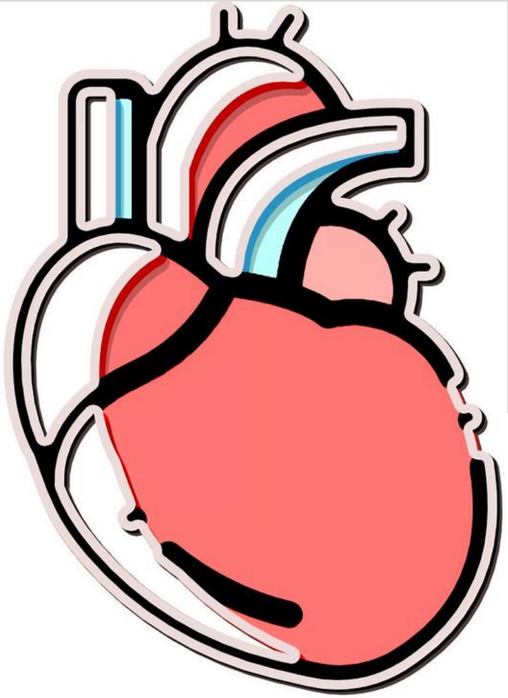
أكاديمية سدايا
SDAIA Academy

# Cardio Disease

## for Classification



Presented by:
Ghadah Alharbi and Rahaf Alyousef

4-11-2021

Cardiovascular disease is the leading cause of death throughout the United States.
In this Dataset "Cardio Disease", We'll use it to build classification models and try to analyze and gather the insights of a dataset and predict the possibility of a person having Cardiovascular disease based on various parameters specified in this dataset.

# Business Objective

# Data

The cardiovascular disease dataset is found on Kaggle.
The data consists of 70,000 patient records and 14 features.

