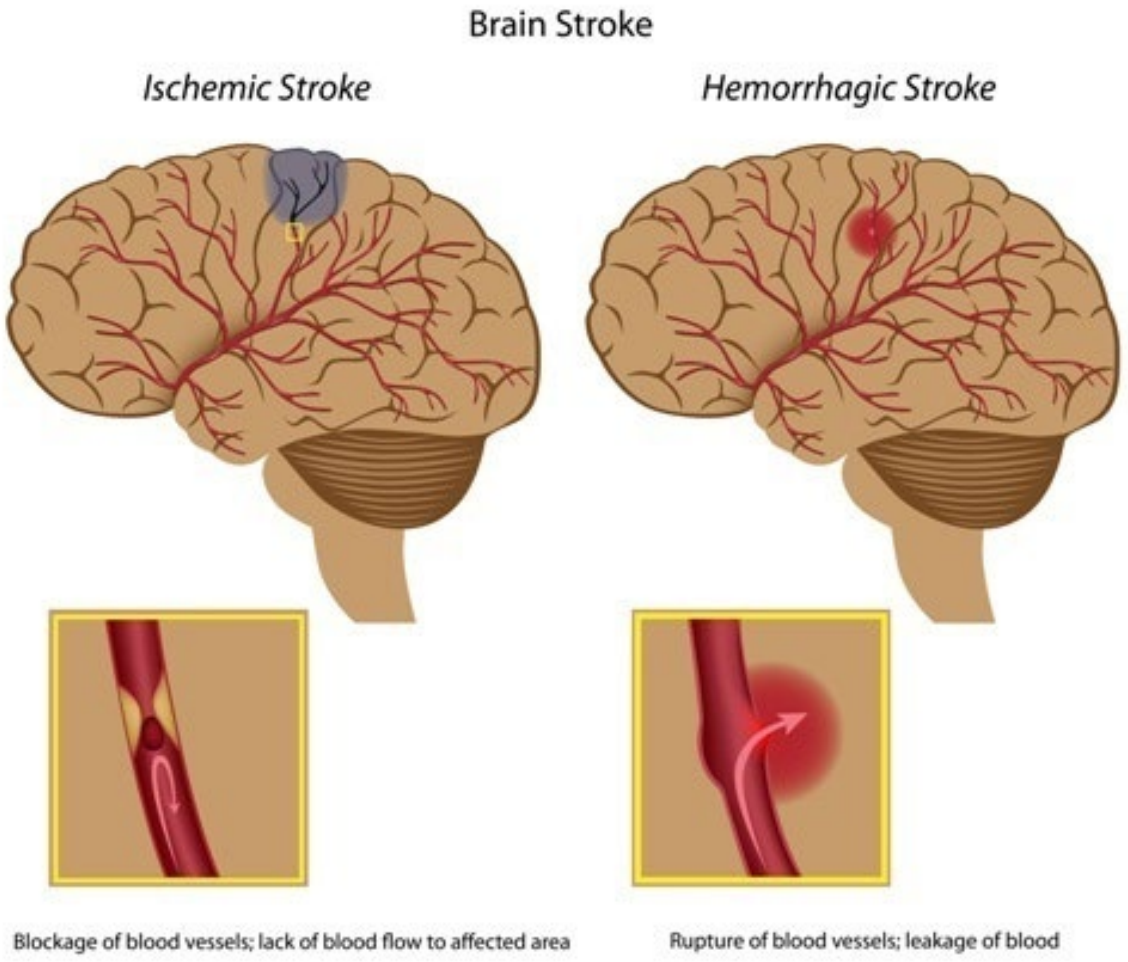
Stroke

Data Wrangling and Model Development with RapidMiner

Avery Jan

9-22-2022



87%

13%

Outline

- Executive Summary
- Introduction
- **Methodology** data source, software, operator, process, models
- **Dataset** composition and attributes
- **Visualization of Dataset** age, hypertension, heart disease, glucose level
- **Data Wrangling**

  Example exclusions, missing values handling, discretization of continuous numerical values, data type changes, attribute reduction, sampling
- **Model Development and Evaluation** five models, confusion matrices, ROC curves
- Conclusion and Appendix

Executive Summary

This project is to showcase (1) the extent of the necessary data preparation for developing models that classify whether or not a patient is likely to have a stroke (2)the types of models that can be developed using a rather imbalanced stroke dataset. That is, examples of stroke are scarce in this dataset. Three categories of issues pertaining to the dataset were addressed using operators provided by RapidMiner (RM). The first category was the imperfection in the dataset, e. g. missing values, underrepresented classes, etc. The second category was containing data inappropriate for model development. For example, incorrectly imported data types, continuous values to be used as discretized input to models. The last category includes an ambiguous, attribute and redundant attributes. Once these data issues were resolved, the processed dataset was sampled to generate a balanced subset with a ratio of stroke to no_stroke closer to 1:1. Then, a correlation analysis was performed on this subset to confirm that attributes were not strongly correlated with one another to meet the requirement of such an assumption made for developing some models. Next, using this subset as the input to RM processes which were formed with interconnected RM operators, five types of classification models (Logistic Regression, Decision Tree, k-Nearest Neighbors, Naïve Bayes, and Neural Network) were constructed. k-folds cross validation was used to test the models. The testing results were organized in confusion matrices and ROC (receiver operating characteristic curve) curves of the models were plotted. Overall, these models performed comparably at most classification thresholds with accuracies between 67% and 73%. Most models performing 70% or above accurate. Also, the area under the curve (AUC) of ROCs appeared to be similar for all models. Lastly, the specifics of three models – Logistic Regression, Decision Tree, and Neural Network, revealed Age, Hypertension, and Glucose Level being the three most significant predictors of stroke.

Introduction

A dataset of 5110 patients who either have had or never had a stroke was analyzed in this project. This dataset contains twelve attributes that cover patients' age, gender, medical history, and various aspects of lifestyle. The data in the dataset are in the form of integer, real number, binary number, or text. The data was mined using the RapidMiner Studio software (RM). Once data was imported to RM, first visualizations were created on four attributes – age, heart disease, hypertension, and glucose level to explore the dataset. These attributes were thought to have linked with the chance of having a stroke. Then, a lengthy data wrangling procedure that involves excluding inappropriate data, dealing with missing data, discretization of continuous-valued data, fixing the data types that were incorrectly interpreted by RM during the import process, defining a new attribute as needed, removing underrepresented classes etc. In addition, it was necessary to create partly balanced subsets by sampling the dataset because the dataset is biased with only 249 stroke cases out of 5110 cases. A completely balanced subset was defined as stroke:no-stroke = 1:1.

Five types of classification models were developed using one of the partly balanced subsets. The models were Logistic Regression, Decision Tree, k-Nearest Neighbors, Naïve Bayes, and Neural Network. All models were evaluated using k-folds cross validation method. The results were represented in a confusion matrix. The accuracy of these models are about 70%. ROC curves (receiver operating characteristic curve) were created to view the performance of these models at different classification thresholds. It was concluded that these models have relatively similar ROC curves. Moreover, the fact that their ROC's are deviated from the diagonal line, which represents random flip, means that these models are generally performing well. Lastly, the specifics of Logistic Regression, Decision Tree, and Neural Network provided additional insights into the factors influencing the chance of having a stroke.

Methodology

Data

Data Source: healthcare-dataset-stroke-data.xlsx https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

Data Preparation:

- A series of data wrangling tasks were performed prior to sampling.
- The cleaned dataset was sampled to generate partly balanced subsets for creating models.
- A fully balanced subset is defined as the ratio (stroke : no-stroke) being 1:1.

Software RapidMinor Studio, 2021, version 9.9 (referred to as RM hereafter)

- **Simple Operator** (represented as a rectangle in a process flowchart): A building block of processes that has input and output ports; it can perform actions like data access, cleaning, modeling, etc.; the action performed on the input ultimately leads to what is supplied to the output. https://docs.rapidminer.com/latest/studio/operators/
- **Stacked Operator** (stacked rectangles): A combination of multiple operators that performs a complex multi-step task
- **Naming of Operator**: Some operators retained their names given by RM while others were renamed by the author.
- **Process** (shown as a flowchart): A set of interconnected operators represented by a flow design to complete a task
- **File Format:** A file with .rmp extension is created for each process performed in RM

Approach Data Overview → Data wrangling → Correlation Analysis → Model Development → Model Evaluation

Dataset Overview

The Data

(5110 Examples; 249 Stroke Examples)

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1
27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1
12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1
5317	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smoked	1
58202	Female	50	1	0	Yes	Self-employed	Rural	167.41	30.9	never smoked	1
56112	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes	1
34120	Male	75	1	0	Yes	Private	Urban	221.29	25.8	smokes	1
27458	Female	60	0	0	No	Private	Urban	89.22	37.8	never smoked	1
25226	Male	57	0	1	No	Govt_job	Urban	217.08	N/A	Unknown	1
70630	Female	71	0	0	Yes	Govt_job	Rural	193.94	22.4	smokes	1
13861	Female	52	1	0	Yes	Self-employed	Urban	233.29	48.9	never smoked	1

Attributes

Special Attribute

id (id):	a unique value for each Example
stroke (label):	0, 1

Personal Data

gender:	Male, Female, Other (1 Example)
age:	0.08 – 19 (966), 20-82

Medical Data

hypertension:	0, 1
heart_disease:	0, 1
avg_glucose_level:	55.12 - 271.74
bmi:	10.3 – 97.6, N/A (201)
smoking_status:	never smoked, formerly smoked, smokes, Unknown (1544)

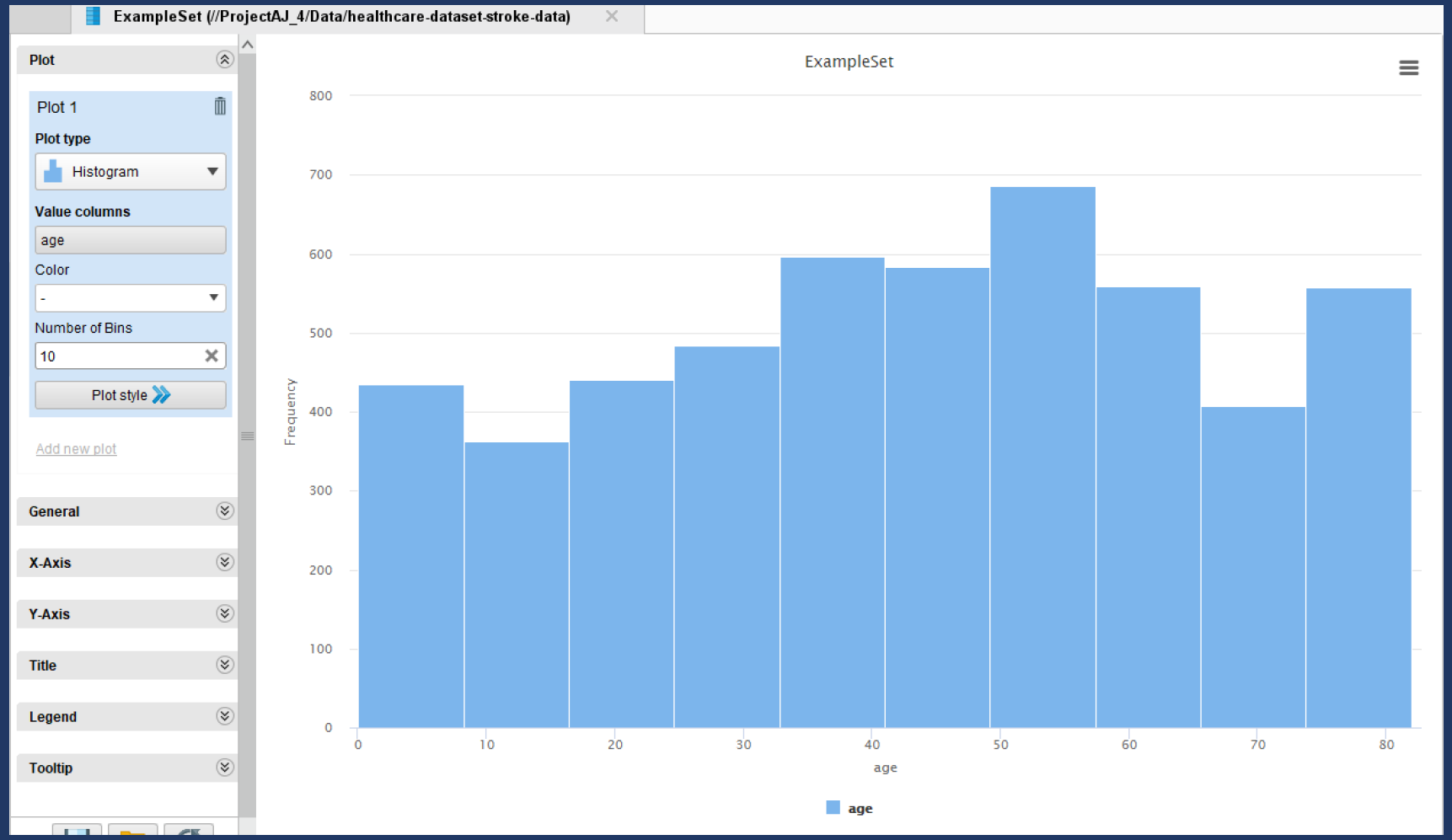
Lifestyle Data

ever_married:	Yes, No
work_type:	Private, Self-employed, Govt_job, Never_worked (22), children (687)
Residence_type:	Urban, Rural

* The label attribute (stroke) will be predicted by the models while the other attributes will be used in the models as the predictors.

age

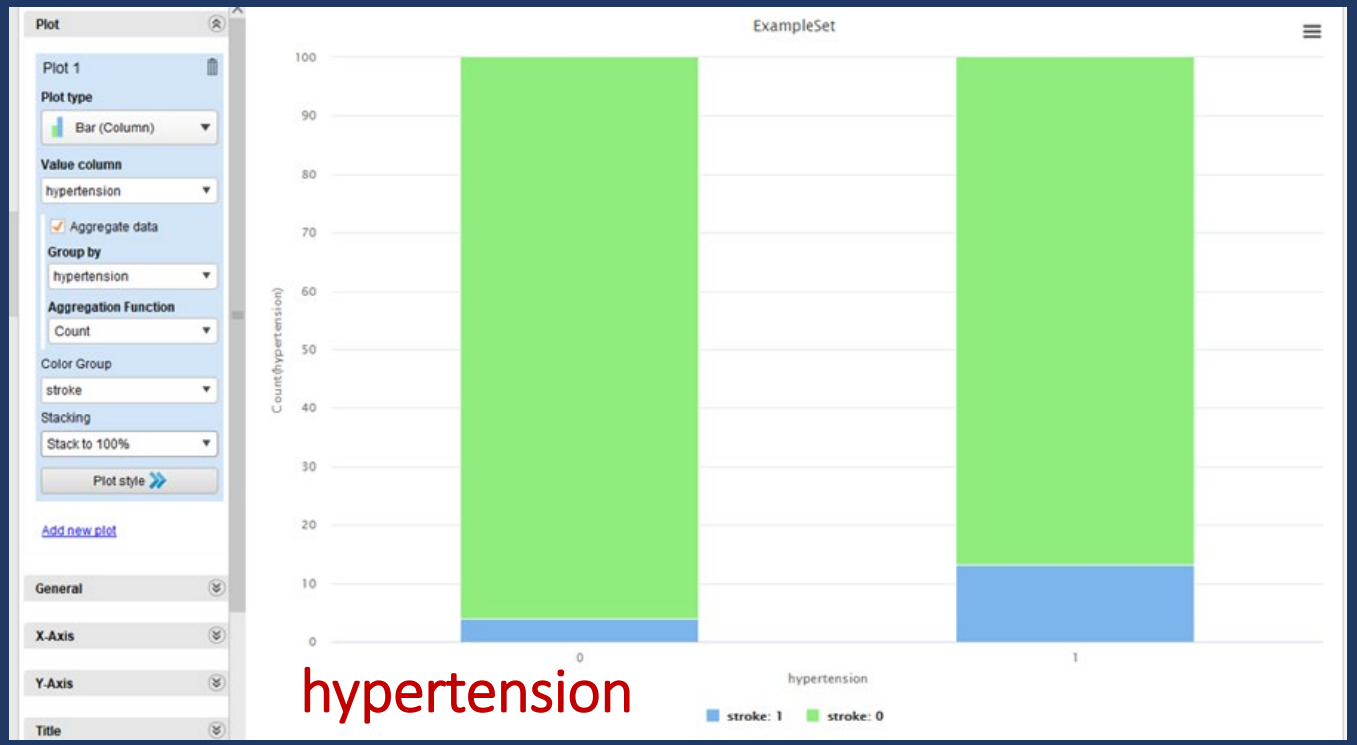
The dataset includes patients from babies to 84 years old. All of the ten bins contains a good number of patients. No extremes are present in any of the bins.



