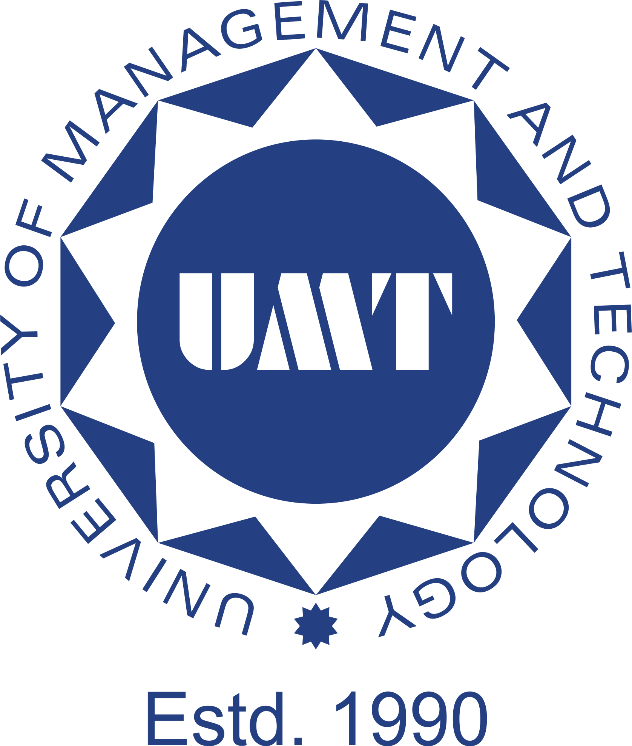
Final Project



*University of Management and Technology*

**Submitted by**:

M. Bilal Iqbal F2019054047

Syed Mohsin Kazmi F2019054048

Mugheera Khan F2019054013

**Date of Submission**:12th June 2023

**Course Module**: Data Warehousing & Mining

**Section:** A

**Prepared for:** Sir Abdul Ghafar

TABLE OF CONTENTS

[ABSTRACT: 2](#_Toc137605119)

[INTRODUCTION: 3](#_Toc137605120)

[FIVE PREVIOUS RESEARCH WORKING ON THE SAME DOMAIN WITH PROPER REFERENCES; 3](#_Toc137605121)

[EXPLAIN YOUR SOLUTION: 5](#_Toc137605122)

[RESULTS \ FINDINGS: 13](#_Toc137605123)

[CONCLUSION: 15](#_Toc137605124)

**CREATING A RELIABLE DRUG PREDICTION MODEL:**

**A DATA MINING APPROACH USING RAPID MINER**

# ABSTRACT:

The dataset is briefly described in this abstract, along with any potential implications for data analysts, academics, and medical experts. In this work, we use the well-known data mining program RapidMiner to create a medication prediction model. Age, sex, blood pressure, cholesterol, sodium-to-potassium ratio (Na\_to\_K), and the relevant prescribed medication are all included in the dataset. Our objective is to create a classification algorithm that can precisely predict whether a patient with particular blood pressure, NA to K ratio, and cholesterol level would receive a prescription for a medicine.

We will use a variety of RapidMiner data mining methods to do this. We will be able to investigate the dataset's patterns and linkages using these algorithms, and we'll be able to give healthcare professionals insights and suggestions based on the patterns, correlations, and risk ratings we find. We can evaluate which algorithm performs best for our particular prediction job by evaluating their results.

Preprocessing the data include addressing missing values and transforming categorical variables (like Drugs) into particular drug categories. This stage makes sure the data is appropriate for model training and assessment. The dataset will then be divided into training and testing sets so that the performance of our models can be assessed. Using the built-in operators of RapidMiner, we will construct popular classification methods including Decision Trees, Naive Bayes, Gradient Boosting, K-Nearest Neighbors, and Random Forests.

We will determine the method that offers the maximum accuracy in forecasting medication kinds for the healthcare industry through iterative testing and model review. On the basis of performance indicators like accuracy, we will rate the models.