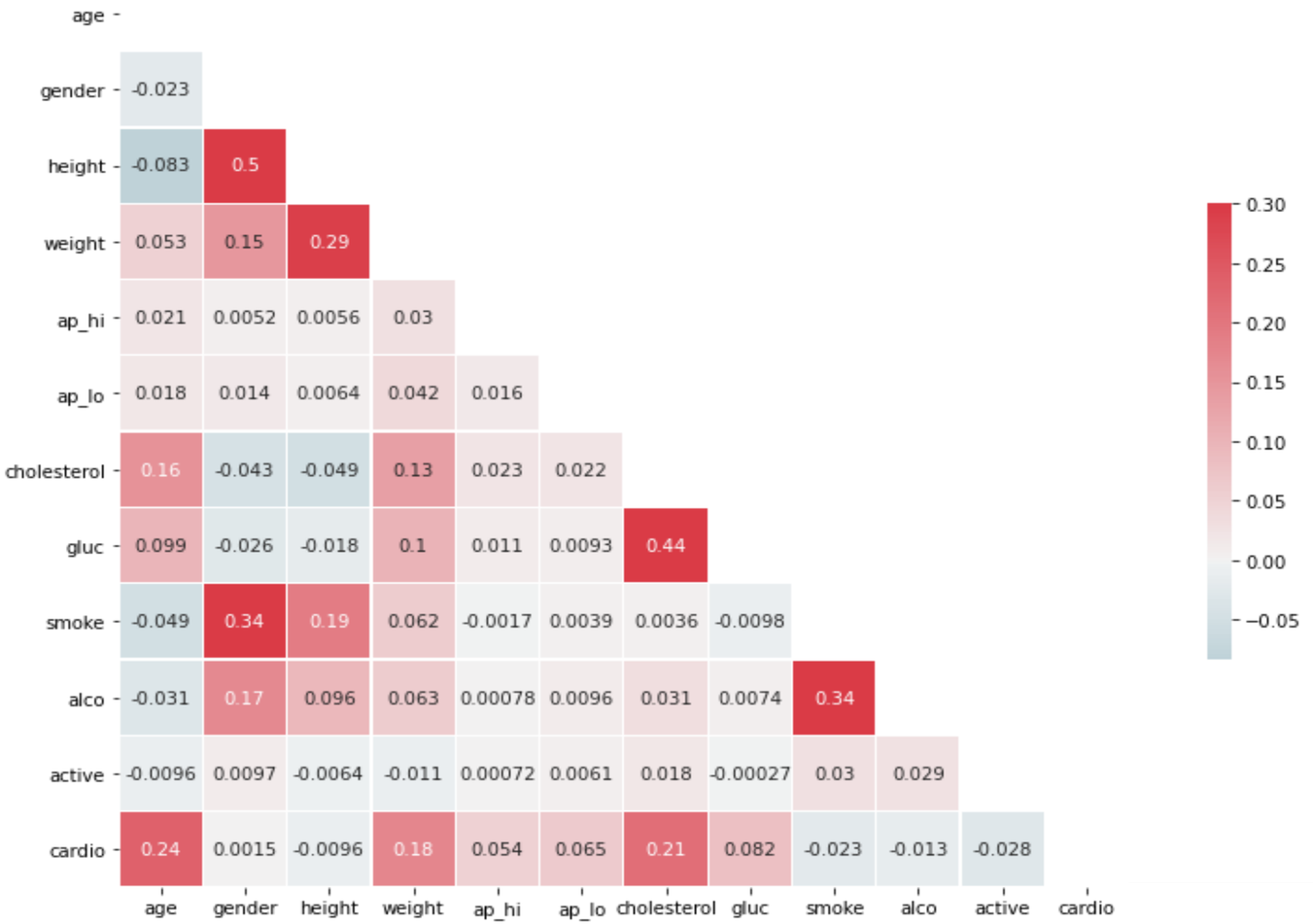
| | id | age_days | age_year | gender | height | weight | ap_hi | ap_lo | cholesterol | gluc | smoke | alco | active | cardio |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 18393 | 50.391781 | 2 | 168 | 62.0 | 110 | 80 | 1 | 1 | 0 | 0 | 1 | 0 |
| **1** | 1 | 20228 | 55.419178 | 1 | 156 | 85.0 | 140 | 90 | 3 | 1 | 0 | 0 | 1 | 1 |
| **2** | 2 | 18857 | 51.663014 | 1 | 165 | 64.0 | 130 | 70 | 3 | 1 | 0 | 0 | 0 | 1 |
| **3** | 3 | 17623 | 48.282192 | 2 | 169 | 82.0 | 150 | 100 | 1 | 1 | 0 | 0 | 1 | 1 |
| **4** | 4 | 17474 | 47.873973 | 1 | 156 | 56.0 | 100 | 60 | 1 | 1 | 0 | 0 | 0 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **69995** | 99993 | 19240 | 52.712329 | 2 | 168 | 76.0 | 120 | 80 | 1 | 1 | 1 | 0 | 1 | 0 |
| **69996** | 99995 | 22601 | 61.920548 | 1 | 158 | 126.0 | 140 | 90 | 2 | 2 | 0 | 0 | 1 | 1 |
| **69997** | 99996 | 19066 | 52.235616 | 2 | 183 | 105.0 | 180 | 90 | 3 | 1 | 0 | 1 | 0 | 1 |
| **69998** | 99998 | 22431 | 61.454795 | 1 | 163 | 72.0 | 135 | 80 | 1 | 2 | 0 | 0 | 0 | 1 |
| **69999** | 99999 | 20540 | 56.273973 | 1 | 170 | 72.0 | 120 | 80 | 2 | 1 | 0 | 0 | 1 | 0 |

