**Drug Classification Report Using ML**

Contents

[**Introduction** 2](#_Toc193885297)

[**Data Exploratory Analysis** 2](#_Toc193885298)

[**Features Exploratory Analysis** 3](#_Toc193885299)

[**Data Cleaning** 5](#_Toc193885300)

[**Prediction Model** 5](#_Toc193885301)

[**Base SVM Model** 5](#_Toc193885302)

[**Random Forest Model (Advanced)** 5](#_Toc193885303)

[**Conclusion** 6](#_Toc193885304)

# **Introduction**

This work employs machine learning methods to classify medication types. The main goal is to create a prediction model able to precisely identify the suitable drug type depending on patient characteristics. Features including age, sex, blood pressure (BP), cholesterol, and the sodium-to---potassium ratio make up the dataset. Target variable is multiclass, denoting several kinds of drugs. Particularly for pharmaceutical corporations, hospitals, and clinical practitioners, the capacity to properly classify medications has major consequences in the healthcare sector. The key business problems this project aims to solve include:

* Enhancing the efficiency of medication assignment based on patient characteristics.
* Minimizing the need for manual selection of drugs, thereby reducing medical errors and improving patient safety.
* Helping pharmacies and hospitals optimize their drug stock by predicting demand for specific drug types.
* Facilitating faster and more accurate treatment decisions, leading to better patient outcomes.

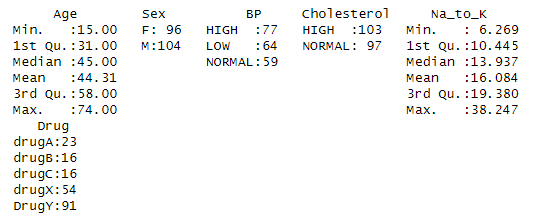
Machine learning models that are used includes a base model: Support Vector Machine (SVM). It is used as a benchmark due to its ability to handle multiclass classification effectively. The advanced model is the Random Forest (RF) model. It is implemented to improve predictive accuracy and robustness.

# **Data Exploratory Analysis**

Effective data preprocessing and cleaning are critical for ensuring high-quality input to the machine learning models. Each record in the dataset represents a patient demographics and the assigned drug. The data is first of all checked for missing values and it can be seen from the output attached below that there are no missing values in the dataset.

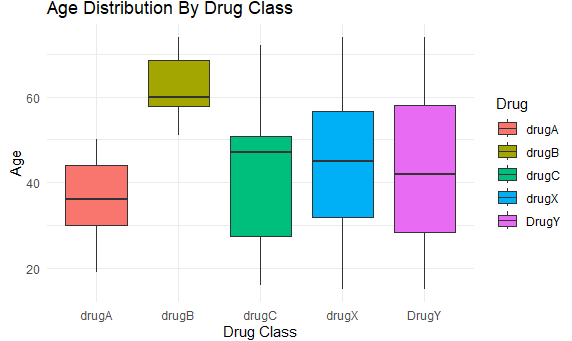


The data types of all of the categorical columns is changed from character to factors for better representation and the summary statistics is obtained. The results shows that the average age is 44.31 years, most of the samples in the data are of male, most have high blood pressure, most have higher chlesterol & the average ratio of na to k is 16.084.

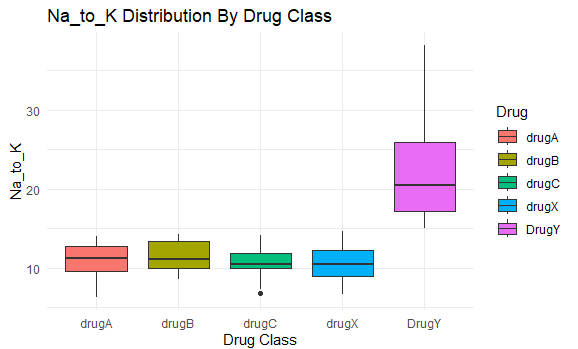


# **Features Exploratory Analysis**

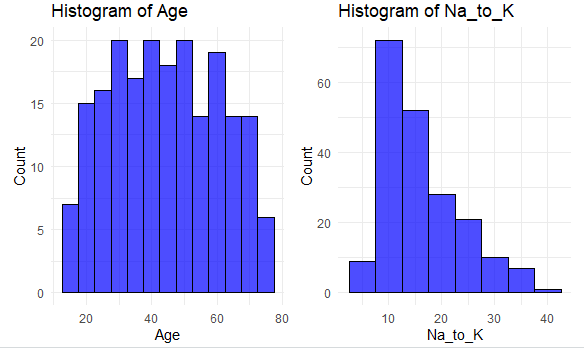
The boxplot of age is created based on the drug type. It can be seen that the average age is higher for those who are assigned drug type B, followed by C, X, Y & A.



The na to k ratio is highest on average for drug type Y and is approximately same on average for all other types of drugs.



The histogram for age & ratio of na to K is attached below. The histogram shows that the age ranges from 15 to 70 years while the ratio ranges from 6 t 40 approximately. The ratio histogram is positively skewed.



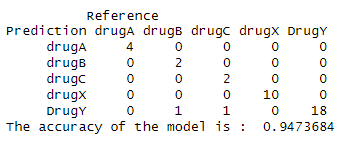
# **Data Cleaning**

One-hot encoding is a technique used to convert categorical variables into a numerical format suitable for machine learning models. In this process, each category of a categorical variable is transformed into a separate binary column, where a value of 1 indicates the presence of that category and 0 indicates its absence. In the provided R code, the dummy\_cols() function from the fastDummies package is used to apply one-hot encoding to the categorical variables BP, Cholesterol, and Sex. The argument remove\_first\_dummy = TRUE ensures that the first category of each variable is removed to prevent multicollinearity, which is essential for models like regression. Additionally, remove\_selected\_columns = TRUE removes the original categorical columns, replacing them with the newly created binary columns. This transformation allows machine learning models to process categorical data efficiently while avoiding issues related to direct categorical encoding.

# **Prediction Model**

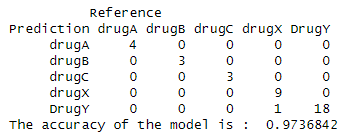
## **Base SVM Model**

The data is divided into training and testing set with 80% of the data used for training while 20% is used for testing the performance of the model. The SVM model is trained and the performance of the model is evaluated on unseen test set of 20%. The results shows that the accuracy of the model is 94.74% which is good. The confusion matrix shows that the two samples which actually belonged to drug Y are misclassified as drugB and drugC. Rest of the classifications are accurate.



## **Random Forest Model (Advanced)**

The random forest model is trained and the performance of the model is evaluated on unseen test set of 20%. The results shows that the accuracy of the model is 97.37% which is much better compared to our base model of SVM. The confusion matrix shows that only one sample is miss classified which actually belonged to drug Y is classified as drugX. Rest of the classifications are accurate.



# **Conclusion**

This work shows that, given patient characteristics, machine learning algorithms can efficiently categorize drug classes. With an accuracy of 94.74%, the base SVM model misclassified two samples of drug Y while properly categorizing most drug kinds. With just one misclassification and an accuracy of 97.37%, the upgraded Random Forest model greatly enhanced performance. These results show how machine learning could help medical practitioners with drug classification, mistake reduction, and patient outcomes enhancement. Additional patient features or hyperparameter optimization for even improved classification accuracy could be part of future studies.