Heart attack prediction

- **Aditya Sumbaraju**
  DSC 680 Applied Data Science
  Dr. Brett Werner
  Feb 27, 2022

Table of Contents

Topic

Heart attack prediction

Background

- Heart attack is the number 1 cause of death compared to other diseases globally, taking an approximate estimation of 18 million lives each year, accounting for 31% of worldwide deaths. Heart failure can be prevented by addressing behavioral risk factors such as unhealthy diet, tobacco use, obesity (overweight concerns), physical inactivity, and heavy use of alcohol using population-wide strategies. If these risk factors are coupled with early treatment, it dramatically impacts its prognosis.
- It is undoubtedly difficult to identify high-risk patients because of several multi-factorial contributory risk factors such as high B.P., diabetes, and high cholesterol. Here comes the need for machine learning and data mining to study, evaluate and predict the disease beforehand.
- Medical researchers, doctors, and scientists are still contributing to machine learning (ML) techniques to develop interactive GUIs to predict the early detection of this disease. This is because of their superiority in classification compared to other traditional statistical approaches and pattern recognition. In this use case, I will be addressing below research questions.

➢ ***Can physicians will be able to predict Cardiovascular disease with the help of patient demographics***

➢ ***Does this prediction reduce the risk and prevent heart attack disease. Is early detection of heart attack possible?***

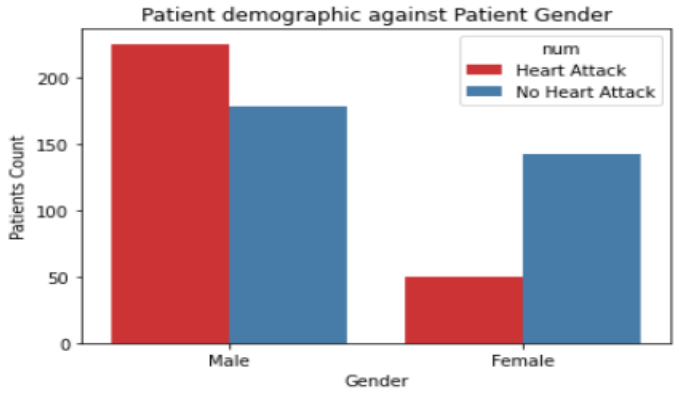
About Data

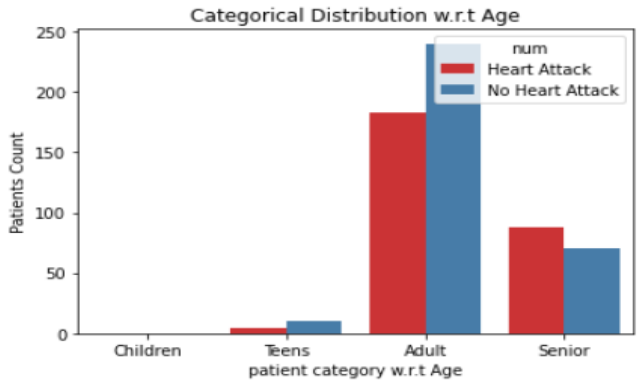
The dataset was gathered from the Machine Learning Repository from the Center for Machine Learning and Intelligent Systems at the University of California, Irvine. This directory contains four datasets concerning heart failure diagnosis. Features are numeric-valued. The databases have 76 raw attributes; only 14 of them are used.

