Project Report: Heart Failure Prediction

By

Azam Jah Al Ahmed

Tejashri Chavan

INDEX

Sr. No.	Topic	Description
1	**Summary of Project**	Overview of the project’s purpose, scope, and key insights and findings.
2	**Business Problem**	Description of the problem or opportunity being addressed (from the proposal) and its relevance/impact on business or healthcare.
3	**Business Objective**	Clear and measurable objectives of the project (from the proposal), outlining how the project aims to solve the problem or capitalize on the opportunity.
4	**Selecting and Gathering Data**	Methods and criteria used to select the data, sources of data, and rationale behind the selection.
5	**Data Exploration and Findings**	Initial analysis, observations, key patterns, trends, and potential outliers identified in the data.
6	**Description of Data Preparation**	Details on data preparation steps: repairs, replacements, reductions, partitions, derivations, transformations, and variable clustering.
7	**Description of Data Modeling/Analyses and Assessments**	Explanation of chosen models, evaluation metrics used, and insights gained from model assessments.
8	**Explanation of Model Comparisons and Model Selection**	Comparative analysis of different models’ performances and reasons for selecting the best model based on evaluation metrics.
9	**Conclusions and Recommendations**	Summary of findings, whether business objectives were met, implications for the business problem, and suggestions for future work.

Project Report: Heart Failure Prediction

Summary

Heart failure is one of the leading causes of hospitalization and mortality worldwide, presenting an enormous challenge to healthcare systems and patient care. The project will apply machine learning to construct a predictive model that pinpoints patients at high risk for heart failure, thus enabling timely intervention and better outcomes.

The project is anchored on anonymized patient data, including clinical, demographic, and lifestyle variables for training and validation of the predictive model. By identifying patterns in risk factors related to heart failure, the proposed solution will enable healthcare practitioners to apply proactive and timely personalized care strategies that decrease late-stage complications and improve patient quality of life.

Key objectives will be to improve early risk detection, enhance operational efficiency of healthcare systems, and enable data-driven decision-making.

Business problem

Cardiovascular-related diseases are the primary cause of death globally and heart failure (HF) is a frequent ultimate manifestation of many heart pathologies. Timely interventions through early diagnosis and control of EHF (End-Stage Heart Failure) condition would greatly improve results. The increasing popularity of machine learning algorithms allows clinicians to develop predictive models that can be used to predict whether or not an individual will develop heart failure using many medical and lifestyle factors. Objective: To develop a predictive model to help healthcare providers determine high-risk (HR) patients who need tailored preventive care. Machine learning can make predictions about heart failure risks from medical and lifestyle factors and thereby aimed at identifying HR patients who will be provided with appropriate preventive care by healthcare workers.

Business Objective

The main business purpose of this analysis is to create an accurate prediction machine learning model identifying if a patient has the risk of Heart Failure due to their sustainable features predicated on medical and lifestyle necessities. Healthcare providers can use the model to decide which patients need urgent attention and how to best allocate medical resources. Furthermore, it has the potential to develop tailored treatments for those at risk.

Data Gathering

We have conducted a thorough search on public data repositories over the internet and

extracted the Heart Failure Prediction Dataset from Kaggle website. The data is of CSV format with approximately 900 number of observations and including labels in appropriate format relevant to heart disease prediction. The data is well structured contains critical features such as age, sex, blood pressure, cholesterol, and ECG results.

Preliminary Data Exploration and Findings

The dataset contains 12 features, including both numerical and categorical variables.

The dataset contains various medical and lifestyle variables that can impact heart disease risk. The following are the independent variables in the dataset:

Age: Age of the patient.

Sex: Gender of the patient (1 = Male, 0 = Female).

ChestPainType: Type of chest pain (4 categories: TA, ATA, NAP, ASY).

RestingBP: Resting blood pressure (in mm Hg).

Cholesterol: Serum cholesterol level (in mg/dl).

FastingBS: Fasting blood sugar (> 120 mg/dl) (1 = True, 0 = False).

RestingECG: Resting electrocardiogram results (Normal, ST, LVH).

MaxHR: Maximum heart rate achieved during exercise.

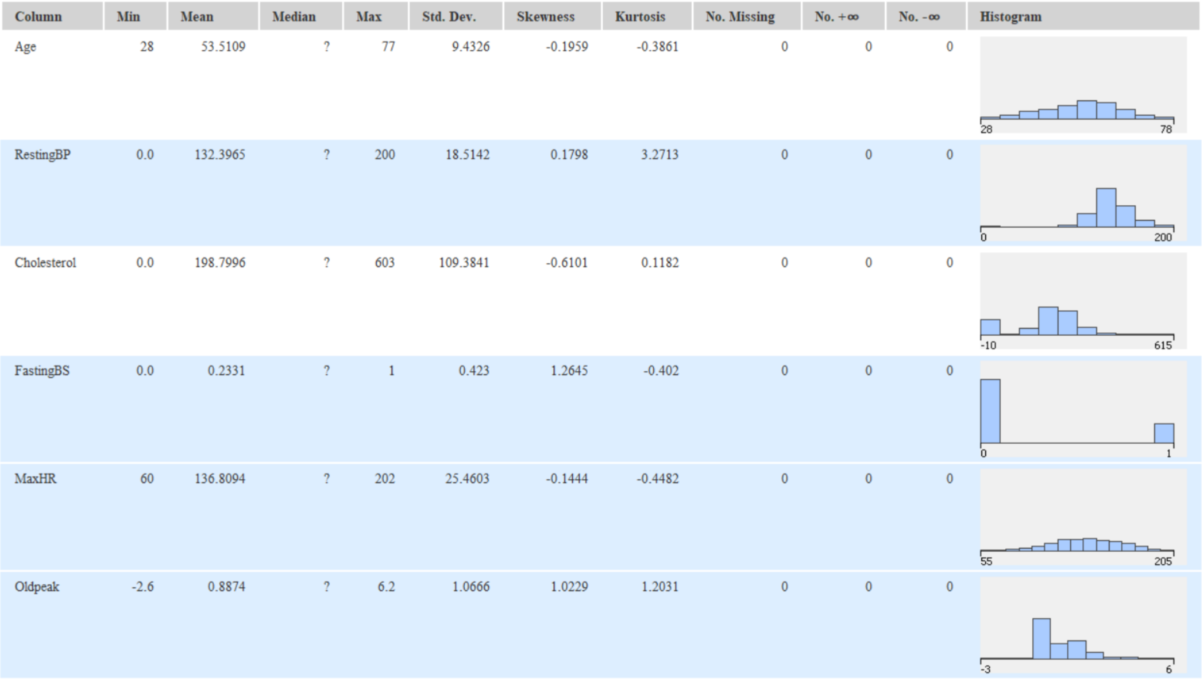
ExerciseAngina: Exercise-induced angina (1 = Yes, 0 = No).

Oldpeak: ST depression induced by exercise relative to rest.

ST_Slope: The slope of the peak exercise ST segment (Up, Flat, Down).