We hope to create a reliable drug prediction model by the conclusion of the trial, which will allow us to identify possible patients who are more likely to take a medicine. Based on the patterns, connections, and risk ratings produced from the information, this model can offer hospitals useful insights and suggestions.

**Keywords:** correlations, sodium-to-potassium ratio, medication prediction model and accuracy.

# INTRODUCTION:

Introduction to the problem:

Following are problems that were identified in our data set;

* **Lack of data:** The dataset may not include enough instances or records to reliably identify patterns or produce insightful conclusions. For data mining tools to accurately detect important trends or patterns, a large volume of data is frequently needed.
* **Data imbalance:** The distribution of the target variable (if existent) may be biased towards one class in the dataset, making it difficult to develop precise prediction models. This may lead to models that are inaccurate or inefficient.
* **Data that is noisy or inconsistent:** The dataset may include items that are inconsistent, missing values, or outliers. These problems may result in erroneous or deceptive patterns and have a negative impact on the quality of the mining results.
* **Irrelevant or redundant features**: The dataset may contain characteristics that are duplicated or that are highly connected with one another yet are irrelevant to the mining process. The effectiveness and precision of the data mining process may be increased by recognizing and removing such traits.
* **Data preprocessing challenges:** Before the dataset can be mined efficiently, it may need to undergo extensive preparation. This covers, among other preprocessing chores, resolving missing values, dealing with outliers, normalizing or scaling features, and handling categorical variables. In accordance with the advice of our resource person, we must rename the medication variables.
* **Bias in the data**: The dataset could be skewed, which means it might not be representative of the general population or might have biases built in from the data gathering procedure. Biased models and unjust or discriminatory outcomes might result from biased data.

# FIVE PREVIOUS RESEARCH WORKING ON THE SAME DOMAIN WITH PROPER REFERENCES;

1. **Research-1: (Implementation of Machine Learning Model to Predict Heart Failure Disease)**

Approximately 26 million people are affected by high blood pressure and other cardiac conditions each year. Medical professionals can benefit from classification and prediction algorithms even if it is challenging to forecast Heart problems correctly. The results of this study show improved ratings for heart disease prediction accuracy. The machine learning model may be connected with medical information systems to aid in anticipating cardiac issues using real-time patient data.

1. **Research-2: (Scalable Predictive Analysis in Critically Ill Patients Using a Visual Open Data Analysis Platform)**

Predictive analytics, which may transform reactive medicine into predictive, preventative, and personalized care, can have an influence on the cost and quality of treatment. Though it is challenging to swiftly transform the data into models with therapeutic significance due to their high dimensionality and complexity. This study attempts to address this issue by emphasizing open, visible environments suitable for medical use. A framework for the effective utilization of data from critical care patients, database integration, and predictive modeling was developed using RapidMiner. Scalable predictive analytics in health research are intriguing due to these trustworthy methods.

1. **Research-3: (Compare and Contrast of Machine Learning Classification Algorithms to Predict Accuracy and Performance of Lung Cancer Disease)**

Lung cancer is a severe disease that affects breathing and is carried on by unchecked cell growth. Medical enterprises provide cutting-edge facilities and potent treatments to combat illness. Machine learning and artificial intelligence are essential in early sickness prediction and data analysis. In this work, lung cancer datasets and diseases are categorized using machine learning classification techniques, which transform the data into binary form for accurate and effective diagnosis.

1. **Research-4: (Using Machine Learning classification methods to detect COVID-19 Status in Selangor District)**

Naive Bayes Classifiers are often used for prediction, even if ML techniques are not yet widely applied. This study focuses on predicting Covid-19 cases using four Rapid Miner algorithms. A decision tree's accuracy rating of 75.47% is good. More study is needed to overcome bias and unequal datasets and incorporate more risk factors like vaccination status.

1. **Research-5: (Data Mining for Cardiovascular Disease Prediction)**

Worldwide, cardiovascular diseases (CVDs) are major contributors to incapacity and early mortality. Clinical data and developments in computational intelligence have made it possible to create recognition algorithms that can find hidden patterns and important information. Rapid Miner's Random Forest Classification is used in conjunction with Data Mining Techniques (DMTs) to predict CVD in medical examination data. The most accurate, precise, sensitive, and particular model was the one that had been optimized.

