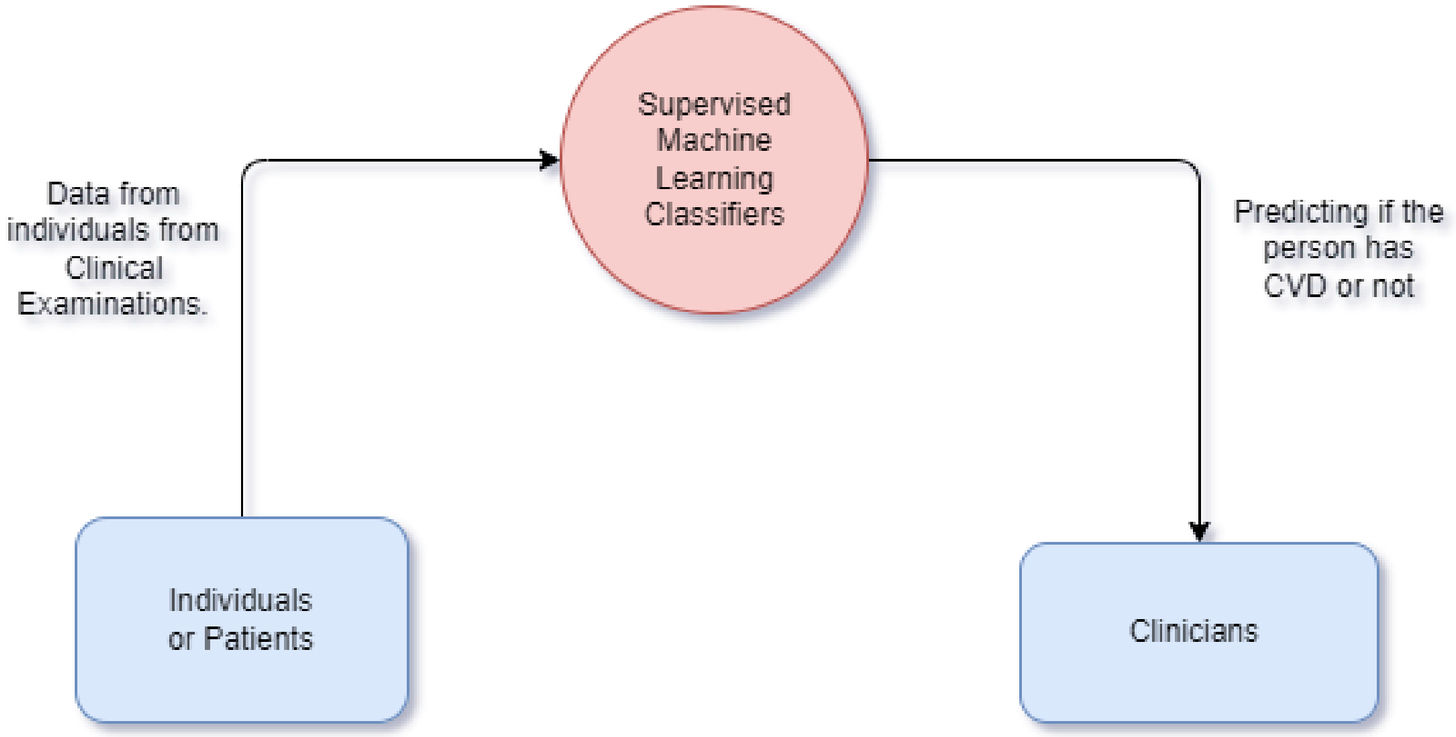
Prediction of CVD using Supervised Learning Classifiers

Under the guidance of Dr. Santwana Sagnika

Major Project ~ 8th Semester

Introduction to Problem Statement



The aim of this proposed supervised learning classifier model is to detect CVD at the earlier stages of clinical examinations in an individual.

Research Work & Data Gathering

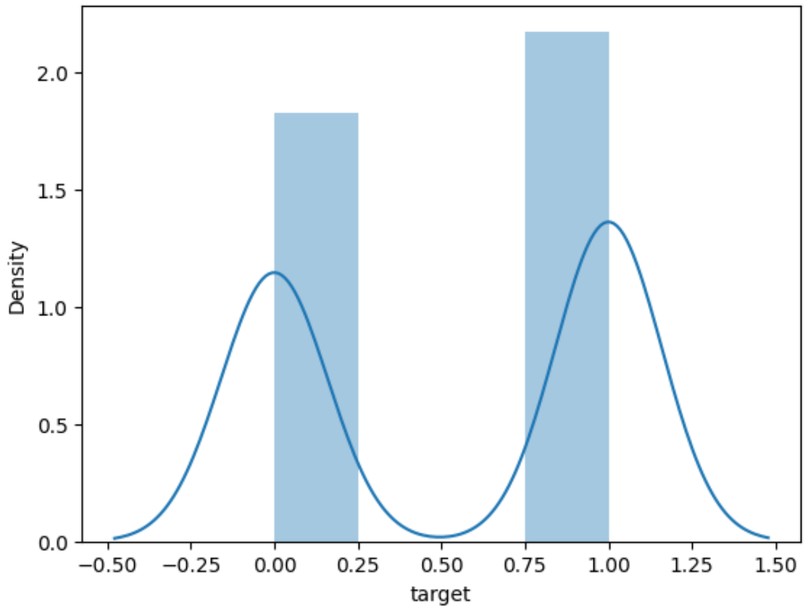
Datasets that were gathered for the scope of this project:

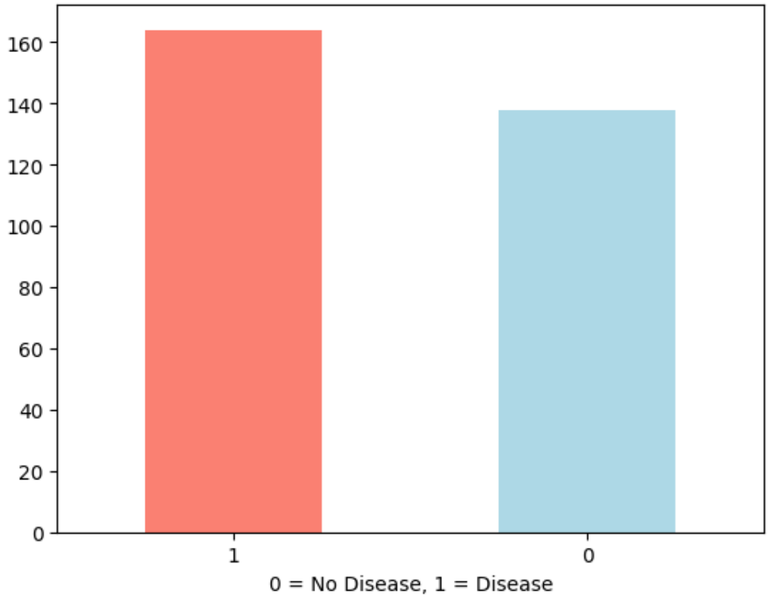
- Kaggle Dataset of 70,000+ records.
- UCI Machine Learning Repository with 1000+ records.
- Cleveland Dataset from UCI Heart Disease Database with 303 records.

Sl. No.	Features	Description
1	age	Age in years
2	sex	Male = 0; Female = 1
3	cp	Chest Pain type
4	trtbps	Resting Blood Pressure (in mmHg) Range: 94-200
5	chol	Serum Cholestrol (in mg/dl) Range: 126-564
6	fbs	Fastin Blood Sugar Range: < and > 120 mg/dl True = 1; False = 0
7	restecg	Resting Electrocardiographic Result
8	thalachh	Maximum Heart Rate 71 to 202
9	exng	Exercise Induced Angina Yes = 1; No = 0
10	oldpeak	ST Depression due to exercise w.r.t. rest: 0 to 0.2
11	slp	The slope of the peak exercise ST segment: 0 to 1
12	caa	Number of major blood vessels: 0 to 3
13	thall	Normal value = 3
14	target	Cardiac Disease Yes = 1; No = 0

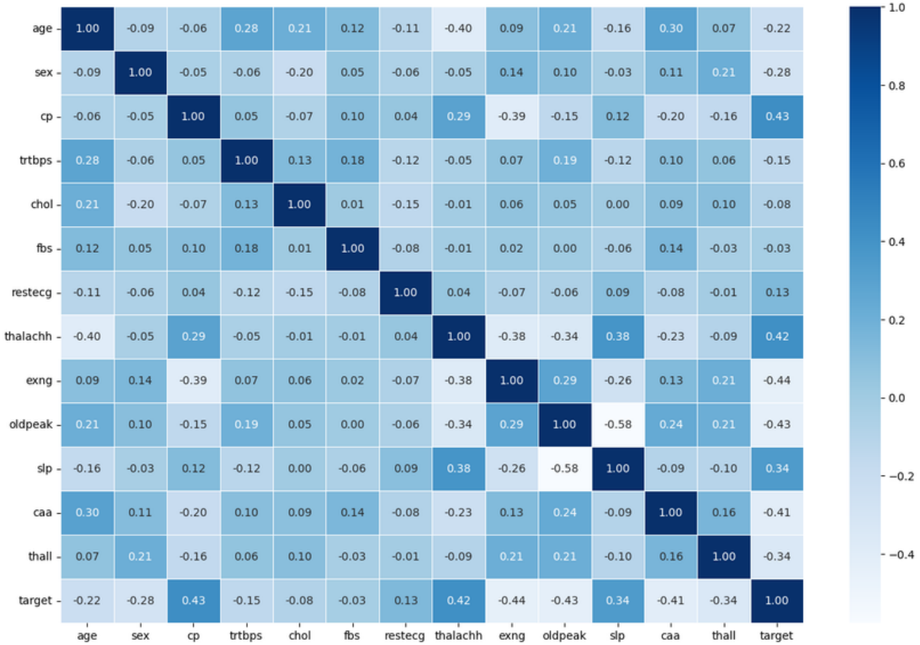
Exploratory Data Analysis

For the EDA, the features are represented using the bars, graphs and plots as the following:

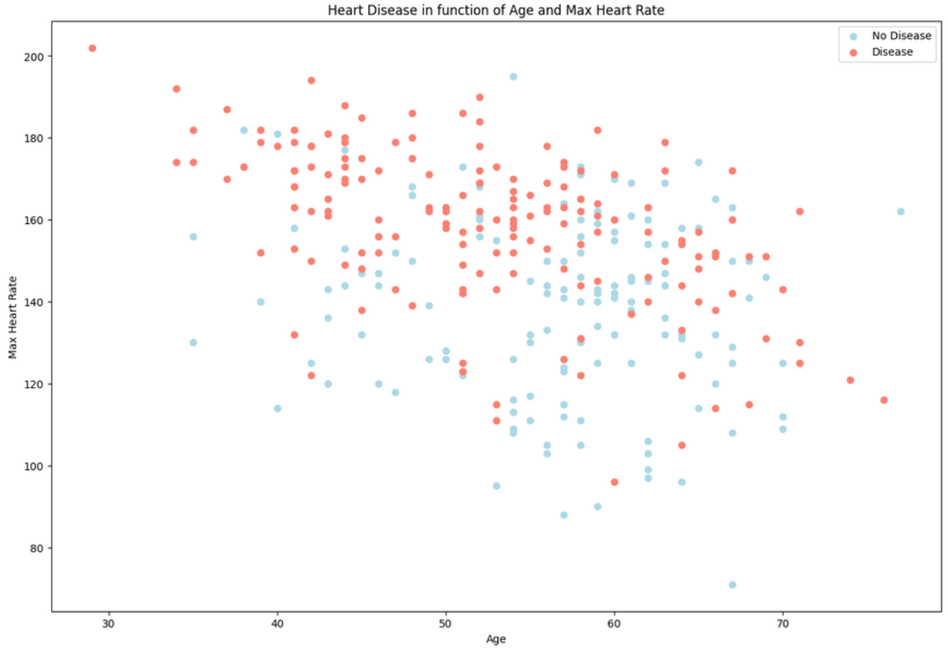




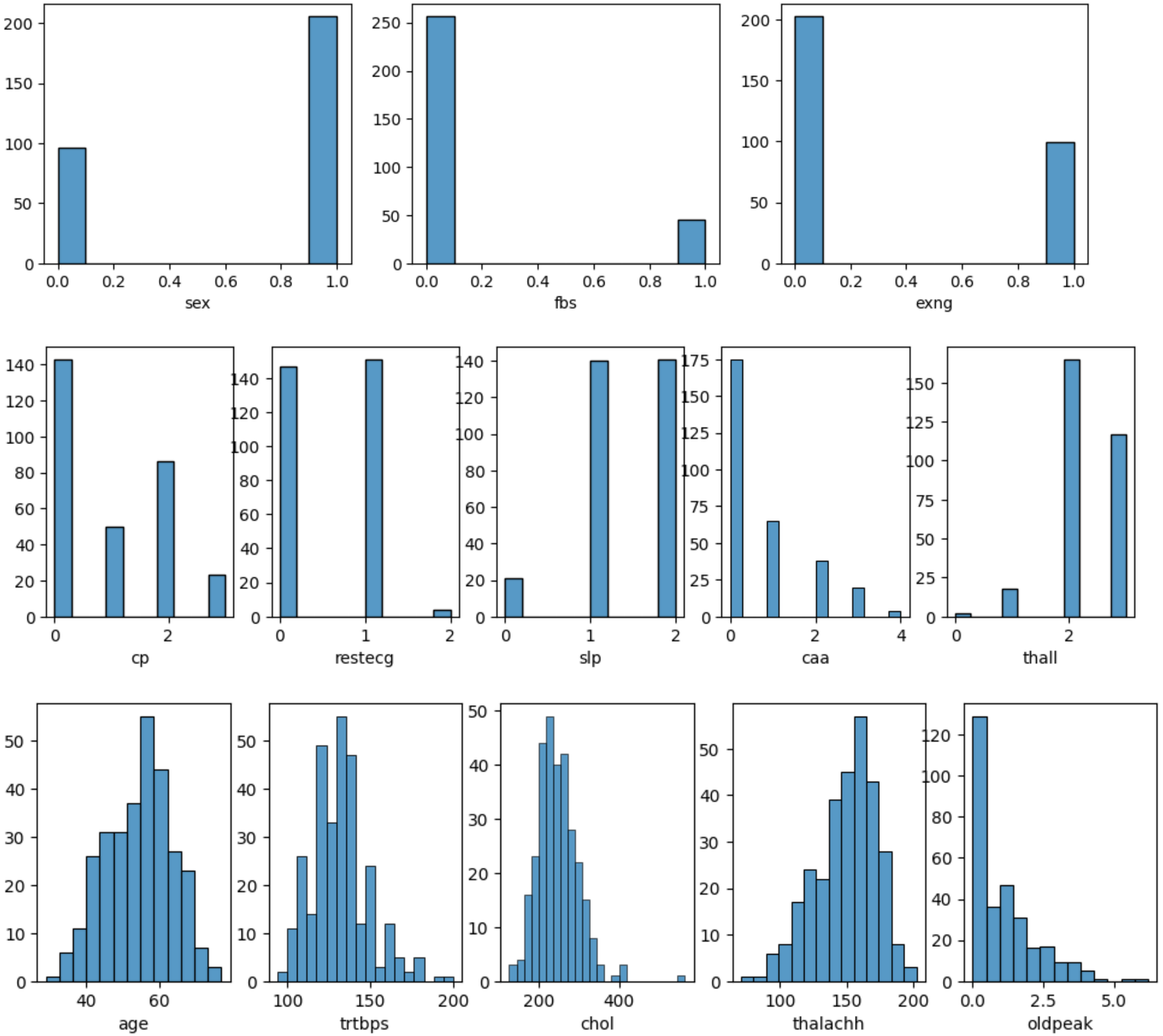
Distribution Plot & Bar Plot of the "target" variable



Correlation Matrix Heatmap for all features



Scatter Plot of variables “age” and “thalachh”



Histogram for all Binary, Categorical & Numerical features

Data Preprocessing

For Data Preprocessing purpose, we cleaned our data and then applied feature scaling through StandardScaler.

