OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

WW FAKULTÄT FÜR WIRTSCHAFTSWISSENSCHAFT

Management Science

Group 12: HCC Survival

Pankaj Tiwari, Chaitanya Upadrashta, Baalakrishnan Aiyer Manikandan, Aikaterini Drizi

Challenge

Predict whether a Patient will survive the diagnosed liver cancer.

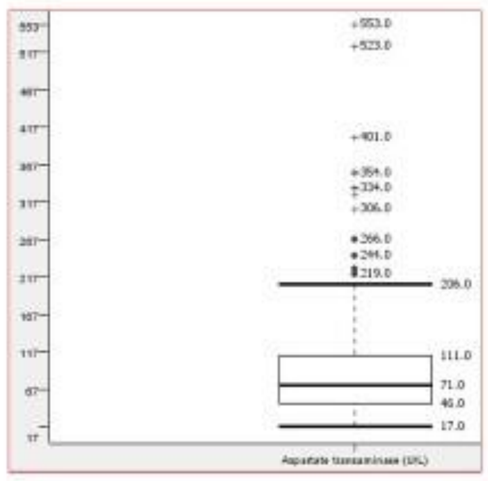
Class 0: patient Dies
Class 1: patient Survives

Dataset

- Contains data of 165 real patients diagnosed with HCC(liver Cancer)
- 49 features + 1 Class Attribute
- Nominal features : 23
- Ordinal features: 3
- Continuous features: 23

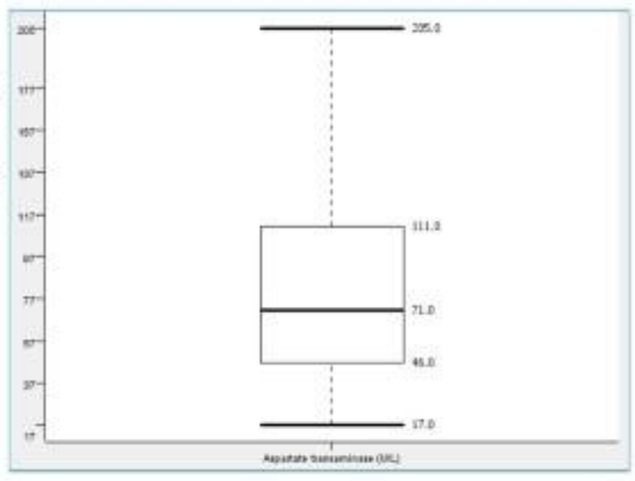
Data Understanding

- 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

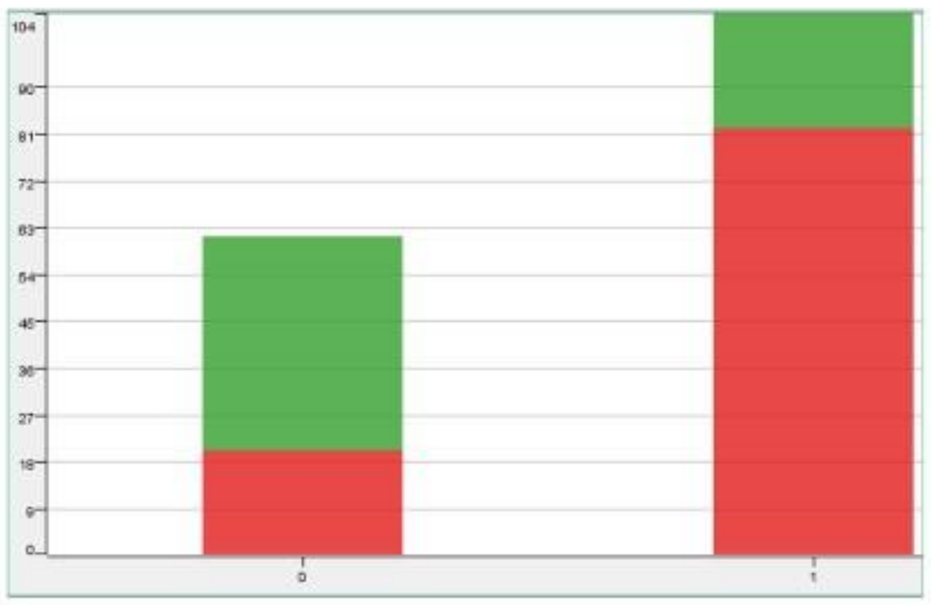
- Class Imbalance was fixed using SMOTE node (minority-Class 0 oversampled)
- X- Partitioner with Leave one- out was used to generate Training and Test data
- Classification Models Tested:-
  - Random Forest
  - Decision Tree
  - K- Nearest Neighbour

Evaluation

- Comparison of Models based on Minimization of False Positives(Diseased patient predicted as "Survived") and Better Accuracy.
- Random Forests excels in both Accuracy and Number of False Positive Predictions.

Results

- Accuracy of Random Forest Model is around 75%
- False Positive Rate is approx. 35%.
- Use of SMOTE reduces the skewed training of the model and increases accuracy of prediction.



Knime-Workflow