Advanced Coursera Capstone Project

Analysis and Prediction of Heart Disease

Bertan Gunes

INTRODUCTION

The aim of this project is to explore the Heart Disease UCI data set and, the construction of machine learning and deep learning model that predicts whether or not the patient has heart disease through the variables in this data set.

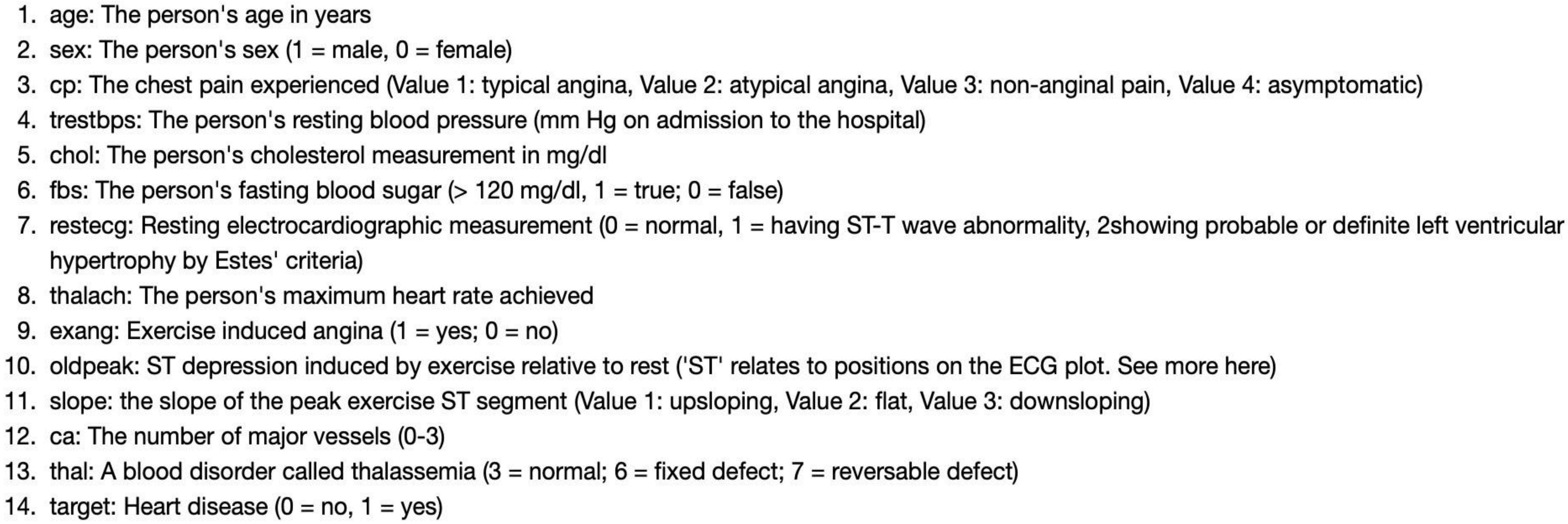
BUSINESS PROBLEM

- Heart disease describes a range of conditions that affect your heart.
- Heart disease is usually based on signs, symptoms and physical examination of the patient.
- There are several factors that increase the risk of heart disease such as age, sex, heredity, high cholesterol, hypertension, physical inactivity, etc

DATA

- The Heart Disease UCI data set from Kaggle
- Data set contains 76 attributes, but all published experiments refer to using a subset of 14 of them.
- Data set has information about 303 people. There are 13 features which are made available and there is one target variable.

