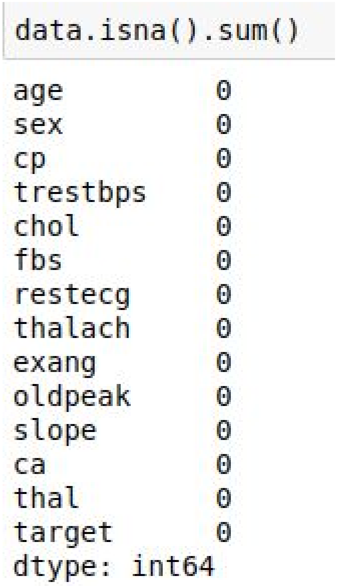
Report on
Heart Disease Dataset

Abstract

In the given dataset, Different types of classifiers were performed i.e SVM, Decision Tree, Random Forest, Adaboosting, and Gradient Boosting. Among them, Gradient Boosting Performed well with accuracy 85%.

1. Data cleaning

There was no NaN and missing value in the dataset.



```
data.isna().sum()
```

```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

2. Data analysis and visualization

The total number of data in the dataset was 302 and 242 data was used for training and 61 was used for testing.

The number of training and testing dataset tabulated below:

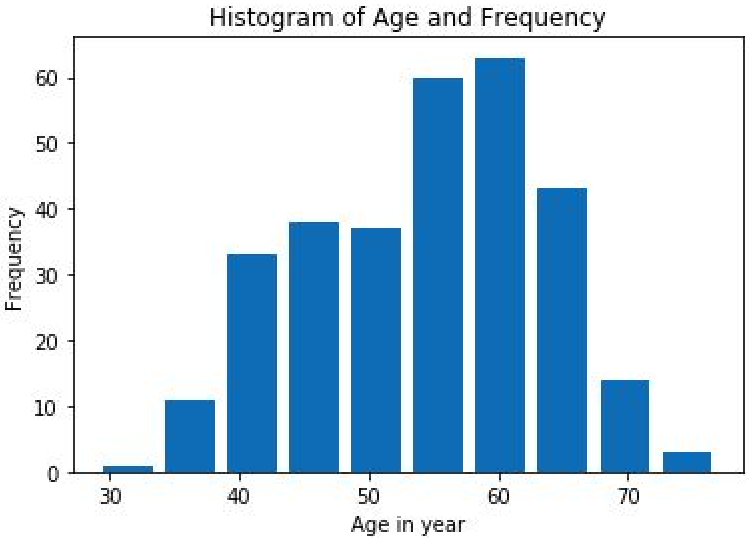
	Number of data
Training	242
Testing	61

There were 13 attributes in the dataset whereas 9 numerical type attributes and 4 categorical type attributes.

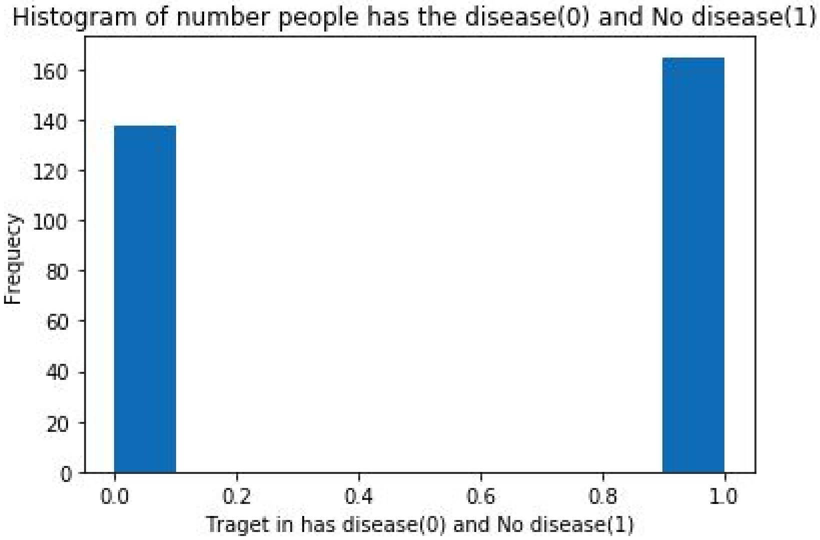
The numerical type attributes and categorical type attributes tabulated below**:**

Numerical TypeAttributes	**Categorical Type Attributes**
'age','sex', 'trestbps', 'chol','fbs', 'restecg', 'thalach','oldpeak', 'ca'	'cp','exang','slope','thal'

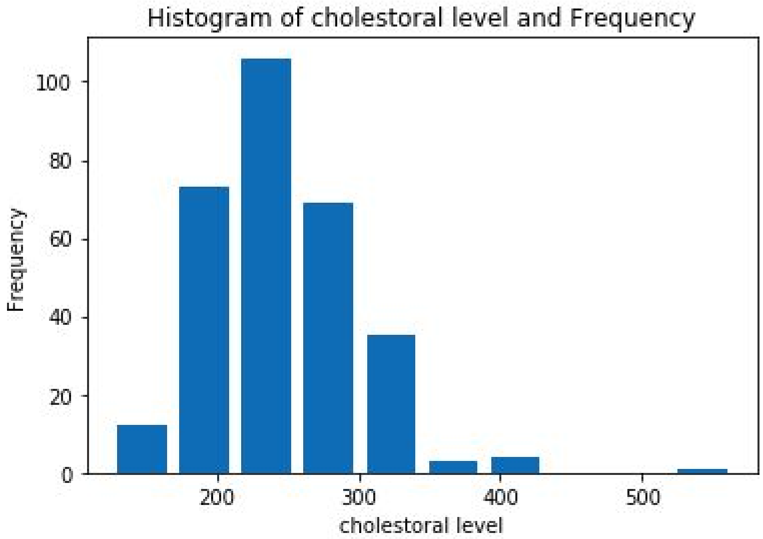
In this dataset, the maximum age was 77, the minimum age was 29, and the maximum number of data was between 55-60 age groups.



In this dataset, 138 people had a disease and 165 people had no disease. That means a higher number of people had no disease.



In this dataset, the maximum number of people had cholesterol levels between 200-300.

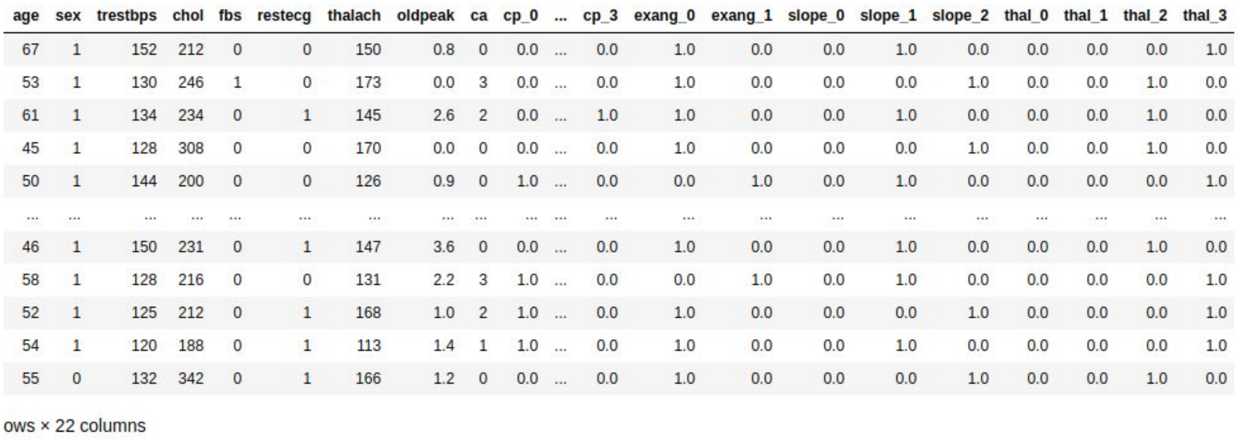


3. Feature extraction and preprocessing

The train dataset was split into x_train and y_train and test dataset was split into x_test and y_test.

Initially, there were 13 features in the dataset. Four categorical columns were encoded using OneHotEncoder.After encoding, the total number of columns became 22.OneHotEncoding applied in both training and testing datasets.

Final dataset looked like this



age	sex	trestbps	chol	fbs	restecg	thalach	oldpeak	ca	cp_0	...	cp_3	exang_0	exang_1	slope_0	slope_1	slope_2	thal_0	thal_1	thal_2	thal_3
67	1	152	212	0	0	150	0.8	0	0.0	...	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
53	1	130	246	1	0	173	0.0	3	0.0	...	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0
