

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

1	Problem Description
2	Problem Statement
3	Analysis
4	Design
5	Output
6	Conclusion and Future Scope

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.

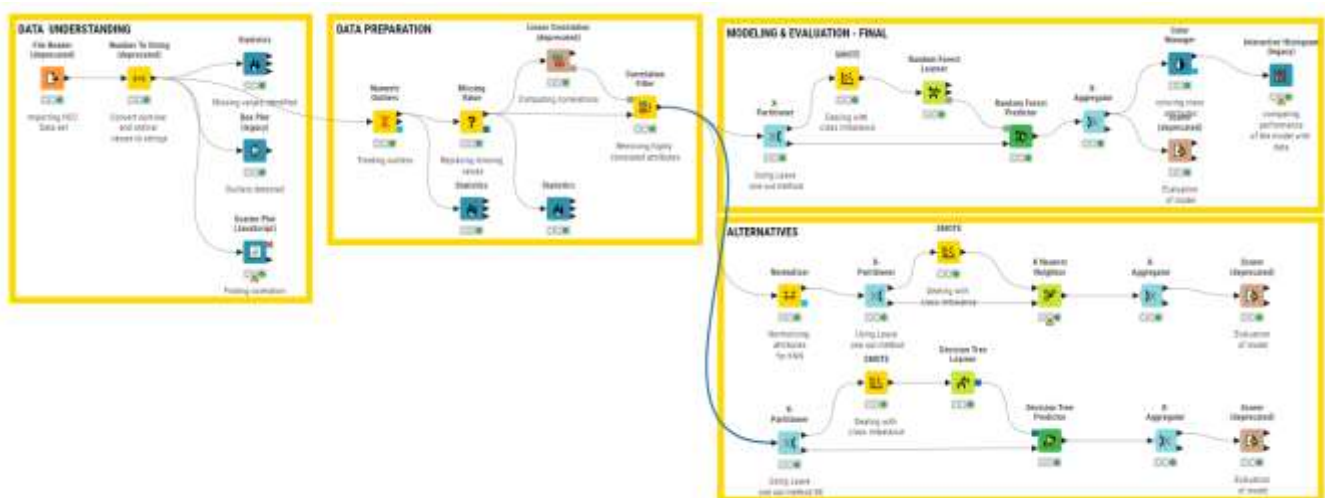
MODELING

- 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

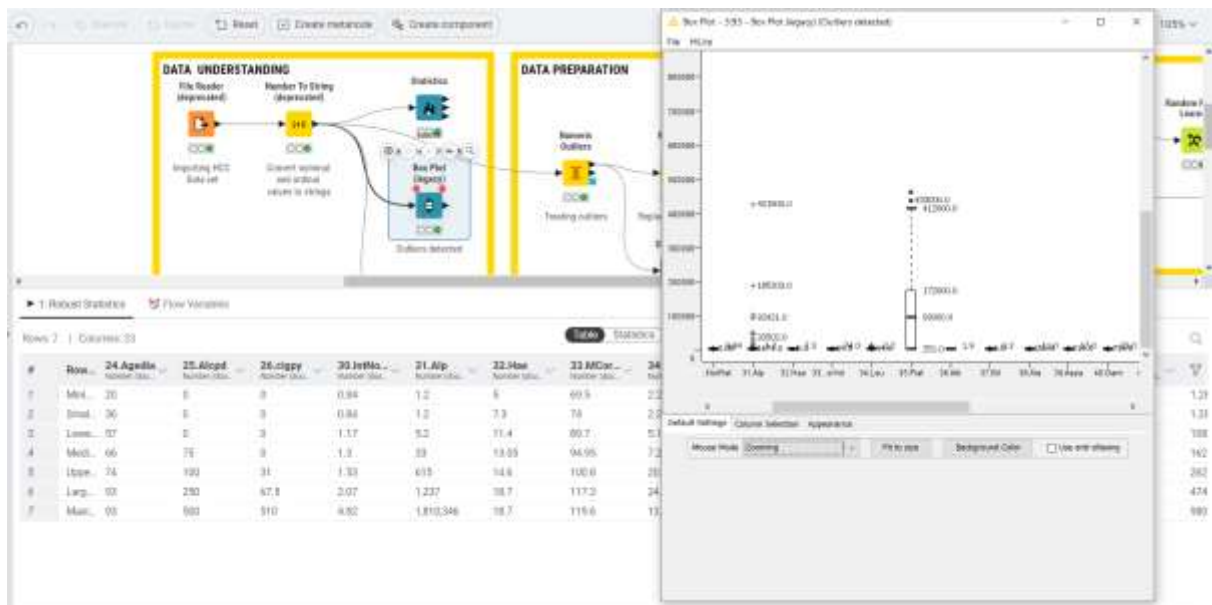
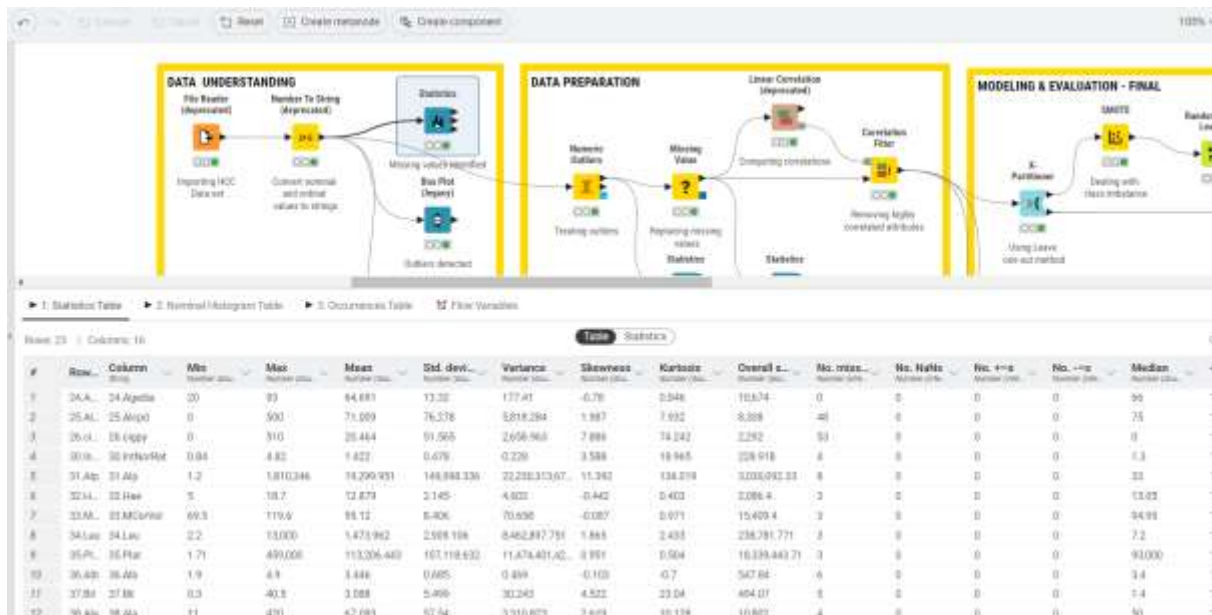
- Comparison of Models based on Minimization of False Positives and better accuracy.
- Random Forests excels in both Accuracy and Number of False Positive Predictions.

KNIME WORKFLOW



OUTPUT

[illegible]



Reset Create metanode Create component

HCC
rt

Convert nominal
and ordinal
values to strings

Box Plot
(legacy)

Outliers selected

Scatter Plot
(JavaScript)

Finding correlation

selection Flow Variables

Dialog - 3/4 - Scatter Plot (JavaScript) (Finding correlation)

File

View Controls Options

Flow Variables Axis Configuration

Job Manager Selection

Memory Policy General Plot Options

☐ Create image at output

Maximum number of rows: 2,500

Selection column name: Selected (Scatter Plot)

