Unveiling Cardiovascular Health Patterns using Statistical and Predictive Analytics

AIT 664-DL3: Represent, Process & Vizualize Applied Information Technology

TEAM 1

Dr. Ebrima N Ceesay

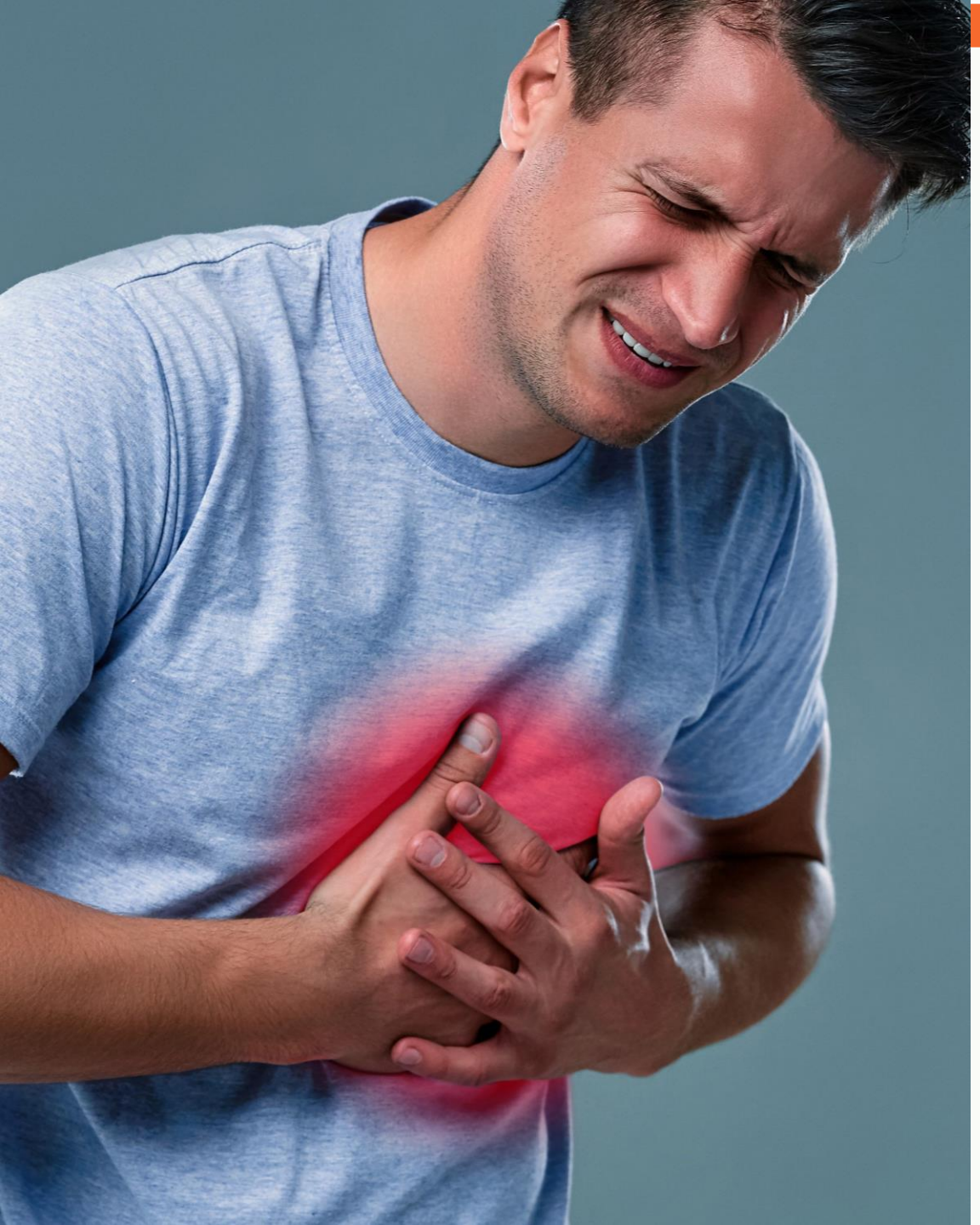
TEAM 1

ABHISHEK ANUMALLA

AAKASH BOENAL

SHASHANK YELAGANDULA

PAVAN TEJAVATH



DESCRIPTION OF THE PROBLEM

Problem Statement:
Cardiovascular disease (CVD) encompasses a spectrum of conditions affecting the heart and blood vessels, including coronary artery disease, stroke, and heart failure.

Global Impact:
Despite advancements in medical science, CVD remains the leading cause of mortality worldwide, responsible for millions of deaths annually.

Magnitude of the Issue:
According to the World Health Organization (WHO), CVD accounts for approximately 31% of all global deaths, highlighting its significant impact on public health.

Rising Incidence:
Factors such as aging populations, sedentary lifestyles, and unhealthy diets contribute to the increasing prevalence of CVD, particularly in developing regions.

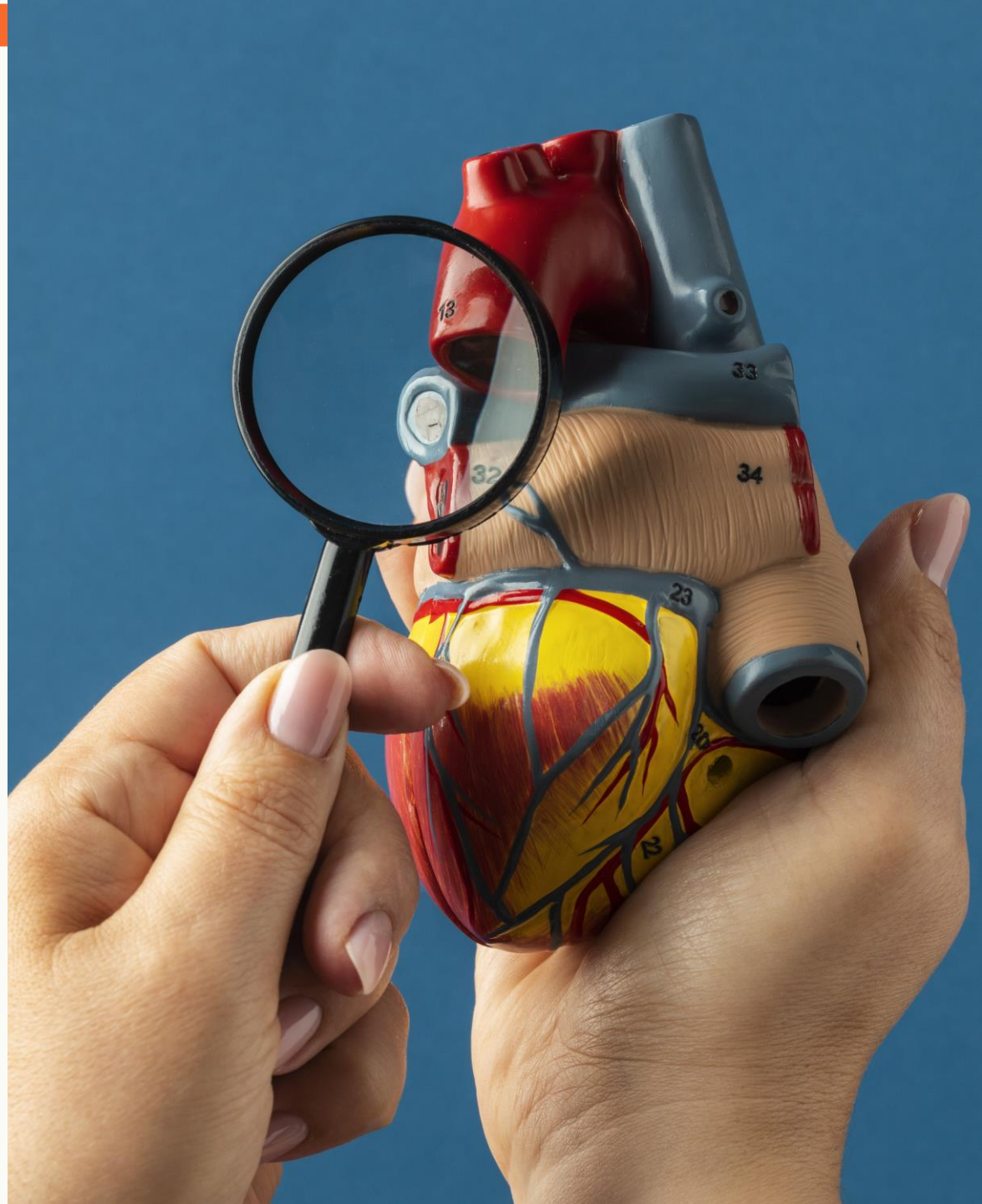
MOTIVATION

Cardiovascular disease (CVD) is the leading cause of global mortality, encompassing conditions like heart disease and stroke.

Need for Action:
Addressing the problem of cardiovascular disease is essential to reduce the burden of preventable deaths and improve public health outcomes worldwide.

Prevention is Key:
Despite advancements in medical science, CVD remains the leading cause of mortality worldwide, responsible for millions of deaths annually.