# EXPLAIN YOUR SOLUTION:

Introduction of Our Data Set:

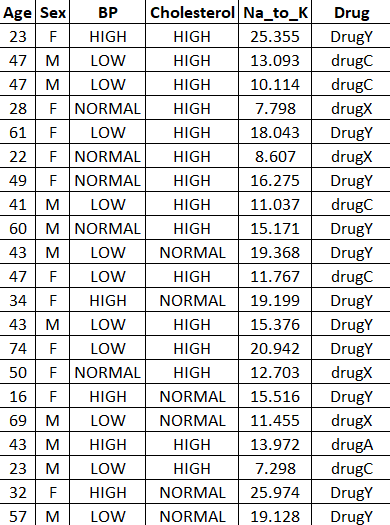
Our dataset appears to contain data on a variety of individual characteristics, including their age, sex, blood pressure, cholesterol, sodium-to-potassium ratio, and the specific medication that was provided to them. It appears to be a dataset of medical or healthcare information.

Here is a quick overview of the dataset:

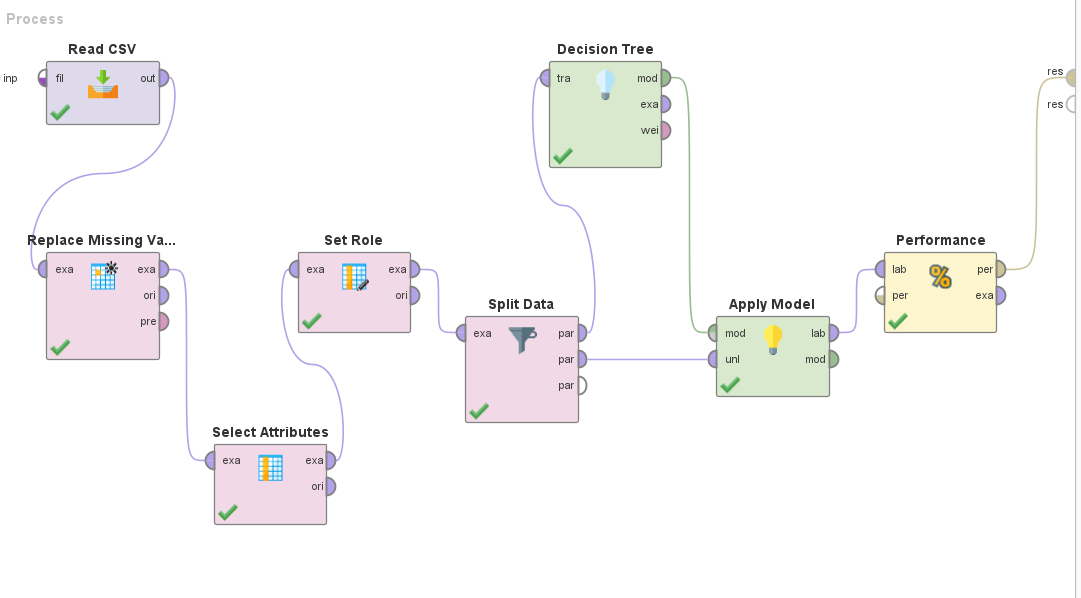
The dataset is made up of a number of entries that represent people and the qualities that go with them. Age, sex, blood pressure (classified as LOW, NORMAL, or HIGH), cholesterol level (classified as LOW, NORMAL, or HIGH), the sodium-to-potassium ratio in the body, and medication prescribed to an individual are all included in each record.

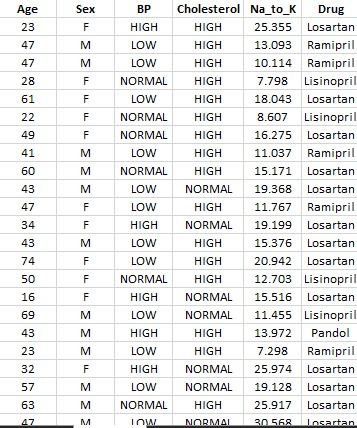
The dataset appears to be geared at examining the connection between various variables and the prescription of particular medications. Based on an individual's features, such as age, sex, blood pressure, cholesterol level, and sodium-to-potassium ratio, the qualities presented may be used to analyze trends and correlations to predict which prescription is frequently recommended.

