Business Intelligence & Analytics (MIS 401)

Dr. Huiyu Qian

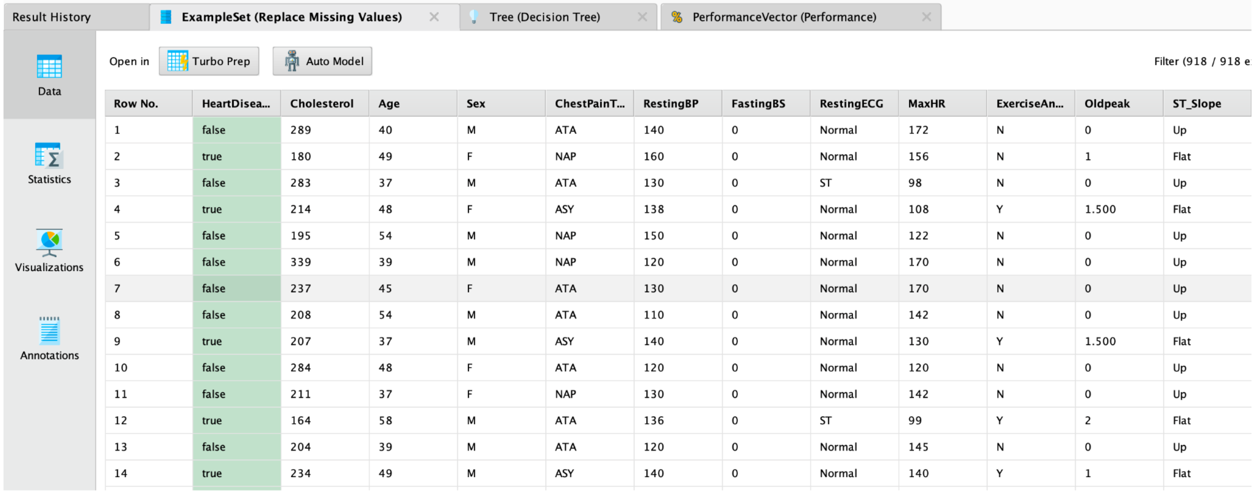
May 11, 2023

Team 5: The Avengers

Final Project: Heart Failure Prediction

By Manuel Coria, James Hurst, Basil Livaditis, Melissa Ortega

Data

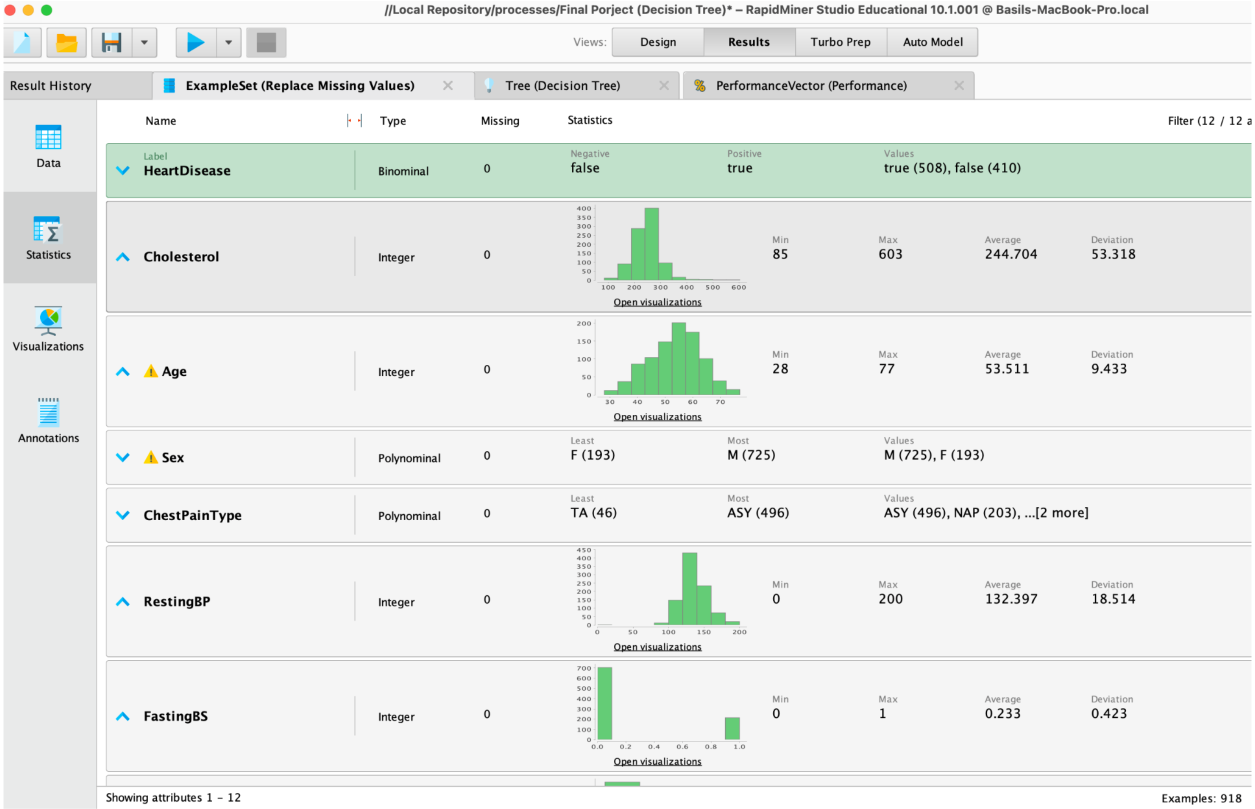


Filter (918 / 918 e

Row No.	HeartDisea...	Cholesterol	Age	Sex	ChestPainT...	RestingBP	FastingBS	RestingECG	MaxHR	ExerciseAn...	Oldpeak	ST_Slope
1	false	289	40	M	ATA	140	0	Normal	172	N	0	Up
2	true	180	49	F	NAP	160	0	Normal	156	N	1	Flat
3	false	283	37	M	ATA	130	0	ST	98	N	0	Up
4	true	214	48	F	ASY	138	0	Normal	108	Y	1.500	Flat
5	false	195	54	M	NAP	150	0	Normal	122	N	0	Up
6	false	339	39	M	NAP	120	0	Normal	170	N	0	Up
7	false	237	45	F	ATA	130	0	Normal	170	N	0	Up
8	false	208	54	M	ATA	110	0	Normal	142	N	0	Up
9	true	207	37	M	ASY	140	0	Normal	130	Y	1.500	Flat
10	false	284	48	F	ATA	120	0	Normal	120	N	0	Up
11	false	211	37	F	NAP	130	0	Normal	142	N	0	Up
12	true	164	58	M	ATA	136	0	ST	99	Y	2	Flat
13	false	204	39	M	ATA	120	0	Normal	145	N	0	Up
14	true	234	49	M	ASY	140	0	Normal	140	Y	1	Flat

Attribute Information (Basil) (Data source: Fedesoriano)

- **Age**: patients age in years
- **Sex**: gender of the patient (M: male or F: female)
- **ChestPainType: TA**, (Typical angina), **ATA** (atypical angina), **NAP** (Non-Anginal) Pain, **ASY** (asymptomatic)
- **RestingBP**: resting blood pressure (mm/HG)
- **Cholesterol:** serum cholesterol (mm/dl)
- **FastingBS**: fasting blood sugar (**1:** if FastingBS > 120 mg/dl, **0:** any other amount)
- **RestingECG**: resting electro cardiogram results (**Normal**: Normal, **ST**: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), **LVH**: showing probable or definite left ventricular hypertrophy by Estes' criteria)
- **MaxHR:** maximum heart rate achieved (Numeric value between 60 - 202)
- **ExerciseAngina:** exercise-induced angina (Y: Yes, N: No)
- **Oldpeak:** oldpeak = **ST** (Numeric value measured in depression)
- **ST_Slope:** slope of the peak exercise ST segment (**Up**: upsloping, **Flat**: flat, **Down**: downsloping)
- **HeartDisease:** output class (**1:** heart disease, **0:** Normal) **Label**



//Local Repository/processes/Final Porject (Decision Tree)* – RapidMiner Studio Educational 10.1.001 @ Basils-MacBook-Pro.local

Views: Design Results Turbo Prep Auto Model
Result History
ExampleSet (Replace Missing Values)
Tree (Decision Tree)
PerformanceVector (Performance)
Data
Statistics
Visualizations
Annotations

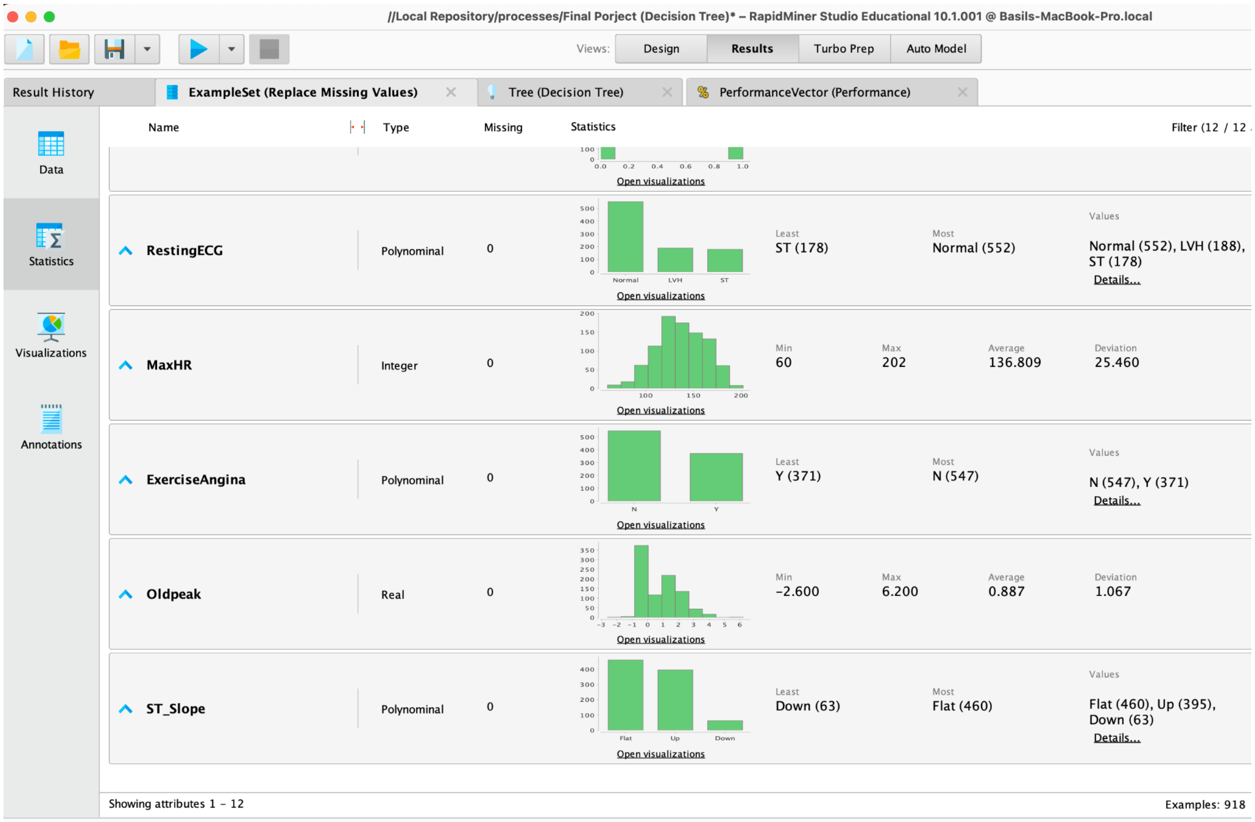
Filter (12 / 12 a

Name	Type	Missing	Statistics			
Label HeartDisease	Binominal	0	Negative false	Positive true	Values true (508), false (410)	
Cholesterol	Integer	0	Min 85	Max 603	Average 244.704	Deviation 53.318
Age	Integer	0	Min 28	Max 77	Average 53.511	Deviation 9.433
Sex	Polynominal	0	Least F (193)	Most M (725)	Values M (725), F (193)	
ChestPainType	Polynominal	0	Least TA (46)	Most ASY (496)	Values ASY (496), NAP (203), ...[2 more]	
RestingBP	Integer	0	Min 0	Max 200	Average 132.397	Deviation 18.514
FastingBS	Integer	0	Min 0	Max 1	Average 0.233	Deviation 0.423

Open visualizations

Showing attributes 1 – 12

Examples: 918



//Local Repository/processes/Final Porject (Decision Tree)* – RapidMiner Studio Educational 10.1.001 @ Basils-MacBook-Pro.local

Views: Design Results Turbo Prep Auto Model
Result History
ExampleSet (Replace Missing Values)
Tree (Decision Tree)
PerformanceVector (Performance)
Data
Statistics
Visualizations
Annotations

Filter (12 / 12

Name	Type	Missing	Statistics			
RestingECG	Polynominal	0	Least ST (178)	Most Normal (552)	Values Normal (552), LVH (188), ST (178) Details...	
MaxHR	Integer	0	Min 60	Max 202	Average 136.809	Deviation 25.460
ExerciseAngina	Polynominal	0	Least Y (371)	Most N (547)	Values N (547), Y (371) Details...	
Oldpeak	Real	0	Min -2.600	Max 6.200	Average 0.887	Deviation 1.067
ST_Slope	Polynominal	0	Least Down (63)	Most Flat (460)	Values Flat (460), Up (395), Down (63) Details...	

