Heart Attack Prediction

Aditya Sumbaraju

Bellevue University

DSC680: Applied Data Science

Dr. Brett Werner

February 13, 2022

https://github.com/adityasumbaraju/aditya_portfolio/tree/main/HeartAttackPrediction

Business Problem

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 challenging 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.

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?

Data Explanation

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 data was collected from below mentioned four locations:

- University Hospital, Zurich, Switzerland (Switzerland.data)
- Cleveland Clinic Foundation (Cleveland.data)
- Hungarian Institute of Cardiology, Budapest (Hungarian.data)
- V.A. Medical Center, Long Beach, CA (long-beach-va.data)

All four database has the same format. The databases have 76 raw attributes; only 14 of them are used.

Metadata:

1. age – Age in Years
2. sex – sex(1-male;0=female)
3. cp- (chest pain type)
    - Value 1: typical angina
    - Value 2: atypical angina

- Value 3: non-anginal pain
- Value 4: asymptomatic

4. trestbps - resting blood pressure (measured in **mm Hg** on admission to the hospital)
5. chol - serum cholestoral in mg/dl
6. fbs - fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7. restecg - resting electrocardiographic results
    - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
    - Value 1: having ST-T wave abnormality (T wave inversions and S.T. elevation or depression of > 0.05 mV)
    - Value 0: normal
8. exang - exercise-induced angina (1 = yes; 0 = no)
9. thalach - maximum heart rate achieved
10. slope - the slope of the peak exercise S.T. segment
    - Value 1: upsloping
    - Value 2: flat
    - Value 3: downsloping
11. ca - number of major vessels (0-3) colored by fluoroscopy
12. oldpeak - S.T. depression induced by exercise relative to rest
13. thal - Thalium Heart Scan - thal: 3 = normal; 6 = fixed defect; 7 = reversible defect
14. num - the predicted attribute

Dataset details:

```
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
```

Sample Data:

```
In [17]: hap_df.head()
```

Out[17]:

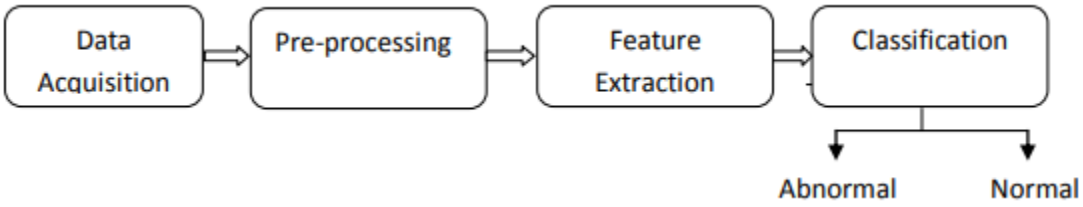
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
0	53.0	1.0	4.0	123.0	282.0	0.0	0.0	95.0	1.0	2.0	2.0	2.0	7.0	1
1	52.0	1.0	4.0	165.0	0.0	NaN	0.0	122.0	1.0	1.0	1.0	NaN	7.0	1
2	60.0	1.0	4.0	132.0	218.0	0.0	1.0	140.0	1.0	1.5	3.0	NaN	NaN	1
3	51.0	1.0	4.0	140.0	0.0	0.0	0.0	60.0	0.0	0.0	2.0	NaN	3.0	1
4	63.0	1.0	4.0	140.0	187.0	0.0	2.0	144.0	1.0	4.0	1.0	2.0	7.0	1

Dataset source URL: https://archive.ics.uci.edu/ml/datasets/heart+disease

Implementation Plan

The implementation plan contains the below steps to accomplish the project

- Business understanding
- Data Preparation
- Exploratory data analysis
- Data prep for Modeling
- Modeling
- Model evaluation



Methods

I have implemented "standardization" on some features using the sklearn library (StandardScalar). I have scaled down features into properties of Standard Normal Distribution where standard deviation = 1 and mean = 0. The takeaway observation is scaling gave a higher performance in classifiers such as Random Forest, Logistic Regression, Decision Tree, Support Vector Classifier. It proves that feature scaling improved the performance of mentioned

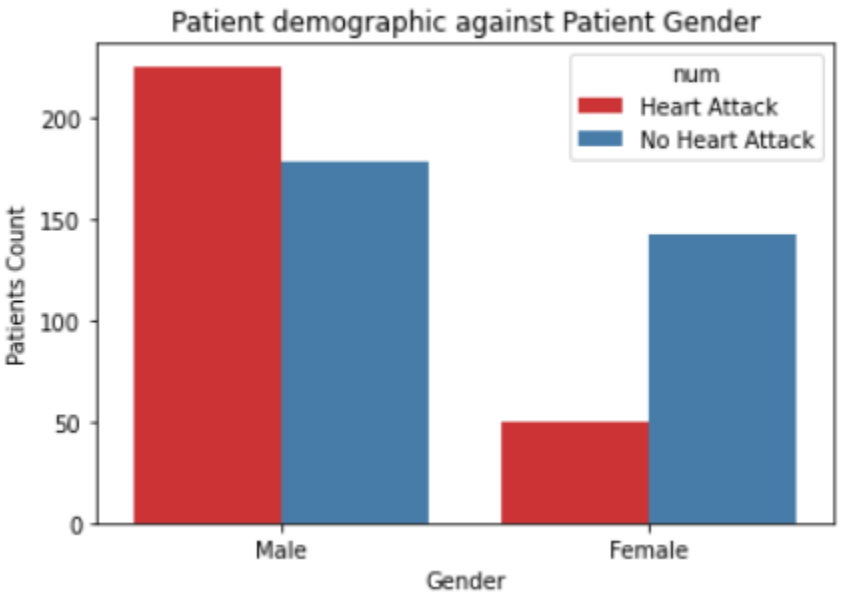
classifiers to predict an accurate model using the CRISP-DM method. Below are the phases I have targeted to achieve a better model.