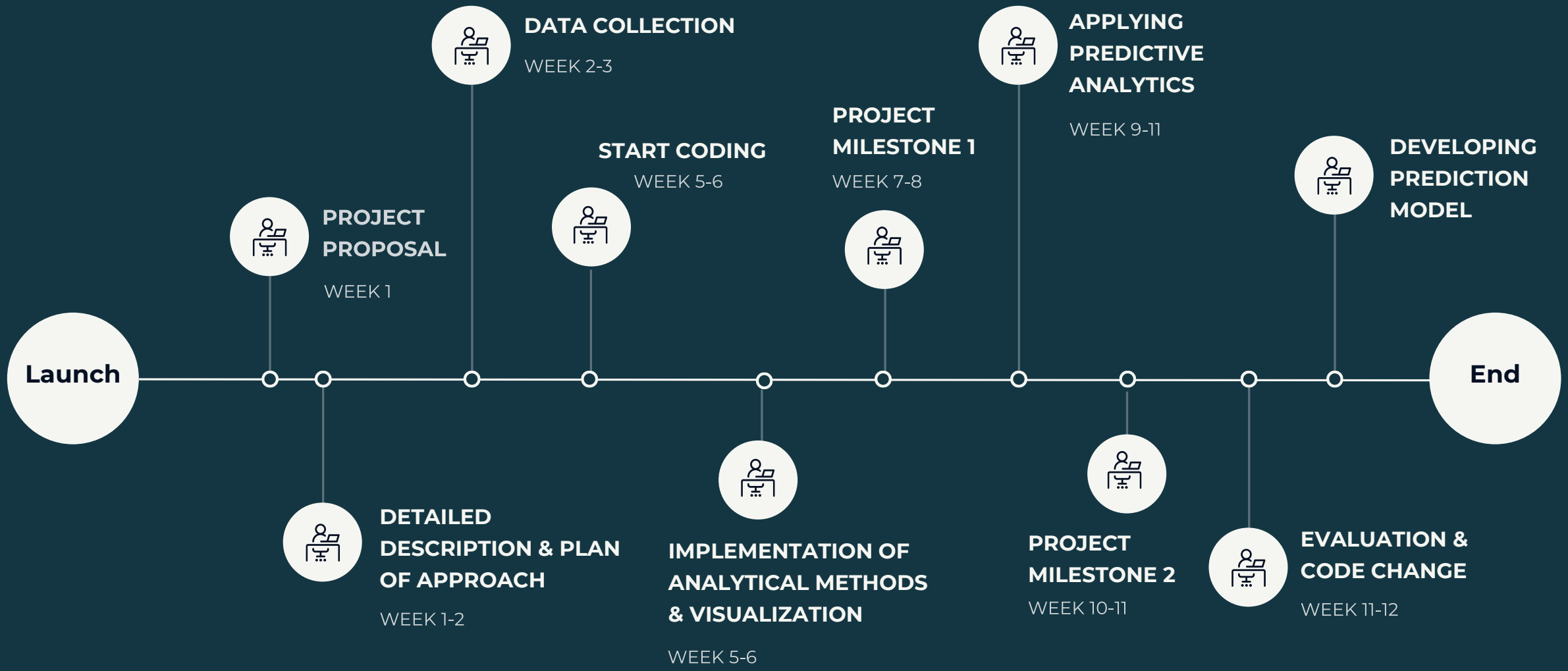
Research and Innovation:
•Investing in research and innovation is essential to develop new treatments, improve diagnostic methods, and implement evidence-based interventions for preventing and managing CVD.



AGENDA

Prediction Model overview	Data Cleaning	Statistical Analysis
Data Visualization	Prediction Tool overview	Analysis of Results

PROJECT TIMELINE

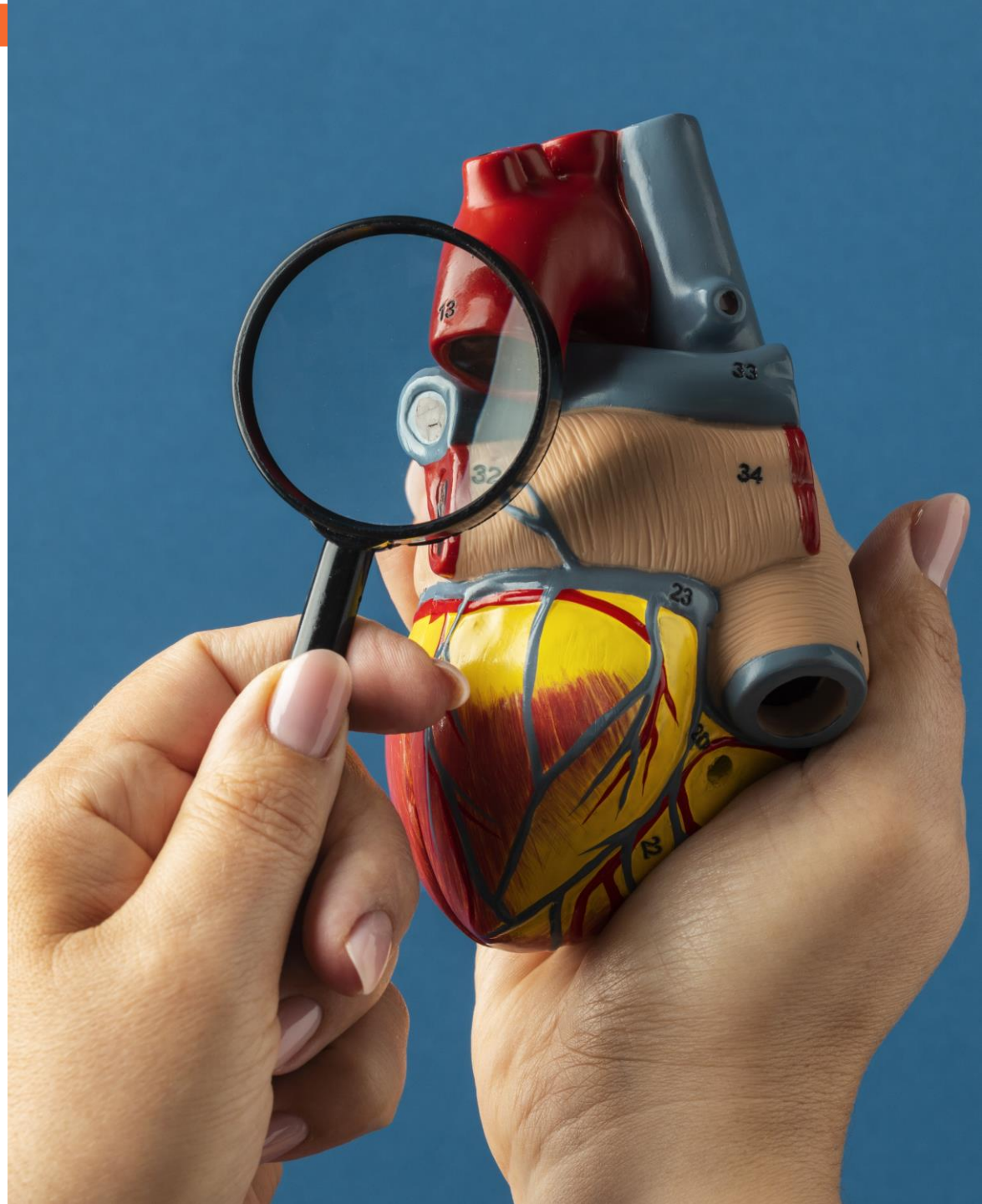


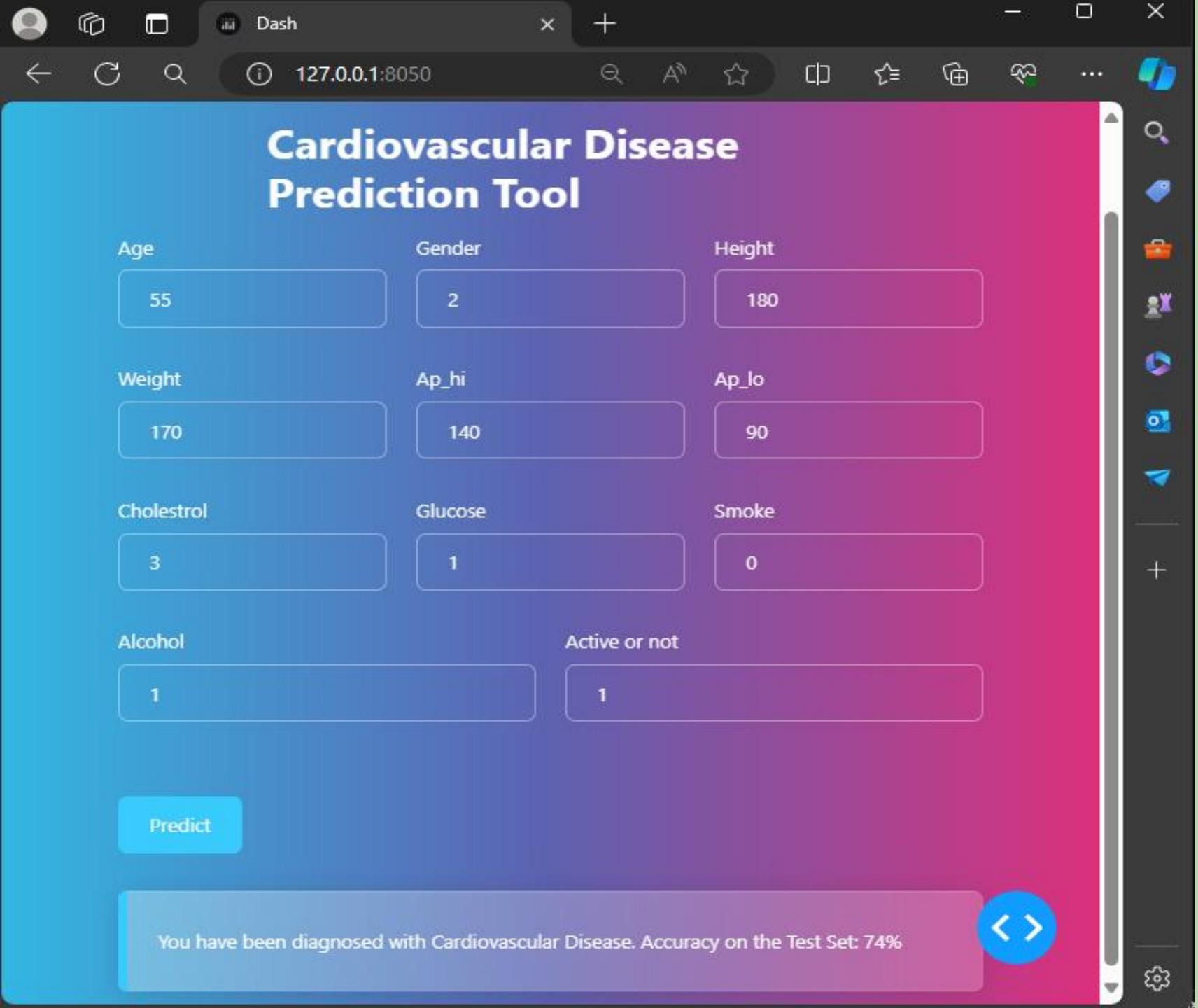
CURRENT WORK

- The current work on cardiovascular health emphasizes specific predictive analytics methods and models.
- These works use varying datasets, often smaller in size or scope. As it is personal information, it is hard to get hold of healthcare data in large amounts.
- Additionally, existing researches may not always prioritize stakeholder engagement or the creation of practical tools for end users, sometimes resulting in outcomes that are more academic than applicable.

