Altair Data Science Contest

By Aishani singh

What is RapidMiner?

RapidMiner is a Data Science platform by Altair that can be used for various task like preparation, text mining ,predictive analytics and machine learning (ML)

Lung cancer Prediction

- we are going to predict the lung cancer and dataset is taken from kaggle -which is an Online platform for data science
- task is to predict the if the person has a cancer or not.
- Link of kaggle : https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer

About the Analysis

- Here , we are using RapidMiner to build a predictive model which will predict
- the lung cancer on dataset
- Key features of RapidMiner is GUI.
- there are 2 method to predict the build model.
  - Automodel
  - Using process Panel
- Another feature is drag and drop feature ,which is more easier for the data analyst
- there are various method and algorithms and operators are available in RapidMiner
  studio.

RapidMiner features

- Drag-and-Drop Interface: Simplifies the creation of data science workflows with an intuitive, code-free, visual design.
- Extensive Operator Library: Provides over 1500 operators for tasks like data preprocessing, machine learning, and model evaluation.
- Auto Model: Automates model selection, hyperparameter tuning, and evaluation to quickly find the best-performing models.
- Real-Time Scoring: Enables real-time predictions and data scoring, facilitating fast deployment in production environments.
- Integration with Big Data Platforms: Connects seamlessly with big data sources like Hadoop and Spark for large-scale data processing.

About Dataset

- Name of the dataset survey lung cancer.csv
- It has 309 records with 16 attributes.

Steps to cleansing the dataset:

- Remove Duplicates
- Handle Missing values

Importing database:

- click the import data
- select the location and file to be used for analysis

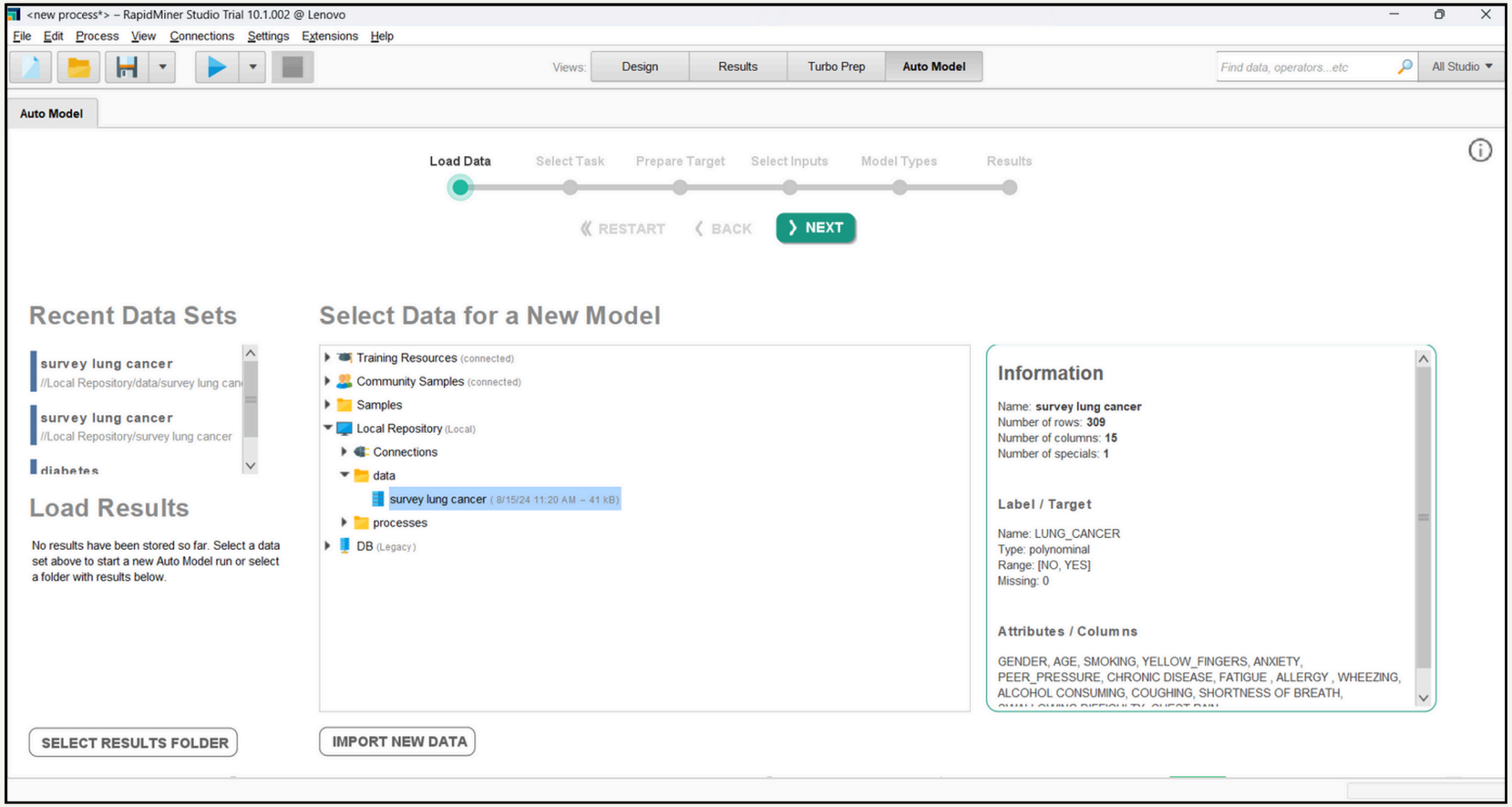
Attributes

- GENDER
- LUNG_CANCER
- CHEST PAIN
- SWALLOWING DIFFICULTY
- COUGHING
- ALCOHOL CONSUMING
- WHEEZING
- GENDER

- ALLERGY
- FATIGUE
- CHRONIC DISEASE
- PEER_PRESSURE
- ANXIETY
- YELLOW_FINGERS
- SMOKING
- AGE

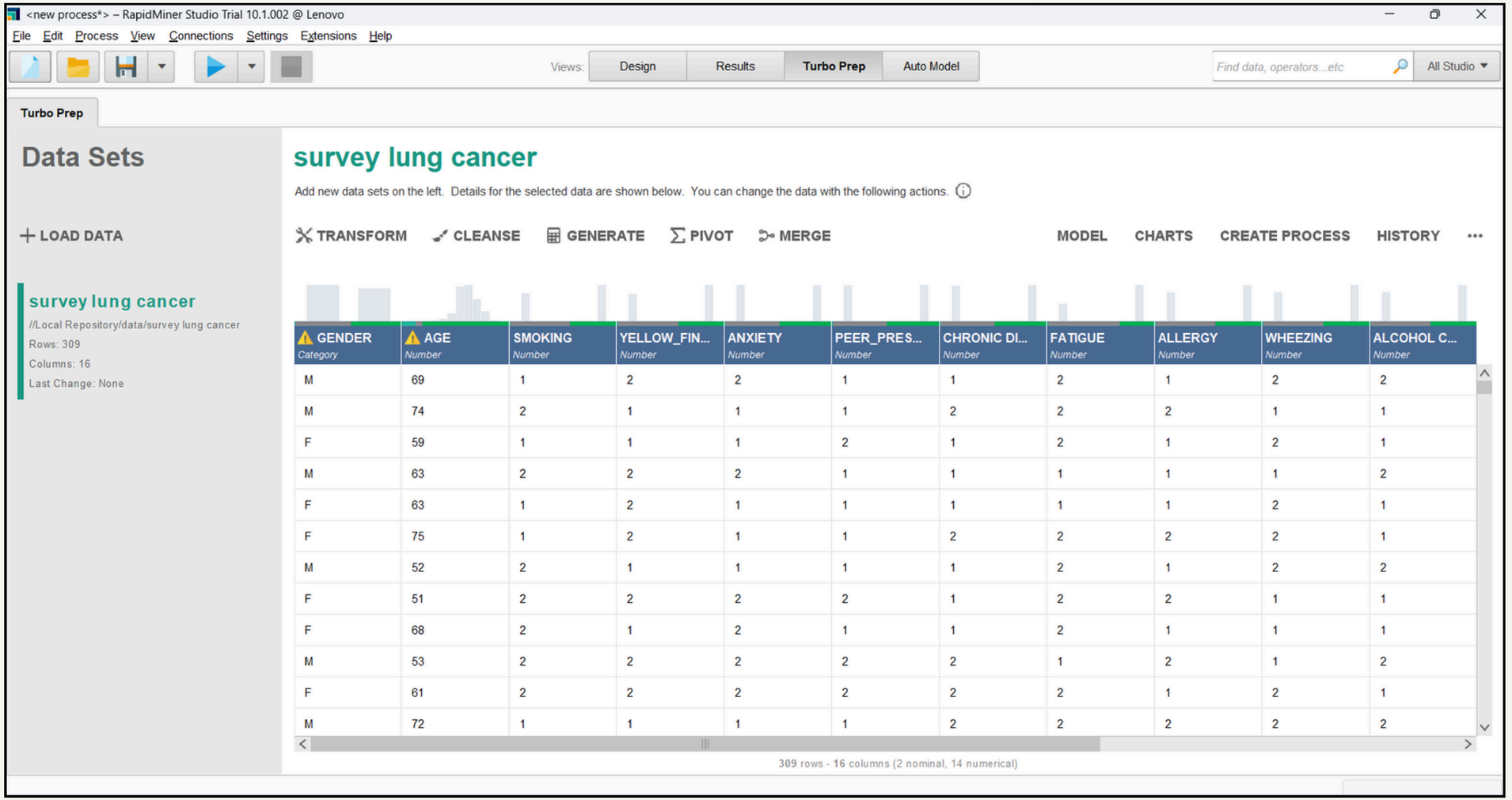
Predictive model using Automodel

Step 1. Load Your Data



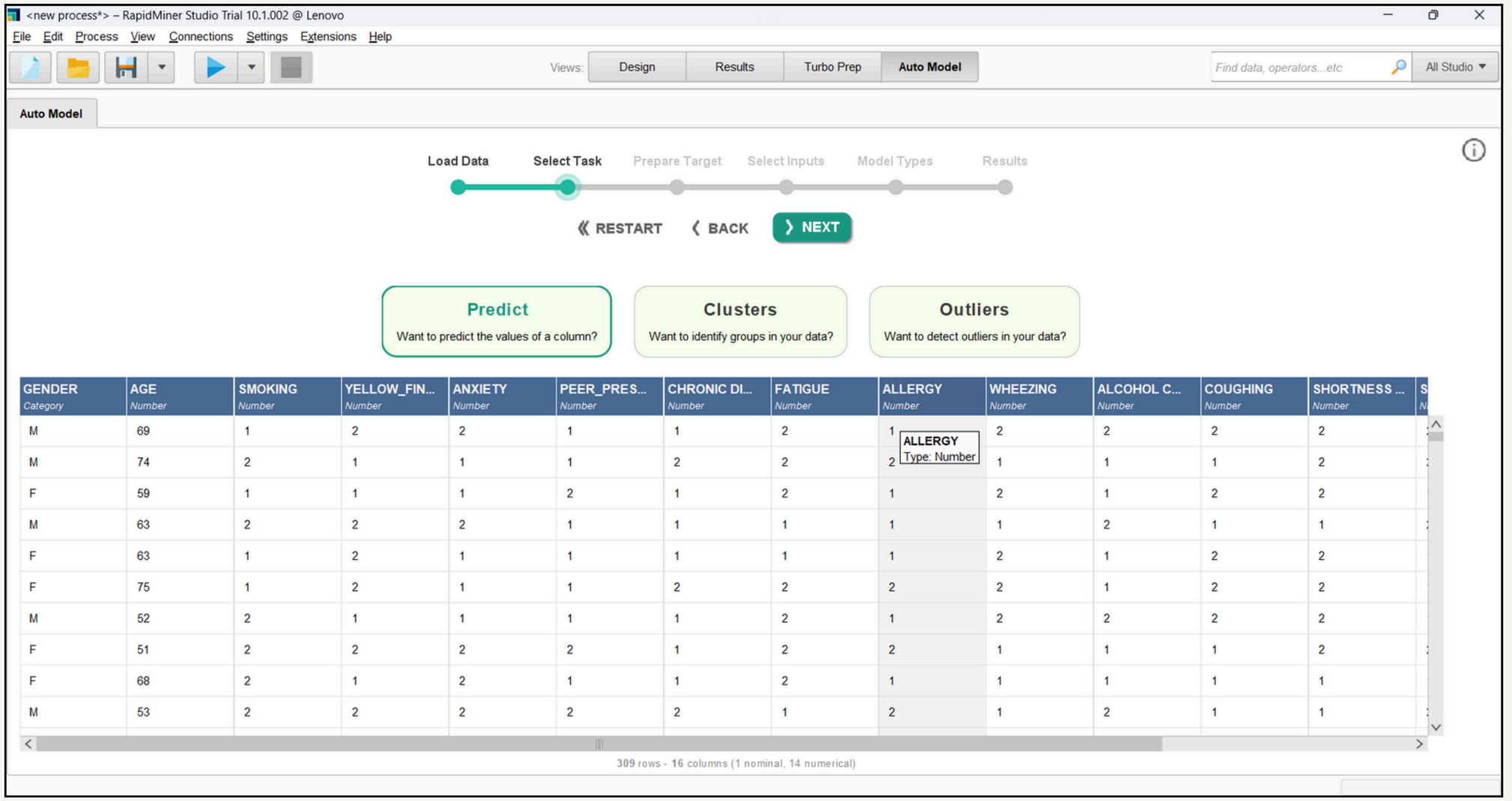
Step 2. Transform the data

Remove attributes which are not contributed to the heart attack risk.



