CBD-2214 Big Data Fundamental Data Storage Networking

Heart Disease Detection with classification ML Algorithms

PROJECT REPORT

Submitted to:
Prof. Aruna Dorai

Submitted by:

Ananya Singh
Harjot Parhar
Mohammad Shaiq
Tony Joseph Sebastian

June 13, 2021

Table of Contents

Introduction

Heart disease is a severe threat to humanity. We strive to detect cardiac disease by inputting a set of 13 numeric features that provide a numerical value for each feature; hence we aim to save lives by identifying early heart conditions. Among the algorithms we used were K-Nearest Neighbors Classifier, Decision Tree Classifier, Random Forest Classifier, and Naive Bayes, Logistic Regression. Regression. The project's main aim is to evaluate the features that have the most relationship with heart disease, compare the accuracy of algorithms, research trends, and determine whether a patient has heart disease.

Dataset and Features

Our dataset is from Kaggle (Kaggle, 2018). We did feature scaling on our dataset and our dataset to change 1s to 0s in the target column and vice versa. The value 1 indicates the presence of heart disease, and 0 indicates the absence of cardiac disease. The following are the column names of the dataset.

1. Age.
2. Sex: 1 = male 0 = female
3. Chest-pain (cp)type displays the type of chest pain experienced by the person using the following format: 1 = typical angina 2 = typical angina 3 = non — anginal pain 4 = asymptotic
4. Resting Blood Pressure (trestbps): displays the resting blood pressure value of an individual in mmHg (unit)
5. Serum Cholesterol(chol): displays the serum cholesterol in mg/dl (unit)
6. Fasting Blood Sugar(fbs): compares the fasting blood sugar value of an individual with 120mg/dl. If fasting blood sugar > 120mg/dl then: 1 (true) else: 0 (false)
7. Resting ECG((restecg): displays resting electrocardiographic results 0 = normal 1 = having ST-T wave abnormality 2 = left ventricular hypertrophy
8. Max heart rate achieved(thalach): displays the max heart rate achieved by an individual.
9. Exercise-induced angina (exang) (1 = yes; 0 = no)
10. ST depression induced by exercise relative to rest (Old peak): displays the value of an integer or floats.
11. Peak exercise ST segment (Slope): 1 = upsloping 2 = flat 3 = down sloping
12. Number of major vessels (0–3) colored by fluoroscopy(ca): displays the value as integer or float.
13. Thal: displays the thalassemia: 3 = normal 6 = fixed defect 7 = reversible defect
14. Target: Displays whether the individual is suffering from heart disease or not : 0 = absence 1, 2, 3, 4 = present.

Exploratory Data Analysis

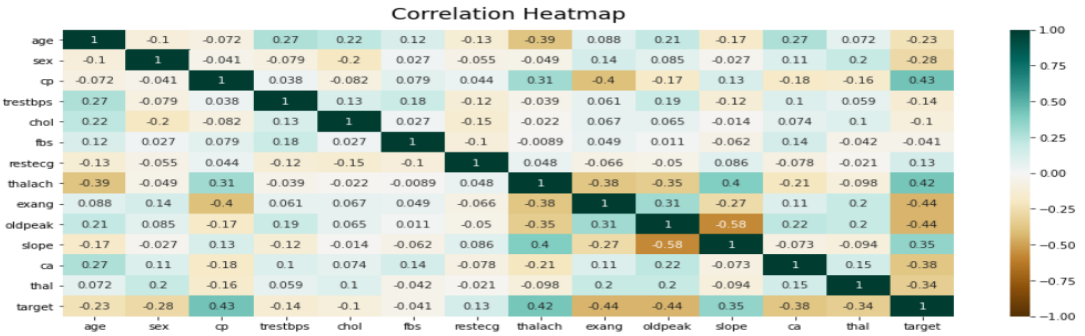
Data Description

```
df.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	1025.000000	1025.000000	1025.000000	1025.000000	1025.00000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000
mean	54.434146	0.695610	0.942439	131.611707	246.00000	0.149268	0.529756	149.114146	0.336585	1.071512	1.385366	0.754146	2.323902	0.513171
std	9.072290	0.460373	1.029641	17.516718	51.59251	0.356527	0.527878	23.005724	0.472772	1.175053	0.617755	1.030798	0.620660	0.500070
min	29.000000	0.000000	0.000000	94.000000	126.00000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	48.000000	0.000000	0.000000	120.000000	211.00000	0.000000	0.000000	132.000000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	56.000000	1.000000	1.000000	130.000000	240.00000	0.000000	1.000000	152.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	275.00000	0.000000	1.000000	166.000000	1.000000	1.800000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.00000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

Correlation

Correlation is the relationship between different features.

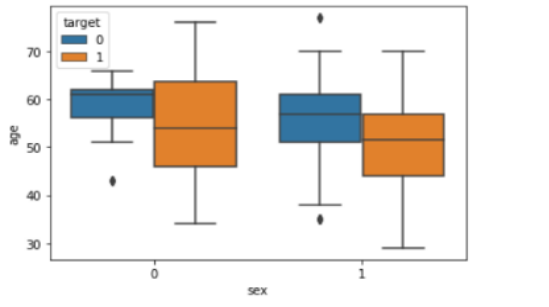


We can see a **positive correlation** between chest pain and heart disease. Since most people who have a cardiac arrest have chest pain, this makes sense. As we exercise more, we see a **negative correlation** between angina and the target. Therefore, the blood flow to the heart is increased, and cardiac arrests are reduced.

Box Plot and Violine Plot

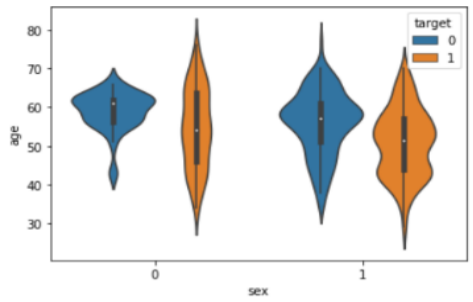
```
#Box Plot
sns.boxplot(x='sex',y='age',hue='target',data=df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7efef6e89850>
```