OUR WORK

- In contrast, our work stands out by leveraging a larger and more diverse dataset of 70,000 records with 12 comprehensive features.
- We employed a variety of advanced machine learning models, including decision trees and gradient boosting classifiers, allowing us to identify patterns and insights across a range of demographic, clinical, and lifestyle factors.
- The development of a user-friendly heart disease prediction tool emphasizes on interpretability and stakeholder engagement. This focus on translating research insights into actionable strategies for healthcare providers and policymakers, as well as creating practical tools for individuals, demonstrates our clear commitment to producing real-world impact and improving public health outcomes.



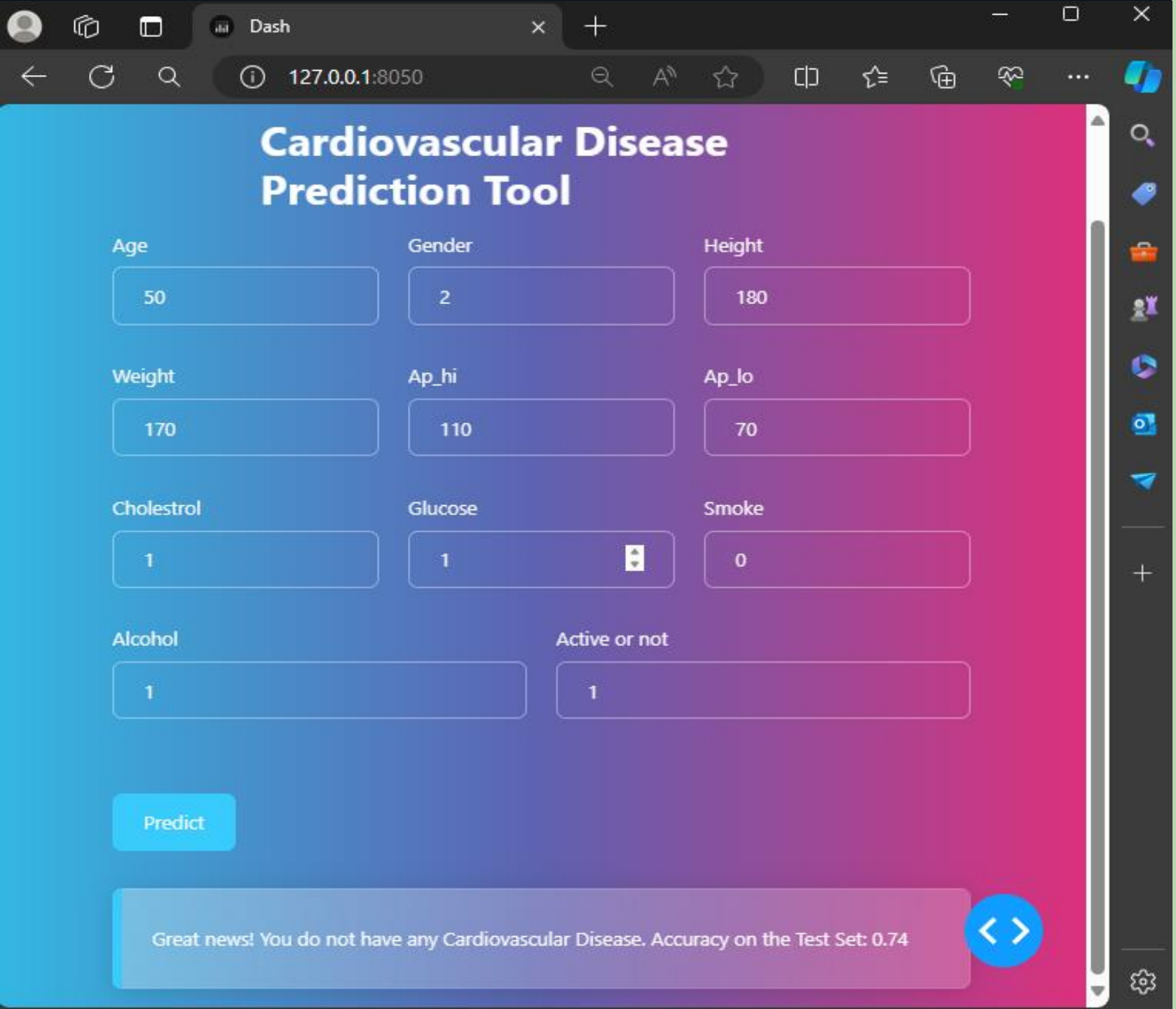


PREDICTION TOOL

The model is designed on the DASH APP Framework of Python.

So, what is Dash ?

- Dash is an open-source Python framework used for building analytical web applications.
- It is basically used by data scientist who don't have in-depth knowledge of Web Applications.
- This framework has a backend and a front-end part which helps us to create interactive analytical websites which show and can interact with the user.



The Prediction Result

DATA CLEANING

- There were no null values in the dataset but there were a few discrepancies in the data and those had to be dealt with.
- Firstly, the age column was recorded in terms of days. That has been converted to years.
- Upon carefully examining the dataset statistics, it was found that ap_hi and ap_lo columns had many outliers and they had to be dealt with. Hence, removed the records that had abnormal values in those columns. Ex: 1000,11000,-150 etc.
- The id column was dropped from the dataset.

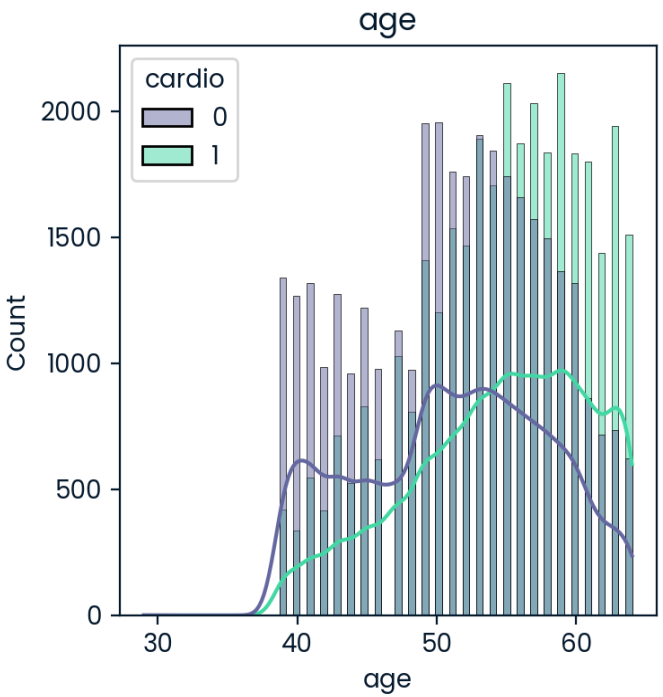
Index	age	ap_hi	ap_lo
count	70000	70000	70000
mean	19468.9	128.818	96.6304
std	2467.25	154.011	188.473
min	10798	-150	-70
25%	17664	120	80
50%	19703	120	80
75%	21327	140	90
max	23713	16020	11000