- **Business understanding**: I would be focusing on case study objectives and requirements from a business perspective. This is a valuable phase to create a preliminary plan and move with the subsequent stages.
- **Data Understanding**: Dataset is already identified to proceed with the case study. All I need to do is identify data anomalies detect interesting subsets from the hypothesis.
- **Data Preparation**: This phase covers the data wrangling scenarios to construct the final dataset from the raw dataset.
- **Modeling and evaluation** involve building and developing various models based on different modeling techniques. We determine the predictive modeling algorithm in this stage and evaluate which models give the best performance. Pending modeling selection and approach, we might need to split the data into training and test sets using the sklearn train_test_split library. Since the project is classification-based (heart attack or no heart attack based on target feature called **num**), I used classification models such as Logistic Regression, Support Vector Classifier, decision tree, Random Forest to predict. After evaluating several operations such as hyper-parameter tuning cross-validation, the highest accuracy was recorded with Random Forest Classifier followed by a decision tree. Check the Model Evaluation section to see how each classifier performed.

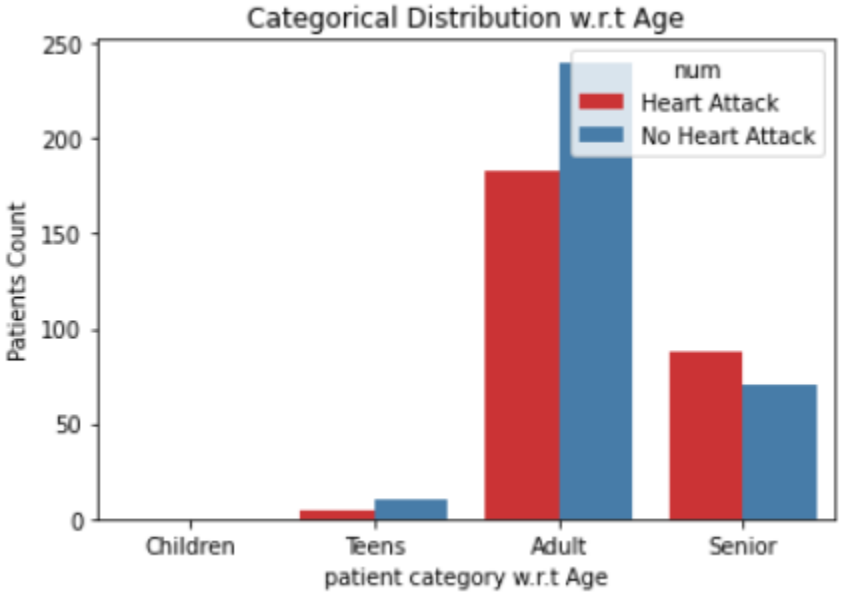
- **Deployment**: Publish final results and conduct a case study retrospect on what went well and required to build the Model better.

Exploratory Data Analysis

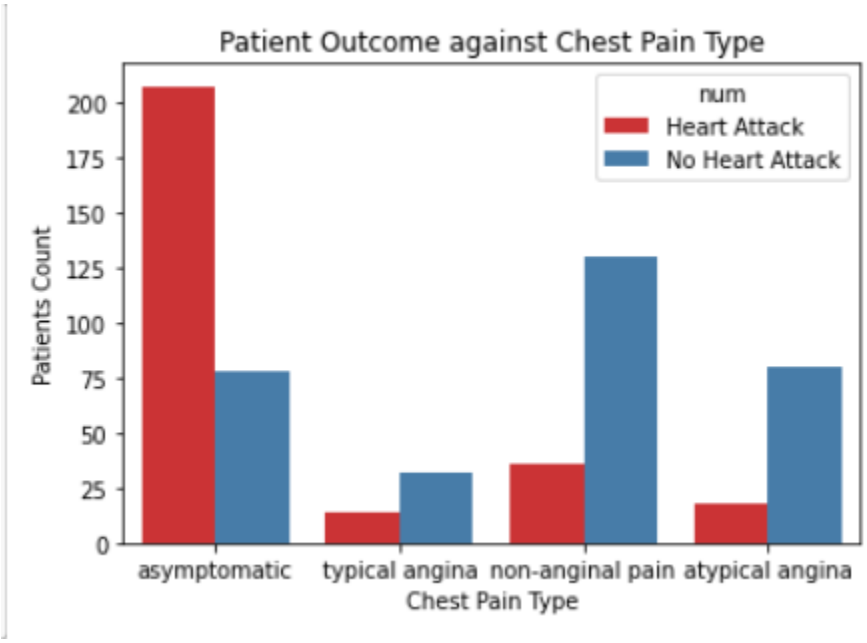
Patient demographic against Patient Gender