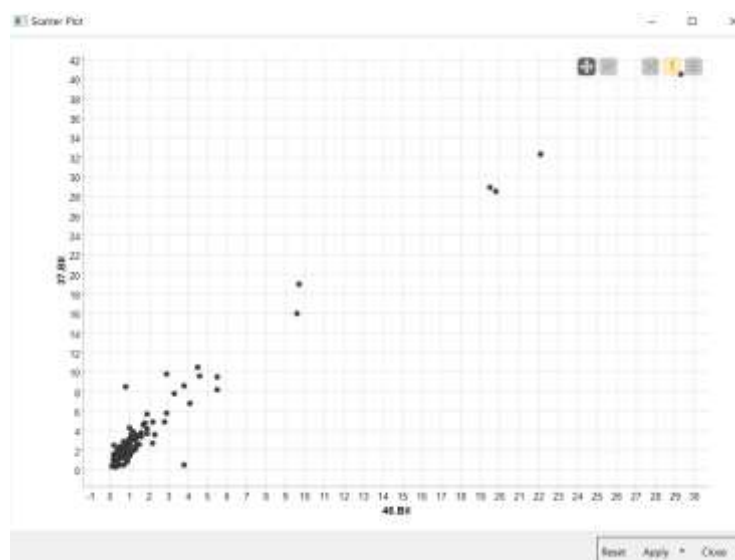
Choose column for x axis:
D: 46.BI

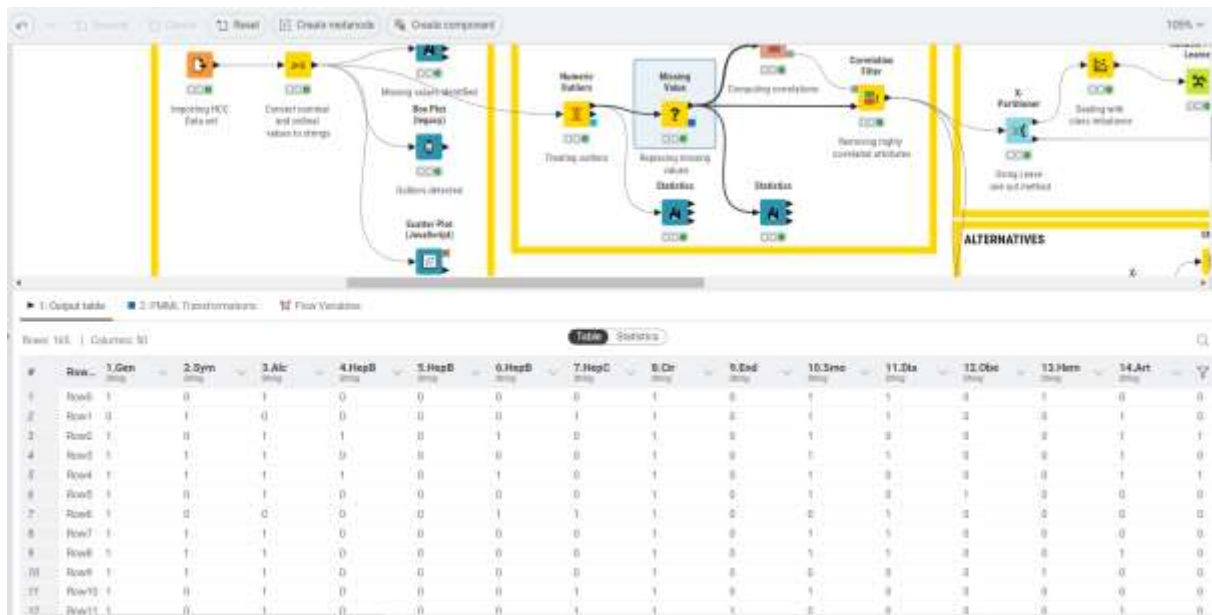
Choose column for y axis:
D: 37.BI

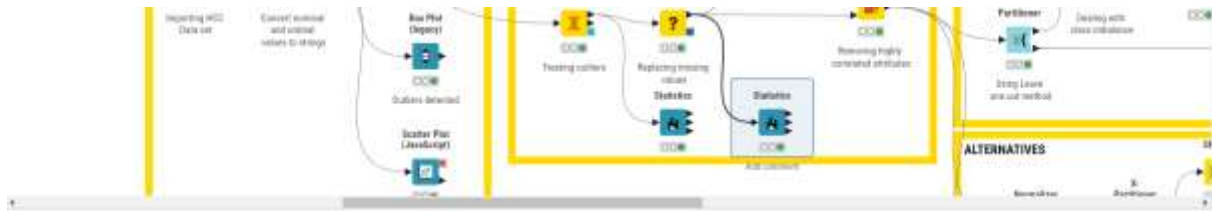
☒ Report on missing values

OK Apply Cancel

This output port is in



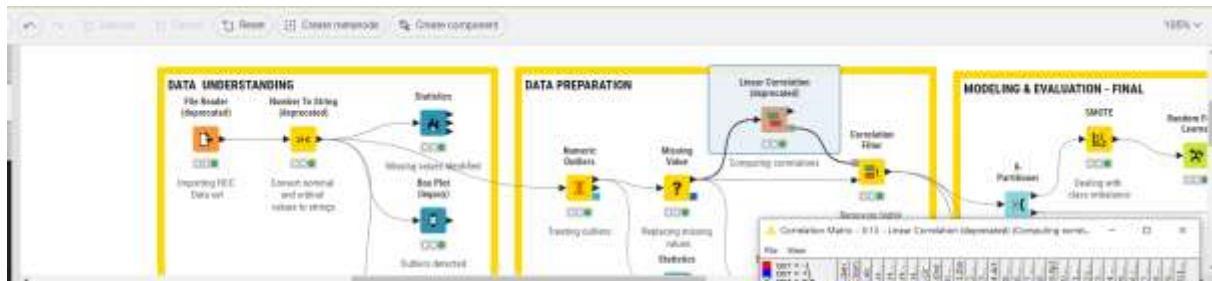




► 1. Statistics Table ► 2. Histogram Table ► 3. Descriptive Table ► Flow Variables

Rows: 13 | Columns: 18

#	Row	Column	Min	Max	Mean	Std. dev.	Variance	Skewness	Kurtosis	Overall s.	No. min.	No. Max	No. +ve	No. -ve	Median
1	24.A.	24.Age	32	93	64.891	12.743	162.39	-0.527	-0.031	16.797	0	0	0	0	66
2	25.A.	25.Age	0	358	10.352	35.393	1250.413	3.8	1.245	11.888	0	3	0	0	75
3	26.A.	26.Age	0	75	11.208	20.003	400.122	1.657	1.945	1.945	0	0	0	0	0
