kaggleX

KaggleX-2023 Cohort 3 Project Showcase

CVD Risk Prediction

By: Ashar Habib
https://www.linkedin.com/in/asharhabib/
https://github.com/habibashar786

Mentor : Misriyah shahul Hameed

Background

- I am a Finance Business Partner, undergraduate in Commerce, Master of Business Administration, CMA, with a Master of Data Science from LJMU UK,
- Currently I am working in the Finance department, Migrating toward a Machine Learning scientist role.

Project Definition

- Cardiovascular diseases (CVD) are a leading cause of mortality globally. The accurate prediction and evaluation of CVD are paramount to implementing preventive measures and treatments effectively. However, there are significant challenges in the current predictive models, including unidentified biases and limited predictive accuracy. The role of Machine learning in cardiovascular risk prediction, I have received the dataset from the National Library of Medicine.

- I have applied Machine learning algorithms and compared the model performance of multiple models like (logistic regression, SVM, XGBoost, decision tree, naive base)

- Logistic Regression: 92.37% XGBoost: 96.31% SVM: 94.64% Decision Tree: 92.60% Naive Bayes: 91.02%

Dataset Description:

- Data set :

```
Data columns (total 18 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   patient_id      22011 non-null  int64
 1   age             22011 non-null  int64
 2   sex             22011 non-null  category
 3   education       22011 non-null  object
 4   marital_status  22011 non-null  object
 5   occupation      22011 non-null  object
 6   sbp_avg         22011 non-null  float64
 7   dbp_avg         22011 non-null  float64
 8   bg_mgdl         22011 non-null  int64
 9   bmi             22011 non-null  float64
 10  smoking         22011 non-null  object
 11  village         22011 non-null  object
 12  areas           22011 non-null  object
 13  cvdrisk         22011 non-null  float64
 14  highrisk        22011 non-null  category
 15  bplt            22011 non-null  category
 16  lltt            22011 non-null  category
 17  aptt            22011 non-null  category
dtypes: category(5), float64(4), int64(3), object(6)
memory usage: 2.5+ MB
```