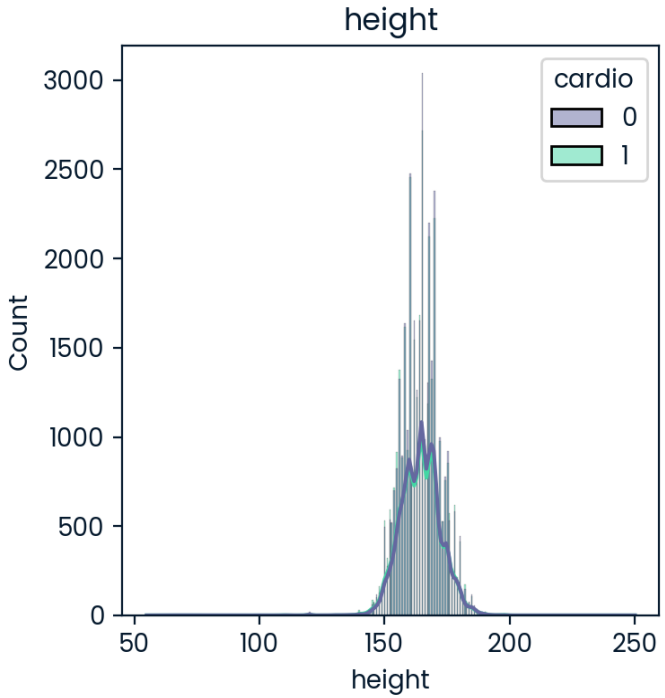
FINAL DATASET

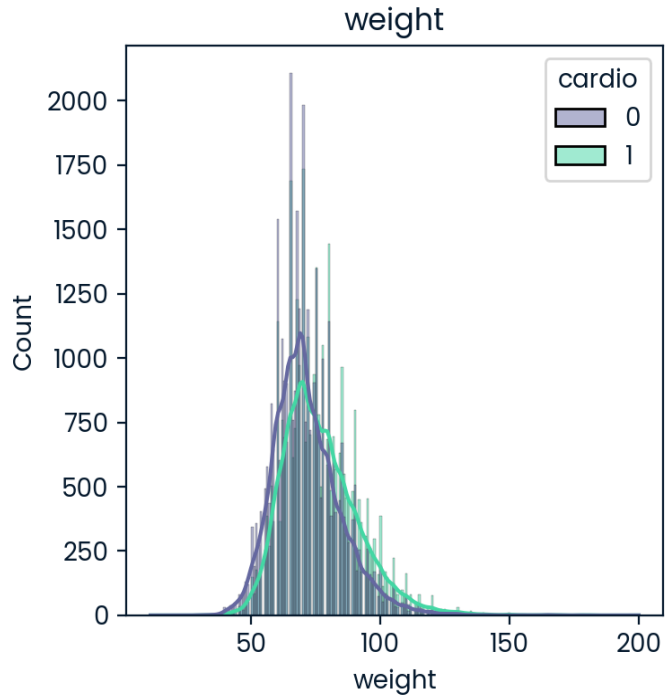
Index	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
count	68692	68692	68692	68692	68692	68692	68692	68692	68692	68692	68692	68692
mean	52.8286	1.34863	164.362	74.1178	126.674	81.2904	1.3646	1.22569	0.0879433	0.0533395	0.803369	0.49473
std	6.76901	0.47654	8.18313	14.3309	16.6951	9.50595	0.678831	0.571536	0.283214	0.224711	0.397454	0.499976
min	29	1	55	11	60	20	1	1	0	0	0	0
25%	48	1	159	65	120	80	1	1	0	0	1	0
50%	53	1	165	72	120	80	1	1	0	0	1	0
75%	58	2	170	82	140	90	2	1	0	0	1	1
max	64	2	250	200	240	182	3	3	1	1	1	1