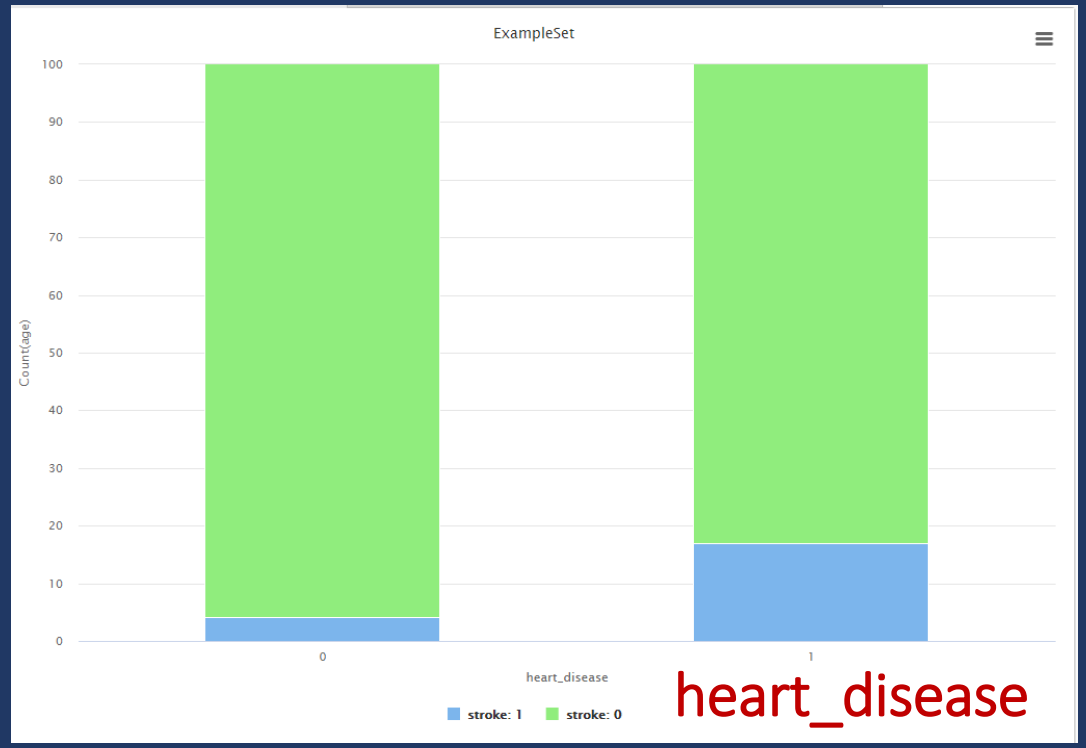
More of the patients with hypertension and/or heart_disease than healthy patients had a stroke

Graphs

- *All columns are normalized to 100.*
- *The columns at the right side of either graph represent patients having the disease.*
- *The blue sections represent the number of patients (normalized) having had a stroke in that group.*

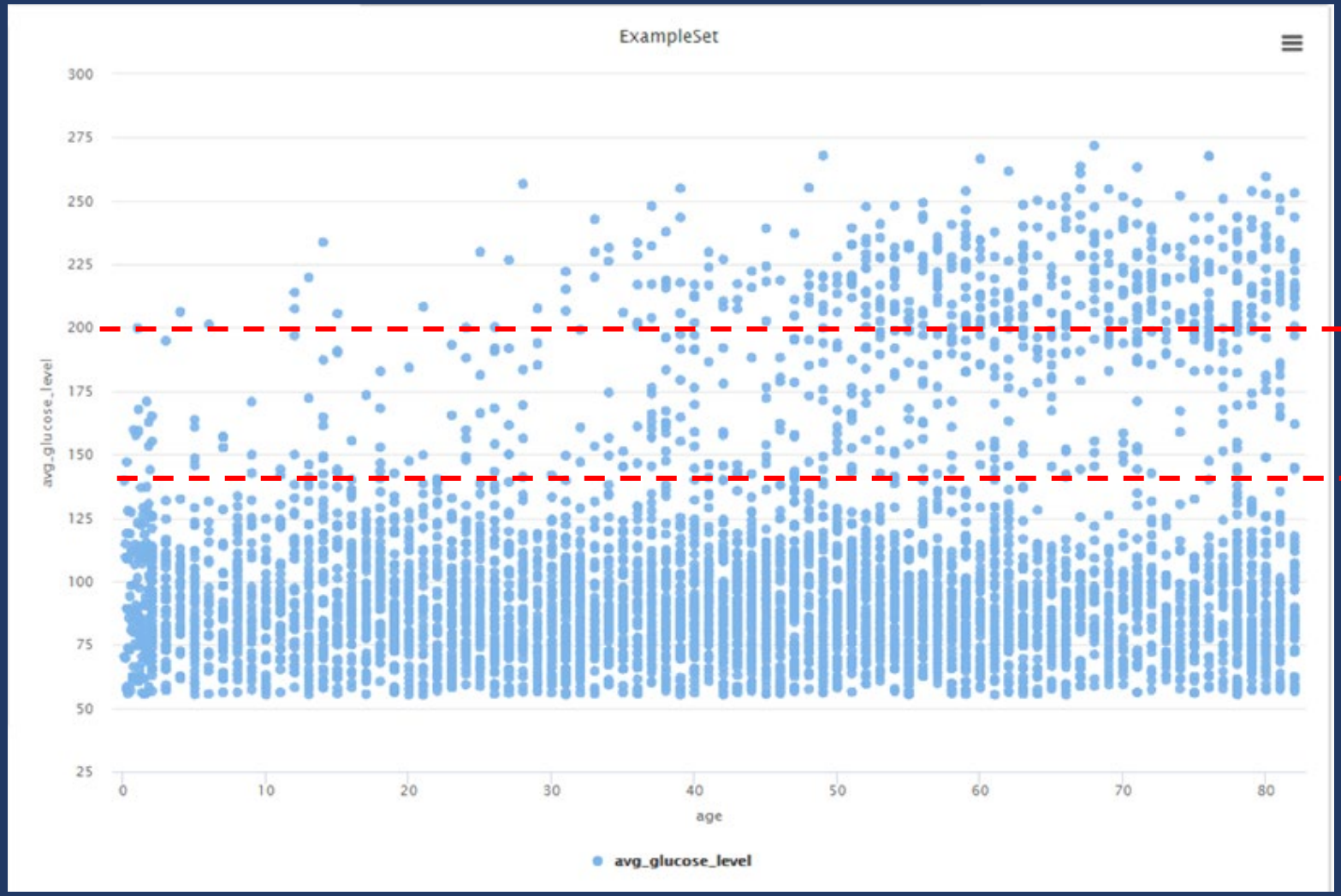


hypertension



heart_disease

avg_glucose_level vs. age



More 50 years old or older patients than younger patients have an average glucose level either above 200 (obese) or between 140 and 200 (overweight).

Data Wrangling

Data Wrangling

- Exclude data unsuitable for model development.
- Handle missing values.
- Discretize continuous values in attributes before they are used to create models.
- Create a new attribute to replace a misleading attribute.
- Convert binary attributes misinterpreted by RM as integer attributes during data import to their intended attribute type.
- Eliminate (1) extra classes generated during the process of converting numerical 0, 1 to binary type. For example, classes in “no” or “false” categories of the attribute (2) the class with the smallest proportion of data of the attribute (3) an underrepresented class of an attribute, if there is any.
- Create subsets from the initial dataset containing 5110 Examples with 249 of them being stroke Examples. Using built-in sampling algorithms to generate more balanced subsets in which the ratio of stroke to no_stroke is close to 1:1.

Exclusions

Children

age: 0.08 – 19 (966 Examples out of 5110 Examples)

(1) Causes for stroke are different for children and adults.

(2) Interpretation of children's bmi involves growth factor.

Extremely underrepresented classes

gender: "Other" (1 Example)

work_type: "Never_worked" (22 Examples)

Non-biased replacement value unavailable

smoking_status: "Unknown" (1544 Examples)

(RM operator: Filter Examples)

Missing Values

"Unknown" of the smoking_status attribute

Justifications: (1) It could mean negligence, privacy, or forgetfulness
(2) A replacement value without bias is unavailable.

Decision: Exclude them from the analysis

"N/A" of the bmi attribute

Justification:
NIH Statistics

	All (Men and Women)	Men	Women
Overweight or Obesity	70.2	73.7	66.9
Overweight	32.5	38.7	26.5
Obesity (including extreme obesity)	37.7	35	40.4
Extreme obesity	7.7	5.5	9.9

Decision: Replace the "N/A" with the average bmi of the data in the dataset.

Discretization (age)

- Nearly three-quarters of all strokes occurs over age of **65.**
- The risk of having a stroke more than **doubles each decade** after the age of **55**.

Five Age Groups

20-44
45-54
55-64
65-74
75-84

(Individual Age Group Study)

(RM Operator: Discretize by User Specification)

class names	upper limit
20-44	44.0
45-54	54.0
55-64	64.0
65-74	74.0
75-84	84.0

Discretization (bmi)

Justification:
NIH definition

BMI of Adults Ages 20 and Older	
BMI	**Classification**
18.5 to 24.9	Normal weight
25 to 29.9	Overweight
30+	Obesity (including extreme obesity)
40+	Extreme obesity

Discretization:
Four classes for bmi

class names	upper limit
Underweight	18.4
Normal Weight	24.9
Overweight	29.9
Obesity	92.0

(RM Operator: Discretize by User Specification)

Discretization (avg_glucose_level)

Justification:
American Diabetes Association

Result	Oral Glucose Tolerance Test (OGTT)
Normal	less than 140 mg/dl
Prediabetes	140 mg/dl to 199 mg/dl
Diabetes	200 mg/dl or higher

Discretization:
Three classes for avg_glucose_level)

class names	upper limit
Normal	139.99
Prediabetes	199.99
Diabetes	279.99

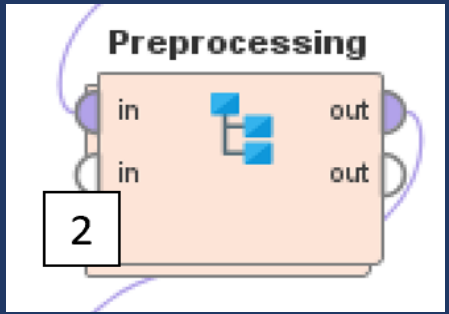
(RM Operator: Discretize by User Specification)

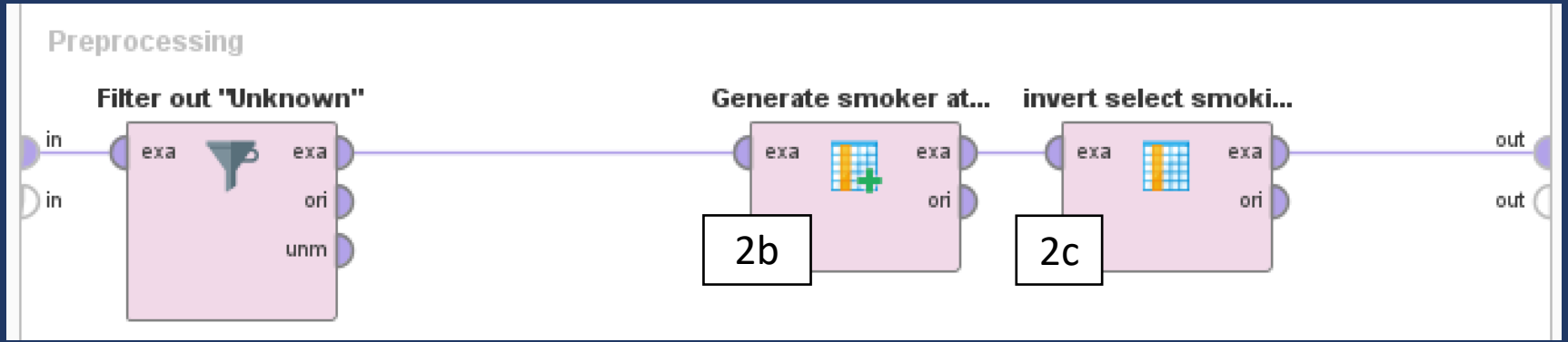
New Attribute “smoker”

smoking_status: smokes formerly smoked never smoked

smoker: 1 0

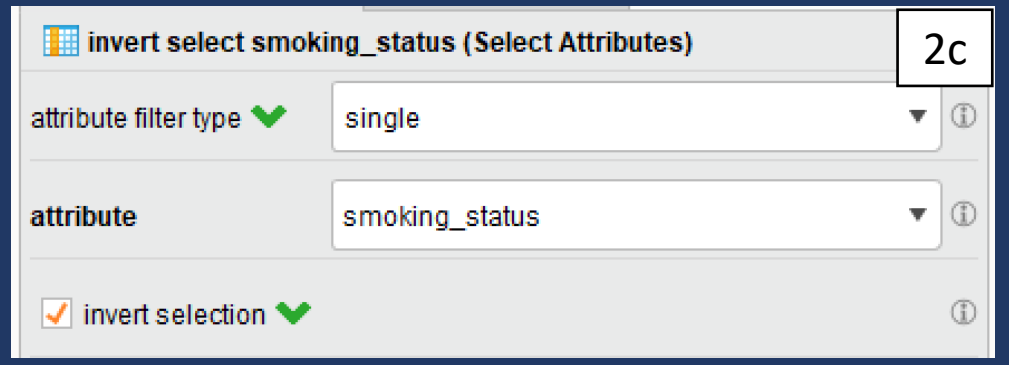
(A binary variable)





2b

attribute name	function expressions
smoker	if(smoking_status=="smokes", 1, 0)

