Diagnosing Cardiovascular Disease

FATIMAH O ALAHMADI

Cardiovascular diseases have always been one of the most common causes of death globally. According to the World Health Organization (An estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths.).

Project Goal

Diagnosing the cardiovascular disease based on several features and symptoms given by the client I will use the features to determine if the disease exists or not in order to be able to warn the client and notify him either way.

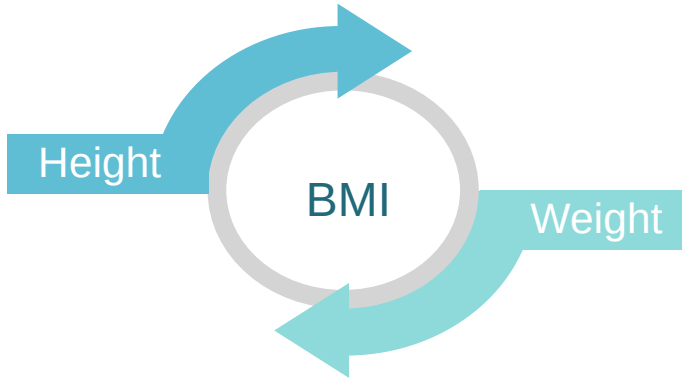
Project Dataset

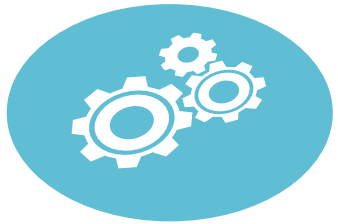
Feature	Type of Feature	name for Feature in data	Data type
ID	Objective	Id	**int**
Age	Objective	age	**int (days)**
Height	Objective	height	**int (cm)**
Weight	Objective	weight	**float (kg)**
Gender	Objective	gender	**categorical**
Systolic blood pressure	Examination	ap_hi	**int**
Diastolic blood pressure	Examination	ap_lo	**int**
Cholesterol	Examination	cholesterol	**1: normal, 2: above normal, 3: well above normal**
Glucose	Examination	gluc	**1: normal, 2: above normal, 3: well above normal**
Smoking	Subjective	smoke	**binary**
Alcohol intake	Subjective	alco	**binary**
Physical activity	Subjective	active	**binary**
Presence or absence of cardiovascular disease	Target	cardio	**binary**

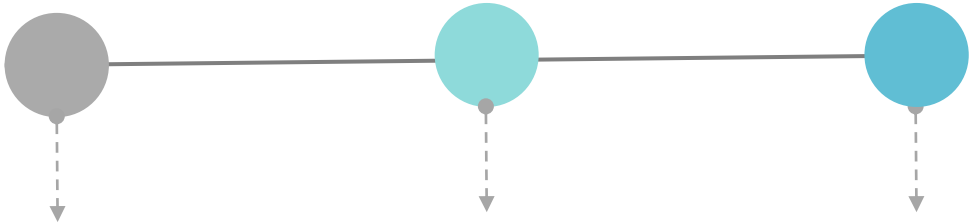
-The dataset from medical examination which were collected at the moment of medical examination.

-Dataset consists of 70000 records of patients data, 12 features + target link

Data Cleaning



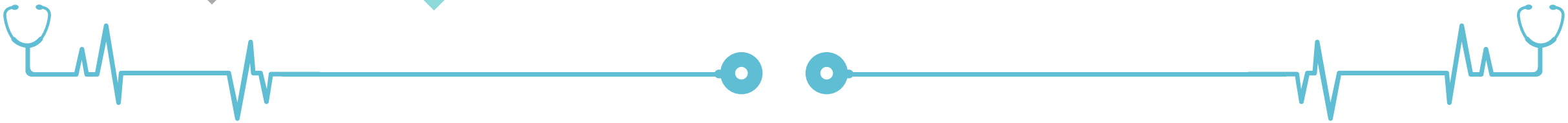




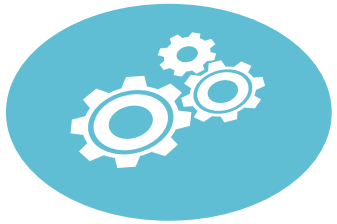
Change age from days to years

Check duplicate

Check missing



TOOLS:

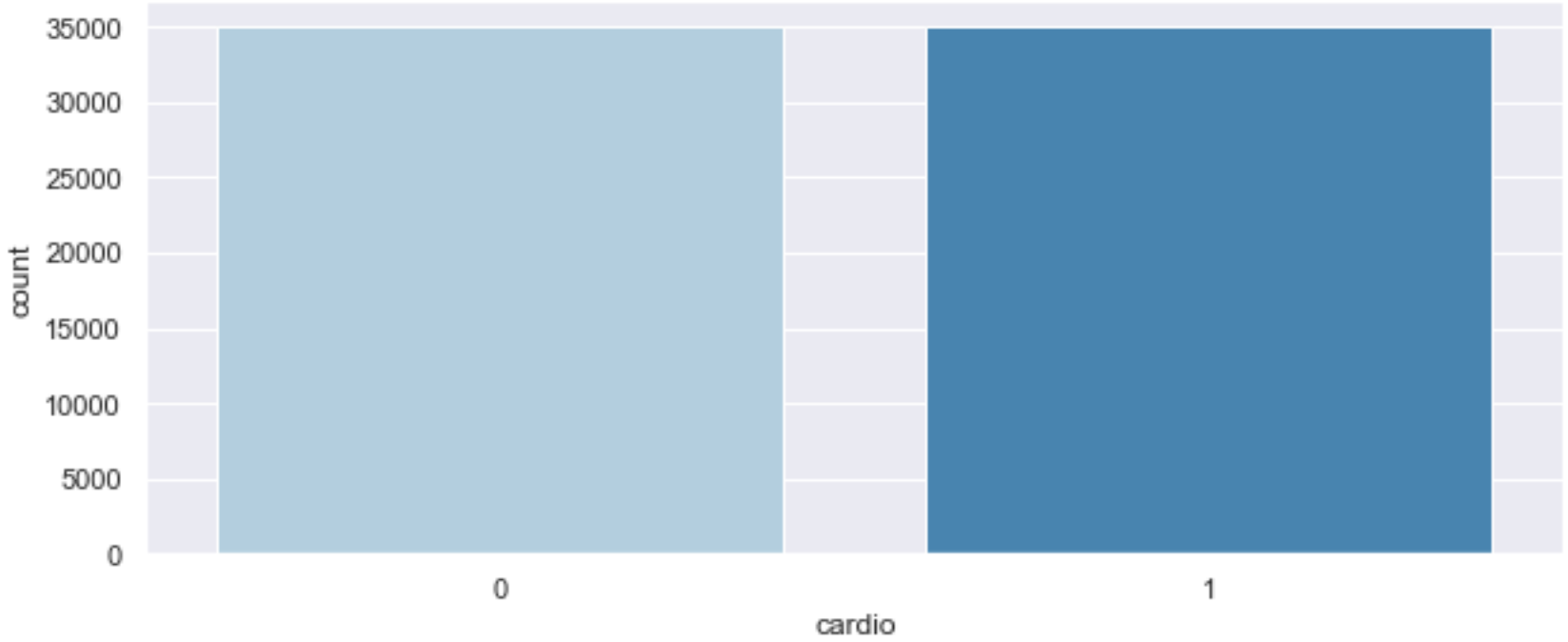


- Programming Language: Python
- Environment: Jupyter notebook
- I used different types of Python libraries for data science :