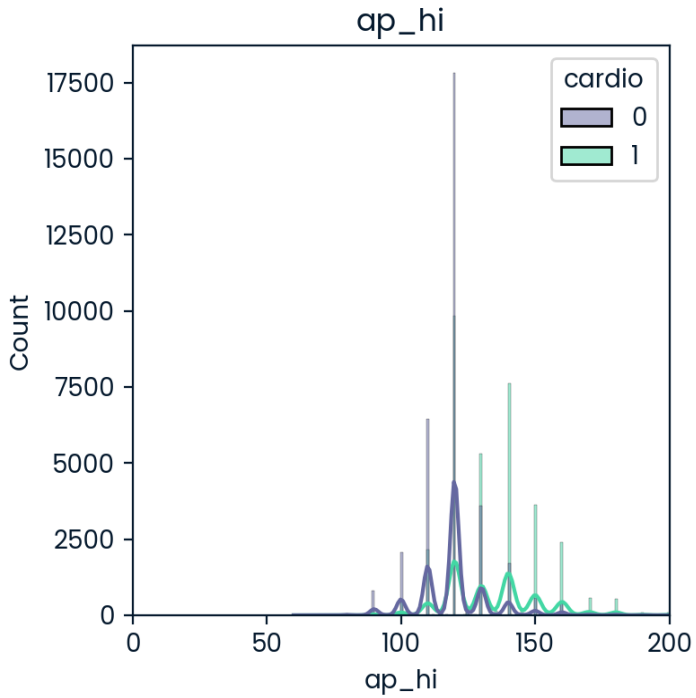
Rows	Columns
68692	12

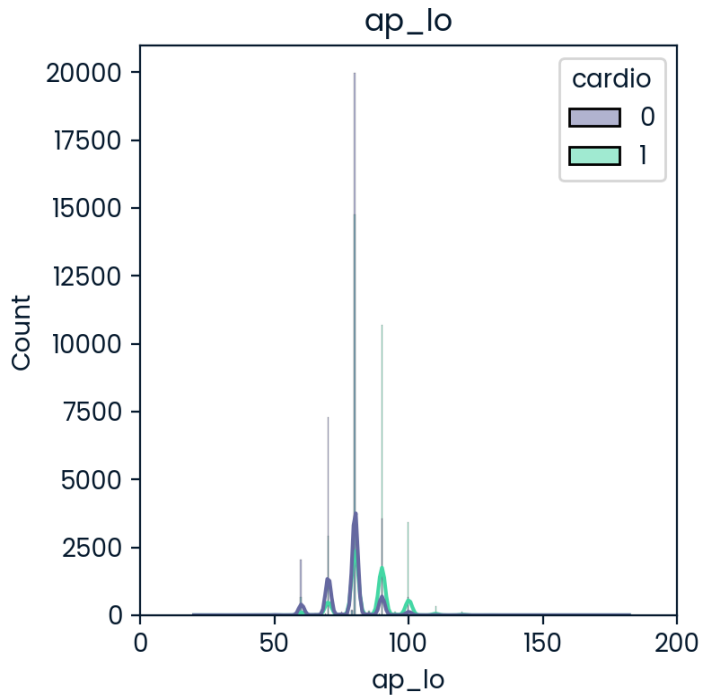
STATISTICAL ANALYSIS DATA DISTRIBUTION

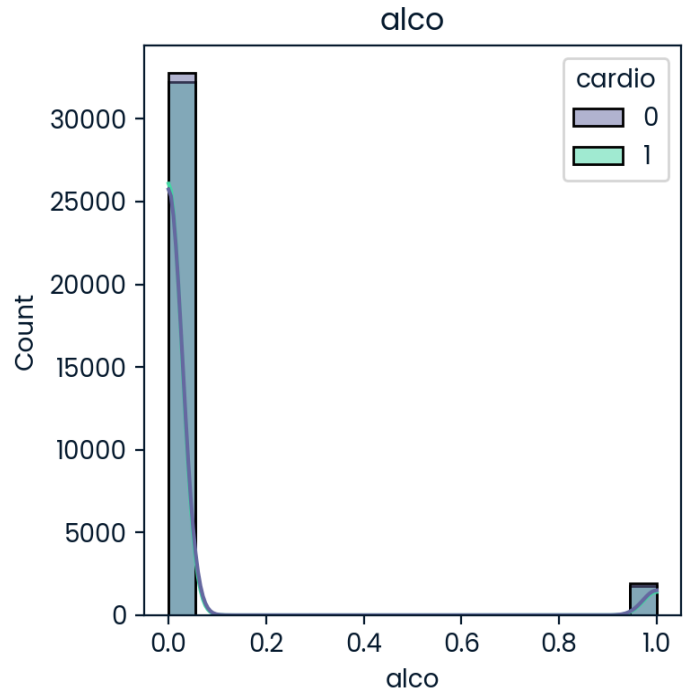




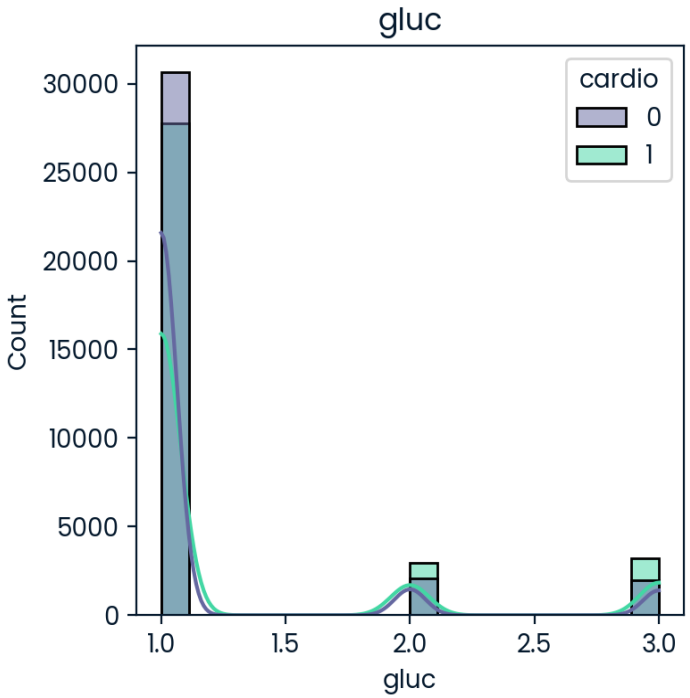


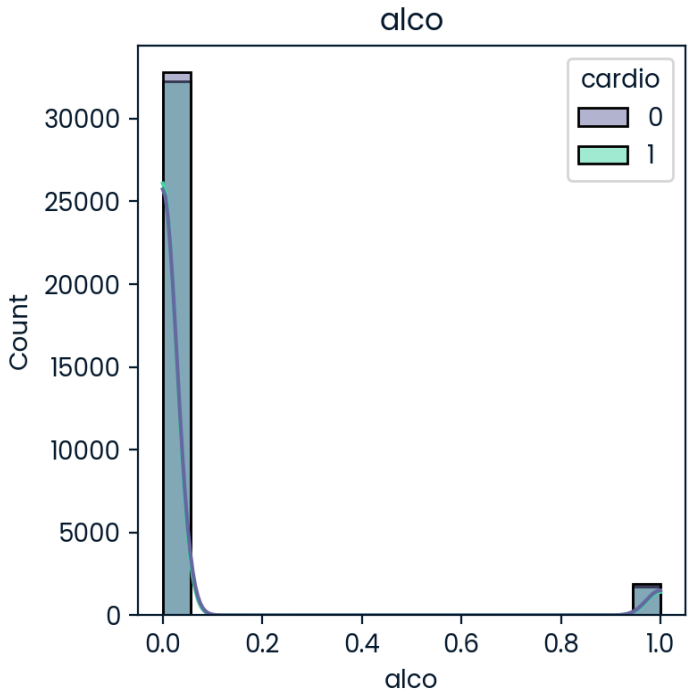


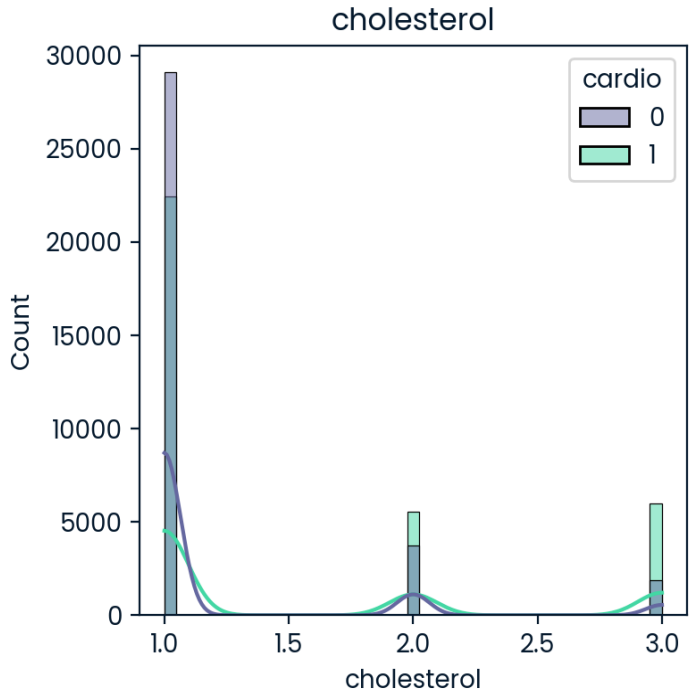


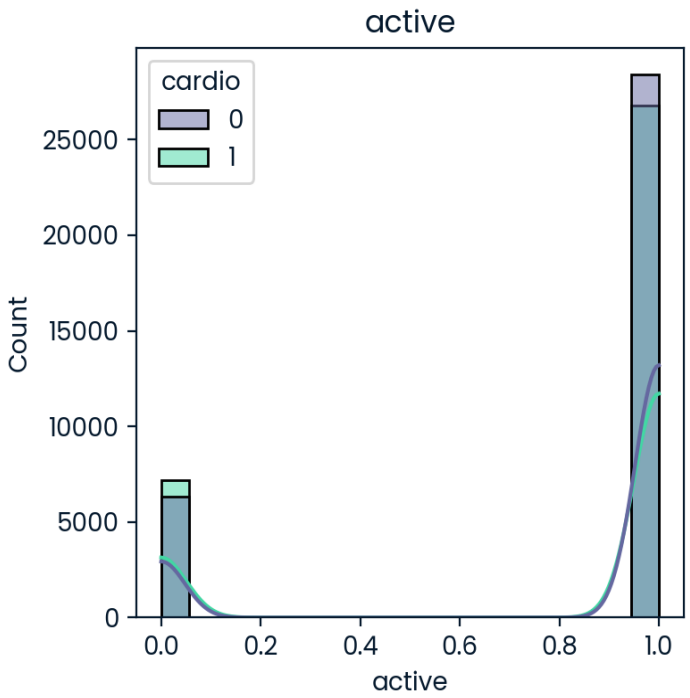


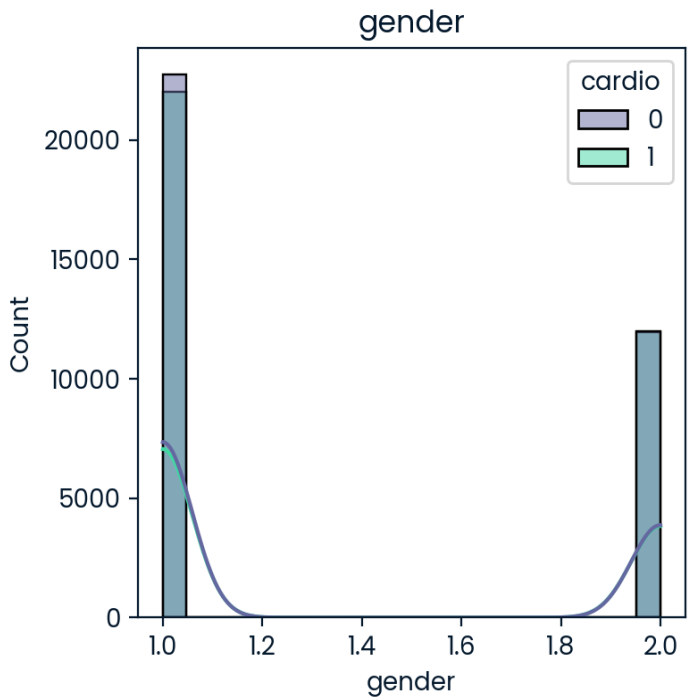
STATISTICAL ANALYSIS DATA DISTRIBUTION

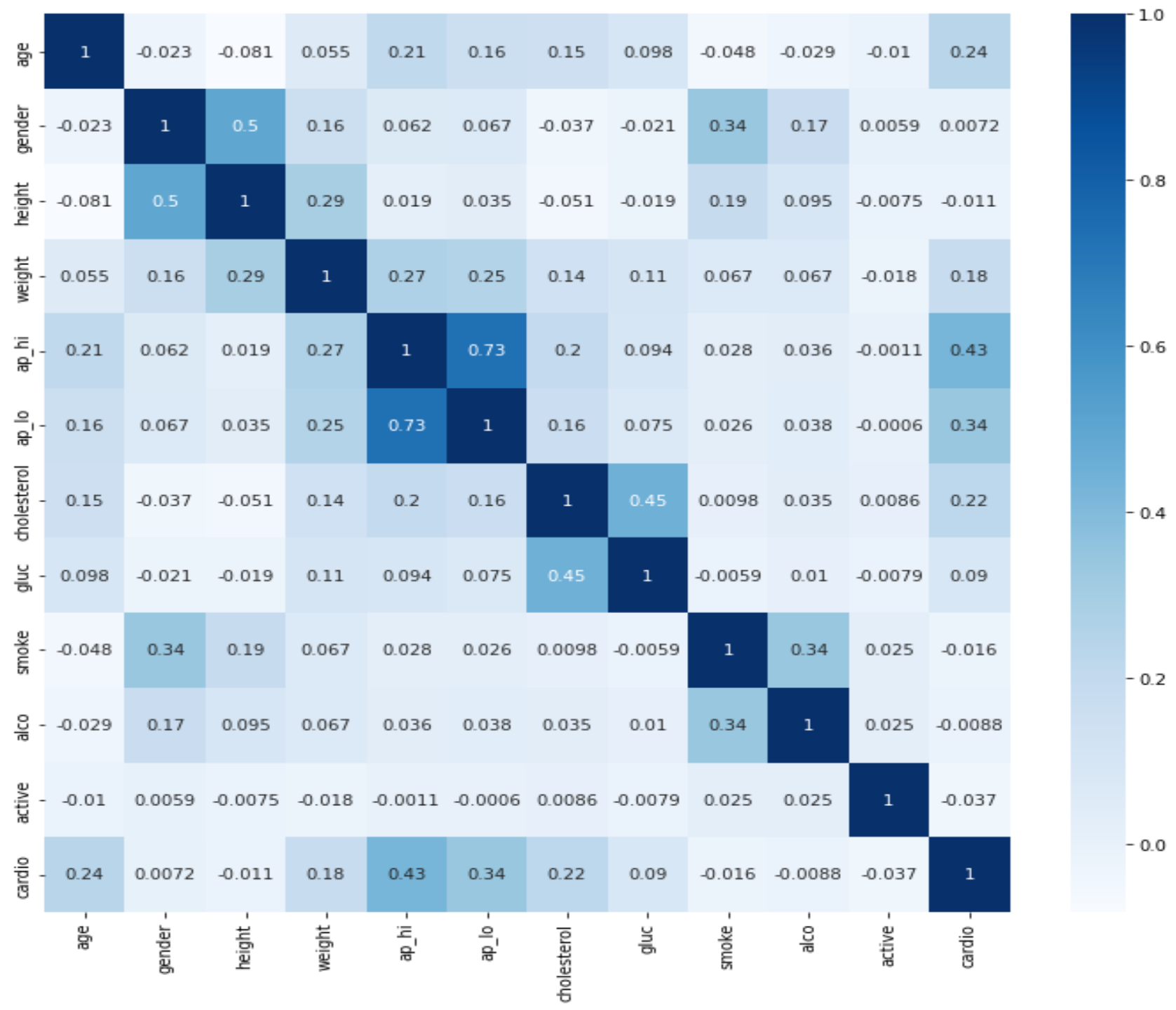






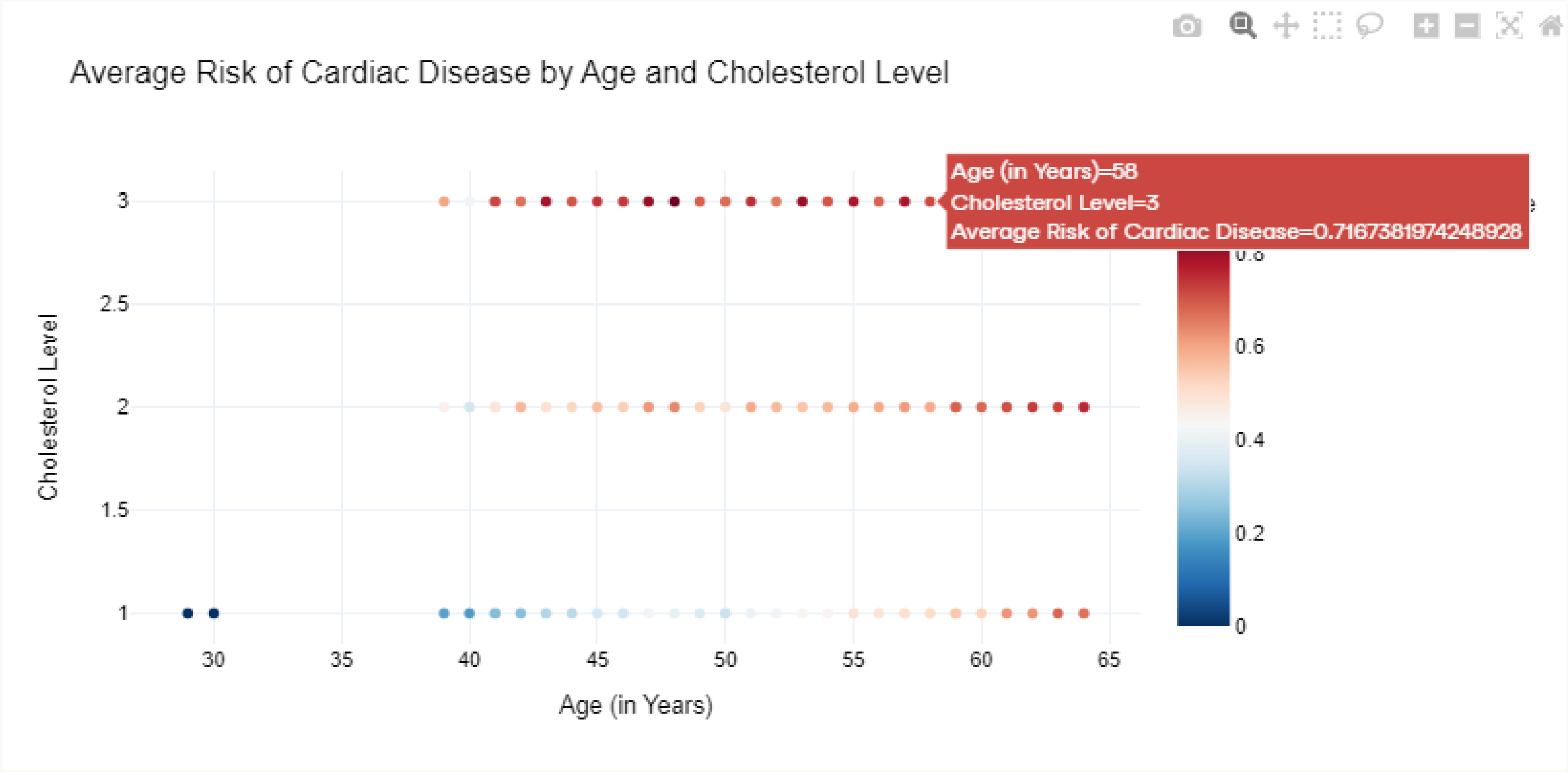


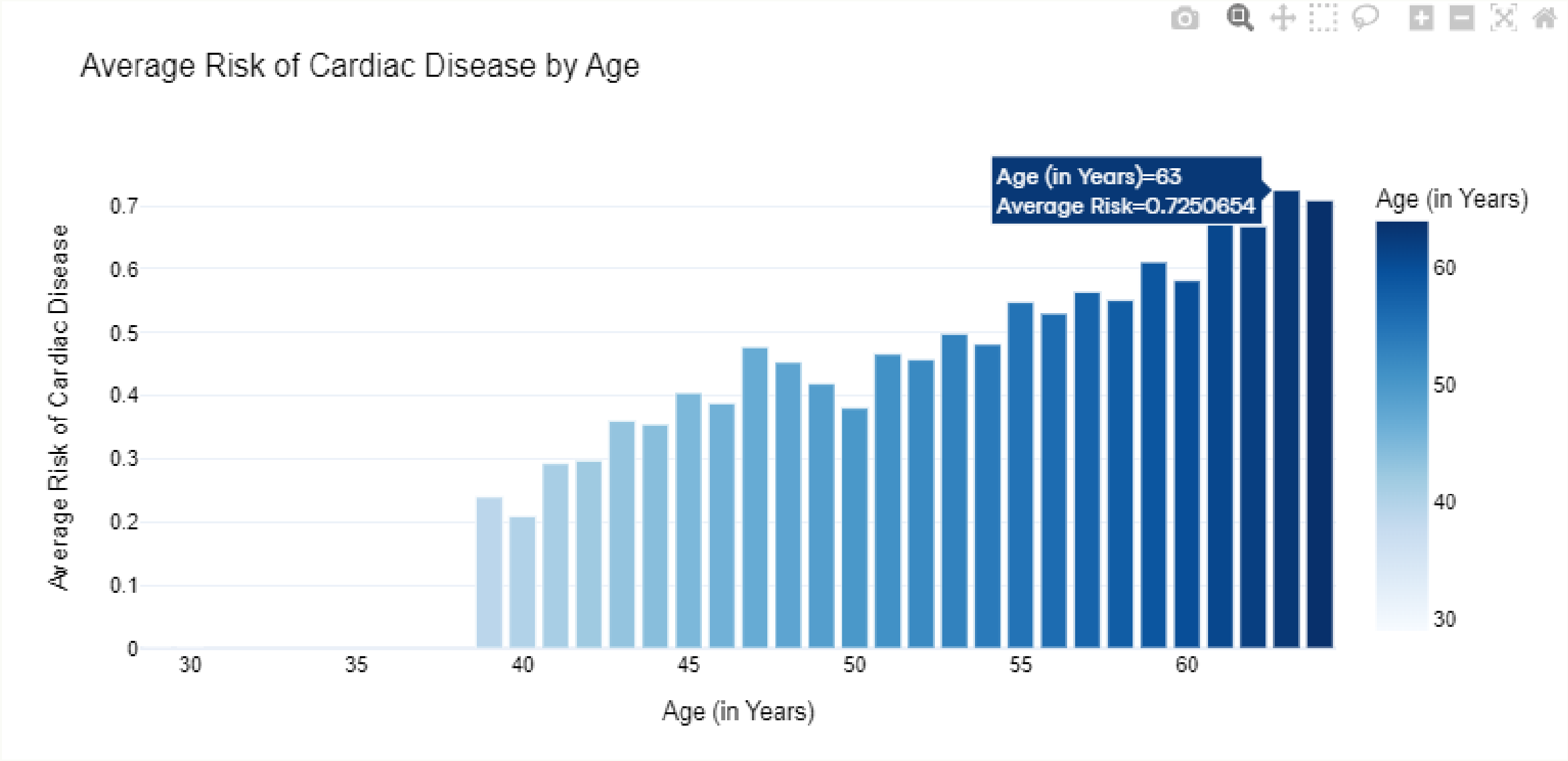


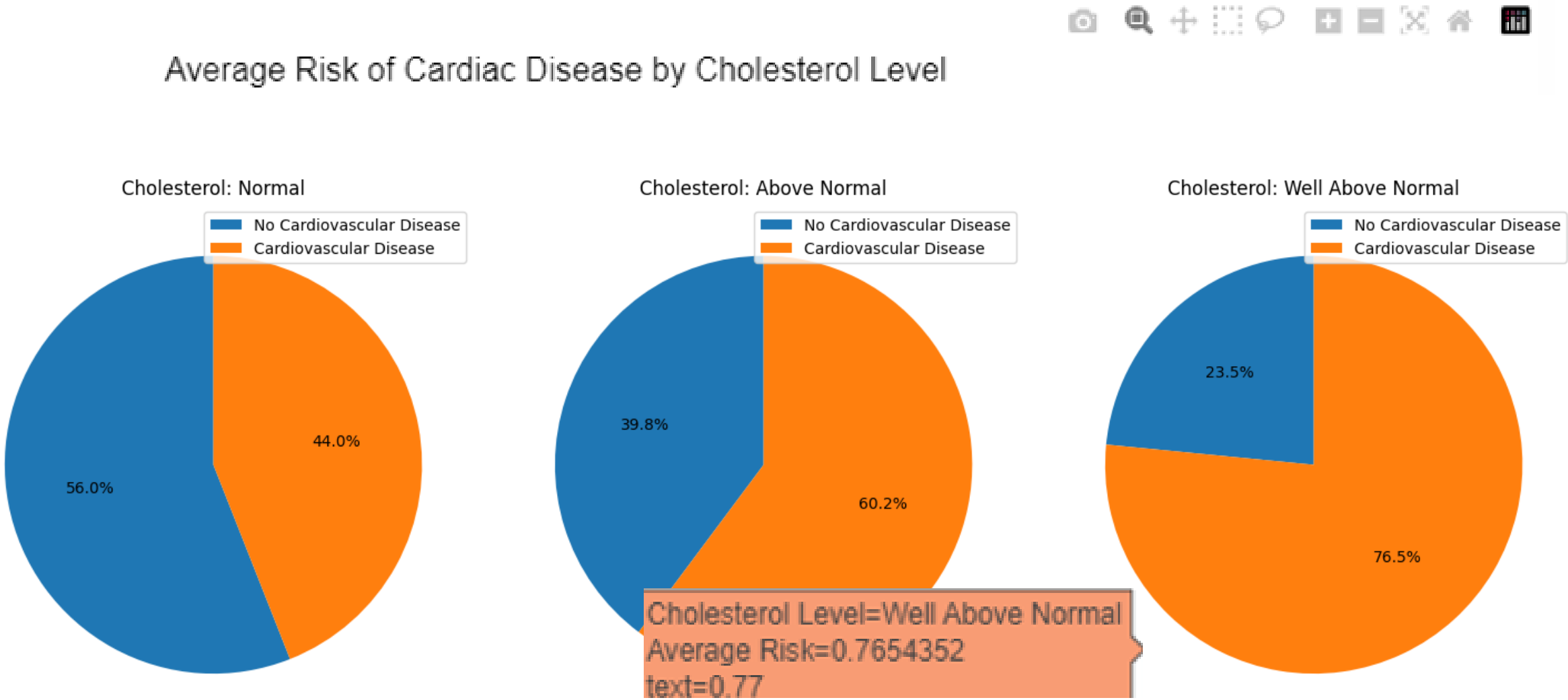


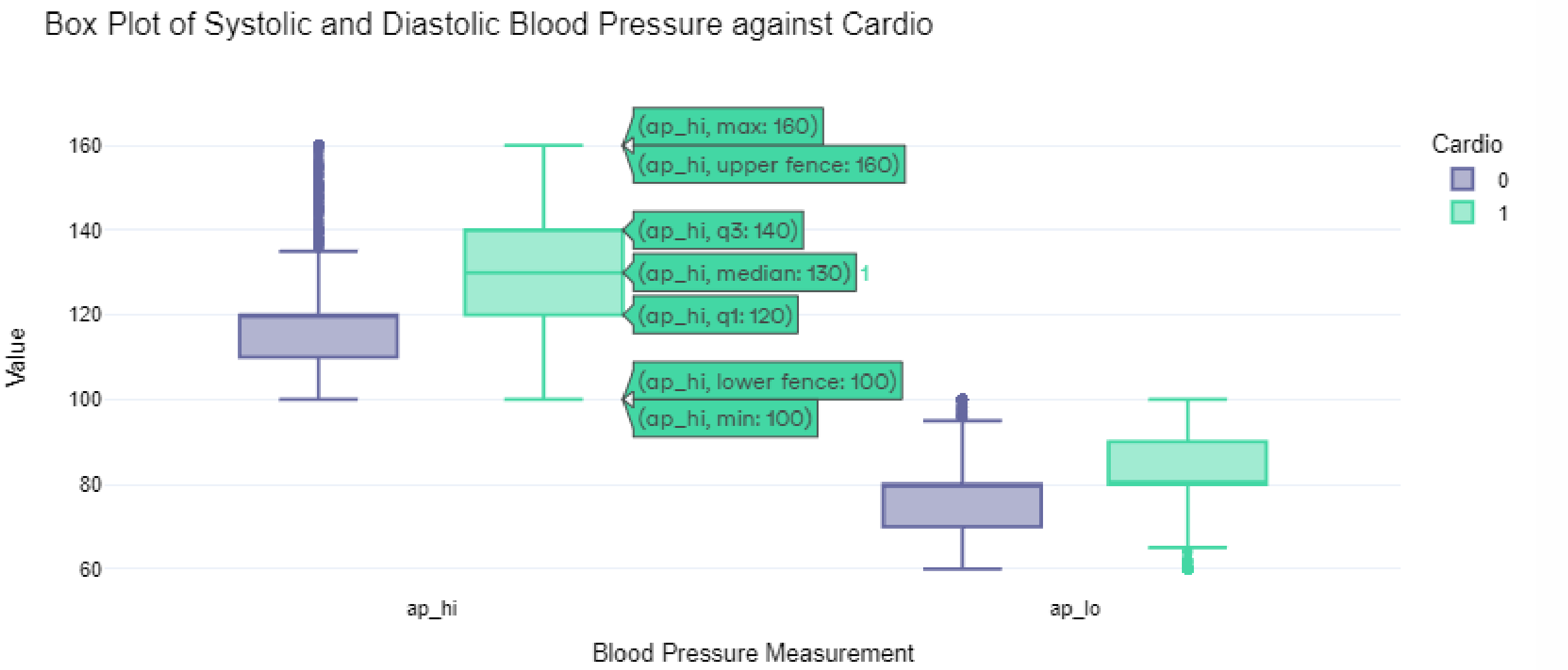
STATISTICAL ANALYSIS – CORRELATION ANALYSIS

- Age, Weight, Ap_hi, Ap_lo, Cholesterol seem to have strong positive correlation.