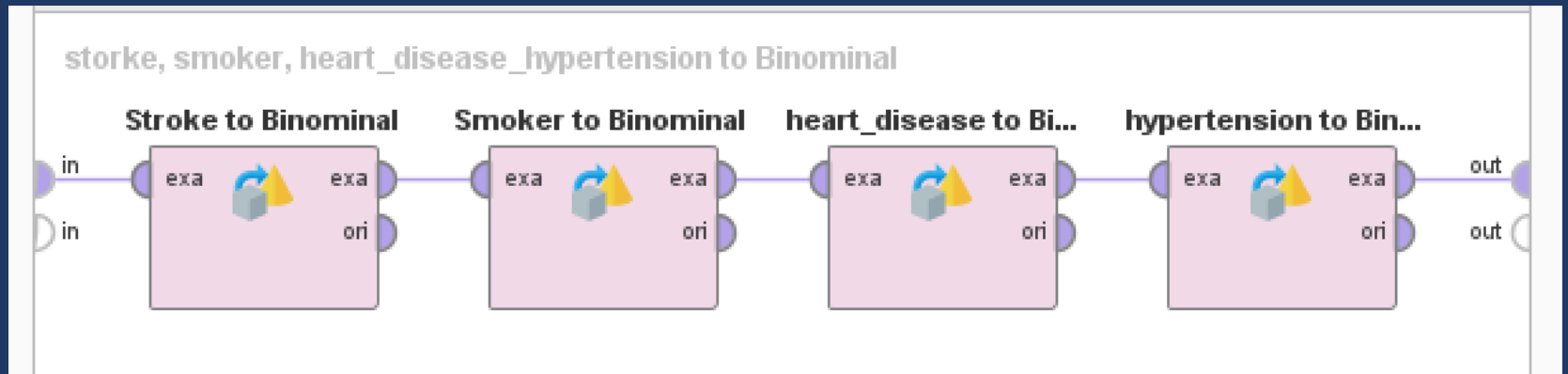


Conversion of Data Types

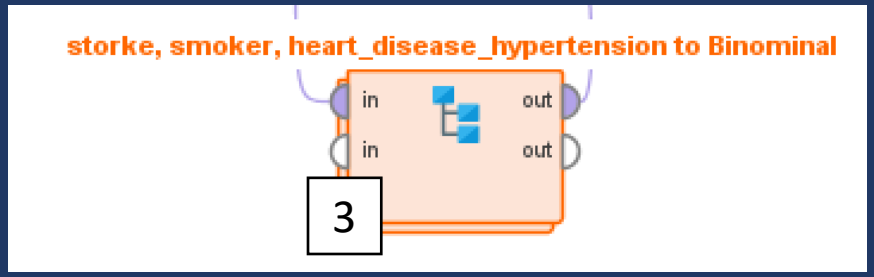
(Convert Binary Variables Misinterpreted as Numerical Variables by RM to Boolean Variables)

stroke (label):	0, 1	**stroke (label):**	false, true
smoker:	0, 1	**smoker:**	false, true
hypertension:	0, 1	**hypertension:**	false, true
heart_disease:	0, 1	**heart_disease:**	false, true

RM Operator: Numerical to Binominal

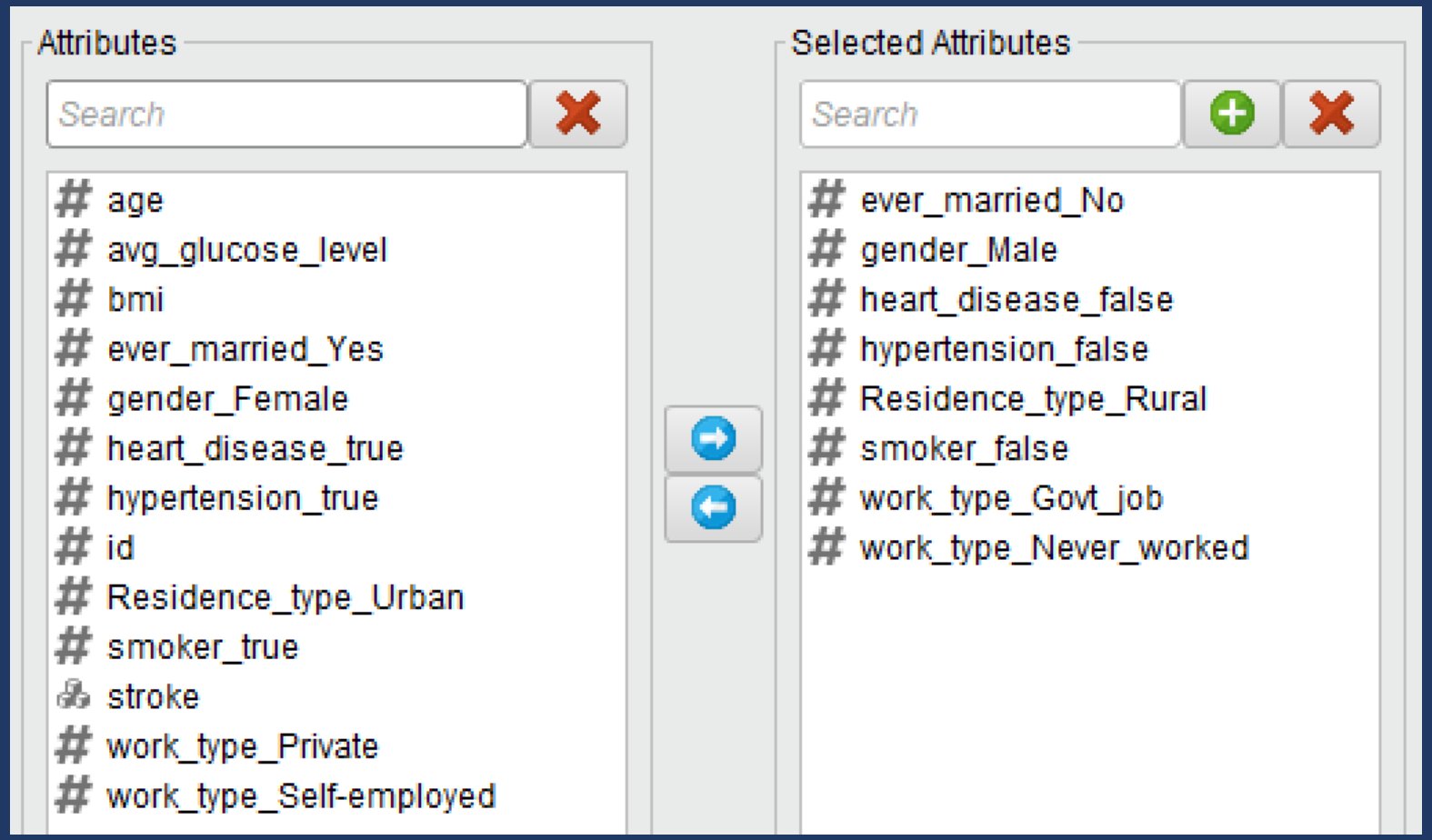


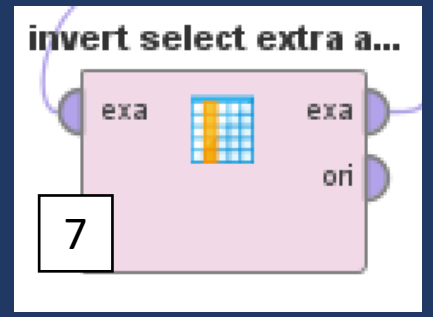
RM Operator: Subprocess



Attribute Reduction

RM Operator: Select "Attributes", then click on "invert select" to remove them!!!





In the original attribute, the attribute is:

1. The smallest proportion
2. "false" or "No"
3. Underrepresented class

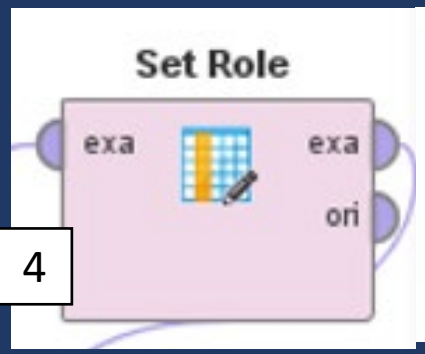
"Inverted select" means selected attributes are removed.

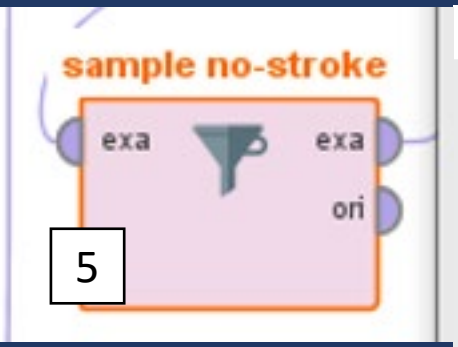
Sampling Method

Sampling method: step 5 in the flowchart of the model

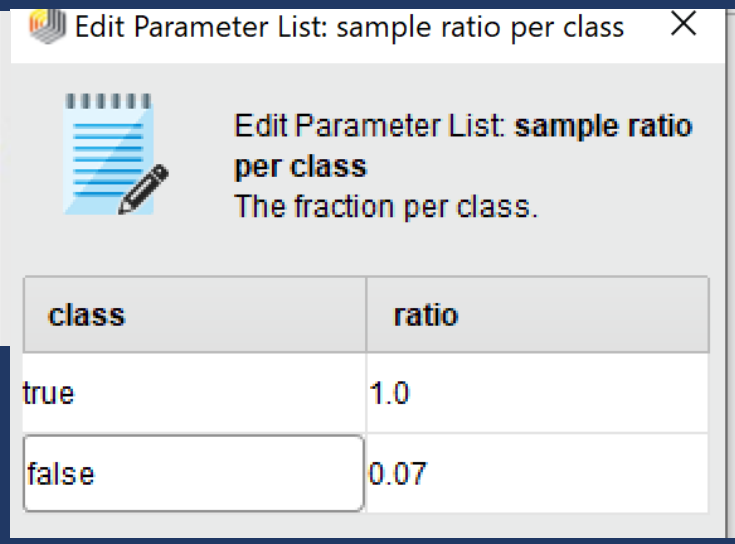
Starting subset: 202 (stroke: true)/3330 (stroke: false)
Resulting subsets: 202 (stroke: true)/218 (stroke: false) 420 dataset
202 (stroke:true)/ 281 (stroke:false) 483 dataset

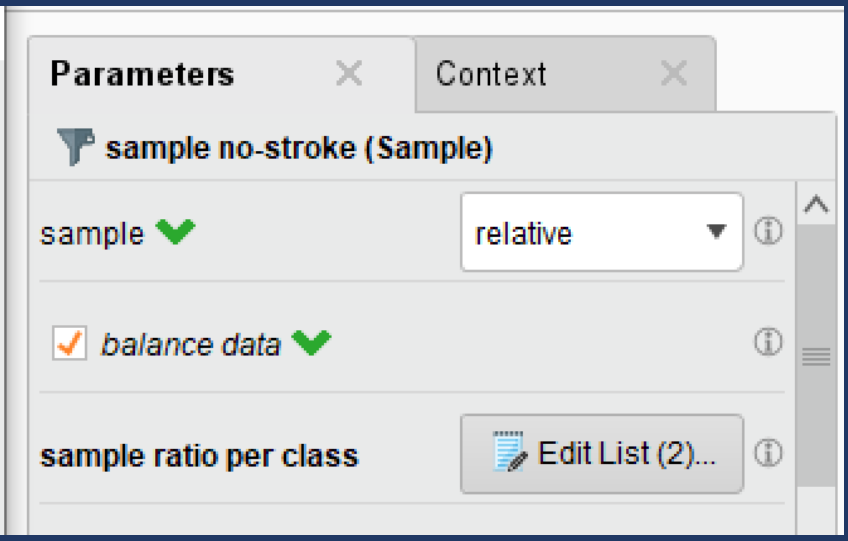
RM Op.: Sample (class: false, ratio varies)





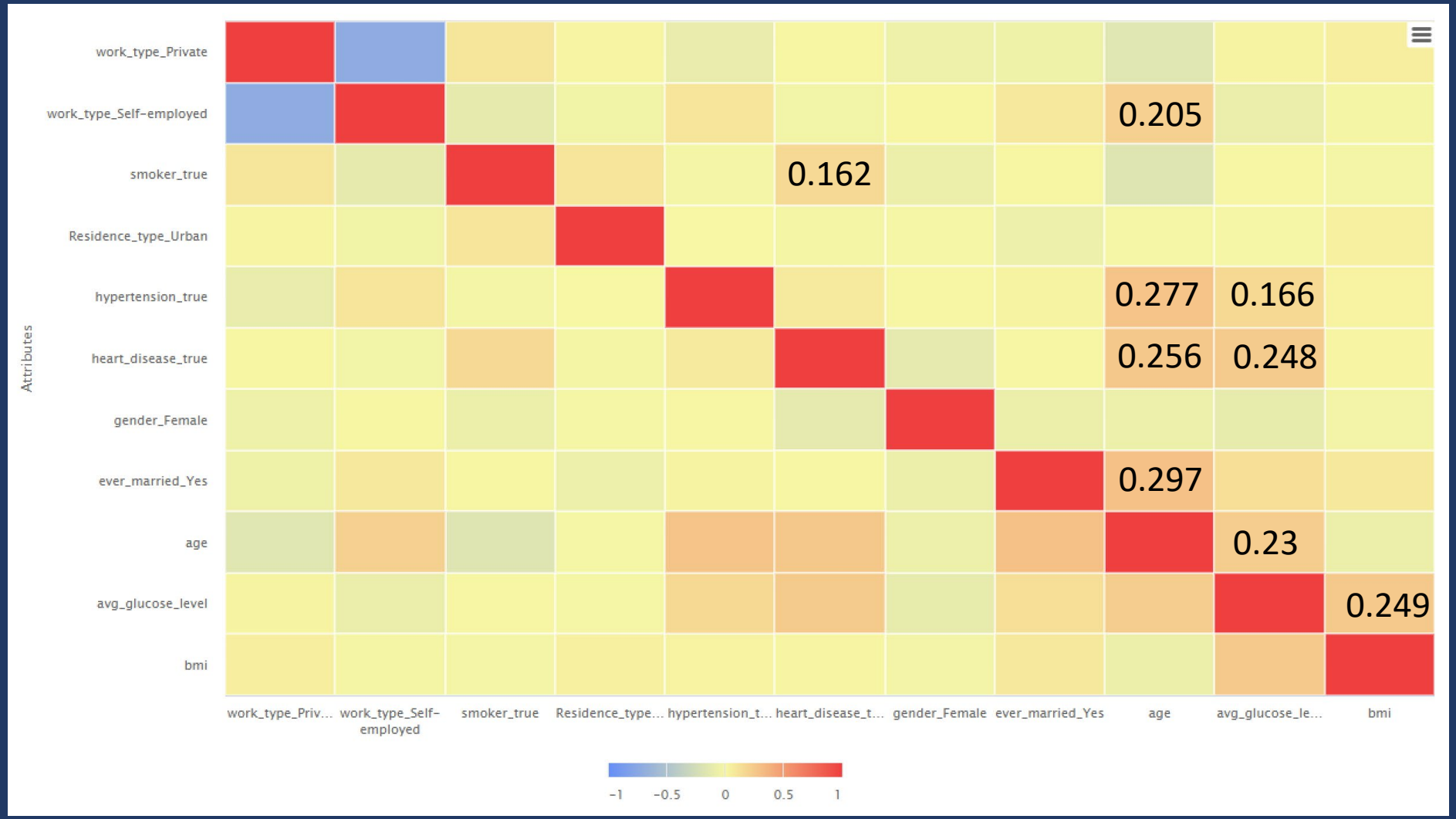
RM Op.: Set Role
(id: id, stroke: label)





Correlation Analysis

Correlation Matrix



The greatest correlation coefficient is 0.297. Thus, the attributes are slightly correlated at most. This is important to consider because when building models like Naïve Bayes model it assumes that the predictors (attributes) are independent from one another.

correlation_no Unknown_smoker_sample0.07.rmp

Model Development

- The process of building a model is shown as a flowchart of RM operators and stacked operators arranged in an appropriate order.
- The RM file containing the process of the model is denoted with the extension of .rmp in the filename.

Types of Classification Models

Logistic Regression

- It assigns probabilities to discrete outcomes (stroke or no_stroke) using the Sigmoid function, which converts numerical results into an expression of probability between 0 and 1.0.
- Attributes that significantly increase the chance of having a stroke can be identified by p-values.

Decision Tree

- It breaks down a dataset into smaller and smaller subsets while at the same time, an associated decision tree is incrementally developed. The tree structure has decision nodes and leaf nodes. The topmost decision node (root node) is the best predictor. A node is an attribute in the dataset.

Naïve Bayes

- It calculates the probability of stroke or no_stroke with a given set of values of the predictors (attributes in the dataset).
- It assumes that any given pair of predictors (attributes) are independent of each other.