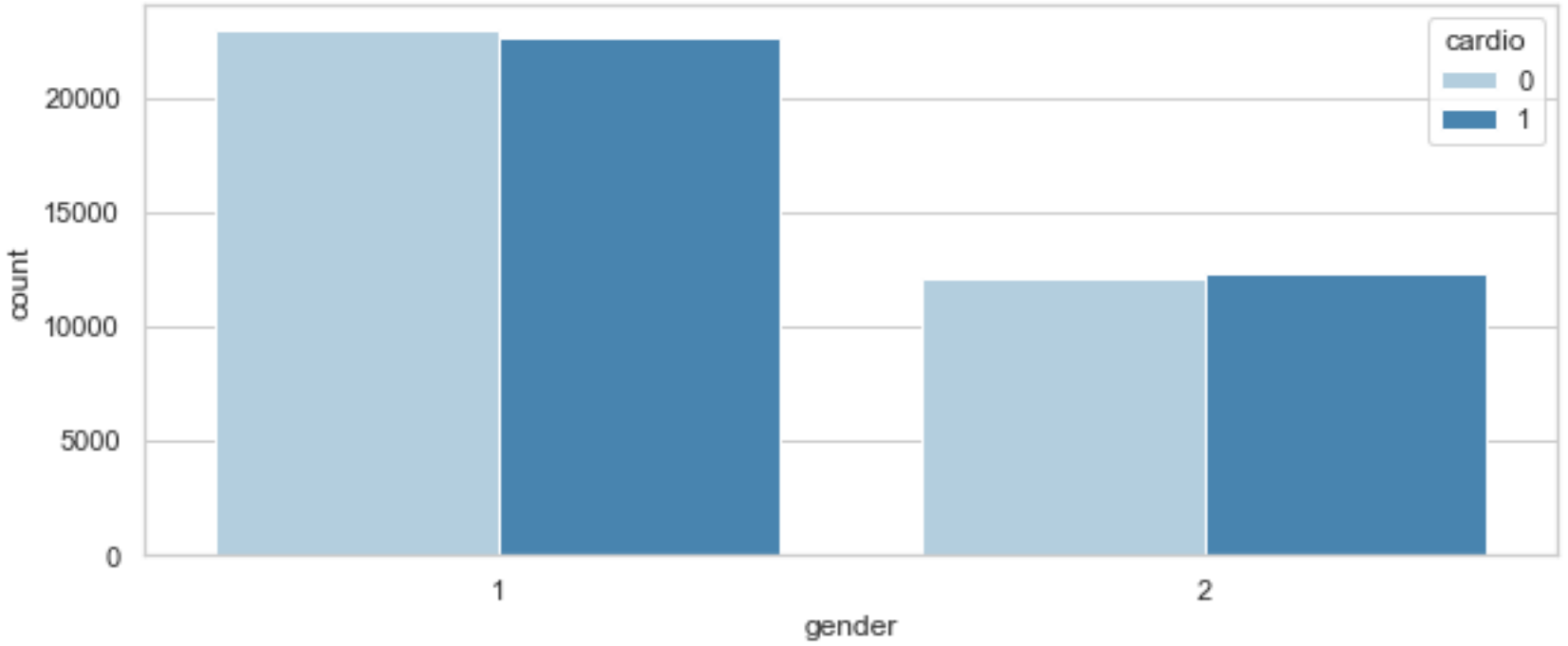
• NumPy
• Seaborn
• Pandas
• Matplotlib
• SciKit-Learn
• Xgboost

