2023-2024

Topic- Heart Disease Prediction

KNIME analytics

Team Members Name – Bhagyashri Pradhan

Roll no.- 21csu313

Introduction

It seems like you've provided a dataset related to heart health. The dataset likely includes various attributes for individuals such as age, sex, chest pain type (cp), resting blood pressure (trestbps), cholesterol level (Chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved (thalach), exercise induced angina (exang), depression induced by exercise relative to rest (oldpeak), slope of the peak exercise ST segment (slope), number of major vessels colored by fluoroscopy (ca), thalassemia type (thal), and a target variable (target) indicating whether the individual has heart disease (1) or not (0).

If you have specific questions about this dataset or if you need assistance with analysis, interpretation, or any other related tasks, please provide more details or ask specific questions, and I'll do my best to help.

Machine learning models

Regression models

1.Linear regression learner –

The primary objective in using linear regression models for heart disease prediction is to develop a predictive model that can accurately estimate the risk of heart disease based on relevant features. The goal is to create a tool that aids in identifying individuals who are more likely to develop heart disease, allowing for early intervention and preventive measures
.

2.decision tree learner-

The objective of using a decision tree learner is to create a predictive model that can accurately classify or predict outcomes based on input features. Decision trees are a machine learning algorithm that recursively partitions the data into subsets, making decisions at each node based on the values of input features.

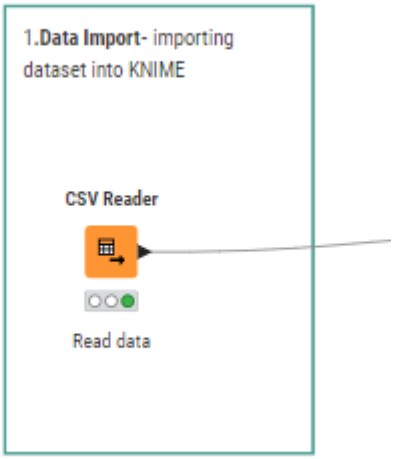
Classification models

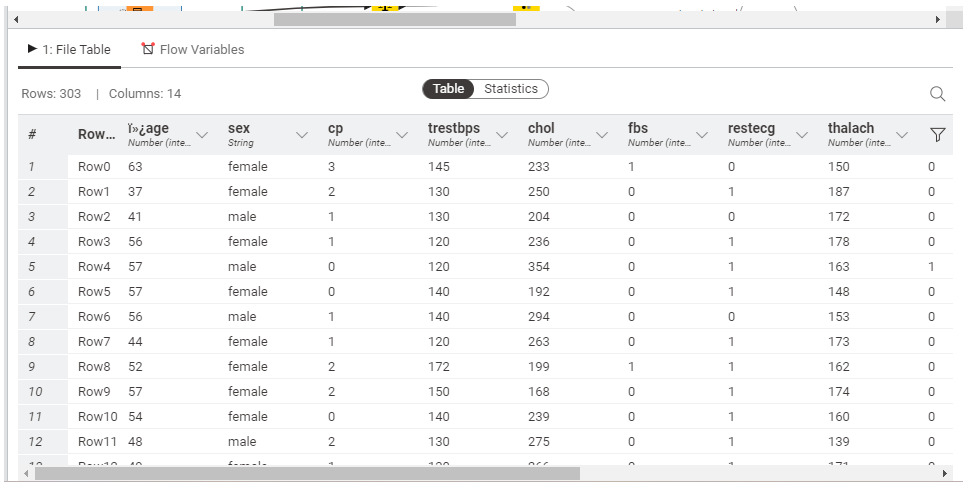
3.Random forest classification

The objective of using a Random Forest learner is to build an ensemble model that combines the predictions of multiple decision trees to improve overall predictive accuracy and generalization. Random Forest is a machine learning algorithm that operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Screenshots

1Import data - importing csv file into knime-he dataset likely includes various attributes for individuals such as age, sex, chest pain type (cp), resting blood pressure (trestbps), cholesterol level (Chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved (thalach), exercise induced angina (exang), depression induced by exercise relative to rest (oldpeak), slope of the peak exercise ST segment (slope), number of major vessels colored by fluoroscopy (ca), thalassemia type (thal), and a target variable (target) indicating whether the individual has heart disease (1) or not (0).





1: File Table | Flow Variables

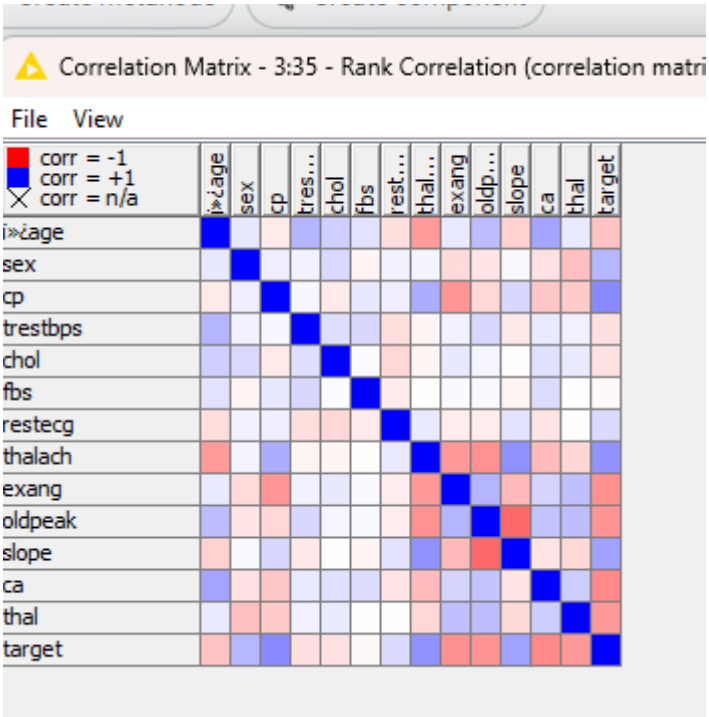
Rows: 303 | Columns: 14 — Table | Statistics

#	Row...	ï»¿age Number (inte...	sex String	cp Number (inte...	trestbps Number (inte...	chol Number (inte...	fbs Number (inte...	restecg Number (inte...	thalach Number (inte...	
1	Row0	63	female	3	145	233	1	0	150	0
2	Row1	37	female	2	130	250	0	1	187	0
3	Row2	41	male	1	130	204	0	0	172	0
4	Row3	56	female	1	120	236	0	1	178	0
5	Row4	57	male	0	120	354	0	1	163	1
6	Row5	57	female	0	140	192	0	1	148	0
7	Row6	56	male	1	140	294	0	0	153	0
8	Row7	44	female	1	120	263	0	1	173	0
9	Row8	52	female	2	172	199	1	1	162	0
10	Row9	57	female	2	150	168	0	1	174	0
11	Row10	54	female	0	140	239	0	1	160	0
12	Row11	48	male	2	130	275	0	1	139	0

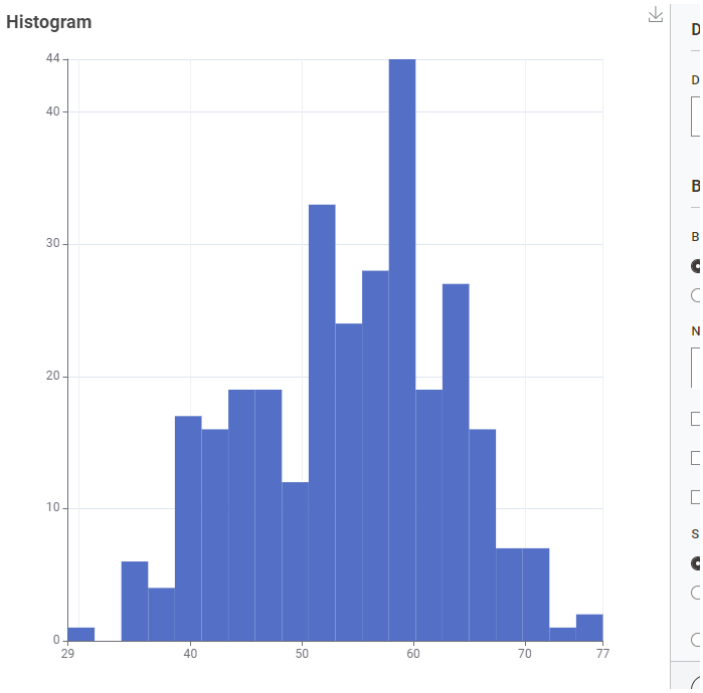
2.Preprocessing –In this process, we are handling data my removing missing value, duplicacy and representing data in more effective way like heat map, histogram etc

Nodes used –

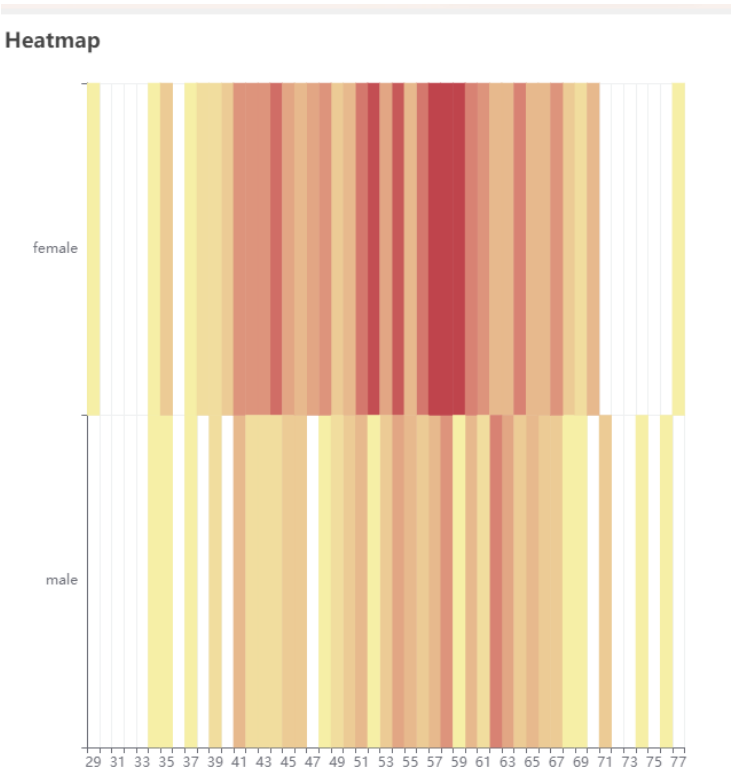
1. duplicate row filter- Either the input data without duplicates or the input data with additional columns identifying duplicates.
2. column filter -Table excluding selected columns.
3. missing value-Table with replaced missing values
4. rank correlation-A table containing the fractional ranks of the columns. Where the rank corresponds to the values position in a sorted table.



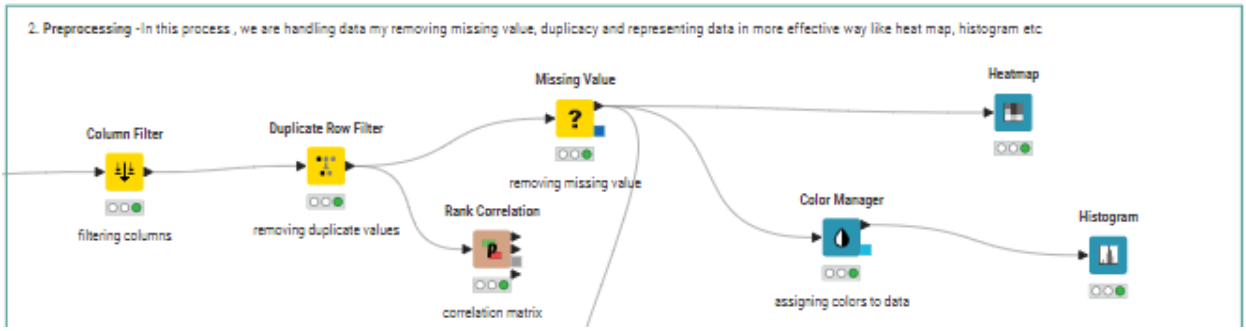
5. Histogram-Data table containing the values to be plotted in a histogram.



6. Color Model-Color model as applied to the input table (if applicable)
7. Heat map- This includes the ability to choose different aggregation methods and the column with the data to color map or the possibility to set a title.

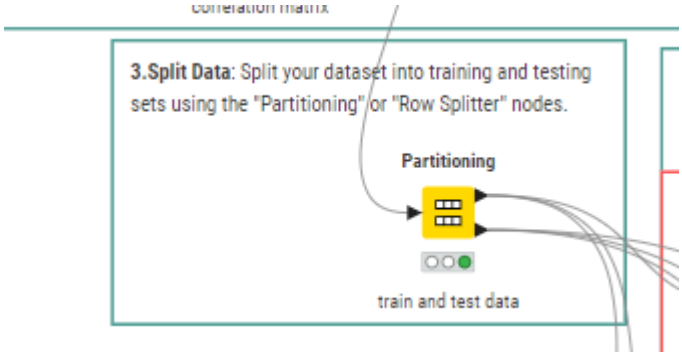


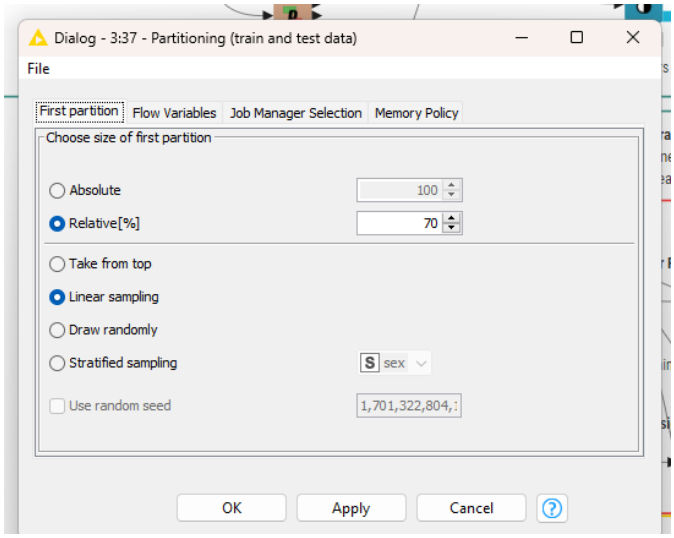
Knime overview -



3.Spliting data-

using partitioning node to split data into train and test



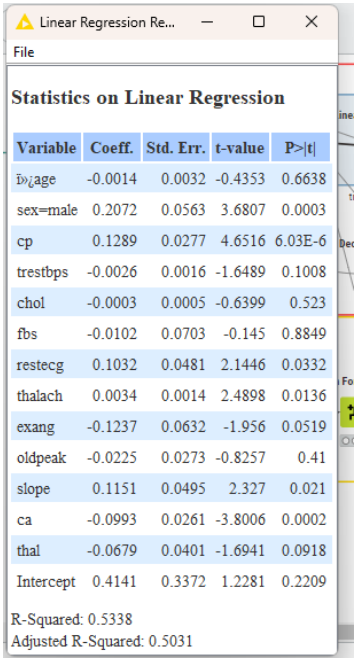


4. Model training

Using linear regression , decision tree and random forest learner to predict the "heart disease" target variable based on selected features

Nodes used – 1.linear regression learner

Regression predictor



Linear Regression Re...

File

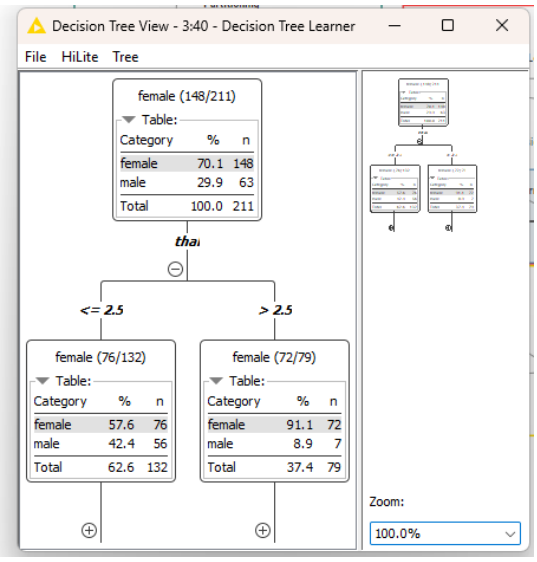
Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
ï»¿age	-0.0014	0.0032	-0.4353	0.6638
sex=male	0.2072	0.0563	3.6807	0.0003
cp	0.1289	0.0277	4.6516	6.03E-6
trestbps	-0.0026	0.0016	-1.6489	0.1008
chol	-0.0003	0.0005	-0.6399	0.523
fbs	-0.0102	0.0703	-0.145	0.8849
restecg	0.1032	0.0481	2.1446	0.0332
thalach	0.0034	0.0014	2.4898	0.0136
exang	-0.1237	0.0632	-1.956	0.0519
oldpeak	-0.0225	0.0273	-0.8257	0.41
slope	0.1151	0.0495	2.327	0.021
ca	-0.0993	0.0261	-3.8006	0.0002
thal	-0.0679	0.0401	-1.6941	0.0918
Intercept	0.4141	0.3372	1.2281	0.2209

R-Squared: 0.5338
Adjusted R-Squared: 0.5031

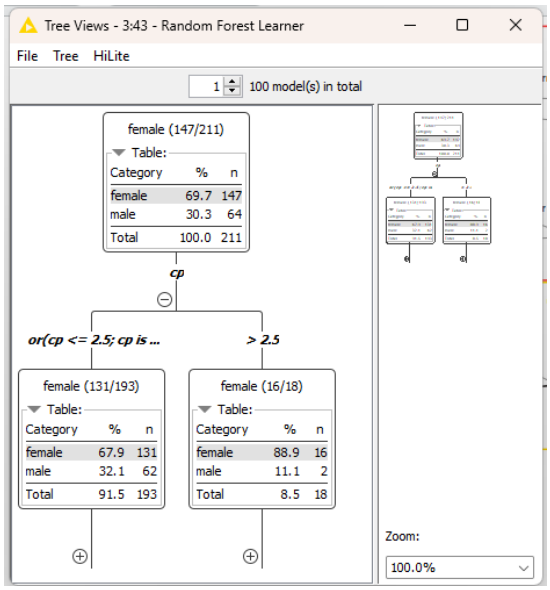
2. decision tree learner

Decision tree predictor

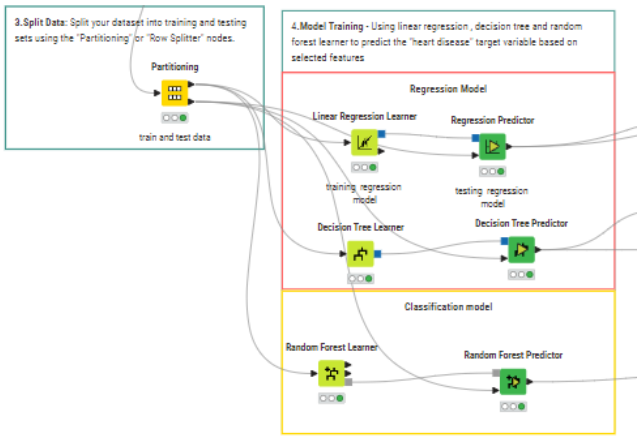


3. random forest learner

Random tree predictor



Knime Overview

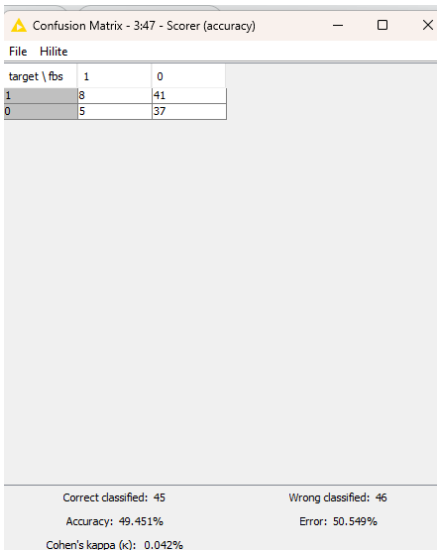


5.Model evaluation and prediction

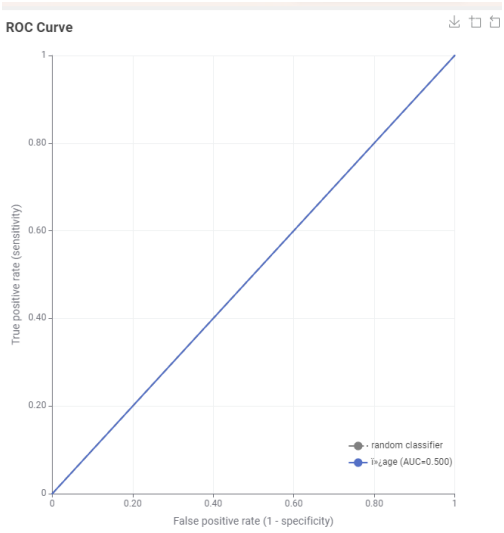
- Evaluating model using scorer and roc curve node to see the correlation matrix and accuracy

NODES USED

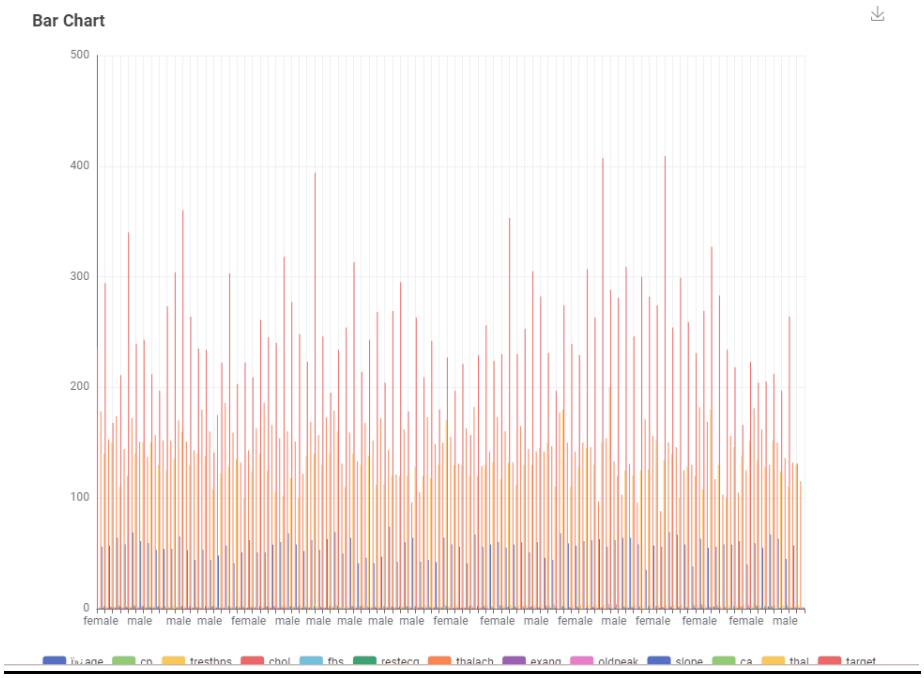
1.Scorer -The output of the node is the confusion matrix with the number of matches in each cell. Additionally, the second out-port reports a number of accuracy statistics such as True-Positives, False-Positives, True-Negatives, False-Negatives, Recall, Precision, Sensitivity, Specificity, F-measure, as well as the overall accuracy



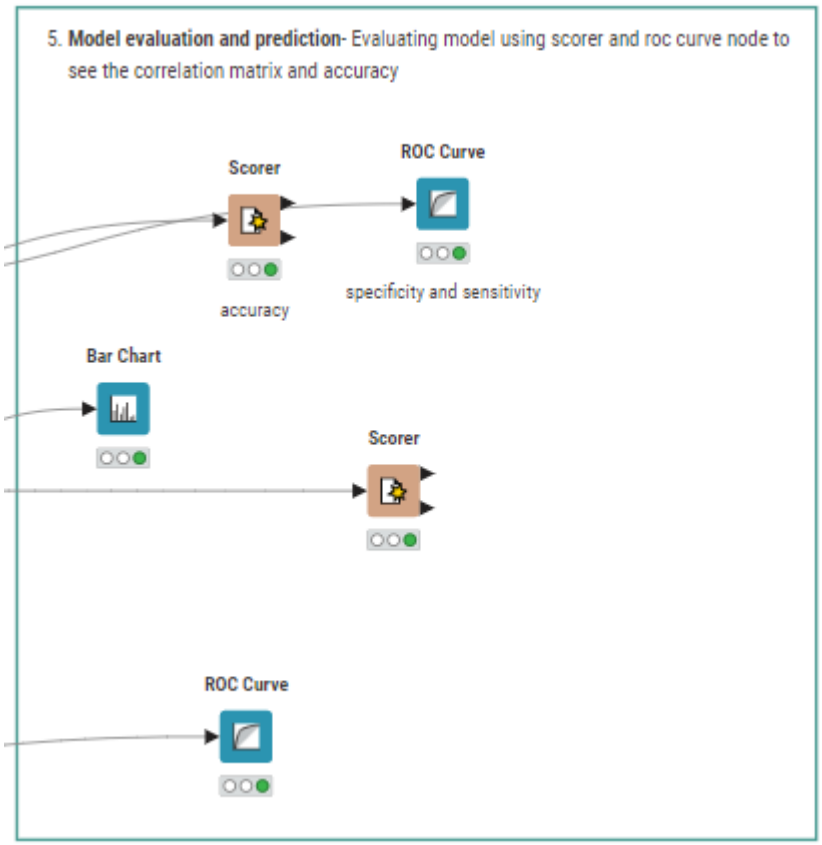
2. ROC curve-The configuration also offers a preview of the view, which should help to get the ROC curve in the desired shape quickly.



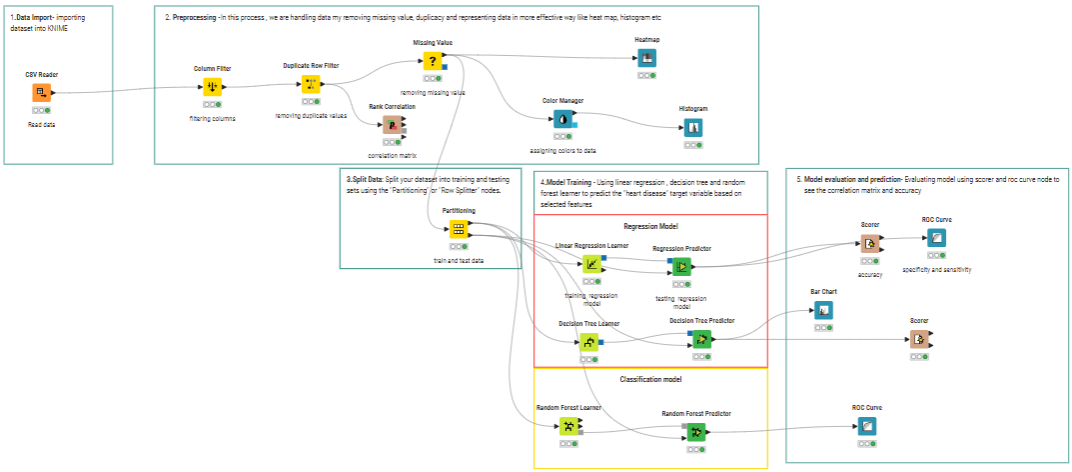
3. Bar chart -The configuration also offers a preview of the view, which should help to get the bar chart in the desired shape quickly.



Knime overview

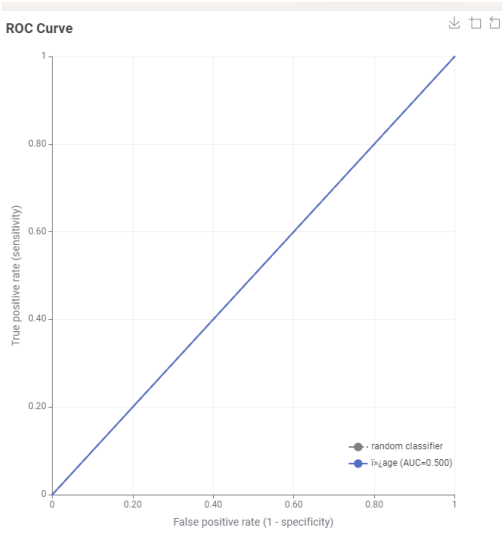


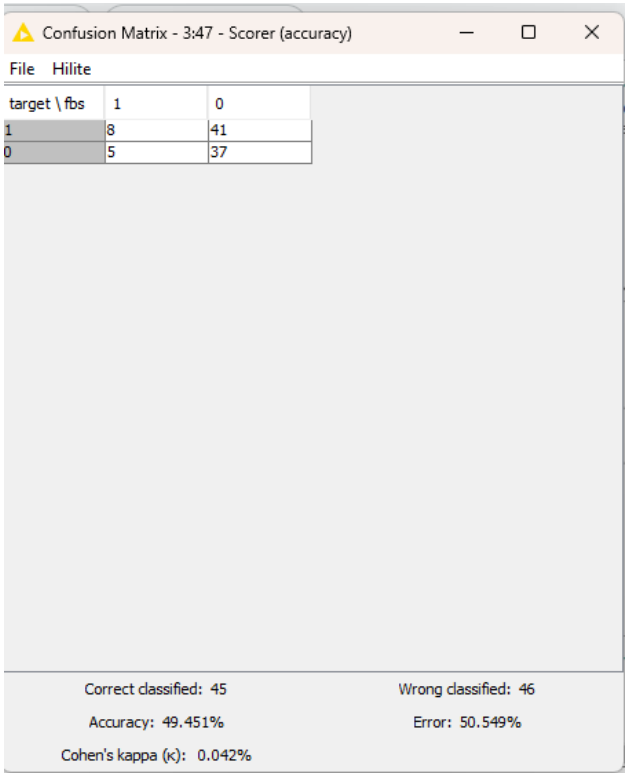
KNIME PROJECT OVERVIEW



Observation

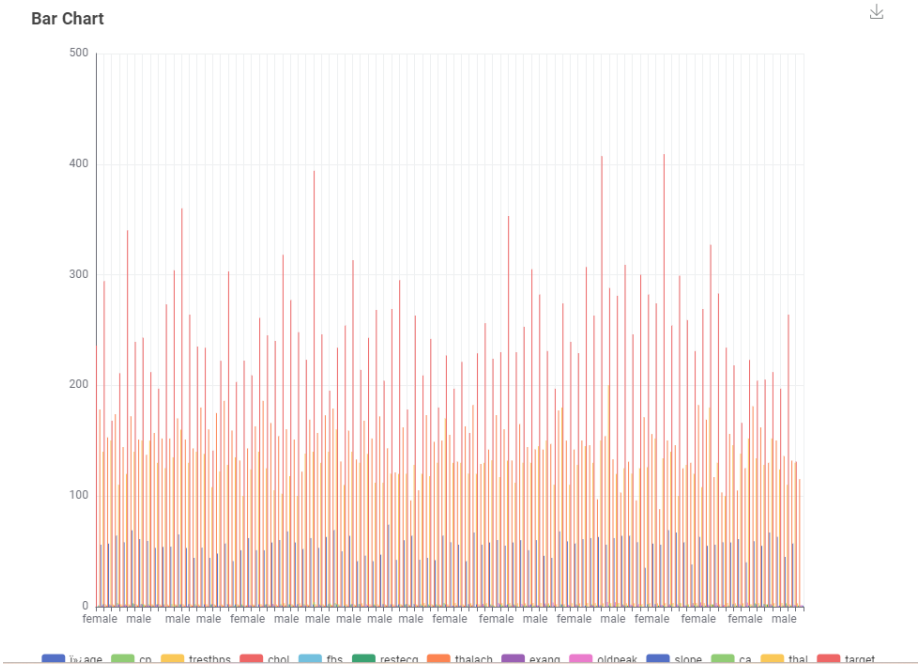
1. linear regression



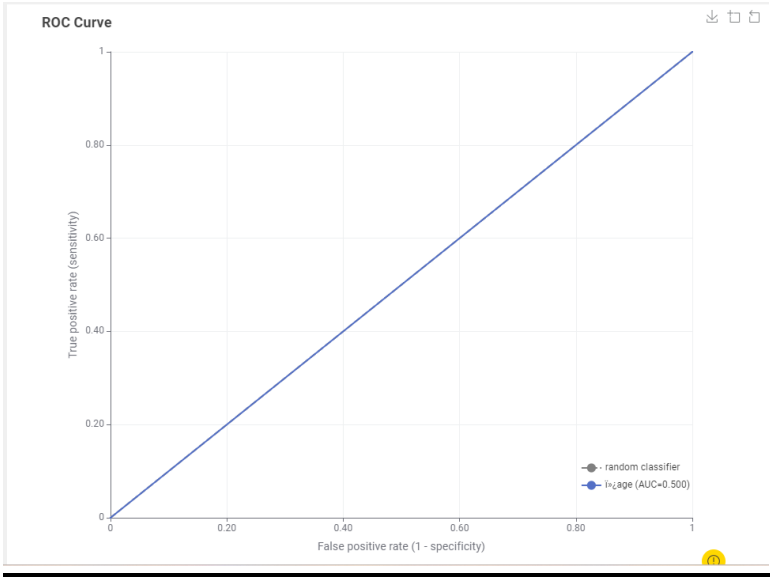


target \ fbs	1	0
1	8	41
0	5	37

2.decison tree



3.random forest classifier



Reference

1. Kaggle
2. Google chrome
3. YouTube
4. Canvas
5. Knime examples