4	30.H.	30.Hist	0.84	2.07	1.374	0.295	0.867	0.885	0.168	215.788	0	0	0	0	1.0
5	31.H.	31.Hist	1.2	1,825.438	575.767	960.18	948,212.714	1.886	-0.092	62,093.893	0	0	0	0	53
6	32.H.	32.Hist	0.475	18.7	12.891	2.095	4.39	-0.331	-0.031	2,127.025	0	0	0	0	13.05
7	33.M.	33.Met	73.425	118.825	85.132	8.099	65.591	-0.040	0.554	15,888.8	0	0	0	0	94.05
8	34.L.	34.Liv	2.2	57.798	14.26	15.485	187.844	1.1	-0.022	2,352.942	0	0	0	0	7.2
9	35.H.	35.Hist	1.71	427,149.75	112,810.565	106,455.582	11,120,881.96	2.931	-0.412	18,581,743.21	0	0	0	0	65,330
10	36.A.	36.Air	1.9	4.0	3.444	0.673	0.452	-0.097	-0.613	964.24	0	0	0	0	3.4
11	37.H.	37.Hist	0.3	6.09	2.094	1.723	2.988	1.284	0.433	345.47	0	0	0	0	1.4
12	38.A.	38.Air	0.0	1.07	0.336	0.677	0.456	-0.047	-0.700	111.64	0	0	0	0	0.0



