HEALTH CARE ANALYTICS

- **Background:**

  Diabetes Mellitus is an increasingly prevalent chronic disease characterized by the body's inability to metabolize glucose. The healthcare industry generates abundant amount of data with respect to patient. So, machine learning can be seen as an effective way to predict diabetes for better awareness and prevention.

- **Problem Objective:**

  To predict if an individual is susceptible to the onset of diabetes by taking into consideration of their lifestyle behaviors, genetic and clinical factors by using supervised learning principle of machine learning algorithm.

- **Methodology:**

  Using a google survey we collected data related to lifestyle habits, certain clinical factors of Indians and built predictive models of machine learning. Predictive modeling and performance analysis is done for the same using Rapid Miner tool for data analytics.

HEALTH CARE ANALYTICS

- **Software used:**

Rapid Miner Data Analytics tool version 9.8

- **Dataset Source:**

Google survey on Lifestyle habits of Indians

- **Machine Learning Algorithm:**

supervised learning model(Knn, Naïve Bayes etc..)

HEALTHCARE ANALYTICS

Attribute description of dataset used

Feature	Description
age	Age in years
Gender	Gender Orientation
Weight	Weight in Kilograms
Height	Height in Centimenters
cholestrol	Any history of having Cholestrol
hypertension	Having hypertension or not
thyroid	Having thyroid or not
vaccine for tuberculosis thrush	Whether having history of vaccination
blood group	Blood group type
hours of sleeping	No. of hours of sleeping
Drinking/smoking habits	Any lifestyle habit of drinking/smoking
When do you take heavy foods	Excessive food intake period
Frequency of restaurant visits	No. of times having food at a restaurant
intake of deep fried foods	How often the consumption of deep fried foods
intake of millets	Intake of healthy foods
intake of sprouts	Intake of healthy foods
Are you at medication for any disease?	History of known illness and medication
Frequency of exercise	No. of times doing exercise
Frequency of yoga	No. of times a person practises yoga
intake of nuts and spices	How often a person consumes healthy spices and nuts
nature of job	Either a sedantary lifestyle or active lifestyle
vegetarian/non vegetarian	Plant based or animal based food consumption
Mode of transport	Active or passive transportation
Diabetes (Class Variable)	Whether the concerned individual has diabetes or not

HEALTHCARE ANALYTICS

Sample Of The Dataset Used

Row No.	Gender	age	weight	height	cholestrol	hypertension	thyroid	vaccine for t...	Blood group
1	Male	42	78	162	No	Yes	No	I do not know	A1B +ve
2	Male	20	69	172	No	No	No	Yes	O+ve
3	Male	21	72	165	No	No	No	I do not know	O+ve
4	Male	21	68	5.800	No	No	No	No	O+
5	Male	21	55	176	No	No	No	Yes	A+
6	Male	21	70	180	No	No	No	I do not know	B+
7	Male	21	100	177	Yes	No	No	No	A+
8	Male	21	63	168	No	No	No	I do not know	O+
9	Male	21	80	178	No	No	No	Yes	B+
10	Male	21	81	184	No	No	No	I do not know	B -ve
11	Male	21	75	178	No	No	No	I do not know	O +ve
12	Female	20	59	161.544	No	No	Yes	Yes	O+ve
13	Male	21	70	175	No	No	No	I do not know	A+
14	Male	39	81	169	No	Yes	No	Yes	A1B+

