DRUG RECOMMENDATION SYSTEM

PROBLEM DEFINITION

Given the data of patients by Pharmaquick, a pharmaceutical company, deliver them the following:

- **Insights into the patient data.**
- **A ML model that can predict which drug type is suitable for each patient, based on the patient details.**
- **Recommendations on how to utilize our model results.**

AGENDA

1. The Dataset
2. Exploratory Data Analysis- EDA
3. Modelling and Comparison
4. Model Deployment
5. Conclusion

THE DATASET

200 **rows** and 6 **columns**

No **null** values

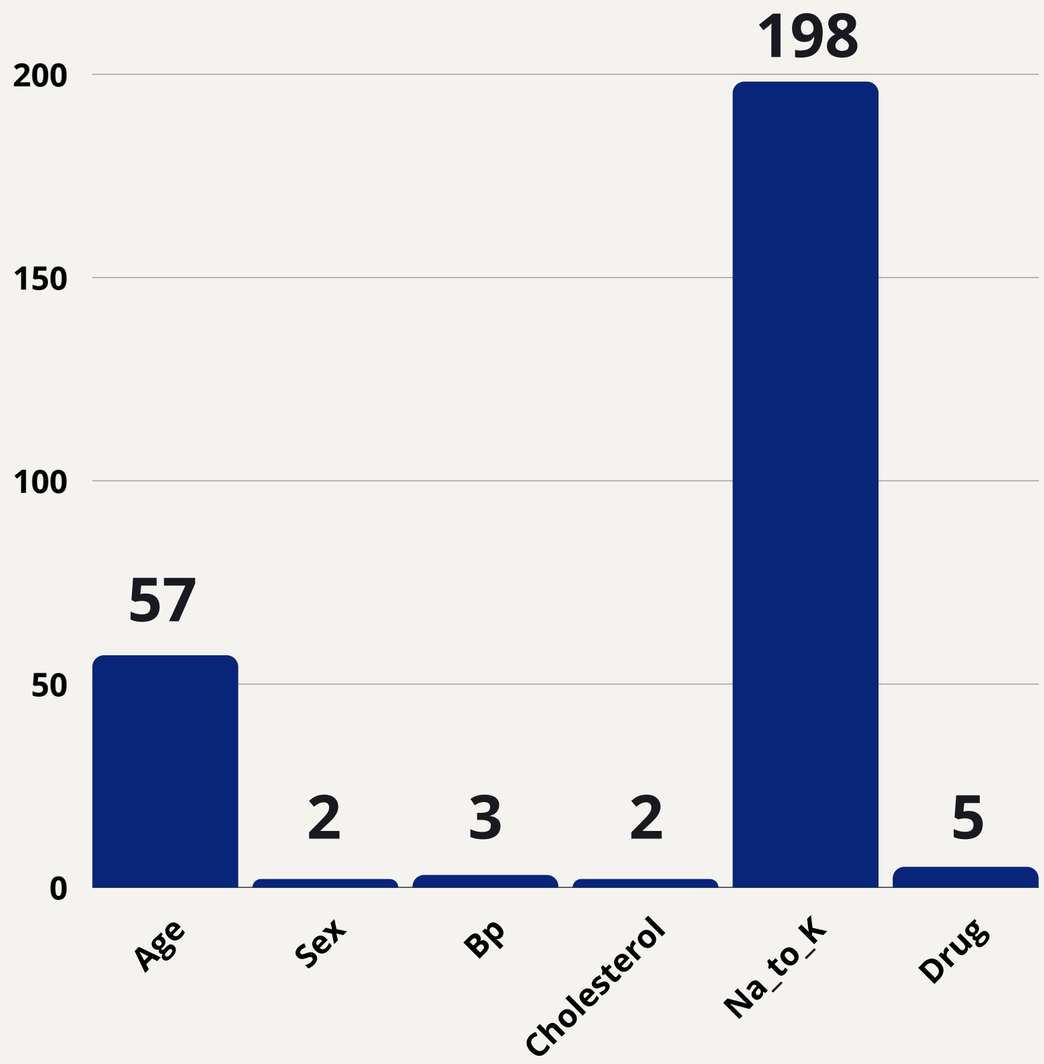
2 **Numeric** columns
- Age
-Na_to_K

4 **Categorical** Columns
- Sex
- BP
- Cholesterol
-Drug

EXPLORATORY DATA ANALYSIS- EDA

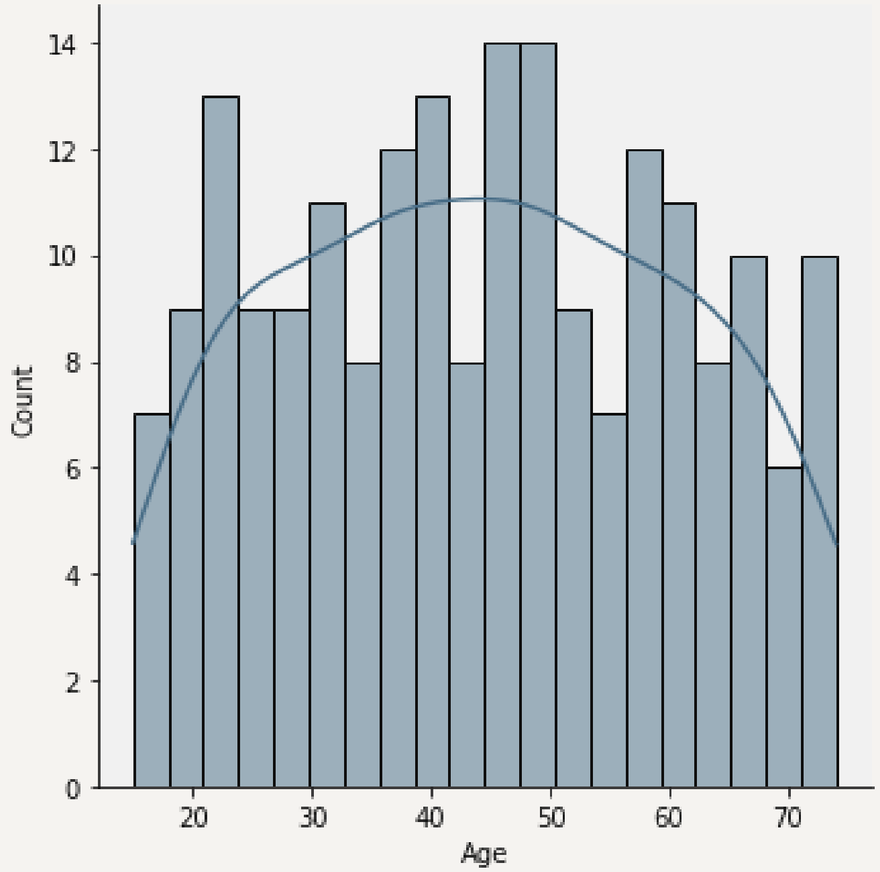
- Unique Elements Graph
- Skewness Graph
- Correlation Analysis