Open visualizations

Showing attributes 1 – 12

Examples: 918

Dataset Explained (Melissa):

The Heart Failure Prediction dataset is a combination of a collection of data from five different sources, featuring 918 observations of patients who have experienced heart failure across a variety of different variables. Each column represents a variable/attribute of the individual patient, and each row symbolizes a patient who has experienced heart failure. Given that cardiovascular disease is the leading cause of death globally, this dataset allows for early detection of individuals at high risk of heart disease, which allows for early intervention, in order to manage the outcome. The dataset consists of 11 key variables that can help in predicting the likelihood of heart disease, and a total of 12 overall attributes including: age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting electrocardiogram results, maximum heart rate achieved, exercise-induced angina, old peak, ST_Slope, and heart disease. Overall, the importance of the dataset comes from its ability to effectively predict the likelihood of heart failure among individuals, which improves outcomes and reduces the number of deaths resulting from cardiovascular disease.

Objective Analysis (Basil)

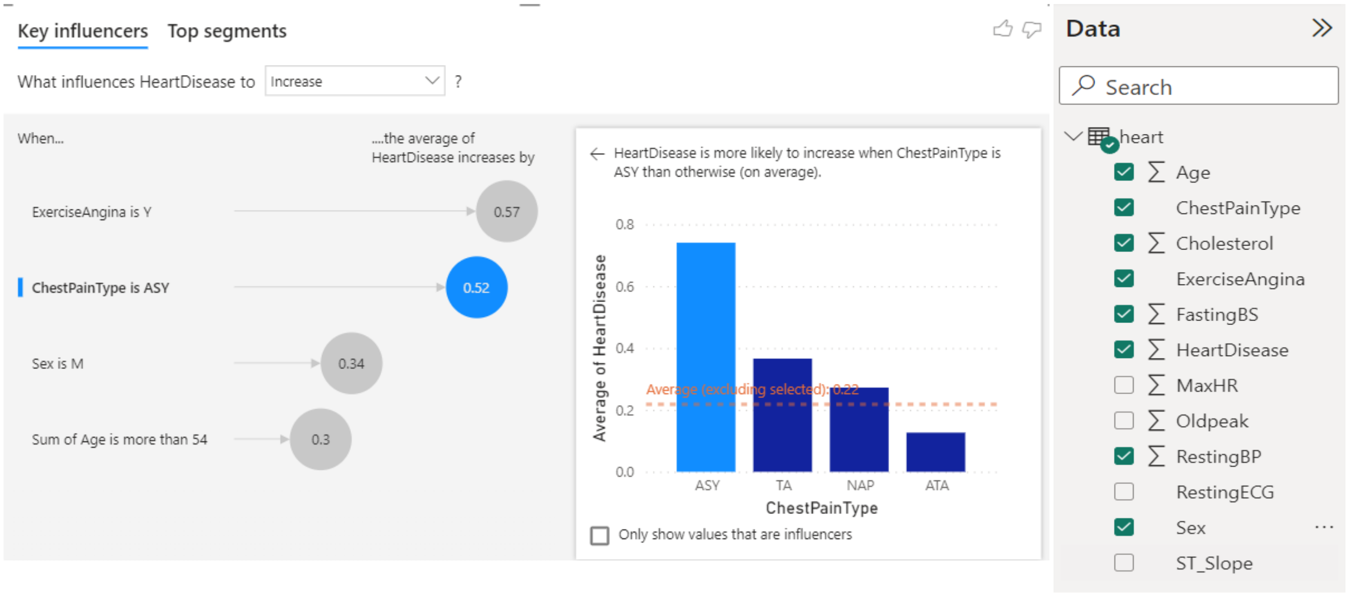
(Data source: Fedesoriano)

To perform analysis to the best of our ability we must first identify and understand the objective. Cardiovascular (heart) disease, the number one cause of death taking 17.9 million lives per year. Heart disease is hard to detect and can go unnoticed for years, detecting the disease early on is key in reducing the number of fatalities it may cause. Through the use of machine learning and predictive modeling we can hope to better predict the likelihood of a patient having cardiovascular disease based on 11 primary features (listed and explained in the attribute information section).

Business/Research Objectives (Manny)

The business/research objective of the Heart Failure Prediction dataset is to further analyze the combined heart failure data from the provided five separate data sources; in order to analyze in depth, accurately predict, and possibly prevent heart failure in patients based on where they land on the provided 12 overall attributes. From a medical standpoint, detecting particular patterns in heart failure with the provided attributes, can help a business create new medical equipment or software. This new equipment or software, can help doctors accurately inform patients of the likelihood of heart failure based on the analyzed demographics.

Descriptive Analysis using Power BI (James)



What influences HeartDisease to Increase ?

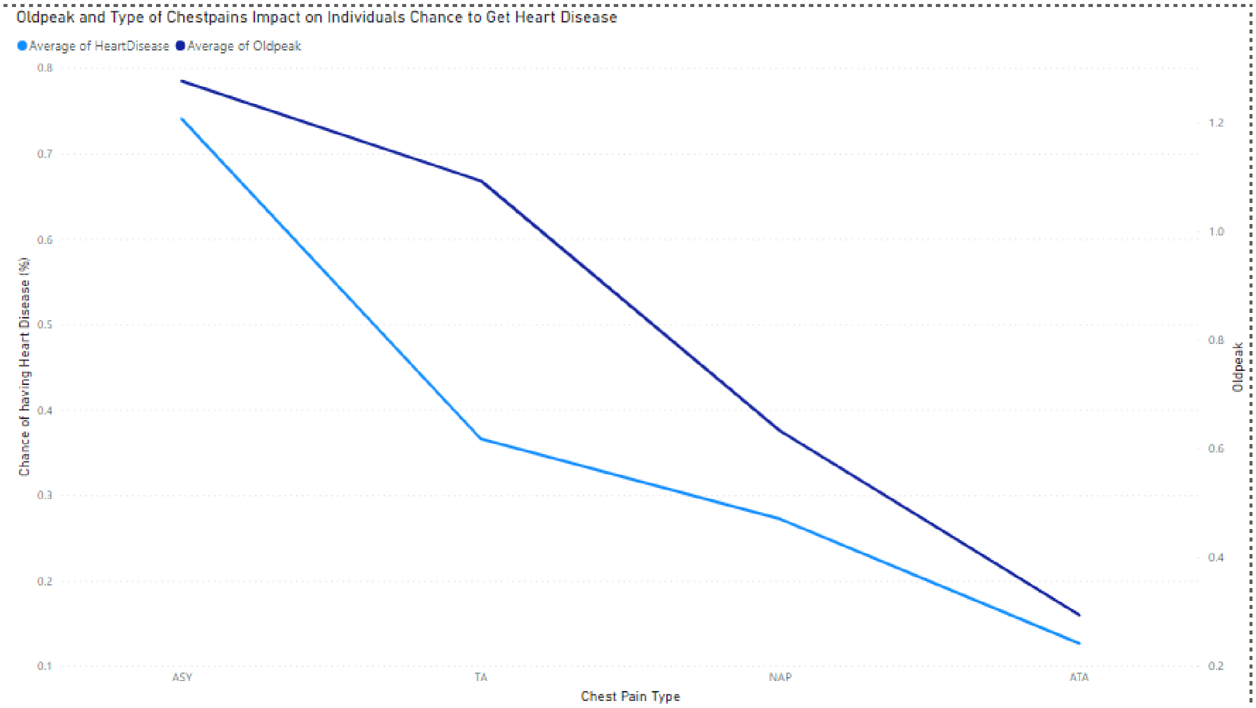
When...
...the average of HeartDisease increases by
ExerciseAngina is Y
0.57
ChestPainType is ASY
0.52
Sex is M
0.34
Sum of Age is more than 54
0.3

HeartDisease is more likely to increase when ChestPainType is ASY than otherwise (on average).

□ Only show values that are influencers

☑ ∑ Age
☑ ChestPainType
☑ ∑ Cholesterol
☑ ExerciseAngina
☑ ∑ FastingBS
☑ ∑ HeartDisease
□ ∑ MaxHR
□ ∑ Oldpeak
☑ ∑ RestingBP
□ RestingECG
☑ Sex
□ ST_Slope

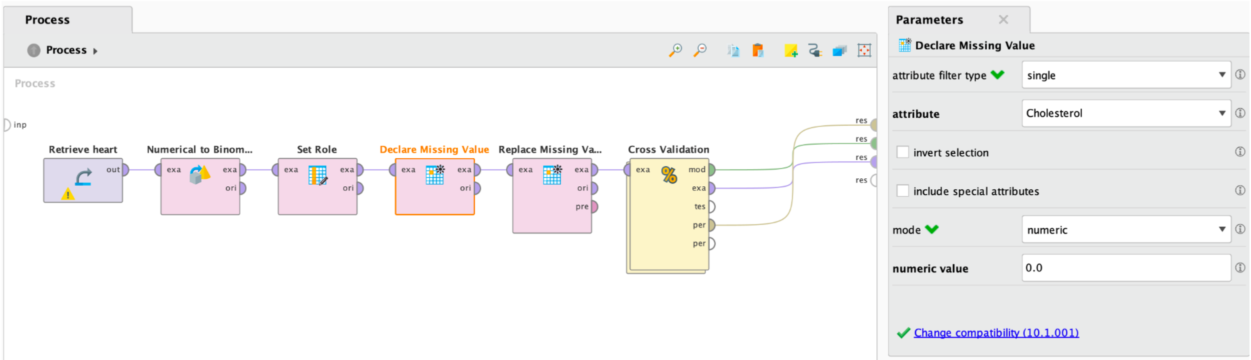
The visualization above is a key influencer's chart measuring what factors have the greatest influence on Heart Disease. The attributes we chose to measure include age, chest pain type, Cholesterol, exercise angina, fasting blood sugar, and resting blood pressure. In PowerBI's data prep, I excluded the missing values of cholesterol because it was swaying the data. In addition to this I chose to change the attributes I was using to explain Heart Disease to 'Do Not Summarize,' so it is not comparing the total quantities of each attribute, instead it is analyzing the true data. As seen by the visualization the biggest influencers to having heart disease include having exercise angina and having asymptomatic chest pain. Other factors that have an impact on having heart disease include, being male and being older than 54 years of age. The biggest influence on whether or not an individual has Heart disease is if they experience Exercise Angina. Because having heart disease is on a 0-1 scale, the impact of the attributes are measured as percentages. So, when someone answered yes to having Exercise Angina, they are 57% more likely to have Heart Disease. As seen in the graph when an individual has asymptomatic chest pain the average of heart disease increases by 52%. This is followed by typical angina chest pain, then non-anginal chest pain.