► 1. Correlation Measure ► 2. Correlation Model ► Flow Variables

Rows: 49 | Columns: 49

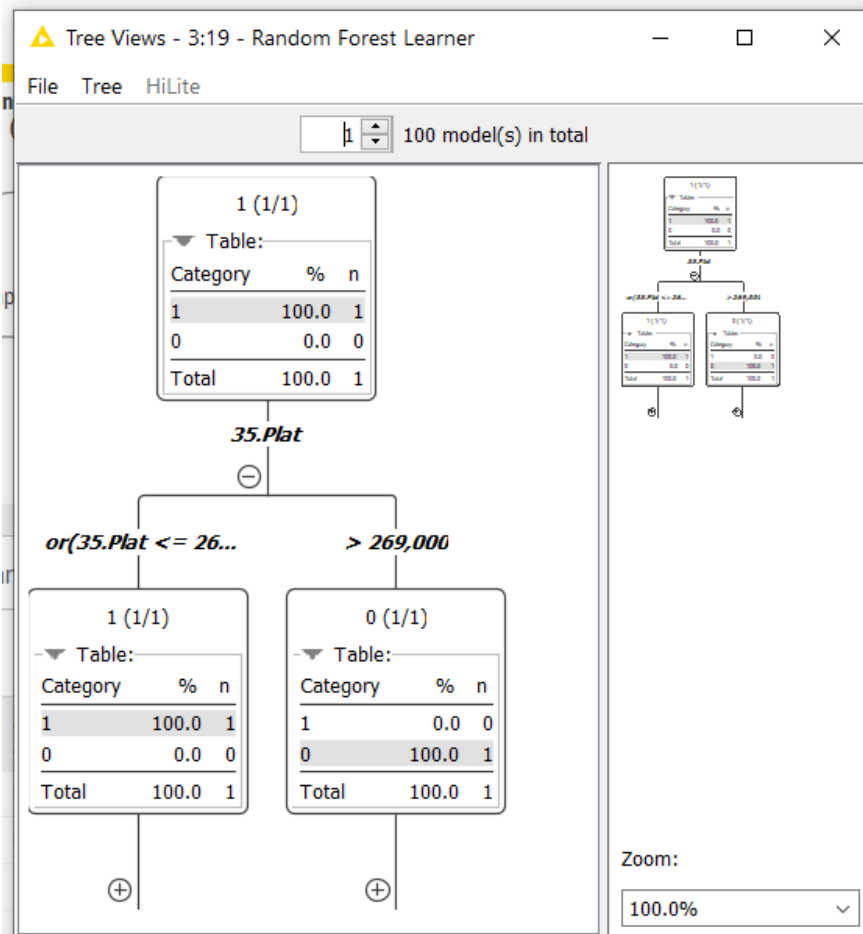
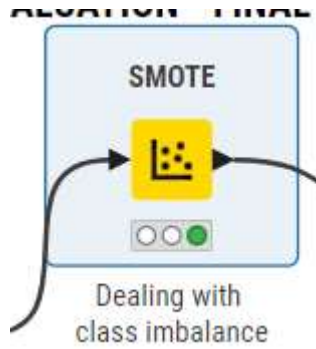
#	Row	1.Gen	2.Sym	3.Air	4.HepB	5.HepC	6.HepB	7.HepC	8.Cir	9.End	10.Smo	11.Dia	12.Obs	13.Hem	14.Art
1	1.Gen	1	0.027	0.442	0.181	0.038	0.05	0.053	0.254	0.034	0.034	0.034	0.034	0.034	0.034
2	2.Sym	0.027	1	0.004	0.094	0.054	0.098	0.083	0.084	0.096	0.096	0.096	0.096	0.096	0.096
3	3.Air	0.442	0.004	1	0.085	0.046	0.089	0.141	0.458	0.081	0.081	0.081	0.081	0.081	0.081
4	4.HepB	0.181	0.094	0.085	1	0.238	0.404	0.075	0.107	0.340	0.340	0.340	0.340	0.340	0.340
5	5.HepC	0.038	0.054	0.046	0.238	1	0.143	0.04	0.025	0.02	0.02	0.02	0.02	0.02	0.02
6	6.HepB	0.05	0.099	0.069	0.404	0.143	1	0.22	0.131	0.102	0.102	0.102	0.102	0.102	0.102
7	7.HepC	0.053	0.083	0.141	0.075	0.04	0.22	1	0.086	0.122	0.122	0.122	0.122	0.122	0.122
8	8.Cir	0.254	0.084	0.458	0.107	0.025	0.131	0.086	1	0.089	0.089	0.089	0.089	0.089	0.089
9	9.End	0.034	0.096	0.081	0.340	0.02	0.102	0.122	0.089	1	0.091	0.091	0.091	0.091	0.091
10	10.S.	0.034	0.096	0.081	0.340	0.02	0.102	0.122	0.089	0.091	1	0.091	0.091	0.091	0.091
11	11.Dia	0.034	0.096	0.081	0.340	0.02	0.102	0.122	0.089	0.091	0.091	1	0.091	0.091	0.091
12	12.Obs	0.034	0.096	0.081	0.340	0.02	0.102	0.122	0.089	0.091	0.091	0.091	1	0.091	0.091
13	13.Hem	0.034	0.096	0.081	0.340	0.02	0.102	0.122	0.089	0.091	0.091	0.091	0.091	1	0.091
14	14.Art	0.034	0.096	0.081	0.340	0.02	0.102	0.122	0.089	0.091	0.091	0.091	0.091	0.091	1



