1. **Executive Summary**

This report outlines the outcomes of a comprehensive data mining analysis conducted to predict the risk of heart attacks.The main goal was to improve the accuracy of heart attack risk assessments through advanced machine learning models, promoting proactive healthcare measures. Various methods were explored, including Gradient Boosted Trees, Random Forests, Neural Networks, Support Vector Machines, k-Nearest Neighbors, and Logistic Regression. After careful feature engineering, model selection, and hyperparameter tuning, Gradient Boosted Trees and Random Forests were identified as the top-performing models.Feature importance analysis revealed that age, cholesterol levels, and blood pressure were crucial factors influencing the accuracy of these predictions.

This executive overview captures the journey from defining the problem to selecting models and analyzing feature importance. The application of sophisticated data mining techniques shows promise in transforming heart attack risk prediction, contributing to a shift in preventive healthcare practices.

**2. Problem Description**

**2.1 Background**

Heart attacks, or myocardial infarctions, are a leading cause of mortality worldwide. Timely identification of individuals at risk of heart attacks is crucial for implementing preventive measures and improving overall health outcomes.

Despite advancements in medical science, the traditional approaches to heart attack risk assessment have relied on simplistic risk factor analysis and statistical models. These conventional methods, while providing valuable insights, often fall short in accurately capturing the complex interplay of factors that contribute to an individual's susceptibility to a heart attack.

**2.2 The Limitations of Traditional Approaches**

In the past, evaluating risk relied on factors like age, gender, cholesterol levels, and family history. Although these aspects are important, traditional models faced limitations in predictive accuracy due to their inability to account for the complex and interconnected nature of various risk elements.

**2.3 The Need for Advanced Predictive Models**

Identifying the limitations in conventional methods, this research aimed to utilize advanced data mining techniques to construct models with the ability to provide more precise and detailed predictions of the risk of heart attacks. Through the incorporation of a wide range of features and the application of machine learning algorithms, our goal was to surpass the constraints of traditional models and open the door to a more refined and individualized approach to assessing the risk of heart attacks.

**3. Brief Description of Data**

**3.1 Data Source**

The dataset used for this analysis was obtained from Kaggle. The dataset, named "Heart Attack Prediction," is one among the many health-related datasets accessible on Kaggle, aligning with the platform's goal of promoting data-driven advancements across diverse fields.

**3.2 Key Characteristics**

The dataset used in this analysis consists of 8763 entries, each corresponding to an individual patient. It encompasses 25 features that offer a thorough representation of diverse aspects related to health and lifestyle. These features encompass a mix of categorical and numerical variables, incorporating demographic details, medical history, lifestyle preferences, and physiological markers.

**3.3 Data Preprocessing**

Prior to the analysis, an extensive data preprocessing stage was undertaken. Notably, the dataset exhibited a commendable level of cleanliness, with no missing values requiring imputation. The dataset includes categorical variables like "Sex," "Country," and "Hemisphere," among others. Several graphs can be explored to gather insights, and a few examples are presented below..