Unique Elements per column



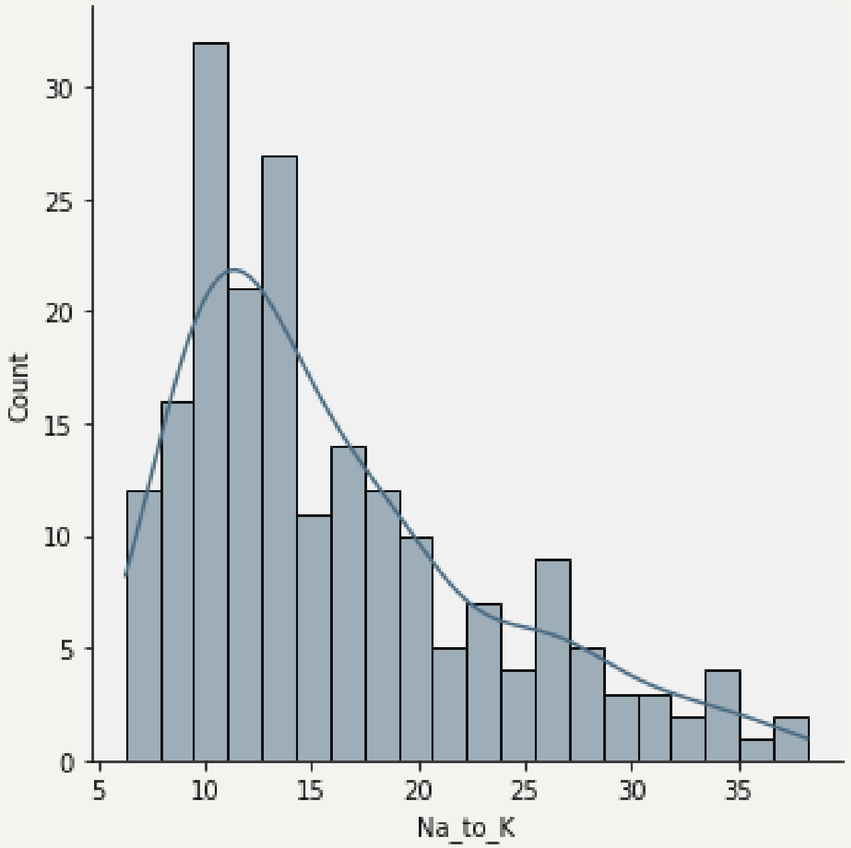
EXPLORATORY DATA ANALYSIS- EDA

Skewness Graph



Age graph: **Normal skewed**, symmetrical

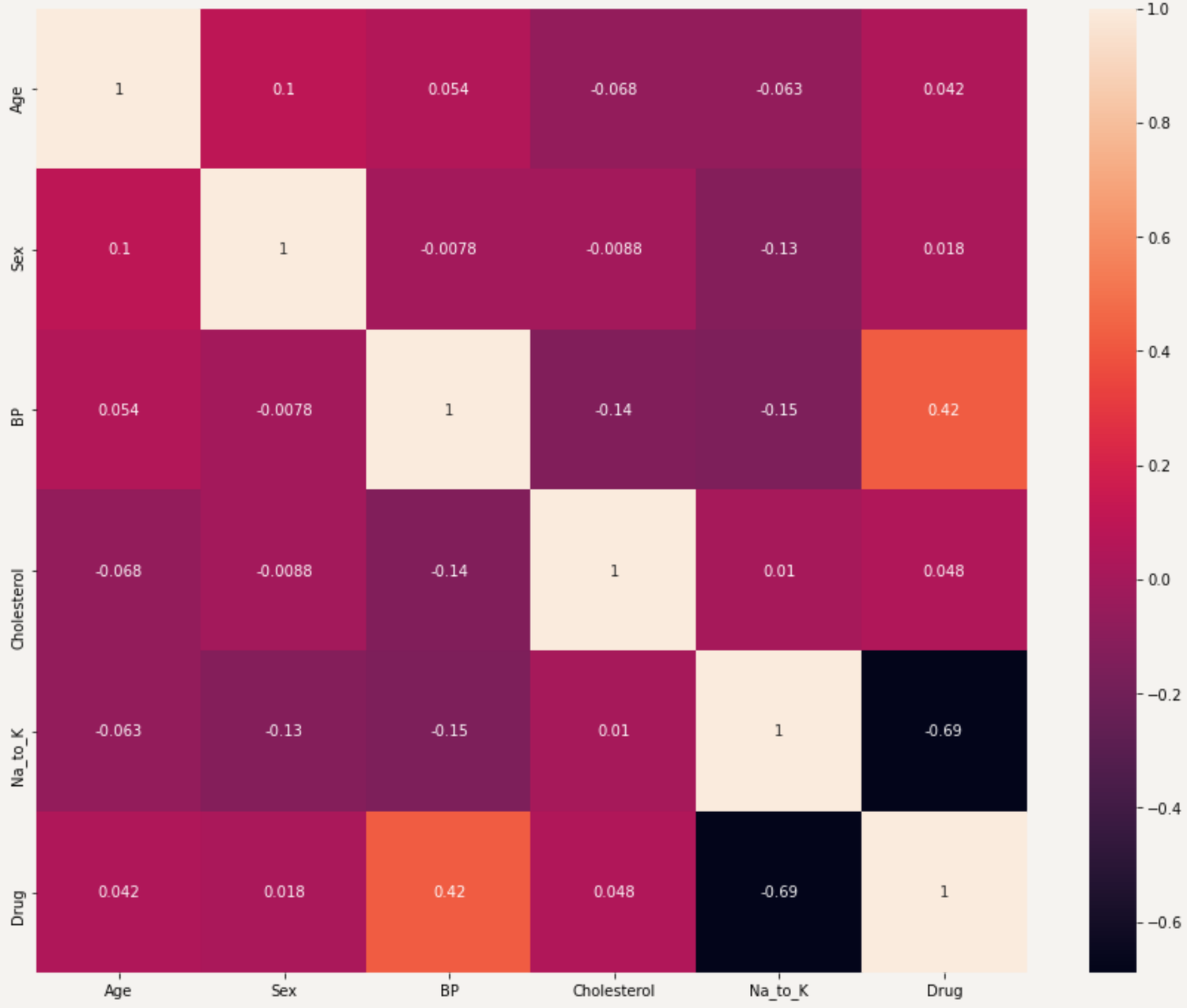
Age skewness: **0.03**



Na_To_K graph: **right- skewed**

Na to K skewness: **1.04**

CORRELATION ANALYSIS



Drug and Na_to_K ratio are **negatively correlated**