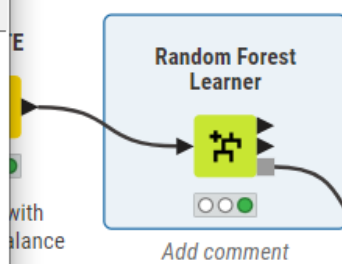
► 1. Filtered data from input ► 2. Flow Variables ► Flow Variables

Rows: 155 | Columns: 49

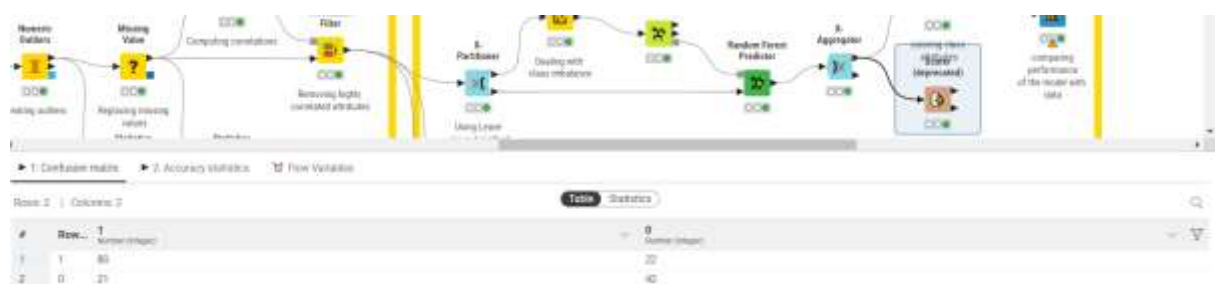
#	Row	1.Gen	2.Sym	3.Air	4.HepB	5.HepC	6.HepB	7.HepC	8.Cir	9.End	10.Smo	11.Dia	12.Obs	13.Hem	14.Art
1	Row1	1	0	1	0	0	0	0	1	0	1	1	0	0	0
2	Row1	0	1	0	0	0	0	1	1	0	1	1	0	0	1
3	Row2	1	0	1	1	0	1	0	1	0	1	0	0	0	1
4	Row3	1	1	1	0	0	0	0	1	0	1	1	0	0	1
5	Row4	1	1	1	1	0	1	0	1	0	1	0	0	0	1
6	Row5	1	0	1	0	0	0	0	1	0	1	0	1	0	0
7	Row6	1	0	0	0	0	1	1	1	0	0	1	0	0	0
8	Row7	1	1	1	0	0	0	0	1	0	1	1	0	0	0
9	Row8	1	1	1	0	0	0	0	1	0	1	1	0	0	1
10	Row9	1	1	1	0	0	0	0	1	0	0	0	0	1	0
11	Row10	1	0	1	0	0	0	1	1	0	1	0	0	0	0