EDA

Cardiovascular Disease Cases count

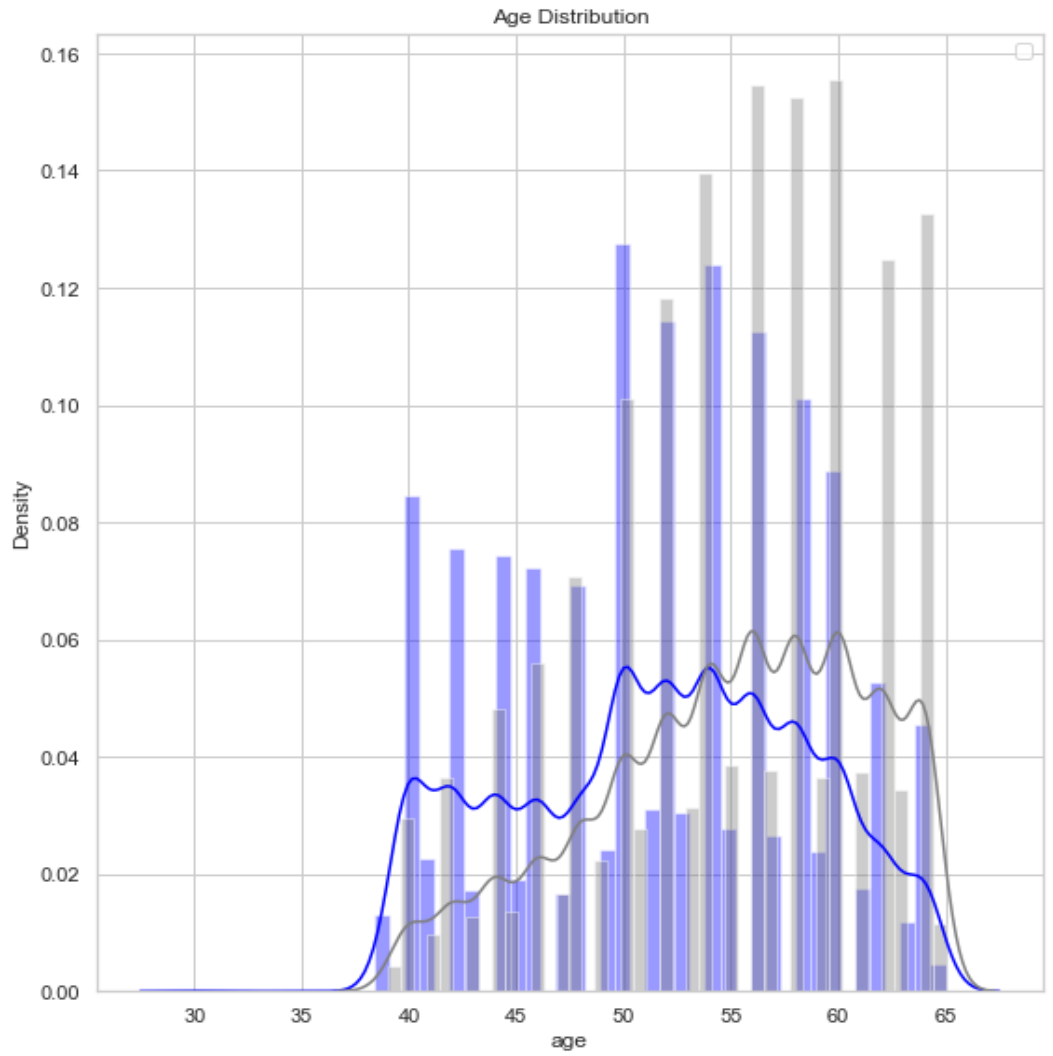