70000 rows × 14 columns

https://www.kaggle.com/raminhashimzade/cardio-disease

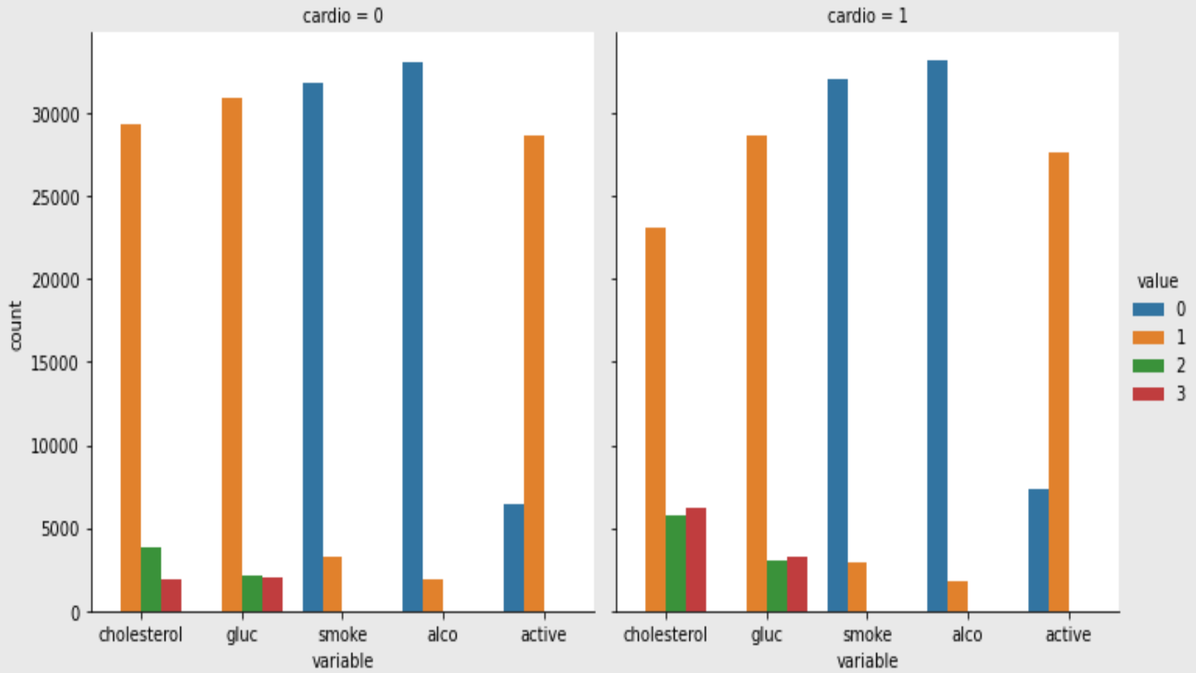
# EDA

Cardio disease is more relative with AGE, WEIGHT and CHOLESTEROL

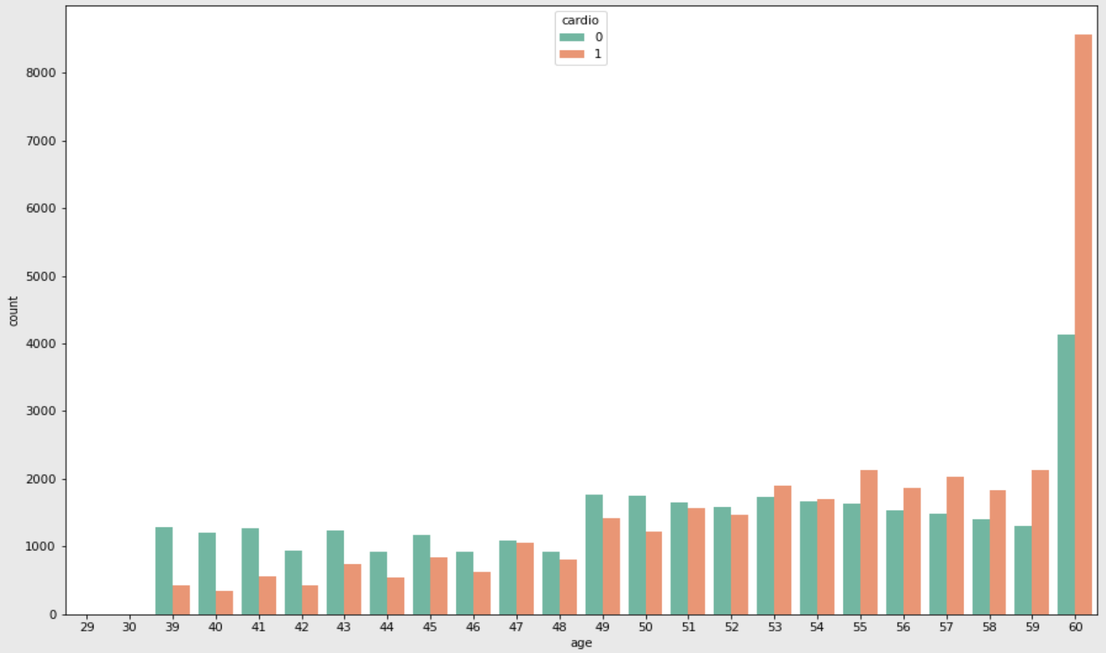


# Analysis

Patients with CVD have higher cholesterol and blood glucose level.