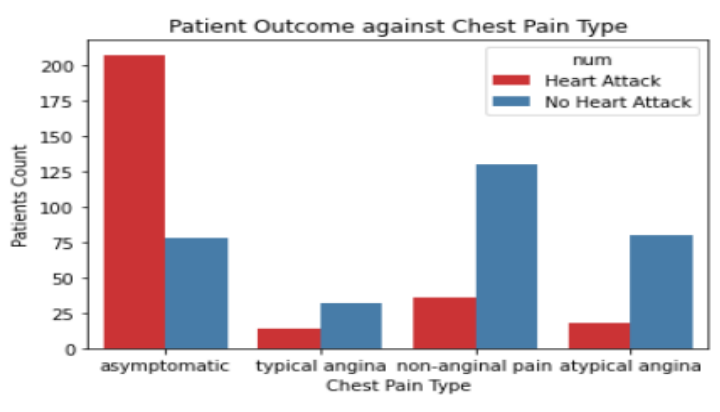
Sl.No	Attribute Name	Attribute Description	Attribute Values
1.	AGE	Age in years	25-75 years
2.	SEX	Male/Female	value 1: Male; value 0 : Female
3.	CHESTPAIN	Chest Pain Type	value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic
4.	RESTBP	resting blood pressure	90-192
5.	CHOLESTEROL	serum cholestoral in mg/dl	160-410
6.	BLOODSUGAR	fasting blood sugar > 120 mg/dl	value 1: > 120 mg/dl; value 0: < 120 mg/dl
7.	ECG	resting electrocardiographic results	value 0: normal; value 1: 1 having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy
8.	MAXHEARTRATE	maximum heart rate achieved	71-202
9.	ANGINA	exercise induced angina	value 1: yes; value 0: no
10.	OLDPEAK	ST depression induced by exercise relative to rest	Continuous
11.	STSLOPE	the slope of the peak exercise ST segment	value 1: unsloping; value 2: flat; value 3: downsloping)
12.	VESSELS	number of major vessels (0-3) colored by flourosopy	value 0 – 3
13.	THAL:	thalac	value 3: normal; value 6: fixed defect; value 7: reversible defect

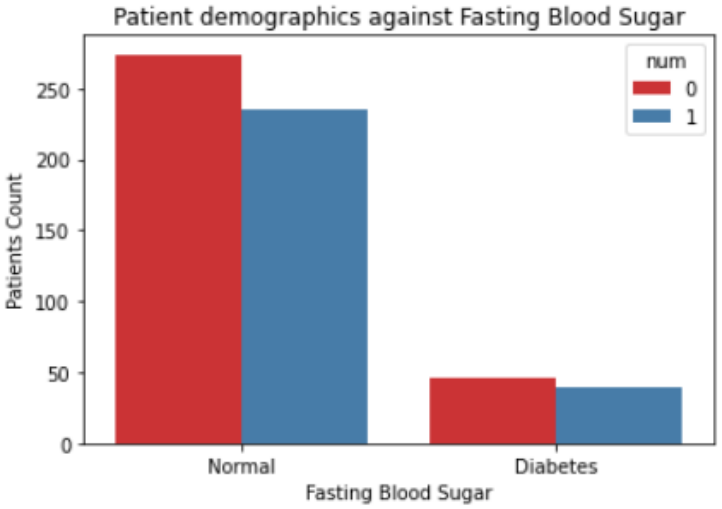
```
In [13]: hap_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 929 entries, 0 to 928
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       929 non-null    float64
 1   sex       929 non-null    float64
 2   cp        929 non-null    float64
 3   trestbps  871 non-null    float64
 4   chol      922 non-null    float64
 5   fbs       847 non-null    float64
 6   restecg   928 non-null    float64
 7   thalach   875 non-null    float64
 8   exang     875 non-null    float64
 9   oldpeak   867 non-null    float64
 10  slope     810 non-null    float64
 11  ca        605 non-null    float64
 12  thal      707 non-null    float64
 13  num       929 non-null    int64
dtypes: float64(13), int64(1)
memory usage: 101.7 KB
```