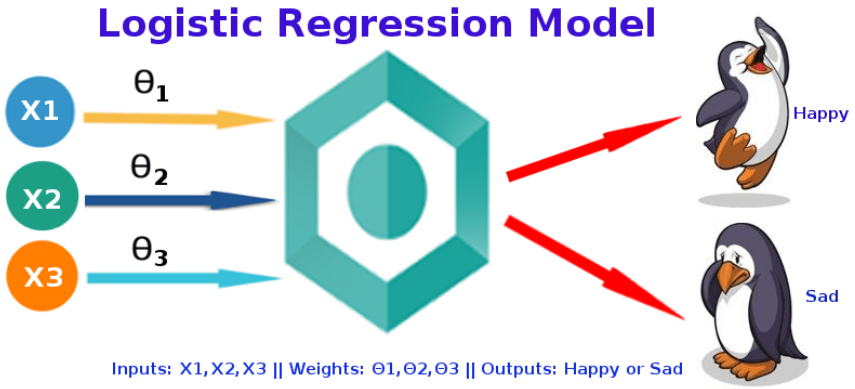
k-nearest neighbors (*k*-NN)

- It stores all available cases first, and then classifies the target to the class of the case to which the target is "most similar". If using a distance function as the similarity measure like it was done in this project, "most similar" means the distance between the case and the target is the shortest.

Neural Network

- It uses the attributes as the input nodes to build a network of artificial neurons (also known as "nodes") that are connected to each other and assigns a value (weight) to indicate the strength of their connections to one another. The greater the weight of an attribute, the higher degree of influence that attribute has on the chance of having a stroke.

Logistic Regression Model

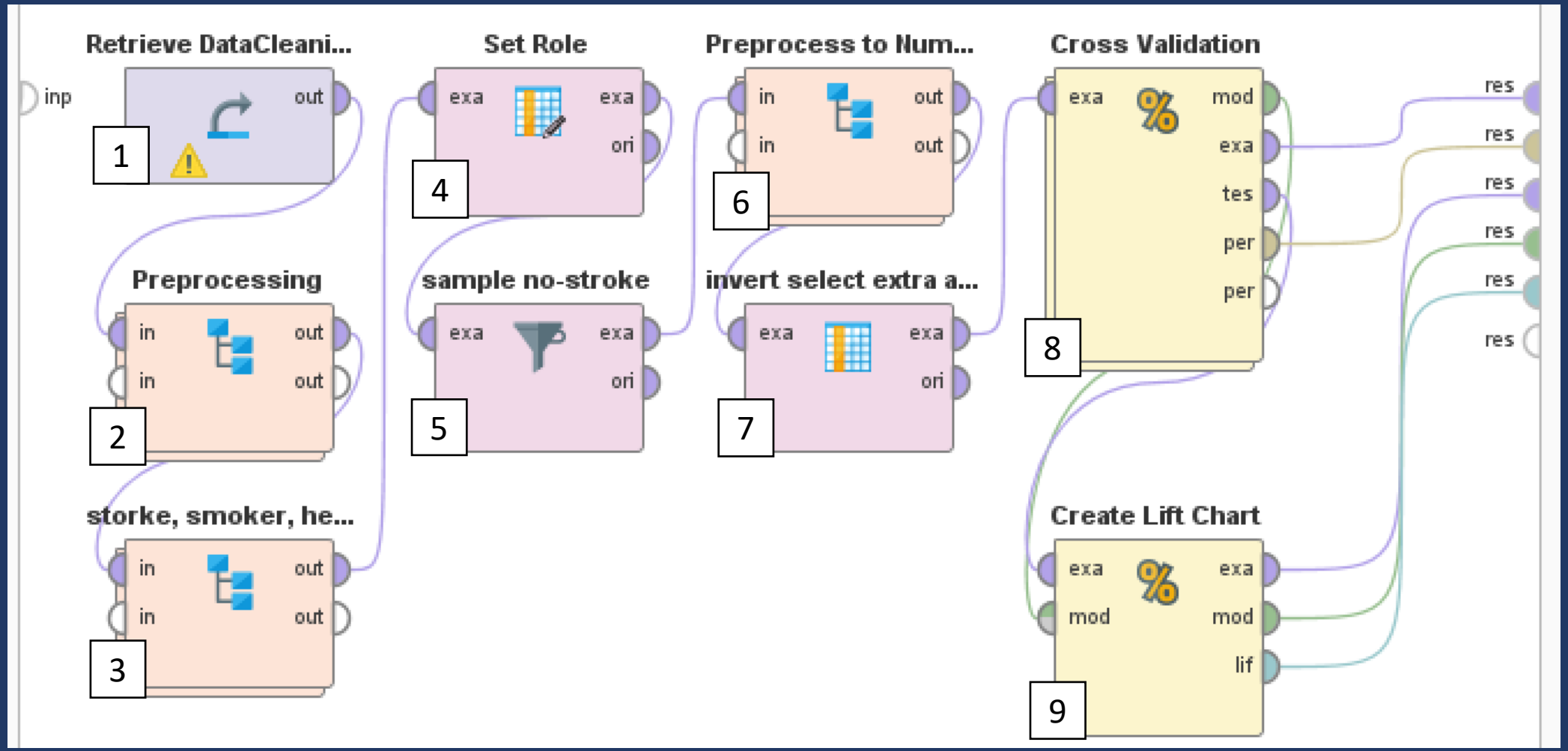


Logistic Regression Model predicts the probability of an outcome that can only have two values (i. e. a dichotomy). The prediction is based on the use of one or several predictors/attributes (numerical and categorical). Attributes significantly increasing the chance of stroke can be identified by the p-value associated with the attributes.

Significant Predictor: A predictor that has a p-value < = 0.05

K-Folds Cross Validation for Model Testing: (i) Split the entire data randomly into K folds. (ii) Fit the model using the K-1 (K minus 1) folds and validate the model using the remaining Kth fold. (iii) Repeat this process until every K-fold serves as the test set. Then take the average of the recorded scores, which is the performance metric of the model that are included in the confusion matrix.

Logistic Regression

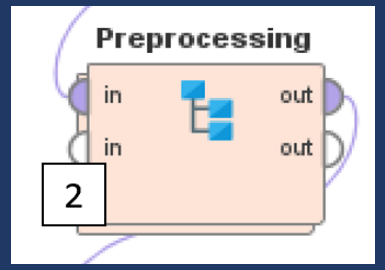


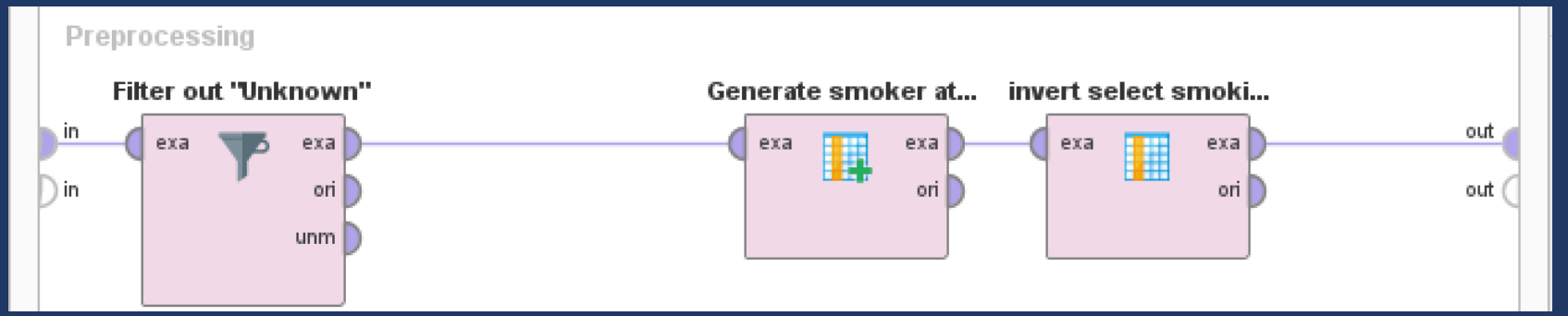
(Numerical Attributes as Input)

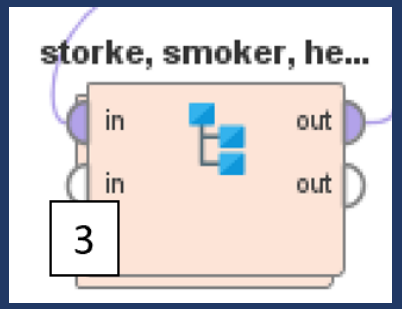
Logistic Regression_no Unknown_smoker_sample0.07.rmp

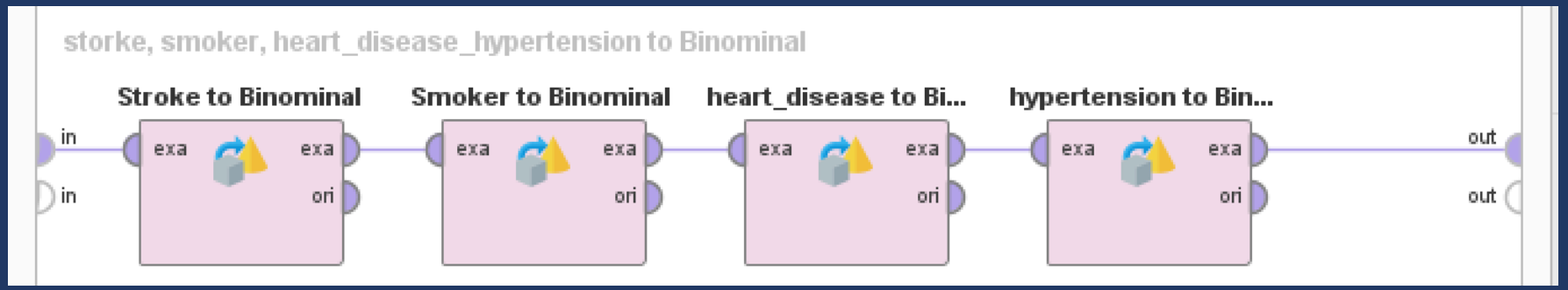
Logistic Regression

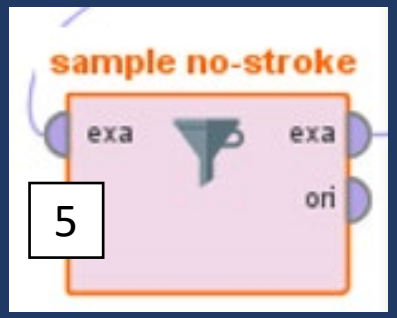
(Details of stacked operators and sampling operators which were necessary for building all types of models)

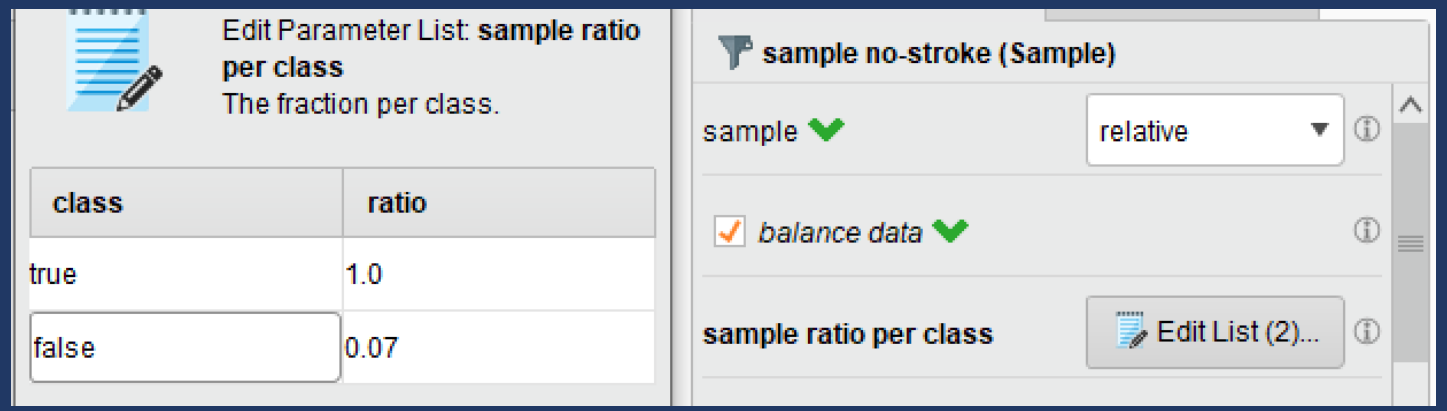








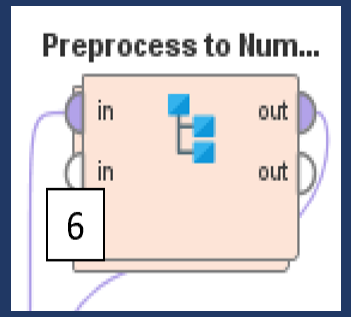




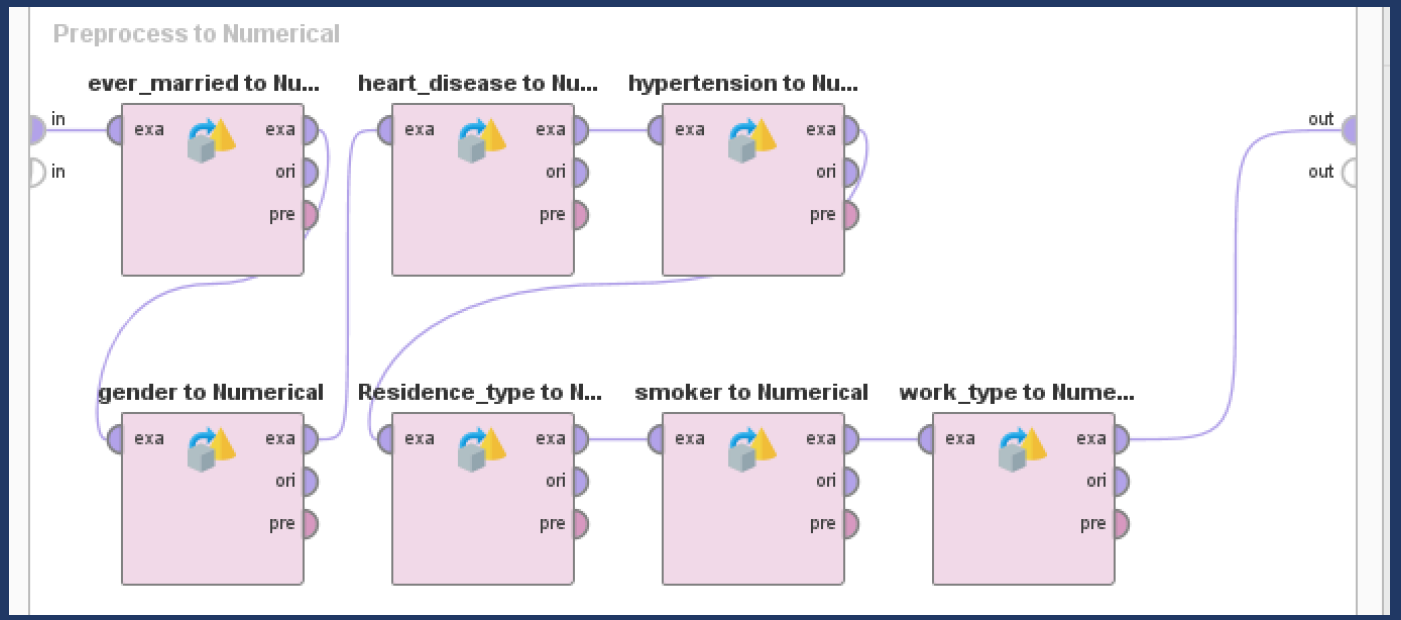
Logistic Regression_no Unknown_smoker_sample0.07.rmp