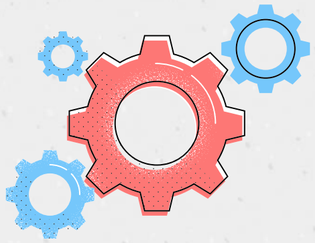

It can be observed that people over 53 of age are more exposed to CVD.



# Data Preparation

## Data

```
Data columns (total 14 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   id           70000 non-null  int64
 1   age_days     70000 non-null  int64
 2   age_year     70000 non-null  float64
 3   gender       70000 non-null  int64
 4   height       70000 non-null  int64
 5   weight       70000 non-null  float64
 6   ap_hi        70000 non-null  int64
 7   ap_lo        70000 non-null  int64
 8   cholesterol  70000 non-null  int64
 9   gluc         70000 non-null  int64
 10  smoke        70000 non-null  int64
 11  alco         70000 non-null  int64
 12  active       70000 non-null  int64
 13  cardio       70000 non-null  int64
dtypes: float64(2), int64(12)
```



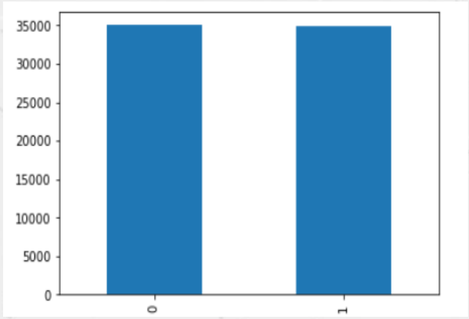
## Cleaning

- Convert the datatype.
- Check if there are any missing/ duplicate/ outlier values.
- Delete the 'id','age_days' feature

## Feature Engineering

- Drop columns (weight - height).
- Add new column (bmi).
- Drop columns that have no correlation with the target feature.

- Check imbalanced data.




- Detecting Multicollinearity with VIF.

$$BMI = \frac{\text{weight in kg}}{(\text{height in m})^2}$$

# Final Dataset

| | age | gender | ap_hi | ap_lo | cholesterol | gluc | bmi | cardio |
|---|---|---|---|---|---|---|---|---|
| **0** | 50 | 2 | 110 | 80 | 1 | 1 | 21.967120 | 0 |
| **1** | 55 | 1 | 140 | 90 | 3 | 1 | 34.927679 | 1 |
| **2** | 51 | 1 | 130 | 70 | 3 | 1 | 23.507805 | 1 |
| **3** | 48 | 2 | 150 | 100 | 1 | 1 | 28.710479 | 1 |
| **4** | 60 | 1 | 120 | 80 | 2 | 2 | 29.384676 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **59302** | 51 | 1 | 170 | 90 | 1 | 1 | 21.604105 | 1 |
| **59303** | 53 | 1 | 130 | 90 | 1 | 1 | 23.661439 | 1 |
| **59304** | 57 | 1 | 150 | 80 | 1 | 1 | 29.384757 | 1 |
| **59305** | 61 | 1 | 135 | 80 | 1 | 2 | 27.099251 | 1 |
| **59306** | 56 | 1 | 120 | 80 | 2 | 1 | 24.913495 | 0 |