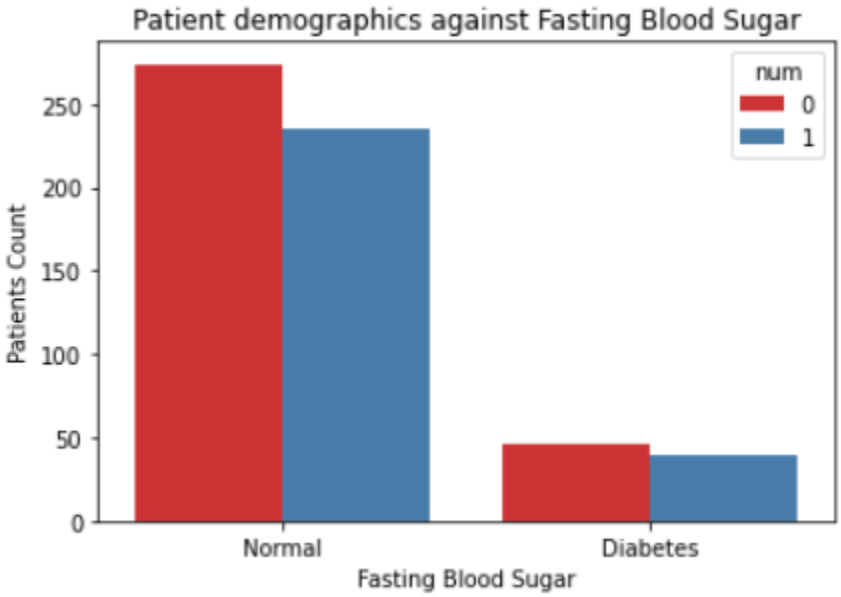
Patient category w.r.t Age



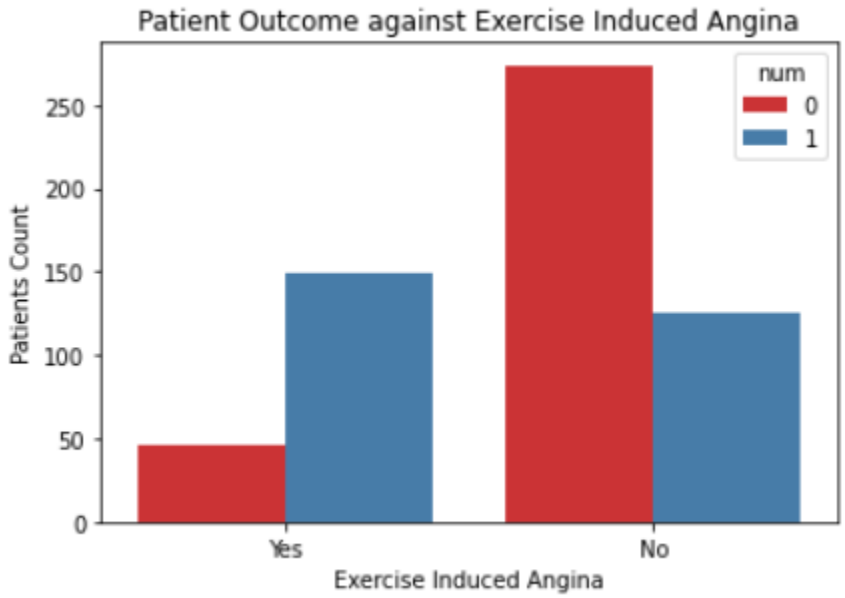
Patient Outcome against Chest Pain Type



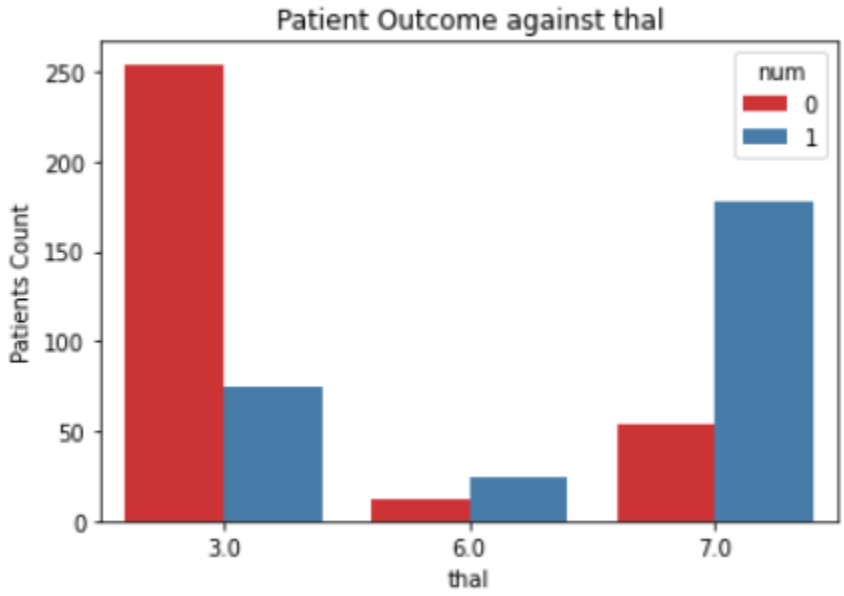
Patient demographics against Fasting Blood Sugar