Researchers or analysts can learn more about the factors affecting medication prescriptions by using data mining techniques to explore this information, and they may be able to anticipate or suggest strategies for similar situations in the future.



**Usage of data preprocessing techniques:**

****

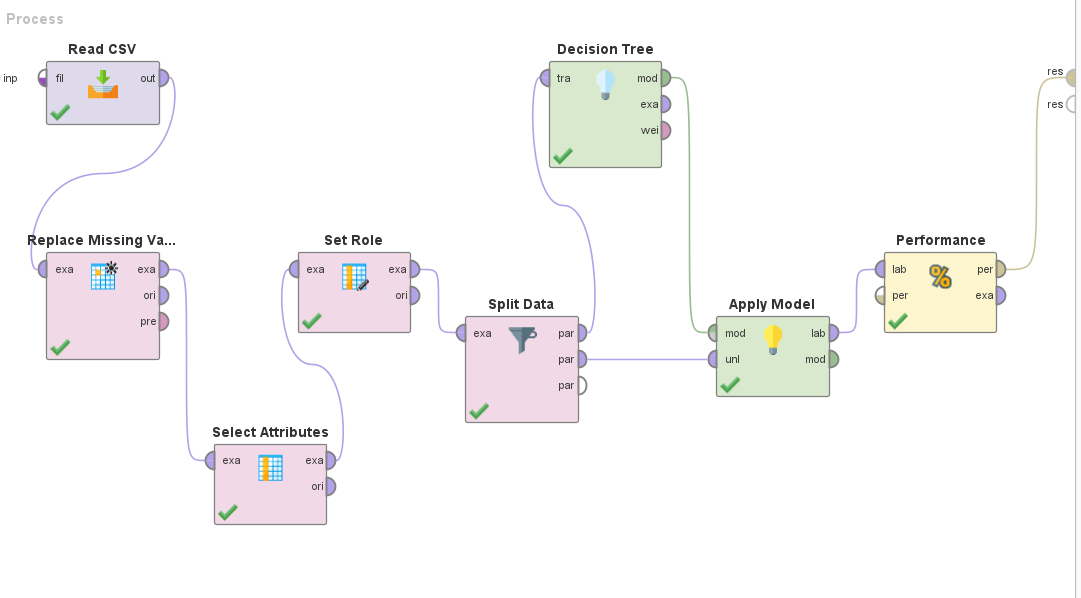


**Data mining algorithms:**

The data mining algorithms were used as follows with implementation and explanation;