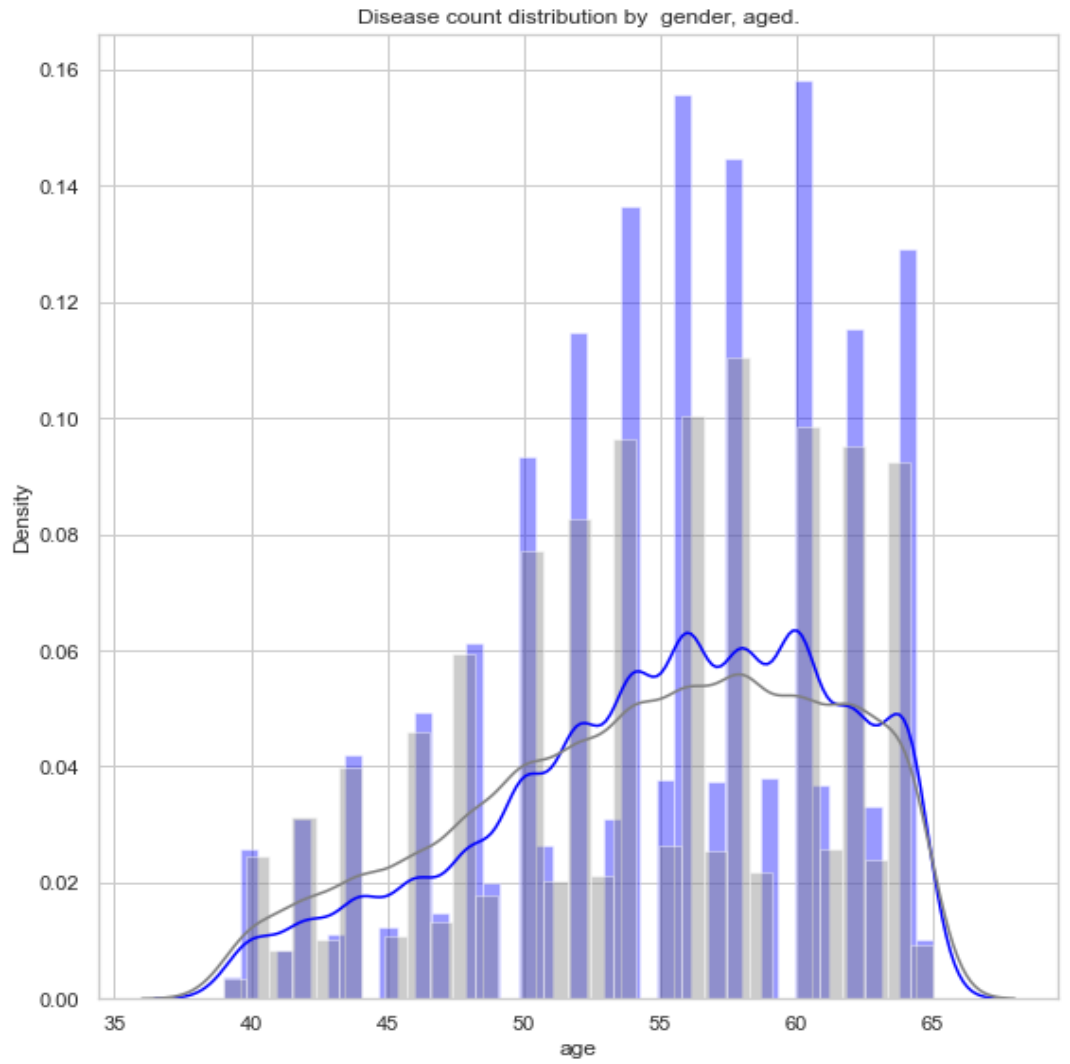
Cardiovascular Disease