kaggle

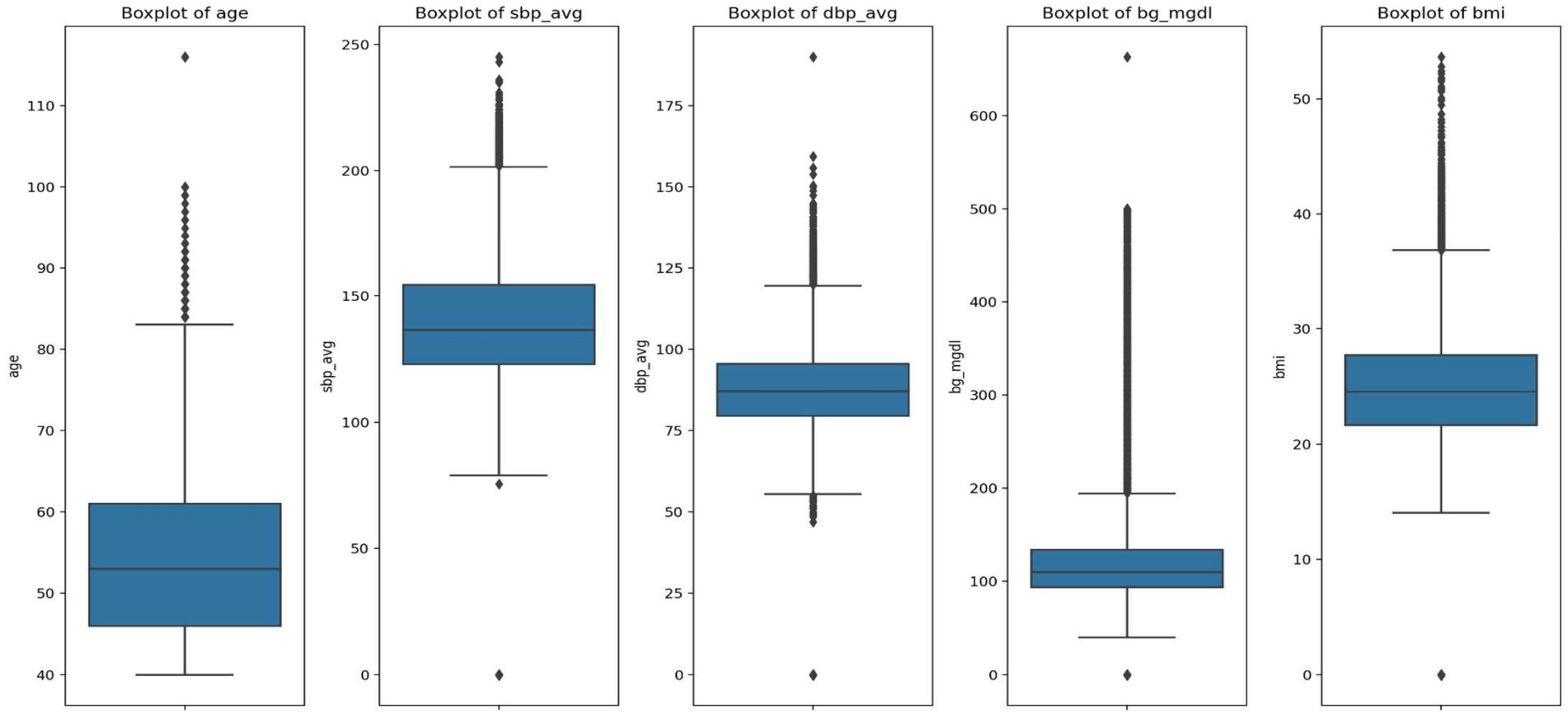
Using the ‘Describe’ Function for the Data set:

```
]:          patient_id           age       sbp_avg       dbp_avg       bg_mgdl  \
count  2.000500e+04  20005.000000  20005.000000  20005.000000  20005.000000
mean   3.187875e+11     54.600350    139.792552     88.169927    117.407848
std    8.500745e+11     10.460096     23.075979     12.702048     38.918047
min    7.709001e+09     40.000000     79.000000     47.000000      0.000000
25%    1.020202e+11     46.000000    123.000000     79.500000     93.000000
50%    1.030301e+11     53.000000    136.000000     86.500000    109.000000
75%    1.040401e+11     61.000000    153.500000     95.500000    130.000000
max    5.691157e+12     87.000000    216.000000    130.500000    308.000000

                bmi       cvdrisk
count  20005.000000  20005.000000
mean      24.889036      0.225444
std        4.518979      0.193450
min       14.073940      0.100000
25%       21.671260      0.100000
50%       24.508950      0.100000
75%       27.630370      0.450000
max       40.723770      0.550000
```

EXPLORATORY DATA ANALYSIS

Box plot for the describe function:



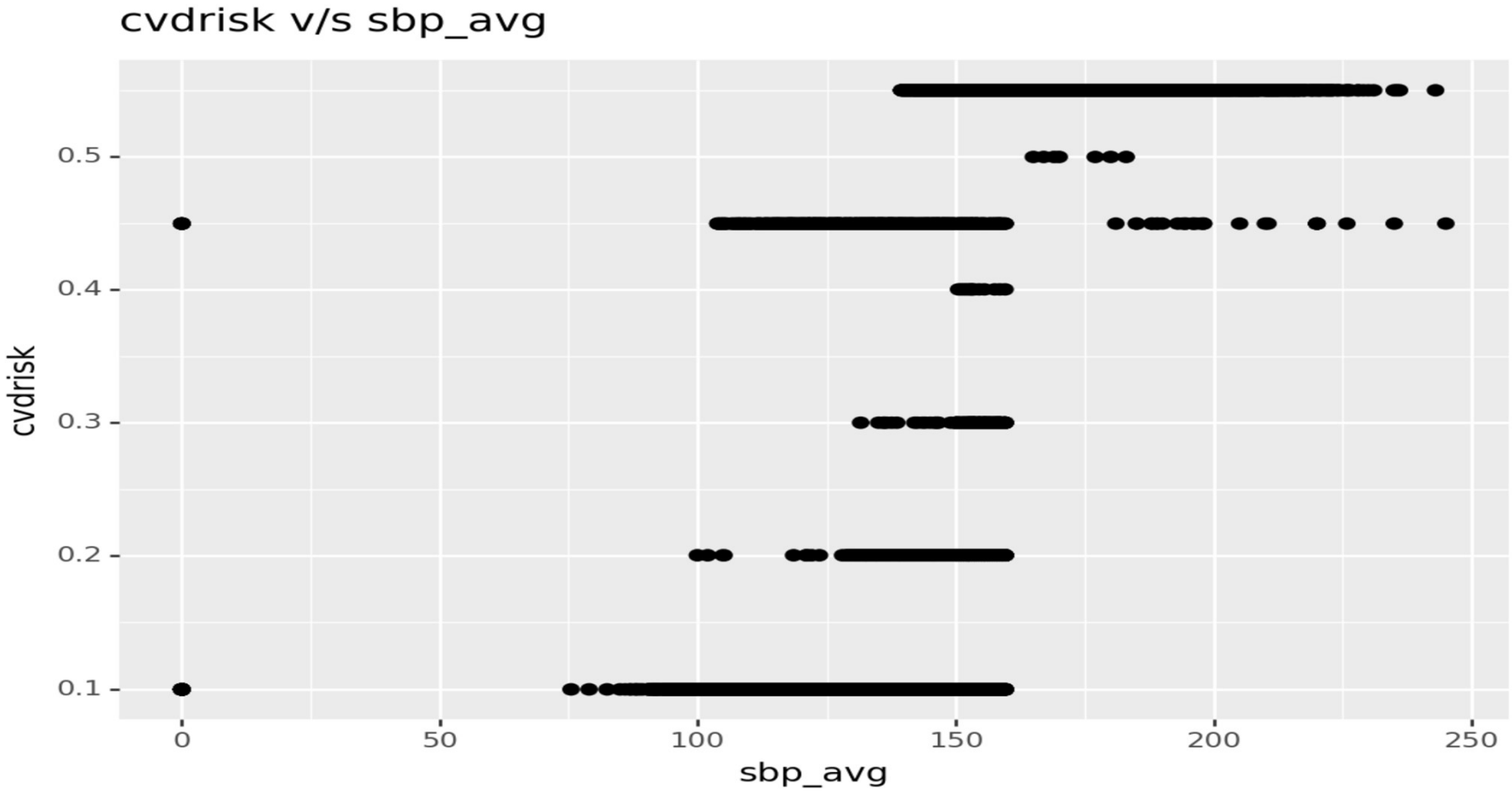
Correlation Matrices highlight relationships:

```
In [19]: db1.corr()
```