Heart Disease: This is a binary dependent variable that indicates whether the patient has been diagnosed with heart disease (1 = Yes, 0 = No)

These variables, reflecting patient demographics, symptoms, and clinical test results, are expected to provide valuable insights for heart disease prediction. During the analysis, we found that there are no missing values for any features. We used the Statistics node from KNIME to further explore the data and assess the distribution across the dataset. From the screenshot, we can see that the distribution curve of most features is normally distributed, except for the FastingBS feature. This is because it contains only binary values (1 or 0), which are expected values for identifying fasting blood sugar spikes. Additionally, this variable shows no unusual spikes at the 0 or 1 values. We also found that the data includes observations from the age group of 28 to 77, with a mean age of 55. This is significant because most heart disease cases occur within mean age group.



Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
Age	28	53.5109	?	77	9.4326	-0.1959	-0.3861	0	0	0	28 78
RestingBP	0.0	132.3965	?	200	18.5142	0.1798	3.2713	0	0	0	0 200
Cholesterol	0.0	198.7996	?	603	109.3841	-0.6101	0.1182	0	0	0	-10 615
FastingBS	0.0	0.2331	?	1	0.423	1.2645	-0.402	0	0	0	0 1
MaxHR	60	136.8094	?	202	25.4603	-0.1444	-0.4482	0	0	0	55 205
Oldpeak	-2.6	0.8874	?	6.2	1.0666	1.0229	1.2031	0	0	0	-3 6

Data preparation

We have converted target variable **Heart Disease** to String to clearly define the two distinct classes which is crucial for the model to understand and learn the relationship between features and the target.

We split data in following way.

Training Data: Approximately 643 (70% of the total observations) will be used for training the model.

Validation Data: Approximately 275 (30% of the total observations) will be used for validating the model's performance

Data Modelling/Analyses and Assessments

In this project we have explored to identify the below models for heart disease prediction.

Decision Trees: The performance of this model will be assessed in terms of its nonlinear relationship between the features and the target. Decision trees have interpretable results, enabling the understanding of the underlying decision-making process for the prediction of heart disease.

Logistic Regression: A simple yet effective model for binary classification tasks. It will be used as a baseline model to predict the presence or absence of heart disease based on the features.

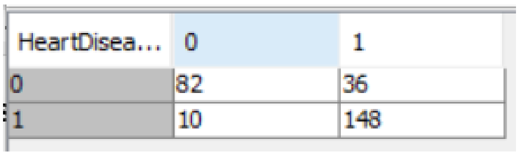
Neural networks: Neural networks are efficient in handling complex and high-dimensional datasets. They consist of interconnected neurons and layers, just like the human brain. Each neuron performs specific mathematical operations, and the output layer produces the final result.

Model Assessments:

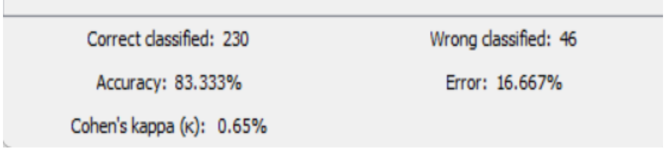
1. Decision Tree:

In this model training, we have used two pruning methods **MDL (Minimum Description Length)** pruning and **No Pruning.**

MDL Pruning: This method works to find a balance between tree complexity and prediction accuracy by pruning unnecessary branches based on the description length principle, thereby reducing overfitting and improving generalization. Using this method, we were able to correctly classify 230 observations for the patient that are diagnosed with heart disease and 46 observations were wrongly classified. Overall, the model is accuracy is approximately 83.33 %.



HeartDisea...	0	1
0	82	36
1	10	148

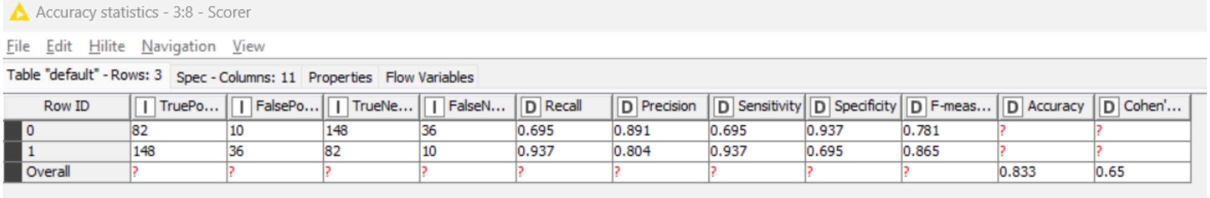


Correct classified: 230 Wrong classified: 46

Accuracy: 83.333% Error: 16.667%

Cohen's kappa (κ): 0.65%

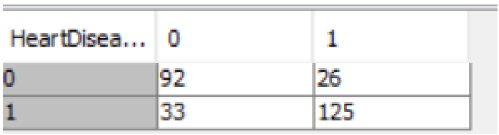
Accuracy statistics:



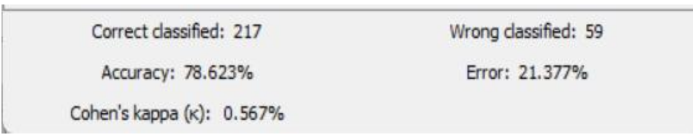
Accuracy statistics - 3:8 - Scorer

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
0	82	10	148	36	0.695	0.891	0.695	0.937	0.781	?	?
1	148	36	82	10	0.937	0.804	0.937	0.695	0.865	?	?
Overall	?	?	?	?	?	?	?	?	?	0.833	0.65

No Pruning: The tree is allowed to grow to its fullest without any constraints, sometimes leading to overfitting when the tree becomes too big for the available data. In this method, we got 217 observations that were correctly classified patients that are diagnosed with heart disease and 59 observations being wrongly classified. The overall accuracy is approximately 78.6%.