1. **Decision Tree**

**Implementation:**

****

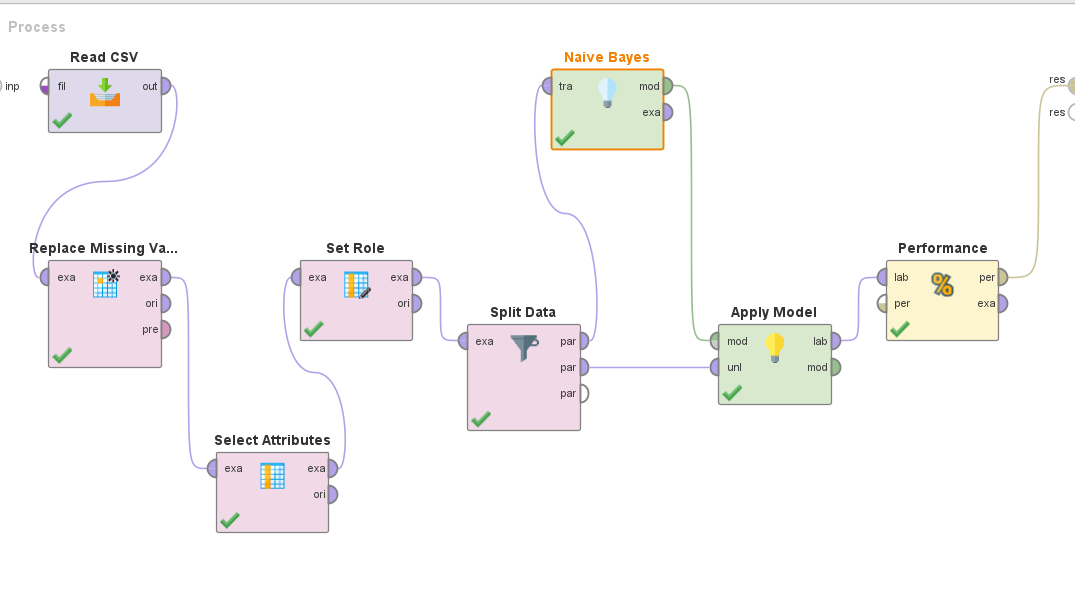
**Explanation:**

A well-liked machine learning approach for classification and regression applications is decision trees. The decision tree model's accuracy in relation to accuracy model was 82.86%.

A decision tree is a flowchart-like structure where each leaf node represents the result or class label, each internal node represents a characteristic or attribute, and each branch represents a decision rule. Recursively dividing the data based on the chosen characteristics up until a halting requirement is satisfied results in the construction of the tree.  
By contrasting the projected class labels with the actual class labels from a test dataset, a decision tree model's accuracy is assessed. The decision tree model in this scenario has an accuracy of 82.86%, which implies that it correctly identified the class labels for about 82.86% of the test dataset instances.

It's important to remember that a model's accuracy might not be sufficient to evaluate its performance in its entirety. In order to get a more thorough knowledge of the model's predictive capabilities, it is frequently advantageous to take into account other assessment metrics including accuracy, recall, and F1 score.

1. **Naïve Bayes:  
     
   Implementation:**

****

**Explanation:**

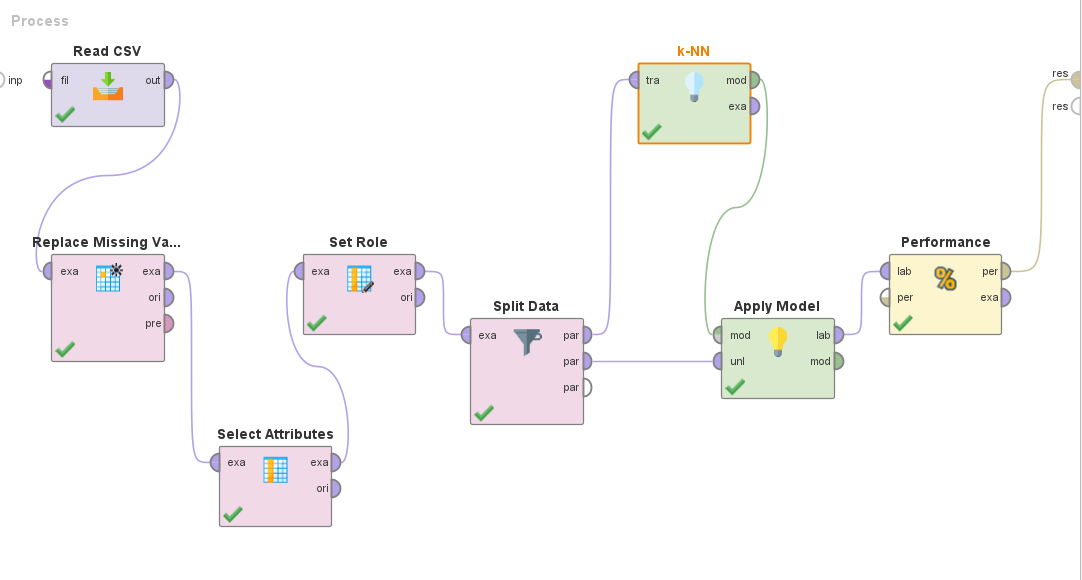
The Naive Bayes method has a 92.14% accuracy rate. Popular machine learning algorithms like Naive Bayes are utilized for classification tasks, especially in text categorization and spam filtering.