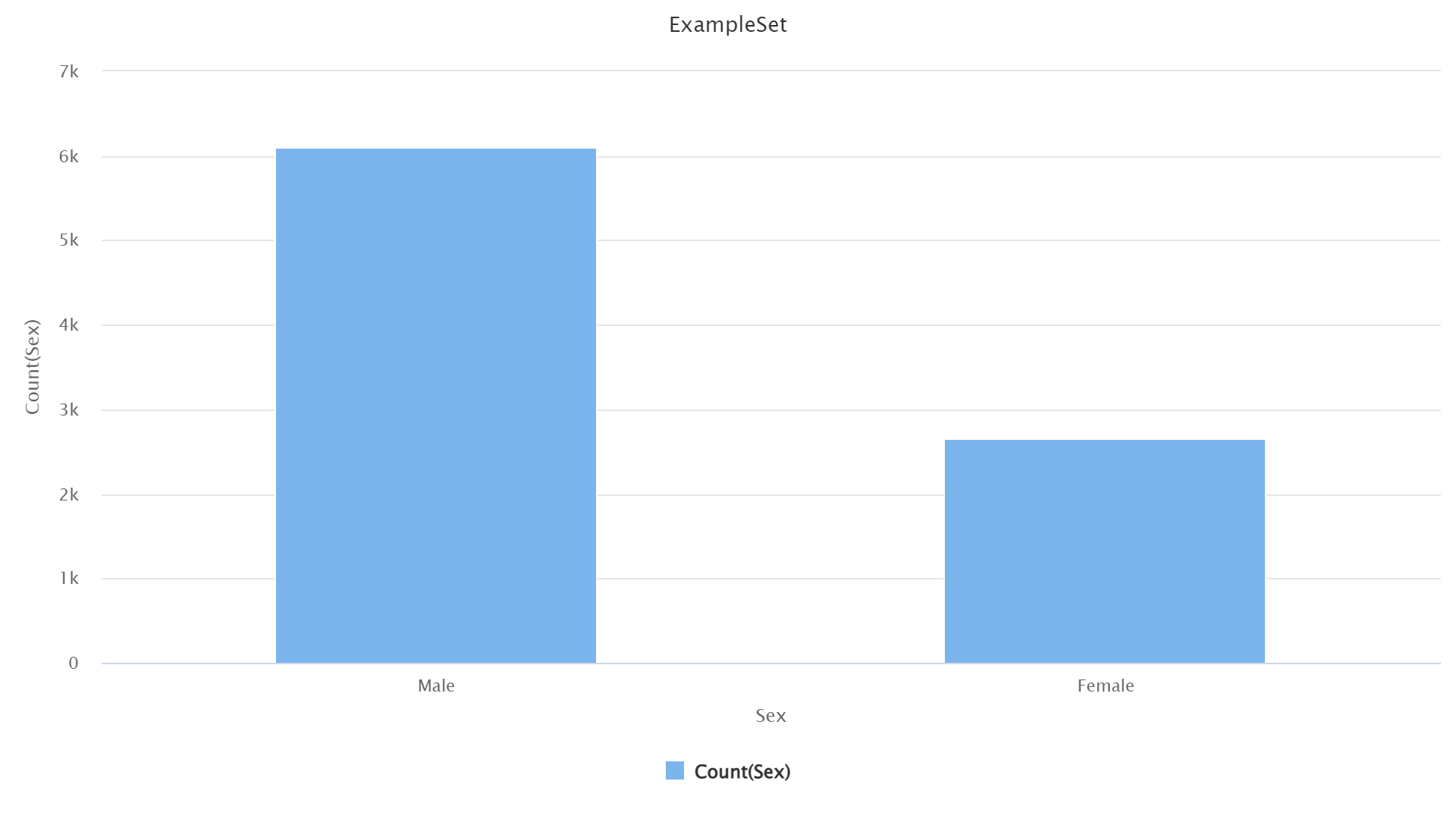


Fig: 3.3.1 This bar chart shows the count of males and females in the given data. Where Male count is nearly 3 times greater than the female data.