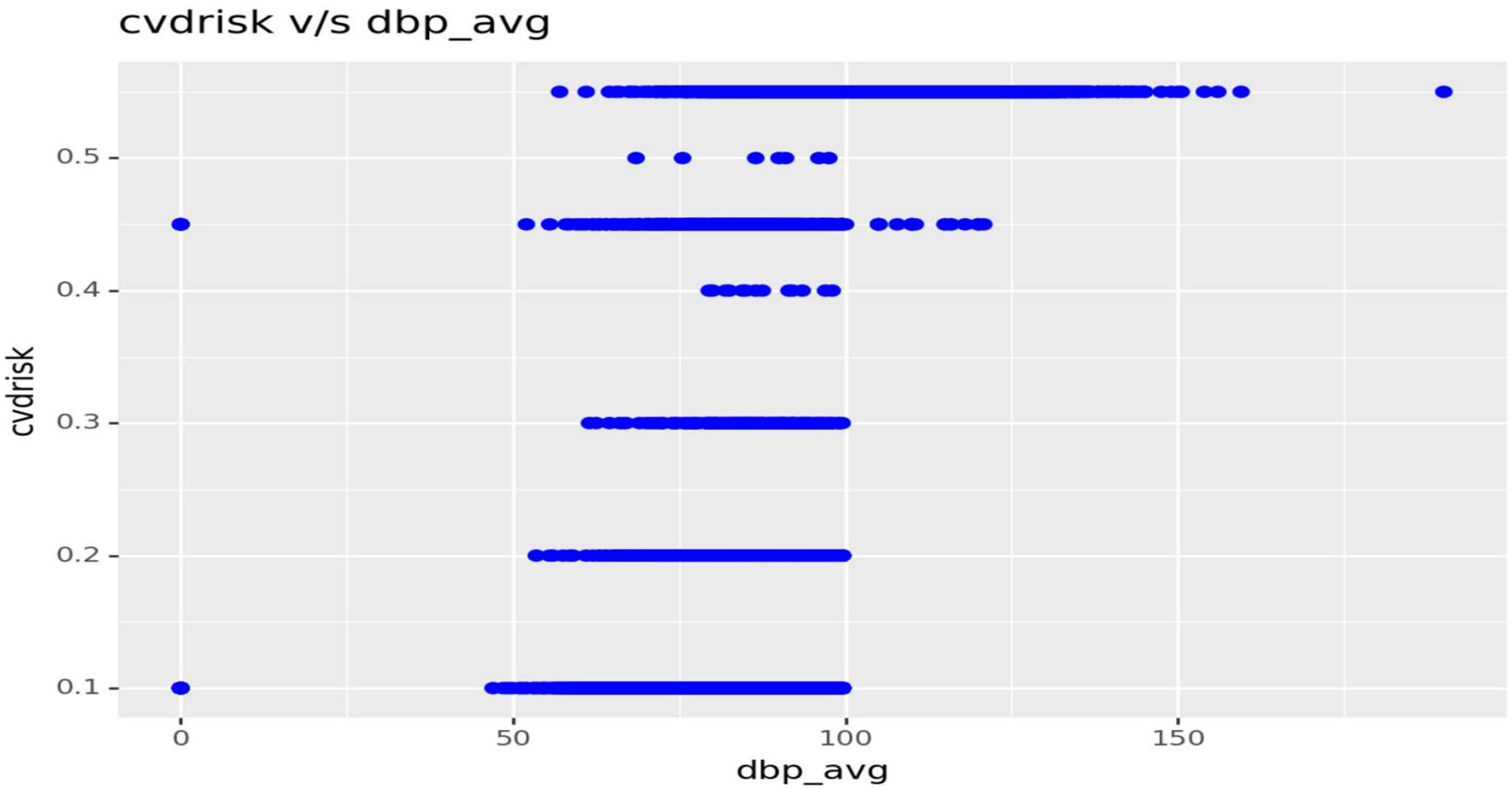
Out[19]:

	patient_id	age	sbp_avg	dbp_avg	bg_mgdl	bmi	cvdrisk
patient_id	1.000000	0.001687	0.007296	0.008360	-0.001498	0.006066	0.006920
age	0.001687	1.000000	0.263526	-0.039284	0.058495	-0.258770	0.241036
sbp_avg	0.007296	0.263526	1.000000	0.792850	0.104097	0.113916	0.733698
dbp_avg	0.008360	-0.039284	0.792850	1.000000	0.055725	0.236623	0.624909
bg_mgdl	-0.001498	0.058495	0.104097	0.055725	1.000000	0.096113	0.108665
bmi	0.006066	-0.258770	0.113916	0.236623	0.096113	1.000000	0.102833
cvdrisk	0.006920	0.241036	0.733698	0.624909	0.108665	0.102833	1.000000