Cases count



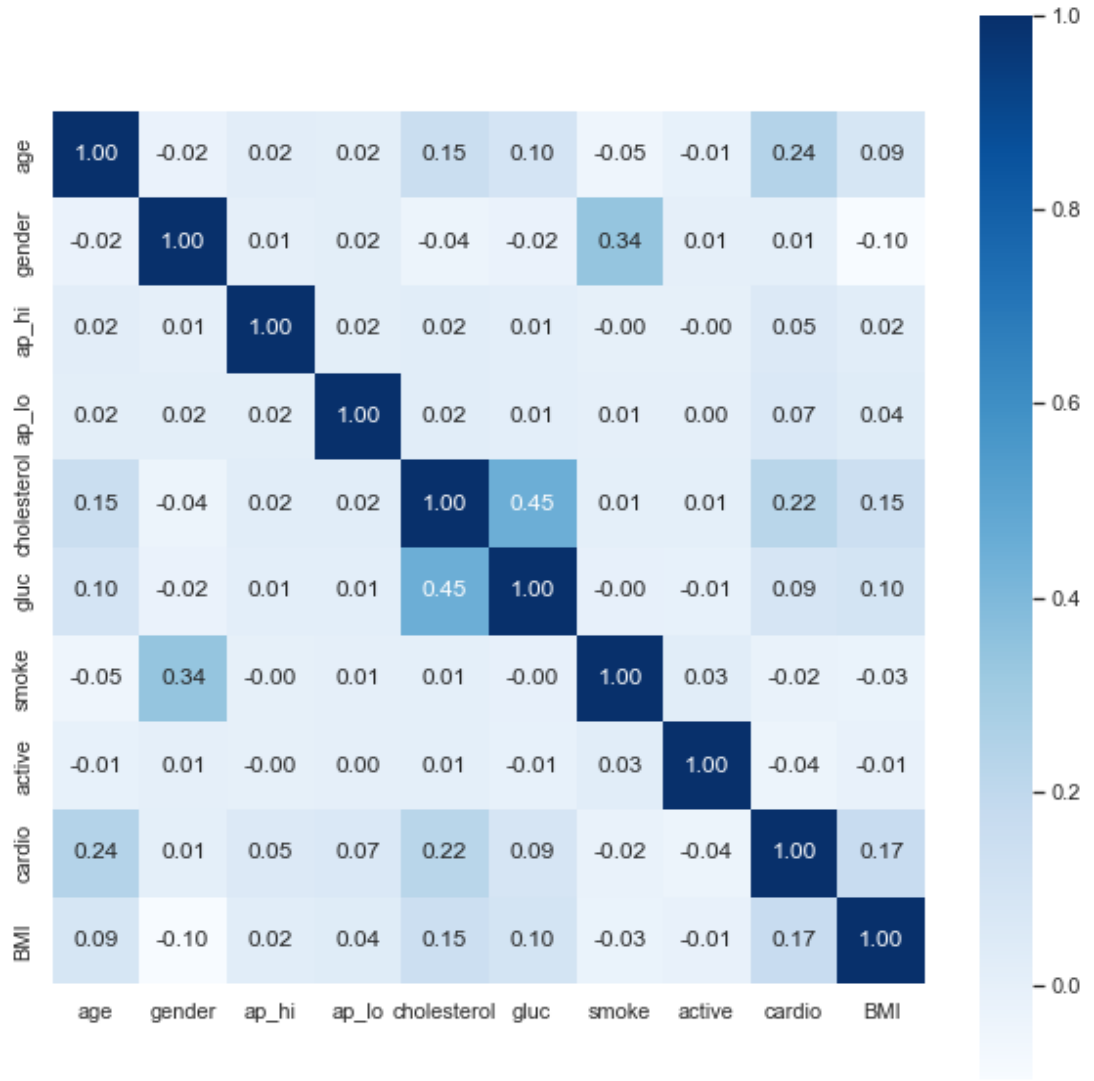
1 for (Female)