59307 rows × 8 columns

# Algorithms

### Logistic Regression
Accuracy: 0.7209
Precision: 0.74917
Recall: 0.67594
F1_Score: 0.71067

### K-nearest neighbors
Accuracy: 0.71787
Precision: 0.74591
Recall: 0.67284
F1_Score: 0.70749

### Decision Tree
Accuracy: 0.61002
Precision: 0.6171
Recall: 0.60856
F1_Score: 0.6128

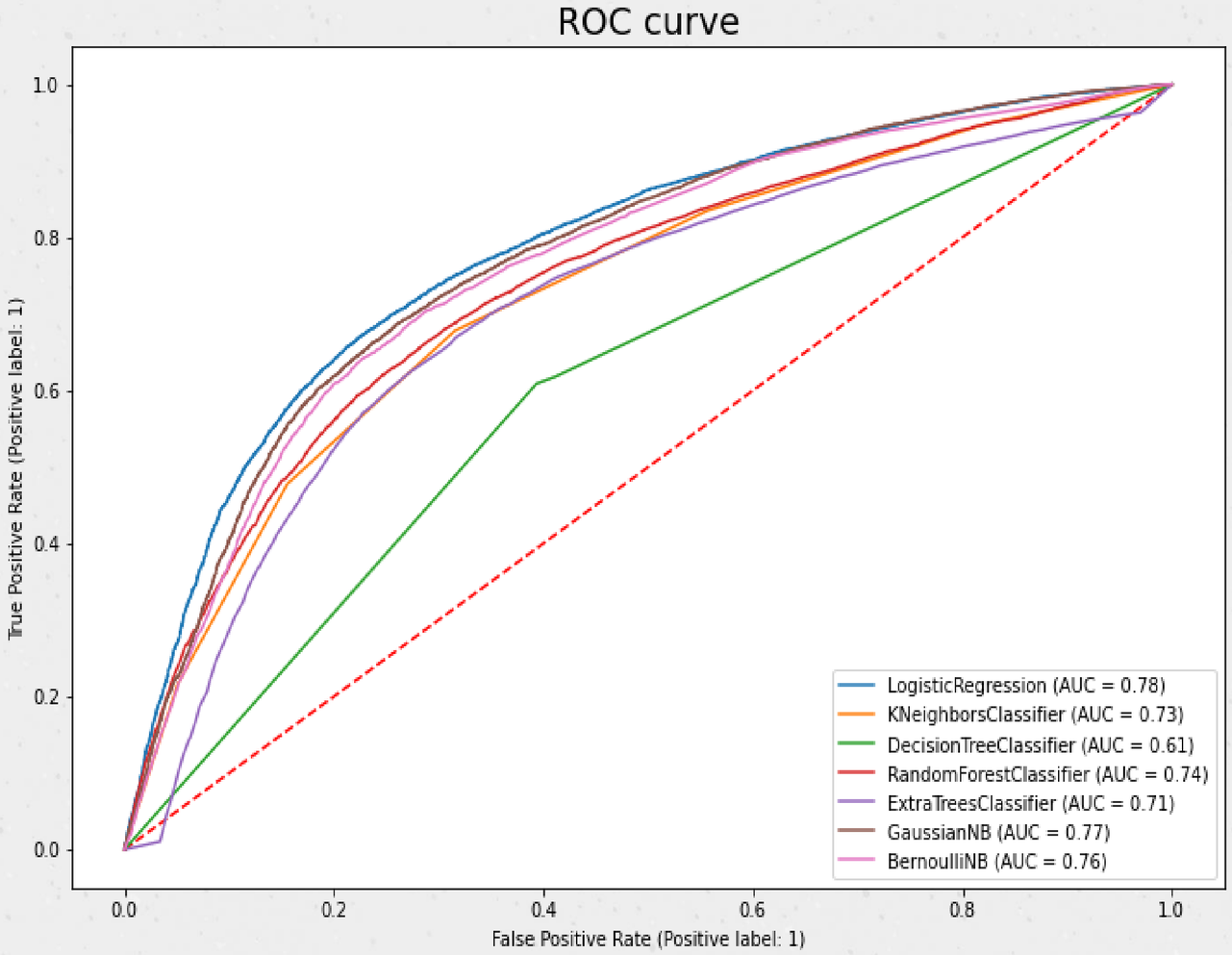
### Extra Trees
Accuracy: 0.64093
Precision: 0.64309
Recall: 0.65599
F1_Score: 0.64948