```
#Violine plot
sns.violinplot(x='sex',y='age',hue='target',data=df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7efef6db7310>
```



Based on the above graphs, it is evident that women around the age of 56-57 have heart disease, and men around the age of 52-53 have heart disease.

Feature Scaling

The 13 features normalized by using Standard Scaler.

```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
x_train = sc.fit_transform(x_train)
x_test = sc.transform(x_test)
```

Methods

1. Logistic regression

Logistic regression is model works based on probability values. The logistic regression model can be derived from linear regression model by applying sigmoid function. y = b0 + b1*x - Linear model, p = 1/(1+e^-y) - Sigmoid function. Applying p to y ln(p/1-p) = b0 + b1*x will result in Logistic regression model.

```
from sklearn.linear_model import LogisticRegression
log_model = LogisticRegression(random_state = 0)
log_model.fit(xlog_train, ylog_train)
```

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=0, solver='lbfgs', tol=0.0001, verbose=0,
                   warm_start=False)
```

2. K Nearest Neighbors

KNN regression is a non-parametric method that, intuitively, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighborhood.

```
from sklearn.neighbors import KNeighborsClassifier
clf = neighbors.KNeighborsClassifier()
clf.fit(x_train,y_train)
```

```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                     metric_params=None, n_jobs=None, n_neighbors=5, p=2,
                     weights='uniform')
```

3. Decision Tree Classifier

Decision Tree Classifiers are a supervised learning algorithm and for classification and regression.

```
from sklearn.tree import DecisionTreeClassifier
dm = DecisionTreeClassifier()
dm.fit(x_train,y_train)
```

```
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
                       max_depth=None, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=None, splitter='best')
```