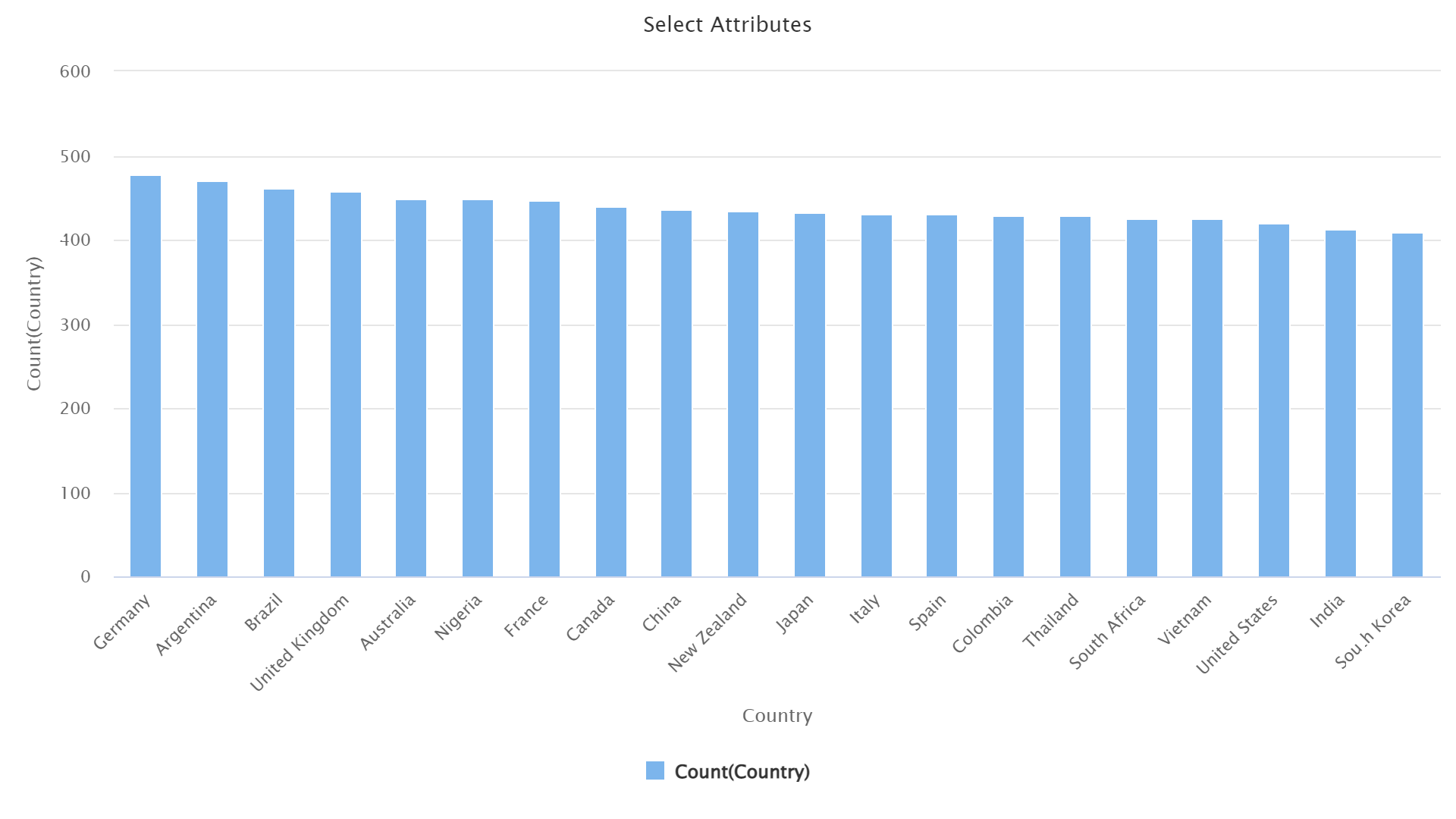


Fig: 3.3.2 This bar chart represents that the patients are from different parts of the world. Which indicates that we are dealing with diversified data.