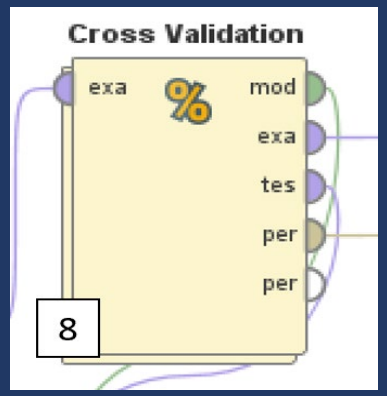
Logistic Regression

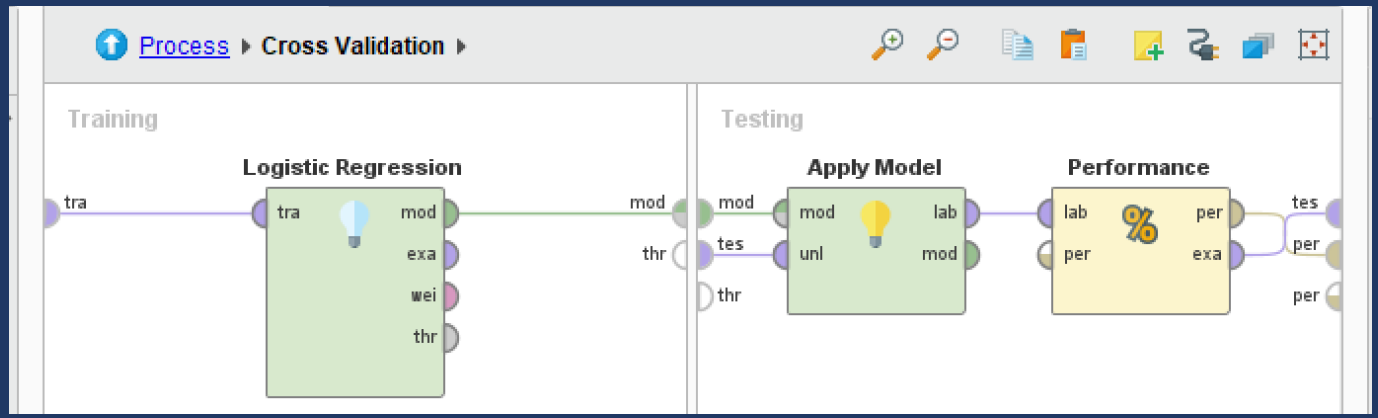
(details of stacked operators that were necessary for all types of models, cont'd)



Operator 6 was used in all other types of models as well.







Logistic Regression_no Unknown_smoker_sample0.07.rmp

Logistic Regression

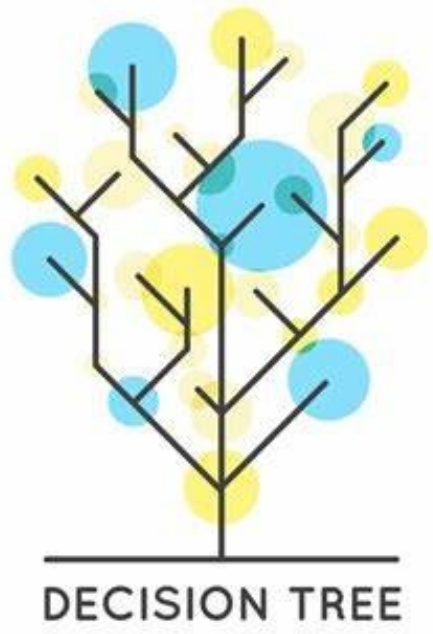
Three significant predictors were identified: age, avg_glucose_level, hypertension (p-value < = 0.05).

Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
age	0.076	1.292	0.010	7.737	0.000
Intercept	-6.147	-0.159	1.025	-5.997	0.000
avg_glucose_level	0.008	0.467	0.002	3.447	0.001
hypertension_true	0.776	0.308	0.322	2.409	0.016
work_type_Self-employed	0.385	0.162	0.419	0.919	0.358
gender_Female	0.225	0.111	0.248	0.907	0.364
smoker_true	0.247	0.105	0.296	0.833	0.405
ever_married_Yes	-0.296	-0.104	0.415	-0.713	0.476
work_type_Private	0.217	0.105	0.365	0.595	0.552
bmi	0.010	0.060	0.021	0.462	0.644
Residence_type_Urban	-0.107	-0.053	0.242	-0.442	0.659
heart_disease_true	-0.028	-0.010	0.371	-0.076	0.939

Subset: Stroke (true/false : 202/218, 1/1.08)

Logistic Regression_no Unknown_smoker_sample0.07.rmp

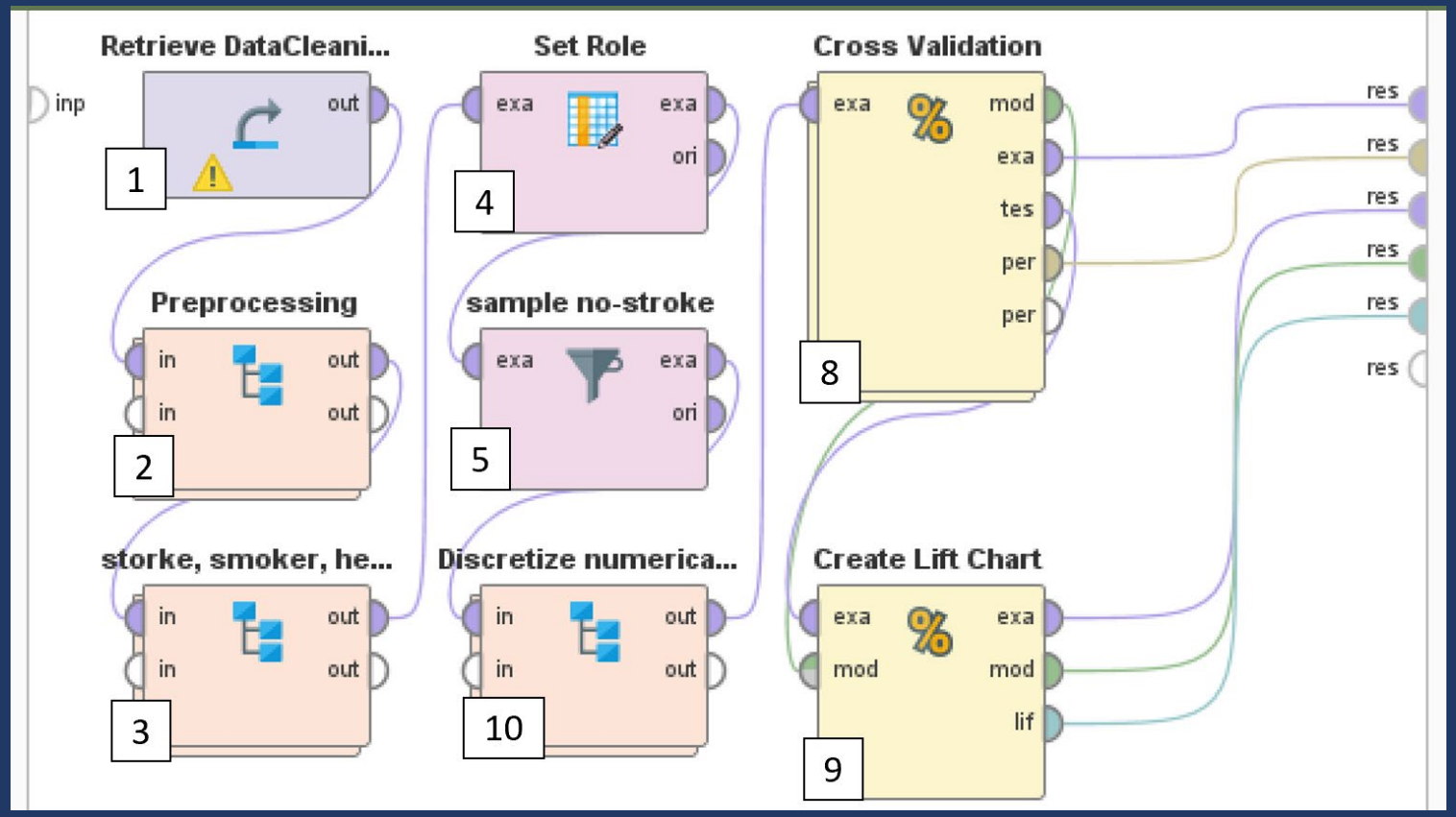
Decision Tree Model

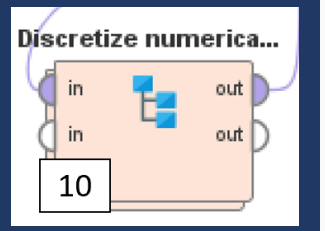


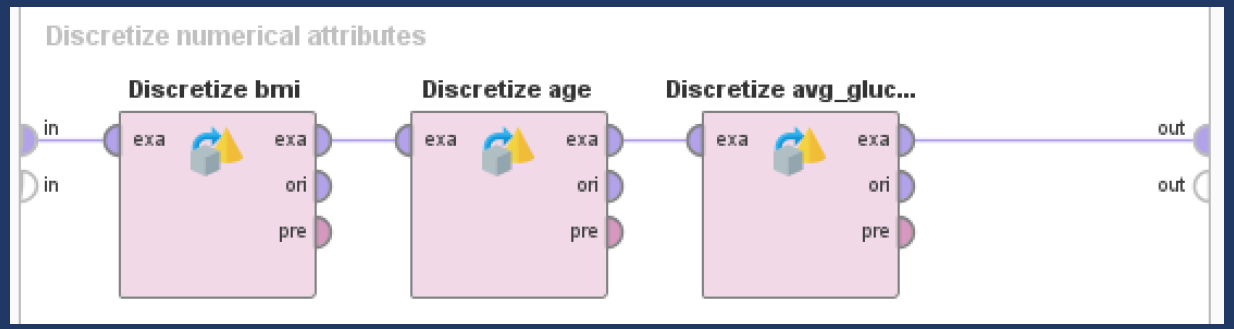
Decision Tree Model builds classification (this project) or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The result is a tree with decision nodes and leaf nodes. The nodes are predictors (the attributes in the dataset). The topmost decision node in a tree (root node) corresponds to the best predictor.

Decision Tree

Input:
Nominal Attributes



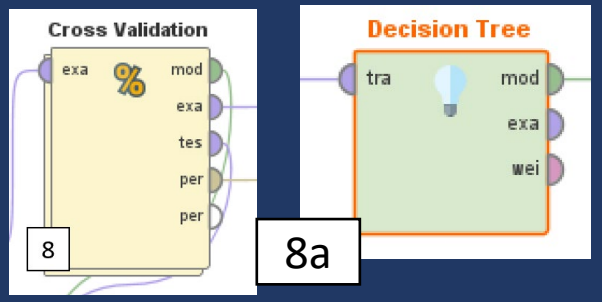




DecisionTree_no Unknown_smoker_sample0.07.rmp

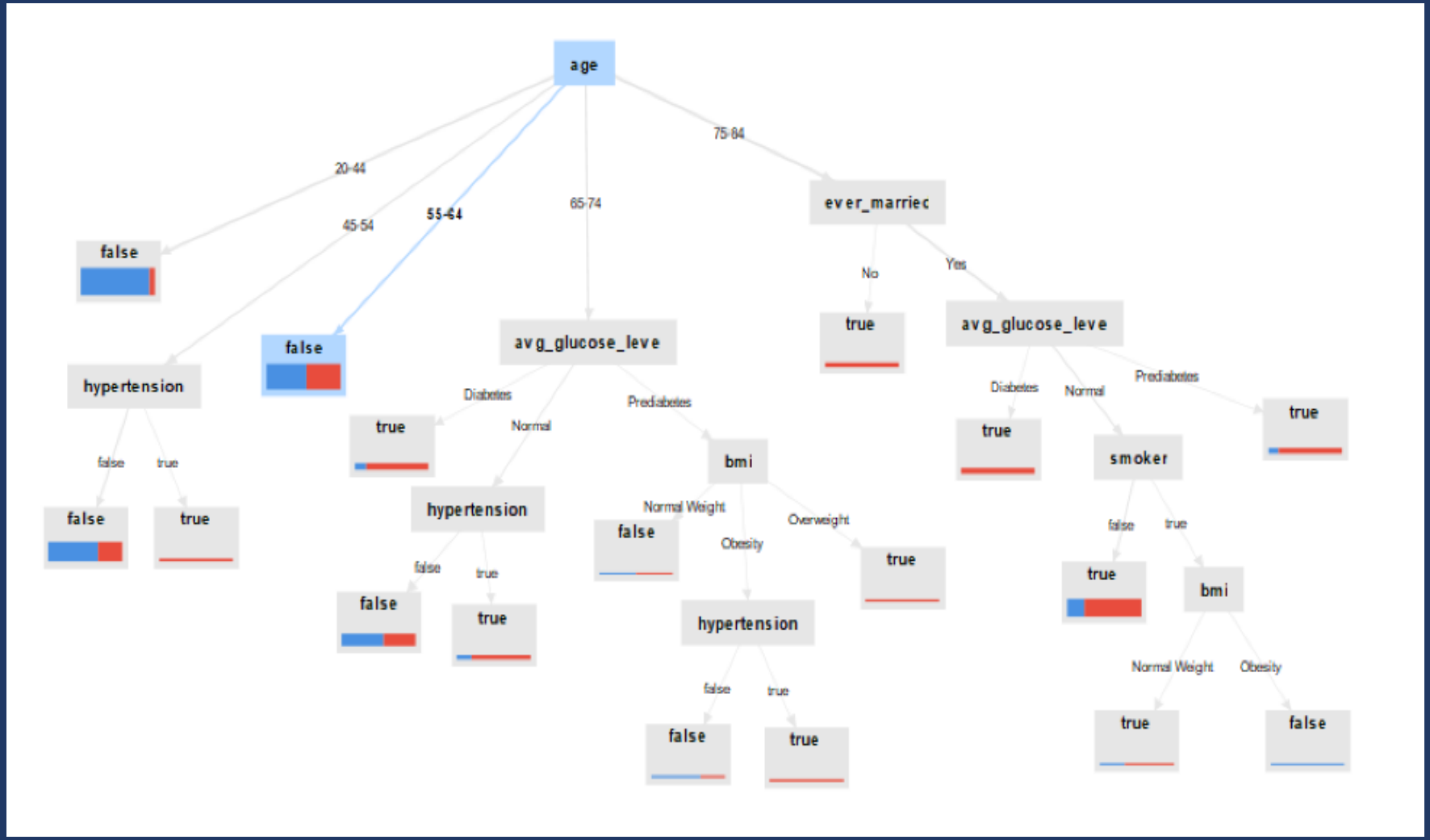
Decision Tree

The three significant predictors identified in the logistic regression model: age, avg_glucose_level, hypertension were found in one of the 65-74 branches as the decision nodes.