Drug and BP are **positively correlated**

EDA USING DATA VISUALIZATION

10 EDA graphs in 2 categories

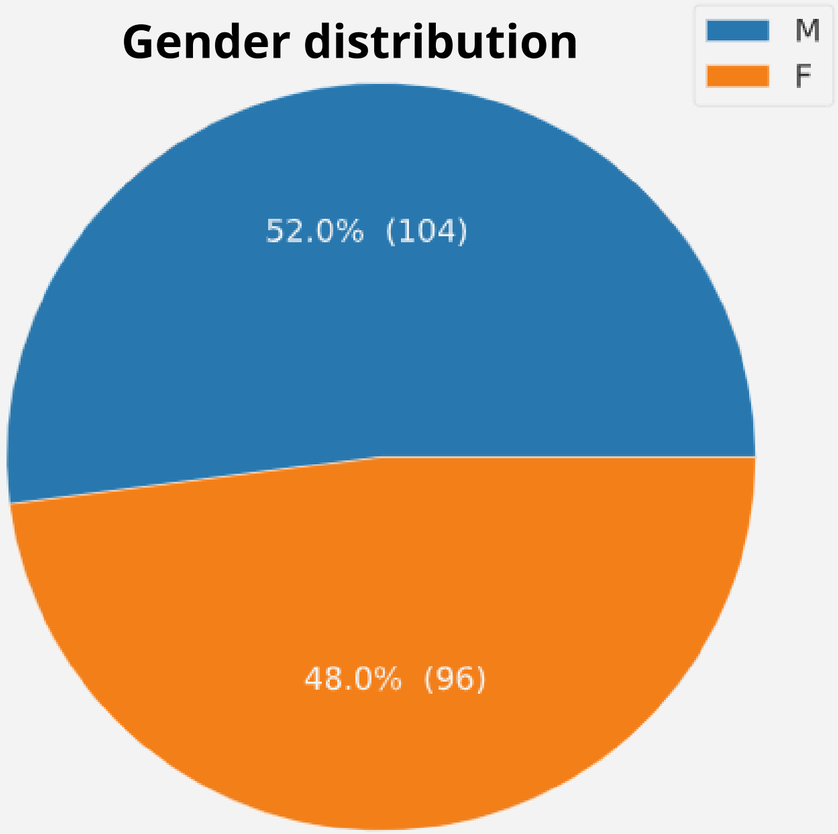
1.Distributaion Graphs

- Gender-wise
- BP-wise
- Cholesterol-wise
- Drugs-wise

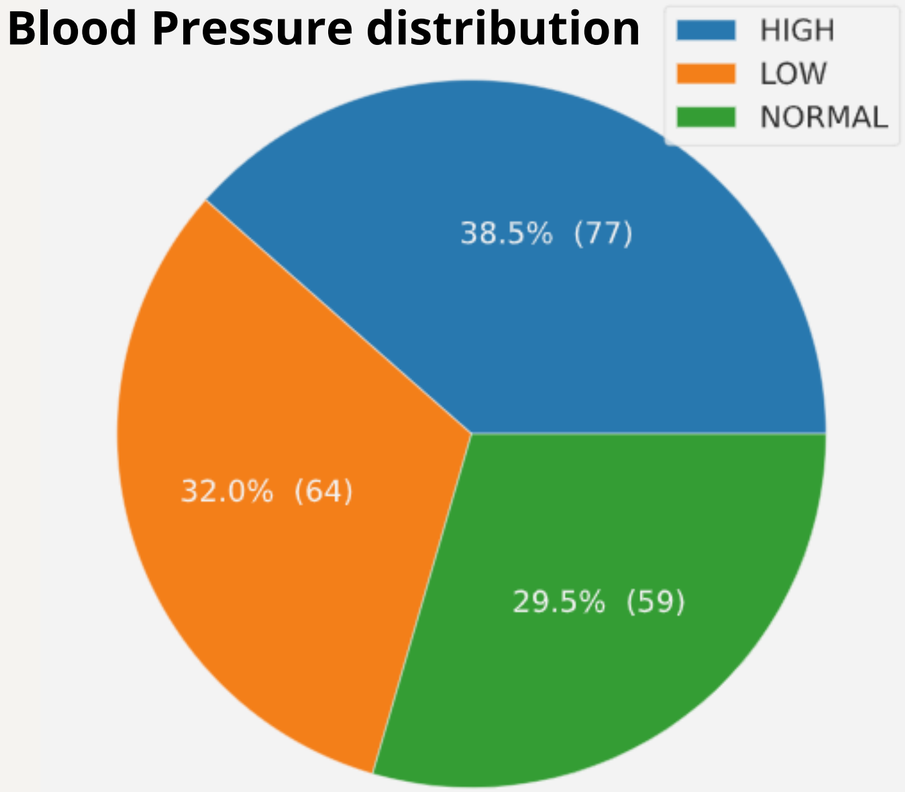
2.Prescription Graphs

- Drugs Vs SEX
- Drugs Vs BP
- Drugs Vs cholesterol
- Drugs Vs Na-to-K
- Drugs Vs Age
- Special- Na vs BP vs Drug