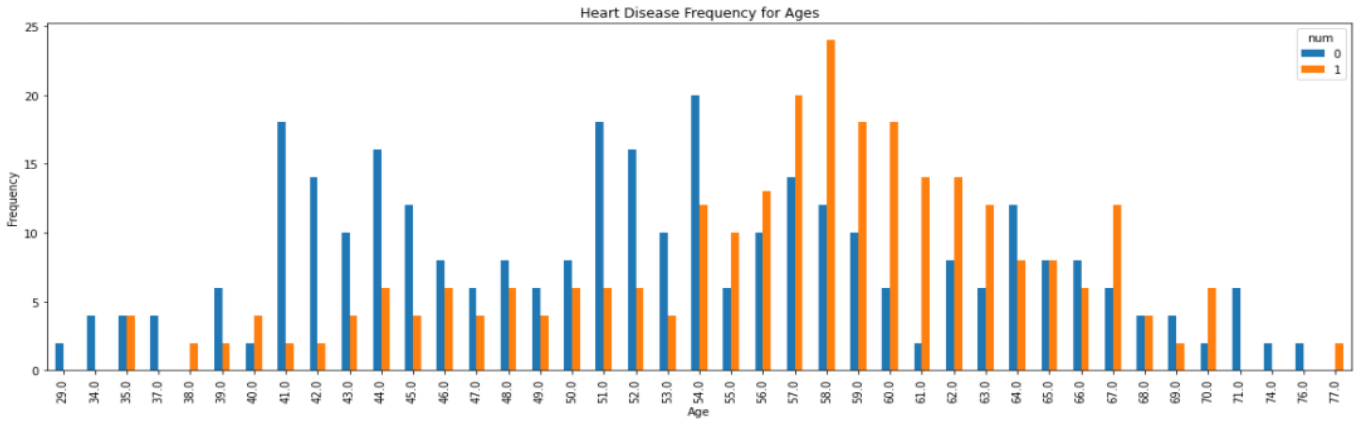
EDA


Patient demographic against Patient Gender
num
Heart Attack
No Heart Attack
Patients Count
0
50
100
150
200
Male
Female
Gender


Categorical Distribution w.r.t Age
num
Heart Attack
No Heart Attack
Patients Count
0
50
100
150
200
250
Children
Teens
Adult
Senior
patient category w.r.t Age


Patient Outcome against Chest Pain Type
num
Heart Attack
No Heart Attack
Patients Count
0
25
50
75
100
125
150
175
200
asymptomatic
typical angina
non-anginal pain
atypical angina
Chest Pain Type


Patient demographics against Fasting Blood Sugar
num
0
1
Patients Count
0
50
100
150
200
250
Normal
Diabetes
Fasting Blood Sugar

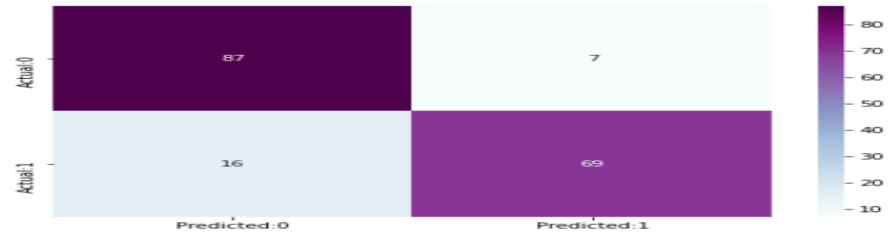

Heart Disease Frequency for Ages
num
0
1
Frequency
0
5
10
15
20
25
Age

Logistic Regression:

Predictive Model and Evaluation

BELLEVUE UNIVERSITY

Logistic Regression



The details for confusion matrix is =

	precision	recall	f1-score	support
0	0.84	0.93	0.88	94
1	0.91	0.81	0.86	85
accuracy			0.87	179
macro avg	0.88	0.87	0.87	179
weighted avg	0.87	0.87	0.87	179

Support Vector Classification

The details for confusion matrix is =

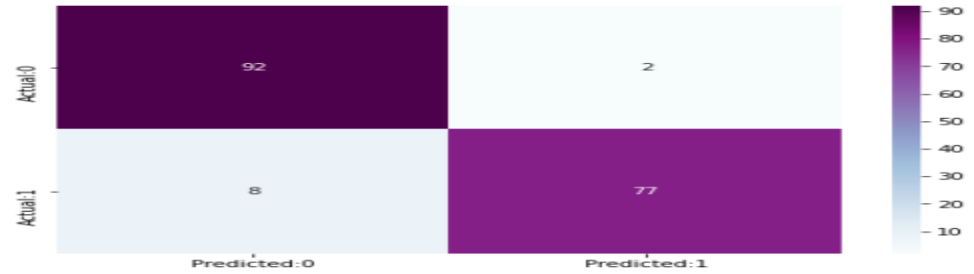
	precision	recall	f1-score	support
0	0.85	0.91	0.88	94
1	0.90	0.82	0.86	85
accuracy			0.87	179
macro avg	0.87	0.87	0.87	179
weighted avg	0.87	0.87	0.87	179

Decision Tree

The details for confusion matrix is =

	precision	recall	f1-score	support
0	0.94	0.94	0.94	94
1	0.93	0.93	0.93	85
accuracy			0.93	179
macro avg	0.93	0.93	0.93	179
weighted avg	0.93	0.93	0.93	179

