Term-Project: Forecasting Heart Disease with Machine Learning Algorithms

Subject : Data Mining
Lecturer : PHAUK Sokkhey & CHAN Sophal
Student : ROS Sereiwathna

Table of Contents

INTRODUCTION

The **Cardiovar disorders** are frequently substituted for **heart disease**. This illness is generally associated with blood vessel blockages or narrowing which can lead to heart attacks, strokes, and angina.

In this research project, we will be working on using machine learning algorithms to help predict whether a patient is having a heart disease or not.

DATASET

This dataset contains of 14 feature attributes such as:

1. age	2. sex	3. cp	4. trestbps	5.chol
6. fbs	7. restecg	8. thalach	9. exang	10. oldpeak
11. slope	12. ca	13. thal	14. target	

This dataset is from UCI Machine Learning Repository where the dataset is available here.

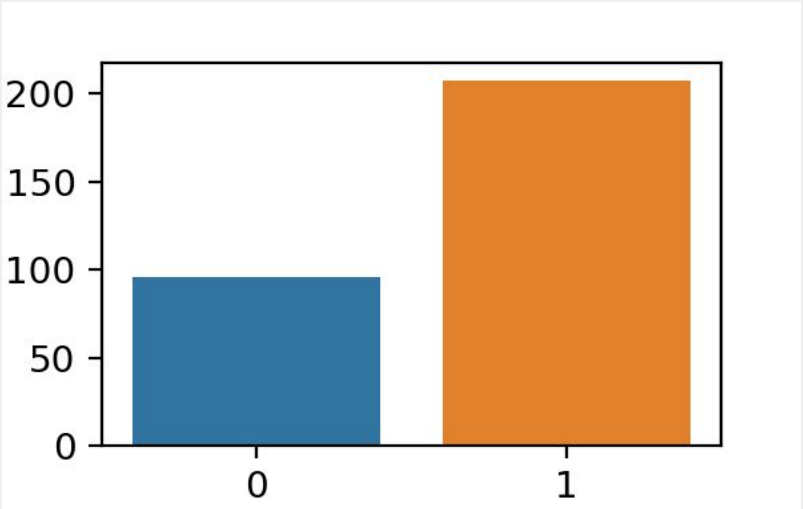
DATASET

OVERVIEW OF THE DATA

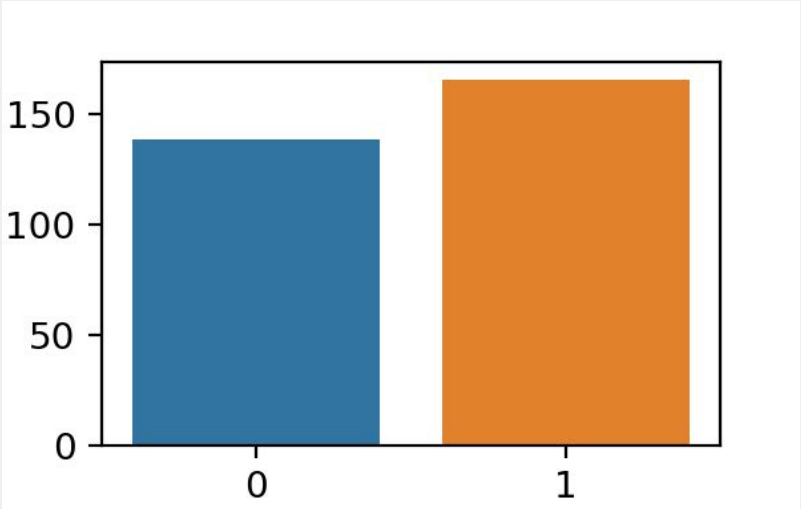
	count	mean	std	min	25%	50%	75%	max
age	303.0	54.366337	9.082101	29.0	47.5	55.0	61.0	77.0
sex	303.0	0.683168	0.466011	0.0	0.0	1.0	1.0	1.0
cp	303.0	0.966997	1.032052	0.0	0.0	1.0	2.0	3.0
trestbps	303.0	131.623762	17.538143	94.0	120.0	130.0	140.0	200.0
chol	303.0	246.264026	51.830751	126.0	211.0	240.0	274.5	564.0
fbs	303.0	0.148515	0.356198	0.0	0.0	0.0	0.0	1.0
restecg	303.0	0.528053	0.525860	0.0	0.0	1.0	1.0	2.0
thalach	303.0	149.646865	22.905161	71.0	133.5	153.0	166.0	202.0
exang	303.0	0.326733	0.469794	0.0	0.0	0.0	1.0	1.0
oldpeak	303.0	1.039604	1.161075	0.0	0.0	0.8	1.6	6.2
slope	303.0	1.399340	0.616226	0.0	1.0	1.0	2.0	2.0
ca	303.0	0.729373	1.022606	0.0	0.0	0.0	1.0	4.0
thal	303.0	2.313531	0.612277	0.0	2.0	2.0	3.0	3.0
target	303.0	0.544554	0.498835	0.0	0.0	1.0	1.0	1.0