Decision Tree

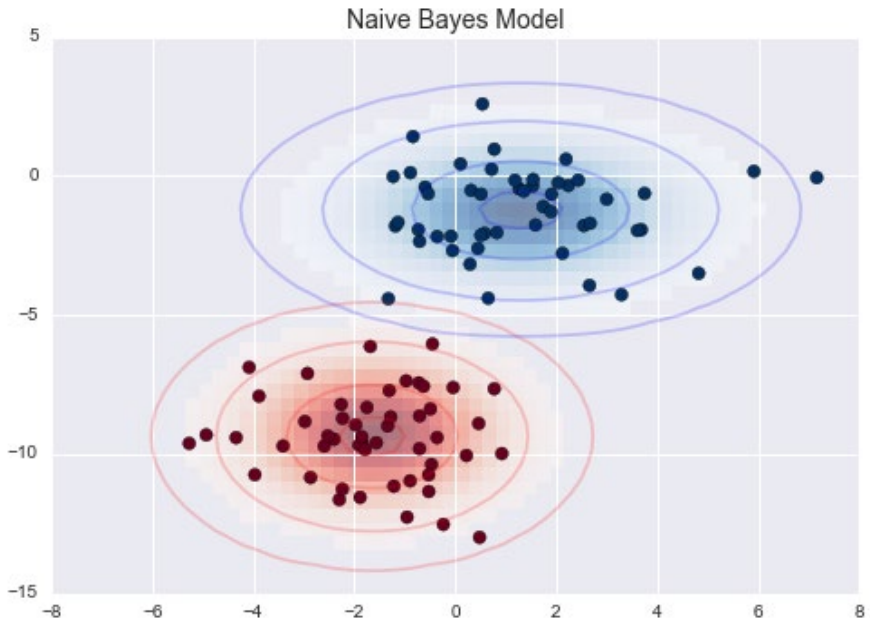
criterion	gain_ratio
maximal depth	20
apply pruning	☑
confidence	0.25
apply prepruning	☑
minimal gain	0.05
minimal leaf size	2
minimal size for split	4
number of prepruning alternatives	3



Subset: Stroke (true/false : 202/218, 1/1.08)

DecisionTree_no Unknown_smoker_sample0.07.rmp

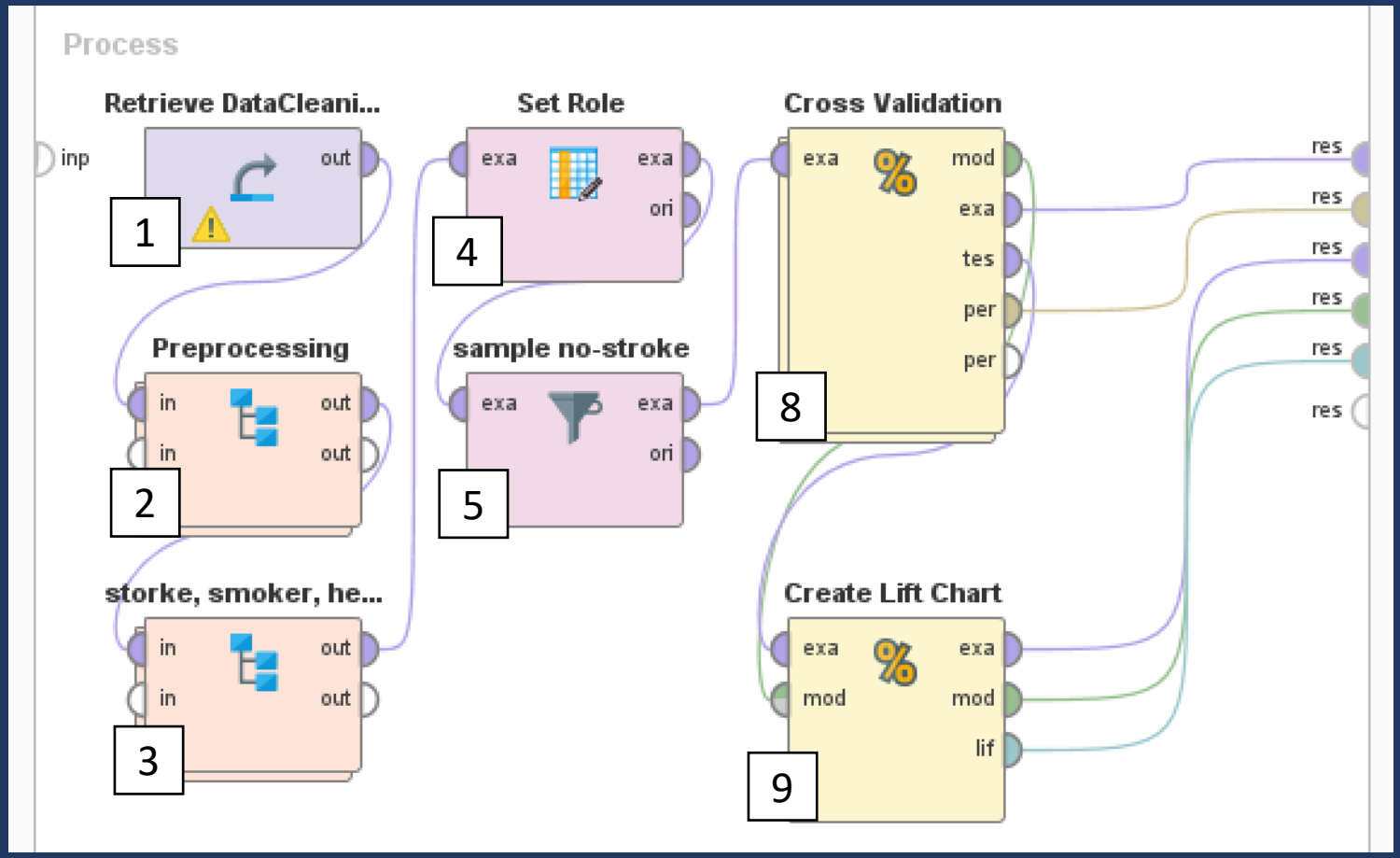
Naïve Bayes Model

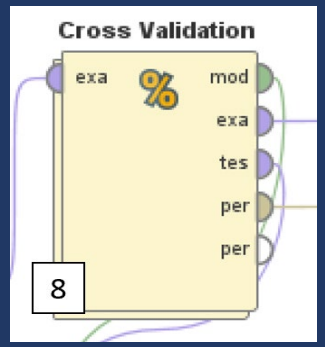


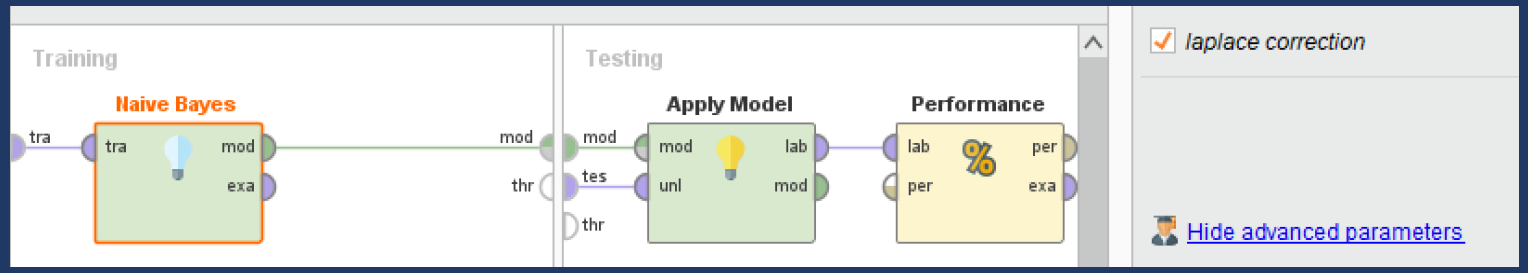
Naive Bayes Model calculates the probability of stroke or no_stroke with a given set of values of the predictors (attributes in the dataset). Naive Bayes Theorem assumes that any given pair of predictors/attributes are independent of each other.

Naïve Bayes

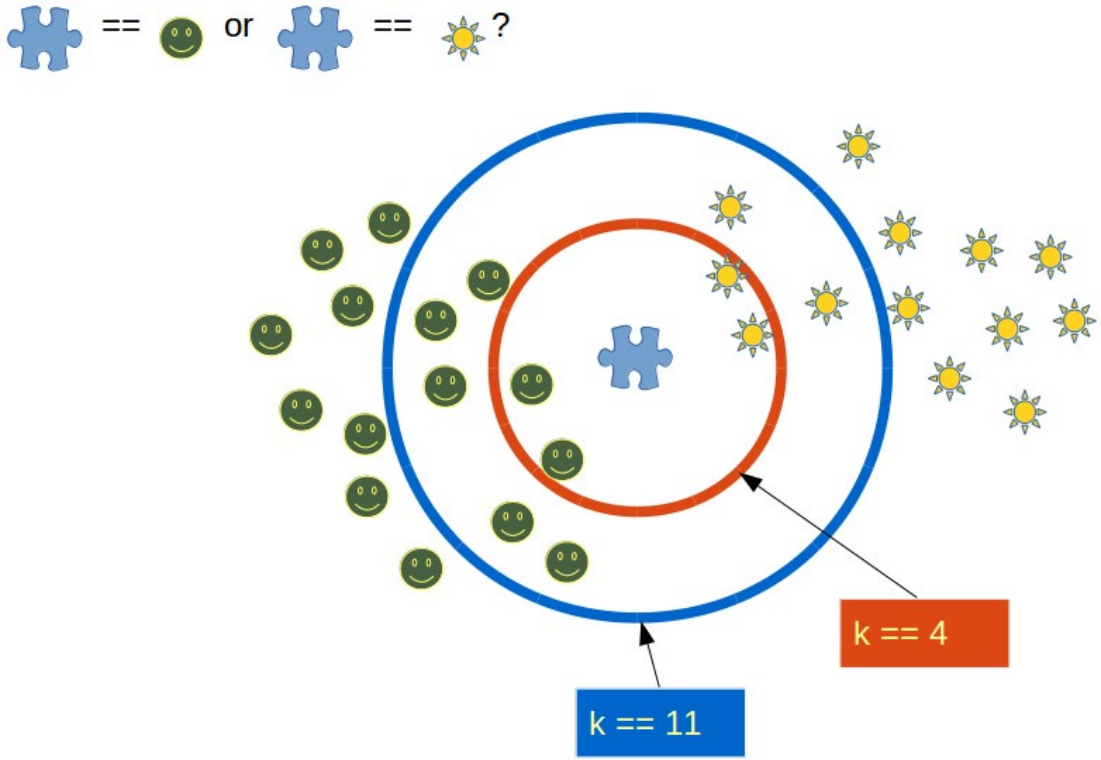
Input:
Mixed Attribute Types







k-nearest neighbors (*k*-NN) Model



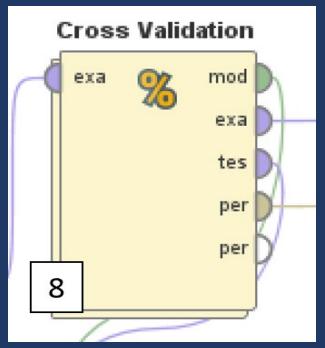
k-nearest neighbors stores all available cases and predict the numerical target based on a similarity measure. The target is classified to the class of the case to which the target is most similar. If the similarity measure is a distance function like the one used in this project, "most similar" means the distance between the case and the target is the shortest. k-NN is a lazy learner and robust to noisy data.

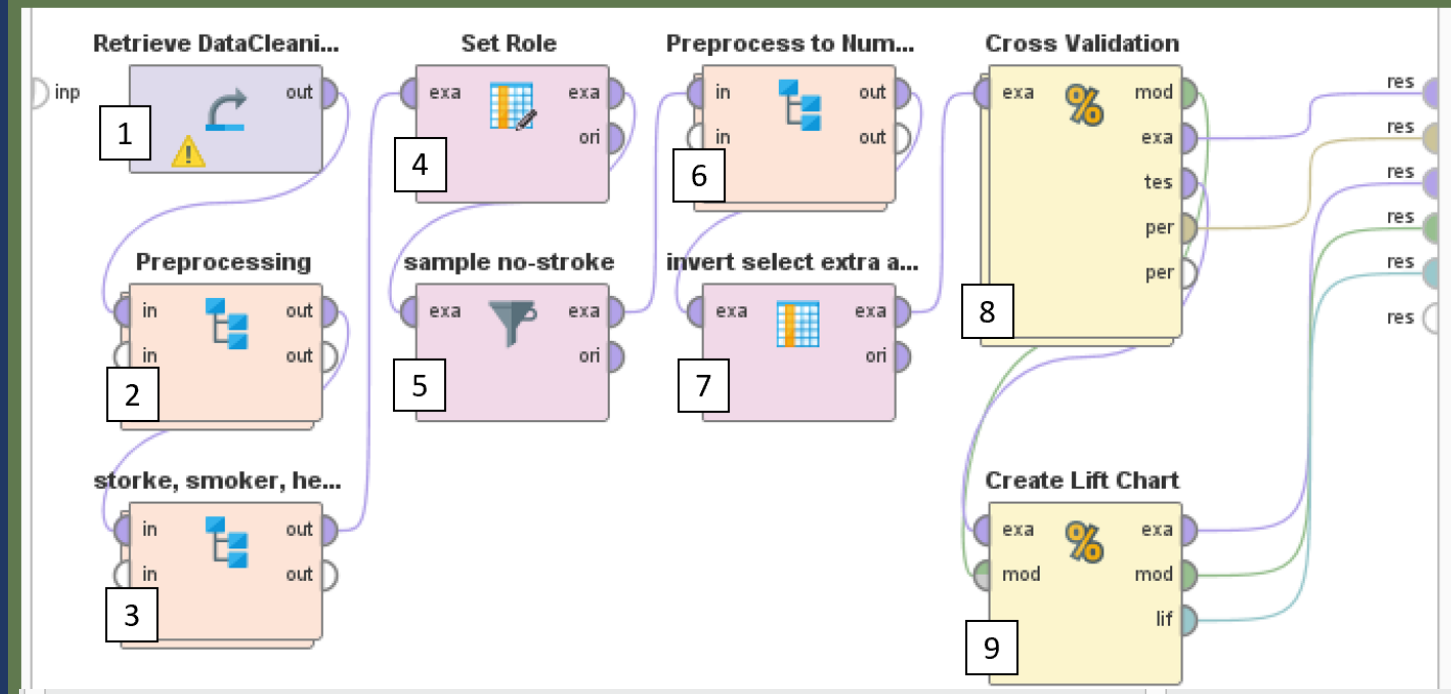
k-nearest neighbors (*k*-NN)

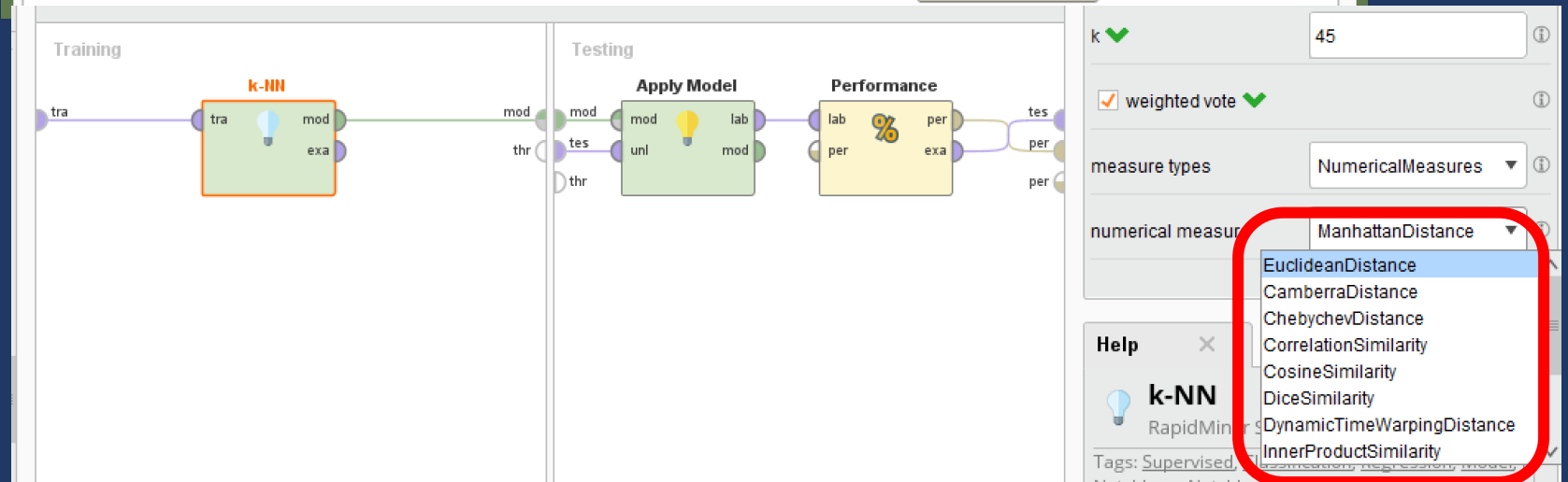
(Numerical Attribute Type as Input)