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	target
0	0.949794	1	3	0.764066	-0.261285	1	0	0.018826	0	1.084022	0	0	1	1
1	-1.928548	1	2	-0.091401	0.067741	0	1	1.636979	0	2.118926	0	0	2	1
2	-1.485726	0	1	-0.091401	-0.822564	0	0	0.980971	0	0.307844	2	0	2	1
3	0.174856	1	1	-0.661712	-0.203222	0	1	1.243374	0	-0.209608	2	0	2	1
4	0.285561	0	0	-0.661712	2.080602	0	1	0.587366	1	-0.382092	2	0	2	1

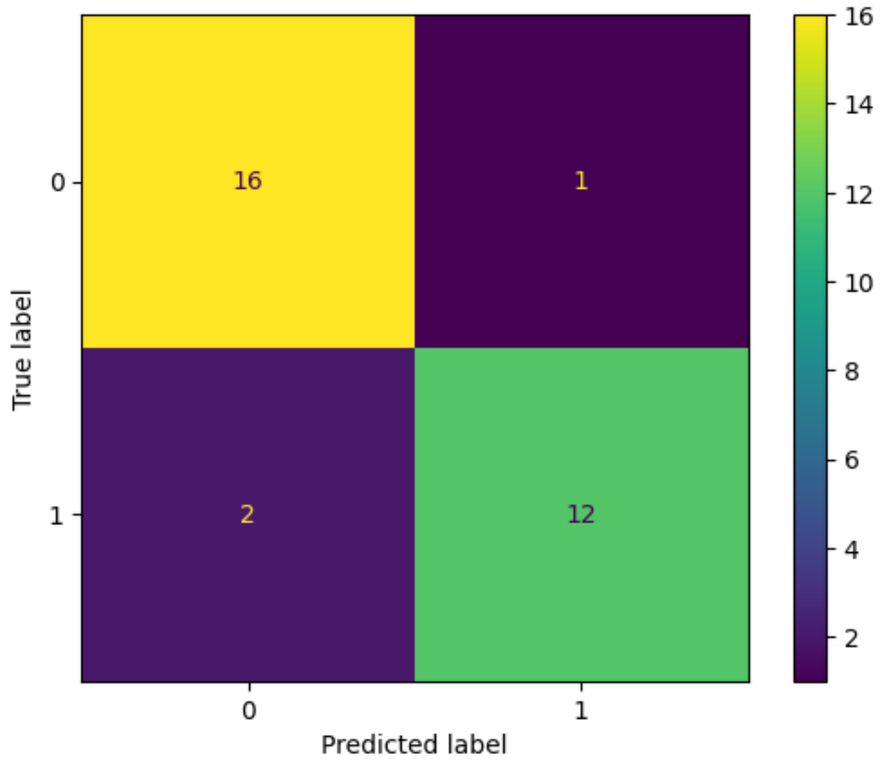
Data after Feature Scaling

Machine Learning Models

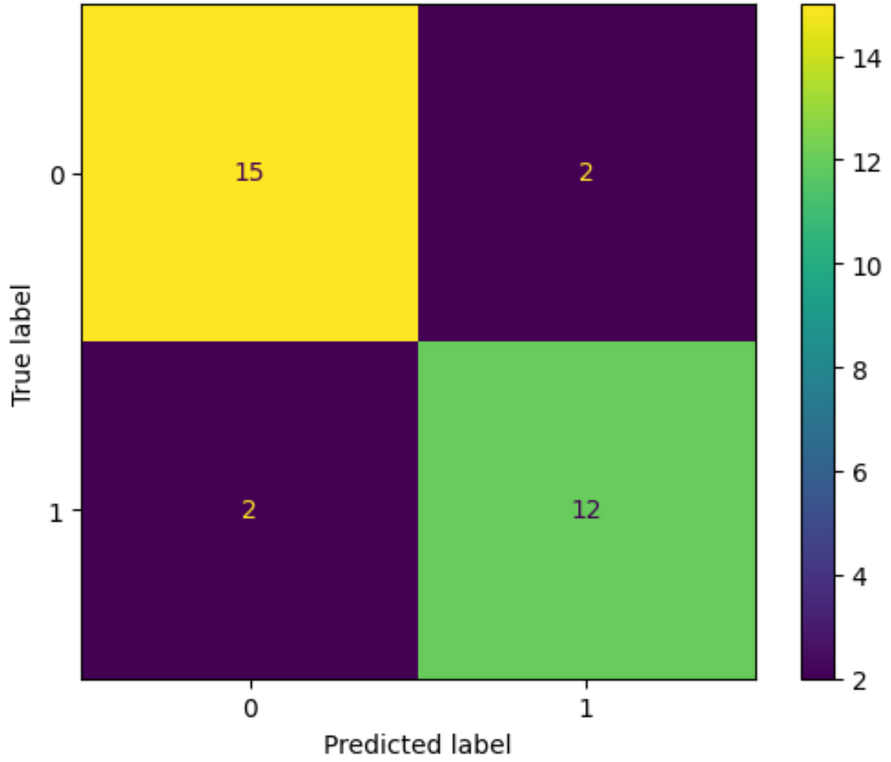
Supervised Machine Learning Classifiers:

- Logistic Regression
- Support Vector Classifier
- K-Neighbors Classifier
- Decision Tree
- Random Forest
- Gradient Boosting
- Gaussian Naive Bayes
- Bernoulli Naive Bayes