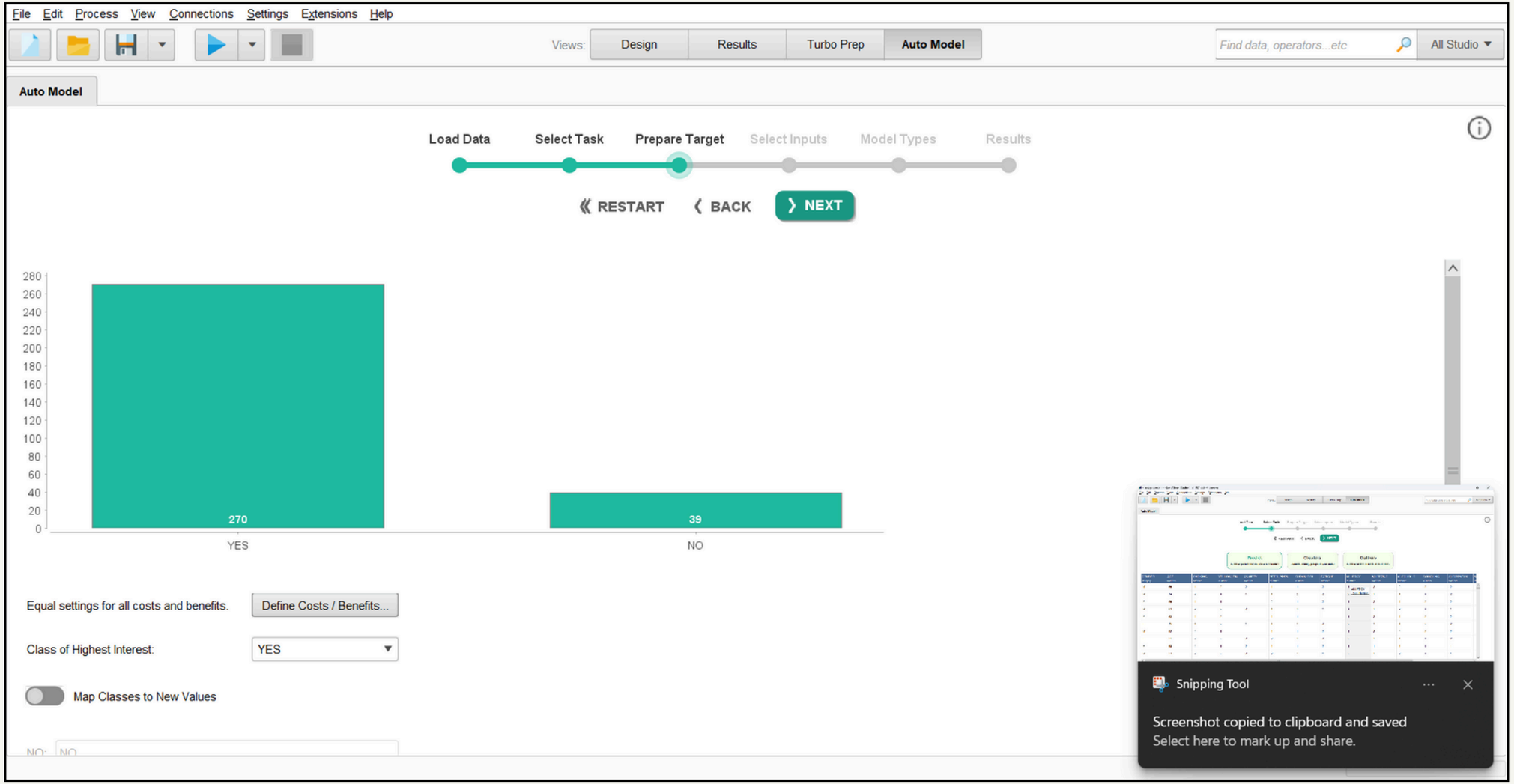
GENDER Category	AGE Number	SMOKING Number	YELLOW_FIN... Number	ANXIETY Number	PEER_PRES... Number	CHRONIC DI... Number	FATIGUE Number	ALLERGY Number	WHEEZING Number	ALCOHOL C... Number
M	69	1	2	2	1	1	2	1	2	2
M	74	2	1	1	1	2	2	2	1	1
F	59	1	1	1	2	1	2	1	2	1
M	63	2	2	2	1	1	1	1	1	2
F	63	1	2	1	1	1	1	1	2	1
F	75	1	2	1	1	2	2	2	2	1
M	52	2	1	1	1	1	2	1	2	2
F	51	2	2	2	2	1	2	2	1	1
F	68	2	1	2	1	1	2	1	1	1
M	53	2	2	2	2	2	1	2	1	2
F	61	2	2	2	2	2	2	1	2	1
M	72	1	1	1	1	2	2	2	2	2