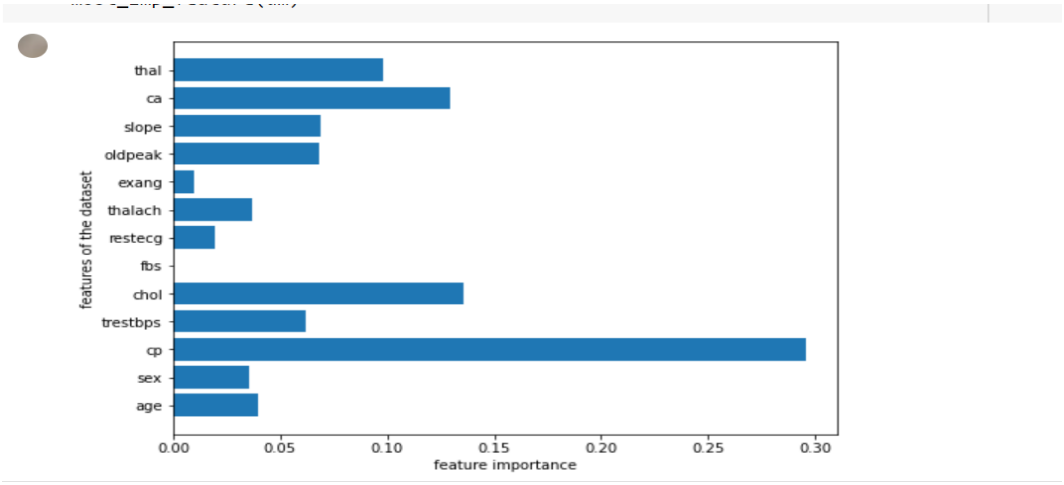
4. Random Forest Classifier

Random forests are a supervised learning algorithm and for classification and regression.

```
from sklearn.ensemble import RandomForestClassifier
random_forest = RandomForestClassifier(max_depth=5)
random_forest.fit(X_train, y_train)
```

```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=5, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=100,
                       n_jobs=None, oob_score=False, random_state=None,
                       verbose=0, warm_start=False)
```

Feature Importance



According to the decision classification model, **chest pain** is the most crucial factor.

Accuracy/Results

Confusion Matrix

Confusion matrix is used for determining the accuracy of the model. If the accuracy score is above 70%, the model is considered as good. Accuracy = (TP + TN)/(TP + TN + FP + FN).

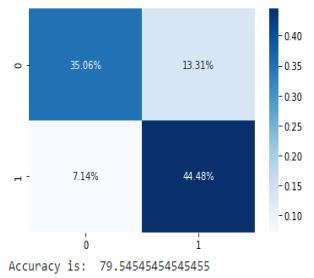


Figure 1. Logistic regression

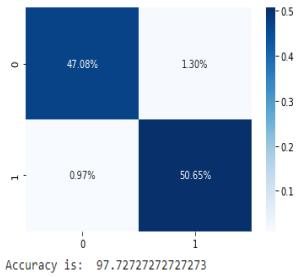


Figure 2. Decision Tree classifier

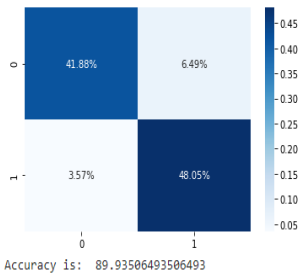


Figure 3. Random Forest Classifier

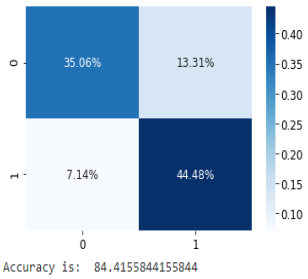


Figure 4. KNN Classifier

Classification Report

	precision	recall	f1-score	support
0	0.83	0.72	0.77	149
1	0.77	0.86	0.81	159
accuracy			0.80	308
macro avg	0.80	0.79	0.79	308
weighted avg	0.80	0.80	0.79	308

Figure 1. Logistic Regression

	precision	recall	f1-score	support
0	0.98	0.97	0.98	149
1	0.97	0.98	0.98	159
accuracy			0.98	308
macro avg	0.98	0.98	0.98	308
weighted avg	0.98	0.98	0.98	308

Figure 2. Decision Tree Classifier

	precision	recall	f1-score	support
0	0.92	0.84	0.88	149
1	0.86	0.93	0.89	159
accuracy			0.89	308
macro avg	0.89	0.88	0.89	308
weighted avg	0.89	0.89	0.89	308

Figure 3. Random Forest Classifier

	precision	recall	f1-score	support
0	0.83	0.85	0.84	149
1	0.85	0.84	0.85	159
accuracy			0.84	308
macro avg	0.84	0.84	0.84	308
weighted avg	0.84	0.84	0.84	308

Figure 4. KNN Classifier

Methods	Accuracy (%)
Logistic Regression	79.5455
Decision Tree classifier	97.7273
KNN Classifier	84.415
Random Forest Classifier	89.9351

Conclusions

- Our tests have examined 13 features, and chest paint has proven to be the most critical feature.
- As a result of our ML algorithms, we can determine whether a patient has heart disease or not, and it gives us high precision in the analysis of the input data.
- We achieved the highest accuracy with our decision tree model, 97%.

References

Kaggle. (2018). *Heart Disease UCI*. Retrieved from Kaggle: https://www.kaggle.com/ronitf/heart-disease-uci