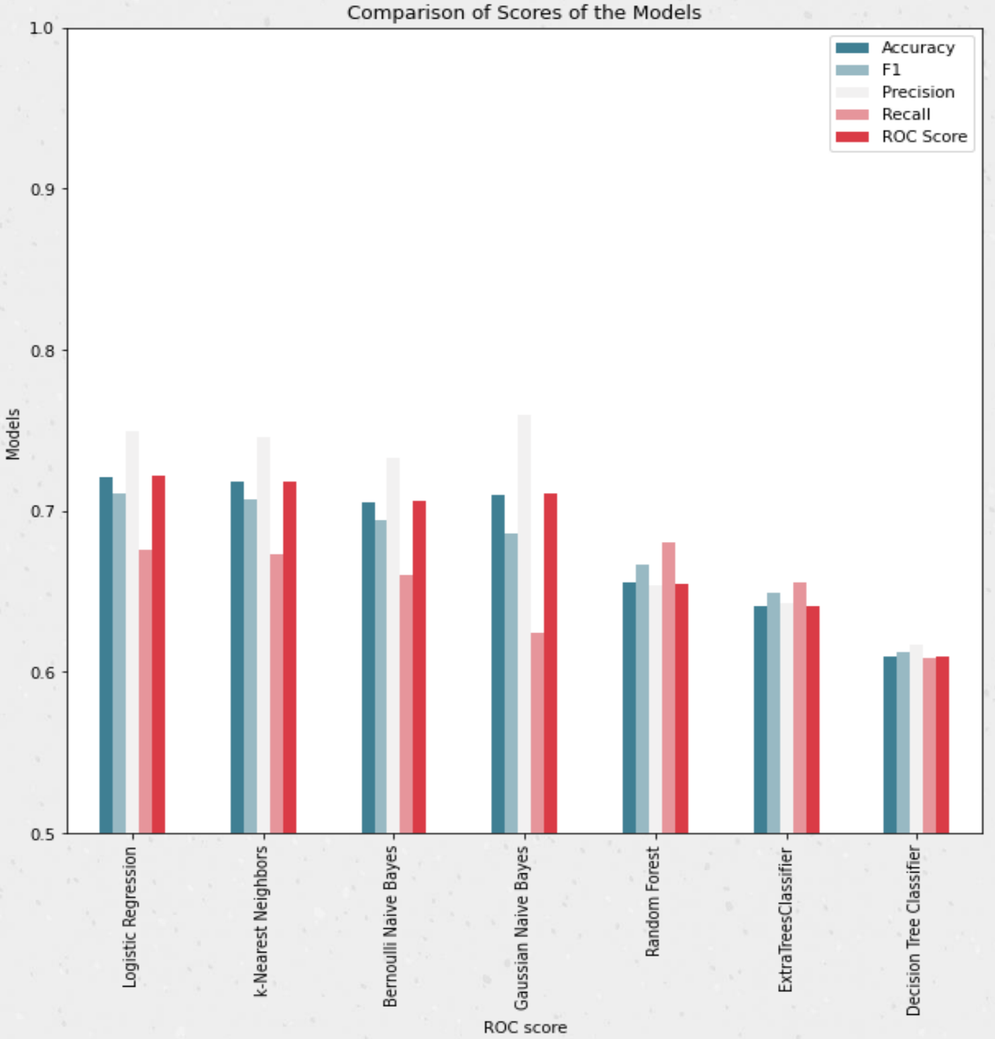
### Random Forest
Accuracy: 0.65537
Precision: 0.65406
Recall: 0.68015
F1_Score: 0.66685