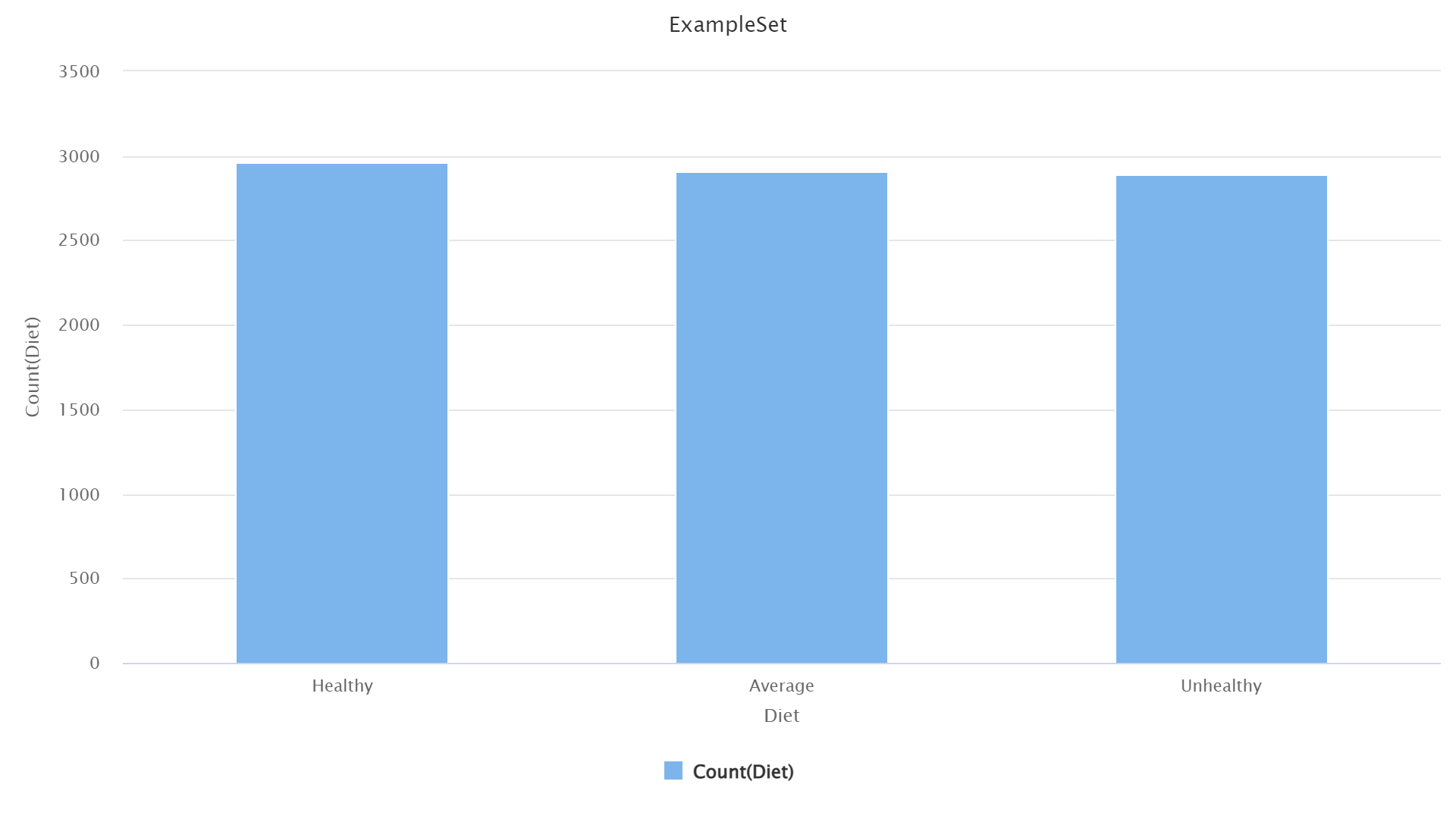


Fig: 3.3.3 Here we can observe that data is evenly spread across different classes which ensures that the model does not overfit.

**3.4 Feature Engineering**

To improve the predictive capabilities of the models, we employed feature engineering techniques. This involved utilizing a Correlation Heat-Map, generated using the correlation matrix operator in RapidMiner. to identify high correlations between two or more features. If such correlations were present, we opted to retain only one of the correlated features. However, in our dataset, this situation did not arise.

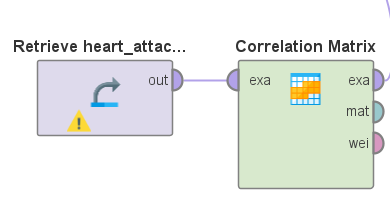


Fig: 3.4.1 Correlation Matrix operator.

Additionally, we found it necessary to eliminate certain features like "Continent," "Country," "Hemisphere," "Income," and "Patient ID" to streamline the dataset and concentrate on the most