HeartDisea...	0	1
0	92	26
1	33	125

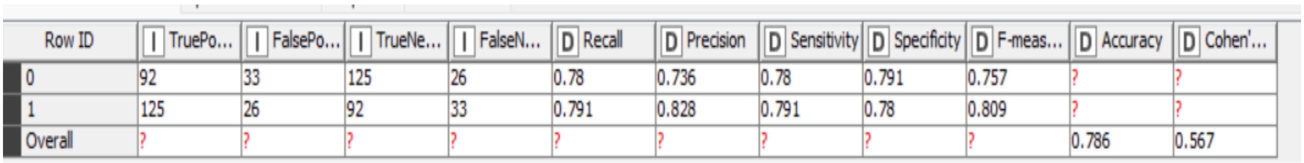


Correct classified: 217 Wrong classified: 59

Accuracy: 78.623% Error: 21.377%

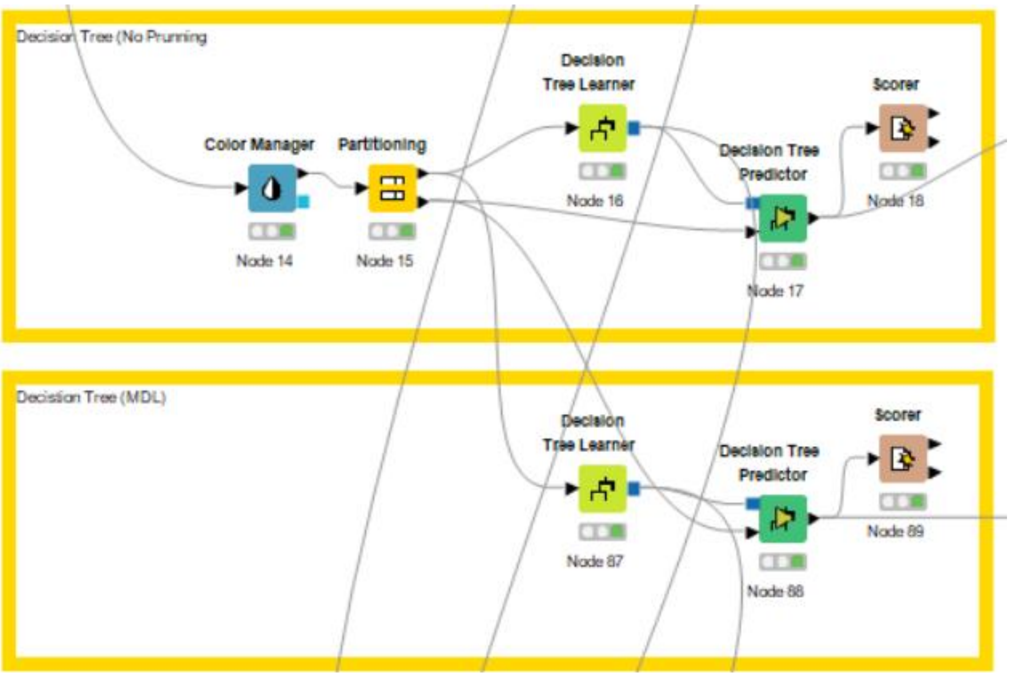
Cohen's kappa (κ): 0.567%

Accuracy statistics:



Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
0	92	33	125	26	0.78	0.736	0.78	0.791	0.757	?	?
1	125	26	92	33	0.791	0.828	0.791	0.78	0.809	?	?
Overall	?	?	?	?	?	?	?	?	?	0.786	0.567

Knime workflow:



2. Logistic regression:

Model Training -

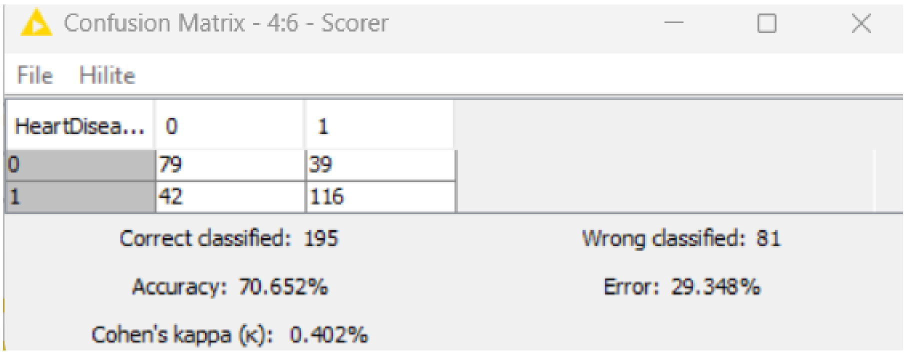
The logistic regression model was trained using the following steps:

Input Variables: Independent variables such as Age, Sex, RestingBP, Cholesterol, MaxHR, etc., were used as predictors.

Output Variable: The target variable HeartDisease was used as the dependent variable.

Training-Test Split: The data was divided into training (70%) and test (30%) sets to evaluate model performance on unseen data.

Performance Evaluation Metrics



HeartDisea...	0	1
0	79	39
1	42	116

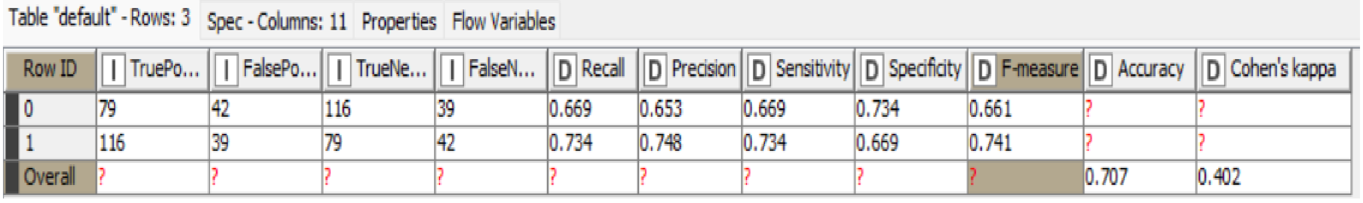
Correct classified: 195

Wrong classified: 81

Accuracy: 70.652%

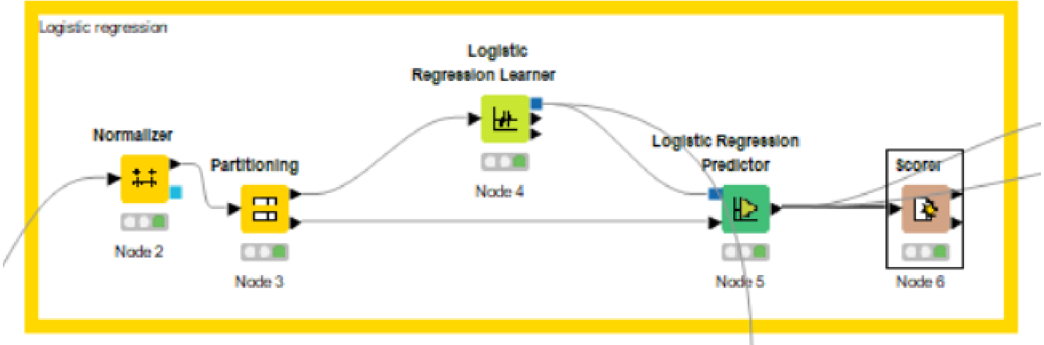
Error: 29.348%

Cohen's kappa (κ): 0.402%



Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-measure	D Accuracy	D Cohen's kappa
0	79	42	116	39	0.669	0.653	0.669	0.734	0.661	?	?
1	116	39	79	42	0.734	0.748	0.734	0.669	0.741	?	?
Overall	?	?	?	?	?	?	?	?	?	0.707	0.402

KNIME Work Flow



Interpretation of Results

- The logistic regression model performed reasonably well, with an accuracy of 70.7% and a strong recall of 73.4% for detecting heart disease.

- The F1-score for HeartDisease = 1 (0.741) indicates that the model balances precision and recall effectively for heart disease cases.
- Cohen’s Kappa (0.402) highlights that while the model shows moderate predictive performance, there is room for improvement, particularly in reducing false negatives and false positives.

Clinical Implications

- The model's **high recall** for heart disease cases ensures it captures most patients at risk, making it useful for screening purposes.
- However, the **moderate precision** (0.748) suggests some false positives, meaning further confirmatory tests would be needed for flagged cases.

3. Neural network:

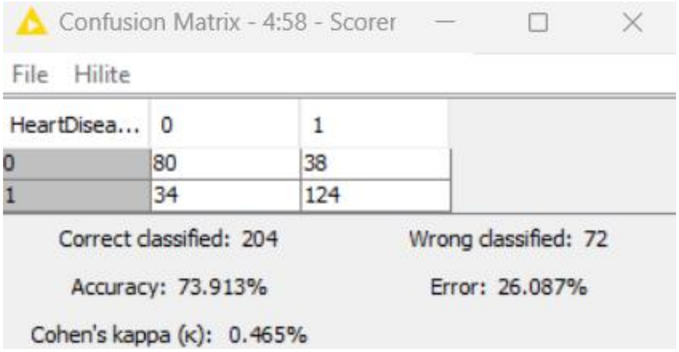
Model Training

Input Variables: Independent variables such as Age, Sex, RestingBP, Cholesterol, MaxHR, etc., were used as predictors.