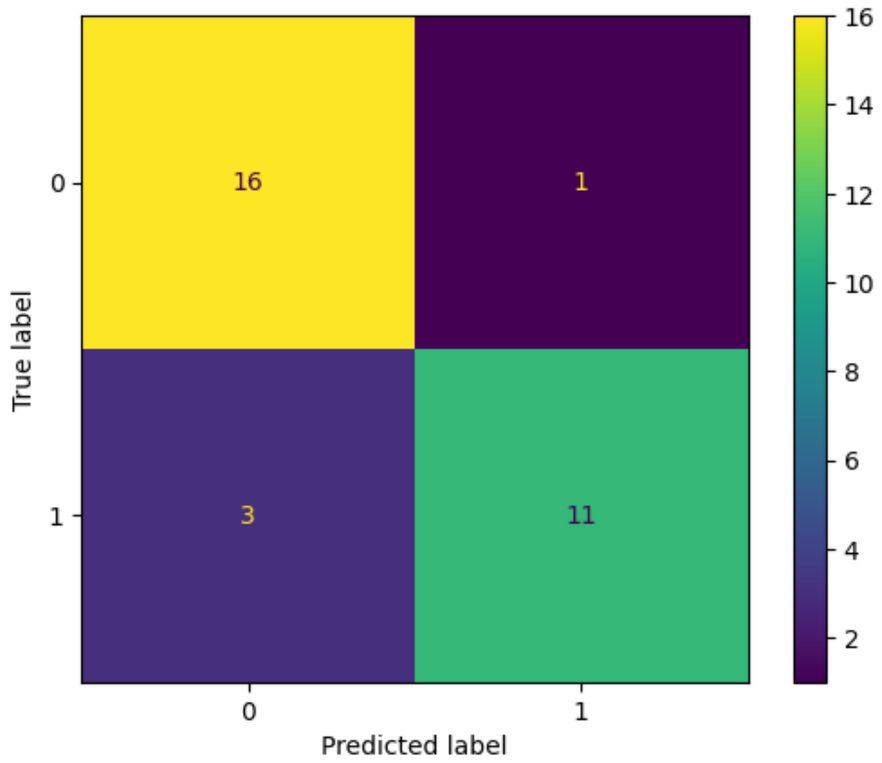
Logistic Regression



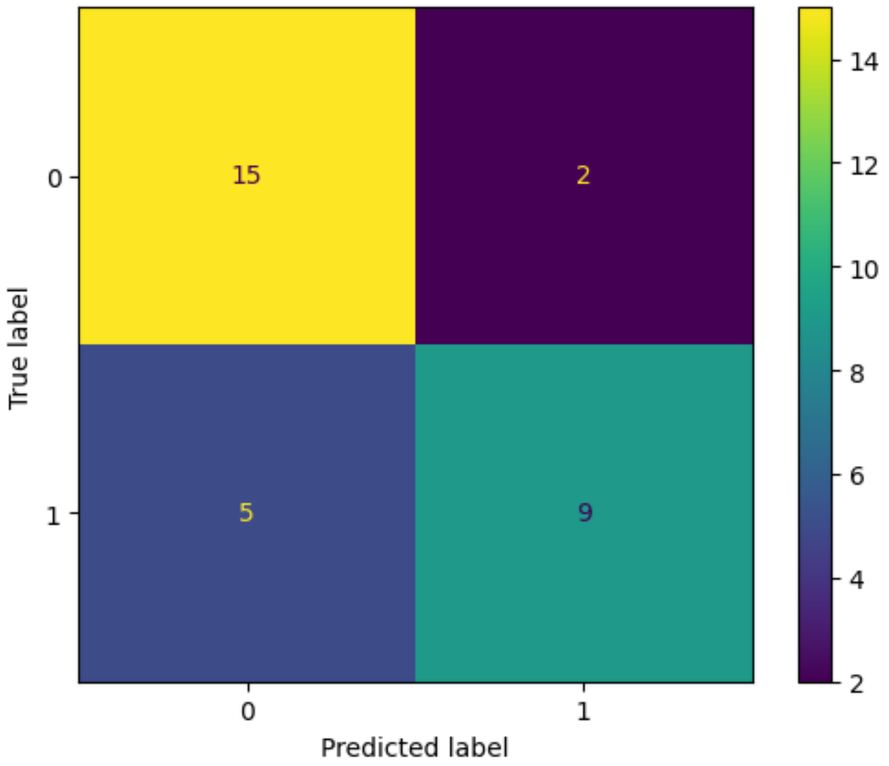
Support Vector Classifier



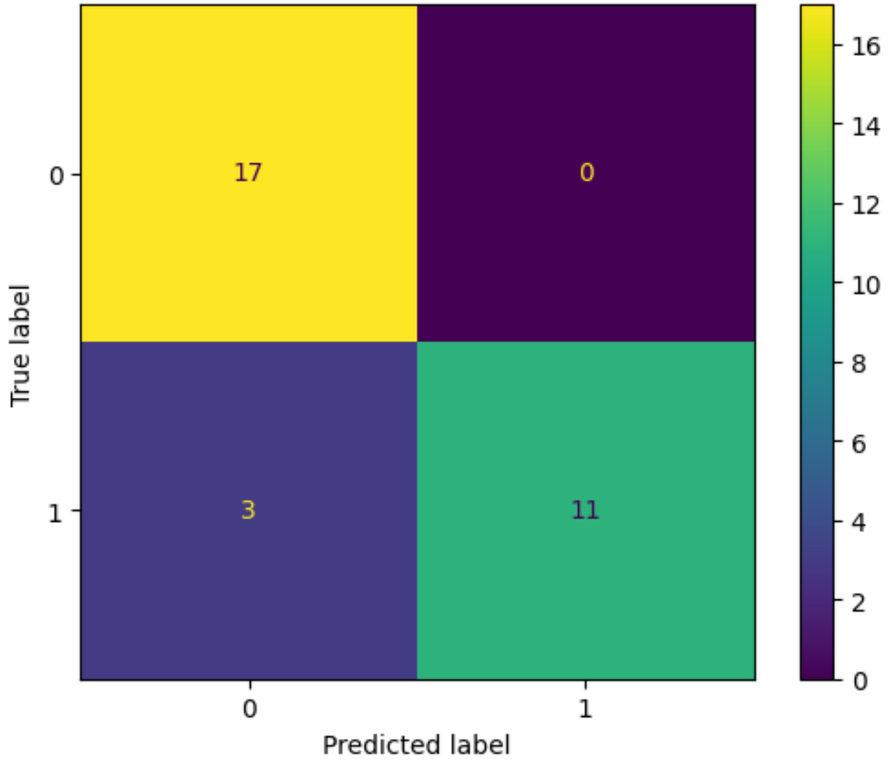
K-Neighbors Classifier



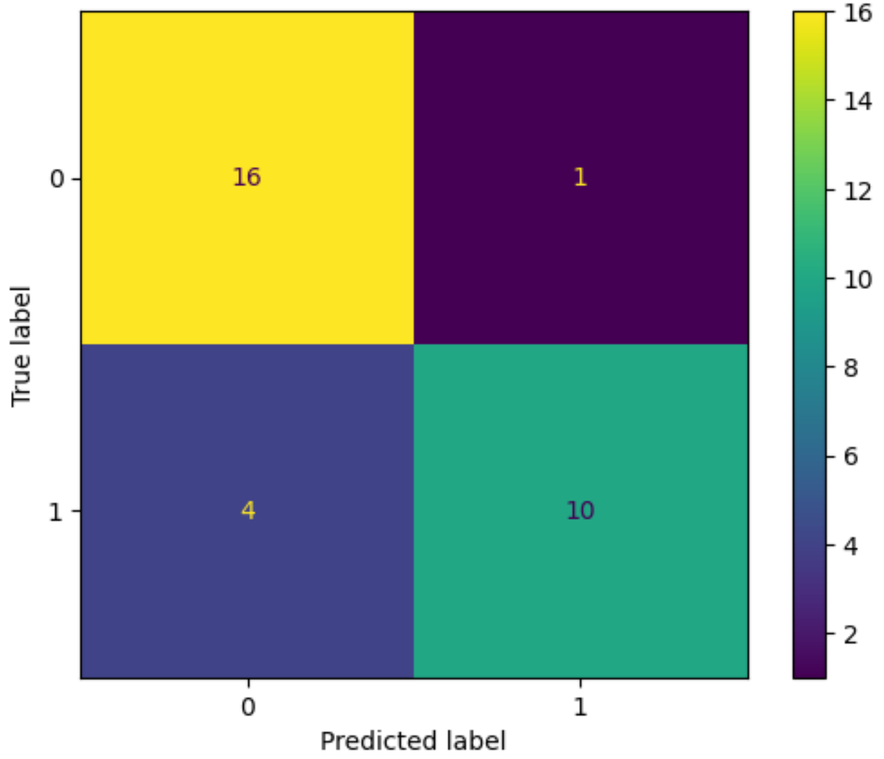
Decision Tree



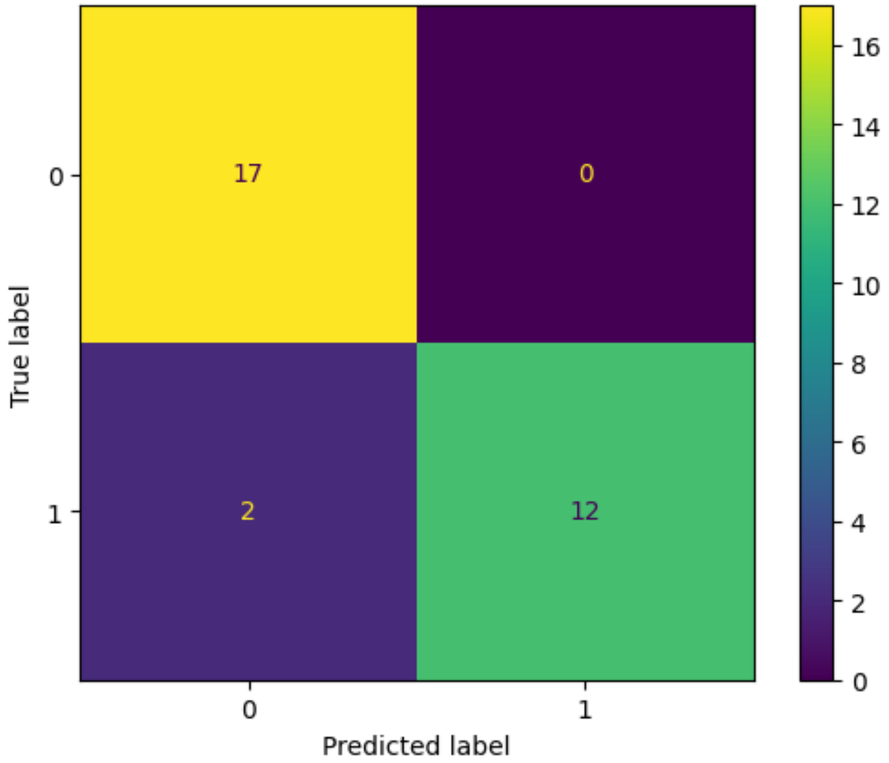
Random Forest



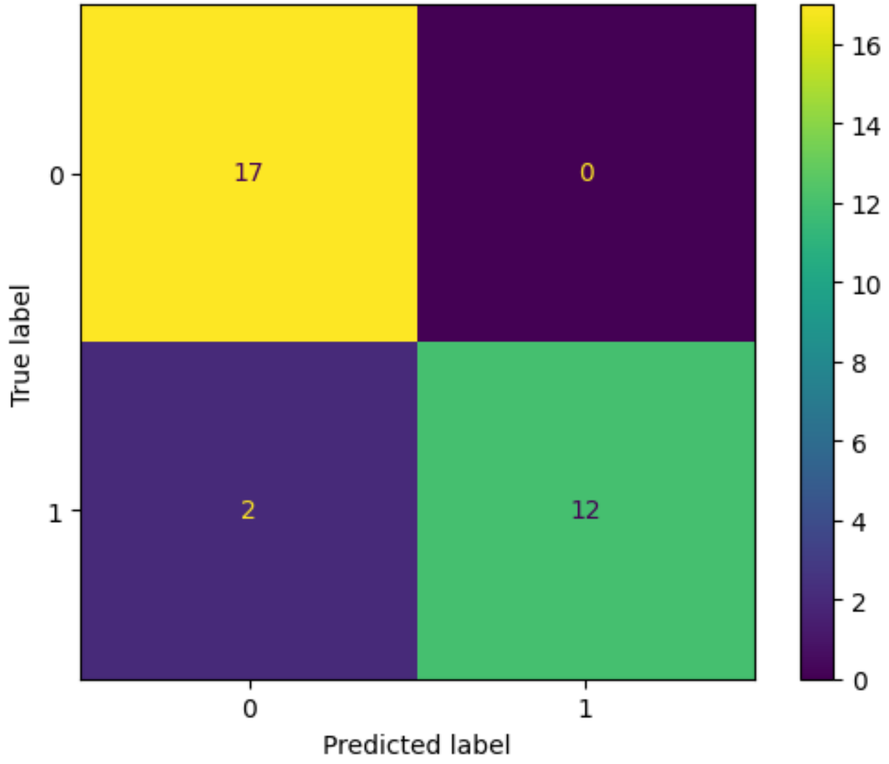
Gradient Boosting



Gaussian Naive Bayes



Bernoulli Naive Bayes



Feature Selection

Feature Selection through Backward Elimination by P-value Approach