Patient Outcome against Exercise-Induced Angina

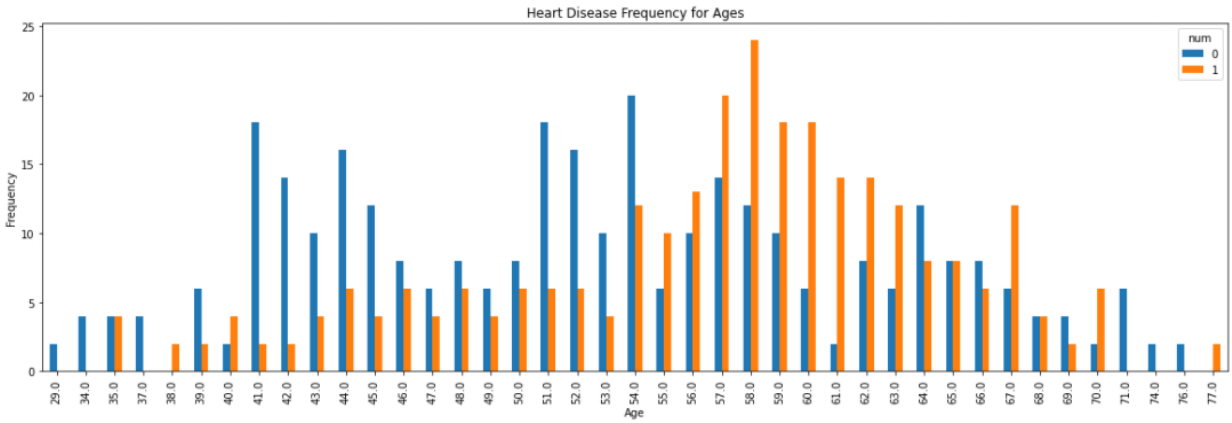


Patient Outcome against thal