The visualization above is a line chart that shows the correlation between chest pain type and Oldpeak and its effect on the percent chance of having Heart Disease. As you can see the average of oldpeak is strongly correlated with the type of chest pain. Your chance of getting heart disease is at its highest when chest pain type is ASY and oldspeak is greater than one, meaning that having both may drastically increase your risk of heart disease. Chest pain types of TA and NAP are less impactful in increasing ones chance of getting heart disease but still raise chances by 25-35%. ATA chest pain has the least impact on heart disease and has heavy correlation with the low value of oldpeak. Lastly it is important to note that when oldpeak increases individuals Chance of getting heart disease also increases.

Machine Learning

Decision Tree: Set up (Basil)



Declare Missing Value

attribute filter type: single

attribute: Cholesterol

□ invert selection

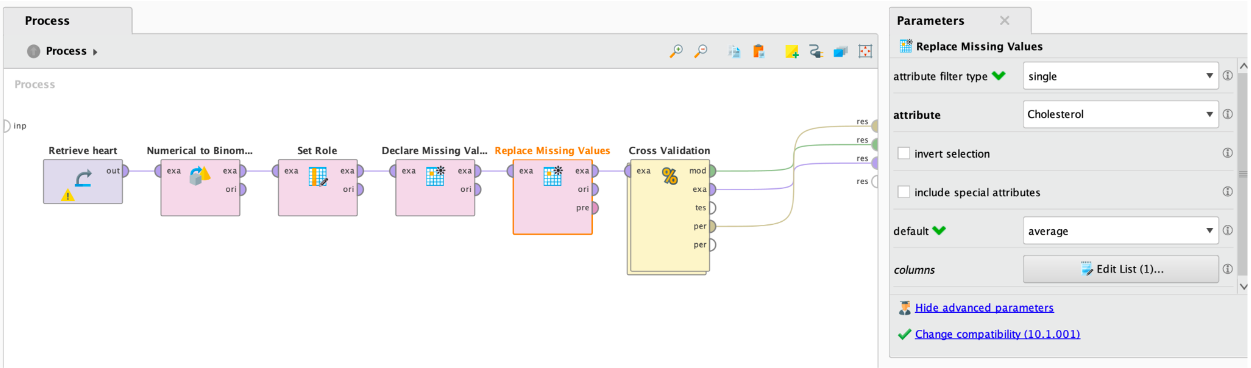
□ include special attributes

mode: numeric

numeric value: 0.0

Change compatibility (10.1.001)

In order to use the data set properly first we need to clean it up. The main issue derived from the raw data set is that the attribute, cholesterol, had over 150 missing values all inputted as 0. This mistake drastically lowered the data sets average cholesterol and skewed the data to fix this issue. I declared any value of 0 in the cholesterol column to be a missing value.



Replace Missing Values

attribute filter type: single

attribute: Cholesterol

□ invert selection

□ include special attributes

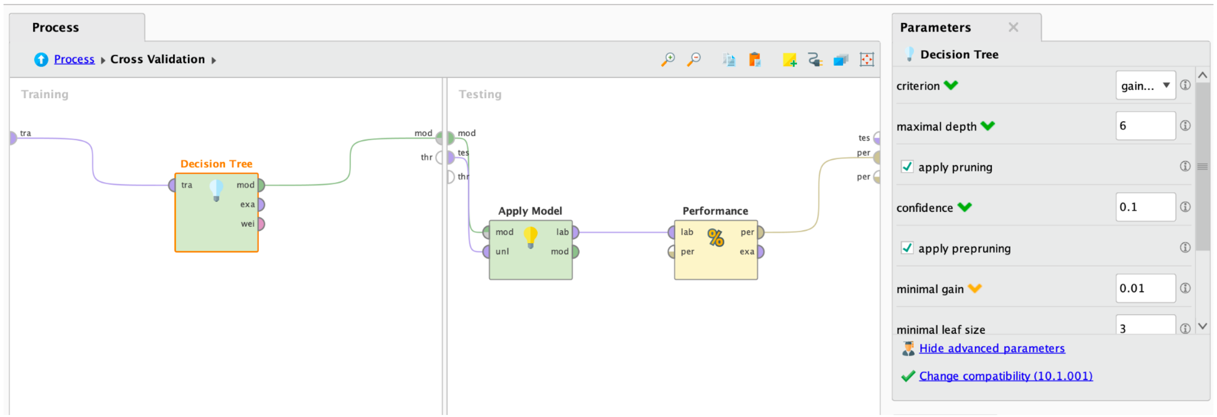
default: average

columns: Edit List (1)...

Hide advanced parameters

Change compatibility (10.1.001)

Now since the values of zero have been declared missing they can be fixed by adding a replace missing value function into the current process. By selecting average, all of the previous values listed as 0 will be replaced with the average cholesterol of the data set excluding the 0s. With that completed the data is now ready to use.



Decision Tree

criterion: gain...

maximal depth: 6

☑ apply pruning

confidence: 0.1

☑ apply prepruning

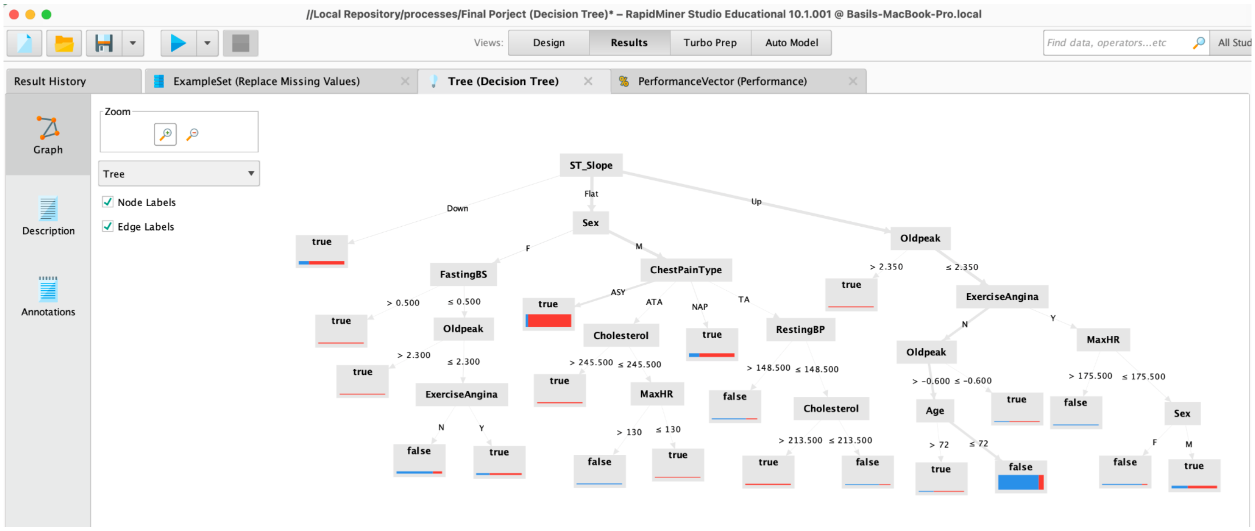
minimal gain: 0.01

minimal leaf size: 3

Hide advanced parameters

Change compatibility (10.1.001)

Decision Tree: Results & Analysis (Basil)



Downward Slope: (6.86% of data set) the least amount of the population fits the category

- If your ST slope is sloped downwards you are more likely to have heart disease, in fact out of 63 cases 49 had heart disease and a downward slope while only 14 didnt have heart disease while having a downward slope. The prediction accuracy of this section 77.78% which could be more accurate, but does not take other attributes into account hence the lower accuracy of prediction.

Flat Slope: (50.1% of the data set) Majority of the population falls into this category
This category holds the most values and has the most detail, branches, and leaves meaning it will also have the most explanation.

- The first split in this section happens with sex splitting the groups into 78 (female) and 385 (male). The female population branches into the attribute of FastingBS (blood sugar), if the value of one fasting blood sugar is greater than 120 mg/dl, any other value is classified as 0. Females who have a fasting blood sugar of 120 mg/dl or more are extremely likely to have heart disease with the model predicting at 100% accuracy for that section.
    - Those who have a fasting blood sugar of less than 120 are then split again with the attribute Oldpeak (numeric value used to measure STdepression). Those who are score higher(have a oldpeak of more than 2.3) are extremely likely to have heart disease with the model predicting at 100% accuracy once again.
    - The other branch containing those with an oldpeak of 2.3 or less, branches with the attribute of exercise angina (chest pain). Those who do not have exercise induced angina, an oldpeak of 2.3 or less and a fasting bs of less than 120 mg/dl are less likely to have heart disease with 28 of the values being false and only 7 being true. This displays a 80% accurate prediction rate for this leaf. On the other hand those who do have exercise angina are more likely to have heart disease than those who don't have angina. This proves true with a 70% accuracy, meaning it is not the best predictor of heart disease, but still can make an impact.

- The other branch stemming from the attribute of sex creates a predictive group for males. The branch is split at the attribute of chest pain which has four different types (and branches), those being TA, (Typical angina), ATA (atypical angina), NAP (Non-Anginal) Pain, ASY (asymptomatic).
    - Those who have ASY and NAP chest pain are predicted to commonly have heart disease. 259 people with ASY chest pain have heart disease while only 15 did not, predicting with 94.5% accuracy (making it one of the most accurate predictors). An important thing to note about ASY chest pain is that 29.85% of the people surveyed have this type of chest pain, making it one of the most populated leaves. Those with NAP chest pain also are more common to have heart disease with 54 people having NAP and heart disease and only 16 not having heart disease.
    - Those who have ATA or TA chest pain branch off into different attributes. The ATA chest pain is split by the attribute of cholesterol. If you have a cholesterol of more than 245.5 or you are extremely likely to have heart disease (11/11 have heart disease).
    - The other branch (less than or equal to 245.5) is split again by the attribute of maximum heart rate, individuals with a maxHR of more than 130 bp/minute are not likely to have heart disease (6/6 for not having heart disease, with 100" accuracy). Whereas individuals with a maxHR of 130 or less are extremely likely to have heart disease (4/4 with 100% accuracy). If the max heart rate your heart can output is below 130 bp/min it is safe to assume your heart is not working at maximum capacity and is more susceptible to failure.
    - The TA chest pain branch is split by the attribute of resting blood pressure. Those who have a restingBP of over 148.5 are not likely to have heart disease with a ¾ ratio and a 75% accurate prediction rate. The other group of individuals with a retsingBP of less than 148.5 are split by cholesterol.
    - Those with a cholesterol of more than 213.5 are extremely likely to have heart disease with a 12/12 ratio and an 100% accurate prediction rate. Cholesterol has shown up multiple times in the tree and is a great predictor of heart disease. Those with 213.5 cholesterol of less are less likely to get heart disease ¾ with a 75% accurate prediction rate.

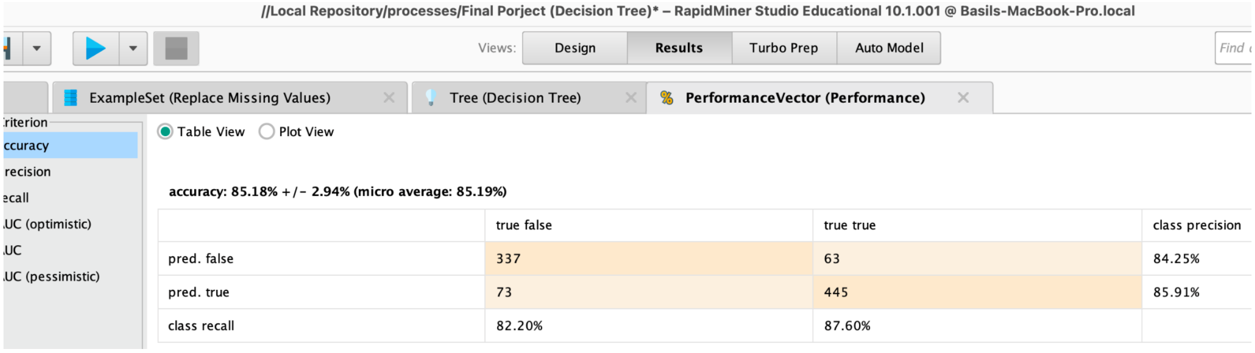
Upward slope: (43.03% of the data set) a large amount falls into this category

- The last branch that stems from Slope is the upward facing slope which is split by oldpeak producing two branches.The first is for those with an oldpeak of more than 2.35 who are extremely likely to have heart disease with 8/8 of the individuals in this category have heart disease. The other branch is those with a oldpeak of less than or equal to 2.35 which is split by exercise angina.
- Exercise angina has two branche stemming from it, one if you have it and one if you don't.
    - Those without exercise angina are split once again by the attribute if oldpeak, splitting them into those who have an old peak of less than or equal -0.6 and those who have one greater than that. Let it be noted that having and oldpeak

less than 0 is a positive in general. Those with an oldpeak of greater than -0.6 are split again by the attribute of age, creating two branches more than 72 years of age and less than or equal to 72 years of age. Individuals over 72 are likely to have heart disease ⅔ individuals, and those younger than 72 are less likely to have heart disease 289/325.