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.3750	0.886	2.680	0.007	0.638	4.112
age	-0.0133	0.212	-0.063	0.950	-0.429	0.402
sex	-1.7509	0.468	-3.740	0.000	-2.669	-0.833
cp	0.8473	0.186	4.566	0.000	0.484	1.211
trtbps	-0.3540	0.182	-1.944	0.052	-0.711	0.003
chol	-0.2319	0.197	-1.179	0.238	-0.617	0.154
fbs	0.0735	0.532	0.138	0.890	-0.970	1.117
restecg	0.4506	0.349	1.293	0.196	-0.232	1.134
thalachh	0.5290	0.239	2.214	0.027	0.061	0.997
exng	-0.9810	0.410	-2.394	0.017	-1.784	-0.178
oldpeak	-0.6071	0.249	-2.441	0.015	-1.095	-0.120
slp	0.5891	0.350	1.684	0.092	-0.097	1.275
caa	-0.8260	0.202	-4.091	0.000	-1.222	-0.430
thall	-0.8872	0.291	-3.052	0.002	-1.457	-0.317

	coef	std err	z	P>|z|	[0.025	0.975]
const	3.2783	0.745	4.398	0.000	1.817	4.739
sex	-1.5153	0.420	-3.604	0.000	-2.339	-0.691
cp	0.8177	0.178	4.604	0.000	0.470	1.166
trtbps	-0.3624	0.172	-2.112	0.035	-0.699	-0.026
thalachh	0.5799	0.208	2.781	0.005	0.171	0.988
exng	-0.9937	0.398	-2.496	0.013	-1.774	-0.213
oldpeak	-0.7976	0.216	-3.689	0.000	-1.221	-0.374
caa	-0.7590	0.186	-4.072	0.000	-1.124	-0.394
thall	-0.8911	0.281	-3.175	0.001	-1.441	-0.341