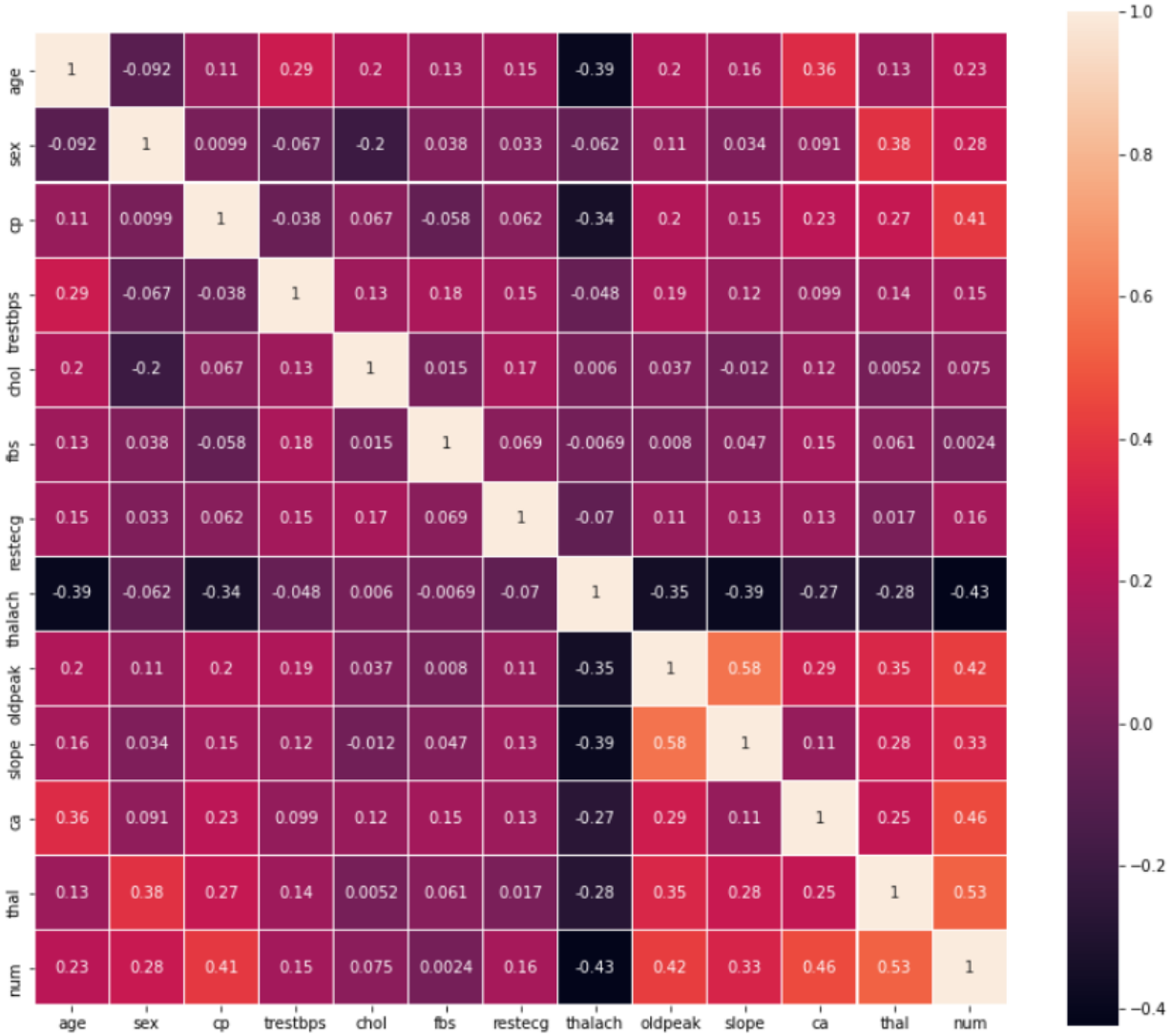


Project 3

Heart Disease Frequency w.r.t Age



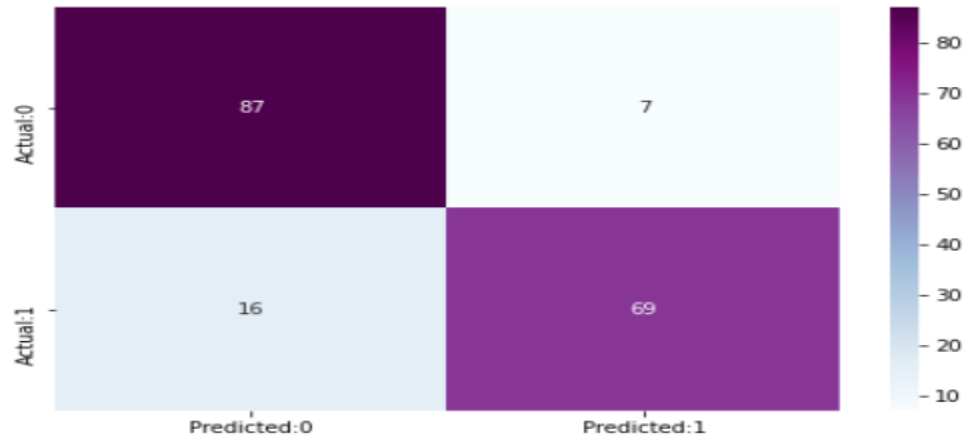
Correlation Matrix:



Modeling Results

Logistic Regression

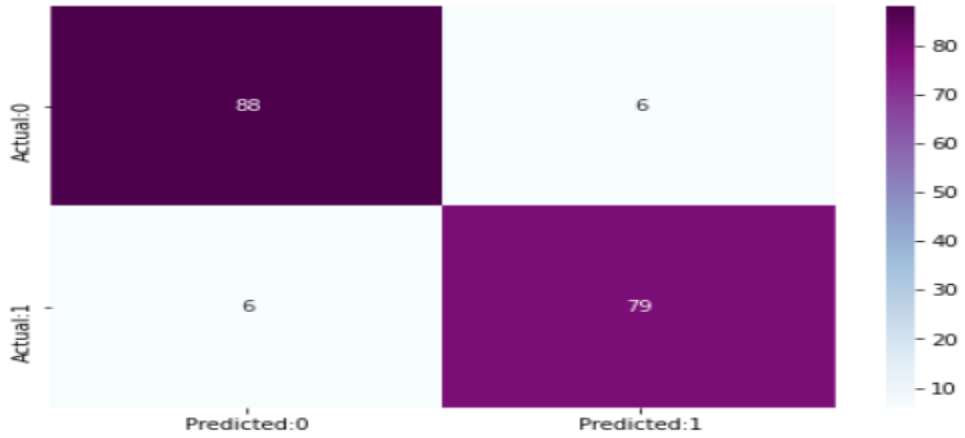
Project 3



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

Decision Tree

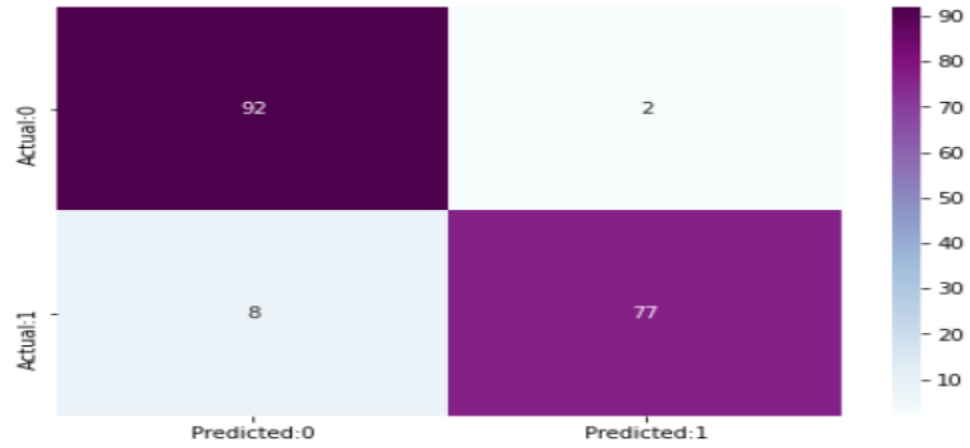


The details for confusion matrix is =

	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

Random Forest

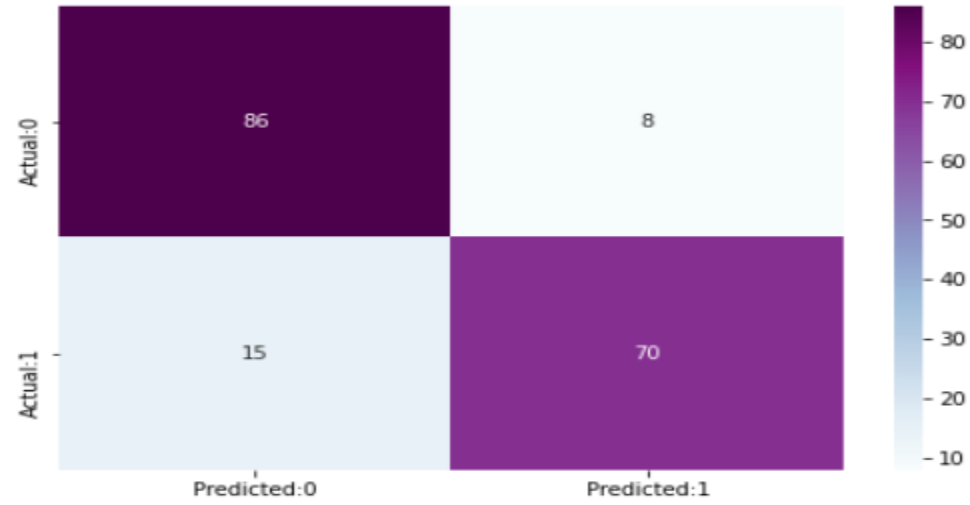
Project 3



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

Support Vector Machine



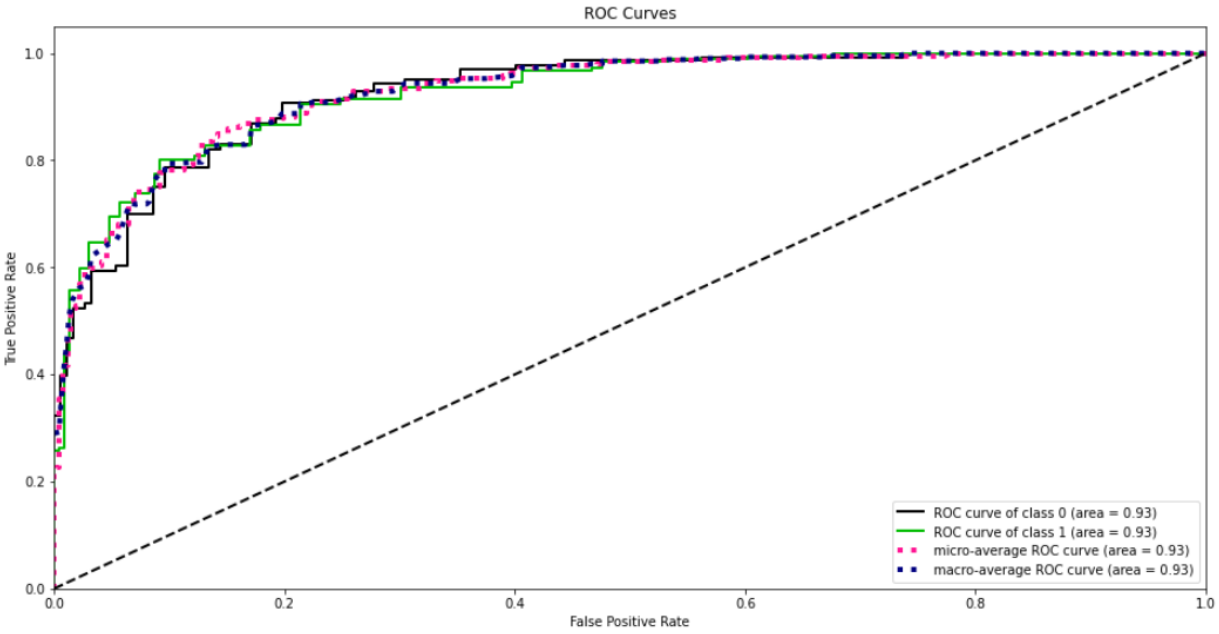
The details for confusion matrix is =

	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

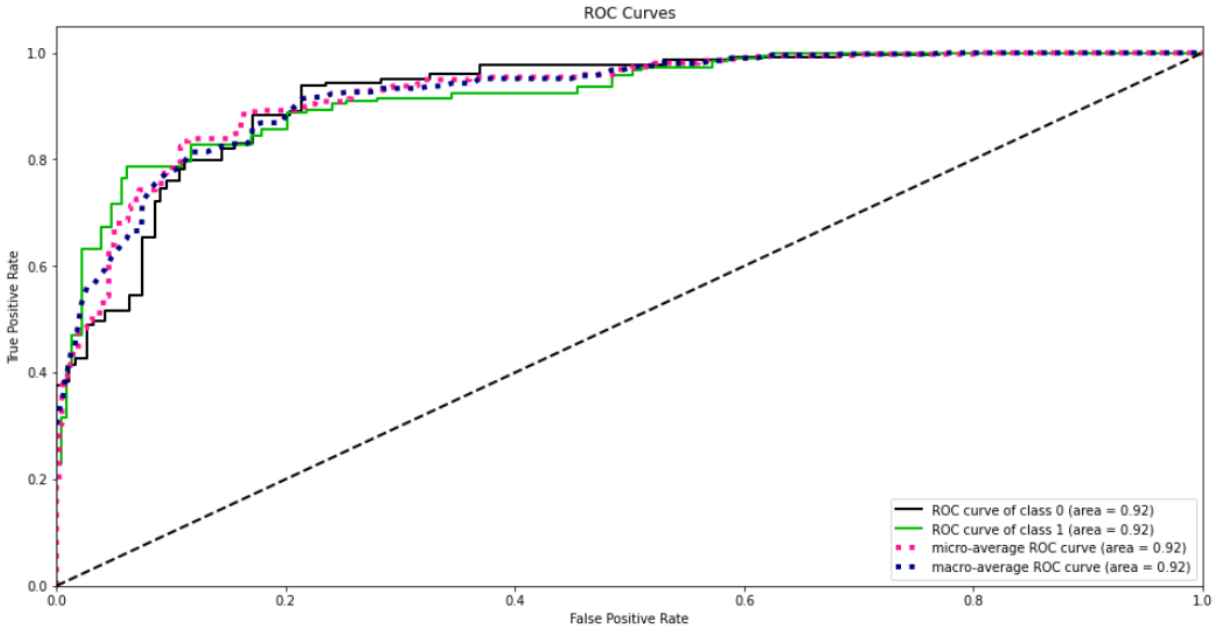
Model Evaluation

Project 3

 # Logistic Regression

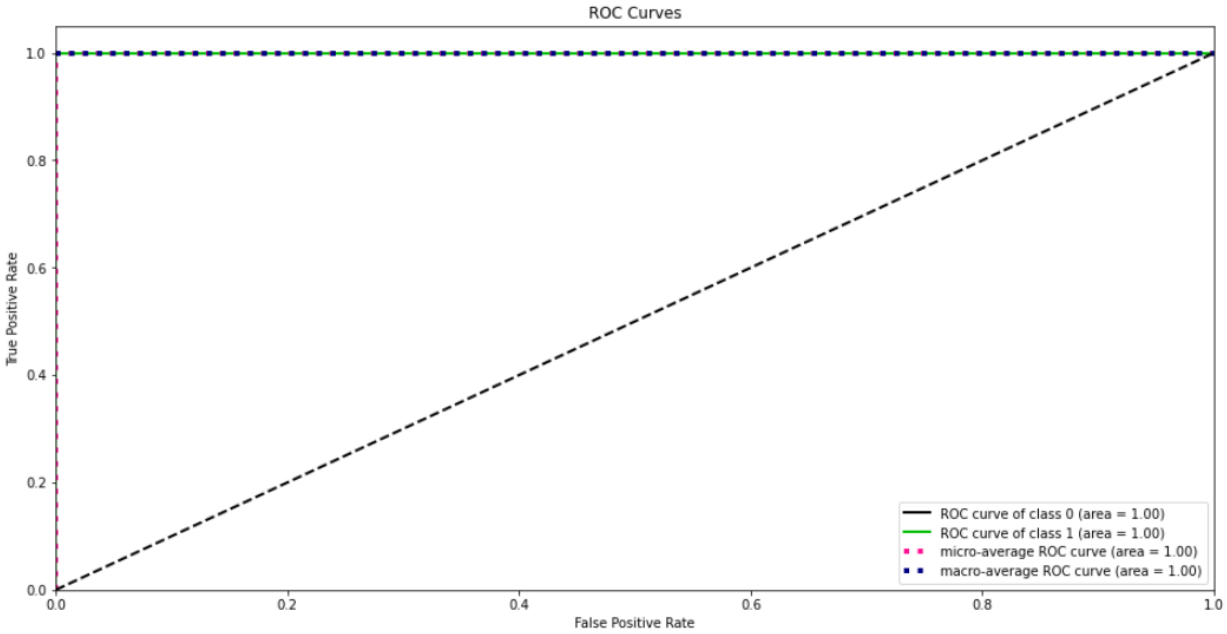


 # SVC

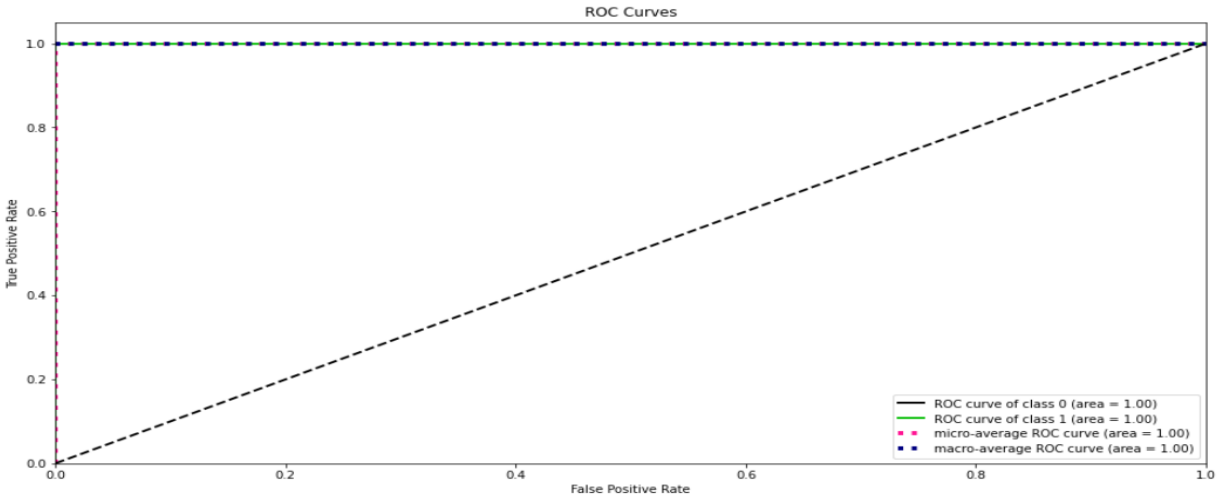


 # Random Forest

Project 3



#Decision Tree



Model Results: Models results match with the confusion matrix

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

Assumptions and Limitations

The medical data constitutes numerous tests necessary to diagnose a particular disease in real-time. Data mining has become inevitable due to the gradual ascent of medical and clinical research data. Data mining can analyze which courses of action prove effective, achieved by differentiating causes, symptoms, and practices of treatments based on the assumption that the effectiveness of medical treatments can be estimated by developing data mining applications. The main limitation with disease detection is that the prediction is derived from numerous factors or symptoms, which is a many-layered subject area