- Those with exercise angina are split by the attribute of maximum heart rate(maxHR). Individuals with a maxHR of over 175 are not likely to get heart disease 3/3. Those with a max hr less than or equal to 175 are split by sex . Males surveyed that fall into this category have heart disease fairly often with a ratio of 29/45. Where as females who also fall into this category are less likely to have heart disease with a ratio of ⅞ not having heart disease.

Confusion Matrix (basil)

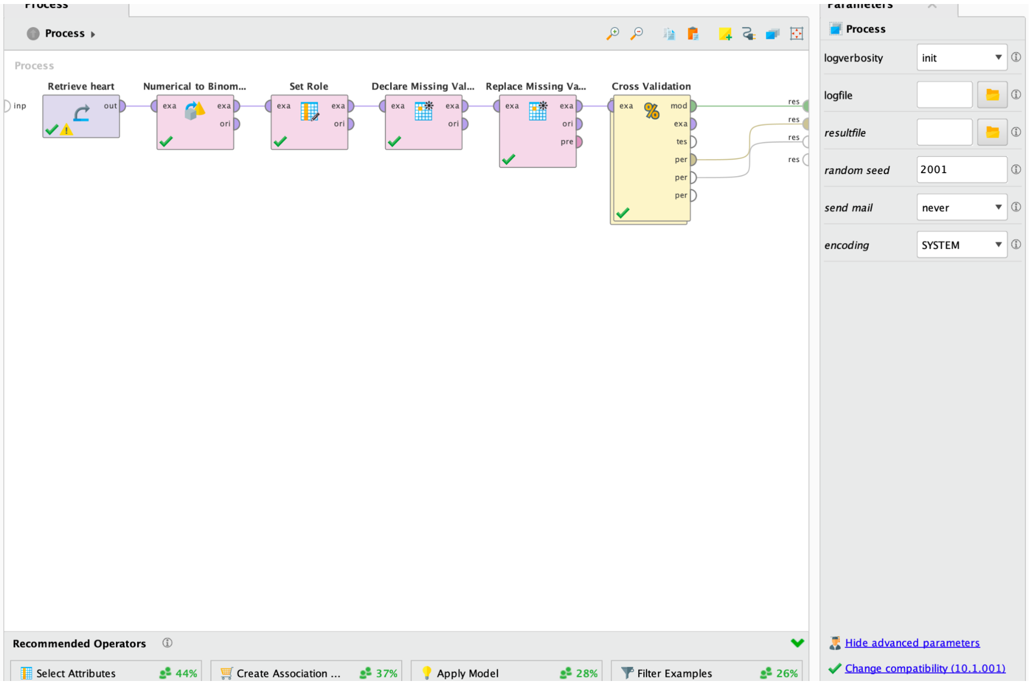

//Local Repository/processes/Final Porject (Decision Tree)* – RapidMiner Studio Educational 10.1.001 @ Basils-MacBook-Pro.local
Views:
Design
Results
Turbo Prep
Auto Model
Find
ExampleSet (Replace Missing Values)
Tree (Decision Tree)
PerformanceVector (Performance)
riterion
ccuracy
recision
ecall
UC (optimistic)
UC
UC (pessimistic)
● Table View
○ Plot View

accuracy: 85.18% +/- 2.94% (micro average: 85.19%)

	true false	true true	class precision
pred. false	337	63	84.25%
pred. true	73	445	85.91%
class recall	82.20%	87.60%	

The final part of the decision tree is the performance of the predictions. The confusion matrix depicts that the model predicted heart disease with a 85.18% prediction rate, which is a good prediction rate (various attempts were made to raise this number, but none were successful). By looking at the confusion matrix it can be seen that 337 values were predicted to be false and were actually false, while 63 values that were predicted to be false were actually true making the precision for false prediction 84.23%. The precision of predicted true vales is 85.91% with 73 predicted false values being true and 445 predicted value true values being true. Recall for false shows 337 predicted false values being true falses while 73 predicted true values were actually false, bringing the recall to a 82.2% sensitivity. For true recall the sensitivity was 87.6% with 63 predicted falses being true and 445 predicted true actually being true. Overall the decision tree is a good machine learning technique to use when trying to predict the likelihood of heart diseases as it breaks down the attributes and allows for various circumstances to be explored.

Logistic Regression (Melissa)



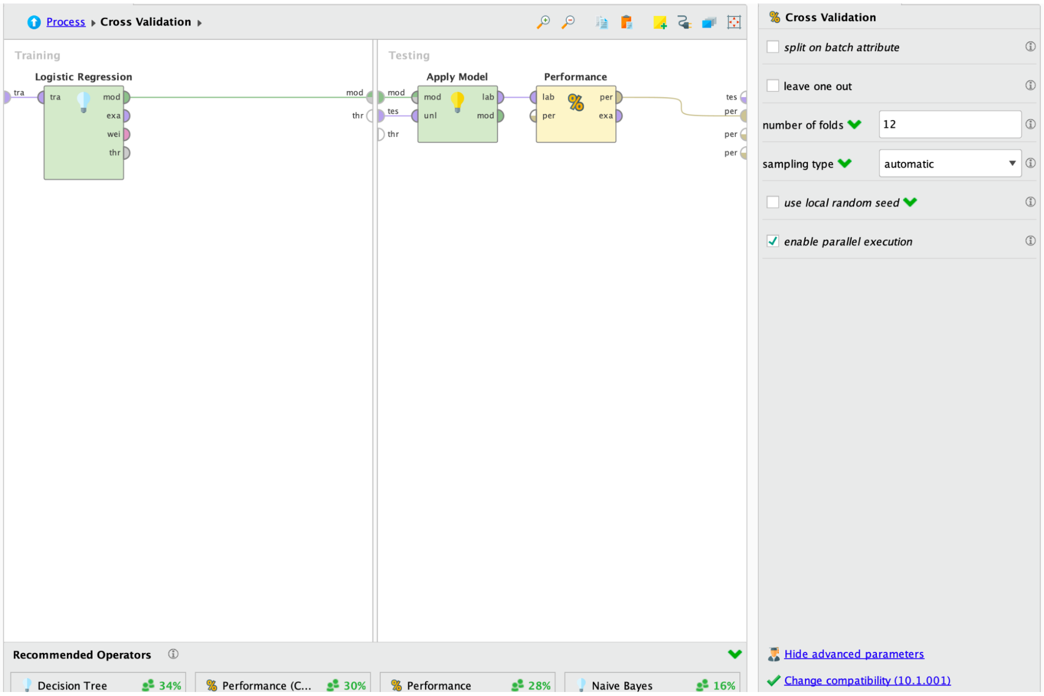
Process

logverbosity	init
logfile	
resultfile	
random seed	2001
send mail	never
encoding	SYSTEM

Hide advanced parameters

Change compatibility (10.1.001)

Select Attributes	44%
Create Association ...	37%
Apply Model	28%
Filter Examples	26%



Cross Validation

□ split on batch attribute

□ leave one out

number of folds	12
sampling type	automatic

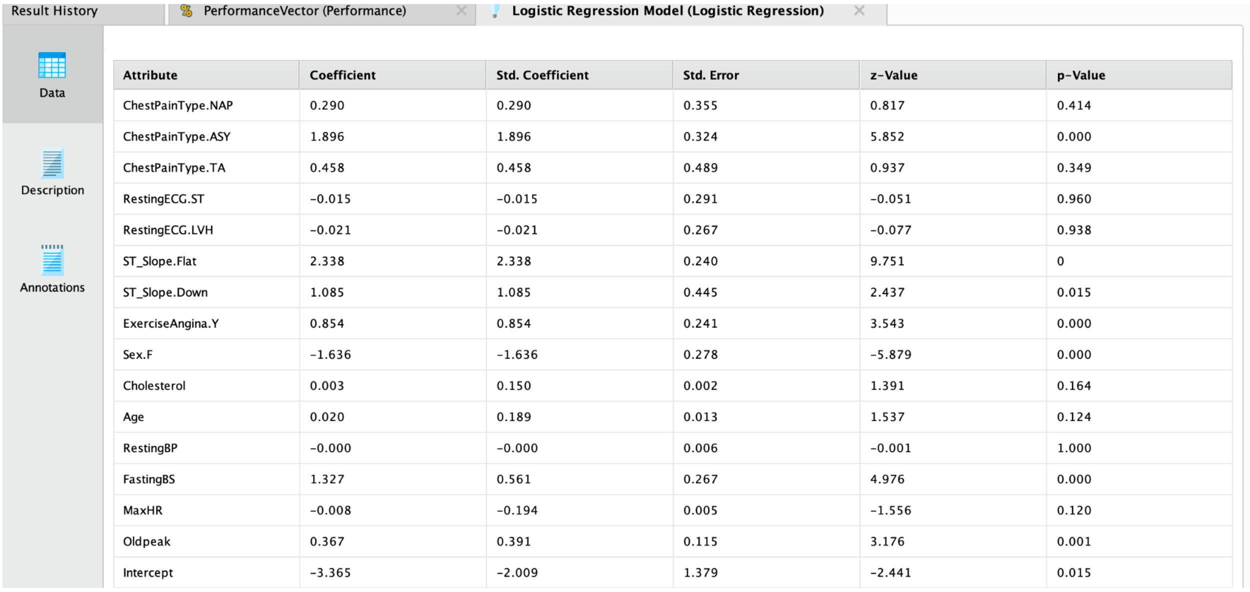
□ use local random seed

☑ enable parallel execution

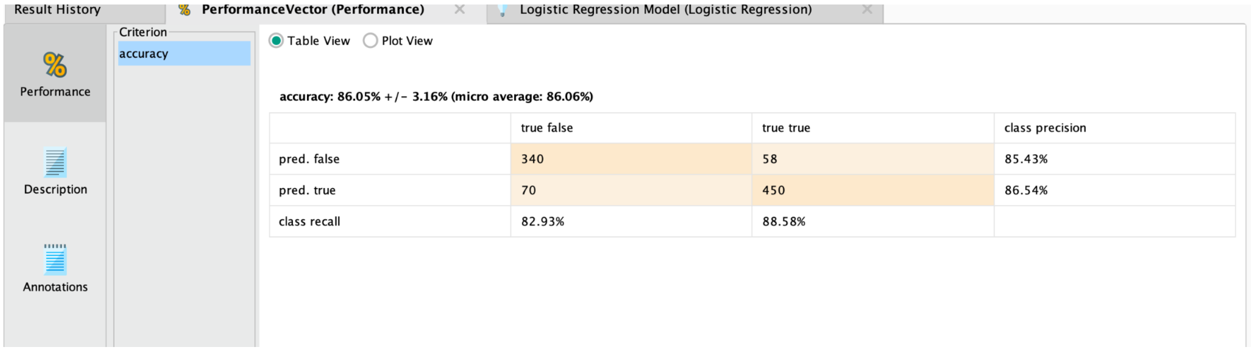
Hide advanced parameters

Change compatibility (10.1.001)

Decision Tree	34%
Performance (C...	30%
Performance	28%
Naive Bayes	16%


Result History
PerformanceVector (Performance)
Logistic Regression Model (Logistic Regression)
Data
Description
Annotations

Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
ChestPainType.NAP	0.290	0.290	0.355	0.817	0.414
ChestPainType.ASY	1.896	1.896	0.324	5.852	0.000
ChestPainType.TA	0.458	0.458	0.489	0.937	0.349
RestingECG.ST	-0.015	-0.015	0.291	-0.051	0.960
RestingECG.LVH	-0.021	-0.021	0.267	-0.077	0.938
ST_Slope.Flat	2.338	2.338	0.240	9.751	0
ST_Slope.Down	1.085	1.085	0.445	2.437	0.015
ExerciseAngina.Y	0.854	0.854	0.241	3.543	0.000
Sex.F	-1.636	-1.636	0.278	-5.879	0.000
Cholesterol	0.003	0.150	0.002	1.391	0.164
Age	0.020	0.189	0.013	1.537	0.124
RestingBP	-0.000	-0.000	0.006	-0.001	1.000
FastingBS	1.327	0.561	0.267	4.976	0.000
MaxHR	-0.008	-0.194	0.005	-1.556	0.120
Oldpeak	0.367	0.391	0.115	3.176	0.001
Intercept	-3.365	-2.009	1.379	-2.441	0.015



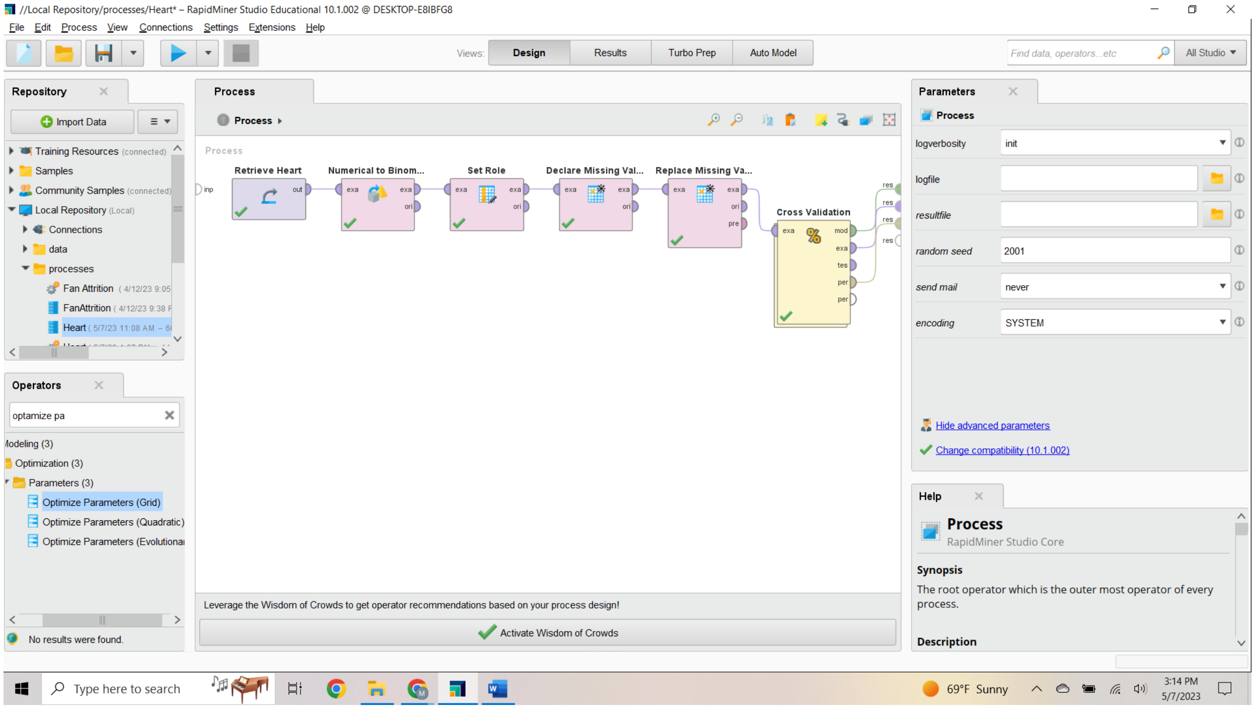
accuracy: 86.05% +/- 3.16% (micro average: 86.06%)

	true false	true true	class precision
pred. false	340	58	85.43%
pred. true	70	450	86.54%
class recall	82.93%	88.58%	

Explanation:

For the analysis above, I used the **logistic regression** method. The steps I took in this method were to first start with retrieving the heart failure prediction dataset, then I added numerical to binomial for a single attribute being HeartDisease, then I set role by selecting the attribute HeartDisease and making the target role label, after that to fix the problem with cholesterol within the data, I added declare missing values and put single attribute and chose that attribute to be cholesterol and to fix the problem I added replace missing value after that and did a single attribute being cholesterol and set the default to be the average in order to get rid of the zeros, after that I added cross validation and found 12 folds to be the most accurate. Moving forward, within cross validation I added logistic regression on the training side, and on the testing side I added apply model and performance (binomial classification). After running this process, the major outputs included the logistic regression model and the performance vector. First, the logistic regression model showed the attributes that had the most significance in terms of predicting heart failure, these predictors include the attributes with greater positive coefficients and smaller p-values. The logistic regression model in this case is showing the ASY chestpaintype to be one of the more significant predictors, alongside ST_Slope.Flat which is where the majority of the population falls which makes it a greater predictor. Similarly, the main predictors continue to be the ones that have high coefficients such as: ST_Slope.Down, and FastingBS. Second, the performance vector (confusion matrix) shows an accuracy of 86.05% with a standard deviation of 3.16% demonstrating the accuracy of predictions made within the data. Starting with the first row it can be seen that 340 values were predicted to be false and were actually false , while 58 were predicted to be false but were actually true, leading to the precision rate for that row to be 85.43%. Moving onto the second row, it can be seen that 70 values were predicted to be true but were actually false, while 450 values that were predicted to be true were actually true which leads to the precision rate for that row to be 86.54%. Further analyzing the confusion matrix, at the bottom the class recall for each column demonstrates the percentage of a certain class correctly identified so for "true false" the class recall can be seen as 82.93%, while for the class "true true" the class recall can be seen as 88.58%.

K-NN (Manny)


//Local Repository/processes/Heart* – RapidMiner Studio Educational 10.1.002 @ DESKTOP-E8IBFG8

Parameters – Process

logverbosity: init

logfile:

resultfile:

random seed: 2001

send mail: never

encoding: SYSTEM

Help – Process – RapidMiner Studio Core

Synopsis

The root operator which is the outer most operator of every process.

Description

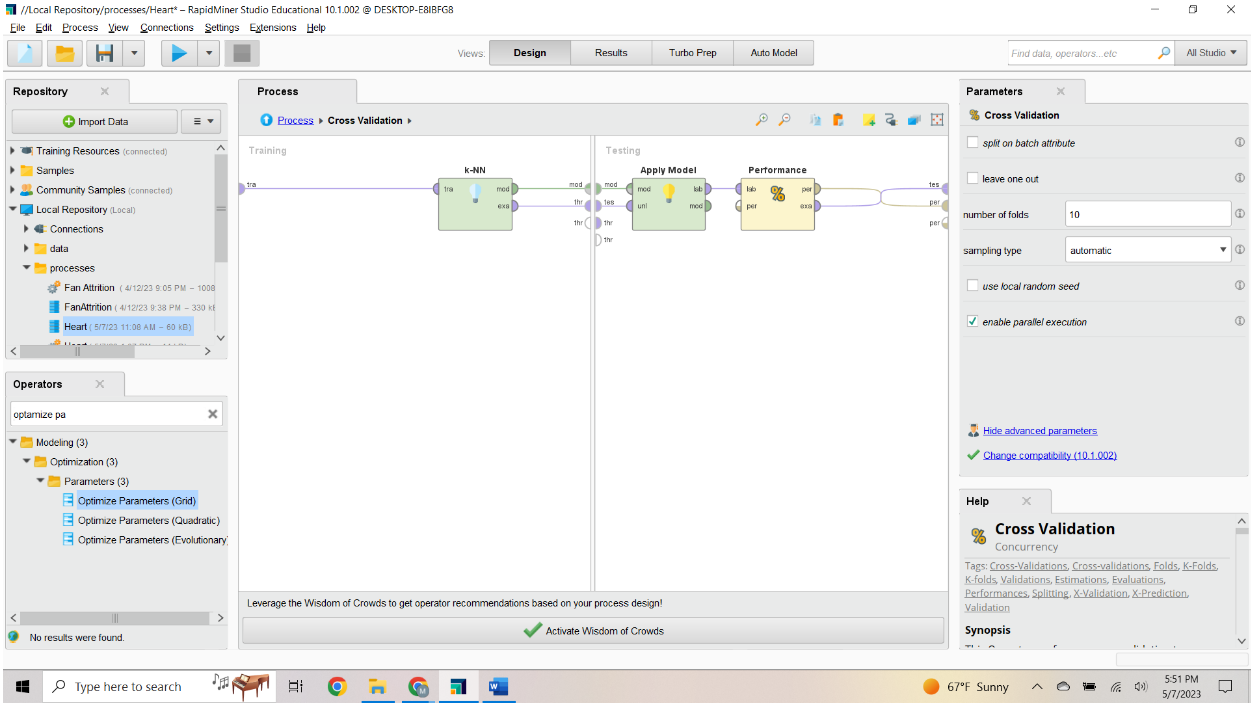
Operators search: optamize pa

No results were found.

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
processes
Fan Attrition (4/12/23 9:05
FanAttrition (4/12/23 9:38
Heart (5/7/23 11:08 AM – 6
Operators
Modeling (3)
Optimization (3)
Parameters (3)
Optimize Parameters (Grid)
Optimize Parameters (Quadratic)
Optimize Parameters (Evolutionar
Process
Retrieve Heart
Numerical to Binom...
Set Role
Declare Missing Val...
Replace Missing Va...
Cross Validation
inp
out
exa
ori
pre
mod
tes
per
res
Hide advanced parameters
Change compatibility (10.1.002)
Activate Wisdom of Crowds
Type here to search
69°F Sunny
3:14 PM
5/7/2023

The machine learning process above is the **K-NN** method. The first step is to import the heart failure prediction data intoRapidMiner . Adding "Numerical to Binomial" as an operator and making HeartDisease as the single attribute. Next, the data reflected a cholesterol of 0 which would not make sense since it is not possible for anyone to have a cholesterol level of 0 . In order to fix this issue, the operator to "Declare Missing Value" was used. Declared missing value pointed out the cholesterol of 0 while the operator "Replaced the Missing Value" replaced the 0 and made them an average. Lastly the operator Cross Validation was added.



Cross Validation

□ split on batch attribute

□ leave one out

number of folds: 10

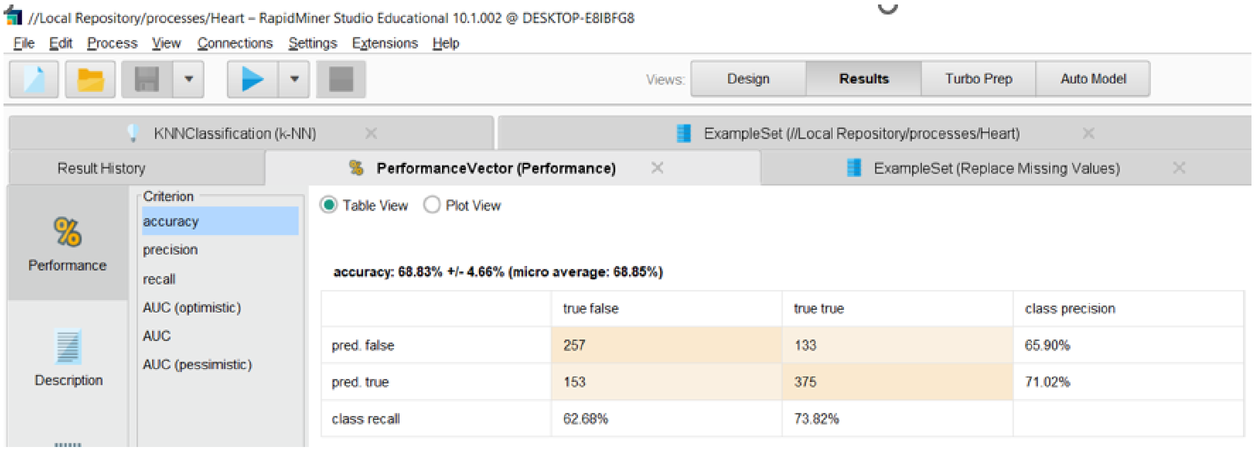
sampling type: automatic

□ use local random seed

☑ enable parallel execution

//Local Repository/processes/Heart* – RapidMiner Studio Educational 10.1.002 @ DESKTOP-E8IBFG8
Repository
Operators
optamize pa
Process
Training
Testing
k-NN
Apply Model
Performance
Parameters
Help
Cross Validation
Concurrency
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds

Inside the cross validation the operators K-NN , Apply Model and Performance were added.