Average Risk of Cardiac Disease by Age and Cholesterol Level
Age (in Years)=58
Cholesterol Level=3
Average Risk of Cardiac Disease=0.7167381974248928
Cholesterol Level
3
2.5
2
1.5
1
30
35
40
45
50
55
60
65
Age (in Years)
0.6
0.4
0.2
0


Average Risk of Cardiac Disease by Age
Age (in Years)=63
Average Risk=0.7250654
Average Risk of Cardiac Disease
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
30
35
40
45
50
55
60
Age (in Years)
Age (in Years)
60
50
40
30


Average Risk of Cardiac Disease by Cholesterol Level
Cholesterol: Normal
No Cardiovascular Disease
Cardiovascular Disease
56.0%
44.0%
Cholesterol: Above Normal
No Cardiovascular Disease
Cardiovascular Disease
39.8%
60.2%
Cholesterol Level=Well Above Normal
Average Risk=0.7654352
text=0.77
Cholesterol: Well Above Normal
No Cardiovascular Disease
Cardiovascular Disease
23.5%
76.5%

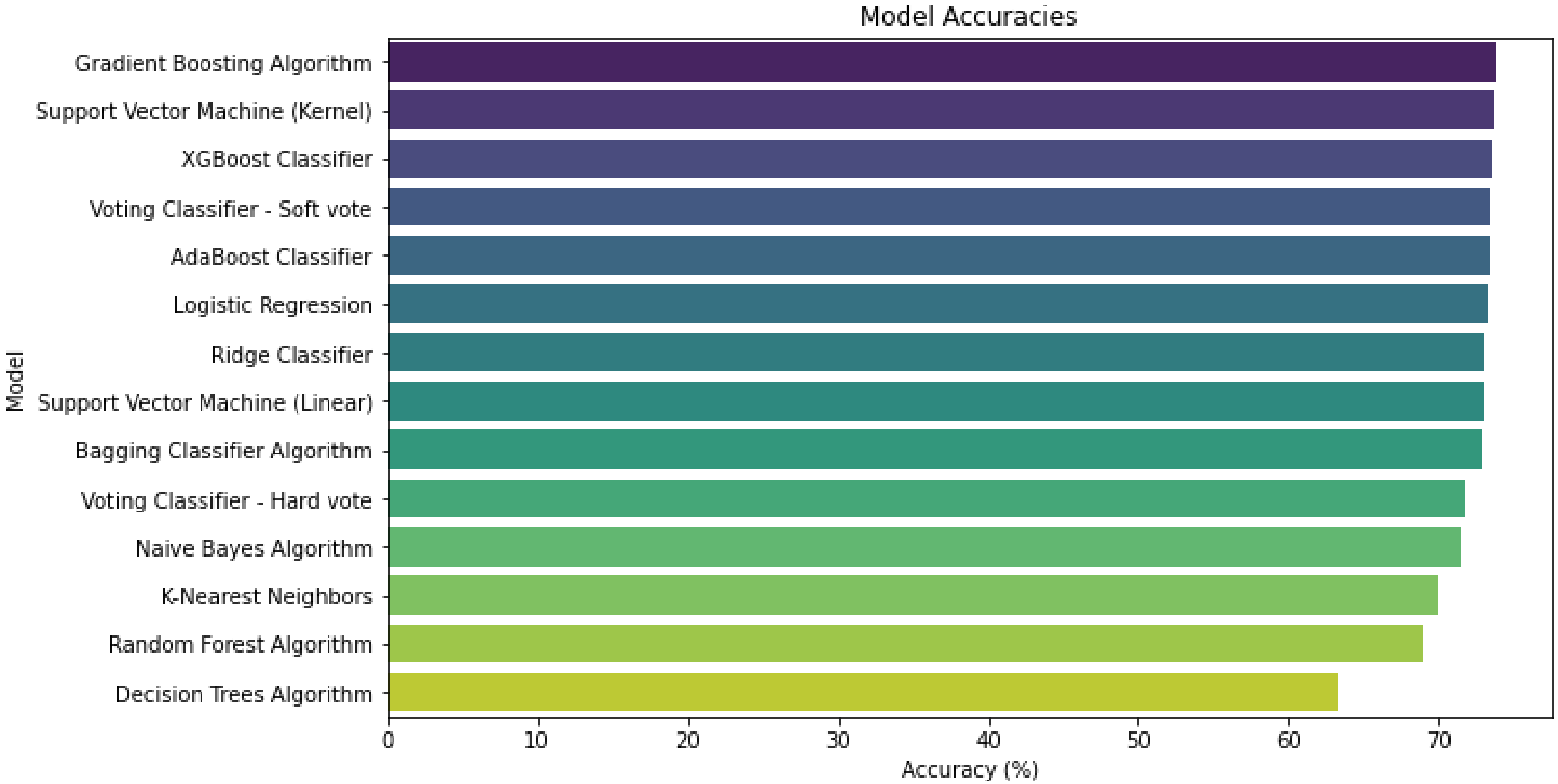

Box Plot of Systolic and Diastolic Blood Pressure against Cardio
(ap_hi, max: 160)
(ap_hi, upper fence: 160)
(ap_hi, q3: 140)
(ap_hi, median: 130)
(ap_hi, q1: 120)
(ap_hi, lower fence: 100)
(ap_hi, min: 100)
1
Cardio
0
1
160
140
120
100
80
60
Value
ap_hi
ap_lo
Blood Pressure Measurement

TABULATION OF 14 MODELS USED

Models	Accuracy
Logistic Regression	73.3%
K-Nearest Neighbors	69.8%
Support Vector Machine (Kernel)	73.7%
Decision Trees Algorithm	63.2%
Grading Boosting Algorithm	74%
Random Forest Algorithm	68.9%
Support Vector Machine (Linear) (Linear)	73%
Naïve Bayes Algorithm	71.4%