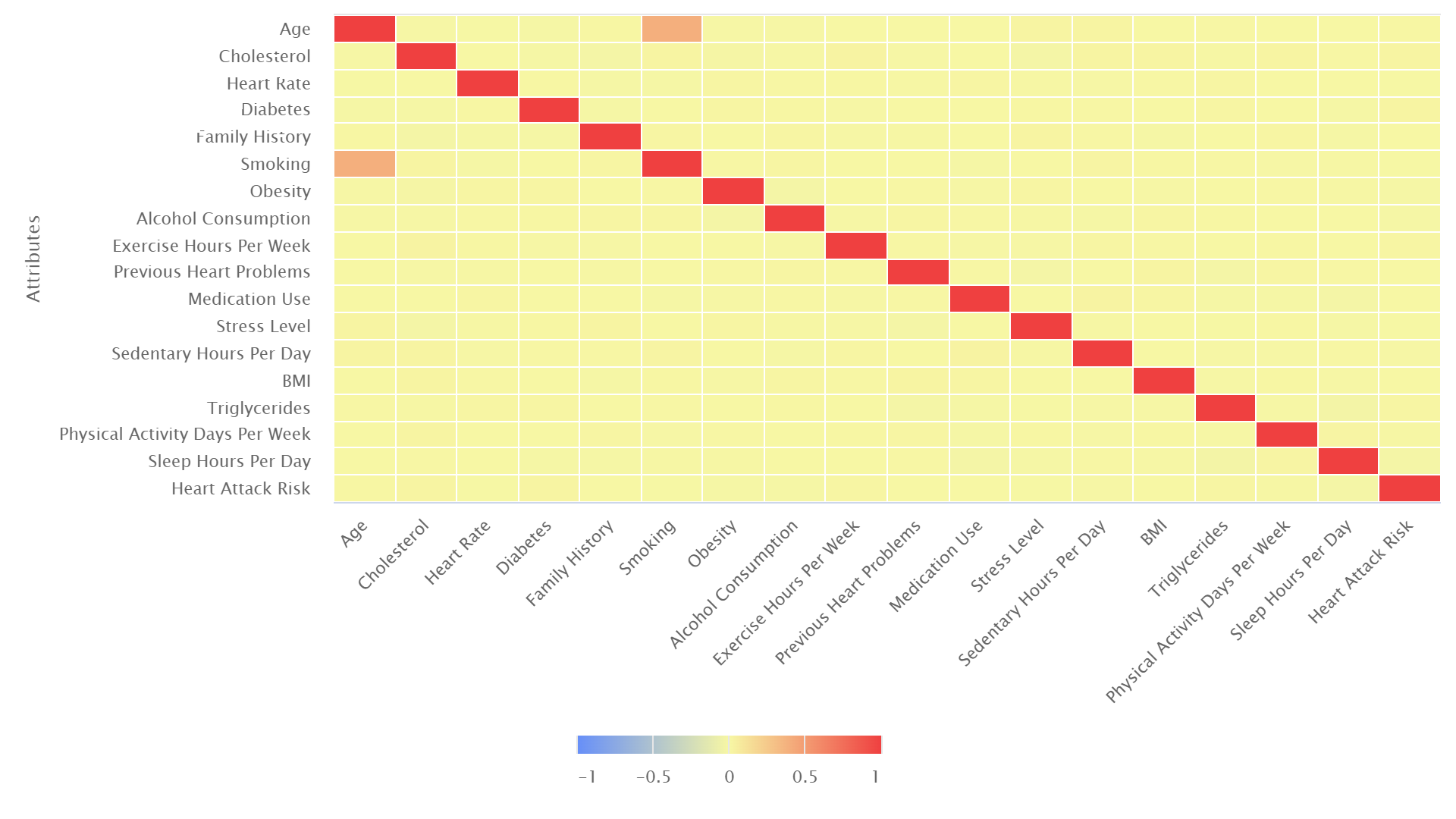


Fig 3.4.2 Correlation Heat-Map (Red shade rectangle holds 0.3 value, which can be neglected)

influential predictors. Analyzing the correlation-weighted graph of features helped us identify the variables that are significant for training purposes.