Step 3.Load the Data and Select the task

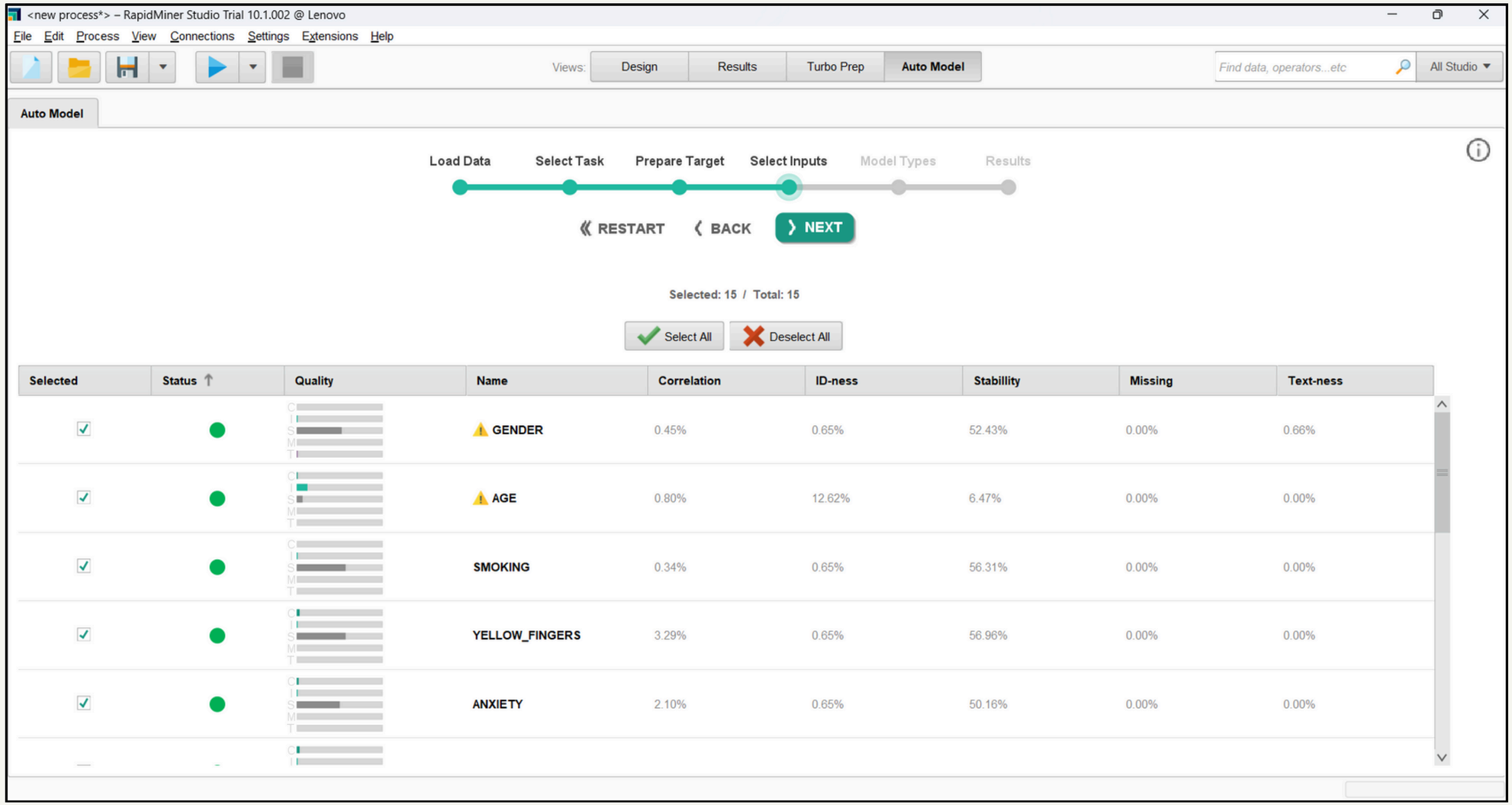


GENDER *Category*	AGE *Number*	SMOKING *Number*	YELLOW_FIN... *Number*	ANXIETY *Number*	PEER_PRES... *Number*	CHRONIC DI... *Number*	FATIGUE *Number*	ALLERGY *Number*	WHEEZING *Number*	ALCOHOL C... *Number*	COUGHING *Number*	SHORTNESS ... *Number*
M	69	1	2	2	1	1	2	1	2	2	2	2
M	74	2	1	1	1	2	2	2	1	1	1	2
F	59	1	1	1	2	1	2	1	2	1	2	2
M	63	2	2	2	1	1	1	1	1	2	1	1
F	63	1	2	1	1	1	1	1	2	1	2	2
F	75	1	2	1	1	2	2	2	2	1	2	2
M	52	2	1	1	1	1	2	1	2	2	2	2
F	51	2	2	2	2	1	2	2	1	1	1	2
F	68	2	1	2	1	1	2	1	1	1	1	1
M	53	2	2	2	2	2	1	2	1	2	1	1