Subset: Stroke (true/false : 202/281, 1/1.39)

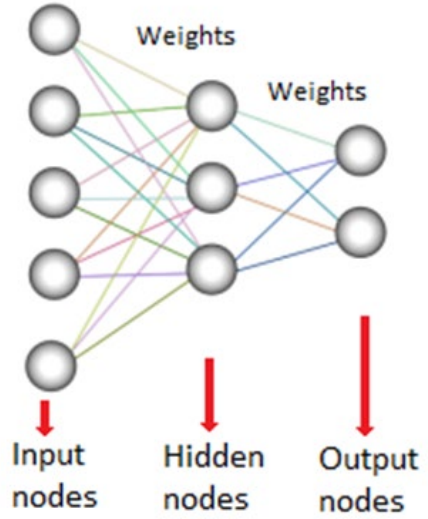
KNN_no Unknown_smoker_sample0.07_numerical input.rmp







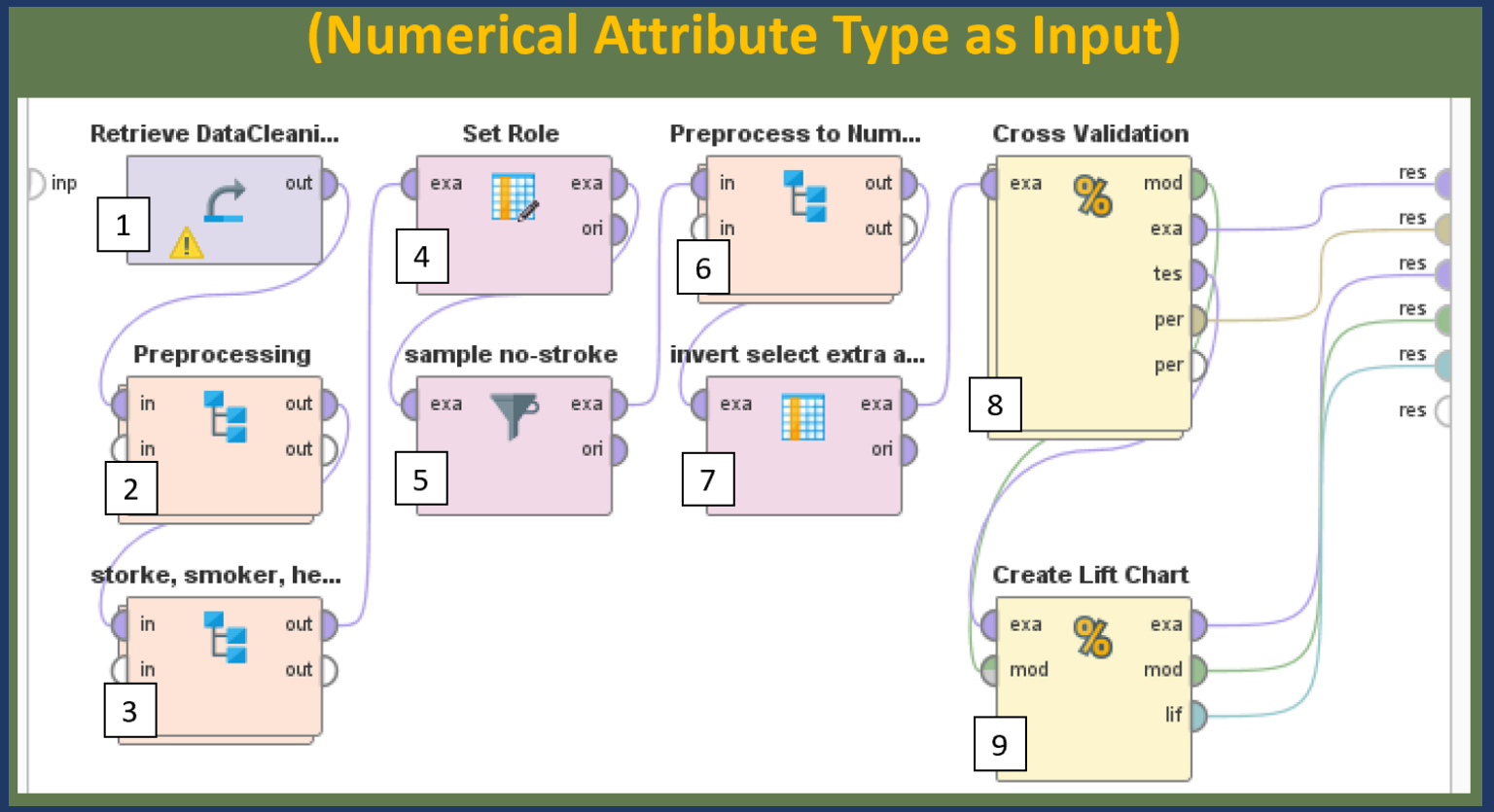
Neural Network Model

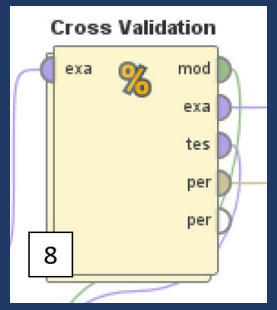


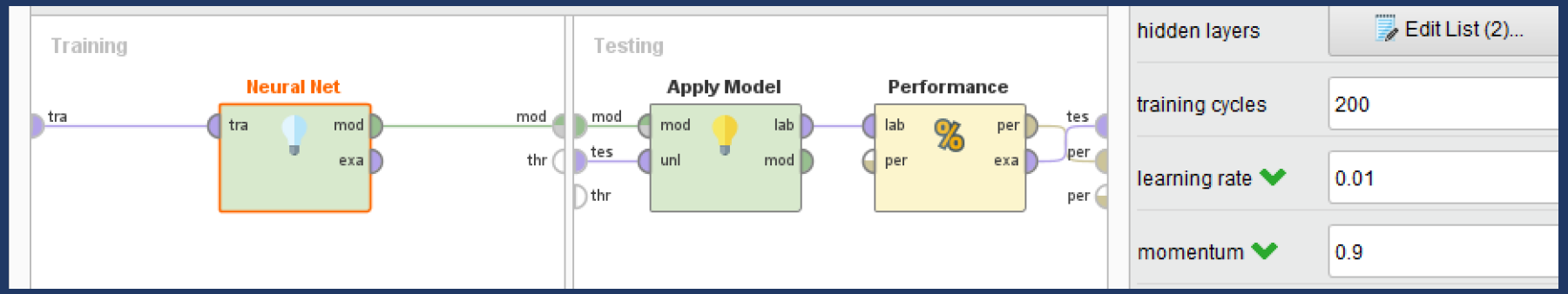
Neural Network Model is comprised of a network of artificial neurons (also known as "nodes"). These nodes are connected to each other, and the strength of their connections to one another is assigned a value based on their strength: inhibition (maximum being -1.0) or excitation (maximum being +1.0). If the value of the connection is high, then it indicates that there is a strong connection. Within each node's design, a transfer function is built in. There are three types of neurons in a neural network, **input nodes**, **hidden nodes**, and **output nodes**. In this project The attributes from the dataset were the input nodes. In RM, the strength of the connections between the nodes is represented by **the width of the line** connecting a pair of nodes.

Neural Network

(Numerical Attribute Type as Input)



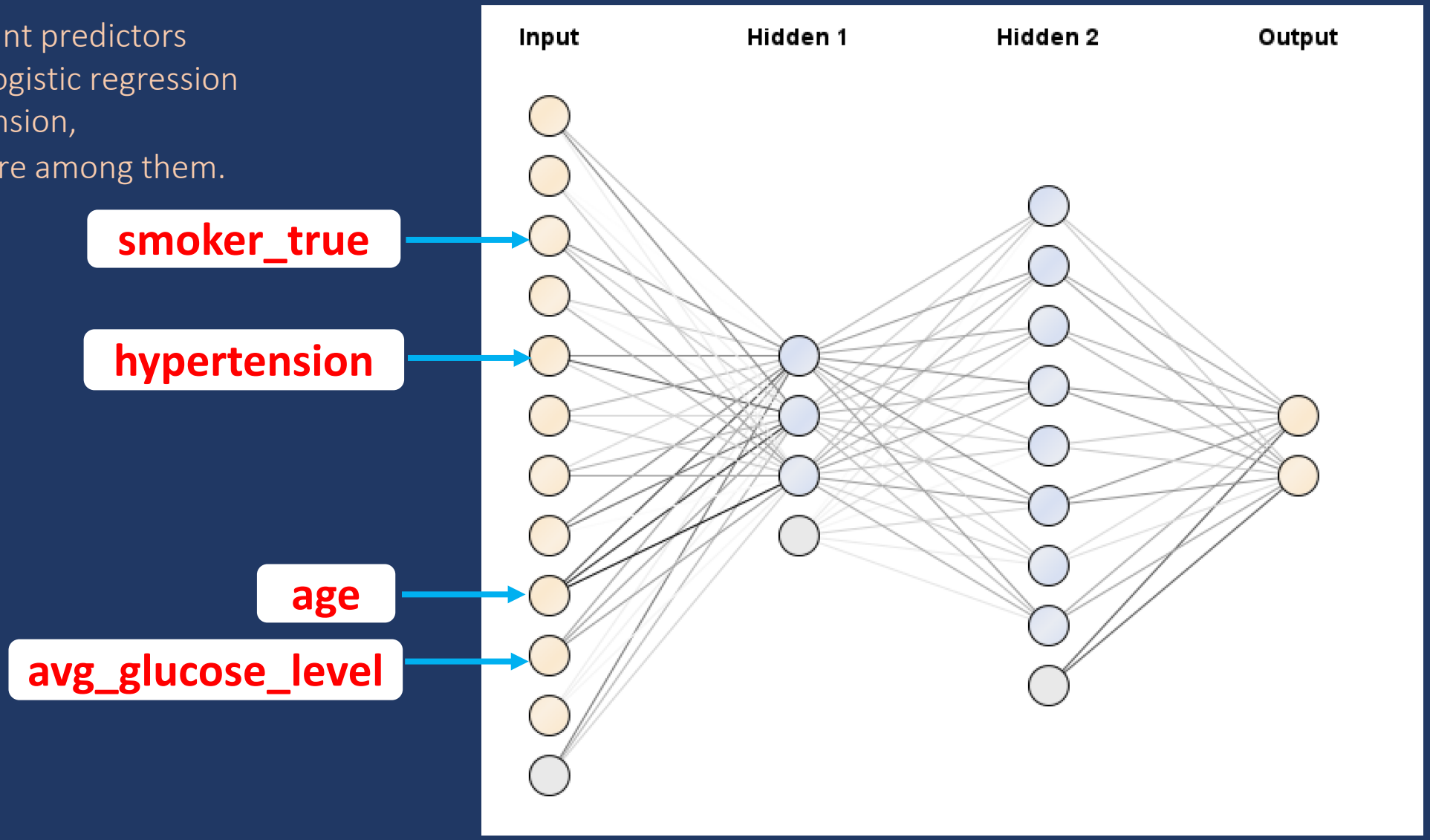




NN_no Unknown_smoker_sample0.07.rmp

Neural Network

1. Four significant nodes connected to hidden layer 1 by thicker lines were identified.
2. The three significant predictors determined via the logistic regression model (age, hypertension, avg_glucose_level) are among them.



Subset: Stroke (true/false : 202/218, 1/1.08)

NN_no Unknown_smoker_sample0.07.rmp

Model Evaluation

Confusion Matrix
ROC Curves

Terminology – Confusion Matrix

Classification Model reads some input and generates an output that classifies the input into some category.

K-folds Cross Validation is where a given data set is split into a K number of sections/folds where each fold is used as a testing set at some point. The model is tested k-times. The averaged results from the k-tests result are organized into a confusion matrix.

Confusion Matrix a table that is often used to describe the performance of a classification model, or "a classifier." A confusion matrix includes the following items.

TP (True Positive)

TN (True Negative)

FP (False Positive)

FN (False Negative)

Accuracy Overall, how often is the classifier correct?

Class Precision For the "True" class, when it predicts "True", how often is it correct?

Class Recall For the "True" class, when it's actually "True", how often does it predict "True"?

Terminology – ROC Curves

Cont'd

- **TPR (Sensitivity, or Recall), True Positive Rate** When it's actually yes, how often does it predict yes?

  TPR = TP/(TP+FN)

- **TNR (Specificity), True Negative Rate** When it's actually no, how often does it predict no?

  TNR = TN/(TN+FP)

- **FPR (False Positive Rate)** When it's actually no, how often does it predict yes?

  FPR = FP/(TN+FP)

- **ROC** (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: TPR vs. FPR. ROC is a helpful diagnostic tool for determining the trade-off between different thresholds.
- **Classification threshold** represents the decision-making boundary that need to be defined before deploying a model. In this project, values above this threshold will be mapped to the stroke category, while those below or at the threshold will be mapped to the no-stroke category.
- **Optimal Threshold** Square Root of (TPR * TNR)
- **ROC AUC** (Area under the ROC Curve) ROC AUC is a useful metric for comparing models based on their overall capabilities. It measures the two-dimensional area underneath the entire ROC curve, providing an aggregate measure of performance across all possible classification thresholds. ROC AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0.

Confusion Matrix

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

A **confusion matrix** is a table that summarizes the performance of a classification model.

Once **class recalls** and **class precisions** have been calculated, they are displayed around the edges of a confusion matrix. Also, after **accuracy** has been calculated, it will appear at the upper left of matrix.

Logistic Regression

accuracy: 72.86% +/- 5.04% (micro average: 72.86%)

	true false	true true	class precision
pred. false	156	52	75.00%
pred. true	62	150	70.75%
class recall	71.56%	74.26%	

Logistic Regression_no Unknown_smoker_sample0.07.rmp

Decision Tree

accuracy: 67.38% +/- 4.35% (micro average: 67.38%)

	true false	true true	class precision
pred. false	163	82	66.53%
pred. true	55	120	68.57%
class recall	74.77%	59.41%	

DecisionTree_no Unknown_smoker_sample0.07.rmp

k-nearest neighbors (*k*-NN)

accuracy: 72.62% +/- 5.64% (micro average: 72.62%)

	true false	true true	class precision
pred. false	153	50	75.37%
pred. true	65	152	70.05%
class recall	70.18%	75.25%	

KNN_no Unknown_smoker_sample0.07_numerical input.rmp

Subset: Stroke (true/false : 202/218, 1/1.08)

Naïve Bayes

accuracy: 72.38% +/- 4.52% (micro average: 72.38%)

	true false	true true	class precision
pred. false	153	51	75.00%
pred. true	65	151	69.91%
class recall	70.18%	74.75%	

NaiveBayes_no Unknown_smoker_sample0.07.rmp

Neural Network

accuracy: 70.48% +/- 5.04% (micro average: 70.48%)

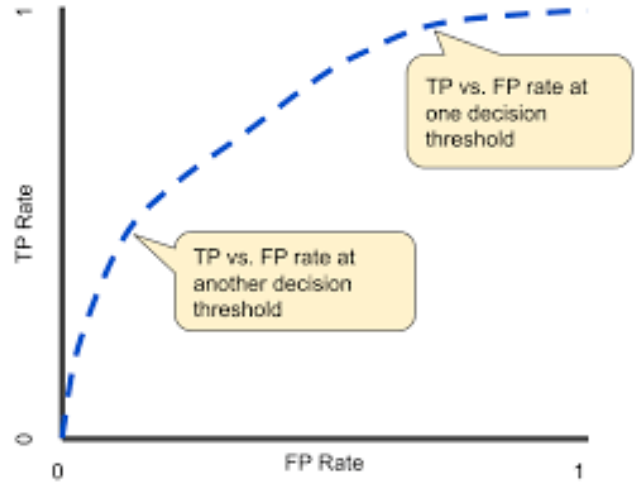
	true false	true true	class precision
pred. false	157	63	71.36%
pred. true	61	139	69.50%
class recall	72.02%	68.81%	

NN_no Unknown_smoker_sample0.07.rmp

Overall, the models have an accuracy ranging from 67% to 73%. Decision Tree model has a lower accuracy, 67.38% and the Neural Network model has the second lowest accuracy of 70.48%. The remaining three models (Logistic Regression, k-NN, and Naïve Bayes) have an accuracy above 72%.

