Cardiovascular Disease Prediction

Table of contents

Introduction
About Cardiovascualr

01

02

Data Analysis
Exploratory data analysis about the data we have

Modeling
Reviewing the results of the models

03

04

Conclusions
Final conclusions in the insights and results we have

1.Introduction

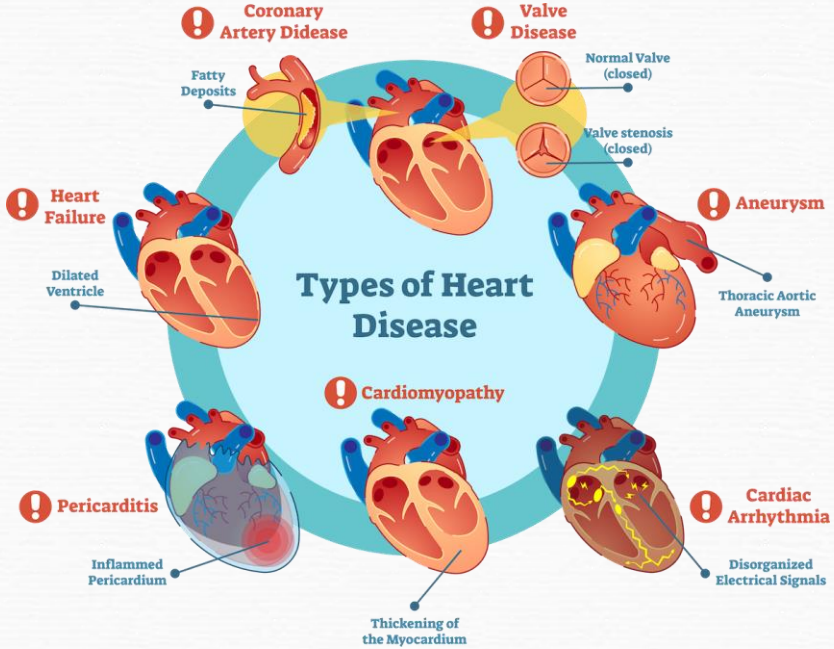
What is cardiovascular, and what are the diseases are related to it?

Cardiovascular:

- Cardiovascular disease (CVD) is a general term for conditions affecting the heart or blood vessels.

Cardiovascular Diseases:

- heart attacks
- Heart failure
- angina
- Strokes

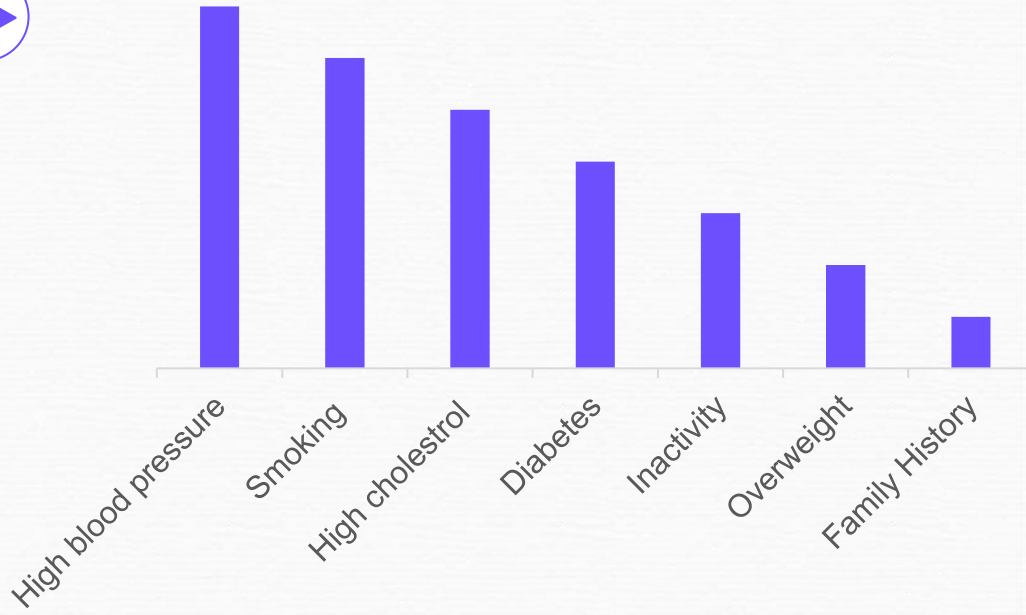


Reference: https://www.nhs.uk/conditions/cardiovascular-disease/

Most Common Causes for Cardiovascular diseases

Causes of cardiovascular diseases:

- Smoking
- Blood Pressure
- Family health history
- Inactivity
- Diabetes
- High cholesterol
- Being overweight



Here are the most common reasons are ranked based on NHS

Reference: https://www.nhs.uk/conditions/cardiovascular-disease/

Ways of Prevention from Cardiovascular diseases

Prevention:

- Stop smoking
- Health diet
- Exercise regularly
- Maintain healthy weight
- Medication

And so on based of each patient with his or her specific life style and what might be applicable treatment for them

There is a different way that might help preventing and help curing the disease before it might it be happening and that solution is :

Machine Learning Model

Machine learning will help using historical data of heart diseases to detect it's pattern and help us identify the diseases before it's actually happen, and start the treatment process early

Reference: https://www.nhs.uk/conditions/cardiovascular-disease/

2.Data Analysis

Before diving in the machine learning process, we will do EDA

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	BMI
1	50.3	0	168	62.0	110	80	1	1	0	0	1	0	21.9
2	55.4	1	156	85.0	140	90	3	1	0	0	1	1	34.9
3	51.6	1	165	64.0	130	70	3	1	0	0	0	1	23.5
4	48.2	0	169	82.0	150	100	1	1	0	0	1	1	28.7
5	47.8	1	156	56.0	100	60	1	1	0	0	0	0	23.0

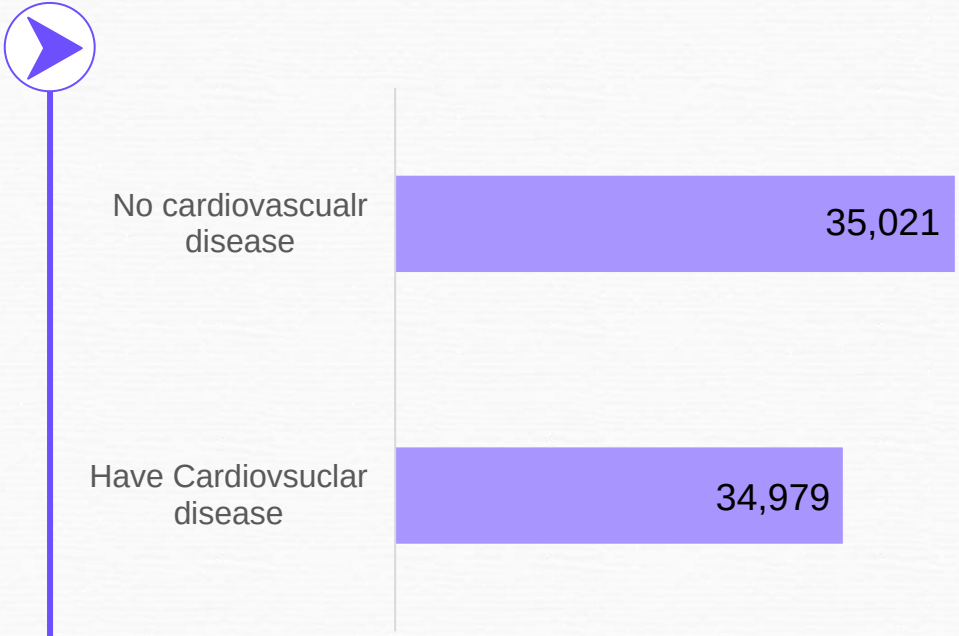
Features

Target Variable

Patients without cardiovascular was higher than patients with it

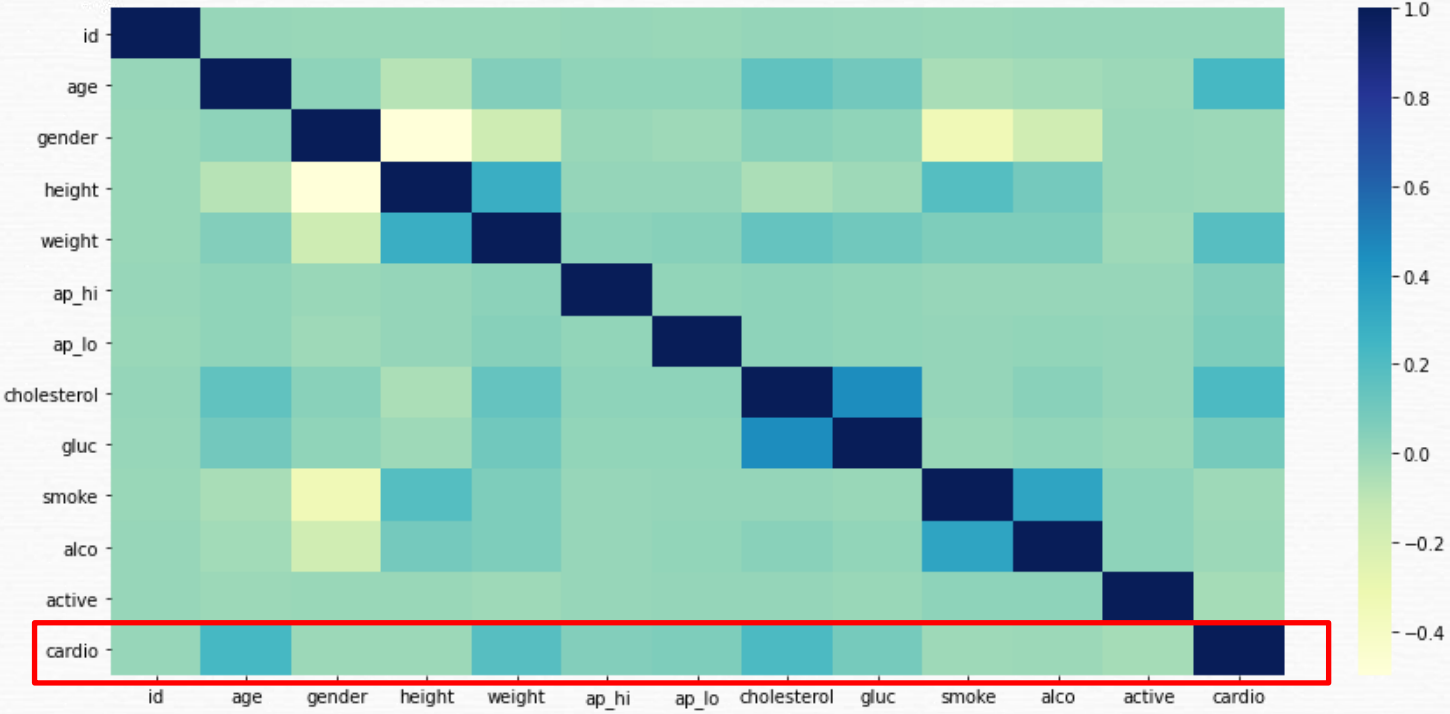
Cardiovascular disease (Our Target):

- Part of the analysis we do is to check number of patients and how many of them are facing diseases, so we can identify if the model and the data are balanced or not
- Also gain better knowledge about the patient data we have

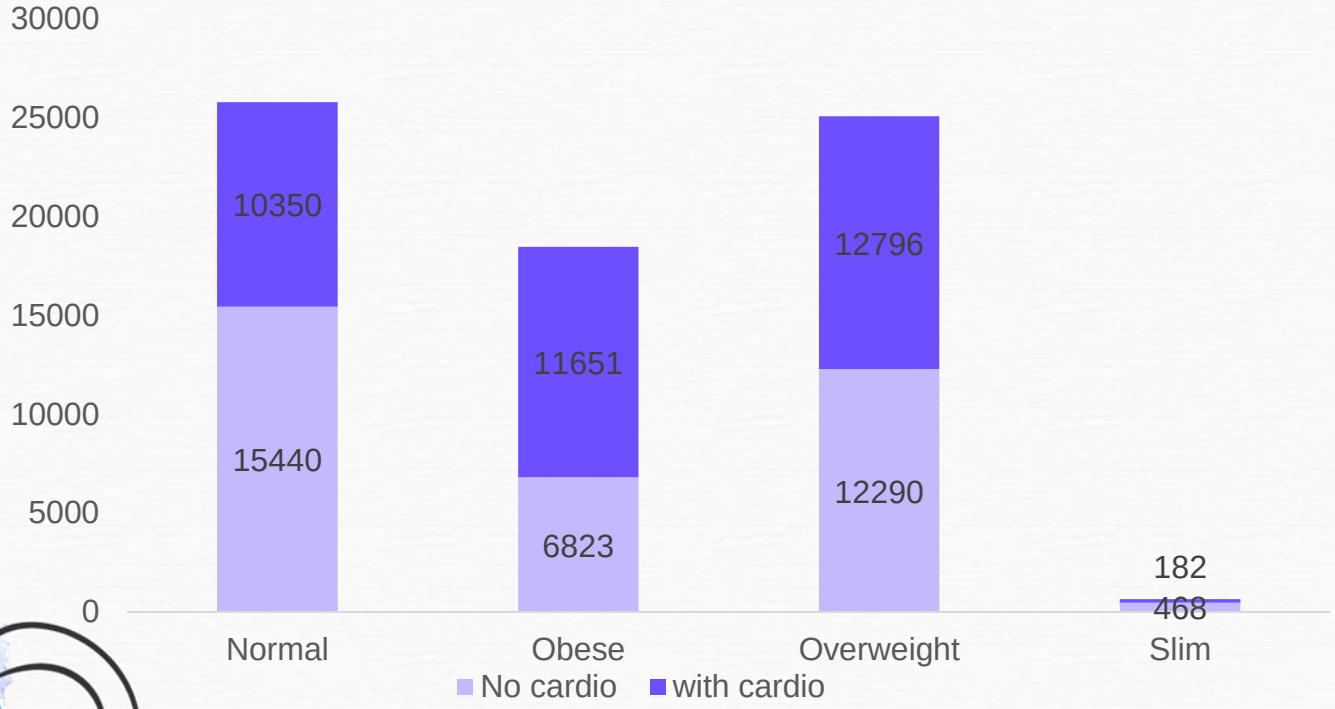


Obviously, we have a balanced dataset

4 Data elements that are positively correlated with cardiovascular



Below shows how patient who are obese and overweight have high cardiovascular disease



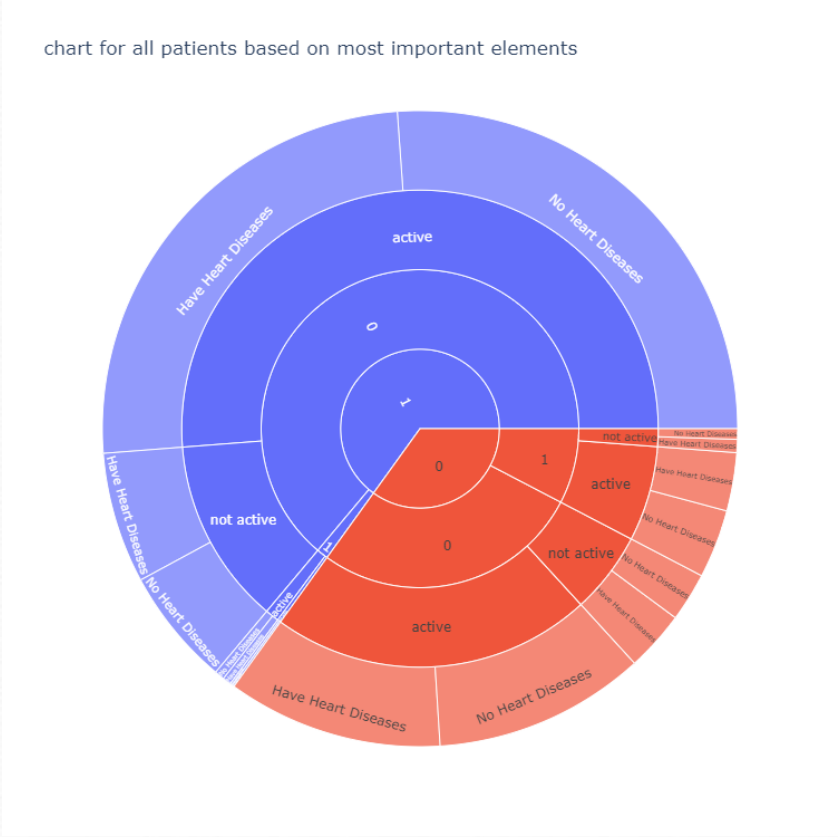
Categorized pie chart

In the right section, we see a pie chart that is categorized for each male and female

And it shows the characteristics of male and female patients and the relation of the data elements with cardiovascular

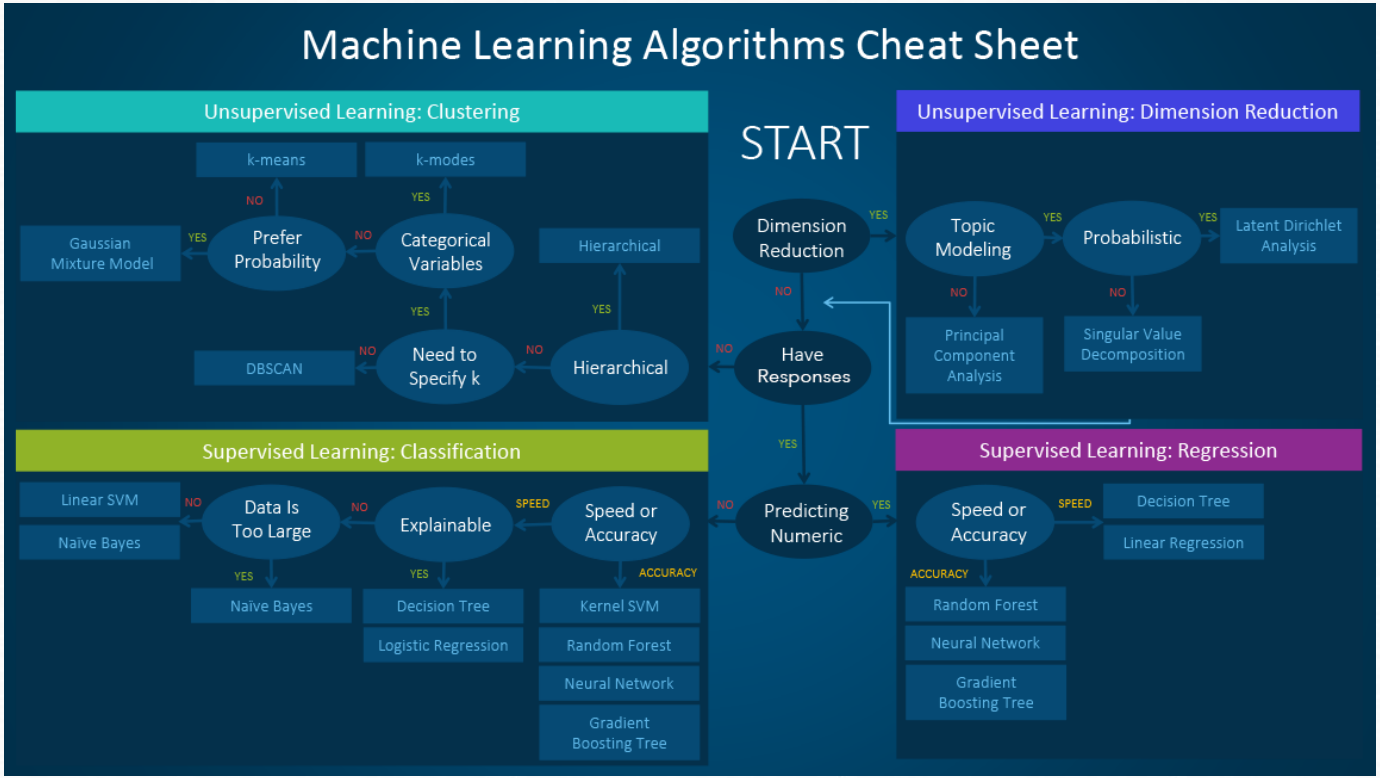
For example:
Female patients, who are not smoking and most of the time are active, most of them are not having heart diseases

We can gain a lot of information from this chart



3.Modeling (ML)

Model Selection



Model Selection

- A multiple algorithms have been built as a based model to be compared and what model can be tuned, the following table shows the models

Model Name	Accuracy %
Logistic Regression	72%
KNN	66%
Random Forrest	71%
SVM	68%
Decision Tree	63%
Gradient Boosting	73%

- The table shows that the logistic regression and Gradient boosting models, performed the best as a based models, so we will be choosing one of them and tune them

Focusing on the best accuracy

- Since we narrowed our models into two models, and they are logistic regression and gradient boosting

Model Name	Accuracy training	Accuracy validation
Logistic Regression	72%	72%
Gradient Boosting	73%	76%

When we compared both models, and checked the metrics we saw that gradient boosting algorithm performs better then logistic regression even with hyperparameter

Gradient Boosting Model

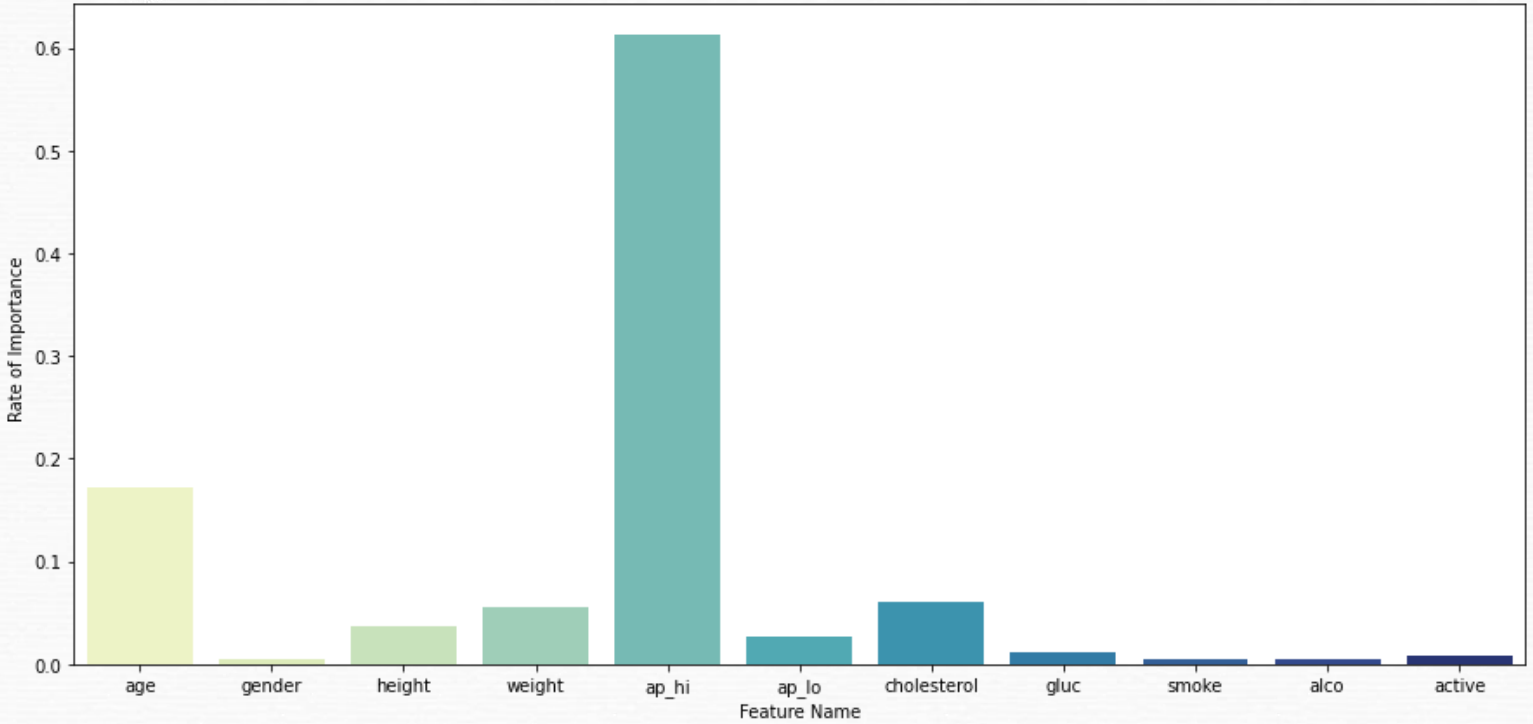
- We found our best fitted model for our case which is to identify patients with cardiovascular based on patient information

Model Name	Accuracy training	Accuracy validation	Accuracy Testing
Gradient Boosting	73%	76%	74%

Lastly we predicted the results using our testing data and from the accuracy in general we can say that the model is very well generalized and learned good and not overfitting

Feature Importance

From below we see that Systolic blood pressure (ap_hi) is the most important feature out of the total elements we have



4.Conclusions

- Our best model performed better than the benchmark we had which is 72%
- Model is very well generalized
- Model accuracy can be increased if we added more data to the model
- Our model can help doctors in diagnoses and check the patient status
- The model can save time and resource consumption sometimes
- Systolic blood pressure is one of the important factors when coming to cardiovascular disease

“In the next 10 years, data science and software will do more for medicine than all of the biological sciences together”
Vinod Khosla

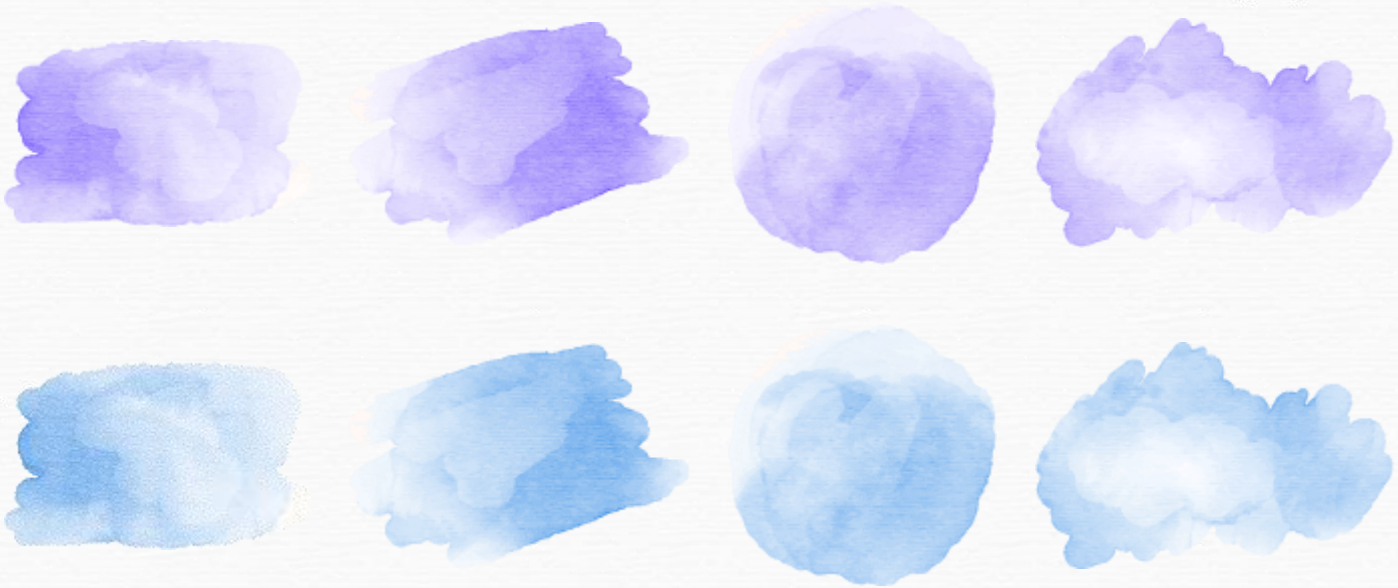
Thank You

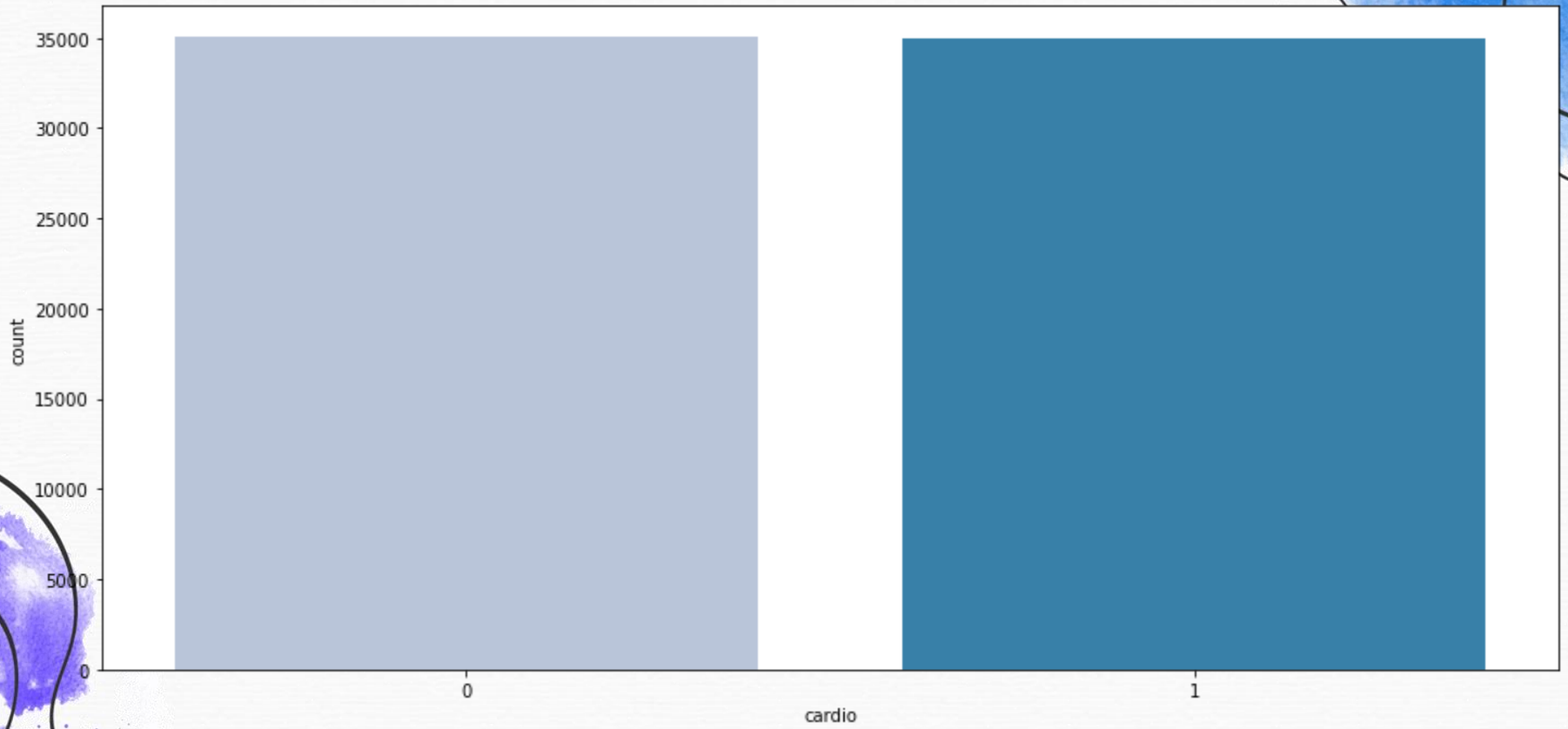
Research resources

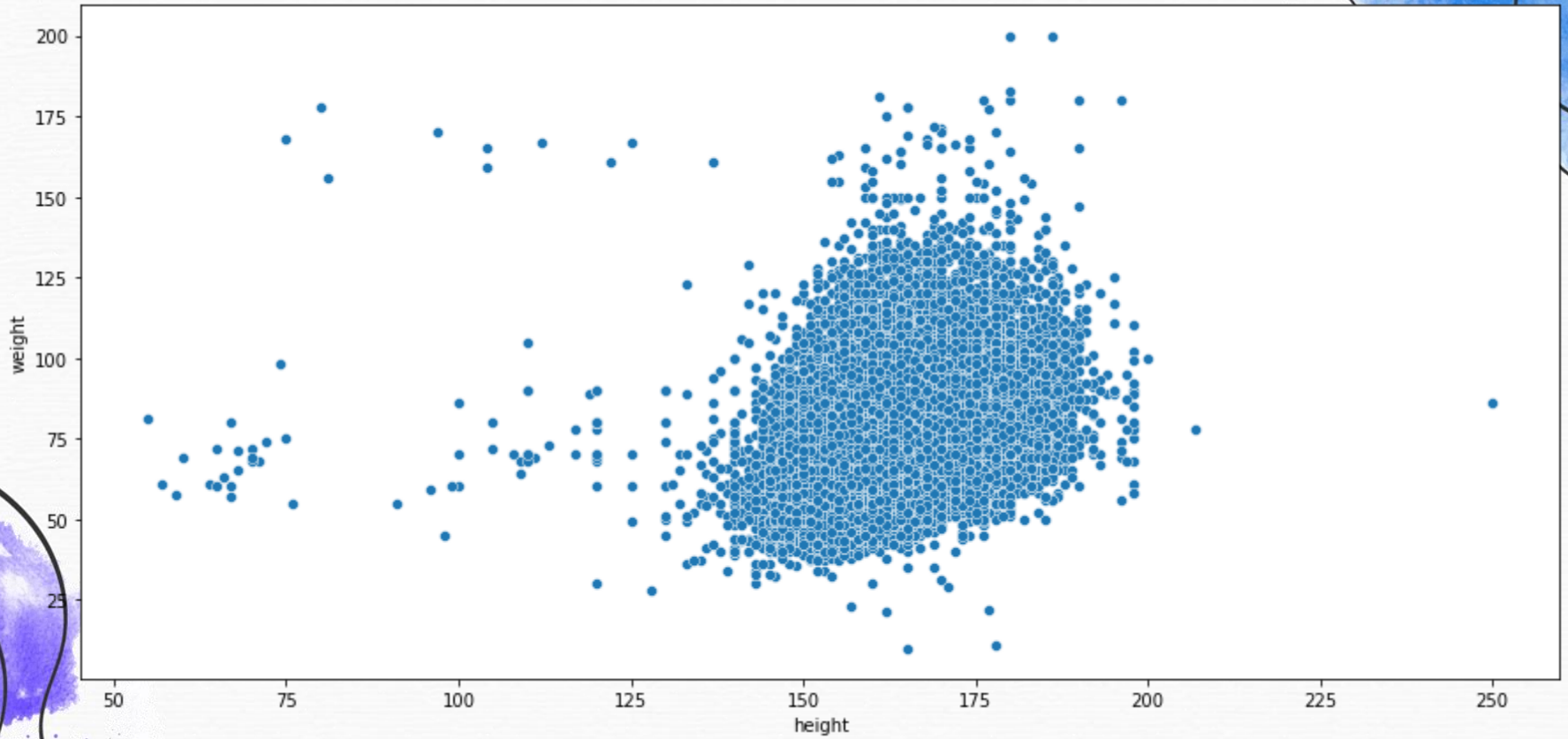
- https://quotefancy.com/quote/1272586/Vinod-Khosla-In-the-next-10-years-data-science-and-software-will-do-more-for-medicine
- https://www.kaggle.com/sulianova/cardiovascular-disease-dataset
- https://www.nhs.uk/conditions/cardiovascular-disease/
- https://plotly.com/python/sunburst-charts/
- https://machinelearningmastery.com/calculate-feature-importance-with-python/

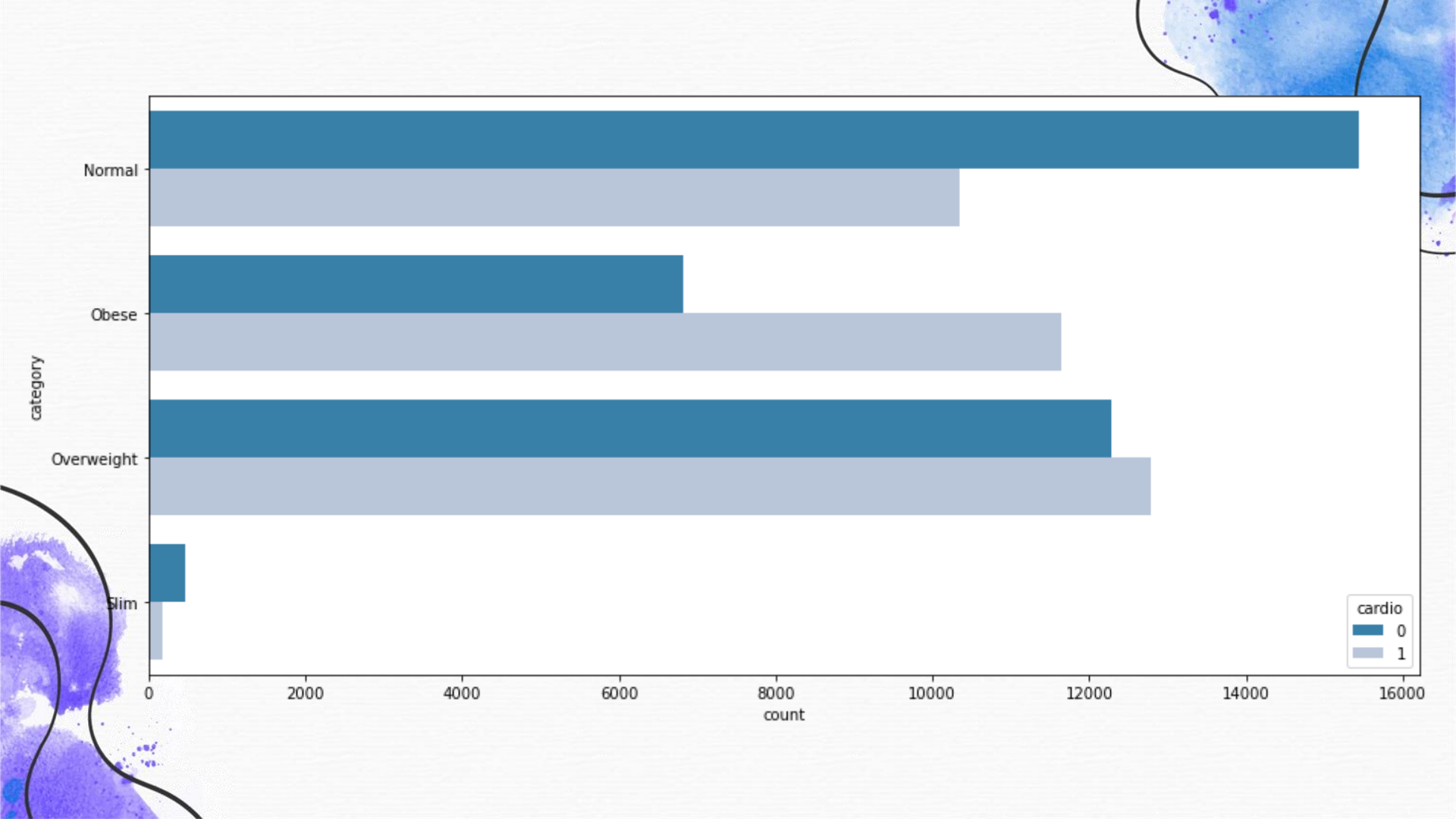
Appendix

Alternative resources




35000
30000
25000
20000
15000
10000
5000
0
count
0
1
cardio


weight
height
200
175
150
125
100
75
50
25
50
75
100
125
150
175
200
225
250


Normal
Obese
Overweight
Slim
category
0
2000
4000
6000
8000
10000
12000
14000
16000
count
cardio
0
1

Colors Used

#000000

#ffffff

#6e4ffd

#348ce7