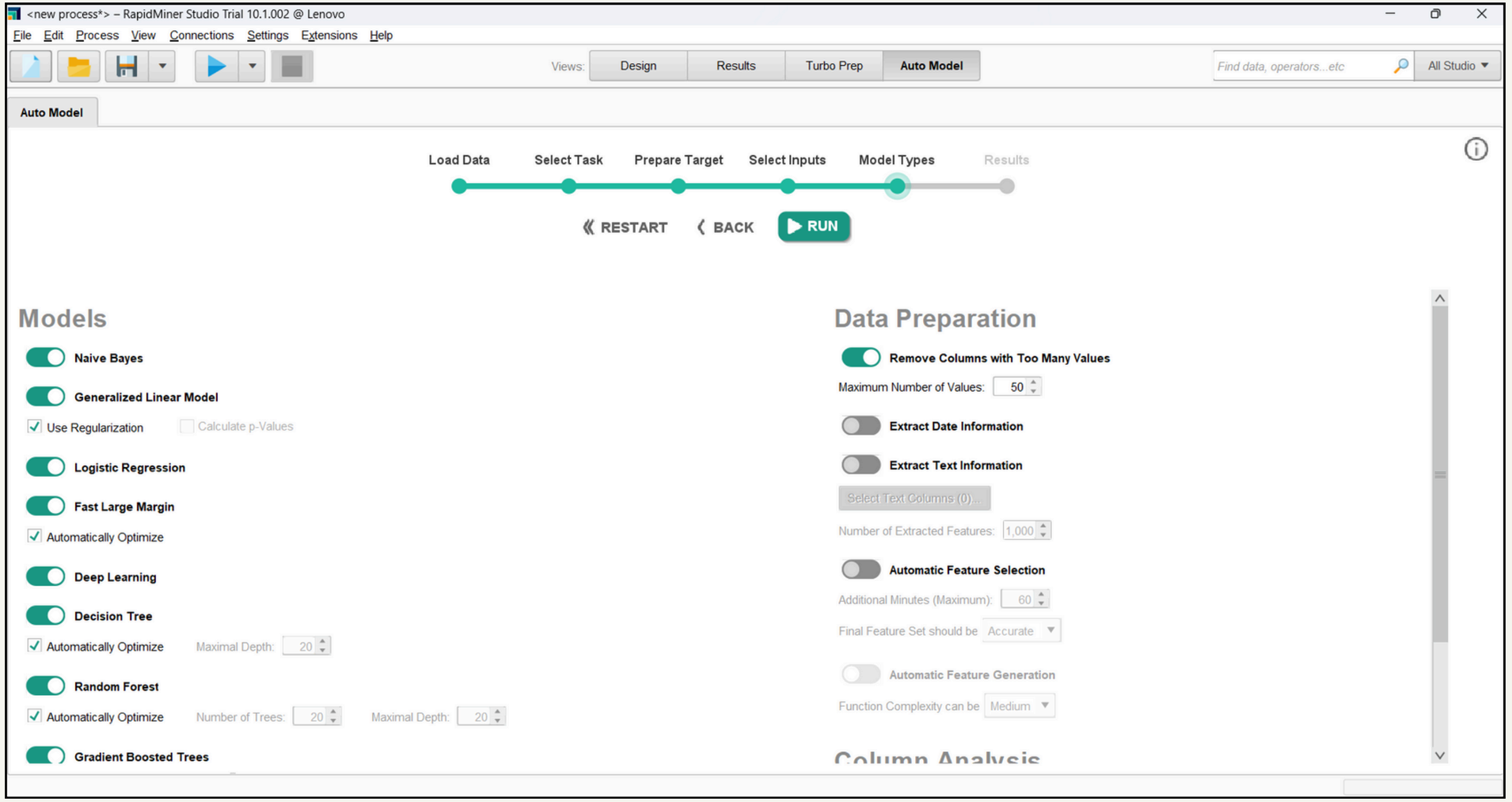
Step 4.Prepare Target



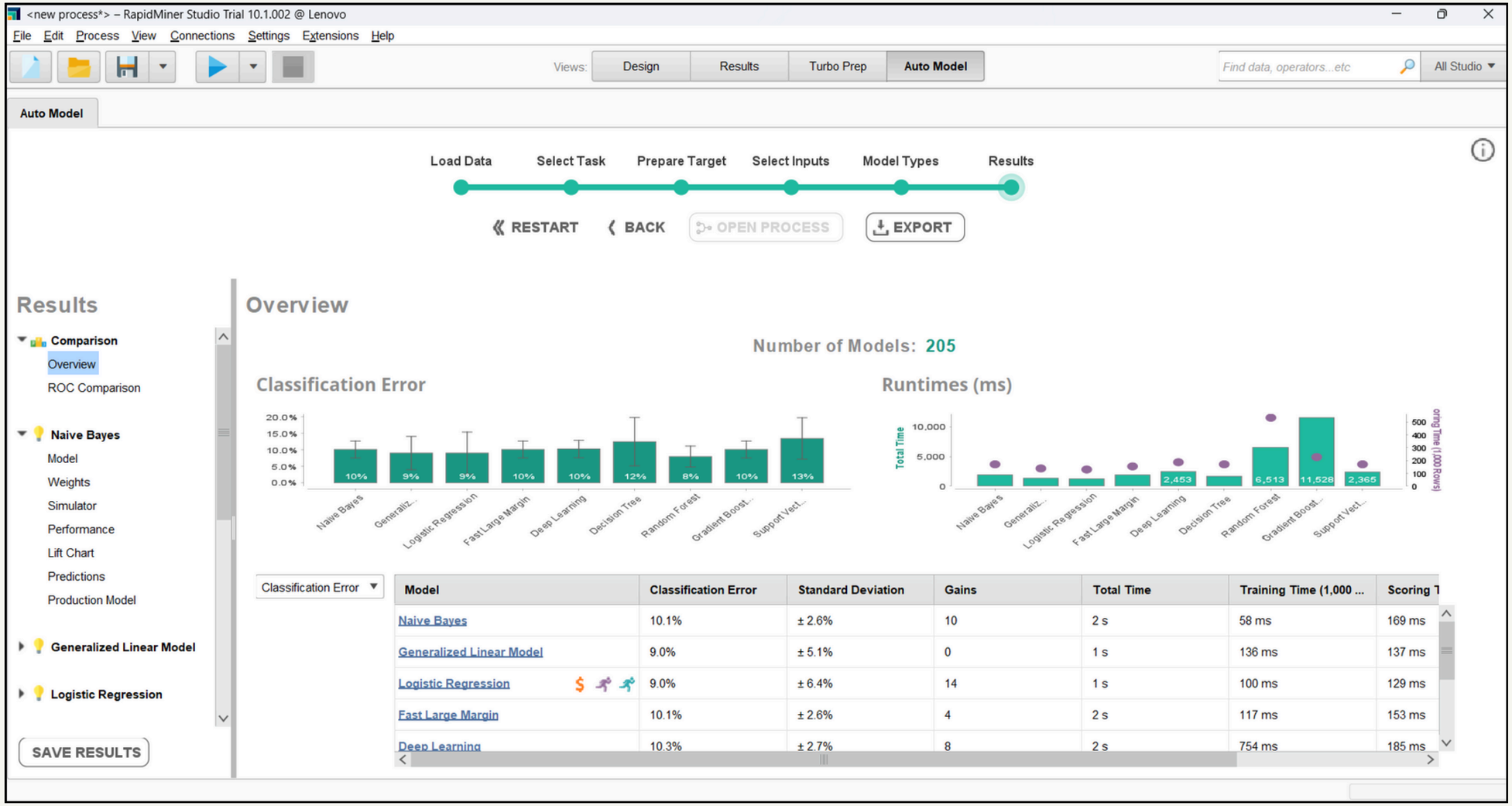
Step 5. Select Inputs



Step 6. Select Model type



Result Page


<new process*> – RapidMiner Studio Trial 10.1.002 @ Lenovo
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
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
Comparison
Overview
ROC Comparison
Naive Bayes
Model
Weights
Simulator
Performance
Lift Chart
Predictions
Production Model
Generalized Linear Model
Logistic Regression
SAVE RESULTS
Overview
Number of Models: 205
Classification Error
20.0%
15.0%
10.0%
5.0%
0.0%
10%
9%
9%
10%
10%
12%
8%
10%
13%
Naive Bayes
Generaliz...
Logistic Regression
Fast Large Margin
Deep Learning
Decision Tree
Random Forest
Gradient Boost...
Support Vect...
Runtimes (ms)
Total Time
10,000
5,000
0
2,453
6,513
11,528
2,365
Classification Error
Model
Classification Error
Standard Deviation
Gains
Total Time
Training Time (1,000 ...
Scoring T
Naive Bayes
10.1%
± 2.6%
10
2 s
58 ms
169 ms
Generalized Linear Model
9.0%
± 5.1%
0
1 s
136 ms
137 ms
Logistic Regression
9.0%
± 6.4%
14
1 s
100 ms
129 ms
Fast Large Margin
10.1%
± 2.6%
4
2 s
117 ms
153 ms
Deep Learning
10.3%
± 2.7%
8
2 s
754 ms
185 ms