Output Variable: The target variable HeartDisease was used as the dependent variable.

Training-Test Split: The data was divided into training (70%) and test (30%) sets to evaluate model performance on unseen data.

Performance Metrics



Confusion Matrix - 4:58 - Scorer

File Hilite

HeartDisea…	0	1
0	80	38
1	34	124

Correct classified: 204 Wrong classified: 72

Accuracy: 73.913% Error: 26.087%

Cohen's kappa (κ): 0.465%

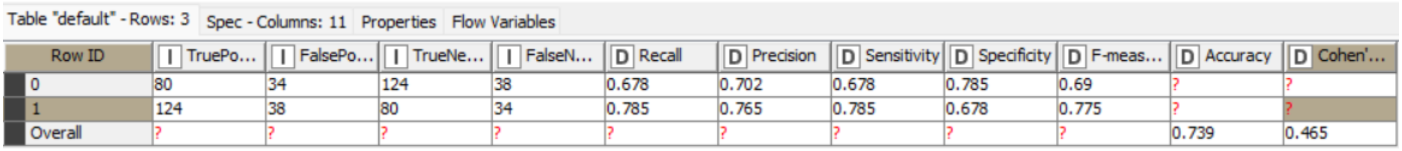
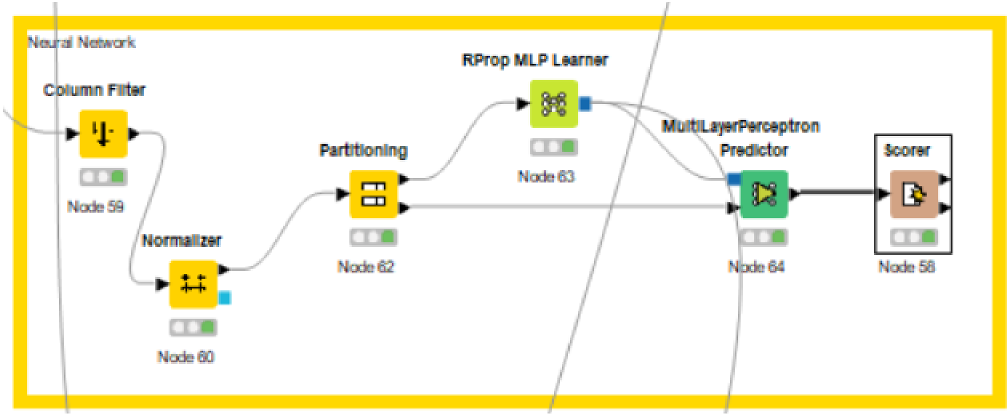


Table "default" - Rows: 3 Spec - Columns: 11 Properties Flow Variables

Row ID	I TruePo…	I FalsePo…	I TrueNe…	I FalseN…	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas…	D Accuracy	D Cohen'…
0	80	34	124	38	0.678	0.702	0.678	0.785	0.69	?	?
1	124	38	80	34	0.785	0.765	0.785	0.678	0.775	?	?
Overall	?	?	?	?	?	?	?	?	?	0.739	0.465

KNIME WORK FLOW



Interpretation of Results

- The neural network achieved an **accuracy of 73.9%**, outperforming the logistic regression model in this case.
- The model demonstrated **higher recall and precision**, particularly for identifying heart disease cases, making it more reliable in reducing false negatives (missed cases).
- The moderate **Cohen's Kappa (0.465)** suggests that while the model performed better than chance, further improvements could be achieved through feature engineering or additional data.

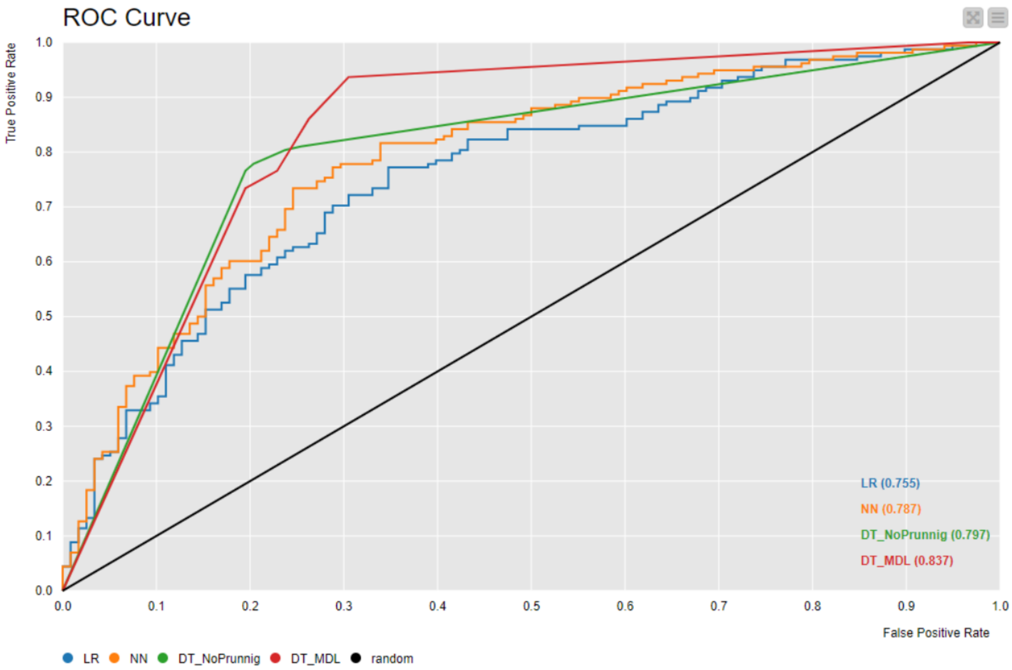
Clinical Implications

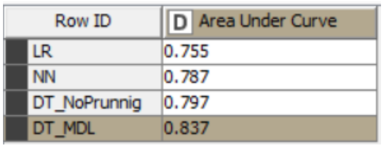
- The neural network's superior recall for heart disease cases makes it highly effective in identifying patients at risk, which is critical in clinical applications.
- However, the presence of false positives (FP = 34) and false negatives (FN = 38) indicates that further refinements are necessary to improve the model’s specificity.

Future Enhancements

- Incorporating additional clinical features or biomarkers could improve predictive performance.
- Experimentation with advanced techniques, such as ensemble learning or deeper neural networks, may yield better results.