Naive Bayes is based on the Bayes theorem, which determines the likelihood of an occurrence based on information already known. The algorithm operates under the naive assumption that all characteristics are independent of one another. Naive Bayes can be computationally effective and frequently performs well in practice despite this oversimplifying assumption.

By contrasting the predicted class labels with the actual class labels from a test dataset, a Naive Bayes model's accuracy is assessed. Naive Bayes model accurately predicted the class labels for around 92.14% of the test dataset instances, according to its accuracy score of 92.14%.

Naive Bayes is renowned for being straightforward, effective, and capable of handling large amounts of data. When the independence assumption is severely broken or when there aren't enough training data, it could not perform as well.

1. **KNN:  
   Implementation:**

****

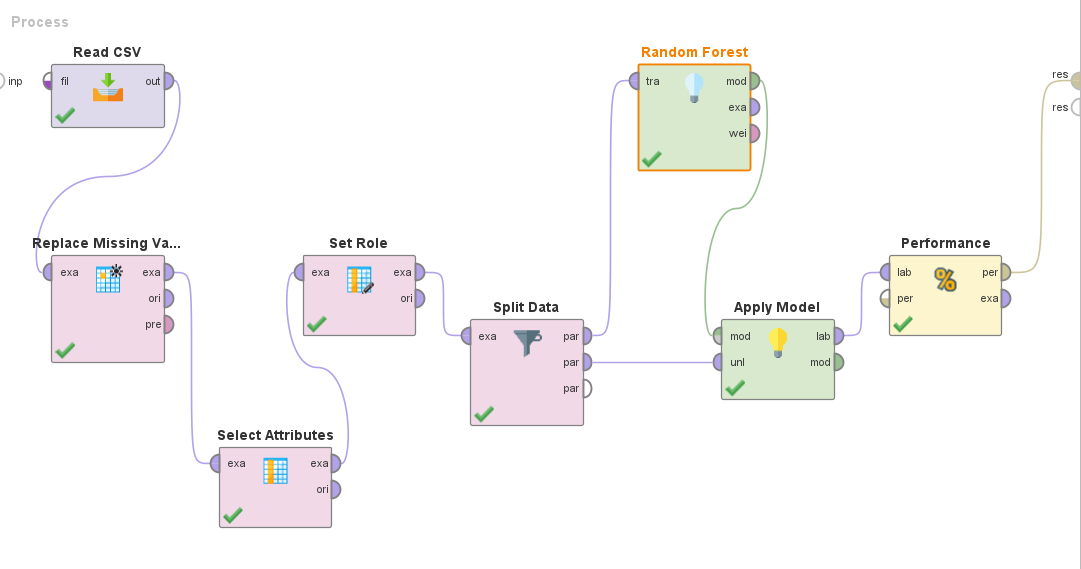
**Explanation:**

The KNN algorithm obtained an accuracy of 57.14% in this accuracy model. The machine learning technique known as KNN, or k-nearest neighbors, is non-parametric and instance-based and is used for classification and regression applications.

By contrasting the predicted class labels with the actual class labels from a test dataset, a KNN model's accuracy is assessed. KNN model accurately predicted the class labels for around 57.14% of the test dataset instances, according to its accuracy score of 57.14%.

KNN is renowned for being straightforward and simple to use. However, the choice of k and the distance measure can have an impact on how well it performs. The training dataset should be sizable and reflective of the underlying data distribution for it to be most successful.