<new process*> – RapidMiner Studio Trial 10.1.002 @ Lenovo

File Edit Process View Connections Settings Extensions Help

Views: Design | Results | Turbo Prep | **Auto Model**

Find data, operators...etc | All Studio

Auto Model

Load Data — Select Task — Prepare Target — Select Inputs — Model Types — Results

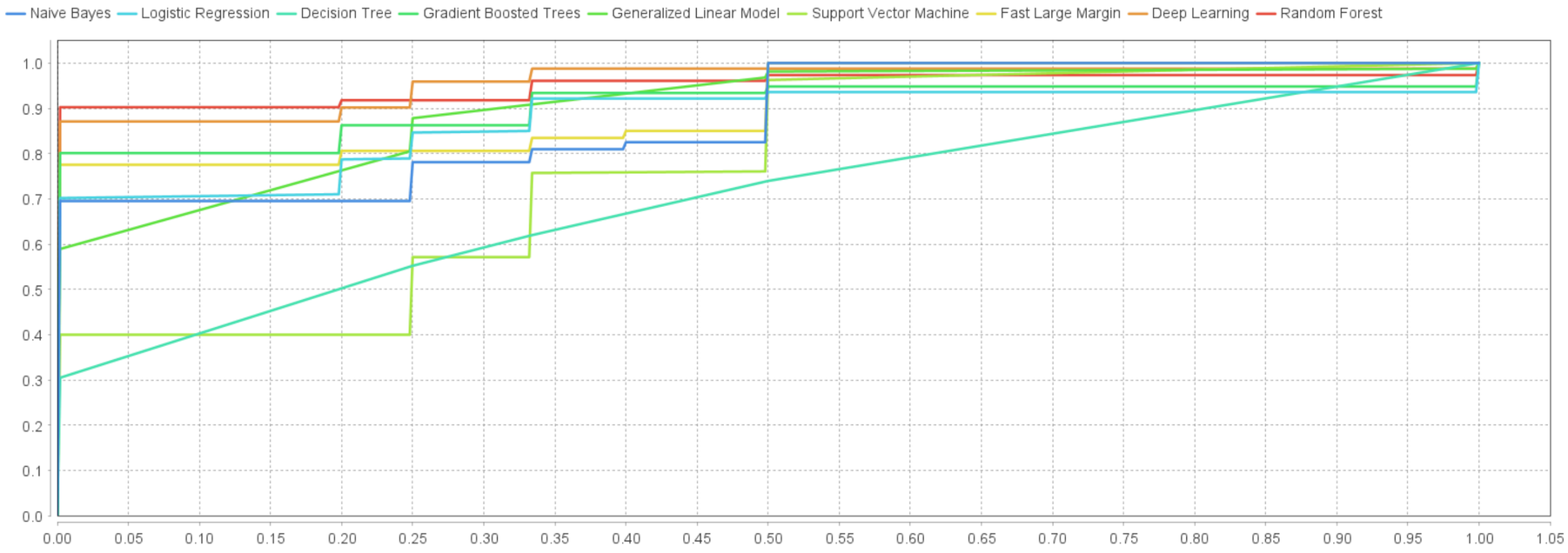
RESTART | BACK | OPEN PROCESS | EXPORT

Results

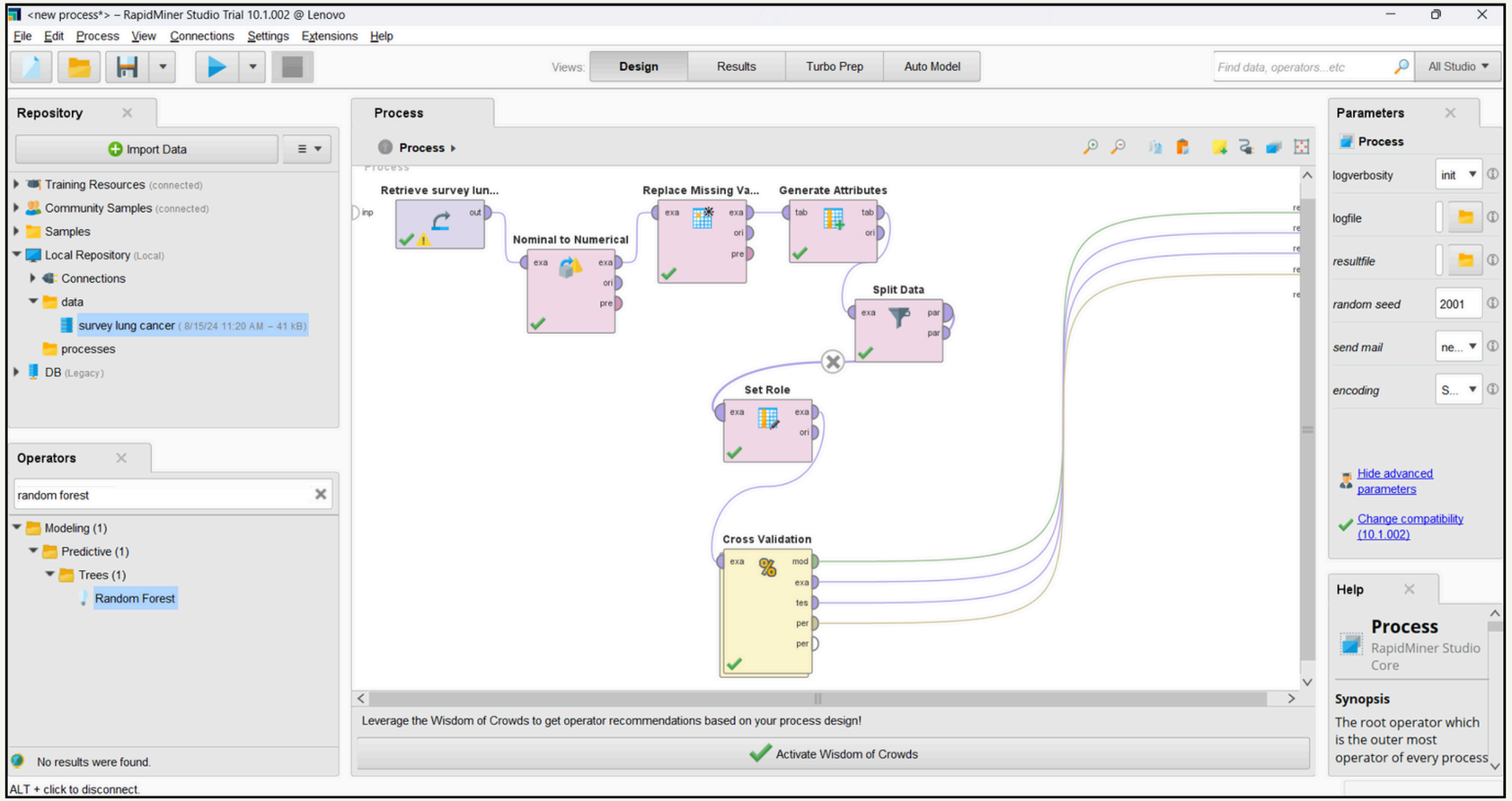
- **Comparison**
  - Overview
  - ROC Comparison
- **Naive Bayes**
- **Generalized Linear Model**
- **Logistic Regression**
- **Fast Large Margin**
- **Deep Learning**
- **Decision Tree**
- **Random Forest**