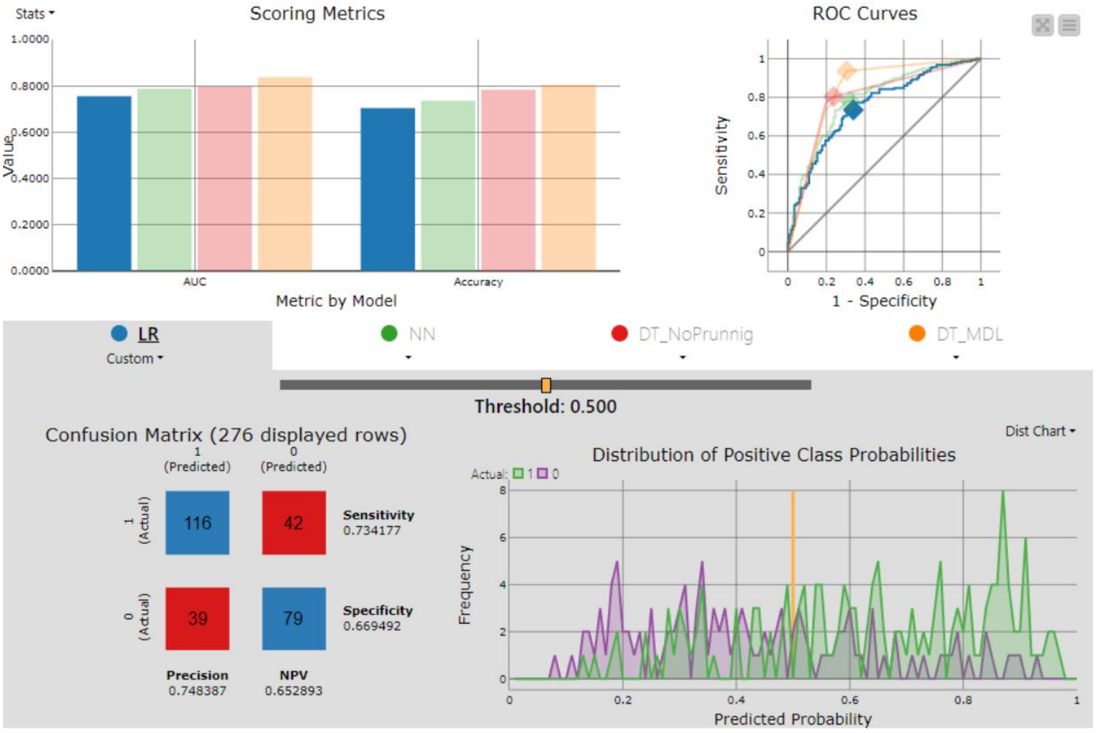
Model comparisons and Model selection





Row ID	D Area Under Curve
LR	0.755
NN	0.787
DT_NoPrunnig	0.797
DT_MDL	0.837

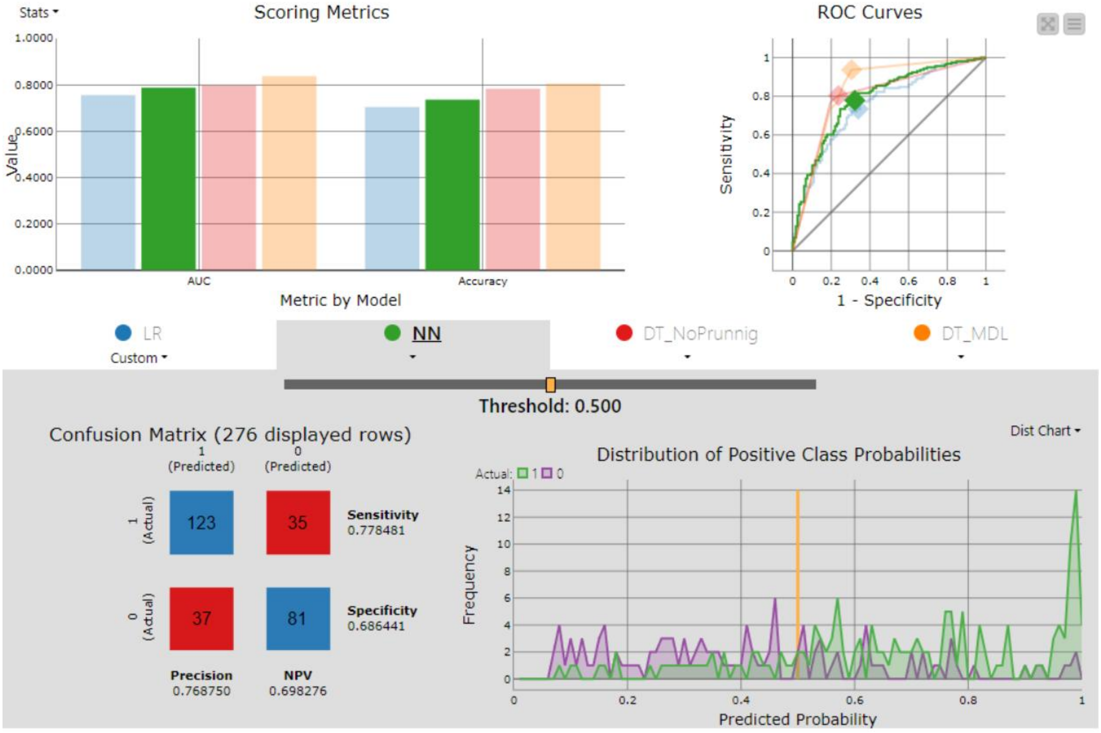
BINARY CLASSIFICATION INSPECTOR



Threshold: 0.500

Confusion Matrix (276 displayed rows)

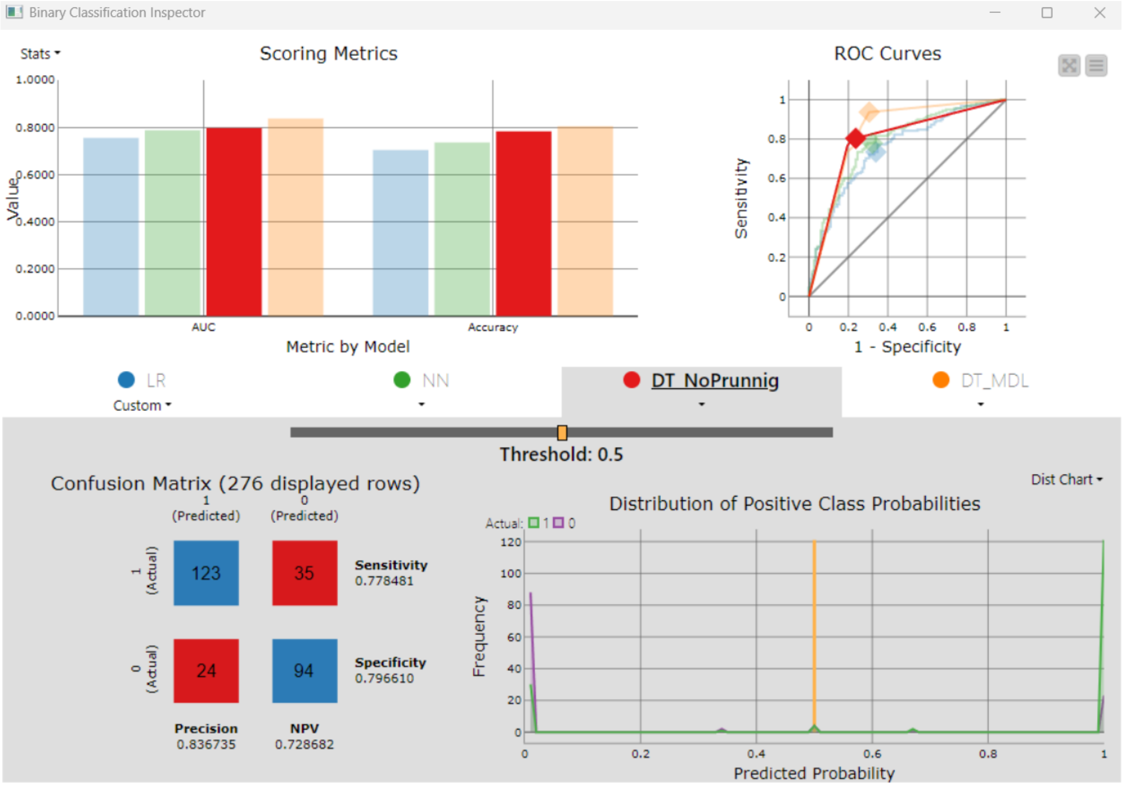
	1 (Predicted)	0 (Predicted)	
1 (Actual)	116	42	**Sensitivity** 0.734177
0 (Actual)	39	79	**Specificity** 0.669492
	Precision 0.748387	**NPV** 0.652893	