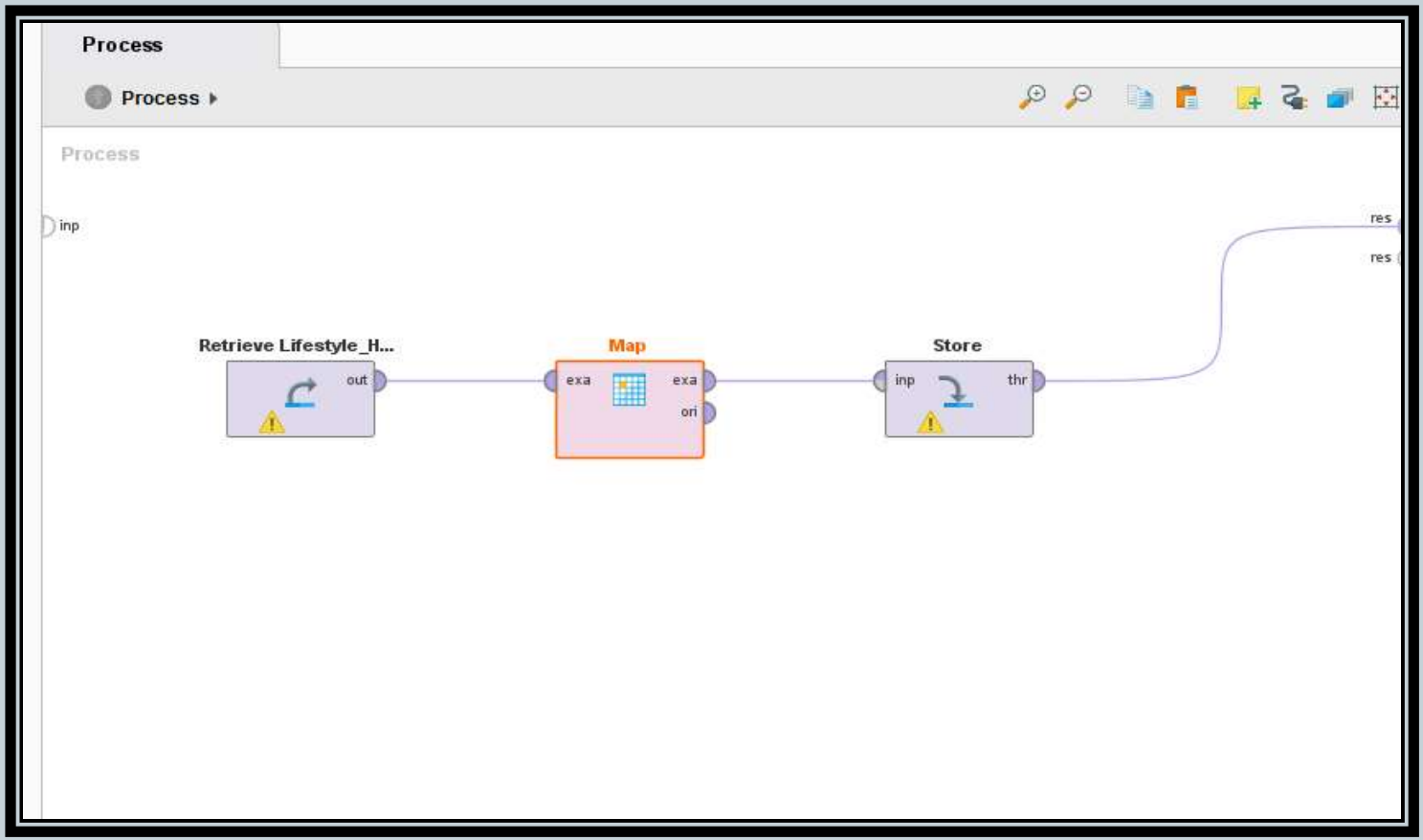
HEALTHCARE ANALYTICS

Sample Of The Dataset Used

intake of de...	intake of mil...	intake of spr...	Are you at m...	Frequency o...	Frequency o...	intake of nut...	nature of job	vegetarian/n...	What is your...	Diabetes
Yes	No	No	No	monthly	rarely	rarely	involves mor...	mixed	bike	No
Yes	Yes	Yes	No	rarely	i dont do yoga	regularly	almost sitting	non-vegetarian	car	No
Yes	No	Yes	No	monthly	twice in a week	regularly	almost sitting	non-vegetarian	bike	No
Yes	rarely	rarely	No	daily	rarely	rarely	physical	non-vegetarian	bus	No
Maybe	rarely	rarely	No	twice in a week	i dont do yoga	rarely	involves mor...	non-vegetarian	car	No
Yes	rarely	rarely	No	rarely	rarely	regularly	almost sitting	mixed	bus	No
Yes	Yes	No	No	twice in a week	rarely	regularly	involves mor...	non-vegetarian	by walk	No
Yes	rarely	Yes	No	i dont do exer...	i dont do yoga	rarely	almost sitting	mixed	bike	No
Yes	Yes	Yes	No	twice in a week	twice in a week	regularly	almost sitting	non-vegetarian	bike	No
Maybe	rarely	rarely	No	daily	rarely	regularly	involves mor...	eggitarian	car	No
Yes	Yes	rarely	No	daily	rarely	regularly	involves mor...	vegetarian	bus	No
Maybe	Yes	rarely	No	rarely	rarely	rarely	almost sitting	vegetarian	by walk	No
Maybe	rarely	rarely	No	twice in a week	rarely	rarely	physical	non-vegetarian	bike	No
Yes	No	No	No	rarely	i dont do yoga	rarely	involves mor...	non-vegetarian	bus	No