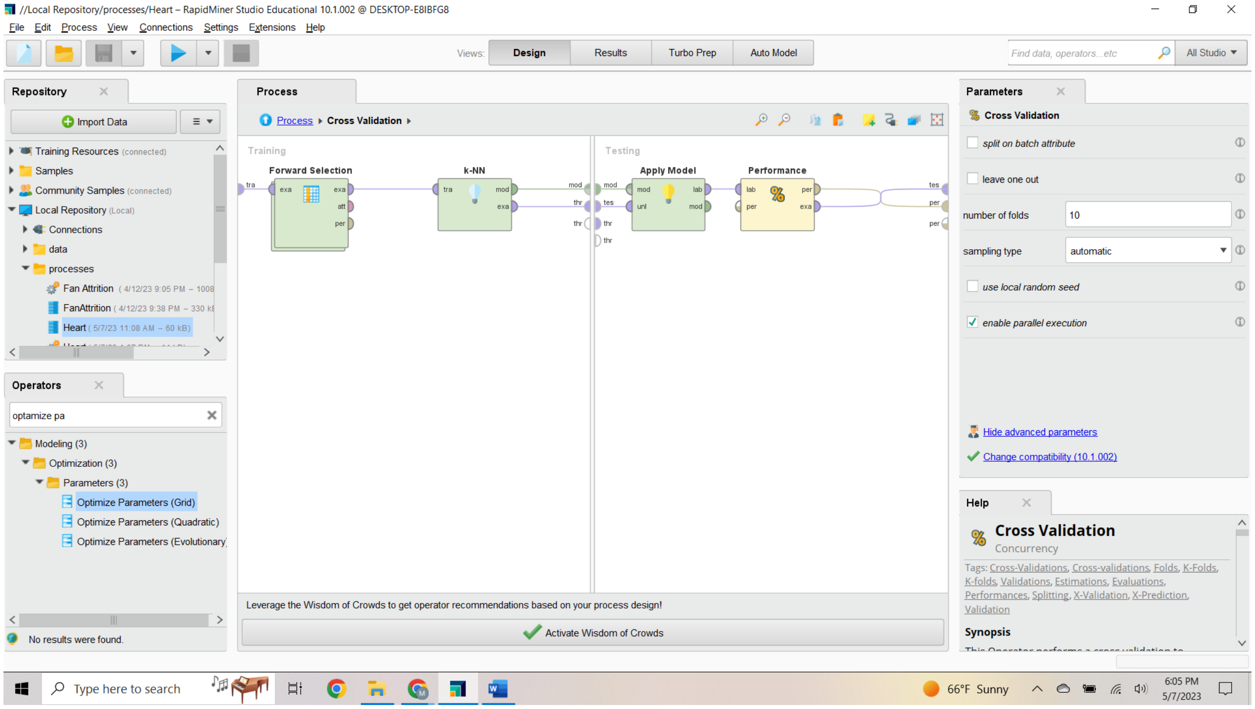


accuracy: 68.83% +/- 4.66% (micro average: 68.85%)

	true false	true true	class precision
pred. false	257	133	65.90%
pred. true	153	375	71.02%
class recall	62.68%	73.82%	

//Local Repository/processes/Heart – RapidMiner Studio Educational 10.1.002 @ DESKTOP-E8IBFG8
Results
KNNClassification (k-NN)
ExampleSet (//Local Repository/processes/Heart)
Result History
PerformanceVector (Performance)
ExampleSet (Replace Missing Values)
Performance
Description
Criterion
accuracy
precision
recall
AUC (optimistic)
AUC
AUC (pessimistic)
Table View
Plot View

After running the current process the result showed a low accuracy of 68.83%


//Local Repository/processes/Heart – RapidMiner Studio Educational 10.1.002 @ DESKTOP-E8IBFG8
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Repository
Process
Process ▸ Cross Validation ▸
Training
Testing
Forward Selection
k-NN
Apply Model
Performance
Operators
optamize pa
Modeling (3)
Optimization (3)
Parameters (3)
Optimize Parameters (Grid)
Optimize Parameters (Quadratic)
Optimize Parameters (Evolutionary)
No results were found.
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Help
Cross Validation
Concurrency
Synopsis

Parameters

Cross Validation

☐ split on batch attribute

☐ leave one out

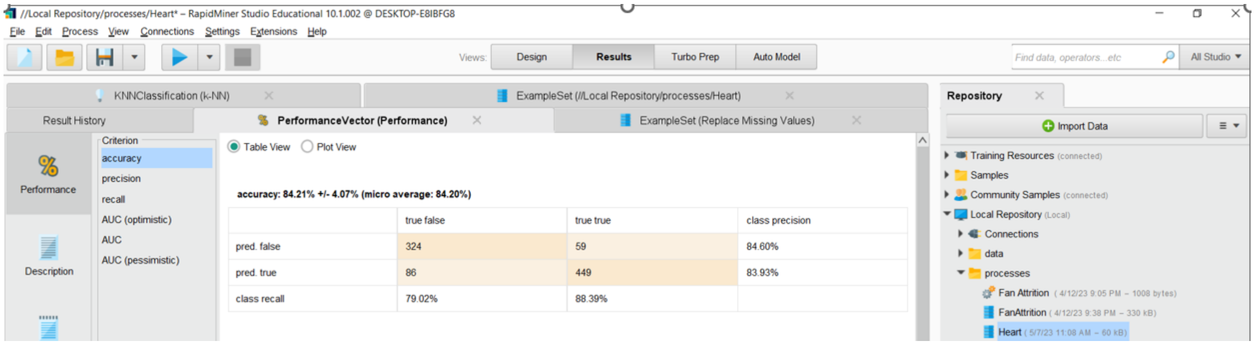
number of folds 10

sampling type automatic

☐ use local random seed

☑ enable parallel execution

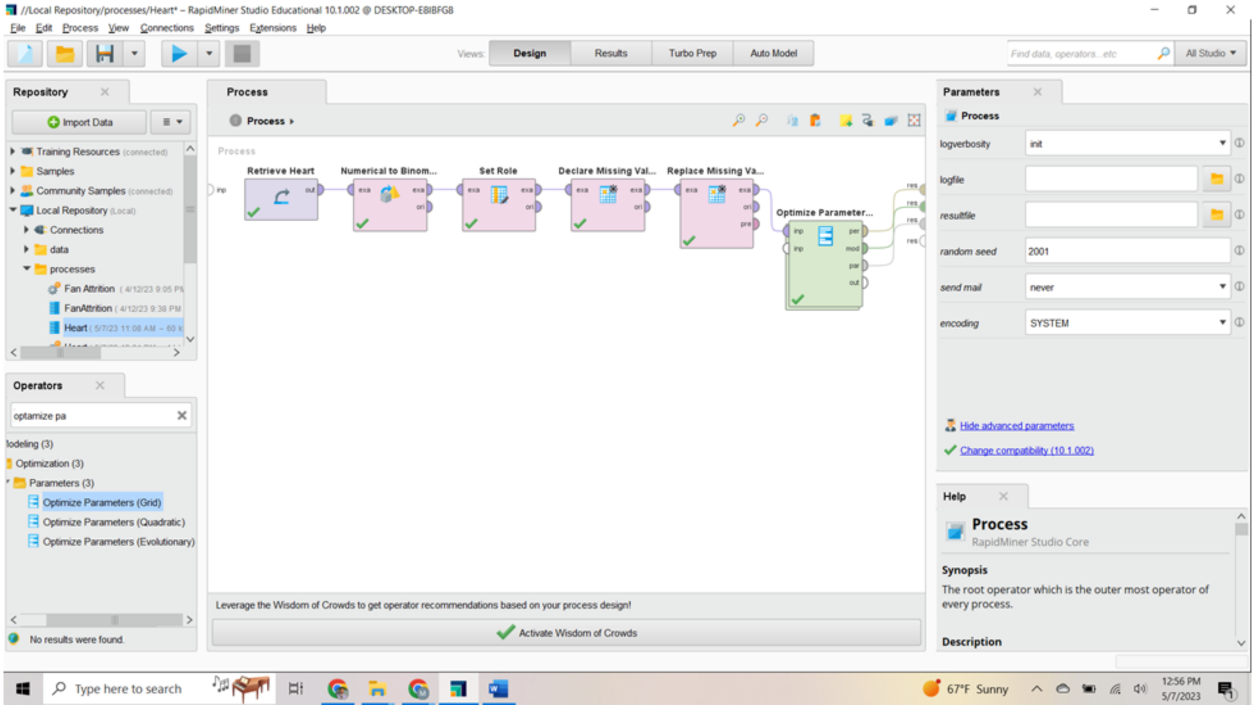
In order to increase the accuracy, Forward Selection was also included within Cross Validation. Within forward selection a second cross validation was added; within the new cross validation a second K-NN, Apply Model and Performance were added.


//Local Repository/processes/Heart* – RapidMiner Studio Educational 10.1.002 @ DESKTOP-E8IBFG8
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
KNNClassification (k-NN)
ExampleSet (//Local Repository/processes/Heart)
Result History
PerformanceVector (Performance)
ExampleSet (Replace Missing Values)
Performance
Description
Criterion
accuracy
precision
recall
AUC (optimistic)
AUC
AUC (pessimistic)
● Table View ○ Plot View
Repository

accuracy: 84.21% +/- 4.07% (micro average: 84.20%)

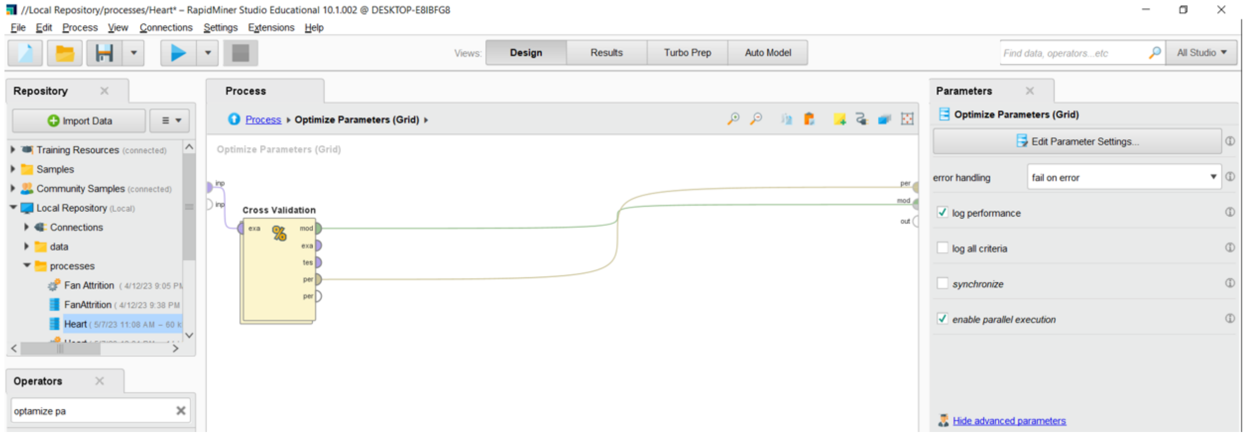
	true false	true true	class precision
pred. false	324	59	84.60%
pred. true	86	449	83.93%
class recall	79.02%	88.39%	