SAVE RESULTS

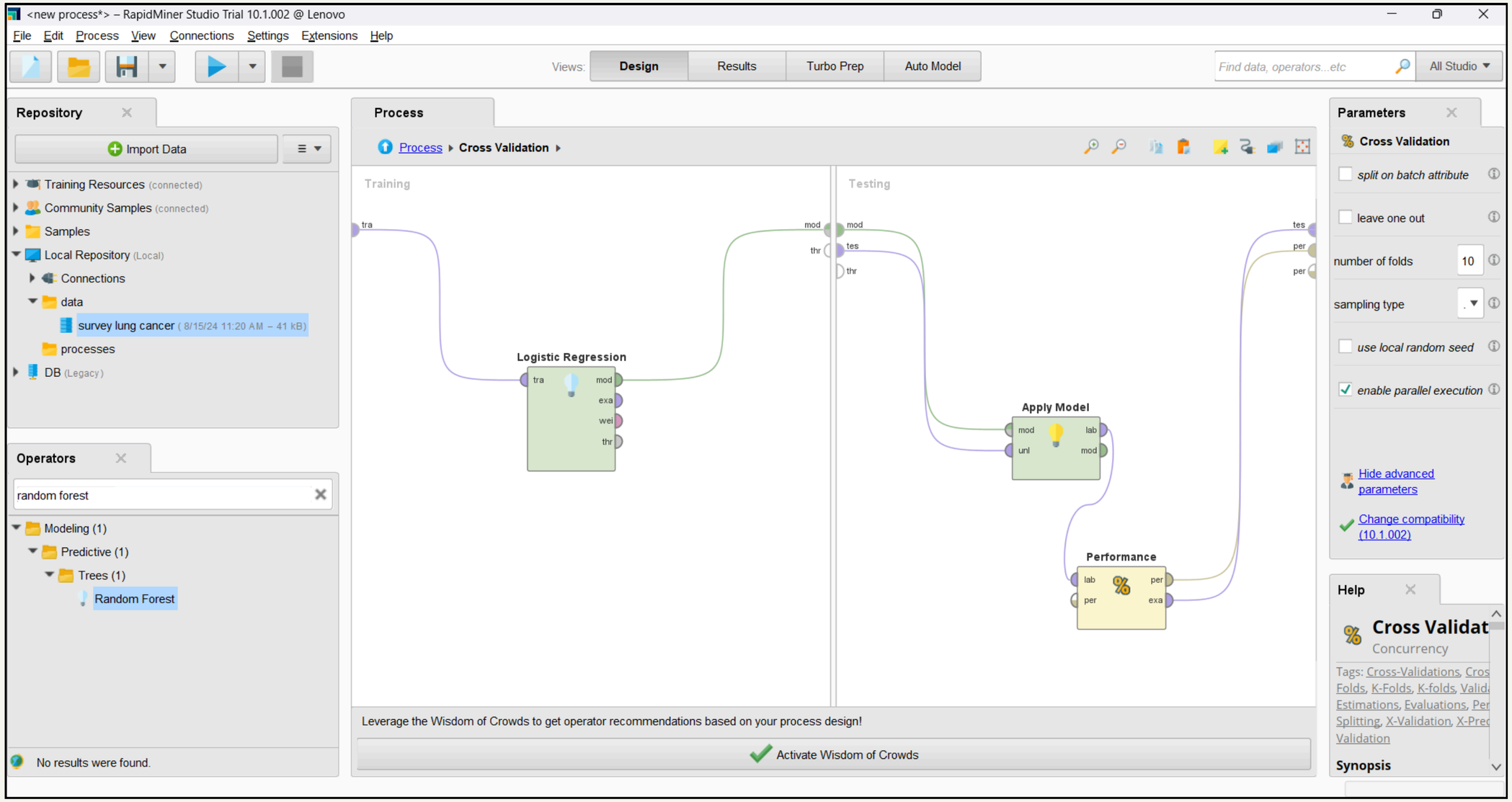
ROC Comparison



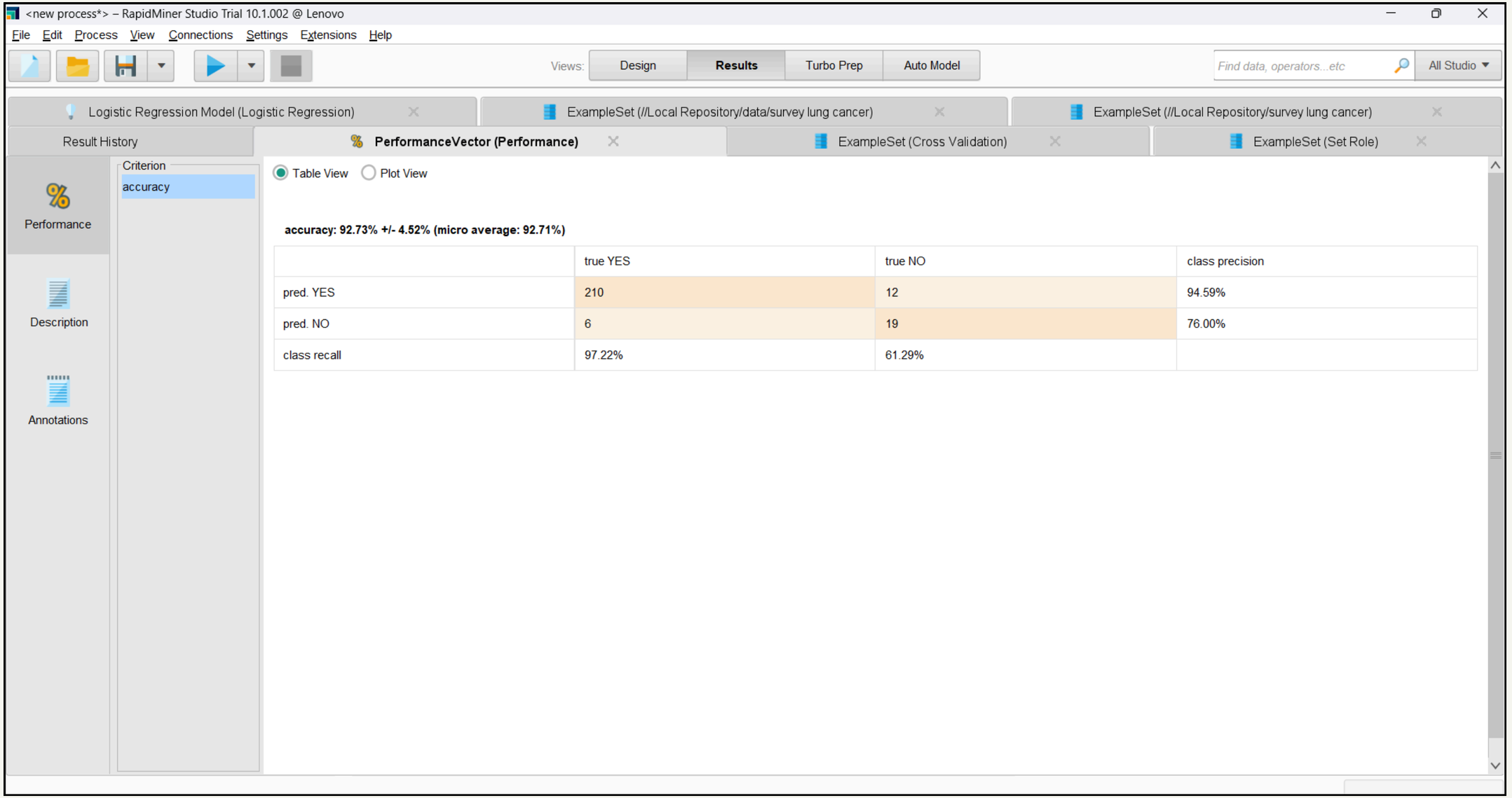
Predictive model using Process Panel

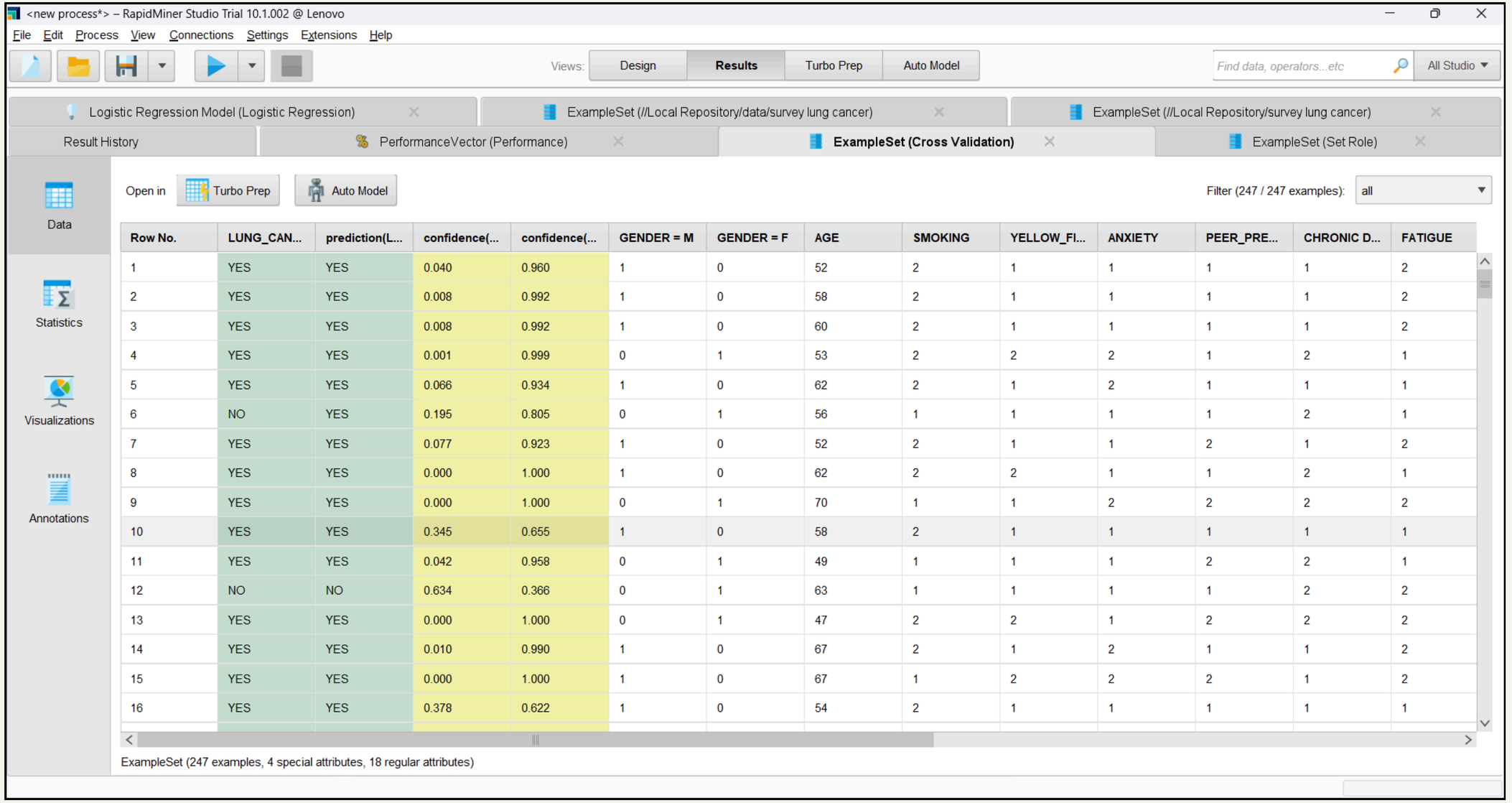


In cross validation



when we Run


<new process*> – RapidMiner Studio Trial 10.1.002 @ Lenovo
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Logistic Regression Model (Logistic Regression)
ExampleSet (//Local Repository/data/survey lung cancer)
ExampleSet (//Local Repository/survey lung cancer)
Result History
PerformanceVector (Performance)
ExampleSet (Cross Validation)
ExampleSet (Set Role)
Performance
Description
Annotations
Criterion
accuracy
Table View
Plot View
accuracy: 92.73% +/- 4.52% (micro average: 92.71%)
true YES
true NO
class precision
pred. YES
210
12
94.59%
pred. NO
6
19
76.00%
class recall
97.22%