Models	Accuracy
Bagging Aggregating Classifier Algorithm	72.9%
Voting Classifier - Hard vote	71.7%
Voting Classifier - Soft vote	70.2%
ADA Boost Classifier	73.4%
XGB Classifier	73.5%
Ridge Classifier	73%

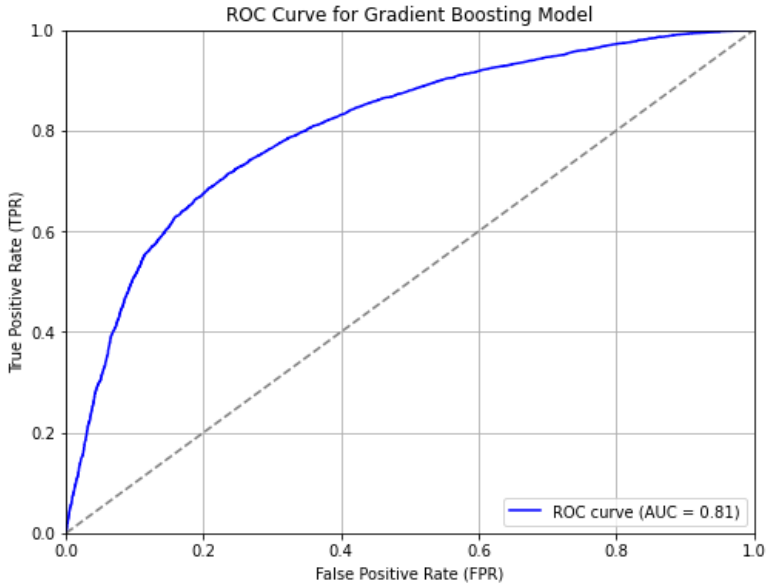
ACCURACY OF ALL MODELS

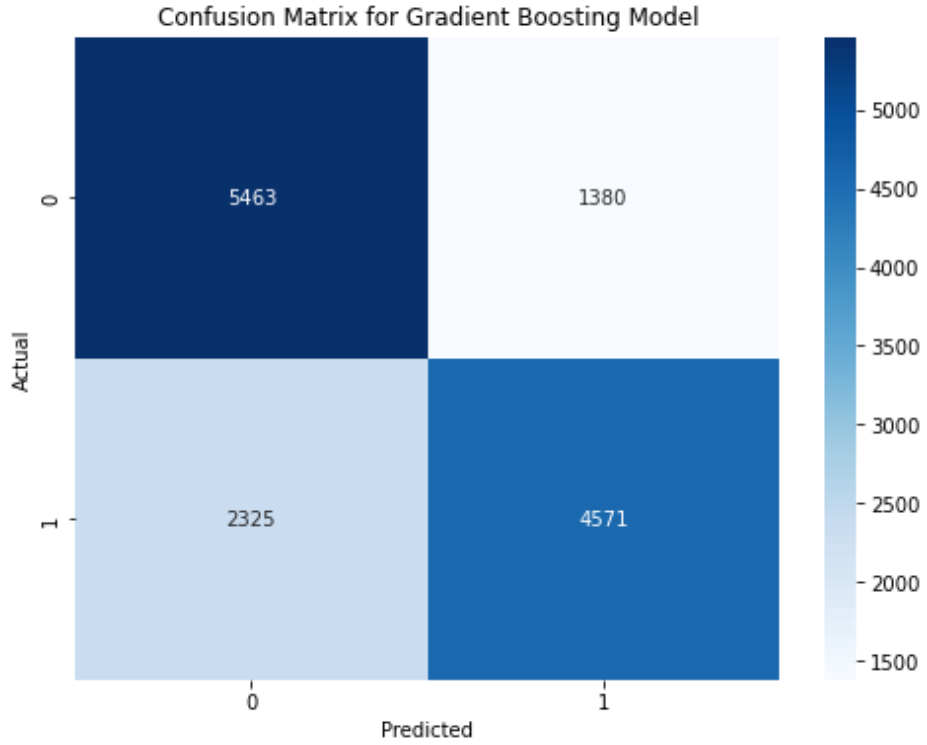


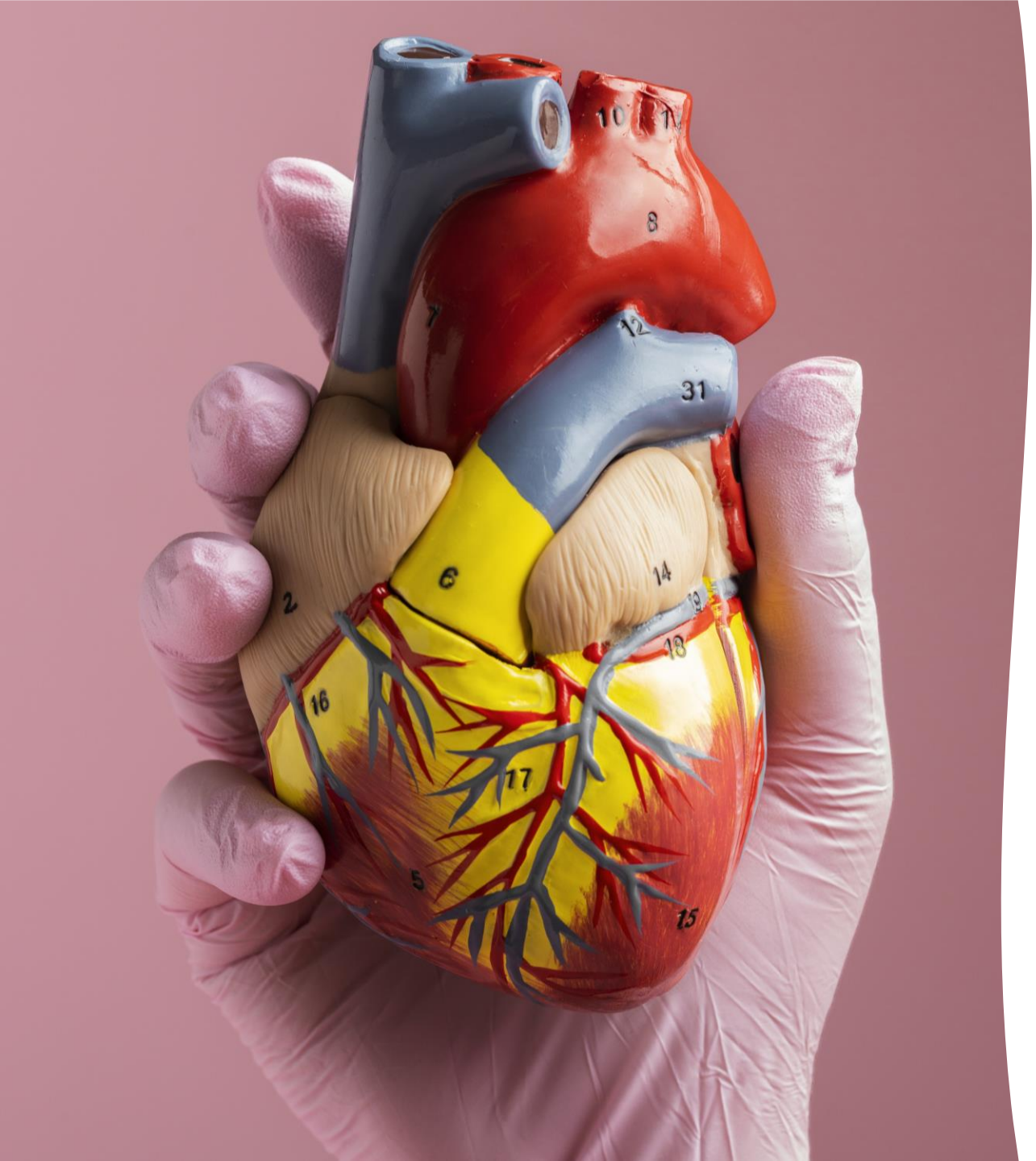
ROC and Confusion Matrix of Best Model

Feature importances sorted in descending order:

	Feature	Importance
4	ap_hi	0.746758
0	age	0.124769
6	cholestrol	0.077272
3	weight	0.019922
5	ap_lo	0.012798
10	active	0.005296
7	glu	0.004637
2	height	0.003528
8	smoke	0.002177
9	alco	0.001970
1	gender	0.000873







CONCLUSION

- To conclude, We can say that the Gradient Boosting Algorithm is the best model with the highest accuracy of 74% and AUC of 81% . We tried several models using different combination of variables but this model with all variables came out to be the best model.

- The main reason for the success of the research is because of the wonderful support and guidance provided by the professor. The lectures given by professors helped us a lot to push beyond visualizations and modeling.

TECHNOLOGIES USED

CODING PLATFORM

The whole project was implemented using Python on Spyder of Anaconda Navigator.

DATA CLEANING

Data Cleaning was done in Python with libraries such as pandas, numpy and math.

STATISTICAL ANALYSIS

The statistical analysis part was done in Python.

PREDICTIVE MODELLING

The predictive modelling part was done in Python.

PREDICTION TOOL

The prediction tool was developed in Python using the Dash App framework. Bootstrap was used for the front-end interface.

FUTURE WORK

- The research can be expanded by trying to improve the accuracy. This may be done with addition of features. With these added features and more information, the accuracy is bound to go up and can become a reliable source to predict Cardiovascular Diseases.

- There can be more records added to the existing dataset. The more records to work with, the better model we can develop in terms of accuracy and efficiency.

- We built the prediction tool on DASH App Framework. The prediction tool can be made more better by building it on angular or similar technologies and making it more efficient by catching the exceptions or errors, if we come across any.

THANKS!
Questions?