DATA VISUALIZATION

SEX

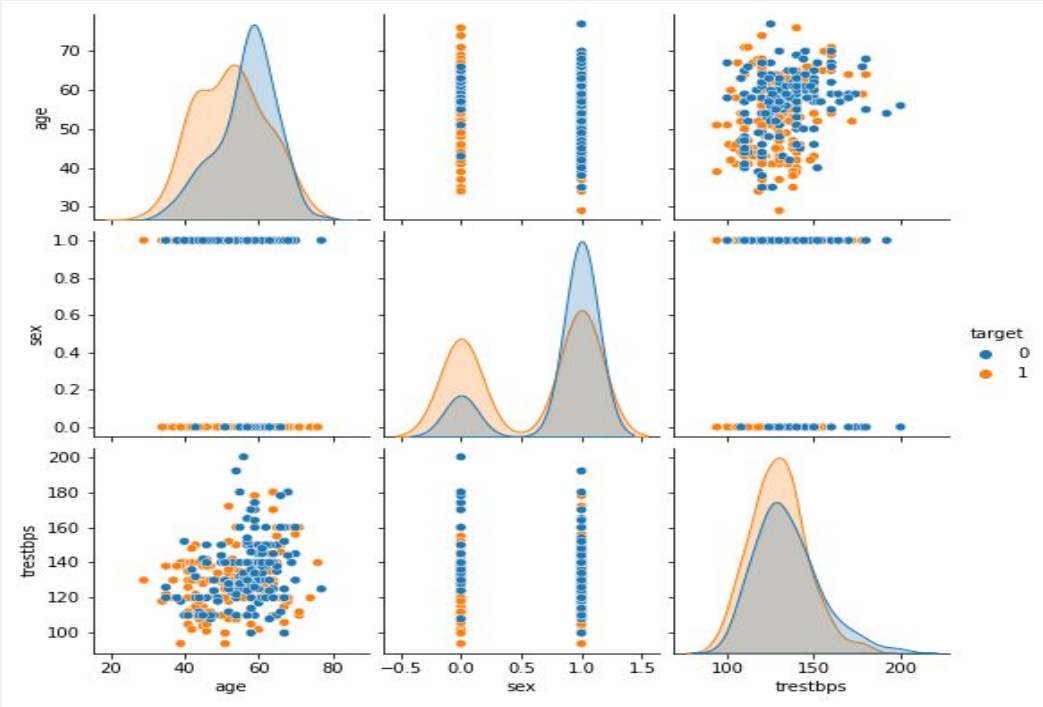


TARGET



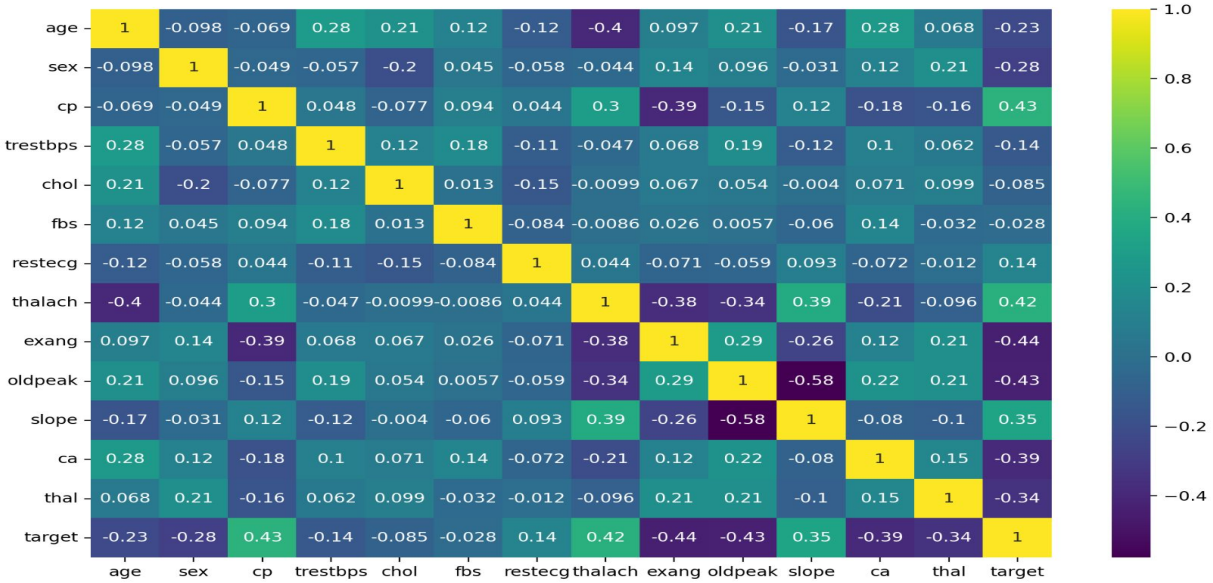
DATA VISUALIZATION

Relation between age, sex, trestbps and target



DATA VISUALIZATION

Correlations between all the features



METHODOLOGY

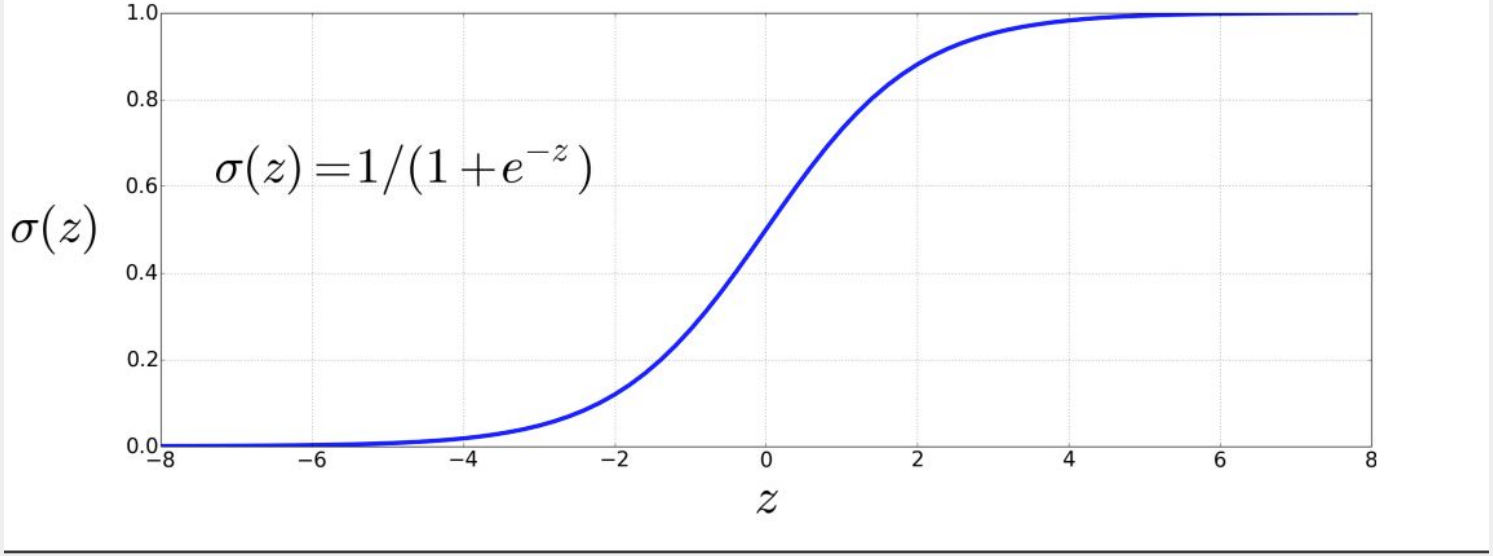
Scaling data with **Standard Score**