DISTRIBUTION GRAPHS



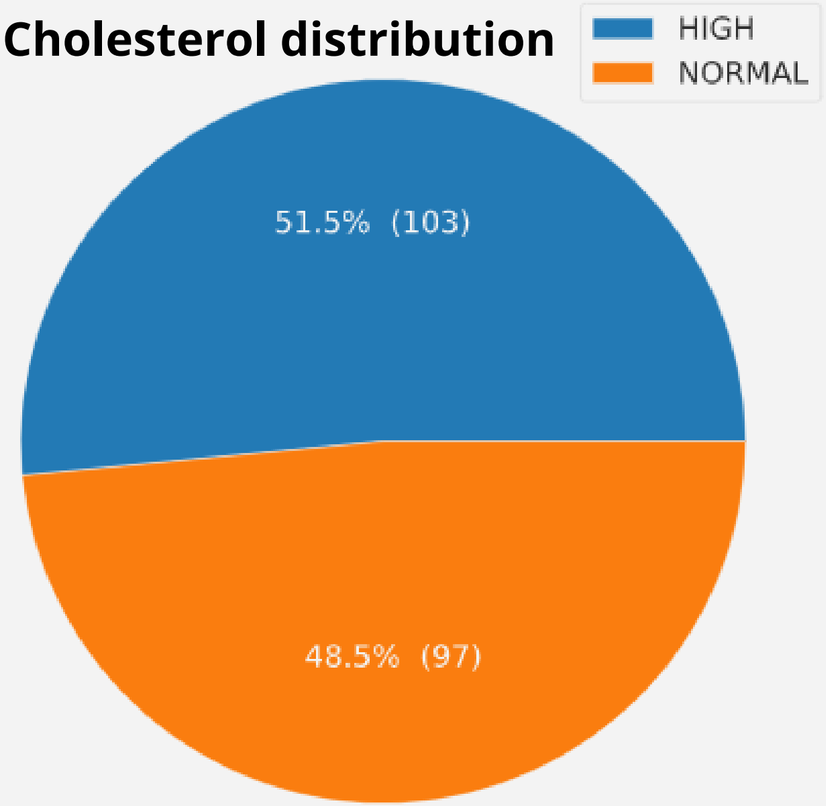
There were **104 Males** and **96 Females** records



Blood Pressure recording includes-**77 High** Bp, **64 Low** Bp and **59 Normal** Bp patients

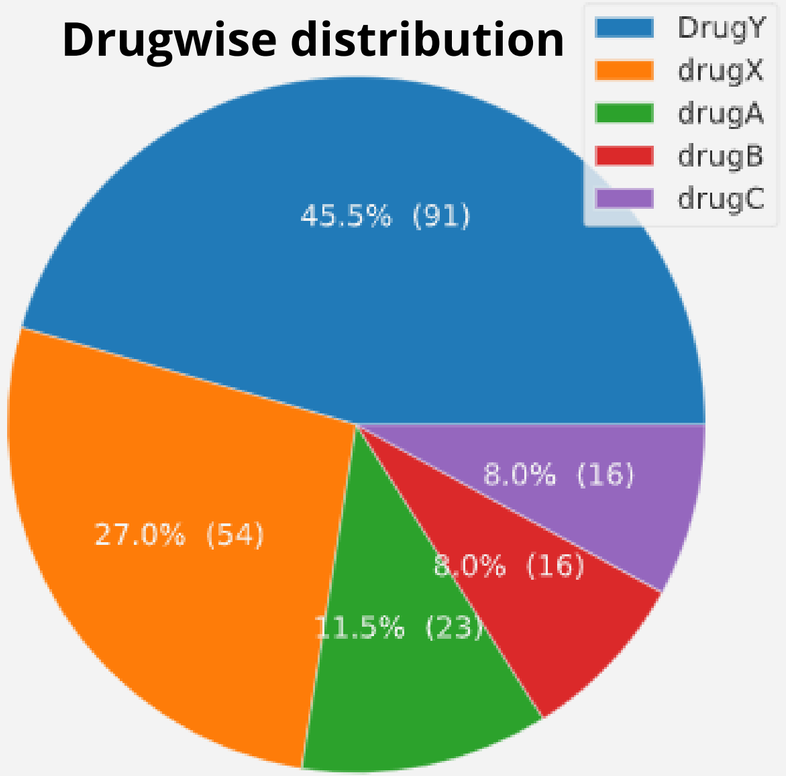
DISTRIBUTION GRAPHS

Cholesterol distribution



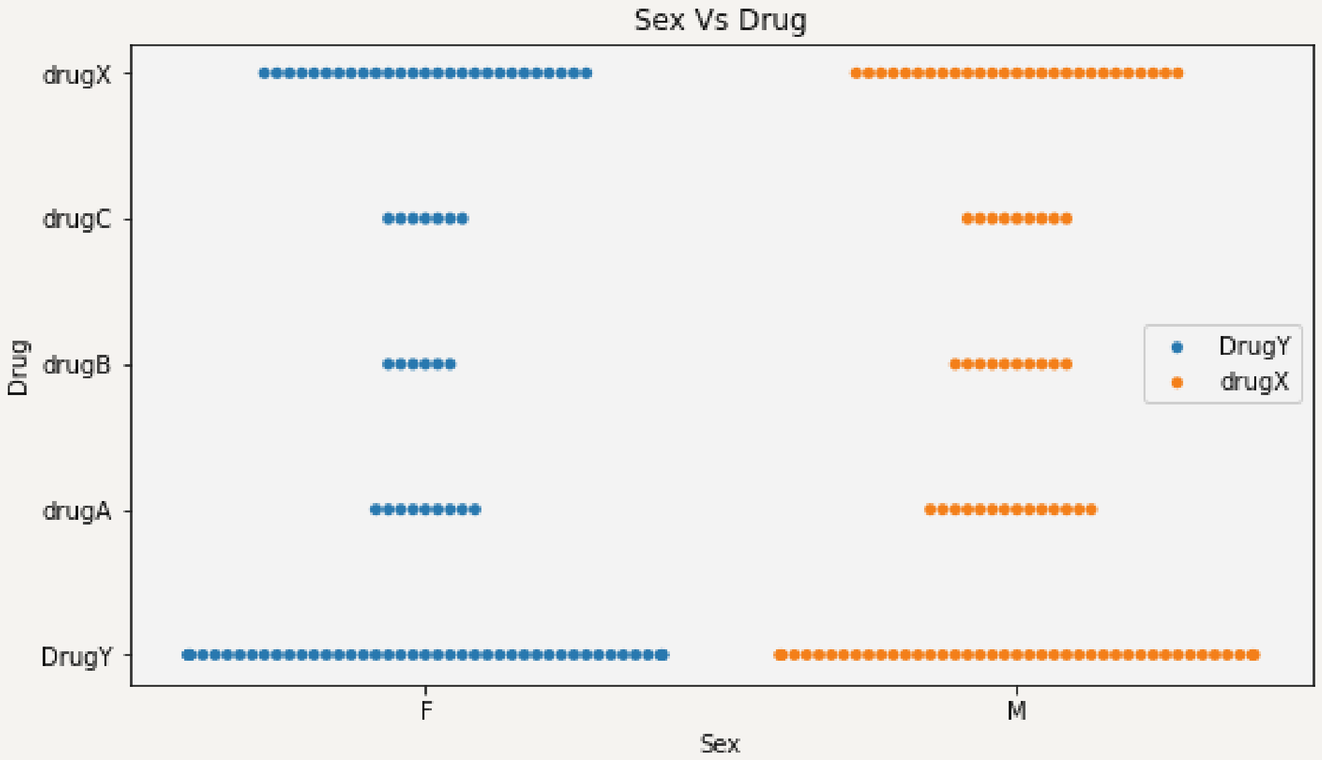
The records with **high cholesterol(103)** were more as compared to **normal (97)**