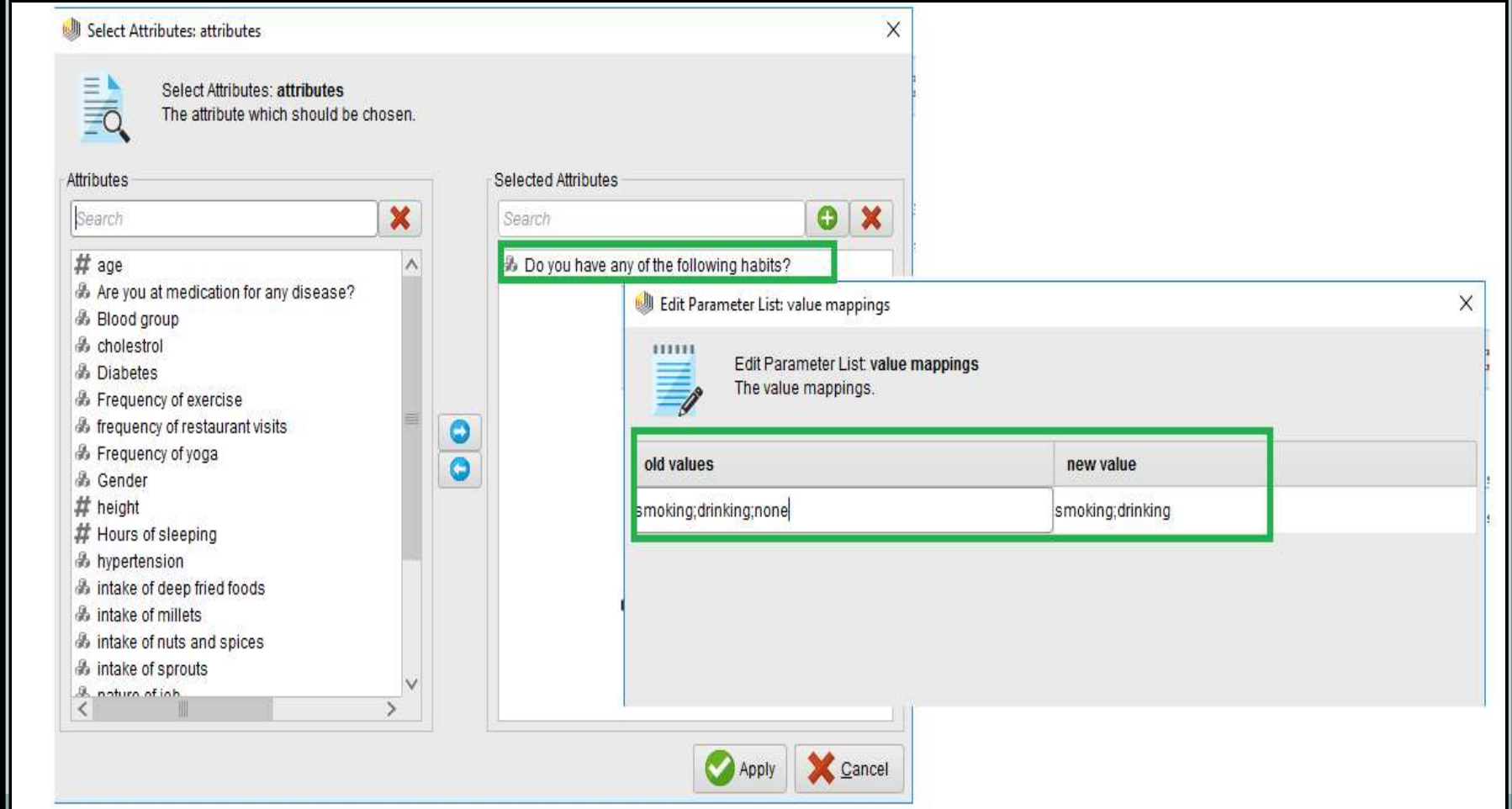
HEALTHCARE ANALYTICS

- **Data Preprocessing**



HEALTHCARE ANALYTICS

- **Data Preprocessing**



HEALTHCARE ANALYTICS

- **Checking for any missing values**

Filter (24 / 24 attributes): Search for Attributes

Name	Type	Missing	Statistics		
height	Real	0	4	590	154.449
cholestrol	Polynominal	0	Least Yes (29)	Most No (256)	Values No (256), Yes (29)
hypertension	Polynominal	0	Least Yes (29)	Most No (256)	Values No (256), Yes (29)
thyroid	Polynominal	0	Least Yes (13)	Most No (272)	Values No (272), Yes (13)
vaccine for tuberculosis	Polynominal	0	Least Yes (86)	Most No (112)	Values No (112), I do not know (87), ...[1 mo
Blood group	Polynominal	0	Least o+ (1)	Most O+ (53)	Values O+ (53), B+ (49), ...[52 more]
Hours of sleeping	Integer	0	Min 4	Max 24	Average 7.544
Do you have any of the followi...	Polynominal	0	Least smoking; [...] ;none (1)	Most none (247)	Values none (247), drinking (19), ...[3 more]

HEALTHCARE ANALYTICS

Why Supervised learning?

Supervised learning is used because the output is given to the model. In supervised learning, both input and output are known. After processing, the actual output is compared with required outputs ie, the algorithm learns on a labeled dataset, providing an answer key that the algorithm can use to evaluate its accuracy on training data.