Threshold: 0.500

Confusion Matrix (276 displayed rows)

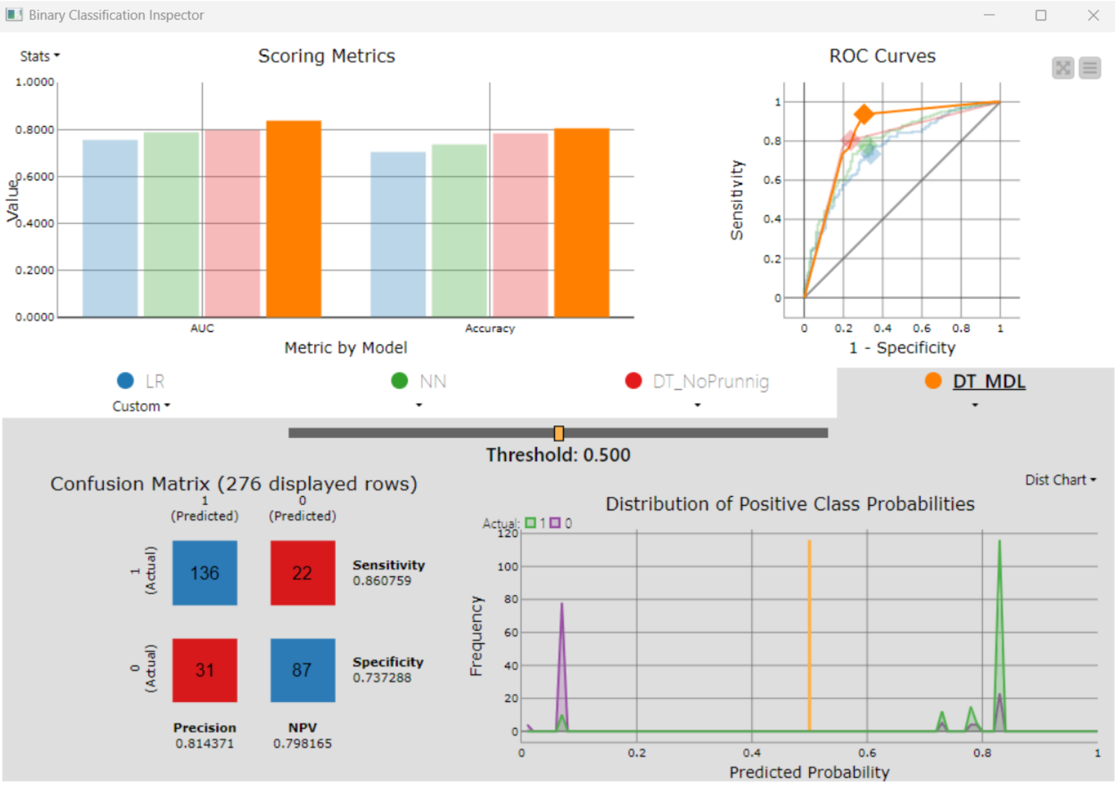
	1 (Predicted)	0 (Predicted)	
1 (Actual)	123	35	**Sensitivity** 0.778481
0 (Actual)	37	81	**Specificity** 0.686441
	Precision 0.768750	**NPV** 0.698276	



Threshold: 0.5

Confusion Matrix (276 displayed rows)

	1 (Predicted)	0 (Predicted)	
1 (Actual)	123	35	Sensitivity 0.778481
0 (Actual)	24	94	Specificity 0.796610
	Precision 0.836735	NPV 0.728682	



Threshold: 0.500

Confusion Matrix (276 displayed rows)

	1 (Predicted)	0 (Predicted)	
1 (Actual)	136	22	Sensitivity 0.860759
0 (Actual)	31	87	Specificity 0.737288
	Precision 0.814371	NPV 0.798165	

Interpretation

1. Logistic Regression (AUC = 0.755):

- Logistic regression performed moderately well in distinguishing between the two classes, with an AUC score of 0.755. This result reflects its capability to detect heart disease but is lower than the more advanced models.

2. **Neural Network (AUC = 0.787):**

- The neural network showed an improvement in AUC compared to logistic regression, achieving 0.787. This highlights the neural network's ability to capture non-linear relationships in the data, contributing to better classification performance.

3. **Decision Tree without Pruning (AUC = 0.797):**

- The unpruned decision tree achieved a slightly higher AUC of 0.797, indicating its strength in making flexible decisions, though it may still be prone to overfitting.

4. **Decision Tree with MDL Pruning (AUC = 0.837):**

- The pruned decision tree with Minimum Description Length (MDL) achieved the highest AUC of 0.837. Pruning effectively reduced overfitting while maintaining strong predictive capabilities, making it the best-performing model among the ones evaluated.