complication. This might result in false assumptions that are often accompanied by erratic effects. Due to its limitations, it is evident in employing the knowledge and experience of numerous specialists collected in databases towards supporting the diagnosis process

Challenges

1. Anticipating whitespaces in data and need to work on data alignment.
2. Incorrect variable types.
3. Python package-related issues.
4. Inaccurate or messy patient demographics

Recommendations and use cases:

A Machine learning model Intelligent Heart Disease Prediction System (IHDPS), was developed with data mining techniques like Bayes, Decision Trees was proposed by *Sellappan Palaniappan et al*. IHDPS was capable of addressing complex queries that the traditional decision support systems were not able to. It illustrated the peculiar strength of each of the methodologies in comprehending the objectives of the specified mining objectives. *Sellappan Palaniappan et al.* addressed that it demonstrated a vital knowledge regarding patterns relationships amid medical factors connected with heart disease. IHDPS subsists well-being web-based, user-friendly, scalable, reliable, and expandable.

Ethical Assessment

Appropriate informed consent is fundamental to the ethical conduct of research in humans. Society has demanded more outstanding efforts to protect the individual rights of patients and human subjects. This is an evolving and complex area. To deal with the current regulatory environment, we must understand and appreciate the historical basis for society's concerns, including physician authority's factual and perceived nature.

Medical Researchers and organizations are well-advised to carefully consider the basis for increasing ethical considerations in conducting research in humans and become familiar with regulations that must be met. Analysts, Scientists, any personnel interacting with patients and volunteer subjects should also understand acquiring consent and authenticating to document their understanding of the issues in obtaining consent from patients, dealing with conflicts of interest, and managing PII data.

Conclusion

In this paper, 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.

Appendix

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

Logistic regression: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

SVM: https://scikit-learn.org/stable/modules/svm.html#:~:text=Support%20vector%20machines%20(SVMs)%20are,than%20the%20number%20of%20samples.

Decision Tree: https://scikit-learn.org/stable/modules/tree.html

Random Forest(RFC): https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

roc_curve: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html

PII – Personal identifiable information

Acknowledgment:

I want to thank the authors mentioned below for providing the dataset.

- Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
- University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
- University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
- V.A. Medical Center, Long Beach, and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

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.