HEALTH CARE ANALYTICS

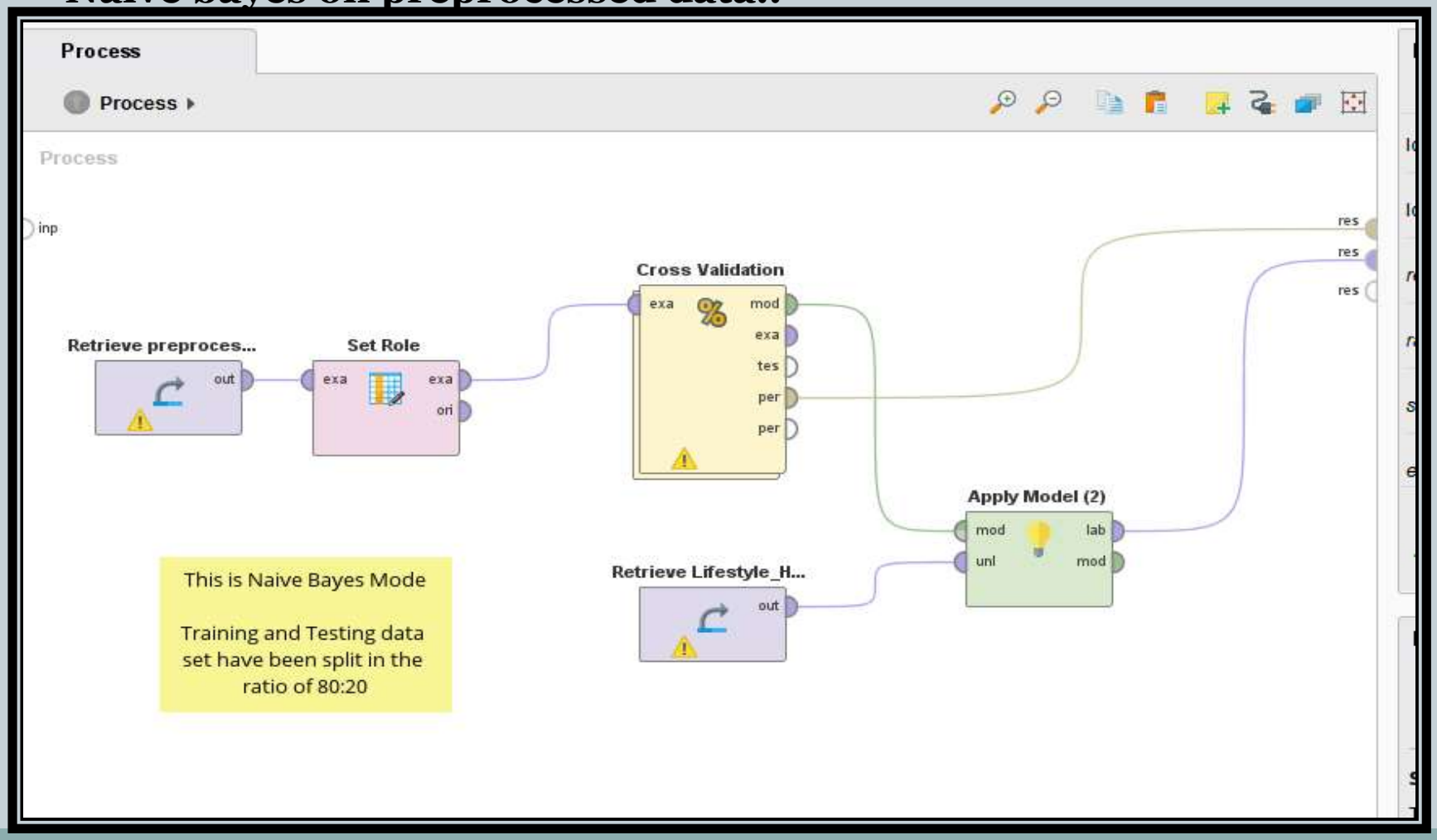
- **<u>Review 1 Implementation:</u>**

  <u>Model 2: Naïve Bayes Model:</u>

  Naive Bayes also called Bayesian theorem is a simple, effective and commonly used machine learning classifier. The algorithm calculates probabilistic results by counting the frequency and combines the value given in data set. By using Bayesian theorem, it assumes that all attributes are independent and based on variable values of classes. In real world application, the conditional independence assumption rarely holds true and gives well and more sophisticate classifier results.