Radom Forest

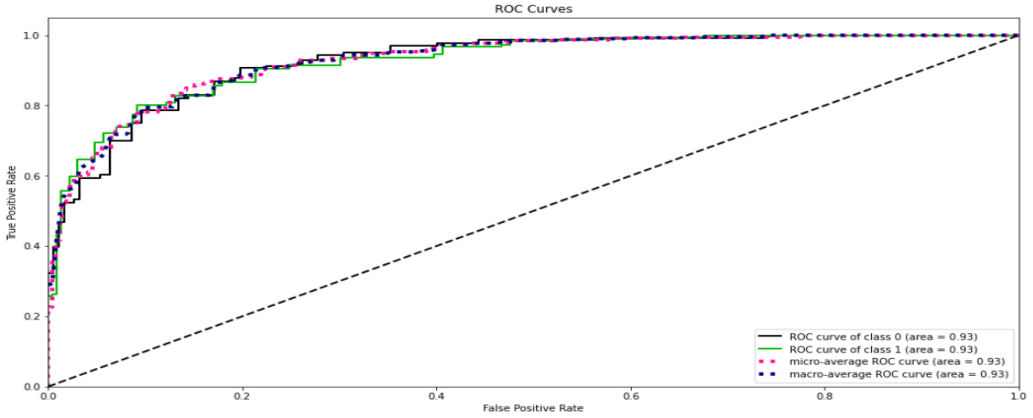


The details for confusion matrix is =

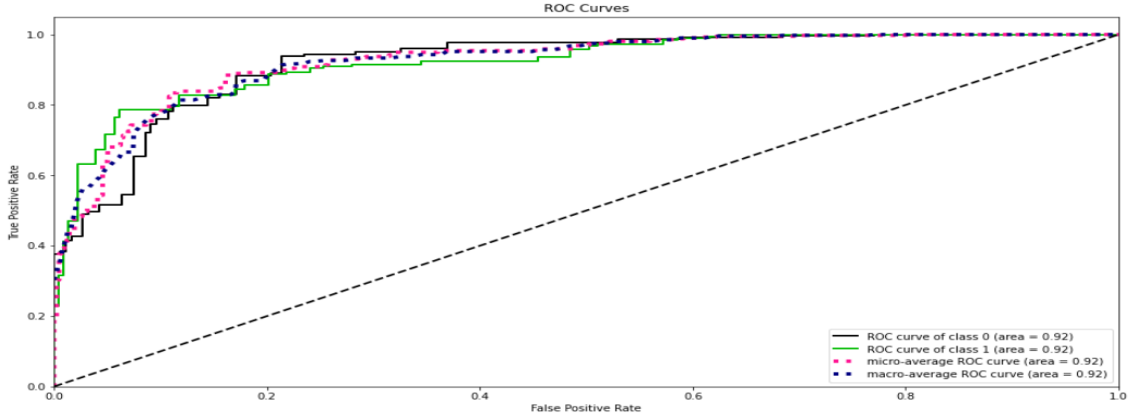
	precision	recall	f1-score	support
0	0.92	0.98	0.95	94
1	0.97	0.91	0.94	85
accuracy			0.94	179
macro avg	0.95	0.94	0.94	179
weighted avg	0.95	0.94	0.94	179

Model Performance

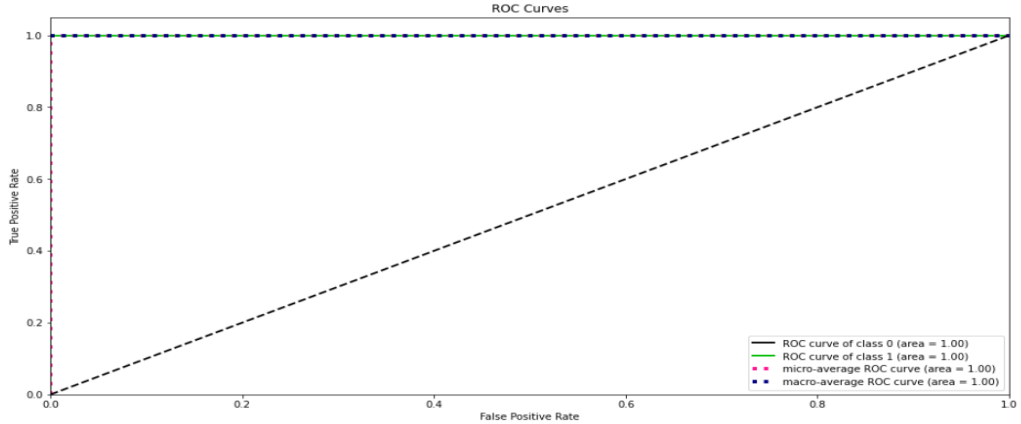
Logistic Regression:



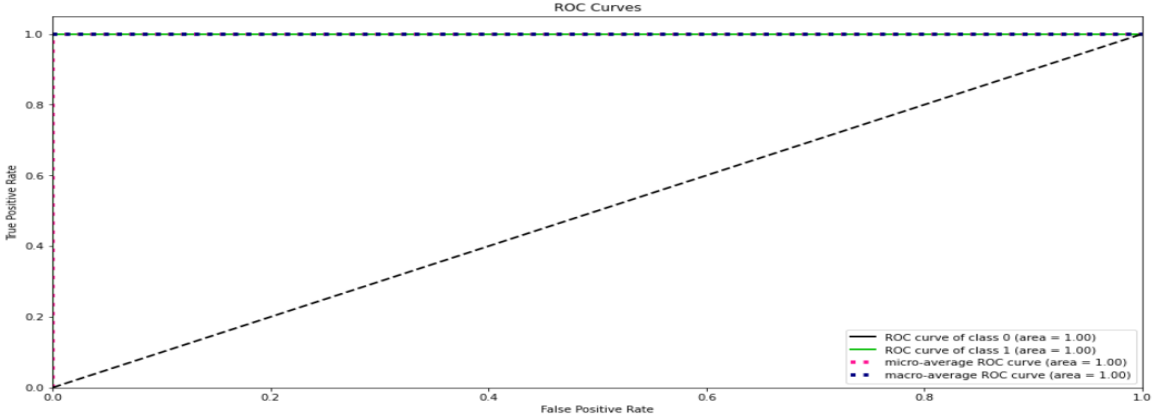
SVC:



Random Forest



Decision Tree



Model results

	MODEL	ACCURACY_SCORE
1	Random Forest Classifier	0.97
2	Decision Tree	0.91
3	Logistic Regression	0.89
4	SVM Classifier	0.88

Conclusion

I have implemented Logistic Regression, Support Vector Classifier, decision tree, and Random Forest to predict a patient's heart attack using the patient demographics collected from several countries. Data pre-processing is done by removing all the null records and duplicate records. In the classification stage, a Logistic Regression, Support Vector Classifier, decision tree, Random Forest are used to label the data as heart disease present or not. The results of the classification experiment, performed over data sets obtained from 929 patients, shows that the Random Forest classifier has achieved better accuracy when compared to Logistic Regression, Support Vector Classifier, and decision tree.

References

- Frye, R. L., Simari, R. D., Gersh, B. J., Burnett, J. C., Brumm, S., Myerle, K., Jaffe, A. S., Holmes, D. R., Lerman, A., & Terzic, A. (2009, November 24). *Ethical issues in cardiovascular research involving humans*. Circulation. Retrieved February 10, 2022, from https://www.ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.107.752766#d3e854
- Shah, D., Patel, S., & Bharti, S. K. (2020, October 16). *Heart disease prediction using machine learning techniques - S.N. computer science*. SpringerLink. Retrieved February 10, 2022, from https://link.springer.com/article/10.1007/s42979-020-00365-y
- Foster, N. (2018, July 23). *The future of heart attack prediction*. Mended Hearts. Retrieved February 10, 2022, from https://mendedhearts.org/story/the-future-of-heart-attack-prediction/
- Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008
- C. Thirumalai, A. Duba and R. Reddy, "Decision making system using machine learning and Pearson for heart attack," 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), 2017, pp. 206-210, doi: 10.1109/ICECA.2017.8212797.
- S. Manikandan, "Heart attack prediction system," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017, pp. 817-820, doi: 10.1109/ICECDS.2017.8389552.
- Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.