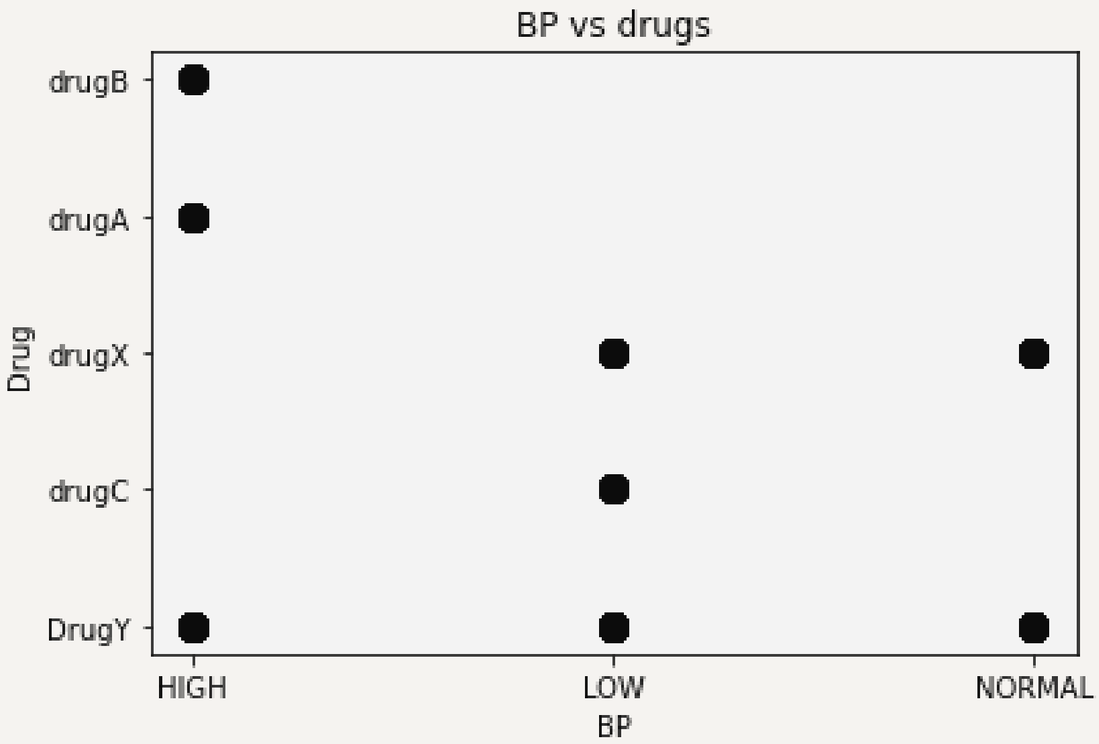
Drugwise distribution



The most prescribed drug is **DrugY -91** and the least prescribed ones are **DrugC** and **DrugB-16**