Subset: Stroke (true/false : 202/218, 1/1.08)

ROC Curves of Models



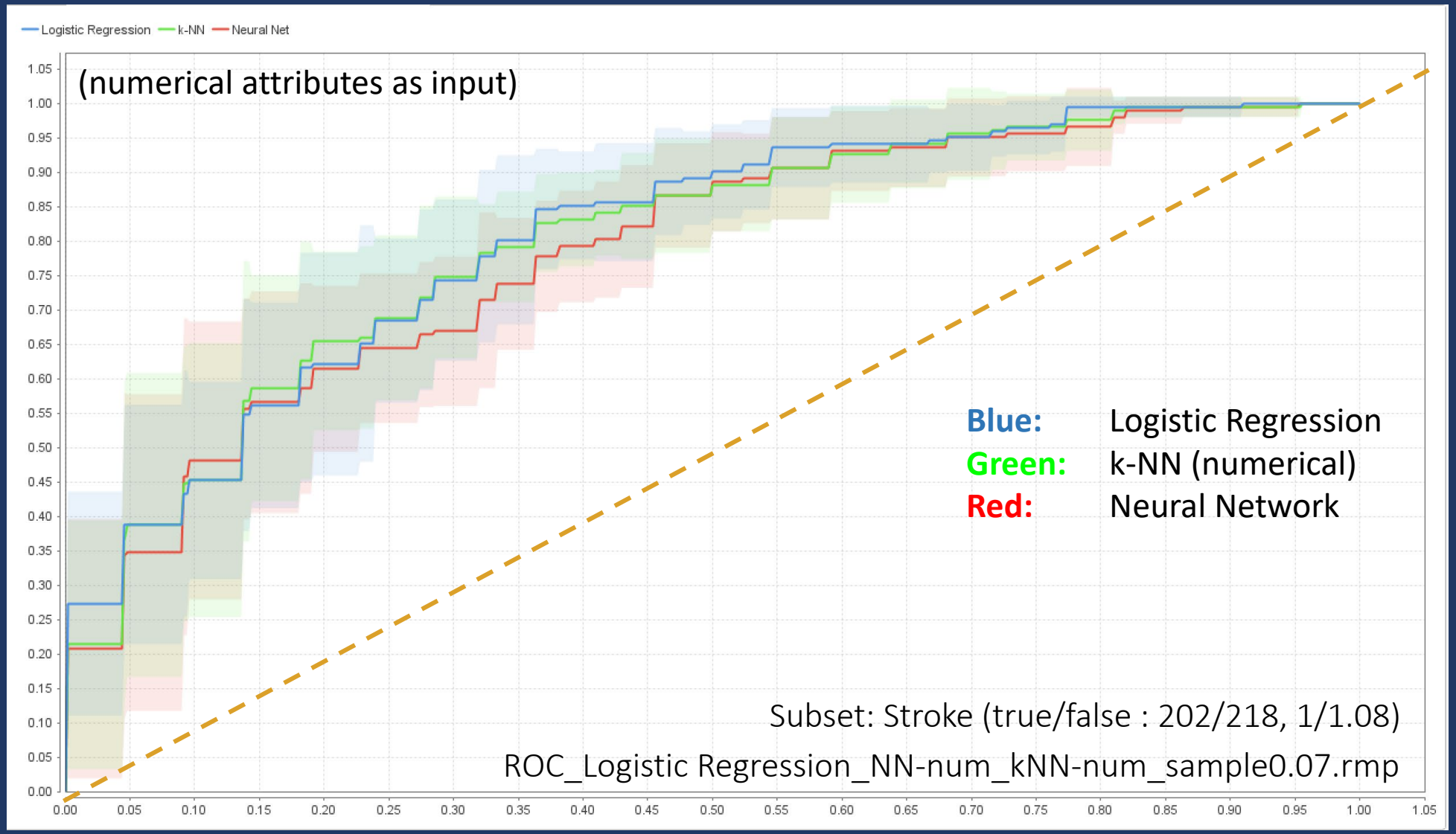
An **ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate (a.k.a., sensitivity or recall), on vertical axis
- False Positive Rate, on horizontal axis

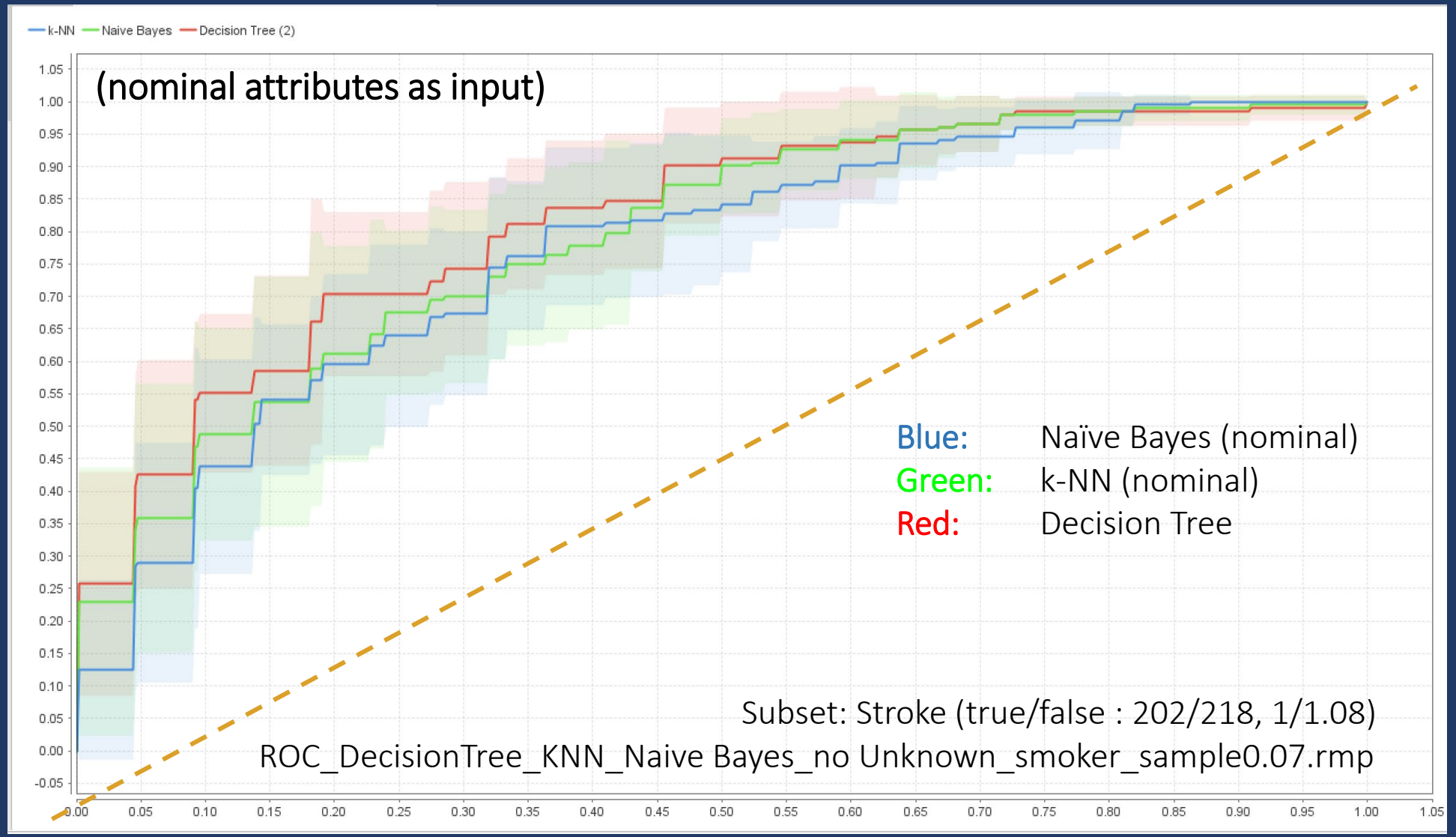
An ROC curve plots TPR vs. FPR at different classification thresholds.The diagonal line in a ROC curve represents a random classifier. In other words, a test that follows the diagonal has no better odds of detecting something than a random flip of a coin. **The area under the curve (AUC)** provides a single number to summarize the performance of a model.

Classification threshold represents the decision-making boundary that need to be defined before deploying a model. In this project, values above this threshold will be mapped to the stroke category, while those below or at the threshold will be mapped to the no_stroke category.

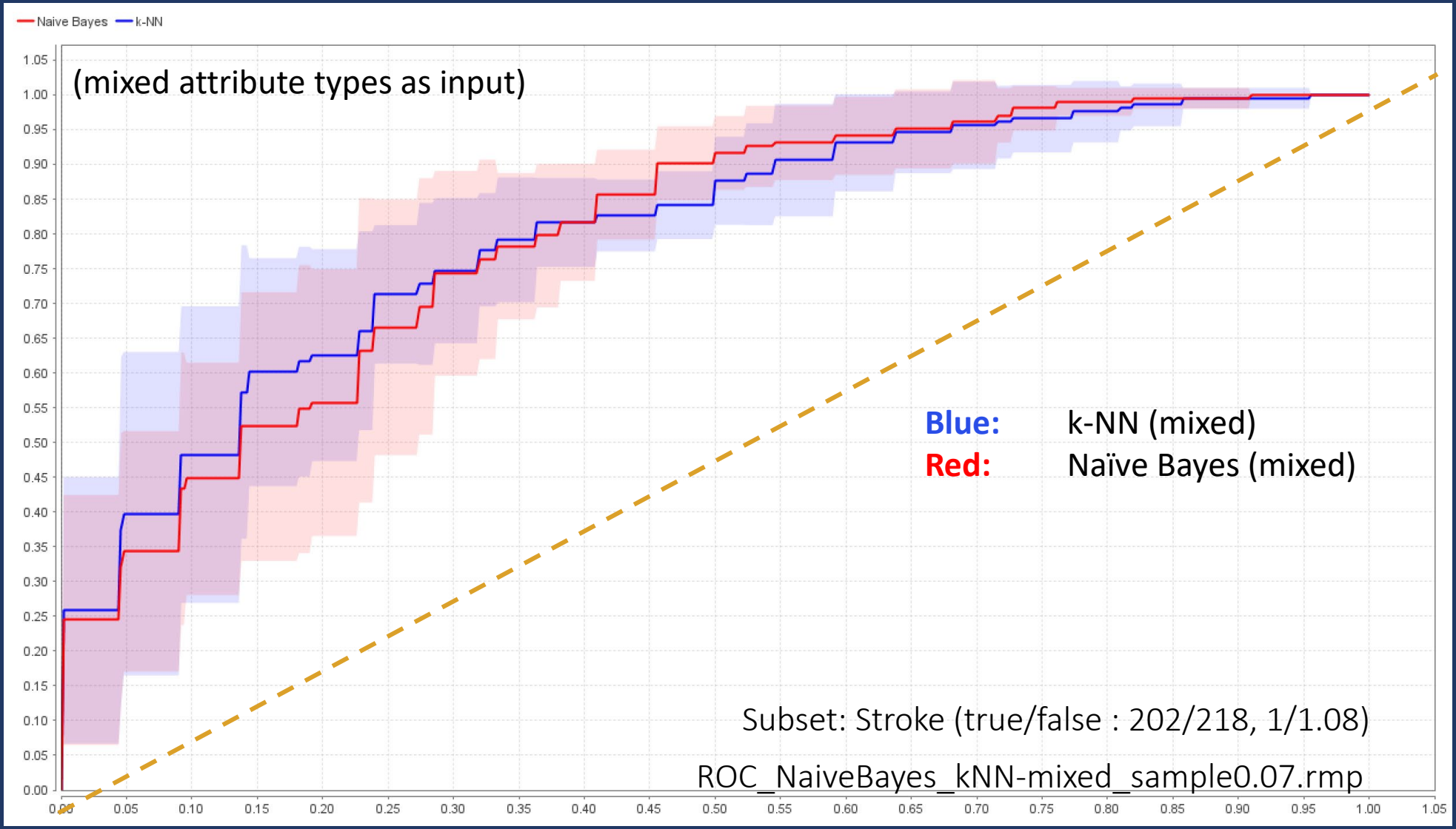
ROC Comparison



ROC Comparison cont'd



ROC Comparison cont'd



Conclusion and Appendix

Conclusion

- The correlation analysis performed on the attributes excluding the “stroke” attribute, confirmed that the attributes are slightly correlated at most. This is important because models like Naïve Bayes assumes that any pair of predictors/attributes are independent of each other.
- The top three significant predictors are age, glucose level, and hypertension. These were first identified in the Logistic Regression model.
- Then, they appeared in the Decision Tree model as the decision nodes in one of the branchs of 65-74 age group and some of them showed up in the branch of 75-84 age group. This makes sense because two-third of stroke occur over age of 65. Thus, the findings from the two models are consistent.
- Later, these three significant predictors are among those input nodes that have strong connections within the Neural Network model. Also, one more strongly connected node in the Neural Network model is the new attribute “smoker.” This is plausible because smoking damages the cells and/or help form blood clots in blood vessels, increasing the risk of stroke.

Conclusion

Cont'd

- K-folds cross validation was carried out on each classification model to test it on some unseen data. The result was summarized in a confusion matrix.
- Based on the confusion matrices, the models are generally 67% to 73% accurate.
- Logistic Regression, k-NN, and Naïve Bayes have an accuracy above 72%.
- Decision Tree and Neural Network have lower accuracies of 67% and 70%, respectively.
- In the confusion matrix of each model, the values of class precision and class recall of the "true" class (i. e. have stroke), for the most part, are of the same magnitude as its accuracy.
- A close examination of the ROC curves of these models indicates that the ROCs of these classification models have a rather similar AUC. In addition, the ROCs are nearly superimposed, meaning that they perform comparably at most classification thresholds.
- The fact that the ROC curves of the models are located on the left side of the diagonal line and far away from it means that the models are performing much better than random classification represented by the diagonal line.
- Before deploying these models to classify whether or not a patient will have a stroke, it is necessary to define the classification threshold as either the geometric mean of TPR and TNR or a threshold desirable for a specific application.

Appendix

RapidMiner Operator https://docs.rapidminer.com/latest/studio/operators/rapidminer-studio-operator-reference.pdf
Classification https://developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative
K-folds Cross Validation https://medium.datadriveninvestor.com/k-fold-cross-validation-6b8518070833
Logistic regression https://www.saedsayad.com/logistic_regression.htm
Decision tree https://www.saedsayad.com/decision_tree.htm#:~:text=Decision%20tree%20builds%20classification%20or,decision%20nodes%20and%20leaf%20nodes.
Naïve Bayes https://www.saedsayad.com/naive_bayesian.htm
K-nearest neighbors https://www.saedsayad.com/k_nearest_neighbors_reg.htm
Neural network https://www.saedsayad.com/artificial_neural_network.htm
Confusion Matrix https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/
ROC curve and AUC https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc