Knime Project

By:

Sakshi (21CSU419)

Under the supervision of

Ms. Poonam Chaudhary

Department of Computer Science and Engineering

School of Engineering and Technology

The NorthCap University

HUDA, Sec-23A, Gurugram, Haryana - 122017

INDEX

Project Description

Build a machine learning model to predict whether a patient will survive the diagnosed liver cancer. • Responsible for processing the big data and applied data mining algorithms to build predictive classification models.

Dataset Description

- Contains data of 165 real patients diagnosed with HCC
- 49 features + 1 class attribute.
- Nominal features: 23
- Ordinal Features: 3
- Continuous features: 23

Problem Statement

Predict whether a patient will survive the diagnosed liver cancer.

CLASS 0: patient does not survive

CLASS 1: patient survives

Analysis

Hardware Requirements

A 64-bit Operating system with at least 34GB RAM and 8 CPU cores as minimum.

Software Requirements

Knime Analytics Platform

DESIGN

DATA UNDERSTANDING

- Converting Nominal and Ordinal Features to Strings.
- Out of 165, only 8 patients have complete information.
- Observed class Imbalance-Data skewed towards class 1.
- Positive Correlation between Total and Direct Bilirubin(mg/dL)
- Outliers detected using Box Plots.

DATA PREPARATION

- Missing Values were replaced.
- Correlations were filtered using a threshold of 0.8.
- Treated Numerical Outliers with Closest Permitted Values.

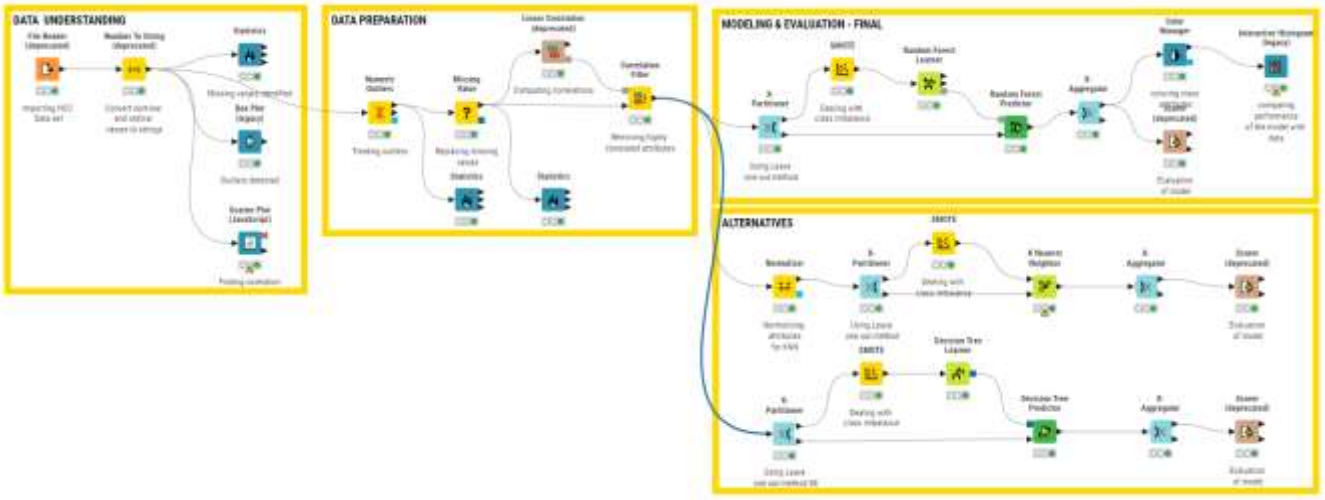
MODELING

- Class imbalance was fixed using SMOTE node (minority class 0 oversampled).
- X-partitioner with Leave one-out was used to generate Training and Test Data.
- Classification Models Tested:
    - Random forest
    - Decision Tree
    - KNN

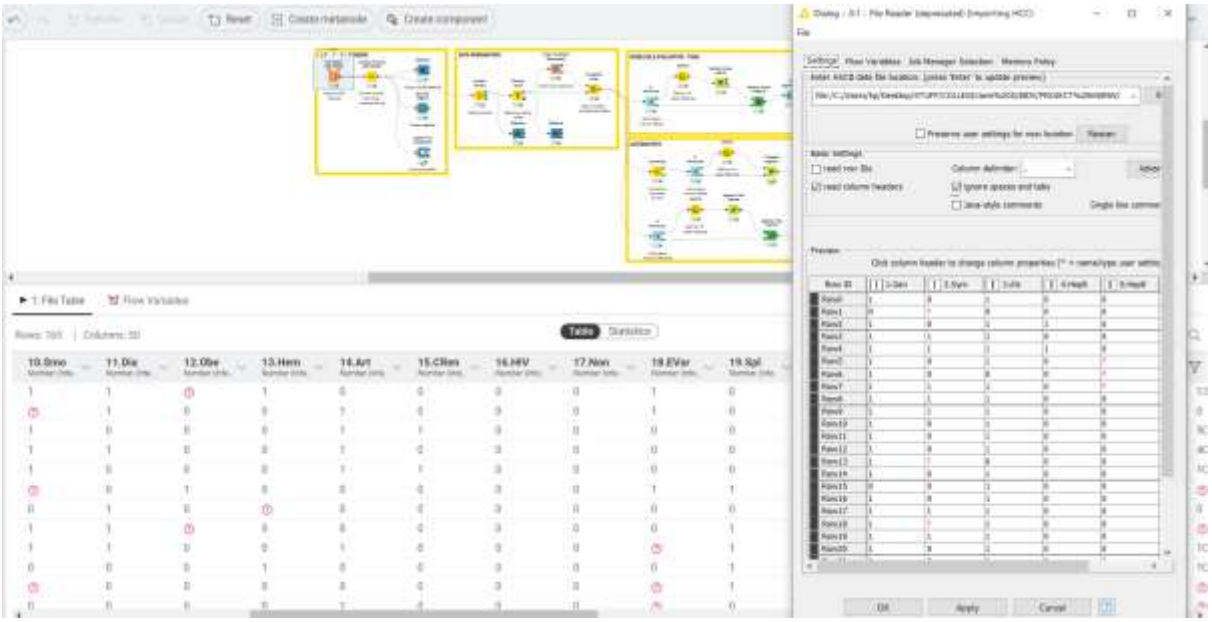
EVALUATION

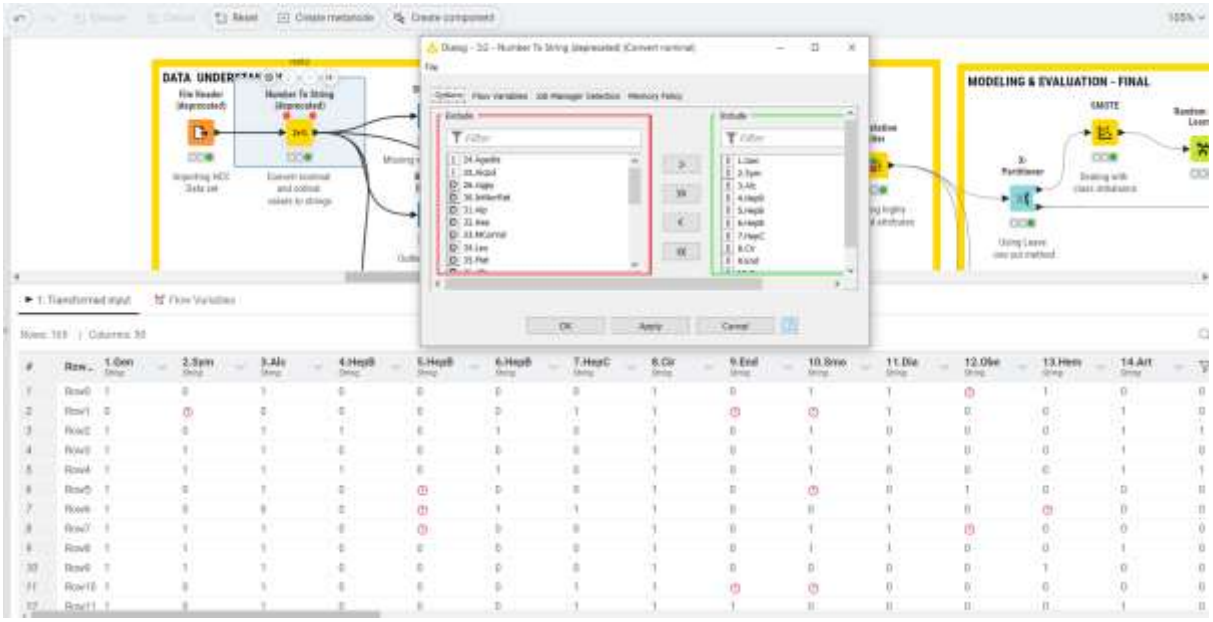
- Comparison of Models based on Minimization of False Positives and better accuracy.
- Random Forests excels in both Accuracy and Number of False Positive Predictions.

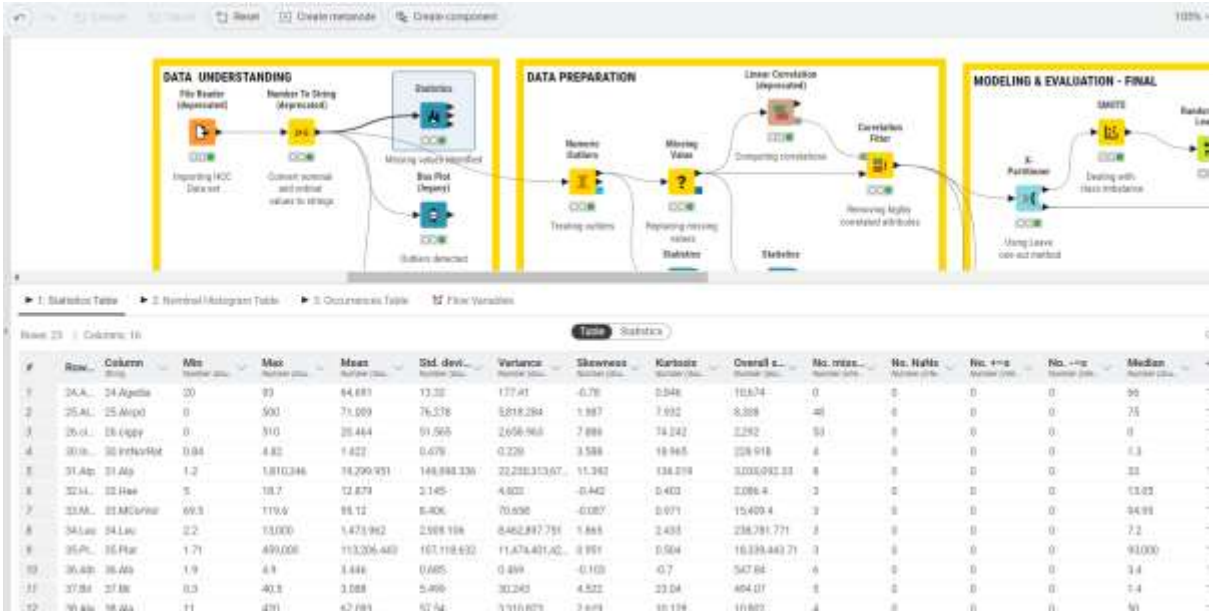
KNIME WORKFLOW

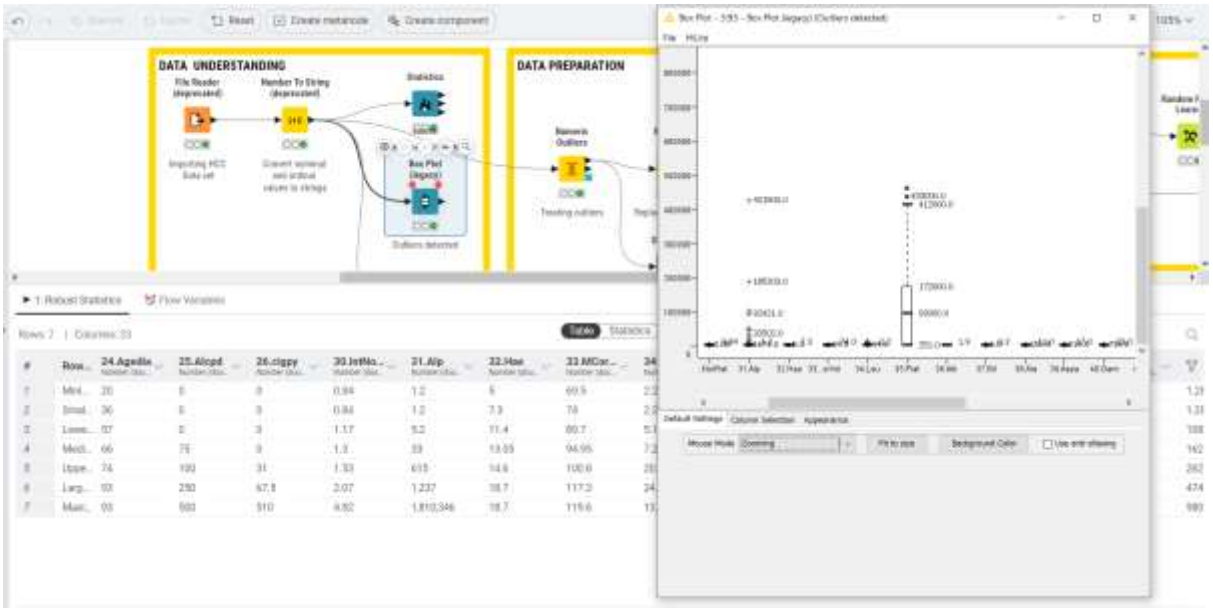


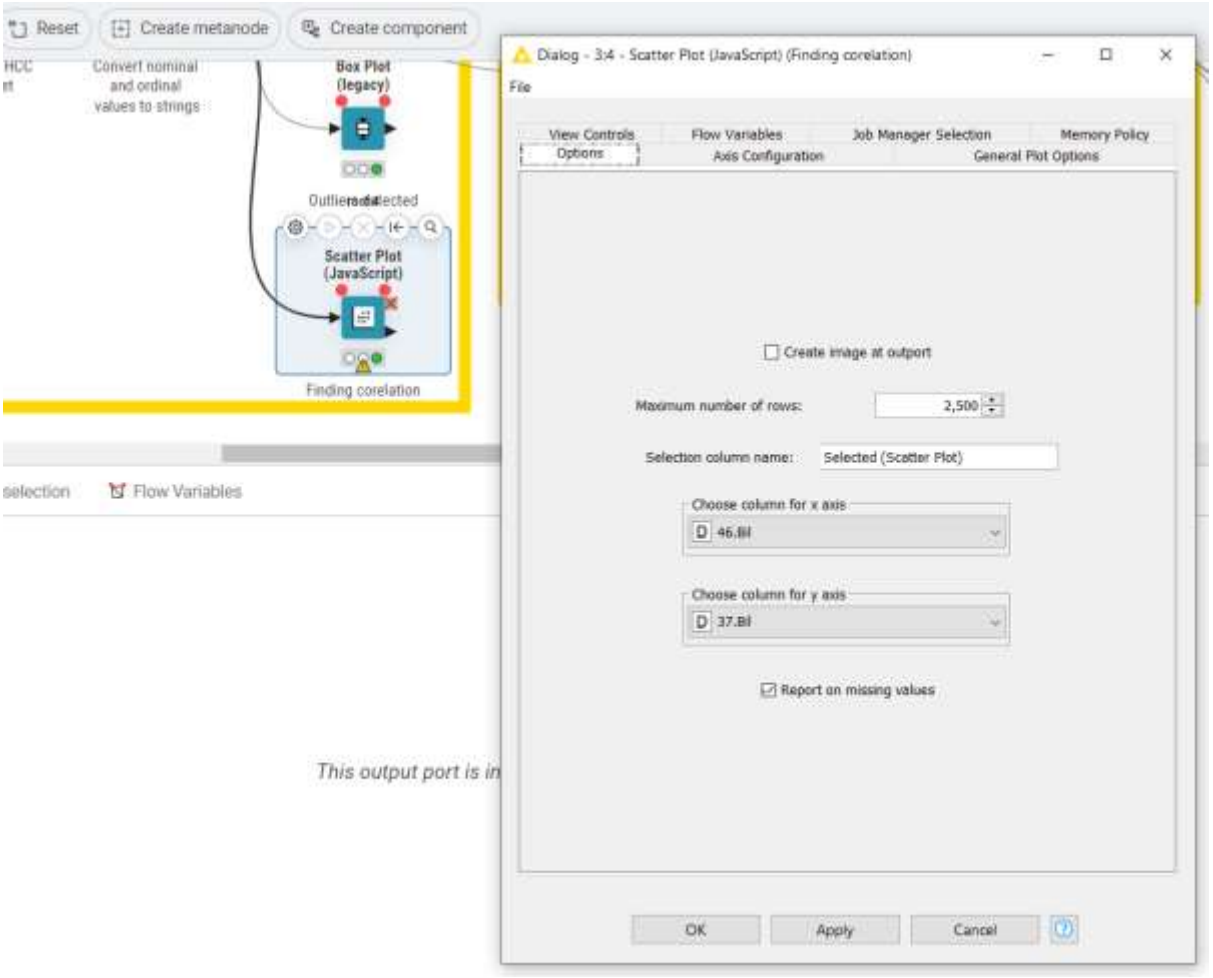
OUTPUT

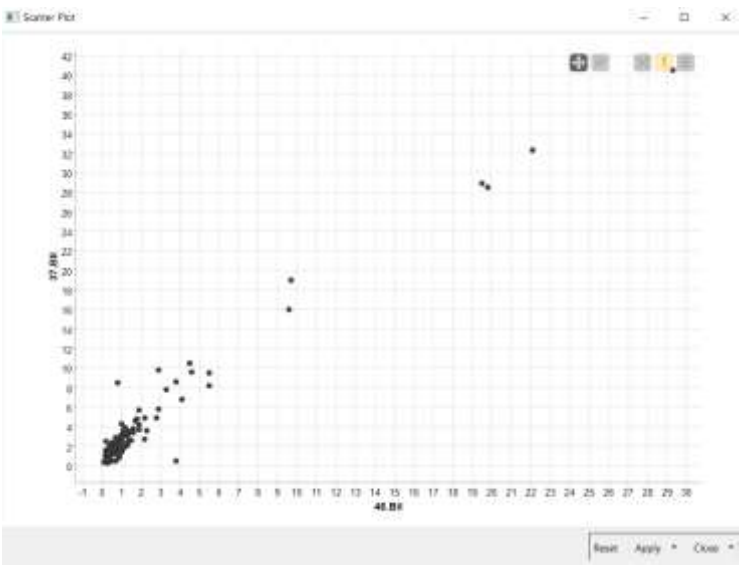


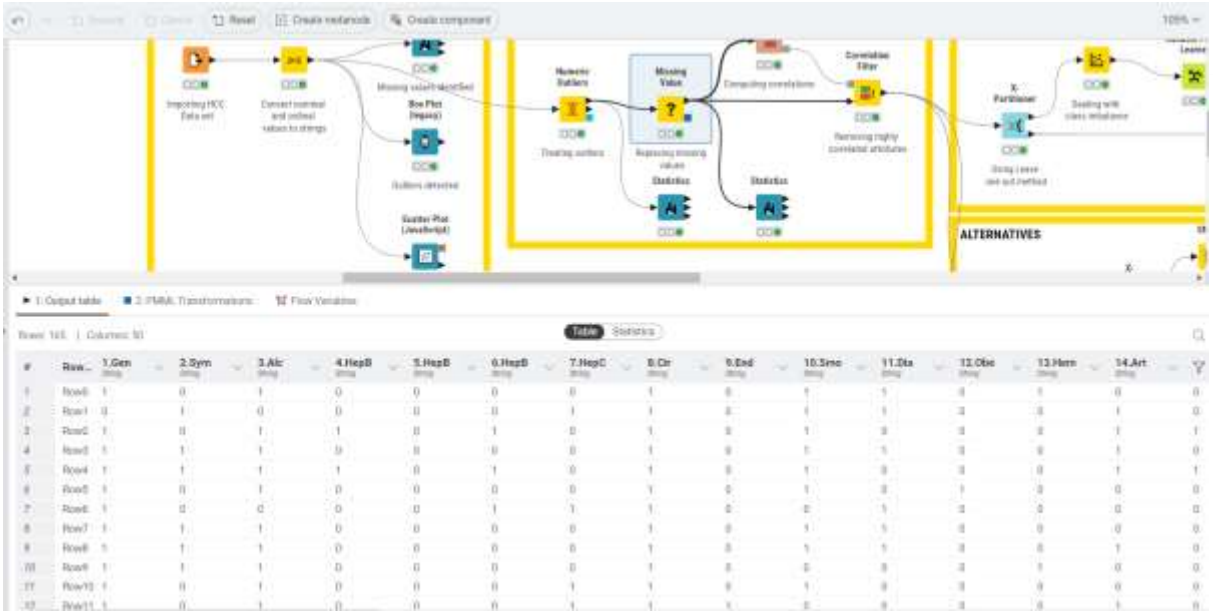



DATA UNDERSTANDING
DATA PREPARATION
MODELING & EVALUATION - FINAL

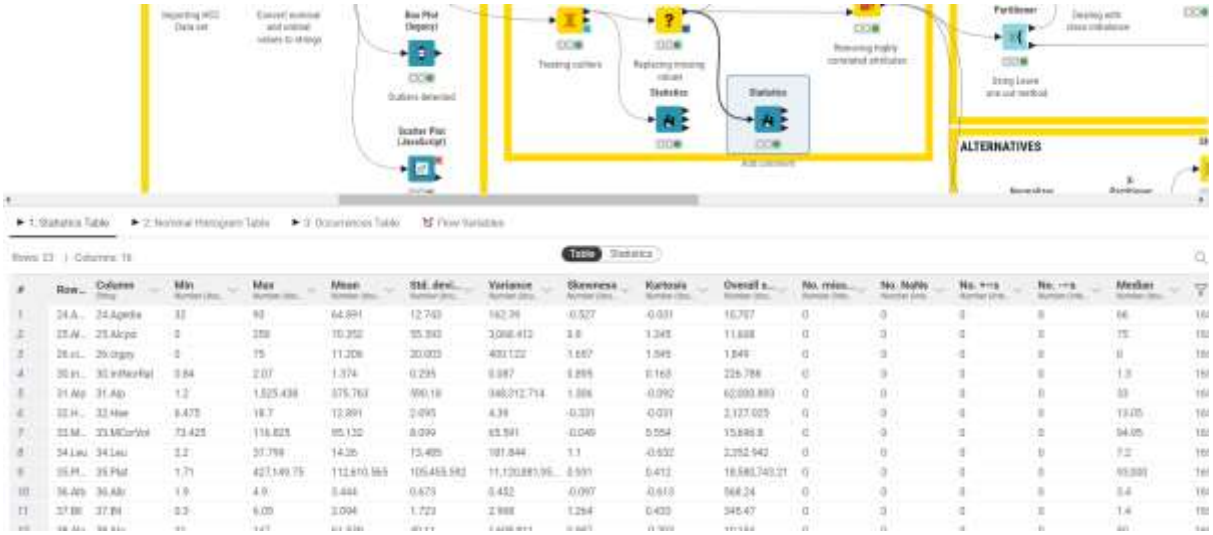

DATA UNDERSTANDING
DATA PREPARATION

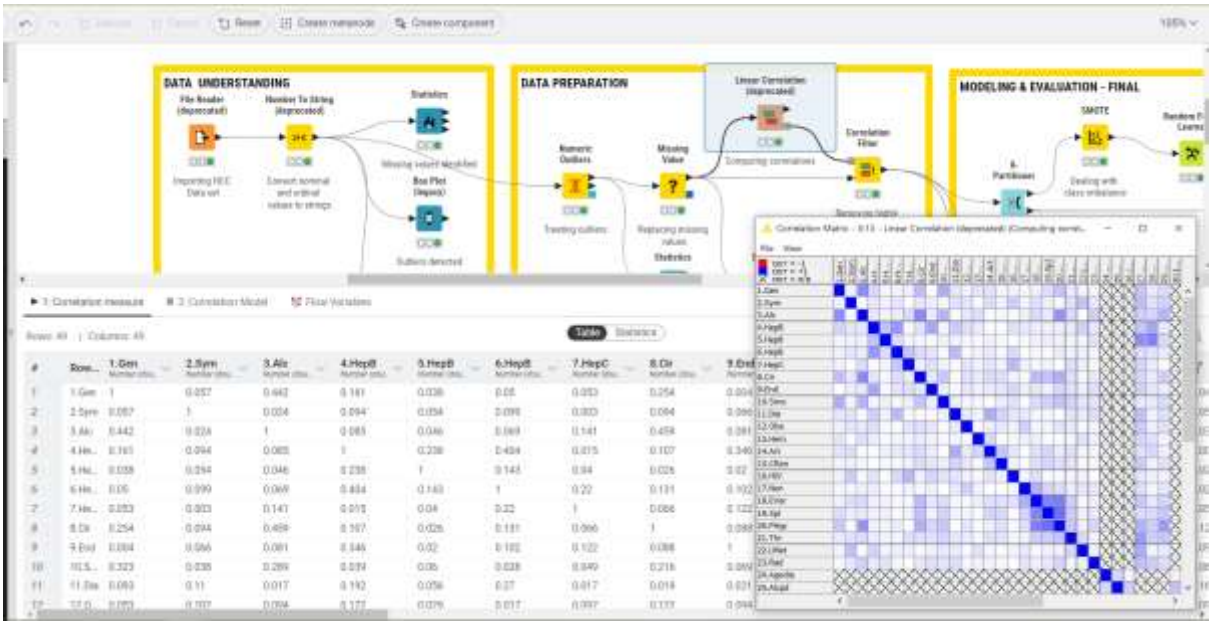

Reset
Create metanode
Create component
Convert nominal and ordinal values to strings
Box Plot (legacy)
Scatter Plot (JavaScript)
Finding corelation
Dialog - 3:4 - Scatter Plot (JavaScript) (Finding corelation)
File
View Controls
Options
Flow Variables
Axis Configuration
Job Manager Selection
General Plot Options
Memory Policy
Create image at outport
Maximum number of rows:
2,500
Selection column name:
Selected (Scatter Plot)
Choose column for x axis
46.Bl
Choose column for y axis
37.Bl
Report on missing values
OK
Apply
Cancel
Flow Variables

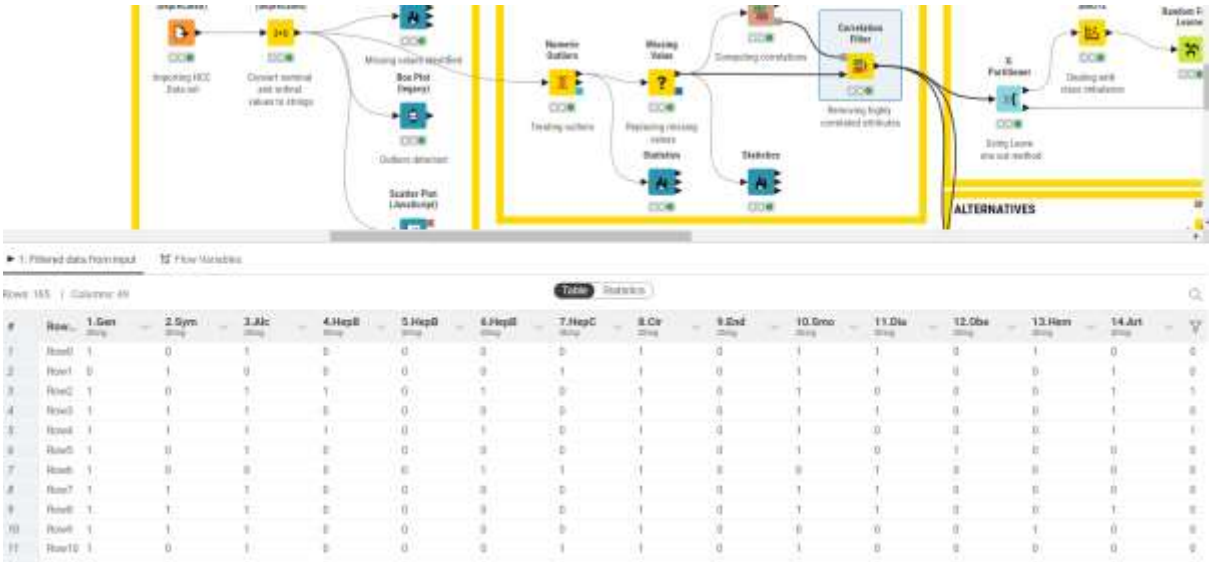


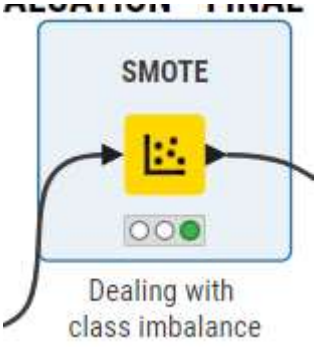

Reset
Create metanode
Create component
Missing value rule/ruleSet
Box Plot (legacy)
Outliers detected
Scatter Plot (JavaScript)
Numeric Outliers
Missing Value
Correlation Filter
Removing highly correlated attributes
Statistics
X-Partitioner
Using Leave-one-out method
ALTERNATIVES
1: Output table
2: PMML Transformations
Flow Variables
Table
Statistics

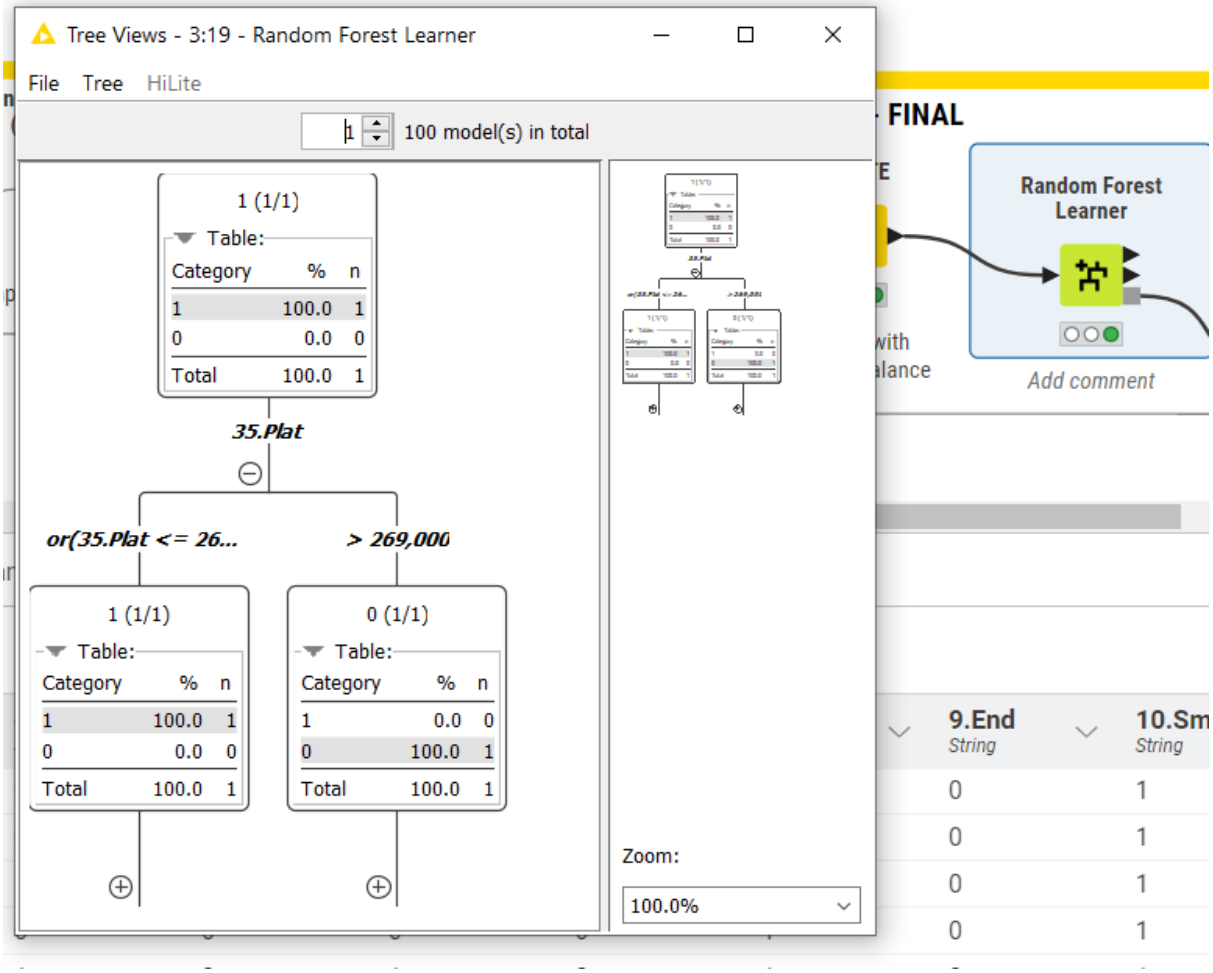

Outliers detected
Scatter Plot (JavaScript)
Statistics
Removing highly correlated attributes
Using Leave-one-out method
ALTERNATIVES
1: Statistics Table
2: Numeric Histogram Table
3: Occurrences Table
Flow Variables
Table
Statistics

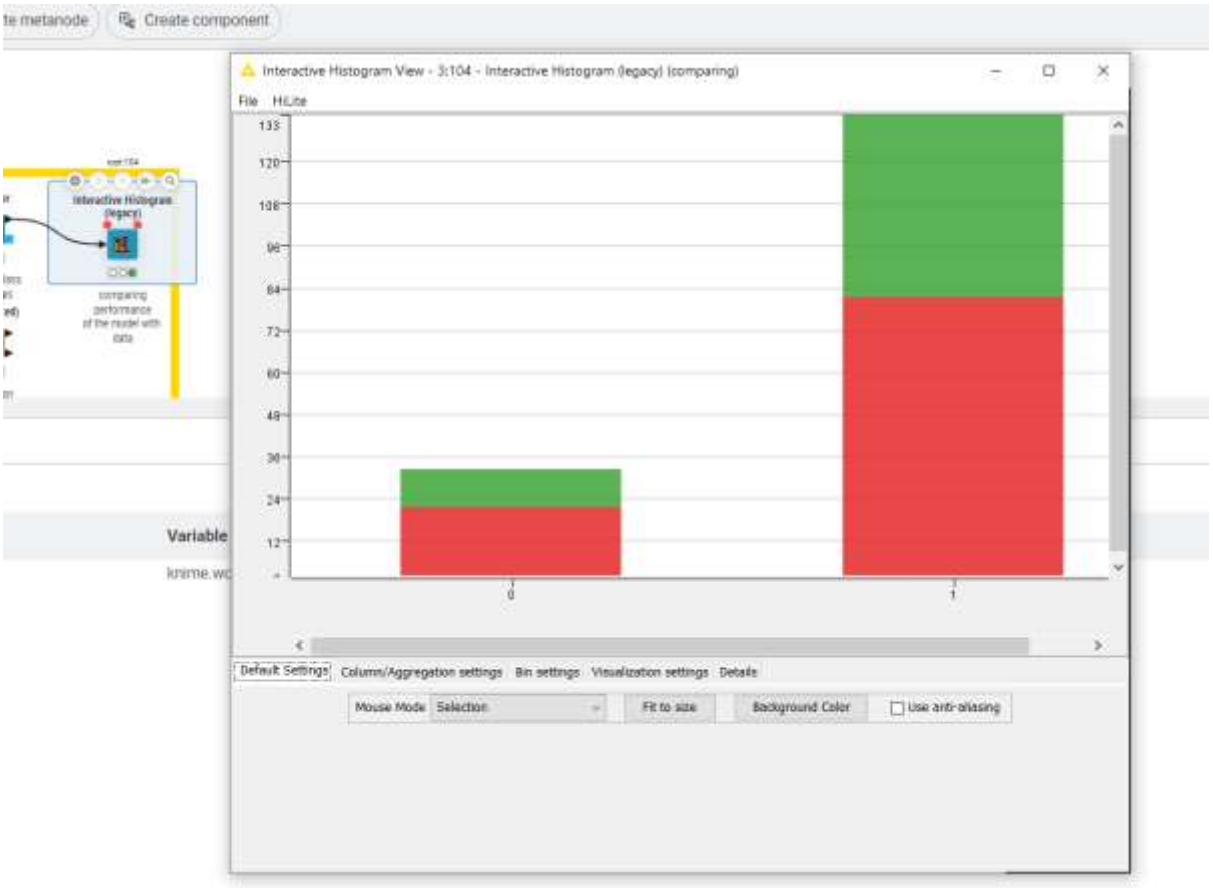

ALTERNATIVES


DATA UNDERSTANDING
DATA PREPARATION
MODELING & EVALUATION - FINAL

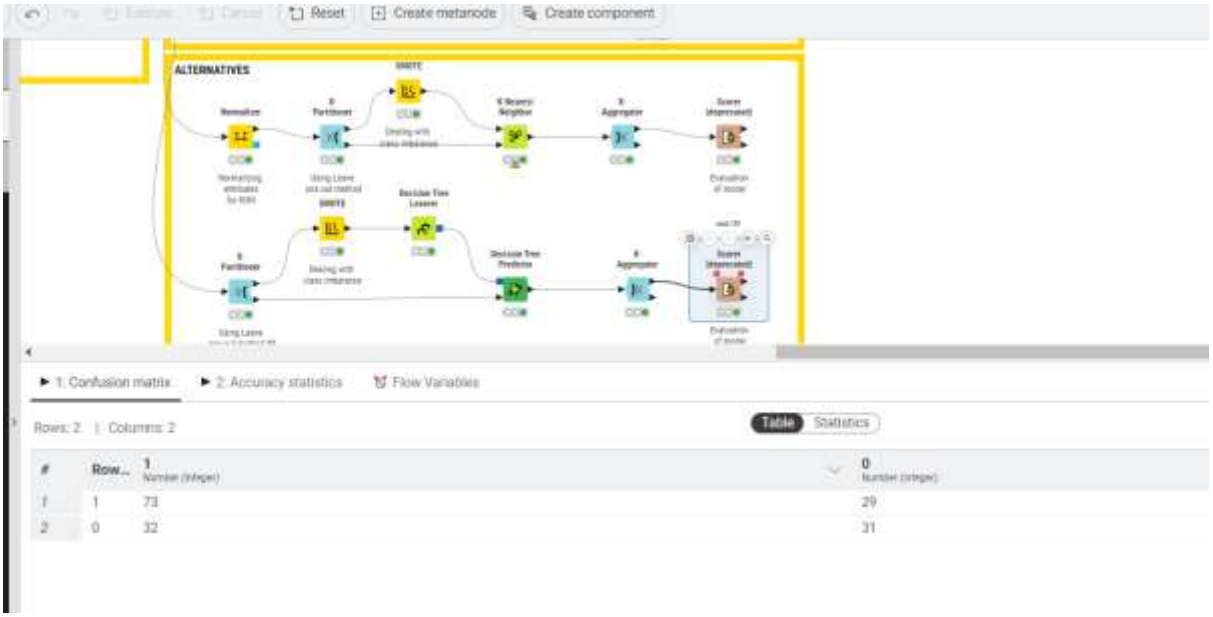

ALTERNATIVES


SMOTE
Dealing with
class imbalance


Tree Views - 3:19 - Random Forest Learner
File Tree HiLite
1
100 model(s) in total
1 (1/1)
Table:
Category % n
1 100.0 1
0 0.0 0
Total 100.0 1
35.Plat
or(35.Plat <= 26...
> 269,000
1 (1/1)
Table:
Category % n
1 100.0 1
0 0.0 0
Total 100.0 1
0 (1/1)
Table:
Category % n
1 0.0 0
0 100.0 1
Total 100.0 1
Zoom:
100.0%
FINAL
Random Forest
Learner
Add comment
9.End
String
10.Sm
String
0 1
0 1
0 1
0 1



#	Row...	1 Number (integer)	0 Number (integer)
1	1	72	30
2	0	23	40



► 1. Confusion matrix ► 2. Accuracy statistics Flow Variables

Rows: 2 | Columns: 2

Table Statistics

#	Row...	1 Number (integer)	0 Number (integer)
1	1	73	29
2	0	32	31

CONCLUSION

- Accuracy of Random Forest Model is around 75%.
- False positive rate is approximate 35%.
- Use of SMOTE reduces the skewed training of the model and increases accuracy of predictions.