### Gaussian Naive Bayes
Accuracy: 0.70944
Precision: 0.75974
Recall: 0.62452
F1_Score: 0.68552

### Bernoulli Naive Bayes
Accuracy: 0.70528
Precision: 0.73244
Recall: 0.65987
F1_Score: 0.69426

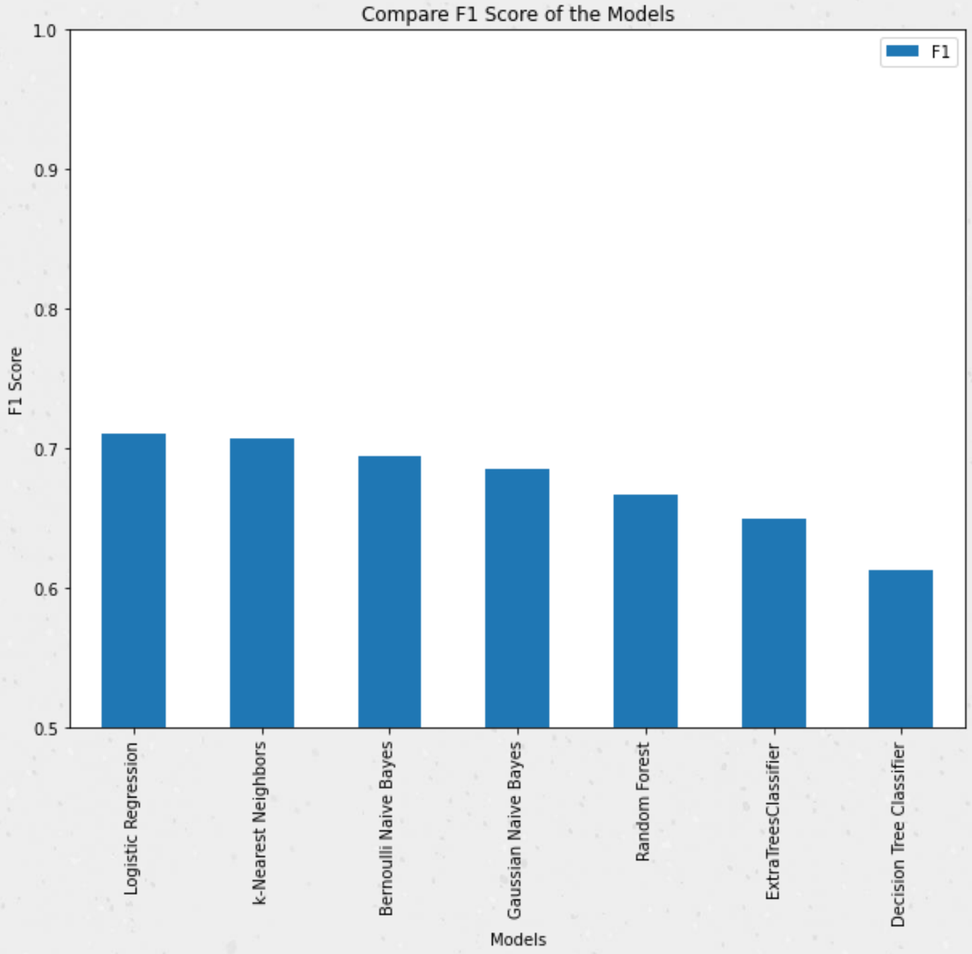
# Results

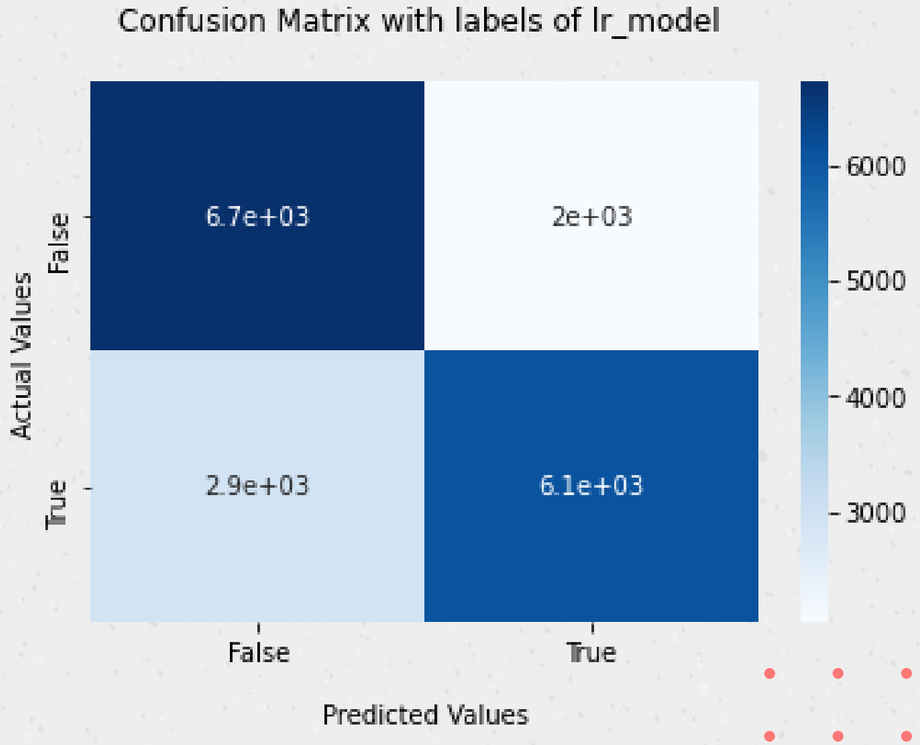

ROC curve
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0
LogisticRegression (AUC = 0.78)
KNeighborsClassifier (AUC = 0.73)
DecisionTreeClassifier (AUC = 0.61)
RandomForestClassifier (AUC = 0.74)
ExtraTreesClassifier (AUC = 0.71)
GaussianNB (AUC = 0.77)
BernoulliNB (AUC = 0.76)


Comparison of Scores of the Models
Accuracy
F1
Precision
Recall
ROC Score
Models
ROC score
1.0
0.9
0.8
0.7
0.6
0.5
Logistic Regression
k-Nearest Neighbors
Bernoulli Naive Bayes
Gaussian Naive Bayes
Random Forest
ExtraTreesClassifier
Decision Tree Classifier

# Best Model

## Logistic Regression



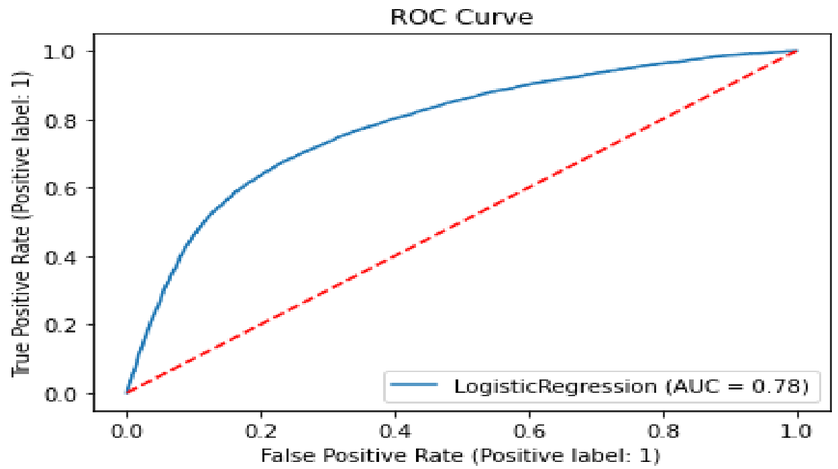



# Conclusion

## Logistic Regression

shows the best results.

| | Model | Accuracy | F1 | Precision | Recall | ROC Score | Score_train | Score_test | Score_diff |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Logistic Regression | 0.72090 | 0.71067 | 0.74917 | 0.67594 | 0.72155 | 71.75 | 72.09 | 0.34 |




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.70 | 0.77 | 0.73 | 8770 |
| 1 | 0.75 | 0.68 | 0.71 | 9023 |
| accuracy | | | 0.72 | 17793 |
| macro avg | 0.72 | 0.72 | 0.72 | 17793 |
| weighted avg | 0.72 | 0.72 | 0.72 | 17793 |

# Recommendation

- It can be observed that people over 53 of age are more exposed to CVD.
- It can be seen that patients with CVD have higher cholesterol and blood glucose level. And, generally speaking less active.
- The percentage is the same in females and men of patients with CVD.

## Improvement:

- Increase Dataset.
- More features like the history of patients.
- Use more other algorithms.

# Tools

- Technologies:

- Python.
- Jupyter Notebook.

- Libraries:

- Pandas
- Pandas_profiling
- Statsmodels
- Pickle
- Matplotlib
- Seaborn
- NumPy
- Sklearn
- sqlalchemy

# Thank you for listening