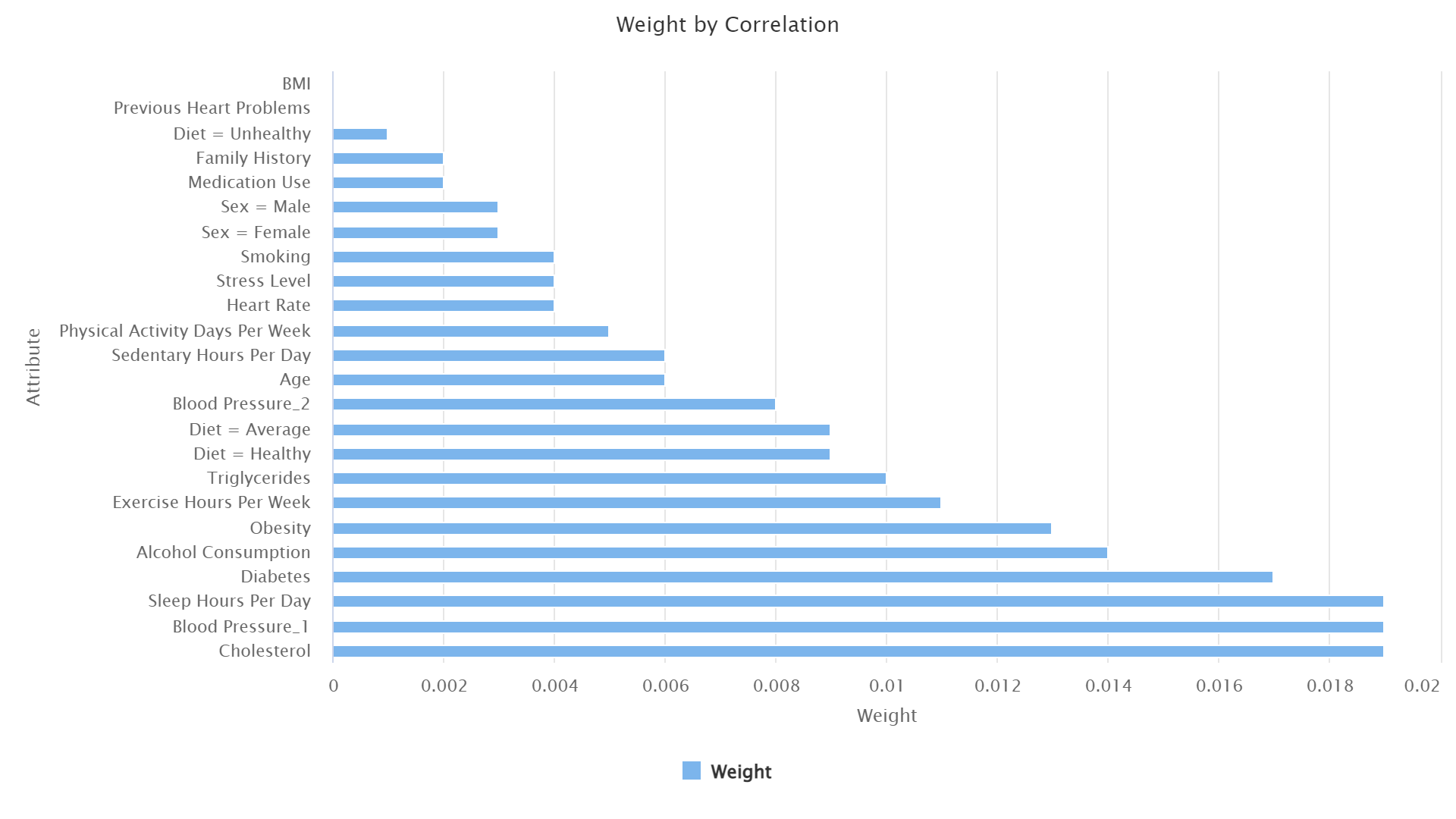


Fig 3.4.2 Here the label feature has more weightage on cholesterol, Blood pressure and few other.

Categorical values underwent conversion into numerical forms using the Nominal to Numerical operator, ensuring their compatibility with machine learning algorithms.Systolic and Diastolic values of Blood Pressure column are separated into two columns by using split operator.

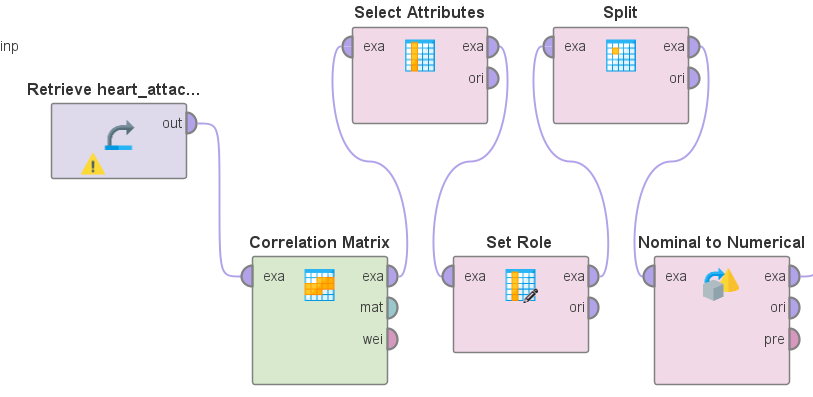


Fig: 3.4.3 The operators that have been used for feature engineering.