2 for (male)


Age Distribution
Density
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
30
35
40
45
50
55
60
65
age


Disease count distribution by gender, aged.
Density
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
35
40
45
50
55
60
65
age

Correlation of variables

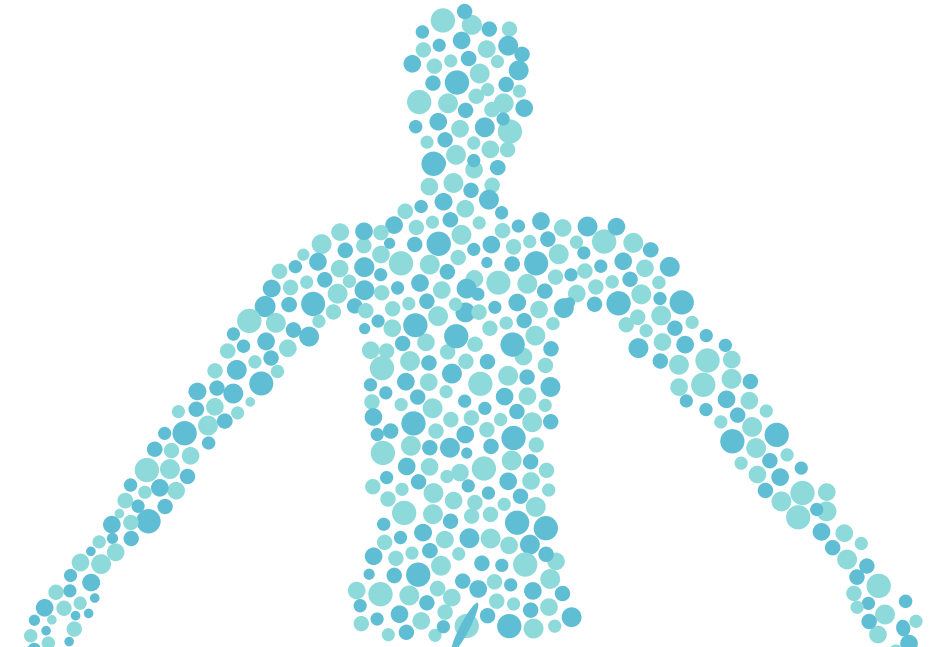


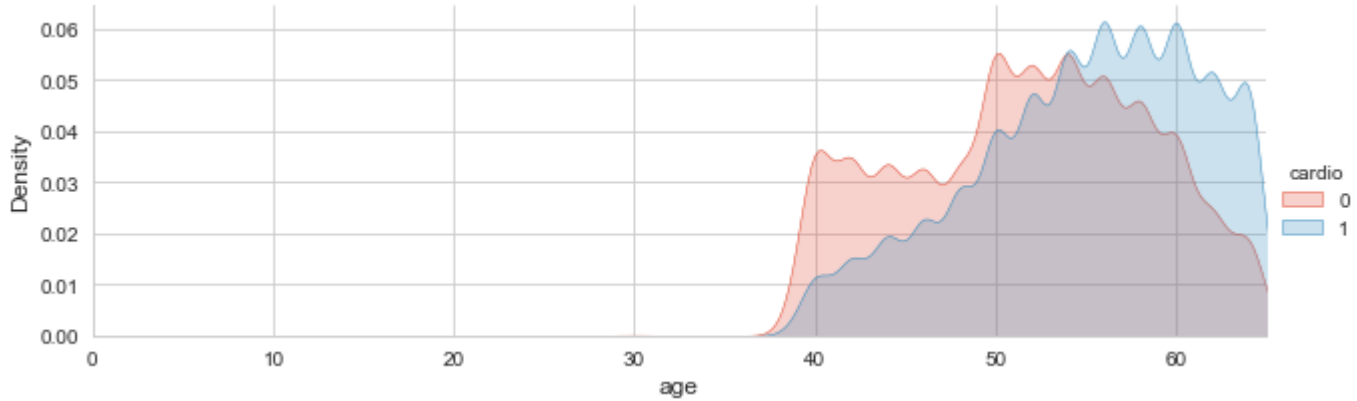
MODELLING

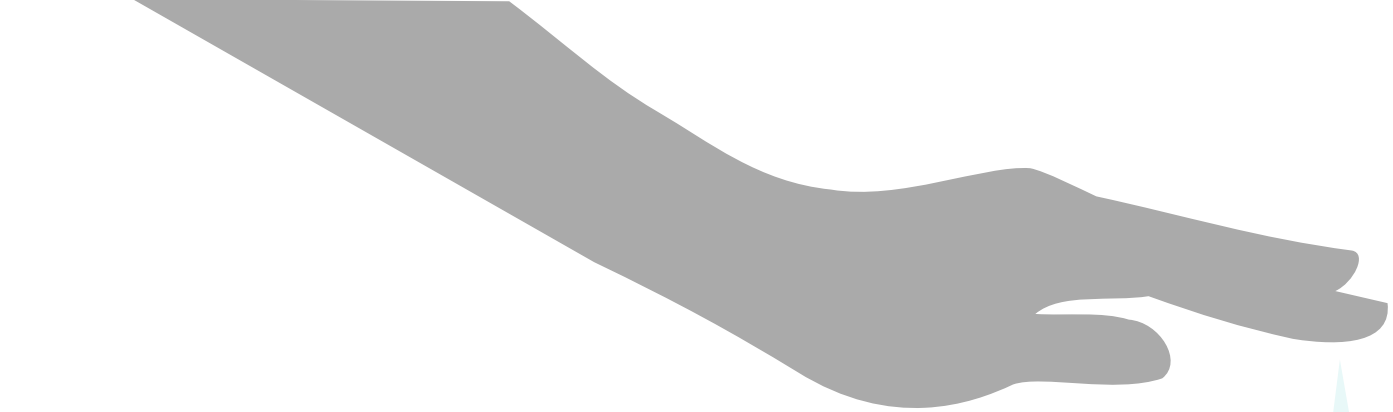
Model	Accuracy
Decision Tree	64%
Bagging	68%
Random Forest	72%
Logistic Regression	72%
XGBoost	74%

CONCLUSION

- The best model was XGBoost
- Knowing the features that cause the disease helps reduce the injury by treating these features.
- In the future, all age groups should be added because of the recent spread of Cardiovascular Disease in age groups under 39.







Thank You