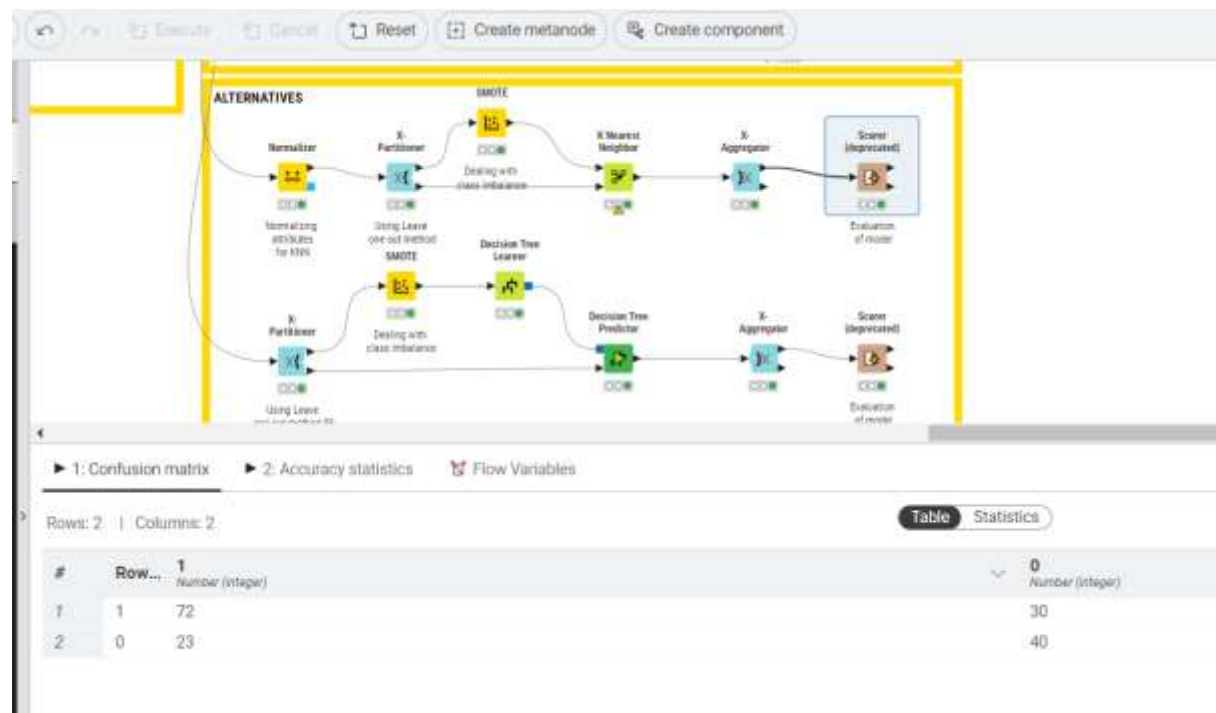
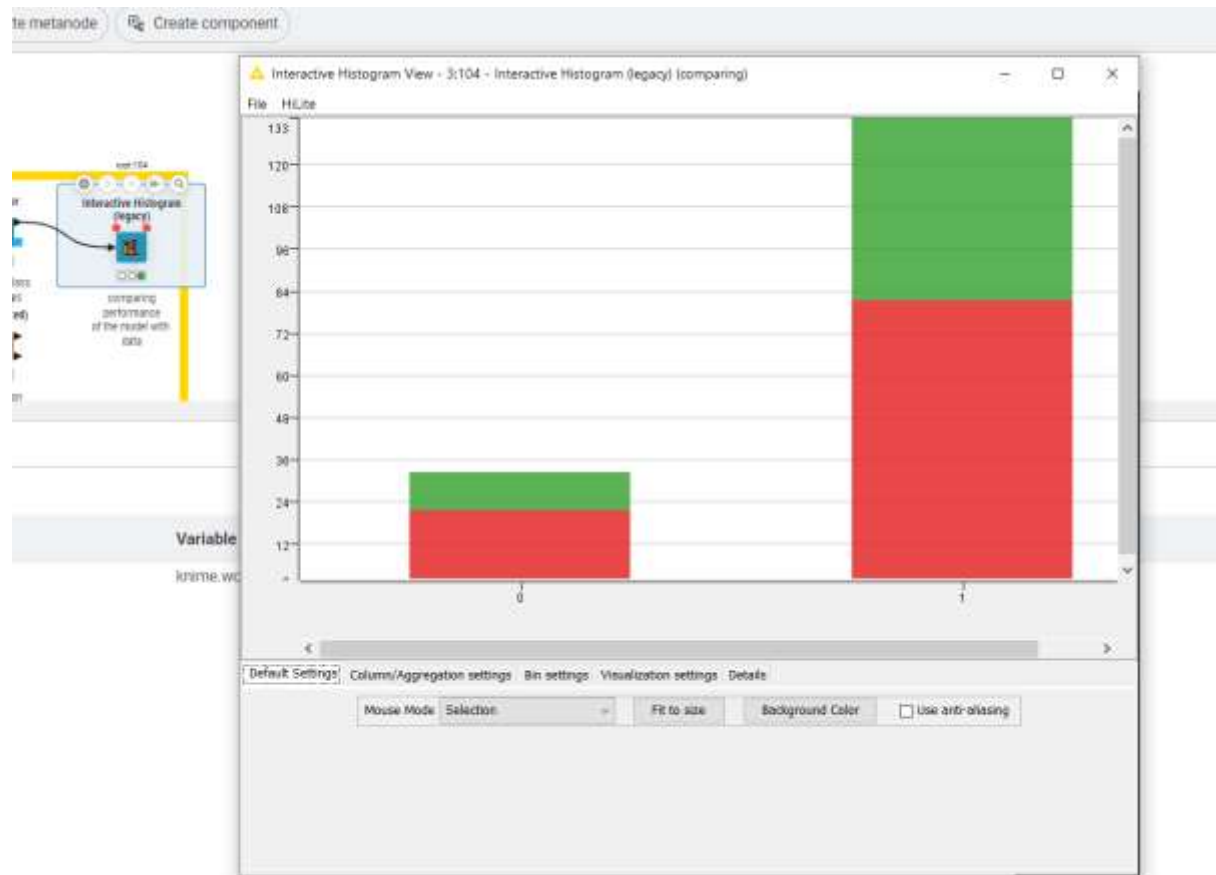