PRESCRIPTION GRAPHS

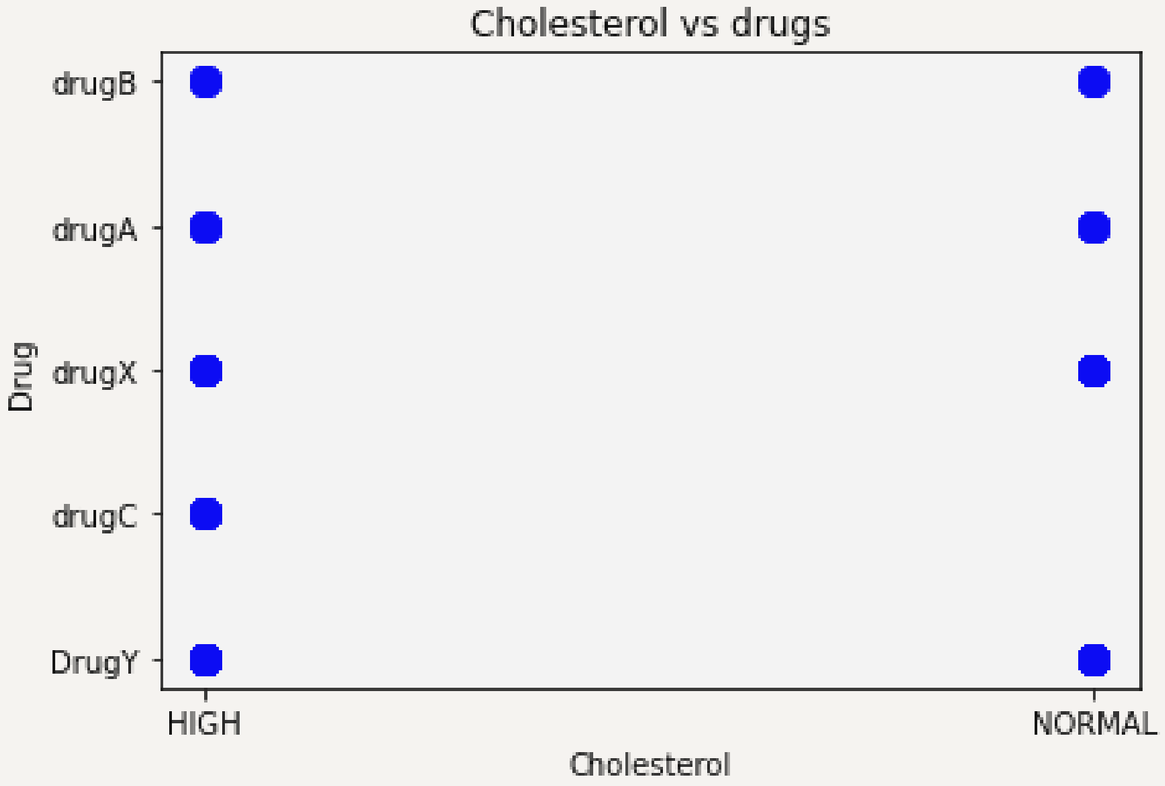


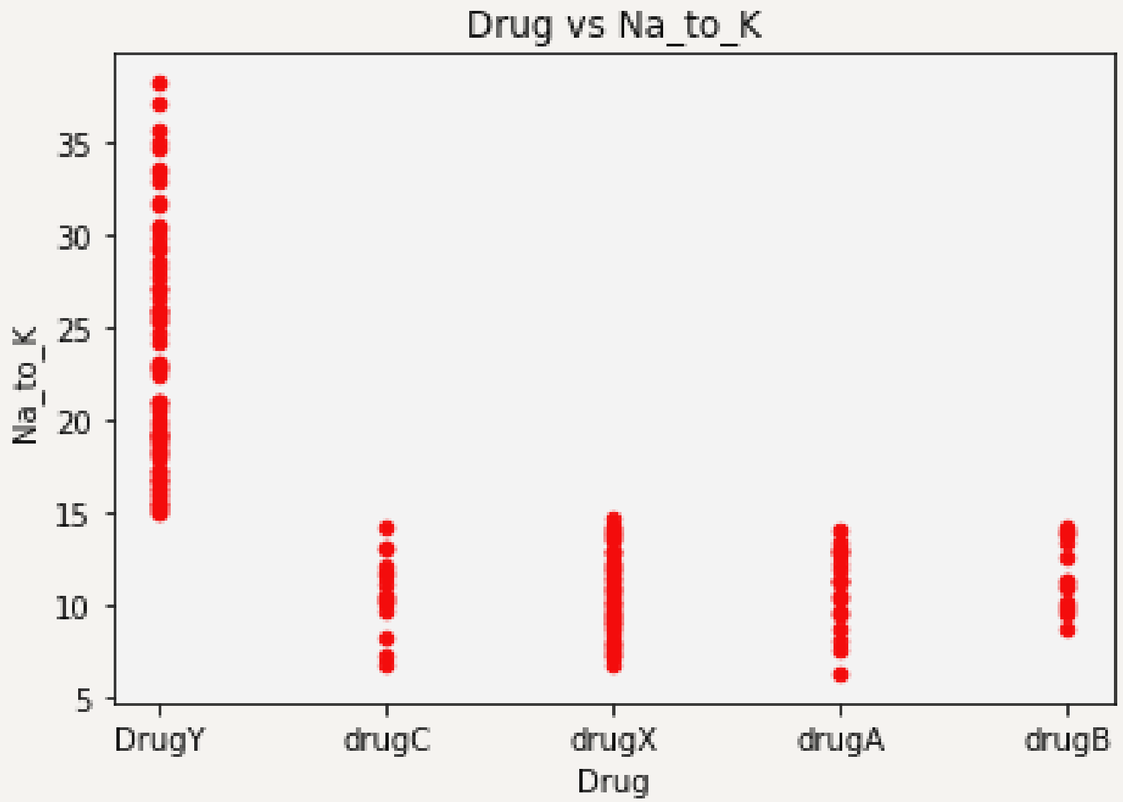


Most prescribed drug to both males and females is **DrugY** while the least one is **DrugB** for females and **DrugC** for males

- **Drug Y** prescribed irrespective of BP
- **Drug A** and **Drug B** only for High BP
- **Drug C** only for Low BP