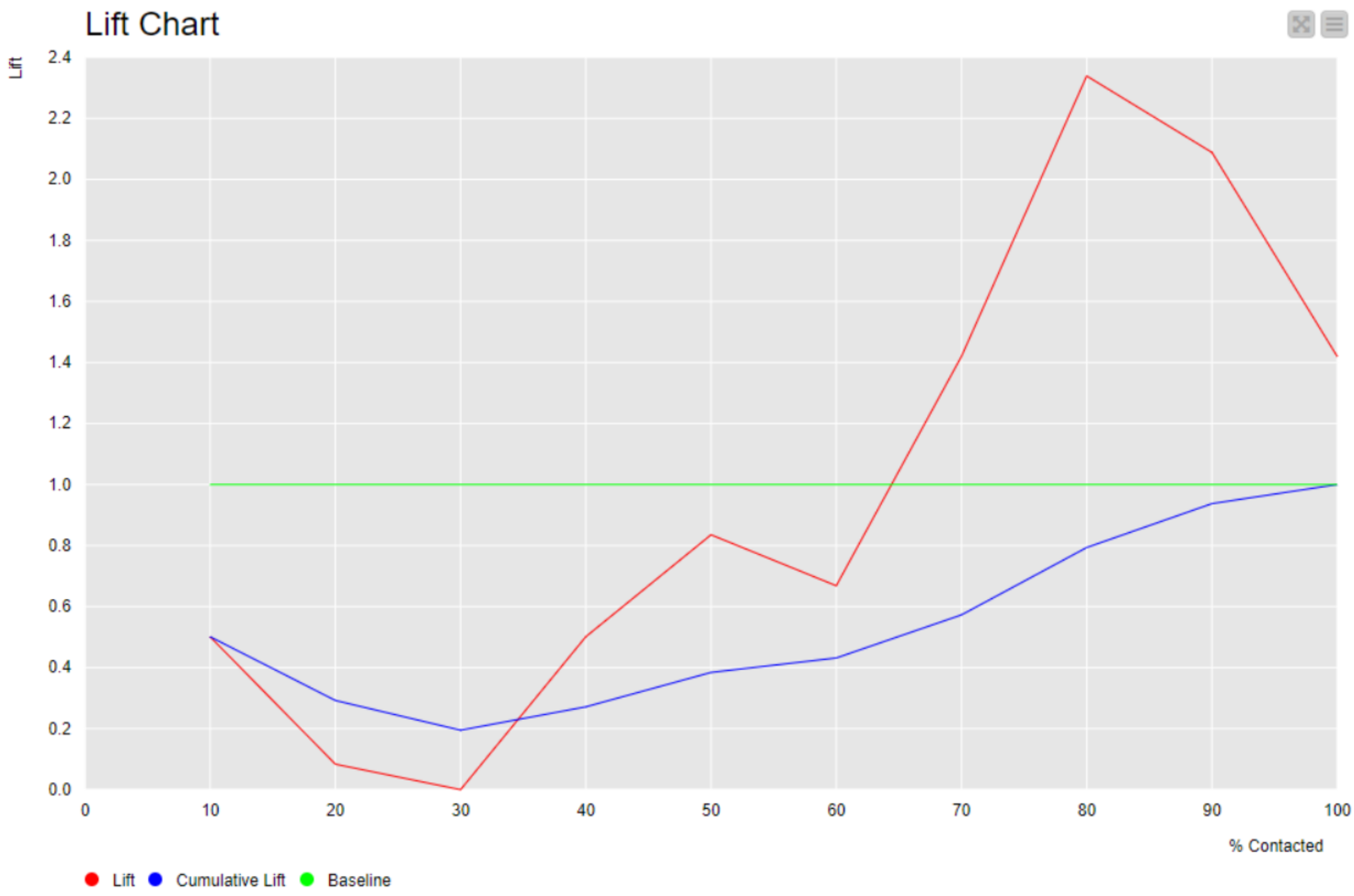
Accuracy Scores:

- **Logistic Regression (LR):** 70.652%
- **Neural Network (NN):** 73.913%
- **Decision Tree (No Pruning):** 78.623%
- **Decision Tree (MDL Pruning):** 83.333%

Recommendation:

Decision Tree with MDL Pruning (83.333%) stands out as the best-performing model based on both accuracy (83.333%) and AUC (0.837). Here's why it should be recommended:

1. **High Accuracy:**
   With an accuracy of 83.333%, this model consistently predicts true positives and true negatives better than all other models.
2. **Strong AUC (0.837):**
   The AUC score of 0.837 shows that this model does an excellent job distinguishing between the positive and negative classes, with better performance than the neural network (0.787) and logistic regression (0.755).
3. **Pruning Reduces Overfitting:**
   The MDL pruning has reduced the model's tendency to overfit, which is crucial for improving its generalizability and performance on unseen data. This makes it a more robust choice compared to the unpruned decision tree.



LIFT CHART OF DT (MDL)

4. **Lift Chart Performance:**
   The pruned decision tree shows strong performance in identifying high-risk cases, particularly in the top-ranked deciles (with lift reaching 2.4), but could be less effective in accurately classifying individuals in the lower-ranked deciles (as shown by the cumulative lift not crossing the baseline). This indicates potential for targeted interventions but also suggests areas for improvement in the model's overall coverage of all risk categories.

Conclusions

Learning from the Analysis:

1. **Model Performance Comparison:**

The analysis revealed the relative performance of four different predictive models: Logistic Regression (LR), Neural Network (NN), Decision Tree (No Pruning), and Decision Tree with MDL Pruning.

- Logistic Regression and Neural Networks performed well but fell short in comparison to decision tree models, particularly in handling non-linear relationships in the data.
- Unpruned Decision Tree showed significant improvement over the baseline models, but it suffered from overfitting, reducing its ability to generalize well on new data.
- The Pruned Decision Tree (MDL) emerged as the best-performing model, achieving the highest accuracy (83.333%) and AUC (0.837), indicating that it was the most reliable in correctly classifying both positive and negative cases.

Did We Meet Our Stated Business Objectives?

Yes, the analysis met its stated business objectives:

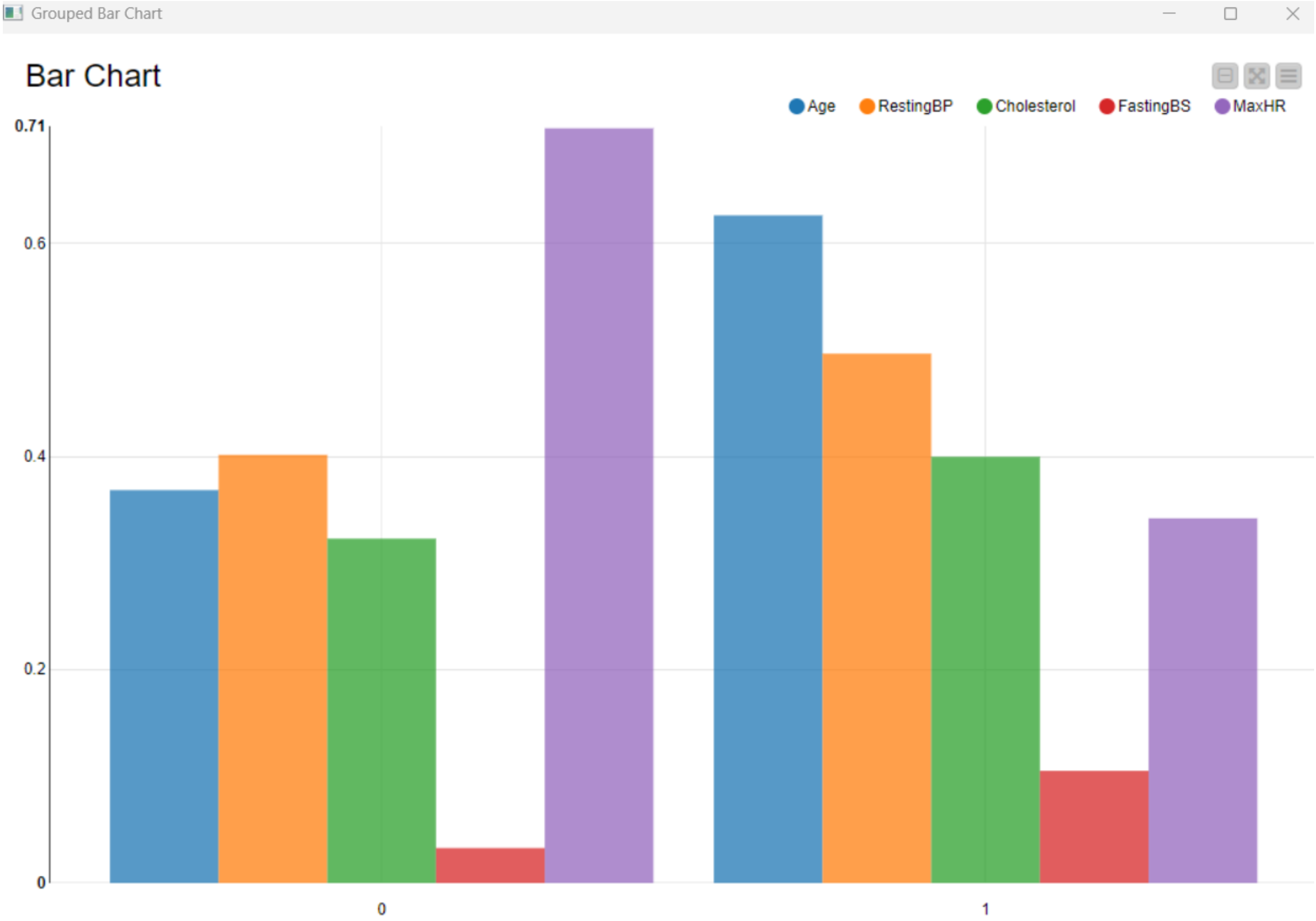
- The **business problem** was to predict heart disease, identify high-risk individuals, and potentially guide interventions.

