Cardio Disease Prediction

Presented by
Diandra MELO

Presentation Outline

Today's Topics

1

CONTEXT

Cardiovascular Disease dataset

Predictive Analysis of the presence or absence of cardiovascular disease using a dataset of the patients features (objective, subjective and examination).

Objective:
create a tool that generates a significant and stable diagnostic accuracy.

DATASET DESCRIPTION

Provided Data

OBJECTIVE FEATURES

Age, Height, Weight, Gender

SUBJECTIVE FEATURES

Smoking, Alcohol intake, Physical activity

EXAMINATION FEATURES

Systolic & Diastolic blood pressures, Cholesterol, Glucose

Added Features

BLOOD PRESSURE LEVEL

'normal', 'pre_hypertension', 'hypertension'

BODY MASS INDEX (BMI)

$BMI = weight / (height)^2$

WEIGHT STATUS

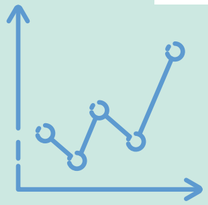
'underweight', 'normal', 'overweight', 'obese'

2

DATASET DESCRIPTION

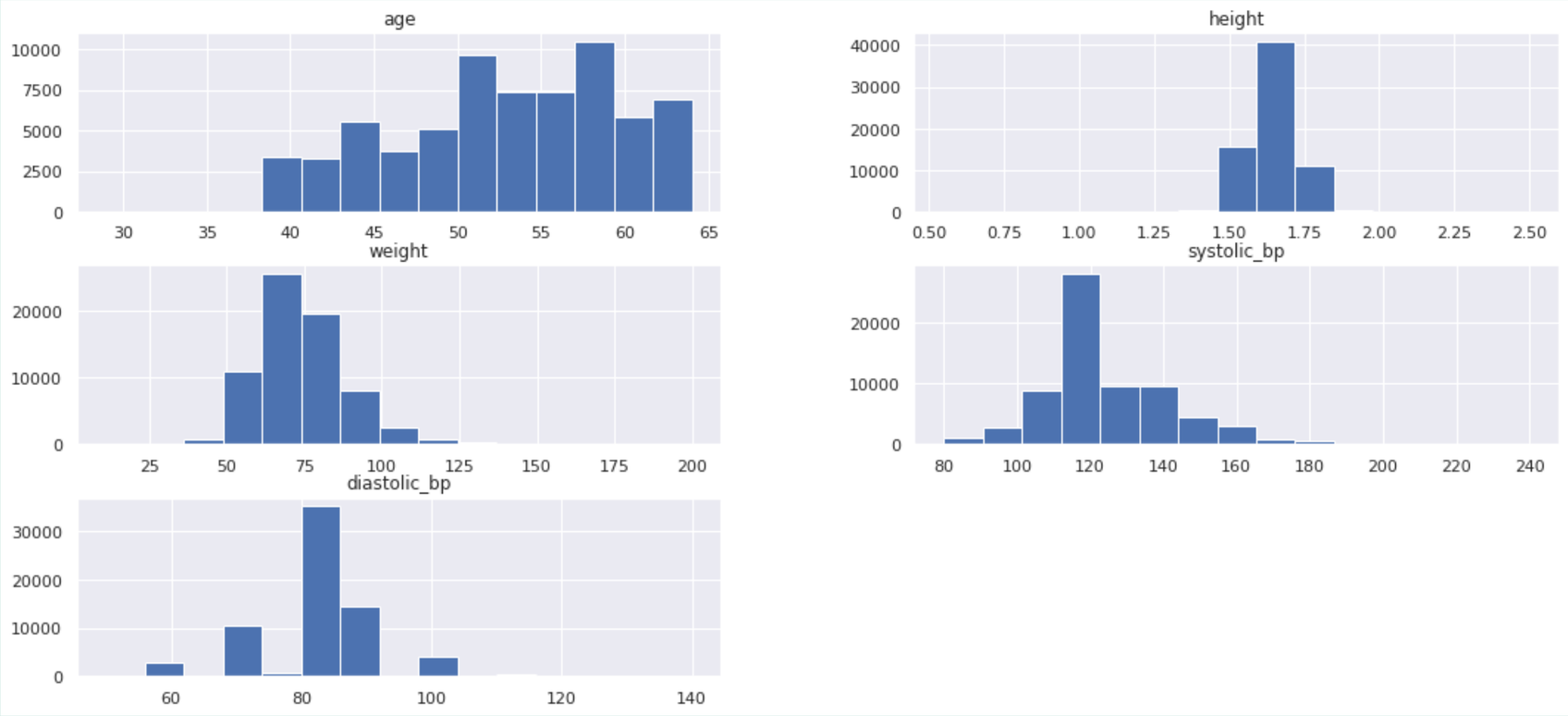
Descriptive Statistics

	mean	median	std	min	max	range	skew	kurtosis
id	49972.419900	50001.5	28851.302323	0.0	99999.0	99999.0	-0.001278	-1.198374
age	19468.865814	19703.0	2467.251667	10798.0	23713.0	12915.0	-0.307055	-0.823447
height	164.359229	165.0	8.210126	55.0	250.0	195.0	-0.642187	7.943653
weight	74.205690	72.0	14.395757	10.0	200.0	190.0	1.012070	2.586825
systolic_bp	128.817286	120.0	154.011419	-150.0	16020.0	16170.0	85.296214	7580.074738
diastolic_bp	96.630414	80.0	188.472530	-70.0	11000.0	11070.0	32.114083	1425.914585



3 EXPLORATORY DATA ANALYSIS

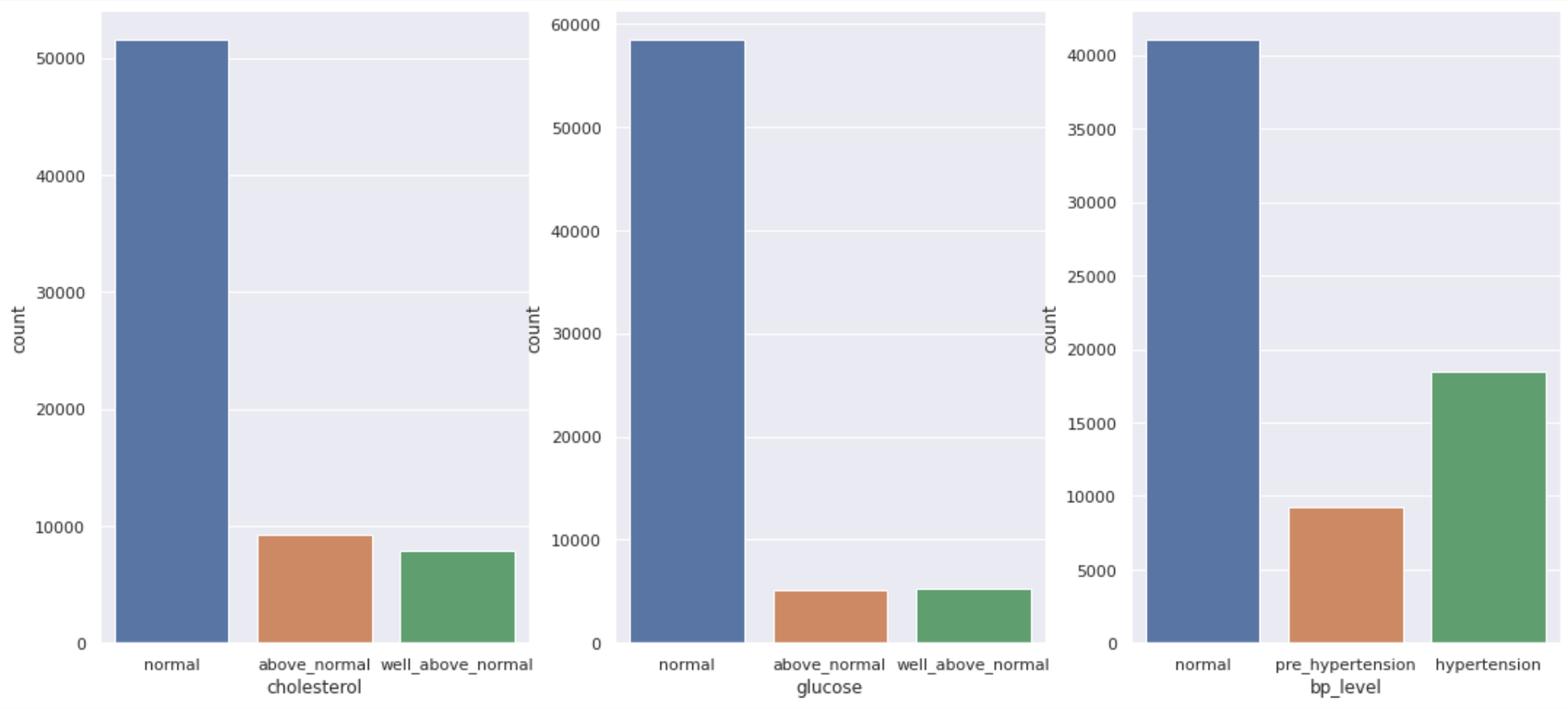
3.1. Univariate Analysis



Numerical Attributes

3 EXPLORATORY DATA ANALYSIS

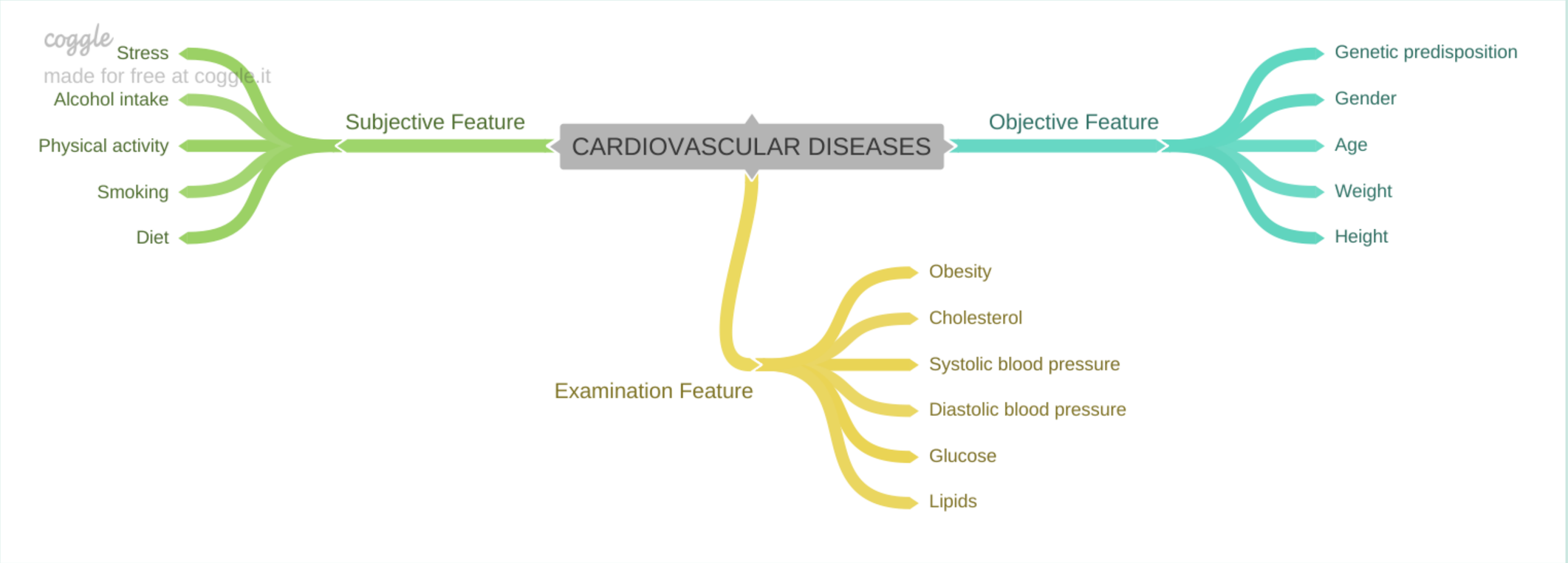
3.1. Univariate Analysis



Categorical Attributes

3 EXPLORATORY DATA ANALYSIS

3.2. Bivariate Analysis



Hypothesis Mind Map

3

EXPLORATORY DATA ANALYSIS

3.2. Bivariate Analysis

Hypothesis Development

H1. Men are **most likely** to present cardiovascular diseases

H2. People with **obesity** are **most likely** to present cardiovascular diseases

H3. Older people are **most likely** to present cardiovascular diseases

H4. People who **drink alcohol** are **most likely** to present cardiovascular diseases

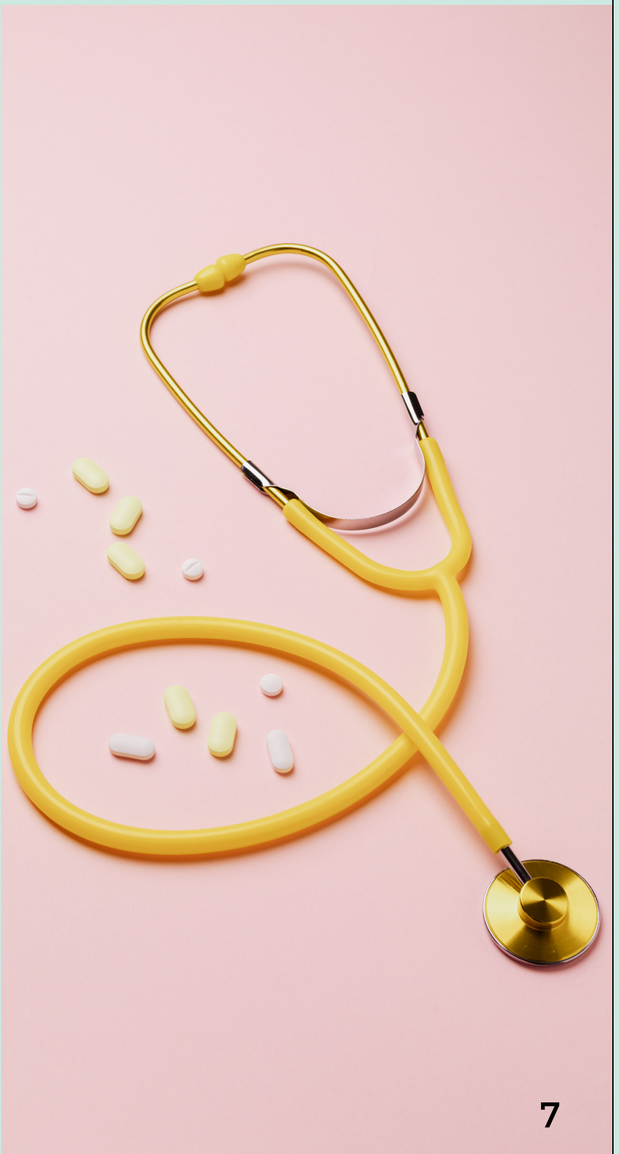
H5. People who **practice regular physical activities** are **less likely** to present cardiovascular diseases

H6. People who **smoke** are **most likely** to present cardiovascular diseases

H7. People who **present high cholesterol levels** are **most likely** to present cardiovascular diseases

H8. High glucose levels people are **most likely** to present cardiovascular diseases

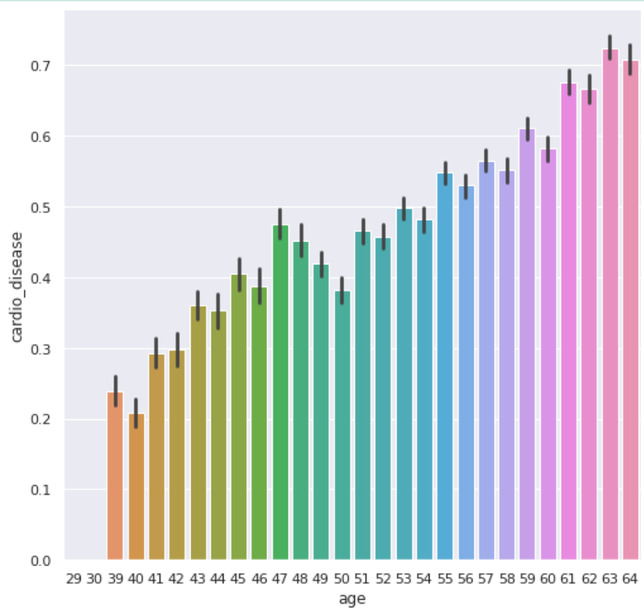
H9. People with **hypertension** are **most likely** to present cardiovascular diseases

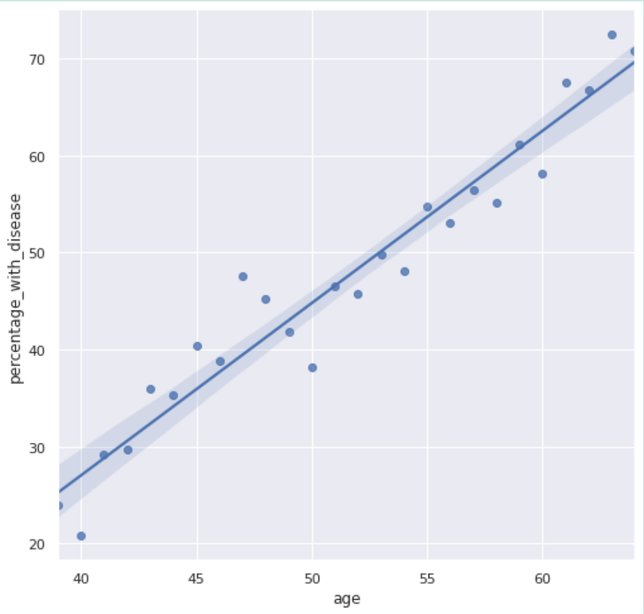


EXPLORATORY DATA ANALYSIS

Hypothesis Validation: H3

"Older people are most likely to present cardiovascular diseases".





This hypothesis turned to be TRUE.

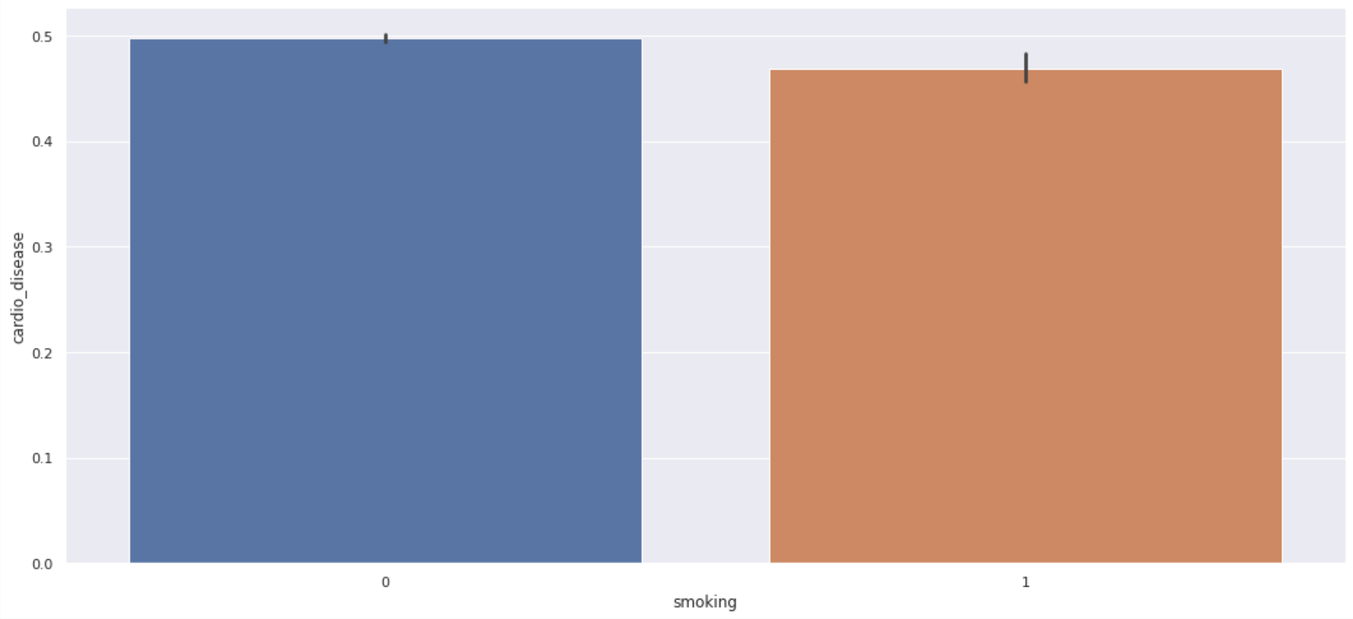
The linear regression slope exhibits well this tendency.

3 EXPLORATORY DATA ANALYSIS

3.2. Bivariate Analysis

Hypothesis Validation: H6

"People who smoke are most likely to present cardiovascular diseases".



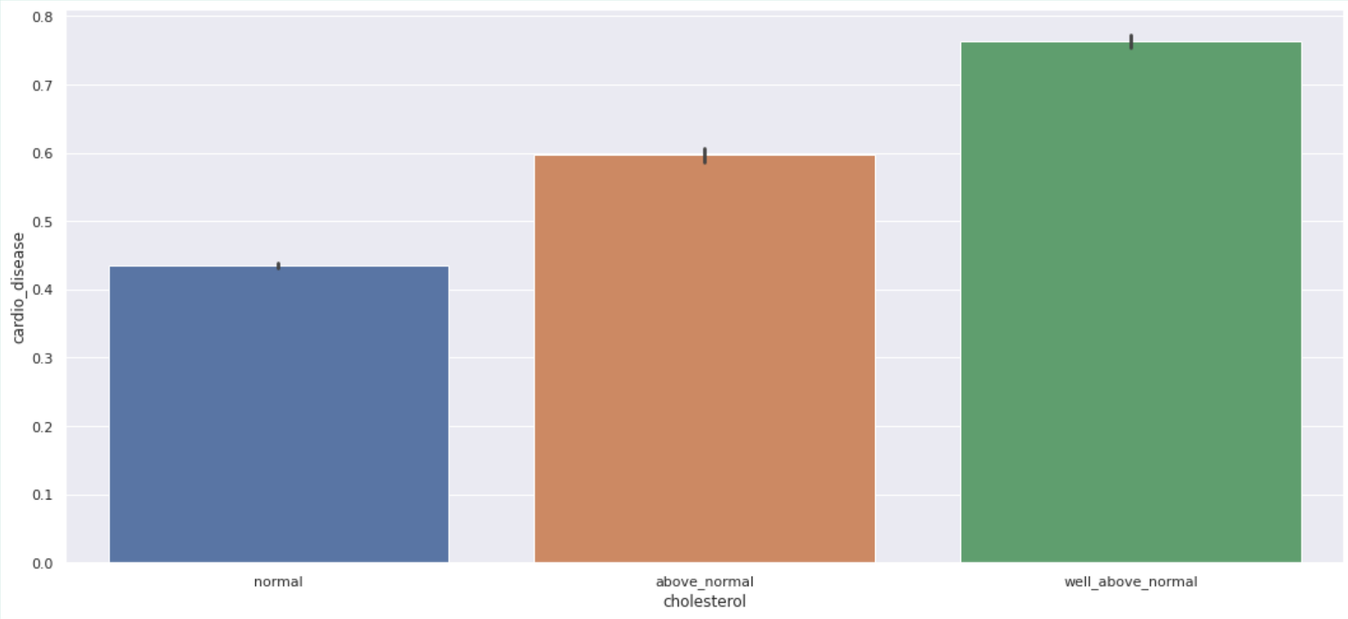
This hypothesis turned to be FALSE.

ALthough this feature does not show a strong correlation with the Target Variable.

EXPLORATORY DATA ANALYSIS

Hypothesis Validation: H7

"People who present high cholesterol levels are most likely to present cardiovascular diseases".



This hypothesis turned to be TRUE.

Plus, this feature shows a strong correlation with the Target Variable.

4

MACHINE LEARNING MODELLING

01

LOGISTIC REGRESSION

Single Performance Accuracy:

72.34 %

Cross-Validation Accuracy:

72.05 %

02

RANDOM FOREST CLASSIFIER

Single Performance Accuracy:

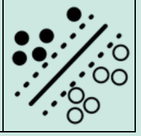
72.22 %

Cross-Validation Accuracy:

72.05 %

03

SVM CLASSIFIER



Single Performance Accuracy:

72.52 %

Cross-Validation Accuracy:

72.53 %

5 CONCLUSION

01.
DIAGNOSIS

The models used were capable of predict well a cardiovascular disease in patients, even with Cross-Validation analysis.

02.
ACCURACY

Not much improvement with the use of fine tuning techniques (less than 0.2 %).

03.
FURTHER STEPS

More data would have to be required to test accuracy's improvement.

THANK YOU!

Diandra Melo

GITHUB

https://github.com/diandramelo

Dataset "**Cardiovascular Diseases**" available at Kaggle