PRESCRIPTION GRAPHS

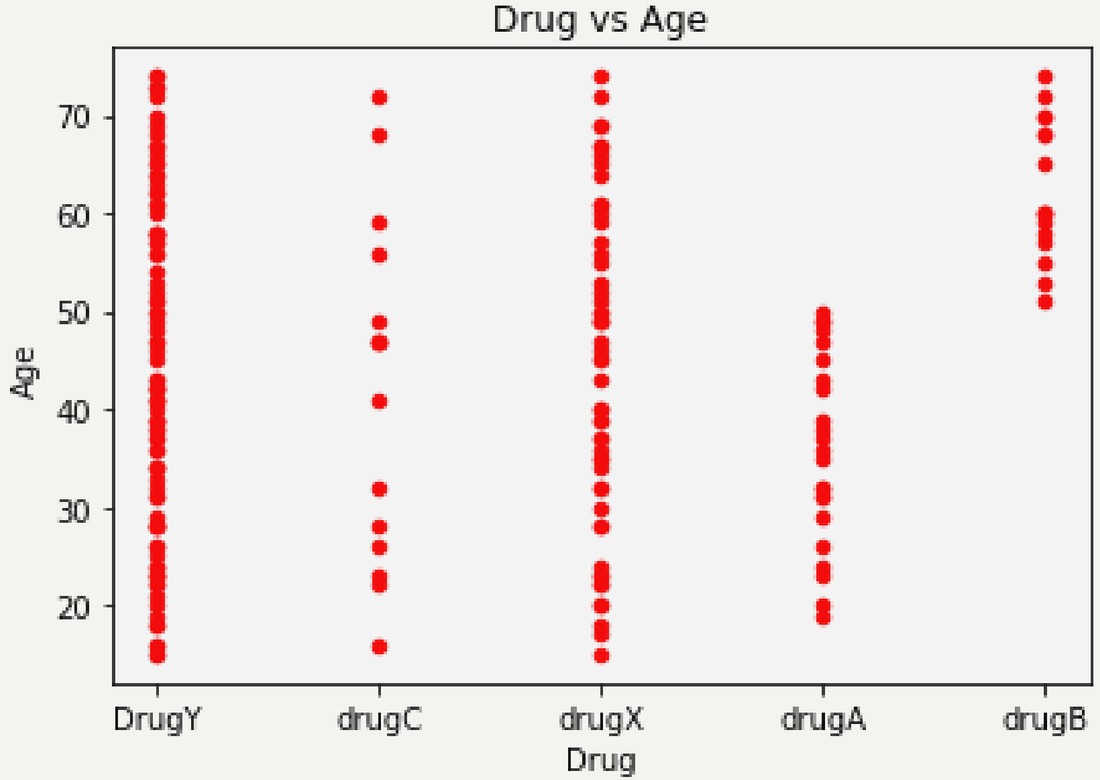




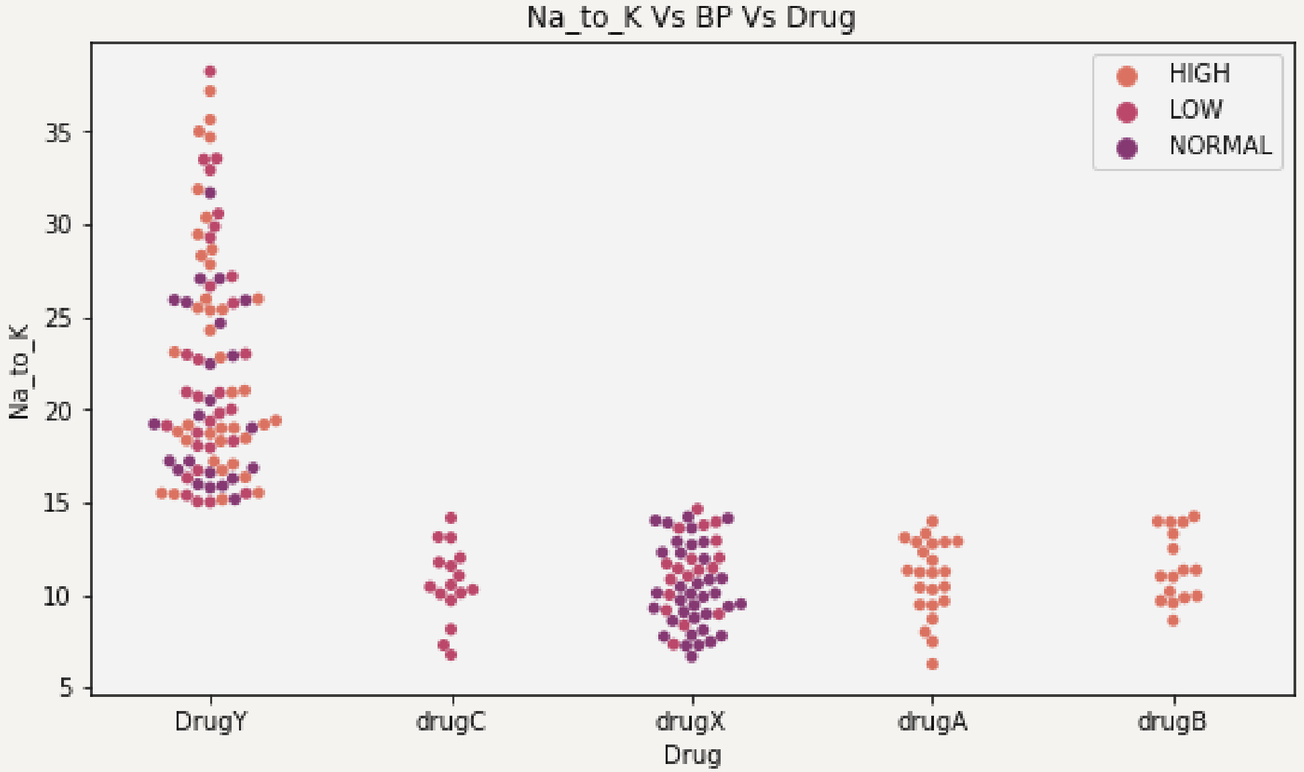
- **All drugs** are prescribed for high cholesterol
- All drugs **except DrugC** are prescribed for normal cholesterol

- **DrugY** is prescribed for Na_to_K level of 15 and above
- **All other drugs** are prescribed for Na_to_K levels below 15

PRESCRIPTION GRAPHS



- **DrugA** is prescribed for age below 50
- **DrugB** is prescribed for age above 50
- **DrugY**, **DrugC, DrugX** are prescribed for all ages



- **DrugA** and **DrugB** is presribed to High BP patients with Na_to_k less than 15
- **DrugC** is prescribed to Low BP patients with Na_to_K less than 15

MODELLING AND COMPARISON

1

Splitting the randomized dataset into train and test sets in the ratio 70:30 with **5-fold stratified cross validation**