After running the updated current process the accuracy increases to 84.21%


//Local Repository/processes/Heart* – RapidMiner Studio Educational 10.1.002 @ DESKTOP-E8IBFG8
Design
Results
Turbo Prep
Auto Model
Repository
Training Resources
Samples
Community Samples
Local Repository
Connections
data
processes
Fan Attrition
FanAttrition
Heart
Process
Retrieve Heart
Numerical to Binom...
Set Role
Declare Missing Val...
Replace Missing Va...
Optimize Parameter...
Operators
optamize pa
Optimize Parameters (Grid)
Optimize Parameters (Quadratic)
Optimize Parameters (Evolutionary)
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Help
Process
RapidMiner Studio Core
Synopsis
The root operator which is the outer most operator of every process.
Description

Parameters — Process

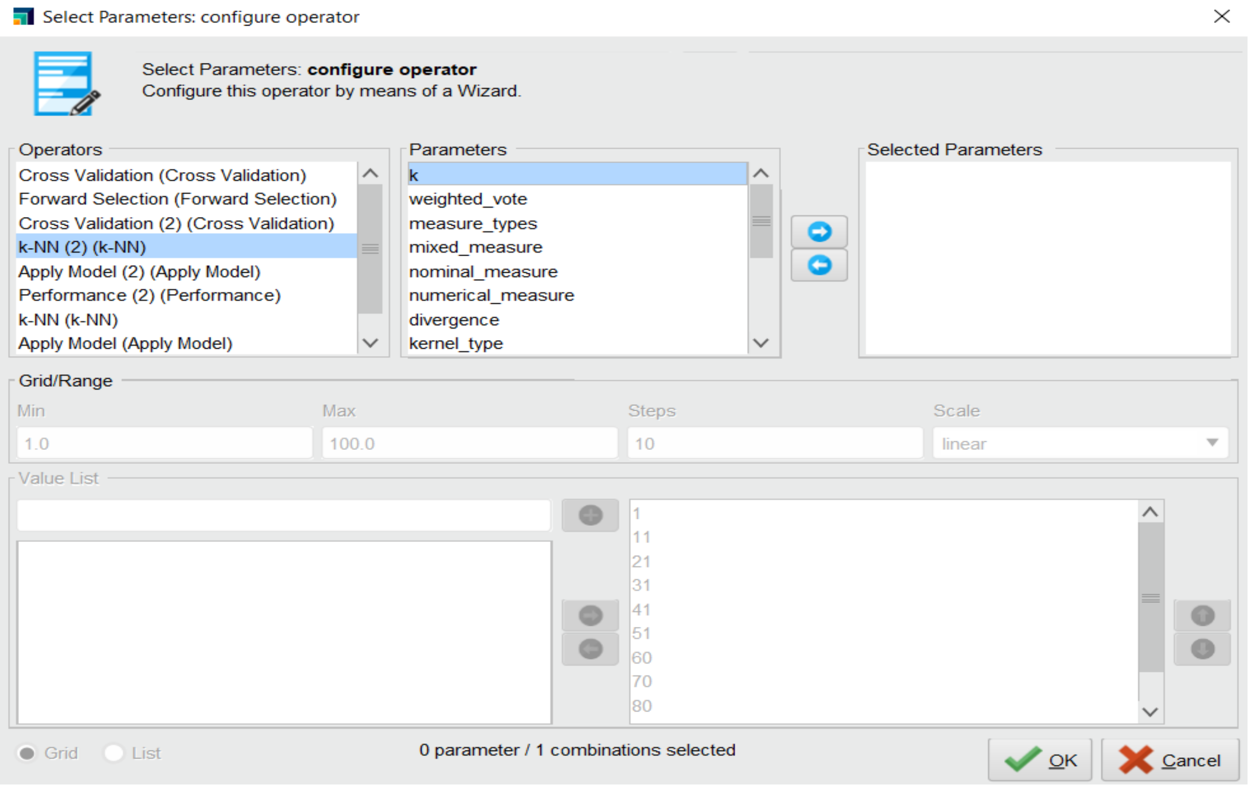
logverbosity	init
logfile	
resultfile	
random seed	2001
send mail	never
encoding	SYSTEM


//Local Repository/processes/Heart* – RapidMiner Studio Educational 10.1.002 @ DESKTOP-E8IBFG8
Process ▸ Optimize Parameters (Grid)
Cross Validation
Edit Parameter Settings...
Hide advanced parameters

Parameters — Optimize Parameters (Grid)

error handling	fail on error
☑ log performance	
□ log all criteria	
□ synchronize	
☑ enable parallel execution	

In order to further optimize the accuracy performance the operator "Optimize Parameters (Grid)" was added. The former cross validation was included in the initial process was cut and paste inside the new optimize parameter (grid).



Select Parameters: configure operator

Select Parameters: **configure operator**
Configure this operator by means of a Wizard.

Operators:
- Cross Validation (Cross Validation)
- Forward Selection (Forward Selection)
- Cross Validation (2) (Cross Validation)
- k-NN (2) (k-NN)
- Apply Model (2) (Apply Model)
- Performance (2) (Performance)
- k-NN (k-NN)
- Apply Model (Apply Model)

Parameters:
- k
- weighted_vote
- measure_types
- mixed_measure
- nominal_measure
- numerical_measure
- divergence
- kernel_type

Selected Parameters

Grid/Range

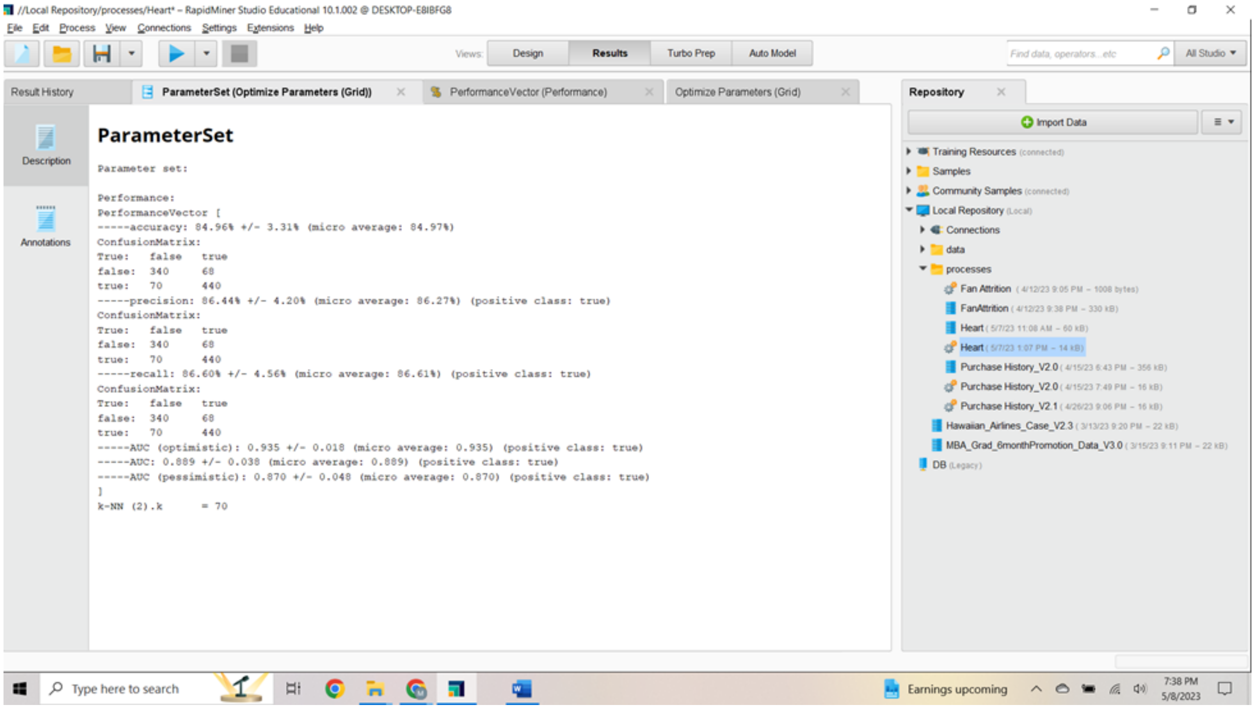
Min	Max	Steps	Scale
1.0	100.0	10	linear

Value List: 1, 11, 21, 31, 41, 51, 60, 70, 80

● Grid ○ List

0 parameter / 1 combinations selected

Under “Edit Parameter Settings”, operator “K-NN(2) (K-NN)” was chosen and under parameters “K” was chosen. as kept. Once these parameters are chosen we can go ahead and run the process.



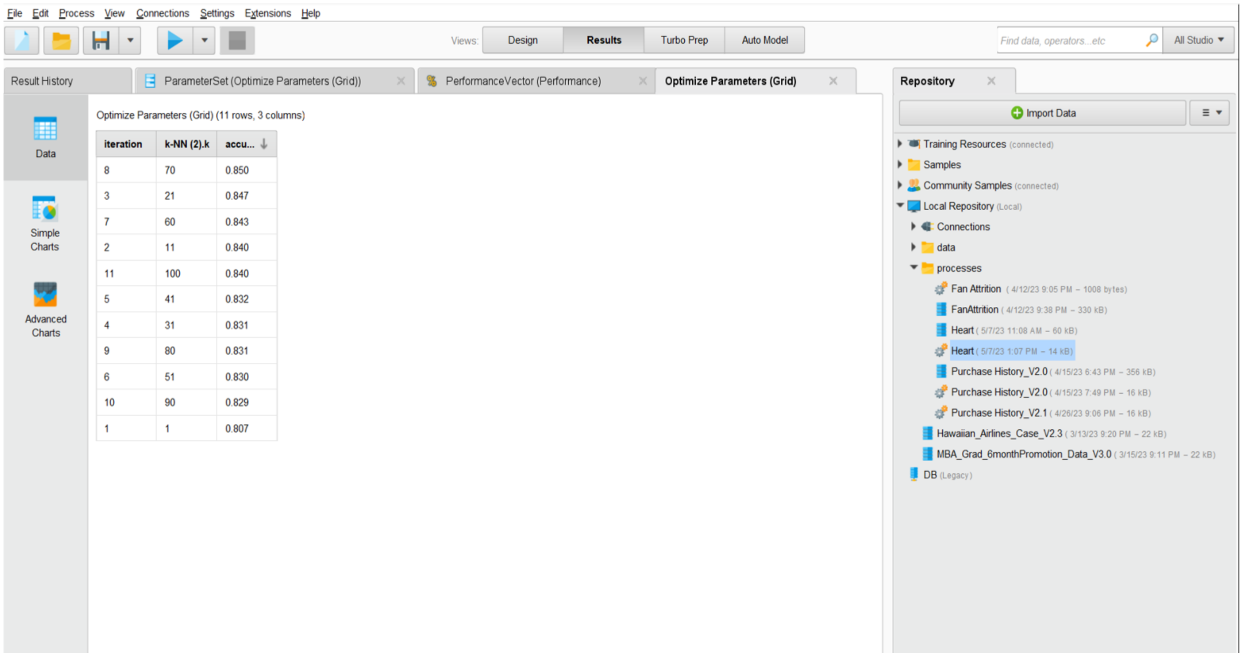
//Local Repository/processes/Heart* – RapidMiner Studio Educational 10.1.002 @ DESKTOP-E8IBFG8

File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Result History
ParameterSet (Optimize Parameters (Grid))
PerformanceVector (Performance)
Optimize Parameters (Grid)
Description
Annotations
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
processes
Fan Attrition (4/12/23 9:05 PM – 1008 bytes)
FanAttrition (4/12/23 9:38 PM – 330 kB)
Heart (5/7/23 11:08 AM – 60 kB)
Heart (5/7/23 1:07 PM – 14 kB)
Purchase History_V2.0 (4/15/23 6:43 PM – 356 kB)
Purchase History_V2.0 (4/15/23 7:49 PM – 16 kB)
Purchase History_V2.1 (4/26/23 9:06 PM – 16 kB)
Hawaiian_Airlines_Case_V2.3 (3/13/23 9:20 PM – 22 kB)
MBA_Grad_6monthPromotion_Data_V3.0 (3/15/23 9:11 PM – 22 kB)
DB (Legacy)
Type here to search
Earnings upcoming
7:38 PM
5/8/2023