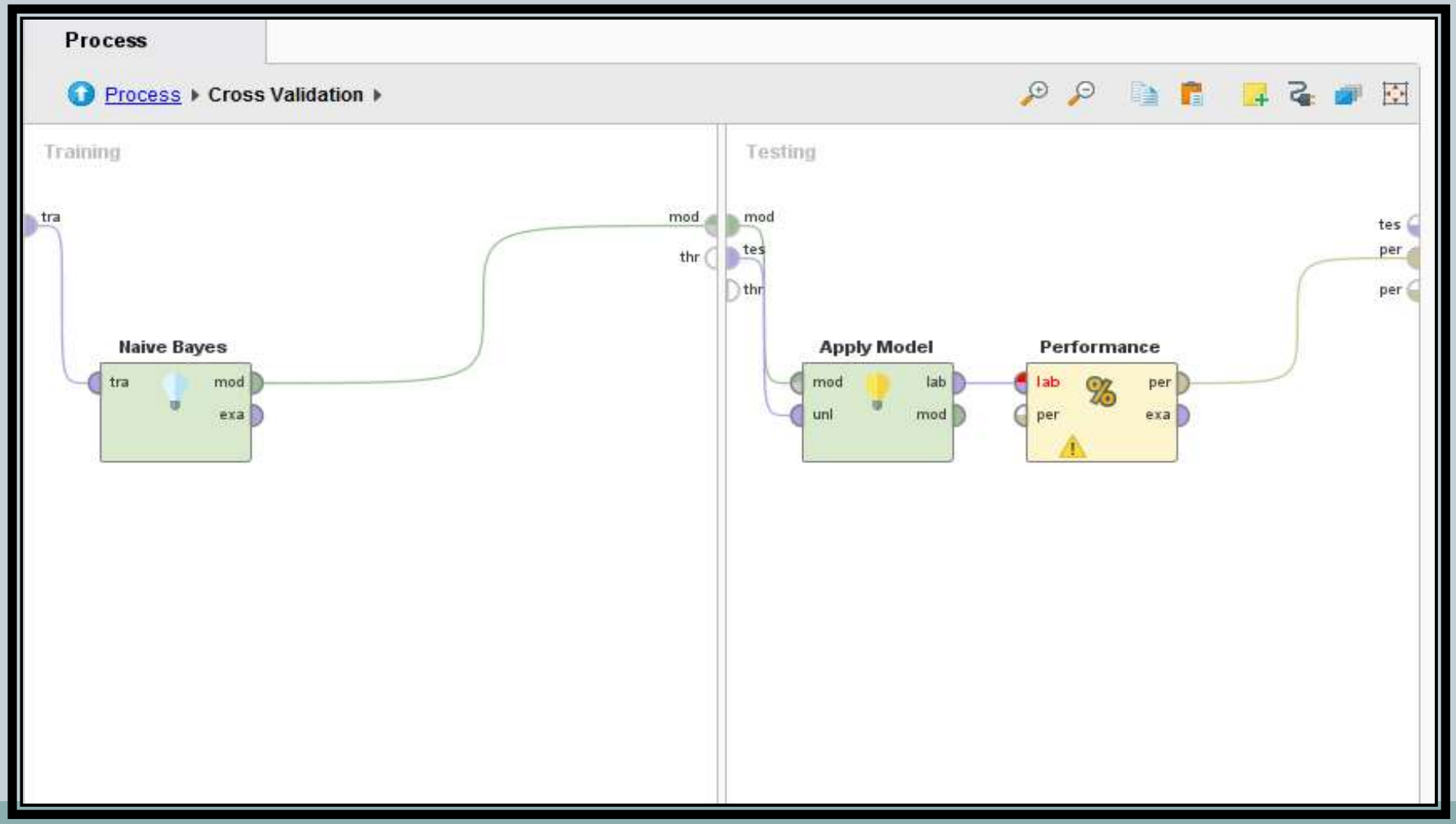
HEALTH CARE ANALYTICS

Naïve bayes on preprocessed data..



HEALTH CARE ANALYTICS

Naïve bayes on preprocessed data..



HEALTH CARE ANALYTICS

Predicted output values with the trained model

Open in Turbo Prep Auto Model

Filter (36 / 36 examples): all

Data

Statistics

Visualizations

Annotations

prediction(Diabetes) (prediction)
Polynominal
Missing: 0
Least: Yes (4)
Most: No (32)

Row No.	prediction(D...	confidence(...	confidence(...	Gender	age	weight	height	cholestrol	hypertension	thyroid
1	No			Male	20	44	150	No	No	No
2	No			Male	43	85	6	No	No	No
3	No			Male	21	67	170.700	No	No	No
4	No	0.998	0.002	Male	22	103	165	No	No	No
5	No	1.000	0.000	Male	19	57	178	No	No	No
6	No	1.000	0.000	Male	19	110	177.800	No	No	No
7	No	1.000	0.000	Male	20	48	173	No	No	No
8	No	1.000	0.000	Female	19	47	155.448	No	No	No
9	No	1.000	0.000	Female	20	46	157	No	No	No
10	No	1.000	0.000	Male	21	70	174	No	No	No
11	No	1.000	0.000	Male	23	74	168	No	No	No
12	No	0.994	0.006	Male	32	95	168	No	No	No
13	No	0.995	0.005	Male	37	86	183	No	No	No
14	No	1.000	0.000	Female	32	72	154	No	No	No

Predicted values

HEALTH CARE ANALYTICS

Performance Table:

Table View Plot View

accuracy: 91.23% +/- 4.73% (micro average: 91.23%)

	true No	true Yes	class precision
pred. No	251	13	95.08%
pred. Yes	12	9	42.86%
class recall	95.44%	40.91%	