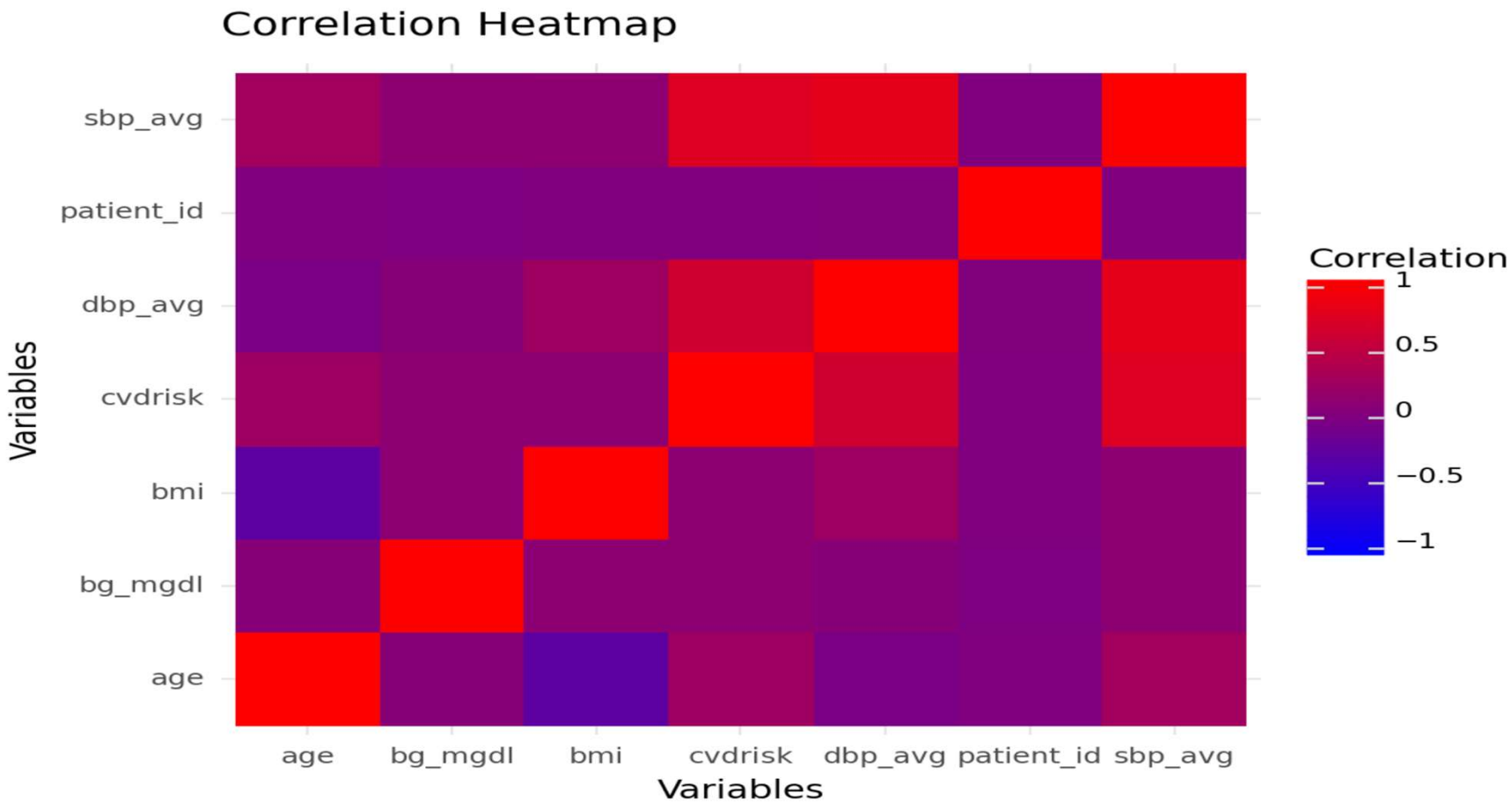
Correlation between Feature and Target variable:



Correlation between Feature and Target variable:



Correlation between Features variable:

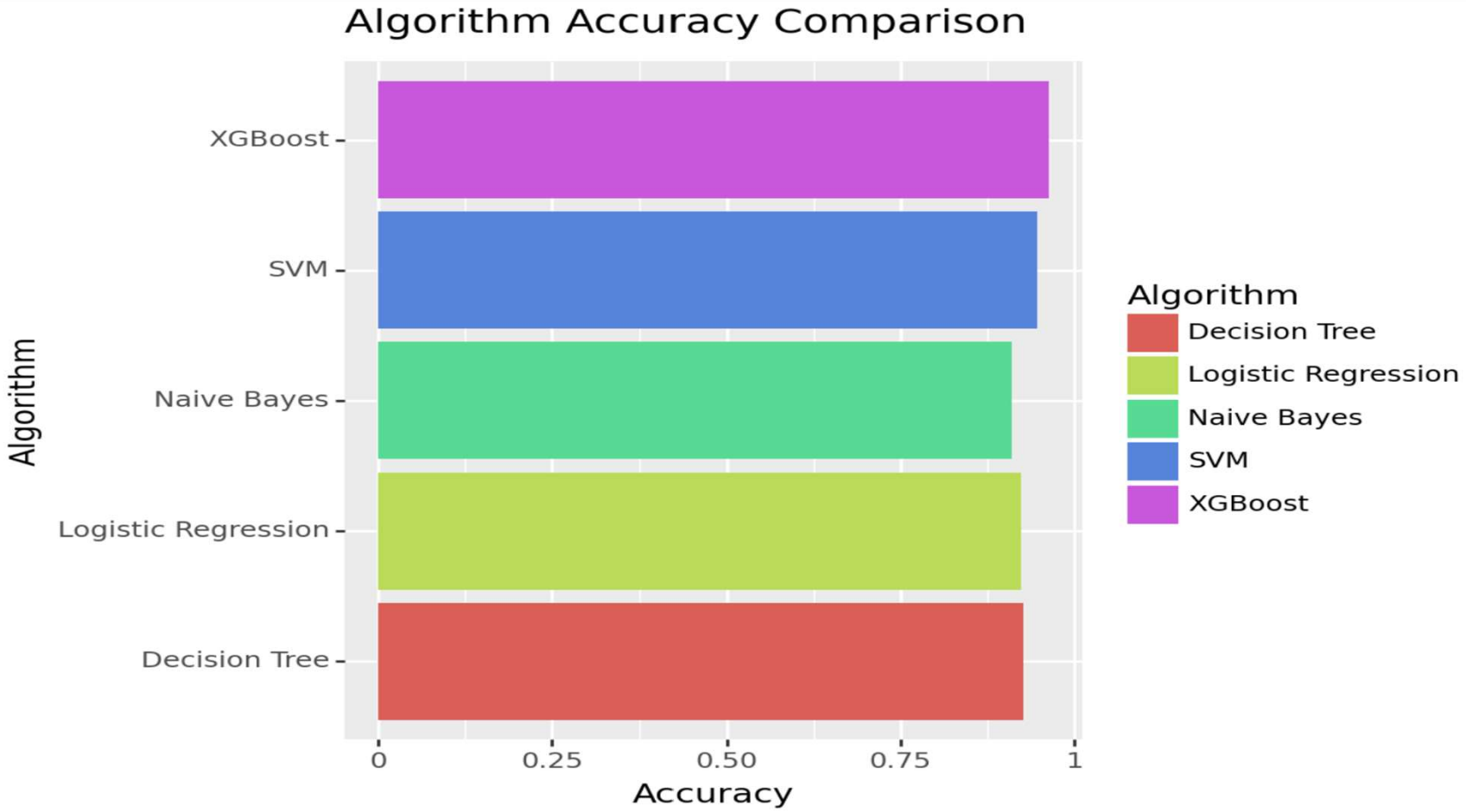


MODELS PREDICTION AND THEIR RESULTS

Comparison of Model predictions :

	Algorithm	Accuracy
0	Logistic Regression	0.923738
1	XGBoost	0.963114
2	SVM	0.946368
3	Decision Tree	0.926001
4	Naive Bayes	0.910161

Visualization for Model performance:



Analysis :

Logistic Regression:

Accuracy: 92.37% Interpretation: This model has performed well, classifying approximately 92% of the test data correctly.

Linear Regression: Accuracy: Not Applicable Interpretation: Since it's a regression model, accuracy is not a suitable metric. we should consider evaluating it using metrics like RMSE, MAE, or R^2.

XGBoost: Accuracy: 96.31% Interpretation: XGBoost has given the highest accuracy among all the tested models. It's a powerful gradient boosting algorithm that performed excellently on your dataset.

SVM (Support Vector Machine): Accuracy: 94.64% Interpretation: SVM also performed well, with an accuracy close to that of XGBoost. It's a reliable model for classification tasks.

Decision Tree: Accuracy: 92.60% Interpretation: The decision tree has a comparable performance to logistic regression. It's a simpler model and can be visualized easily, but it might be prone to overfitting.

Naive Bayes: Accuracy: 91.02% Interpretation: Naive Bayes has the lowest accuracy among the classifiers tested, but it's still above 90%, which is quite good. It's a fast and simple algorithm, especially suitable for high-dimensional datasets.

Conclusion :

Best Model for Prediction: The XGBoost model has emerged as the most accurate model with an accuracy of 96.31%. It has outperformed other classification models in predicting the CVD risk based on the features provided.

The XGBoost model, with its high accuracy, stands as a valuable tool for predicting CVD risk. Its integration into clinical practice can revolutionize CVD management, promoting preventive healthcare, reducing morbidity and mortality, and enhancing the quality of life for individuals. It underscores the pivotal role of machine learning in transforming healthcare, making it more proactive, personalized, and precise.

THANK YOU


kaggle