61	1	134	234	0	1	145	2.6	2	0.0	...	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
45	1	128	308	0	0	170	0.0	0	0.0	...	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0
50	1	144	200	0	0	126	0.9	0	1.0	...	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
46	1	150	231	0	1	147	3.6	0	0.0	...	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
58	1	128	216	0	0	131	2.2	3	1.0	...	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
52	1	125	212	0	1	168	1.0	2	1.0	...	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
54	1	120	188	0	1	113	1.4	1	1.0	...	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
55	0	132	342	0	1	166	1.2	0	0.0	...	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0

ows × 22 columns

4.Grid search

Grid-search is the process of scanning the data to configure optimal parameters for a given model.

The different of classifiers were performed in given dataset they were:

- Support vector machine
- Decision Tree
- Random Forest
- Adaboosting
- Gradient Boosting

The different types of Classifier, Hyperparameter, Best Hyperparameter, and Best Score tabulated below**:**

Classifier	**Hyperparameter**	**Best Hyperparameter**	**Best Score**
SVM	kernel:('linear', 'rbf') C:(1, 10,20)	Kernel:linear C:20	0.8264
Decision Tree	max_depth': (3, 5, 7, 9, 11, 13), min_samples_split: (2, 4, 6, 8, 10)	max_depth: 13 min_samples_split: 10	0.7784
Random Forest	max_depth: (3, 5, 7, 9, 11, 13) min_samples_split: (2, 4, 6, 8, 10) n_estimators:(10,50,100)	max_depth: 3 min_samples_split: 2 n_estimators: 50	0.8390
Adaboosting	learning_rate:(0.001,0.1) n_estimator':(10,50,100)	learning_rate: 0.1 n_estimators: 50	0.8470
Gradient Boosting	max_depth: (3, 5, 7, 9, 11, 13) min_samples_split: (2, 4, 6, 8, 10) n_estimators:(10,50,100)	max_depth: 3 min_samples_split: 2 n_estimators: 10	0.7980

5.Model Evaluation and Comparison

The different types of Classifier their Accuracy and F1-score tabulated below:

Classifier	Accuracy
SVM	0.84
Decision Tree	0.70

Random Forest	0.80
Adaboosting	0.84
Gradient Boosting	0.85

7.Conclusion

Among all the classifiers, the Gradient Boosting performed well with accuracy 85%**.** Hence Gradient Boosting was chosen as the best classifier.

8.code link

https://github.com/rachanakafle/Heart-Disease-Prediction/blob/master/Heart%20Disease%20Dataset.ipynb