1. **Random Forest:  
   Implementation:**

****

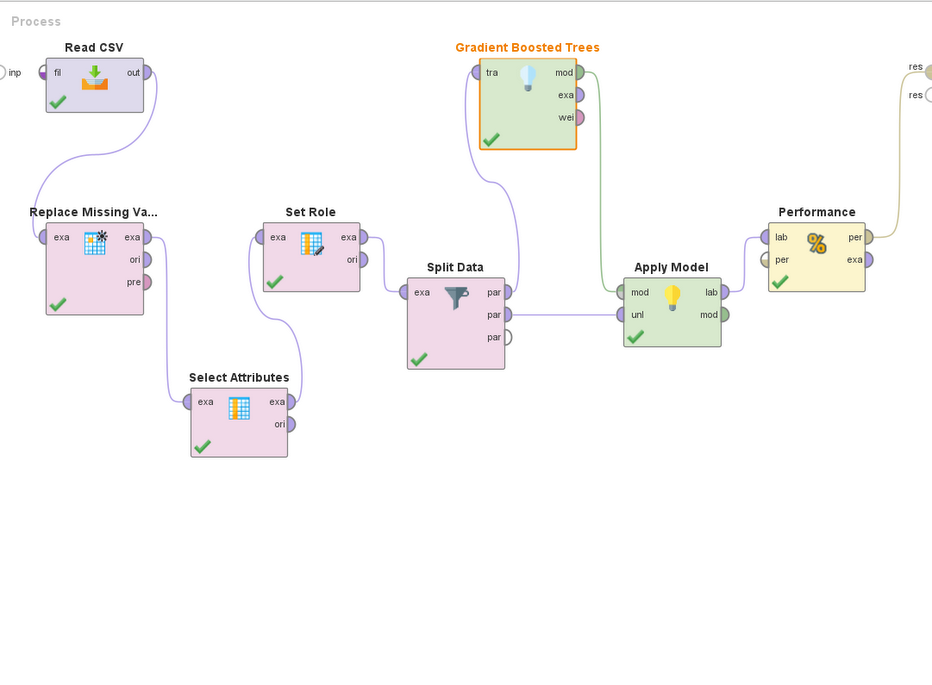
**Explanation:**

The accuracy of the Random Forest method was 97.86%. An ensemble learning technique called Random Forest uses many decision trees to provide predictions. It is commonly used in machine learning for both classification and regression problems.

The Random Forest technique is renowned for its capacity to manage big datasets, handle high-dimensional data, and prevent overfitting. It can produce reliable and accurate forecasts by fusing the predictions of various decision trees. The Random Forest's individual decision trees provide diversity, and the voting mechanism aids in lowering variance and raising overall accuracy.

The Random Forest model accurately predicted the class labels for around 97.86% of the cases in the test dataset in this scenario, achieving an accuracy of 97.86%.

1. **Gradient Boosting:  
     
   Implementation:**



**Explanation:**

The accuracy of the Gradient Boosting method was 91.43%. Gradient Boosting is a technique used in ensemble learning that combines a number of weak prediction models (usually decision trees) to produce a strong predictive model.

Gradient boosting is renowned for its capacity to manage diverse data sources and intricate interactions between features and target variables. When tackling regression and classification issues where high accuracy is sought, it is very successful.

The accuracy of the Gradient Boosting model was 91.43%, meaning that it correctly identified the class labels for around 91.43% of the test dataset instances.

# RESULTS \ FINDINGS:

In our research, we were able to get data that showed the accuracy levels attained by several machine learning models when used on the provided dataset. An explanation of each model's accuracy is provided below:

**KNN (K-Nearest Neighbors) - 57.14% accuracy:**

KNN is a straightforward algorithm that divides data points into groups according to the dominant class among their close neighbors. The accuracy of the KNN model in this example was 57.14%, which implies that for around 57.14% of the dataset's cases, the medication was accurately predicted.

**Naïve Bayes - 92.14% accuracy:**

The Nave Bayes algorithm is a probabilistic classifier that relies on the Bayes theorem and the condition of feature independence. The Naive Bayes model has a 92.14% accuracy rate, which means that it correctly predicted the medicine in 92.14% of the cases.

**Decision Tree - 82.86% accuracy:**

A hierarchical structure known as a decision tree makes judgments based on the values of several attributes. The Decision Tree model has an accuracy of 82.86%, which means it categorized the medication correctly in about 82.86% of cases.