**4. Findings from Data Mining Models**

**4.1 Model Performance Overview**

Different machine learning approaches were employed to determine the optimal model for predicting the risk of heart attacks, with evaluation conducted using three metrics: Root Mean Square Error (RMSE), Absolute Error, and Relative Error. The performance scores are summarized as follows:

| Model | RMSE | Absolute Error | Relative Error |
| --- | --- | --- | --- |
| DNN | 0.494 | 0.467 | 61.29% +/- 11.81 |
| SVM | 1.127 | 0.903 | 95.08% +/- 72.03 |
| Logistic Regression | 0.481 | 0.460 | 64.37% +/- 2.72 |
| Gradient Boosted Tree | 0.481 | 0.461 | 64.35% +/- 2.00 |
| Random Forest | 0.482 | 0.461 | 64.35% +/- 3.57 |
| Random Forest(Hyper-parameter Tuning) | 0.476 | 0.457 | 63.43% +/- 3.03 |
| Neural Network | 0.508 | 0.464 | 65.95% +/- 15.95 |
| KNN | 0.522 | 0.455 | 64.16% +/- 22.03 |

In our model evaluation, the Deep Neural Network (DNN) showed strong performance in capturing complex relationships within the data. In contrast, the Support Vector Machine (SVM) provided valuable insights but had higher error rates compared to other models. Logistic Regression consistently performed well with low relative error. The Gradient Boosted Tree (GBT) performed similarly to Logistic Regression, and Random Forest demonstrated consistent accuracy, matching the performance of GBT. Fine-tuning the Random Forest model resulted in a slight performance improvement. The Neural Network, while competitive, displayed higher variability, and K-Nearest Neighbors (KNN) exhibited increased variability with a slightly higher relative error. Overall, Random Forest, especially after tuning, emerged as a promising option for accurate and reliable predictions of heart attack risk.

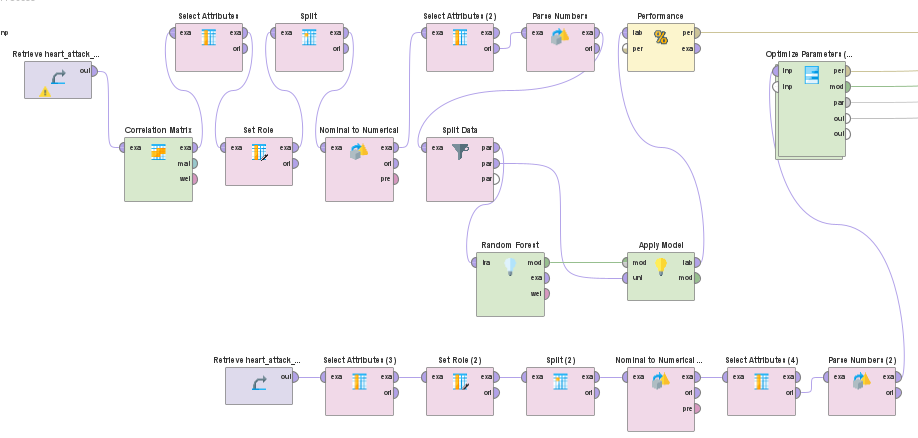


Fig: 4.1.1 The entire process of constructing and optimizing a Random Forest model using RapidMiner.

**4.2 Conclusion on Model Selection**

After a thorough examination, it is clear that both Random Forest and Gradient Boosted Tree models have proven to be highly effective, displaying competitive accuracy and consistency. The Random Forest model, following some fine-tuning of hyperparameters, exhibited a slight enhancement, positioning it as a robust candidate for predicting heart attack risks. On the other hand, the Gradient Boosted Tree model was computationally demanding. Despite each model having its distinct advantages and drawbacks, the results indicate that Random Forest, particularly with tuning, stands out as a promising option.

**5. Conclusion**

This thorough examination of various data mining models for predicting the risk of heart attacks has yielded valuable insights into their performance and applicability in real-world scenarios. The study encompassed a range of models, including Deep Neural Networks and traditional methods like Logistic Regression and Support Vector Machines. Notably, after fine-tuning, Random Forest and Gradient Boosted Tree emerged as top-performing models, demonstrating competitive accuracy and reliability. Logistic Regression consistently showed robust performance, highlighting its effectiveness in this domain. These results underscore the potential of advanced machine learning techniques in improving the accuracy of heart attack risk assessments, paving the way for more precise and personalized preventive healthcare strategies. The investigation into model performance, feature importance, and hyperparameter tuning establishes a solid foundation for future research and implementation in clinical settings, emphasizing the significance of data-driven approaches in healthcare decision-making.