FEATURE SET

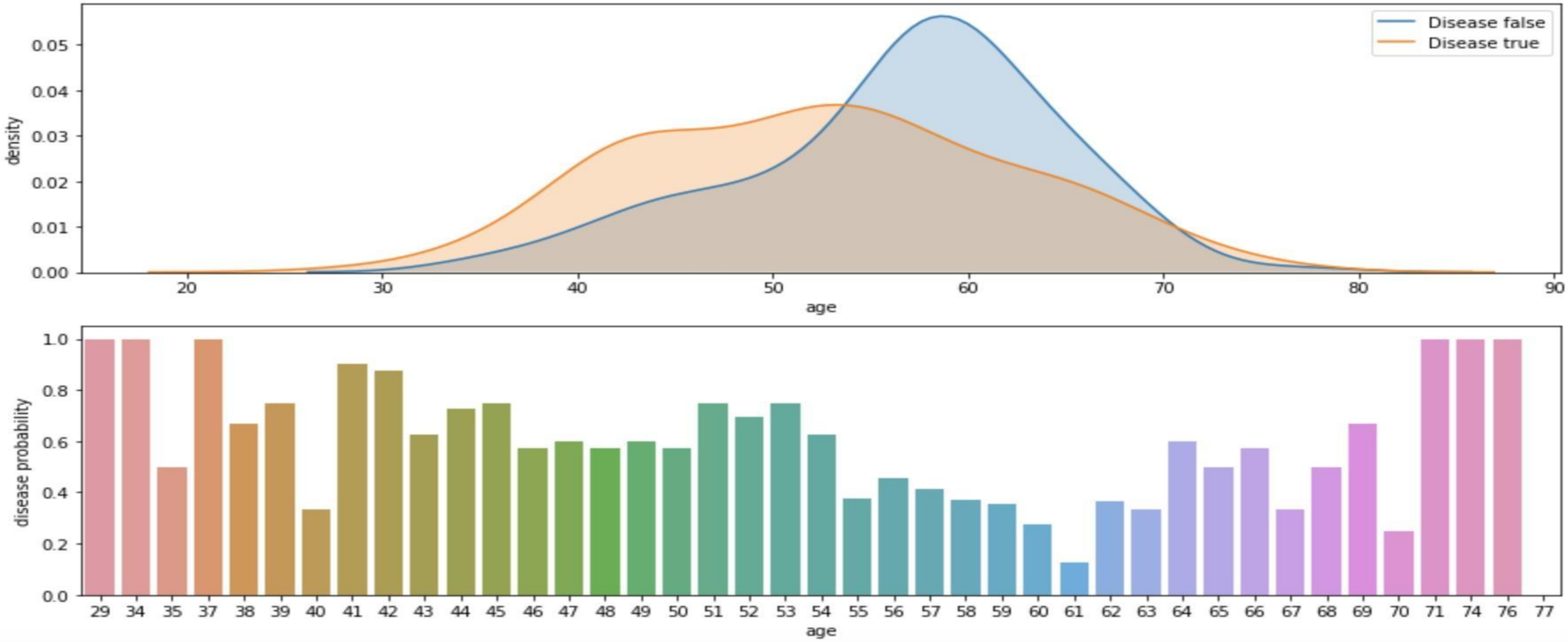


1. age: The person's age in years
2. sex: The person's sex (1 = male, 0 = female)
3. cp: The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
4. trestbps: The person's resting blood pressure (mm Hg on admission to the hospital)
5. chol: The person's cholesterol measurement in mg/dl
6. fbs: The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
7. restecg: Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2showing probable or definite left ventricular hypertrophy by Estes' criteria)
8. thalach: The person's maximum heart rate achieved
9. exang: Exercise induced angina (1 = yes; 0 = no)
10. oldpeak: ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot. See more here)
11. slope: the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
12. ca: The number of major vessels (0-3)
13. thal: A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)
14. target: Heart disease (0 = no, 1 = yes)

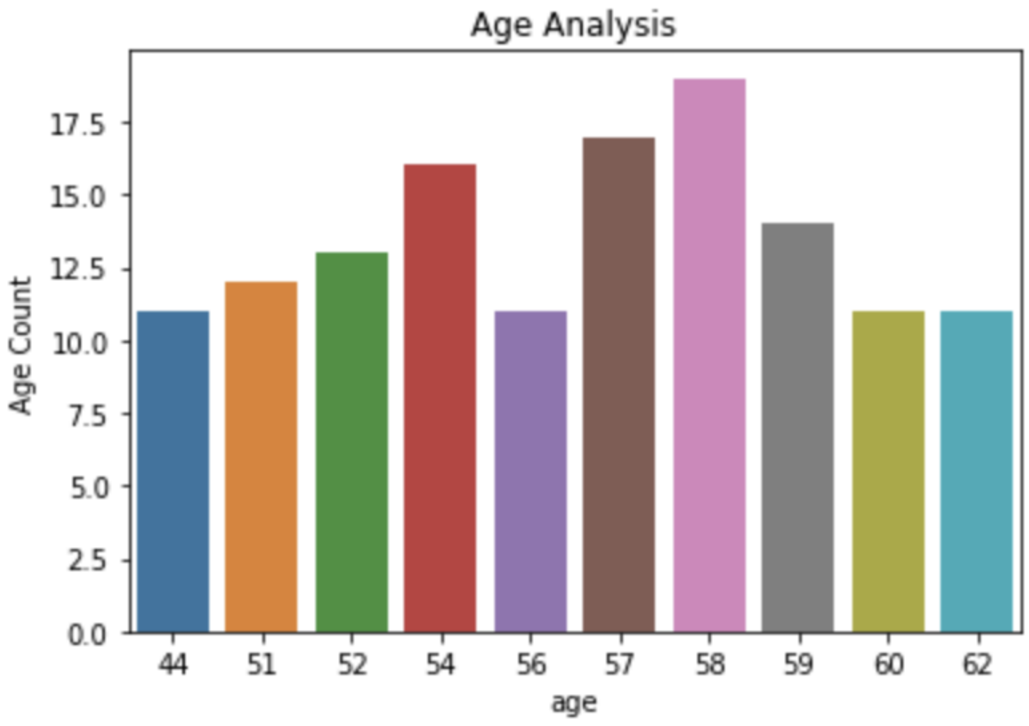
ARCHITECTURAL CHOICES

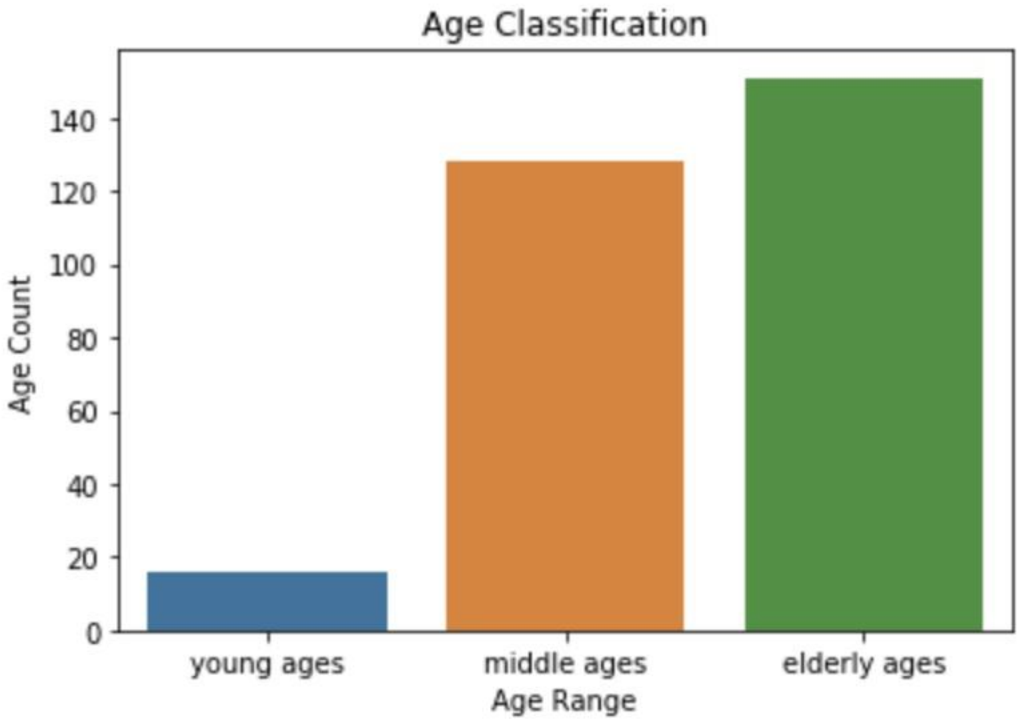
Component	Technology
Development	IBM Watson Studio
	Apache Spark Framework (Spark 2.3)
	Jupyter Notebook with Python 3.5
File Repository	IBM Cloud Object Storage
File Storage	IBM Cloud Object Storage
Data File Format	CSV,Parquet

EXPLORATORY DATA ANALYSIS

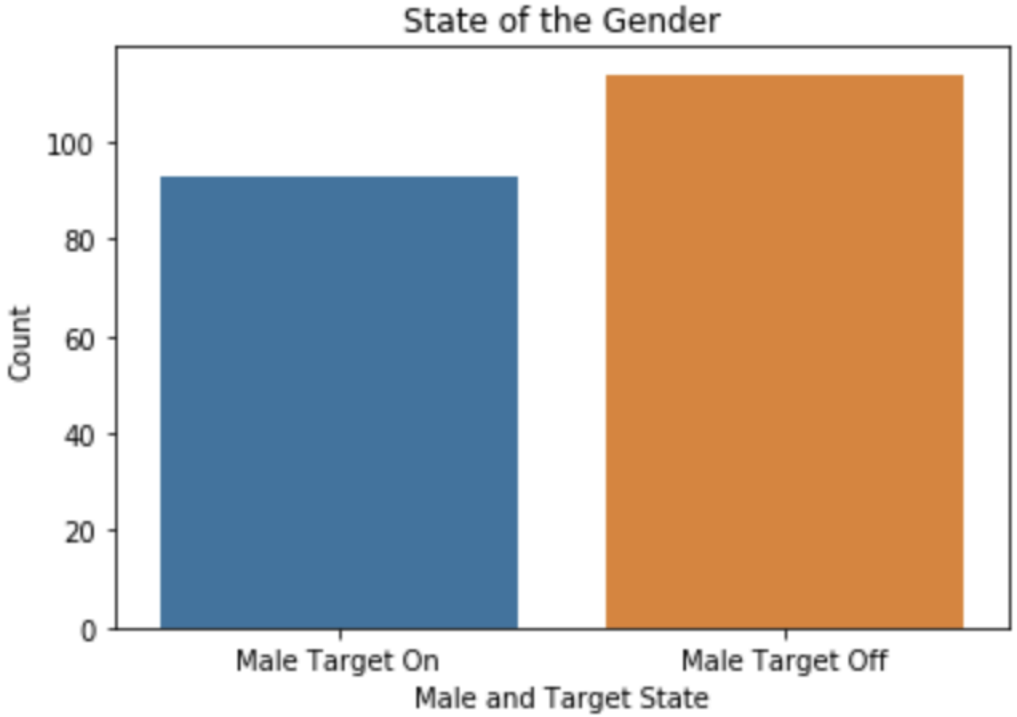


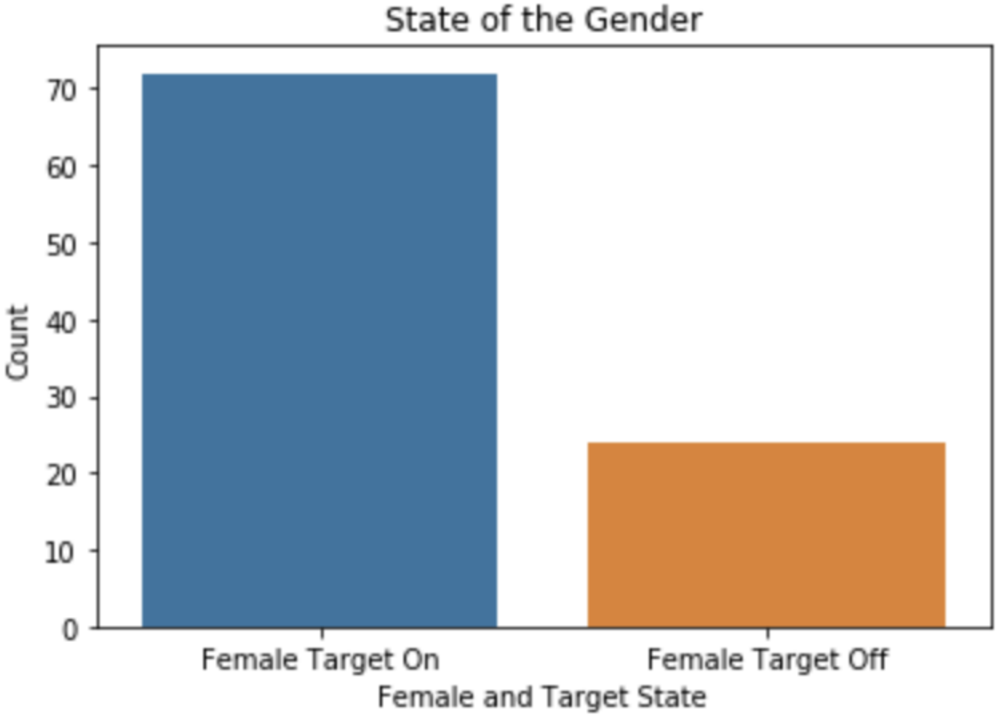
EXPLORATORY DATA ANALYSIS



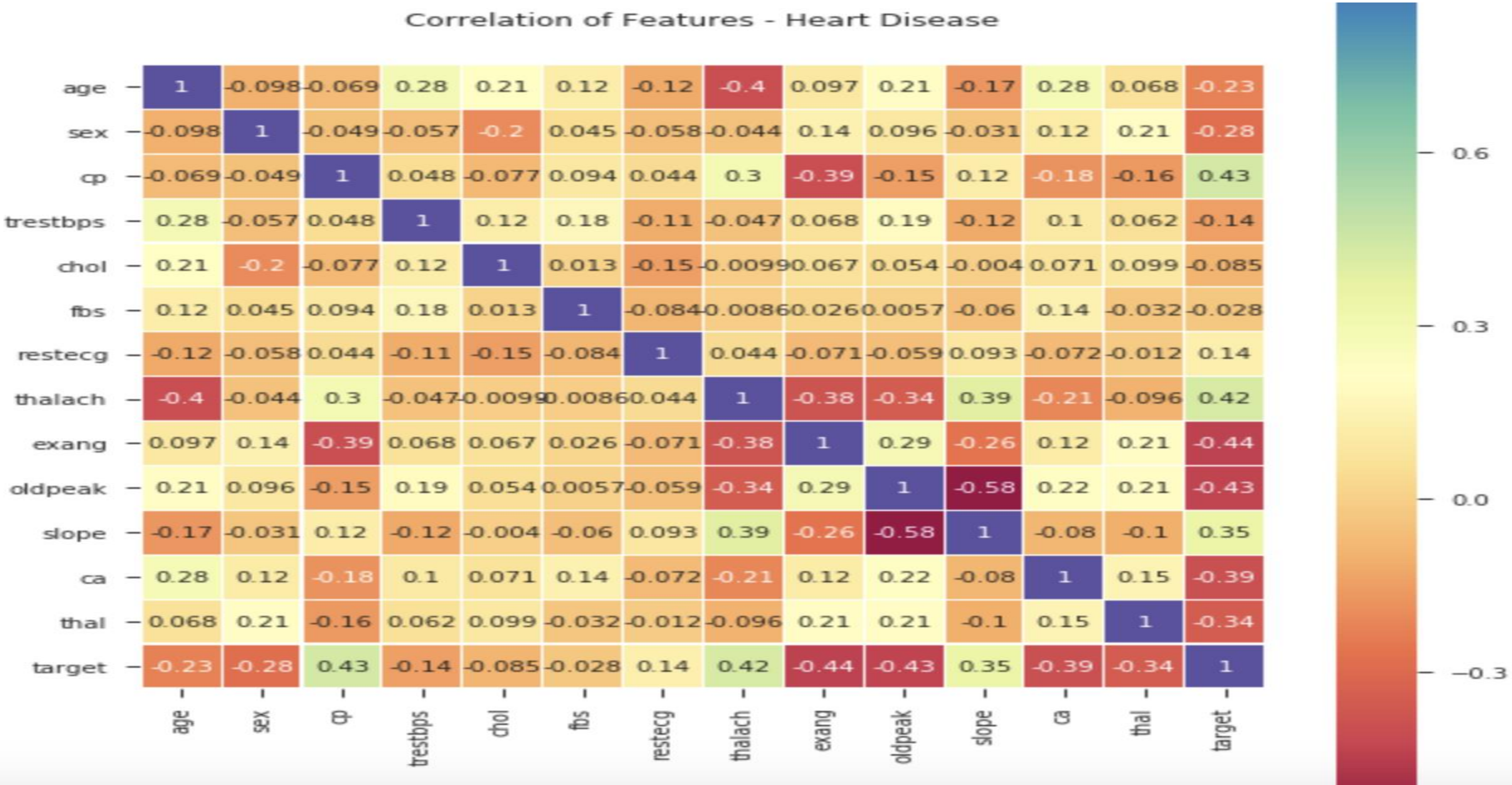


EXPLORATORY DATA ANALYSIS





EXPLORATORY DATA ANALYSIS



Correlation of Features - Heart Disease

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1	-0.098	-0.069	0.28	0.21	0.12	-0.12	-0.4	0.097	0.21	-0.17	0.28	0.068	-0.23
sex	-0.098	1	-0.049	-0.057	-0.2	0.045	-0.058	-0.044	0.14	0.096	-0.031	0.12	0.21	-0.28
cp	-0.069	-0.049	1	0.048	-0.077	0.094	0.044	0.3	-0.39	-0.15	0.12	-0.18	-0.16	0.43
trestbps	0.28	-0.057	0.048	1	0.12	0.18	-0.11	-0.047	0.068	0.19	-0.12	0.1	0.062	-0.14
chol	0.21	-0.2	-0.077	0.12	1	0.013	-0.15	-0.0099	0.067	0.054	-0.004	0.071	0.099	-0.085
fbs	0.12	0.045	0.094	0.18	0.013	1	-0.084	-0.0086	0.026	0.0057	-0.06	0.14	-0.032	-0.028
restecg	-0.12	-0.058	0.044	-0.11	-0.15	-0.084	1	0.044	-0.071	-0.059	0.093	-0.072	-0.012	0.14
thalach	-0.4	-0.044	0.3	-0.047	-0.0099	-0.0086	0.044	1	-0.38	-0.34	0.39	-0.21	-0.096	0.42
exang	0.097	0.14	-0.39	0.068	0.067	0.026	-0.071	-0.38	1	0.29	-0.26	0.12	0.21	-0.44
oldpeak	0.21	0.096	-0.15	0.19	0.054	0.0057	-0.059	-0.34	0.29	1	-0.58	0.22	0.21	-0.43
slope	-0.17	-0.031	0.12	-0.12	-0.004	-0.06	0.093	0.39	-0.26	-0.58	1	-0.08	-0.1	0.35
ca	0.28	0.12	-0.18	0.1	0.071	0.14	-0.072	-0.21	0.12	0.22	-0.08	1	0.15	-0.39
thal	0.068	0.21	-0.16	0.062	0.099	-0.032	-0.012	-0.096	0.21	0.21	-0.1	0.15	1	-0.34
target	-0.23	-0.28	0.43	-0.14	-0.085	-0.028	0.14	0.42	-0.44	-0.43	0.35	-0.39	-0.34	1

FEATURE CREATION

Feature Engineering Method	Description
One Hot Encoder	*Create one hot encoders for the categorical features*
Vector Assembler	Combine a given list of columns into a single vector column
Normalizer	Transform a dataset of Vector rows

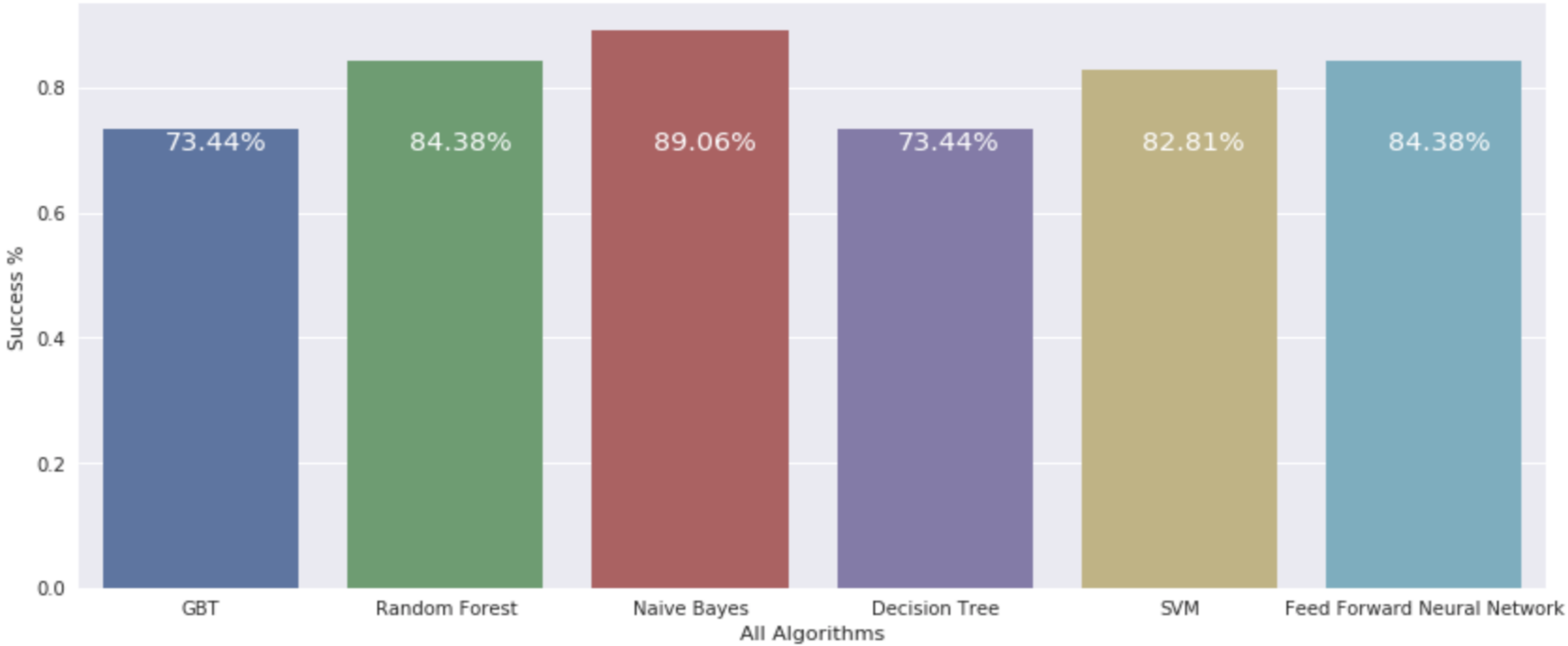
SOLUTION

- The model is an example of binary classification
- Heart Disease (0 = no, 1 = yes)
- Algorithms used for prediction model :
  - Supervised Machine Learning
    - Gradient Boosted Trees, Random Forest Classification,
    - Naive Bayes Algorithm, Decision Tree, SVM
  - Deep Learning Algorithm
    - Feed Forward Neural Network (Multi Layer Perceptron)

MODEL PERFORMANCE

- The Highest Machine Learning Algorithm
    - Naive Bayes Algorithm
        - Validation Accuracy :%89
- Deep Learning Algorithm
    - Multi Layer Perceptron
        - Validation Accuracy :%84

RESULT



Success %
0.0
0.2
0.4
0.6
0.8
73.44%
84.38%
89.06%
73.44%
82.81%
84.38%
GBT
Random Forest
Naive Bayes
Decision Tree
SVM
Feed Forward Neural Network
All Algorithms