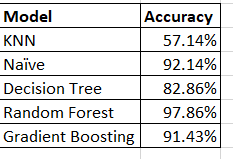
**Random Forest - 97.86% accuracy:**

An ensemble learning technique called Random Forest uses many decision trees to provide predictions. The Random Forest model successfully predicted the medicine in 97.86% of the cases, achieving a remarkable accuracy of 97.86%.

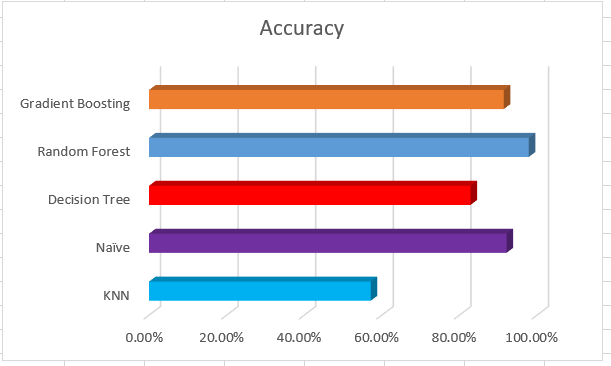
**Gradient Boosting - 91.43% accuracy:**

Another ensemble learning approach called gradient boosting combines weak prediction models to produce a stronger model. The accuracy of the Gradient Boosting model was 91.43%, indicating that it correctly identified the medication in about 91.43% of the cases.

The accuracy scores of our model are as follows;

****

**Visual representation of our results:**

****

These accuracy numbers show the effectiveness of each model on the provided dataset. Higher accuracy scores often imply higher predictive performance, but it's crucial to take other aspects into account when choosing the best model for a given job, such as model complexity, interpretability, and potential overfitting.

# CONCLUSION:

We've detailed a medical dataset in this work that includes information on a patient's age, sex, blood pressure, cholesterol level, sodium-to-potassium ratio (Na\_to\_K), and any prescription medications. With the purpose of precisely forecasting whether a patient with given blood pressure, Na\_to\_K ratio, and cholesterol levels will be prescribed a specific drug, our goal was to create a medication prediction model using the data mining program RapidMiner.

We sought to find patterns, correlations, and risk ratings within the dataset by utilizing several data mining techniques offered by RapidMiner. For healthcare practitioners, this analysis would offer insightful advice that would help them decide what medications to prescribe for specific patients based on their profiles.

We conducted preprocessing procedures, such as dealing with missing values and categorizing categorical variables (such as Drugs) into discrete drug categories, to get the dataset ready for model training and assessment. The dataset's appropriateness for modeling purposes was therefore guaranteed. Additionally, in order to appropriately evaluate the performance of our models, we split the dataset into training and testing sets.

We used well-known classification techniques such as Decision Trees, Naive Bayes, Gradient Boosting, K-Nearest Neighbor, and Random Forests using the built-in operators of RapidMiner. We identified the approach that provided the greatest accuracy in predicting medicine kinds for the healthcare business through iterative testing and model review. The models were evaluated using performance metrics, such as accuracy, to determine which was the best fit for the specific prediction job at hand.

This study's ultimate objective was to develop a trustworthy drug prediction algorithm that might pinpoint possible patients who are more likely to need a certain medication. This program would offer hospitals insightful information and recommendations by extracting patterns, linkages, and risk assessments from the data. These discoveries may help medical practitioners choose drug prescriptions wisely, resulting in better patient outcomes and individualized healthcare treatments.

In summary, this research offers a thorough procedure for creating a medicine prediction model from a medical dataset. In-depth preprocessing, model validation, and the usage of rapid miners’ data mining algorithms all work together to produce predictions and insights that are both accurate and insightful for the healthcare sector. The effective creation of such a model might revolutionize medicine prescription and lead to more efficient and individualized patient care. Moreover, the study mentioned in the assertion exemplifies a methodical approach to creating a medical prediction model. The project seeks to deliver precise predictions and useful insights that can revolutionize drug prescription and enhance patient care in the healthcare industry by combining meticulous preprocessing, model validation, and the use of data mining methods.