2

Using **AutoML** for automatic model training and hyperparameter tuning

3

Comparing the performance and reporting the best model

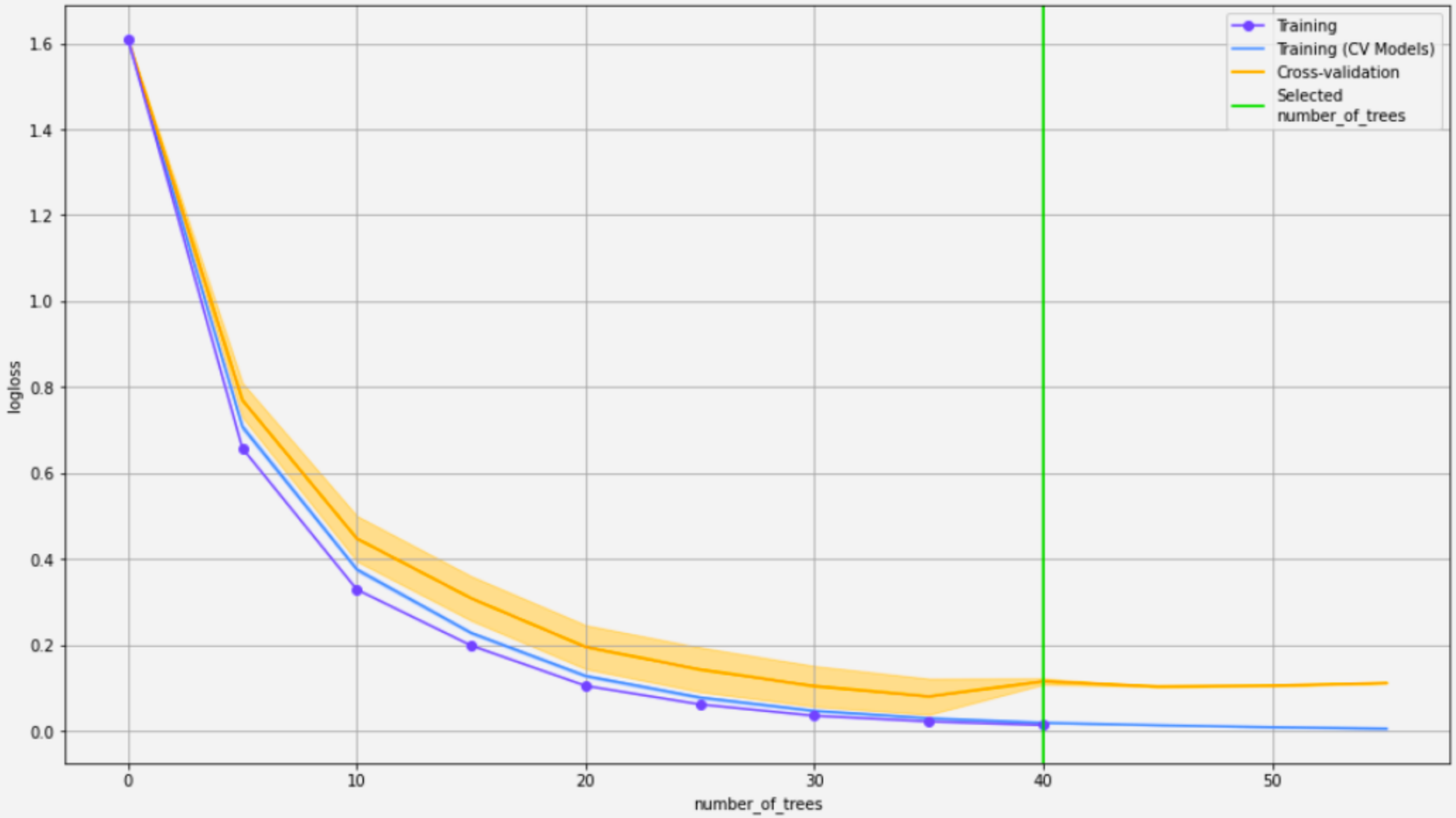
MODELS TRAINED

Algorithms	Accuracy(%)	Precision(%)	Recall(%)
Gradient Boosting Machine (GBM)	**100**	**100**	**100**
Extreme Gradient Boosting (XGB)	**100**	**100**	**100**
Random Forest (RF)	98.3	98	98
Generalised Linear Models (GLM)	96.6	97	97
Extremely Randomized Trees (XRT)	95	96	95
Deep Neural Network (3 Layers)	90	91	90
* Naive Bayes	90	93	90
* Logistic Regression	86.7	88	87
* Decision Tree	86.7	95	87
* Support Vector Classifier	80	67	80
* K-NearestNeighbour	76.7	74	77
Deep Neural Network (5 Layers)	73.3	74	73
* Stochastic Gradient	58.3	57	58
*** Manually Trained Algorithms**			

LEARNING CURVE FOR GBM

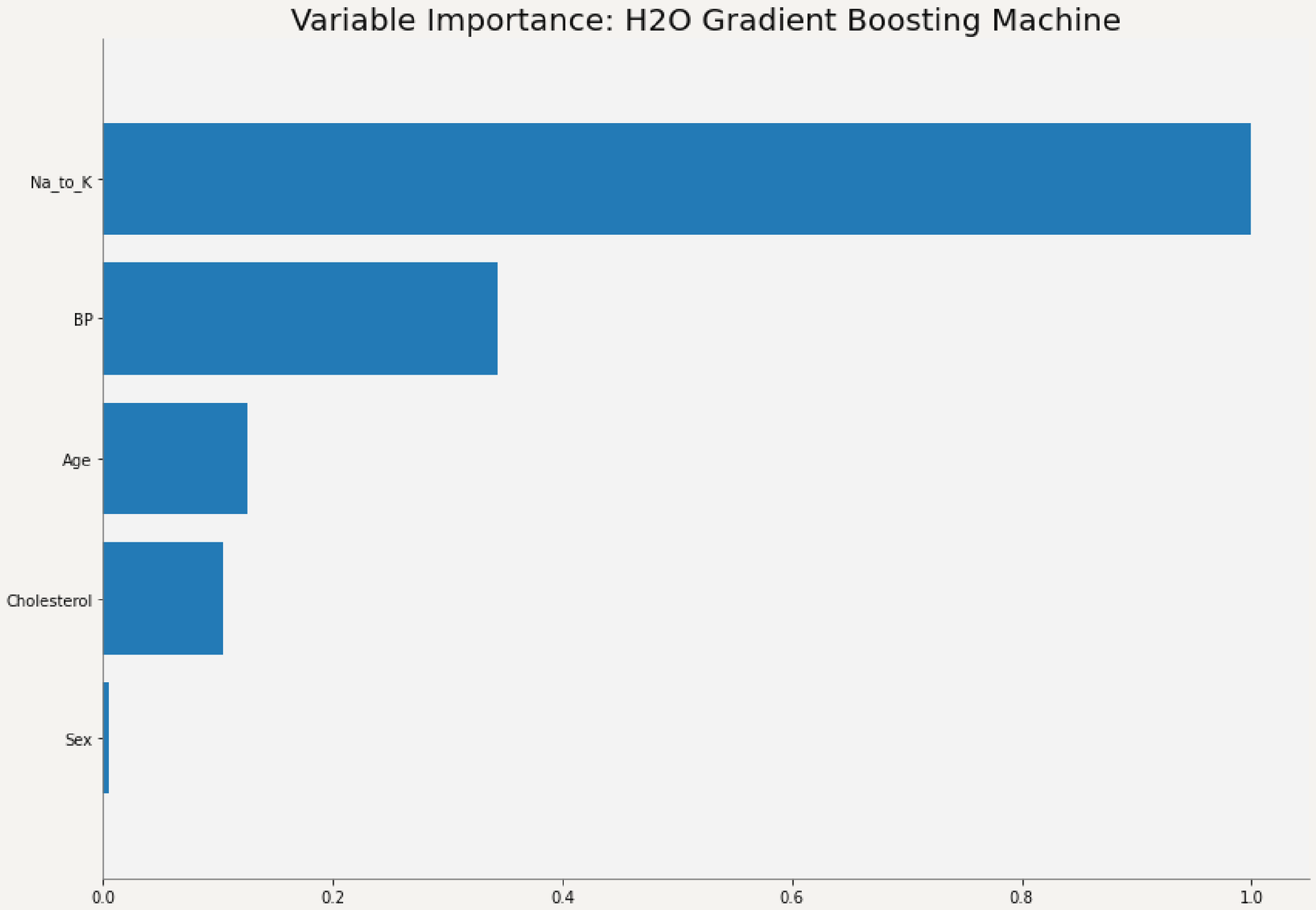
- With an accuracy of 100% **GBM** is the best performing algorithm .

- **The number of trees** is taken as the point where the training and validation curves converge (ideally).



VARIABLE IMPORTANCE PLOT FOR GBM

NA_to_K ratio is the most significant feature in GBM while **Sex** is the least significant



MODEL DEPLOYMENT: LIVE DEMONSTRATION

CONCLUSION

- The best fit model for the respective problem is **Gradient Boosting Machine with 100% accuracy**
- The model has perfect accuracy, however, it should be noted that the train and validation sets are very small. Hence, it is **not guaranteed** to perform perfectly at test time.
- In order to automate the process of model selection and hyperparameter tuning, we opted for **AutoML.**
- The actual deployment of our model has been demonstrated via the **GUI.**

THANK YOU