$$Z = \frac{x - u}{s}$$

METHODOLOGY

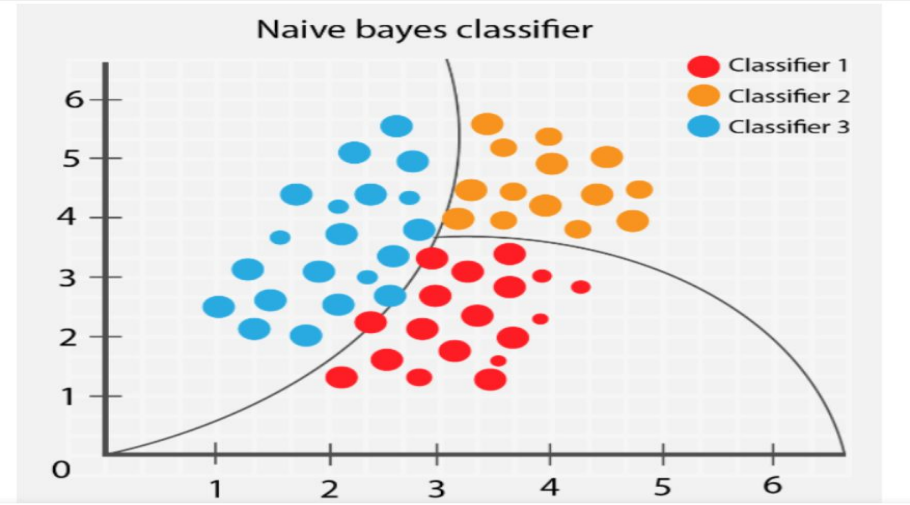
Logistic Regression Classifier

$$z = wx + b$$



METHODOLOGY

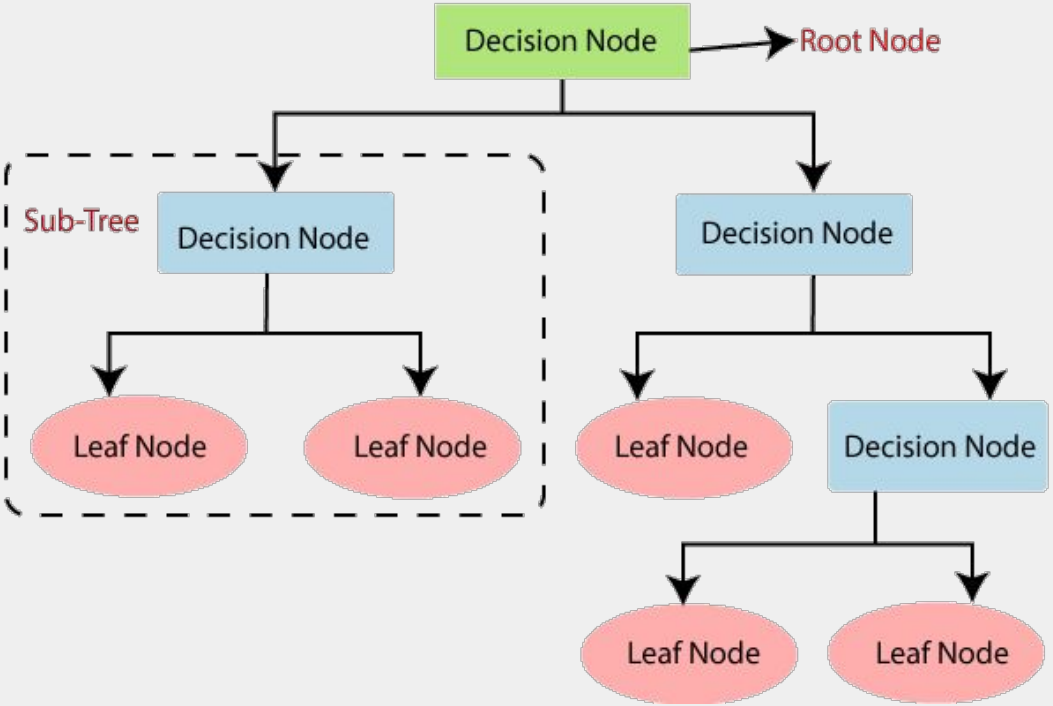
Naïve Bayes Classifier



$$P(x|target) \propto P(x)P(target|x)$$

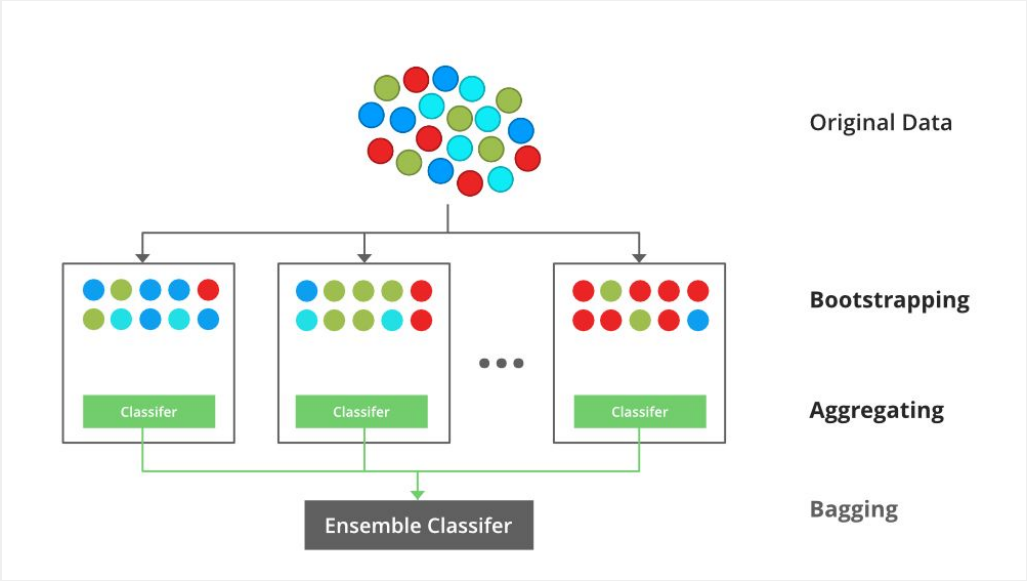
METHODOLOGY

Decision Tree Classifier



METHODOLOGY

XGBoost Classifier



METHODOLOGY

Evaluation Metrics

Precision = $\frac{TP}{TP + FP}$

Recall = $\frac{TP}{TP + FN}$

F1-score = $\frac{TP}{TP + \frac{1}{2}(FP + FN)}$

Accuracy = $\frac{TP + TN}{TP + TN + FP + FN}$

RESULT

Models Evaluation

	Precision	Recall	F1-Score	Accuracy
Logistic Regression	87%	87%	87%	87%
Naïve Bayes	80%	80%	80%	80%
Decision Tree	85%	85%	85%	85%
XGBoost	80%	81%	80%	80%

CONCLUSION

Even though Logistic Regression Classifier is non-complicated and modest than other architectures, however its performance is still best among the others.

We have seen that these models perform with acceptable results.