Group 10: HCC Survivals

Ashish Soni, Darshit Paresh Shah, Momtaj Anar Monni, Roland Schnee

Challenge

- Predict whether a patient will survive the diagnosed liver cancer

Positives (1): Survives
Negatives(0): Death

Dataset

- Number of Instances: 165
- Number of Attributes: 49 + class attribute
- 24 nominal features with 2 values each

Age @ diagnosis - Gender,
Grams of alcohol per day - Oxygen saturation
Major dimension of nodules (cm) - Haemoglobin

Data Understanding

- Only 8 patients have complete data.
- Non-Ignorable Missing Values.
- No Semantic and Syntactic Errors
- Outlier Treatment – **Closely Permitted Value**
- Missing Value Substitution – **Median / Most Frequent Value**

Data Preparation

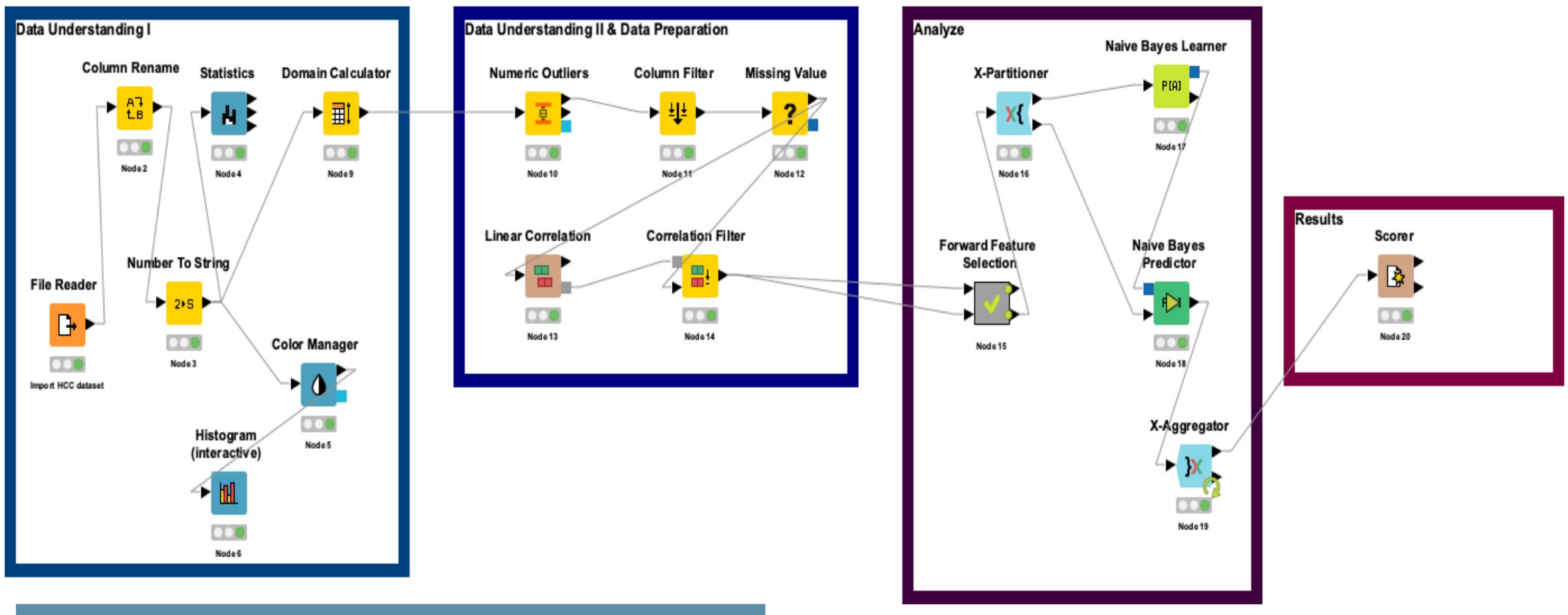
- Remove columns with missing values **>60**
- Forward Feature Selection :
  7 attributes ~ 79%
- Correlation Threshold – **0.7**

Selected Attributes

- Hepatitis B Core Antibody
- Hepatitis C Virus Antibody
- Endemic Countries
- International Normalized Ratio
- Ascites degree
- Alpha-Fetoprotein (ng/mL)
- Mean Corpuscular Volume(fl)
- Alkaline phosphatase (U/L)

Modelling

- Partitioning
  Cross Validation-10 fold
- Classification Model
  Naive Bayes



Evaluation

- Naive Bayes
  Accuracy ~ 77%

Results

7 attributes are enough to classify the given instances.
Dataset is too small and has too many missing values, therefore achieving a better/perfect accuracy is difficult.