Result Analysis

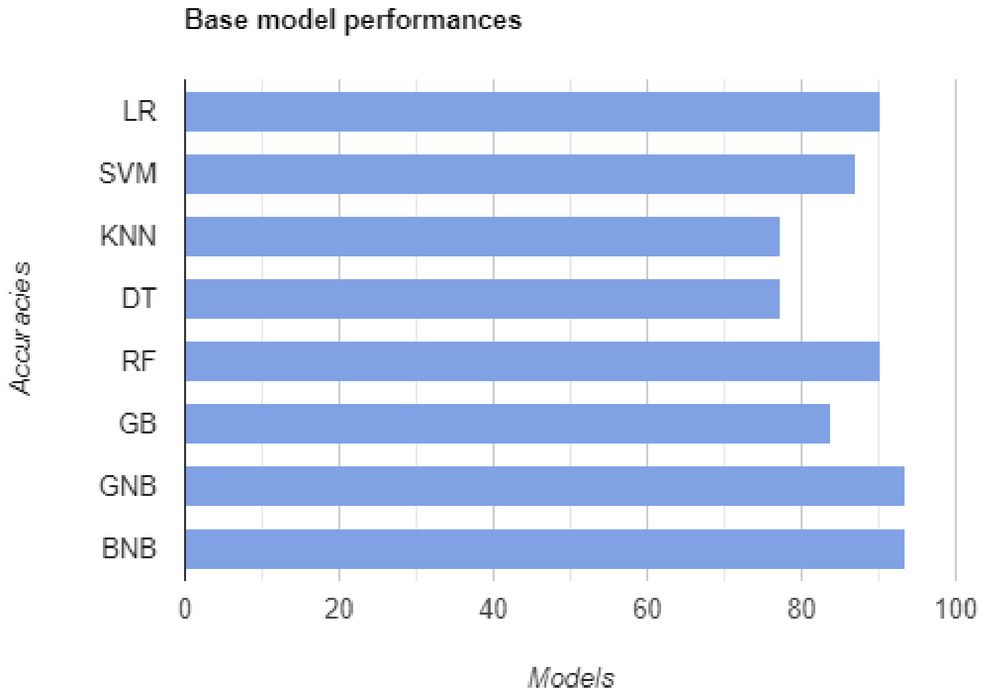
Base Models Performance Metrics

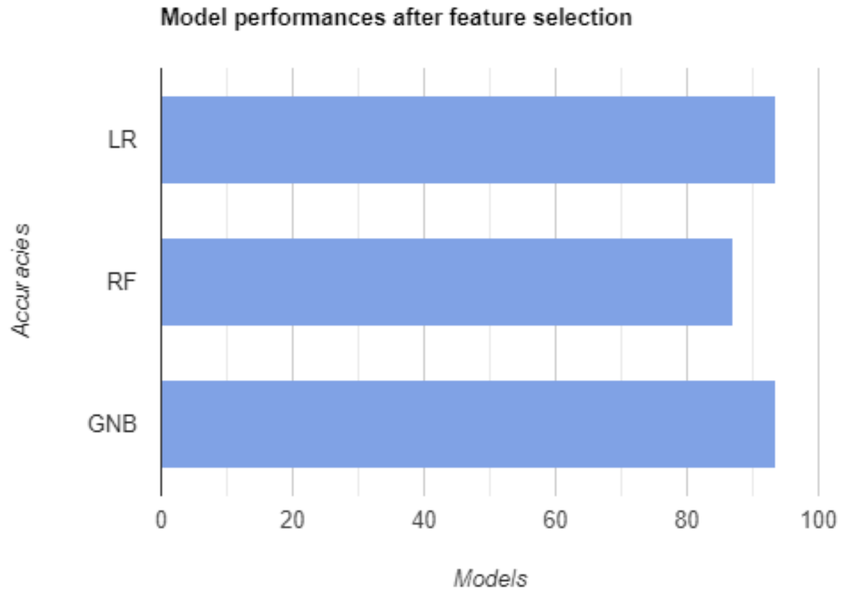
	Method	Accuracy	TP	FP	TN	FN	Precision	Recall	ROC_AUC	PR_AUC	F1_Score	F2_Score
0	Logistic Regression	0.903226	12.0	1.0	16.0	2.0	0.923077	0.857143	0.941176	0.949937	0.888889	0.869565
1	Support Vector Classifier	0.870968	12.0	2.0	15.0	2.0	0.857143	0.857143	0.936975	0.947750	0.857143	0.857143
2	K-Neighbours Classifier	0.774194	8.0	1.0	16.0	6.0	0.888889	0.571429	0.852941	0.859639	0.695652	0.615385
3	Decision Tree Classifier	0.774194	9.0	2.0	15.0	5.0	0.818182	0.642857	0.762605	0.811165	0.720000	0.671642
4	Random Forest Classifier	0.903226	11.0	0.0	17.0	3.0	1.000000	0.785714	0.951681	0.956746	0.880000	0.820896
5	Gradient Boosting Classifier	0.838710	10.0	1.0	16.0	4.0	0.909091	0.714286	0.882353	0.910915	0.800000	0.746269
6	Gaussian Naive Bayes	0.935484	12.0	0.0	17.0	2.0	1.000000	0.857143	0.928571	0.943933	0.923077	0.882353
7	Bernoulli Naive Bayes	0.935484	12.0	0.0	17.0	2.0	1.000000	0.857143	1.000000	1.000000	0.923077	0.882353

After Feature Selection

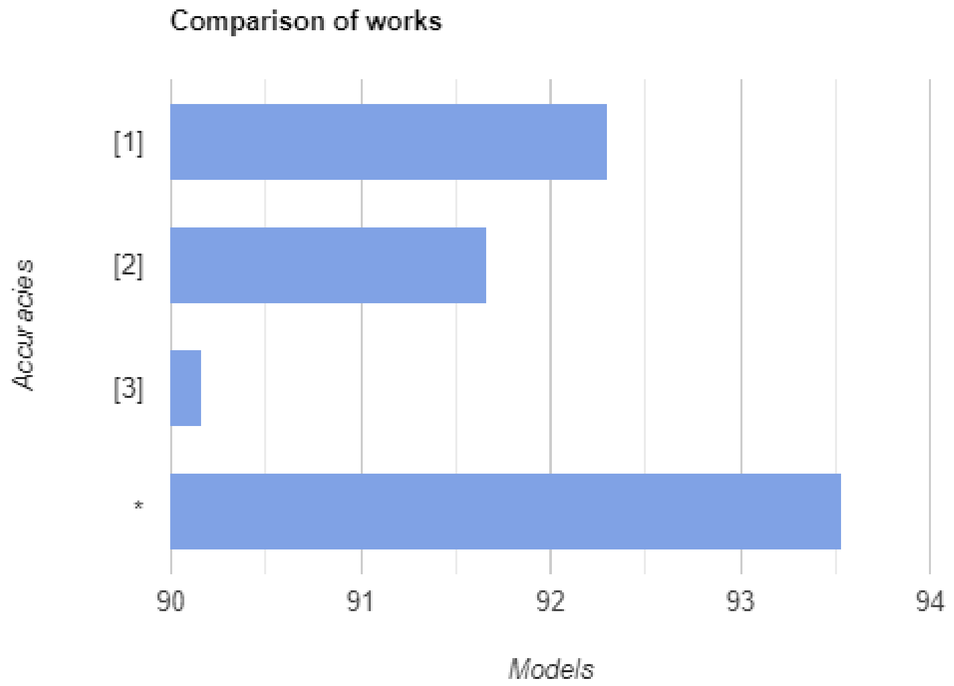
	Method	Accuracy	TP	FP	TN	FN	Precision	Recall	ROC_AUC	PR_AUC	F1_Score	F2_Score
0	Logistic Regression	0.935484	12.0	0.0	17.0	2.0	1.0	0.857143	0.924370	0.941700	0.923077	0.882353
1	Random Forest Classifier	0.870968	10.0	0.0	17.0	4.0	1.0	0.714286	0.951681	0.952597	0.833333	0.757576
2	Gaussian Naive Bayes	0.935484	12.0	0.0	17.0	2.0	1.0	0.857143	0.932773	0.946408	0.923077	0.882353

Accuracy Comparison of Models





Conclusion



Logistic Regression Model in our work beats the performances of all the other models in the worked we refered.

THANK YOU

Deeksha Khera ~ 1906171
Sritishna Sarangi ~ 1906217
Abhinav Deep ~ 1906452