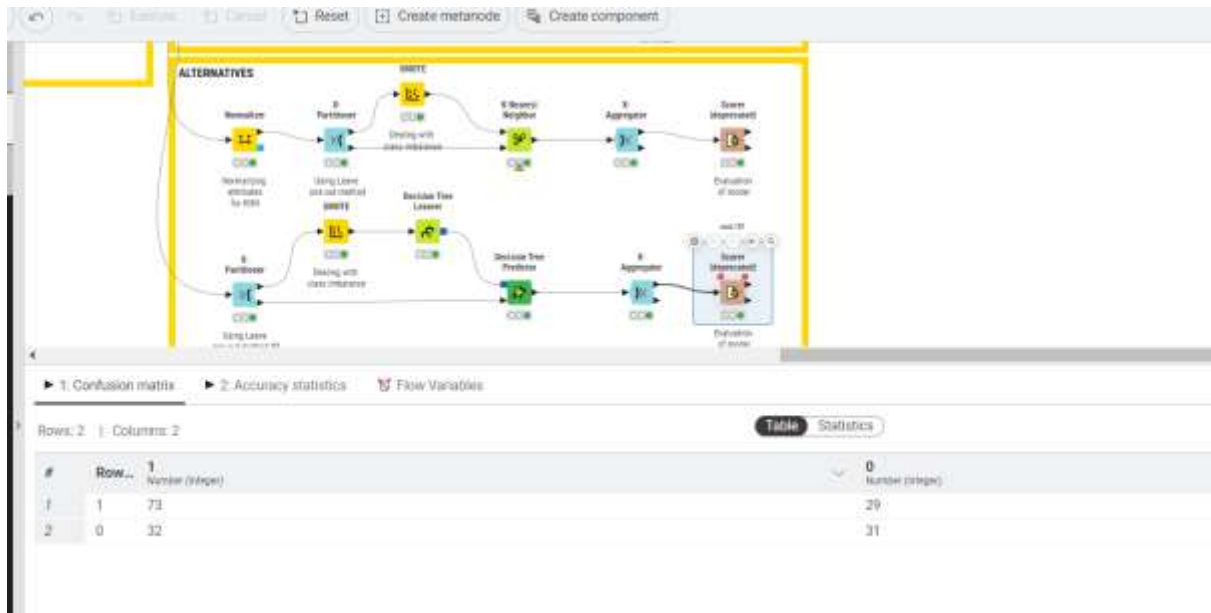
FINAL



9.End String	10.Sm String
0	1
0	1
0	1
0	1







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.