ParameterSet

```
Parameter set:

Performance:
PerformanceVector [
-----accuracy: 84.96% +/- 3.31% (micro average: 84.97%)
ConfusionMatrix:
True:   false   true
false:  340     68
true:   70      440
-----precision: 86.44% +/- 4.20% (micro average: 86.27%) (positive class: true)
ConfusionMatrix:
True:   false   true
false:  340     68
true:   70      440
-----recall: 86.60% +/- 4.56% (micro average: 86.61%) (positive class: true)
ConfusionMatrix:
True:   false   true
false:  340     68
true:   70      440
-----AUC (optimistic): 0.935 +/- 0.018 (micro average: 0.935) (positive class: true)
-----AUC: 0.889 +/- 0.038 (micro average: 0.889) (positive class: true)
-----AUC (pessimistic): 0.870 +/- 0.048 (micro average: 0.870) (positive class: true)
]
k-NN (2).k      = 70
```

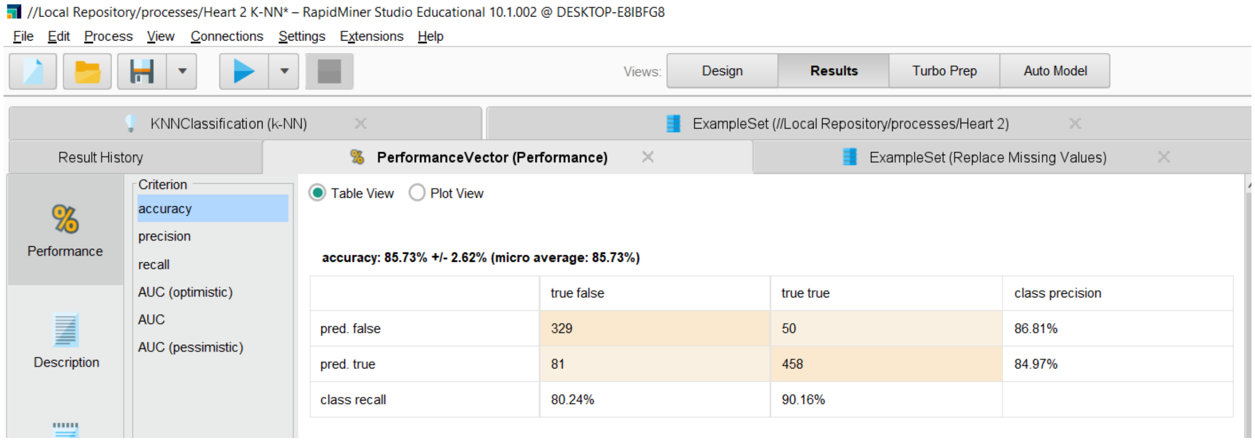

File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Result History
ParameterSet (Optimize Parameters (Grid))
PerformanceVector (Performance)
Optimize Parameters (Grid)
Data
Simple Charts
Advanced Charts
Repository
Import Data

Optimize Parameters (Grid) (11 rows, 3 columns)

iteration	k-NN (2).k	accu...
8	70	0.850
3	21	0.847
7	60	0.843
2	11	0.840
11	100	0.840
5	41	0.832
4	31	0.831
9	80	0.831
6	51	0.830
10	90	0.829
1	1	0.807

Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
processes
Fan Attrition (4/12/23 9:05 PM – 1008 bytes)
FanAttrition (4/12/23 9:38 PM – 330 kB)
Heart (5/7/23 11:08 AM – 60 kB)
Heart (5/7/23 1:07 PM – 14 kB)
Purchase History_V2.0 (4/15/23 6:43 PM – 356 kB)
Purchase History_V2.0 (4/15/23 7:49 PM – 16 kB)
Purchase History_V2.1 (4/26/23 9:06 PM – 16 kB)
Hawaiian_Airlines_Case_V2.3 (3/13/23 9:20 PM – 22 kB)
MBA_Grad_6monthPromotion_Data_V3.0 (3/15/23 9:11 PM – 22 kB)
DB (Legacy)

After running the process, the results of the parameter set and the optimize parameter grid both reflect that when K equals to 70 accuracy is at its highest. The accuracy level shows 85% accuracy.

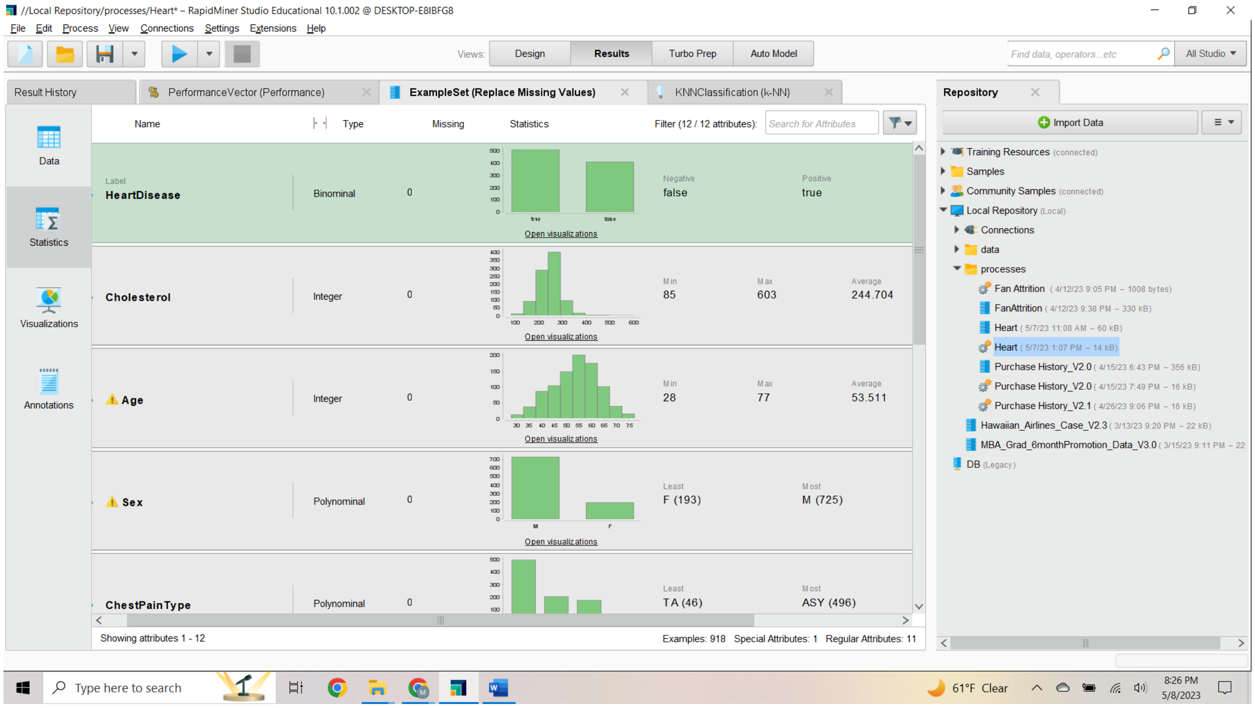

//Local Repository/processes/Heart 2 K-NN* – RapidMiner Studio Educational 10.1.002 @ DESKTOP-E8IBFG8
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
KNNClassification (k-NN)
ExampleSet (//Local Repository/processes/Heart 2)
Result History
PerformanceVector (Performance)
ExampleSet (Replace Missing Values)
Performance
Description
Criterion
accuracy
precision
recall
AUC (optimistic)
AUC
AUC (pessimistic)
● Table View ○ Plot View

accuracy: 85.73% +/- 2.62% (micro average: 85.73%)

	true false	true true	class precision
pred. false	329	50	86.81%
pred. true	81	458	84.97%
class recall	80.24%	90.16%	

For the final step we restored cross validation to the original process and on both K-NN we entered K at 70 and these are the results. Accuracy increase to now 85.73%

The confusion matrix depicts that the model predicted heart disease with a 85.73% prediction rate. The confusion matrix shows that 329 values were predicted to be false and were actually false, while 50 values that were predicted to be false were actually true making the precision for false prediction 86.81%. The precision of predicted true values is 84.97% with 81 predicted false values being true and 458 predicted value true values being true. Recall for false shows 329 predicted false values being true falses while 81 predicted true values were actually false, bringing the recall to a 80.24% sensitivity. For true recall the sensitivity was 90.16% with 50 predicted false being true and 458 predicted true actually being true.



//Local Repository/processes/Heart* – RapidMiner Studio Educational 10.1.002 @ DESKTOP-E8IBFG8

ExampleSet (Replace Missing Values)

Filter (12 / 12 attributes):

Name	Type	Missing	Statistics		
Label HeartDisease	Binominal	0	Negative false	Positive true	
Cholesterol	Integer	0	Min 85	Max 603	Average 244.704
Age	Integer	0	Min 28	Max 77	Average 53.511
Sex	Polynominal	0	Least F (193)	Most M (725)	
ChestPainType	Polynominal	0	Least TA (46)	Most ASY (496)	

Showing attributes 1 - 12

Examples: 918 Special Attributes: 1 Regular Attributes: 11

File
Edit
Process
View
Connections
Settings
Extensions
Help
Views:
Design
Results
Turbo Prep
Auto Model
Find data, operators...etc
All Studio
Result History
PerformanceVector (Performance)
KNNClassification (k-NN)
Data
Statistics
Visualizations
Annotations
Search for Attributes
Open visualizations
true
false
M
F
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
processes
Fan Attrition (4/12/23 9:05 PM – 1008 bytes)
FanAttrition (4/12/23 9:38 PM – 330 kB)
Heart (5/7/23 11:08 AM – 60 kB)
Heart (5/7/23 1:07 PM – 14 kB)
Purchase History_V2.0 (4/15/23 6:43 PM – 356 kB)
Purchase History_V2.0 (4/15/23 7:49 PM – 16 kB)
Purchase History_V2.1 (4/26/23 9:06 PM – 16 kB)
Hawaiian_Airlines_Case_V2.3 (3/13/23 9:20 PM – 22 kB)
MBA_Grad_6monthPromotion_Data_V3.0 (3/15/23 9:11 PM – 22
DB (Legacy)
Type here to search
61°F Clear
8:26 PM
5/8/2023

Based on the results averages or values of K-NN these are the characteristics that most likely put you at risk of having heart failure.

1) Cholesterol : at an average of 245.
2) Age: at an average of 54 (mid 50's).
3) Sex: Male (Out out 918 in the study 725 were male)
4) Chest Pain: Asymptomatic (Surprisingly out of 918 in the study the majority 496 were from people who were asymptomatic compared to those who had atypical or typical angina pain).
5) Resting Blood Pressure: at an average of 132.
6) Fasting BS: Fasting (Out out 918 in the study about 700 showed heart failure while fasting)
7) Resting Electrocardiogram Results: (Out out 918 in the study 552 showed normal levels in their electrocardiogram results)
8) Max Heart Rate achieved: at an average of 137
9) Exercise Angina: (Out out 918 in the study 547 were from people who did not exercise)
10) Old Peak: This measures the level of depression. Old peak shows an average of 0.887. The average is closer to 1 which would reflect that you are more likely to have heart failure if you show higher levels of depression.
11) ST Slope: (Out out 918 in the study 460 were from people whose ST Slope was flat)

Summary of Analysis Results and Research Purpose (James)

The goal of the data analysis is to advance the knowledge of the causes of Heart Disease in individuals. With our analysis we can deeper understand the data and gain better insight into how we should research this matter in the future. The most important knowledge gained from our data analysis is how we can detect the probability of heart disease early on with the use of machine learning tools. This can help reduce the number of fatalities caused by heart disease (the #1 cause of death in the world) by getting individuals proper treatment before it is too late. Our machine learning can be useful to doctors to focus on attributes like exercise angina, or asymptomatic chest pain when screening individuals for heart disease. Through the decision tree processes it is concluded that looking at an individual's cholesterol along with max heart rate is a strong and accurate predictor of heart disease, Exercise Angina and oldpeak are also very strong predictors. Overall, our research shows the importance of data quality, which is crucial in making accurate predictions. Improving data will in turn improve the accuracy of predictions and lead further advancements in business or research goals.

Work Cited

Fedesoriano. “Heart Failure Prediction Dataset.” *Kaggle*, 10 Sept. 2021, www.kaggle.com/datasets/fedesoriano/heart-failure-prediction.