61.29%



Row No.	LUNG_CAN...	prediction(L...	confidence(...	confidence(...	GENDER = M	GENDER = F	AGE	SMOKING	YELLOW_FI...	ANXIETY	PEER_PRE...	CHRONIC D...	FATIGUE
1	YES	YES	0.040	0.960	1	0	52	2	1	1	1	1	2
2	YES	YES	0.008	0.992	1	0	58	2	1	1	1	1	2
3	YES	YES	0.008	0.992	1	0	60	2	1	1	1	1	2
4	YES	YES	0.001	0.999	0	1	53	2	2	2	1	2	1
5	YES	YES	0.066	0.934	1	0	62	2	1	2	1	1	1
6	NO	YES	0.195	0.805	0	1	56	1	1	1	1	2	1
7	YES	YES	0.077	0.923	1	0	52	2	1	1	2	1	2
8	YES	YES	0.000	1.000	1	0	62	2	2	1	1	2	1
9	YES	YES	0.000	1.000	0	1	70	1	1	2	2	2	2
10	YES	YES	0.345	0.655	1	0	58	2	1	1	1	1	1
11	YES	YES	0.042	0.958	0	1	49	1	1	1	2	2	1
12	NO	NO	0.634	0.366	0	1	63	1	1	1	1	2	2
13	YES	YES	0.000	1.000	0	1	47	2	2	1	2	2	2
14	YES	YES	0.010	0.990	1	0	67	2	1	2	1	1	2
15	YES	YES	0.000	1.000	1	0	67	1	2	2	2	1	2
16	YES	YES	0.378	0.622	1	0	54	2	1	1	1	1	1