- The **business objective** was to select the best model for this prediction task that maximizes accuracy and generalization while providing insights into high-risk individuals for targeted interventions.

- Based on the **pruned decision tree's performance** (high accuracy and strong lift in the top deciles), we can confidently say that the objectives were met. The model identifies individuals at high risk, which is useful for health interventions such as early diagnosis and preventative care.

How the Results Address the Business Problem/Opportunity:

- **High Risk Identification**:
  The pruned decision tree model can be leveraged to identify individuals at high risk of heart disease based on various features such as age, cholesterol levels, and exercise habits. By focusing on the top-ranked predictions (high-risk individuals), healthcare providers can intervene early, which could result in better outcomes for patients and lower healthcare costs.



The model shows the factors majorly contributing to the heart failure

- **Decision-Making Support**:
  The model provides actionable insights that can support healthcare practitioners in making data-driven decisions for heart disease diagnosis and treatment planning. By targeting the highest-risk individuals first, resources can be allocated more effectively.

Further Analyses and Future Work:

1. **Feature Engineering**:

- **New Features**: More predictive power could be gained by introducing additional features such as lifestyle factors (e.g., smoking, diet) or genetic data (if available). Incorporating more granular data can enhance the model's ability to predict heart disease risk more accurately.
- **Feature Selection**: A deeper analysis could be done to explore which features have the most significant impact on the predictions and whether certain irrelevant or redundant features can be removed to improve model performance.

2. **Model Tuning**:

- Further **hyperparameter tuning** of the decision tree model (such as adjusting the max depth or the minimum samples per leaf) could improve performance further.

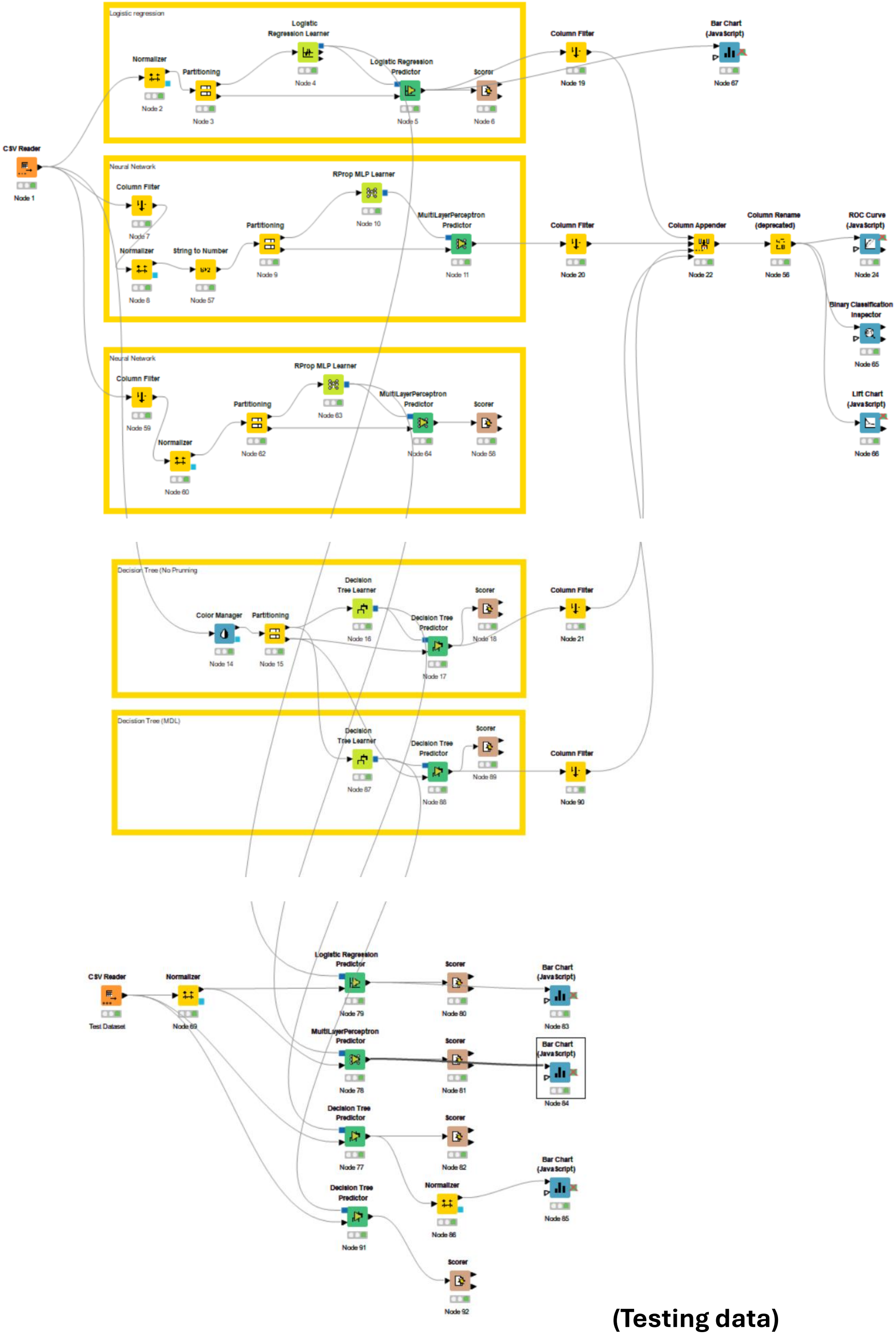
- Trying different pruning techniques or ensemble methods like **Random Forests** or **Gradient Boosting Machines** could yield even better results, especially in terms of reducing overfitting and increasing accuracy.

Final Recommendation:

Given the analysis, the **pruned decision tree (MDL)** model is the most accurate and robust choice for the current problem. It should be the model of choice for predicting heart disease risk and identifying high-risk individuals for targeted interventions. Future work should focus on enhancing the model with better features, more sophisticated tuning, and addressing any class imbalance, with an emphasis on deploying it effectively in real-world settings for predictive healthcare.

APPENDIX -

OVERALL KNIME WORKFLOW –



(Testing data)