification: By analyzing the relationships between the attributes and the presence of heart disease, researchers can identify potential risk factors or indicators of heart disease. This information can be valuable for healthcare professionals in understanding and managing the disease.

Dataset Bias and Generalization: Researchers should also consider potential biases in the dataset, such as overrepresentation of certain demographics or medical conditions in specific databases. Ensuring that the models generalize well to diverse populations is crucial for their applicability in real-world settings.

Further Research Directions: Depending on the findings from the initial analysis, researchers may identify areas for further investigation, such as exploring interactions between different attributes, investigating the impact of lifestyle factors on heart disease risk, or evaluating the effectiveness of different treatment strategies.

Overall, the conclusion drawn from the dataset would depend on the specific analyses performed and the goals of the research project.