HEALTH CARE ANALYTICS

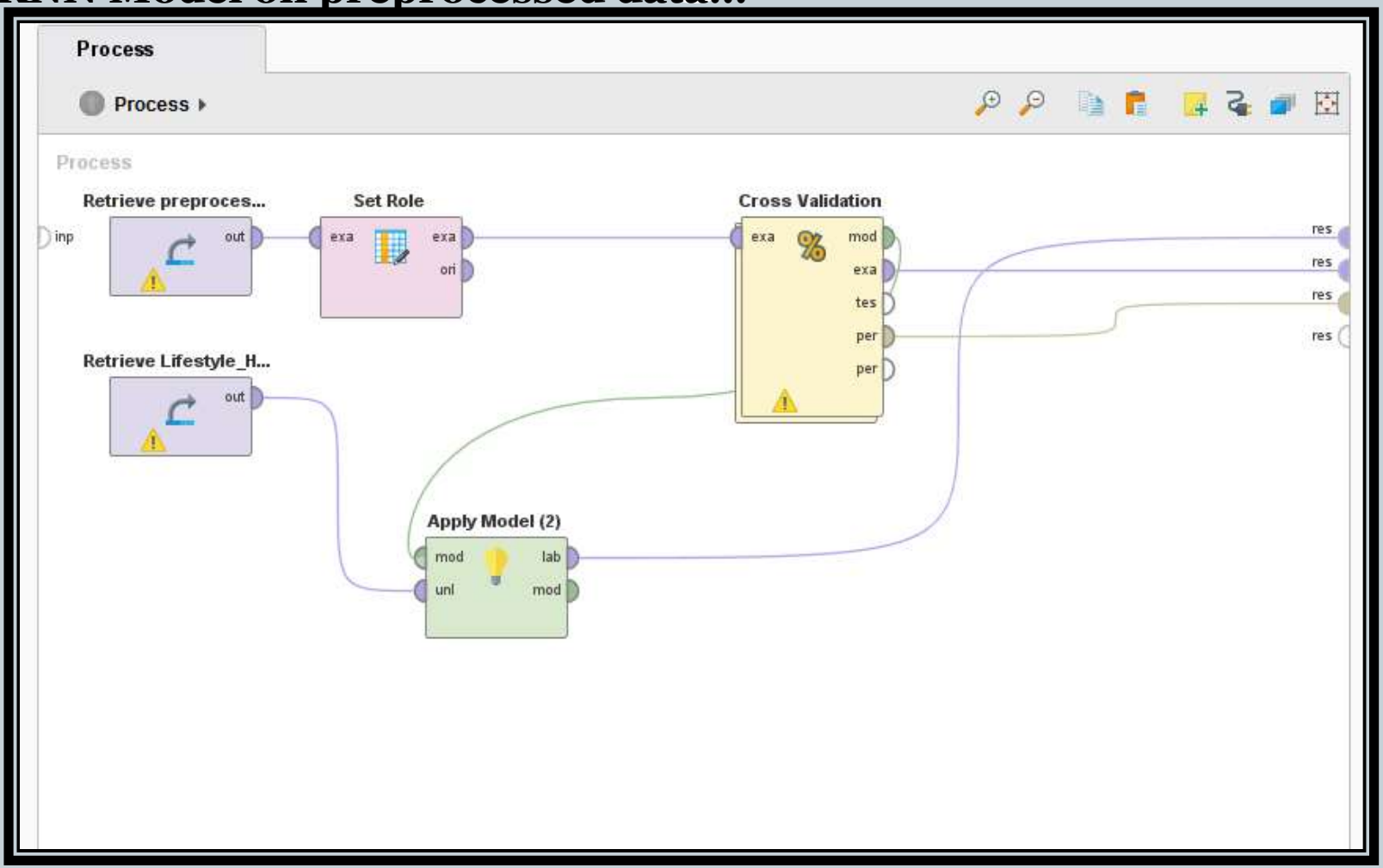
- **<u>Review 2 Implementation:</u>**

  <u>Model 2: KNN:</u>

  K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

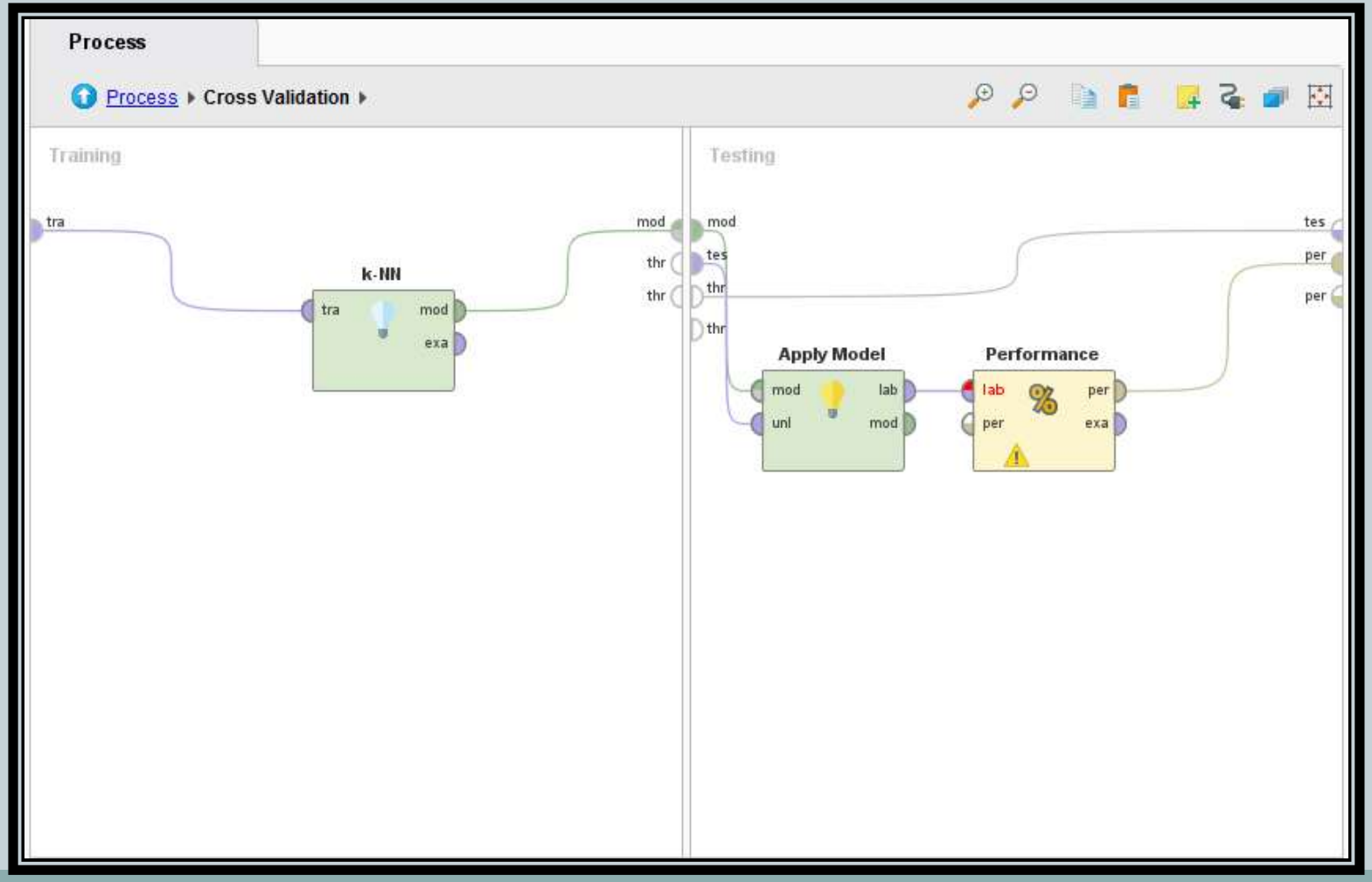
HEALTH CARE ANALYTICS

KNN Model on preprocessed data…



HEALTH CARE ANALYTICS

KNN Model on preprocessed data…



HEALTH CARE ANALYTICS

Predicted output values with the trained model

Open in Turbo Prep Auto Model

Filter (36 / 36 examples): all

Row No.	prediction(D...	confidence(...	confidence(...	Gender	age	weight	height	cholestrol	hypertension	thyroid
1	No			Male	20	44	150	No	No	No
2	No			Male	43	85	6	No	No	No
3	No			Male	21	67	170.700	No	No	No
4	No	1	0	Male	22	103	165	No	No	No
5	No	1	0	Male	19	57	178	No	No	No
6	No	1	0	Male	19	110	177.800	No	No	No
7	No	1	0	Male	20	48	173	No	No	No
8	No	1	0	Female	19	47	155.448	No	No	No
9	No	1	0	Female	20	46	157	No	No	No
10	No	1	0	Male	21	70	174	No	No	No
11	No	1	0	Male	23	74	168	No	No	No
12	No	1	0	Male	32	95	168	No	No	No
13	No	0.773	0.227	Male	37	86	183	No	No	No
14	No	1	0	Female	32	72	154	No	No	No

prediction(Diabetes) (prediction)
Polynominal
Missing: 0
Least: Yes (0)
Most: No (36)

Predicted values

ExampleSet (36 examples, 3 special attributes, 24 regular attributes)

HEALTH CARE ANALYTICS

Performance Table:

PerformanceVector (Performance) | ExampleSet (Set Role)

Table View | Plot View

accuracy: 91.22% +/- 3.01% (micro average: 91.23%)

	true No	true Yes	class precision
pred. No	254	16	94.07%
pred. Yes	9	6	40.00%
class recall	96.58%	27.27%	