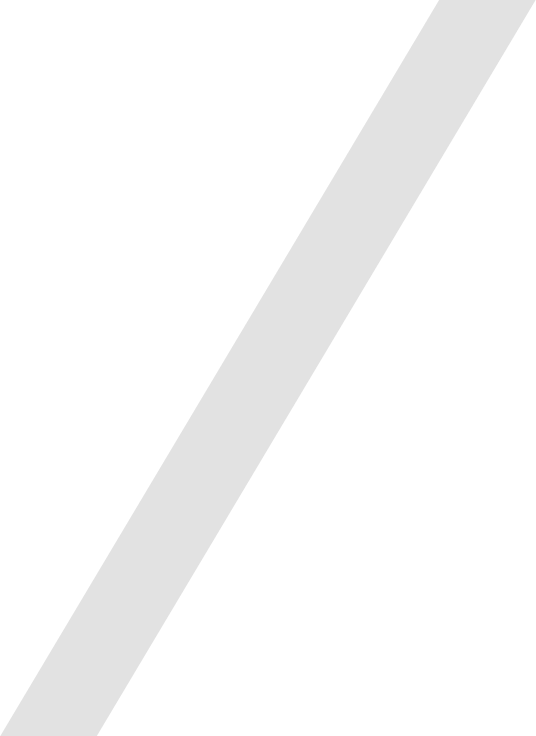
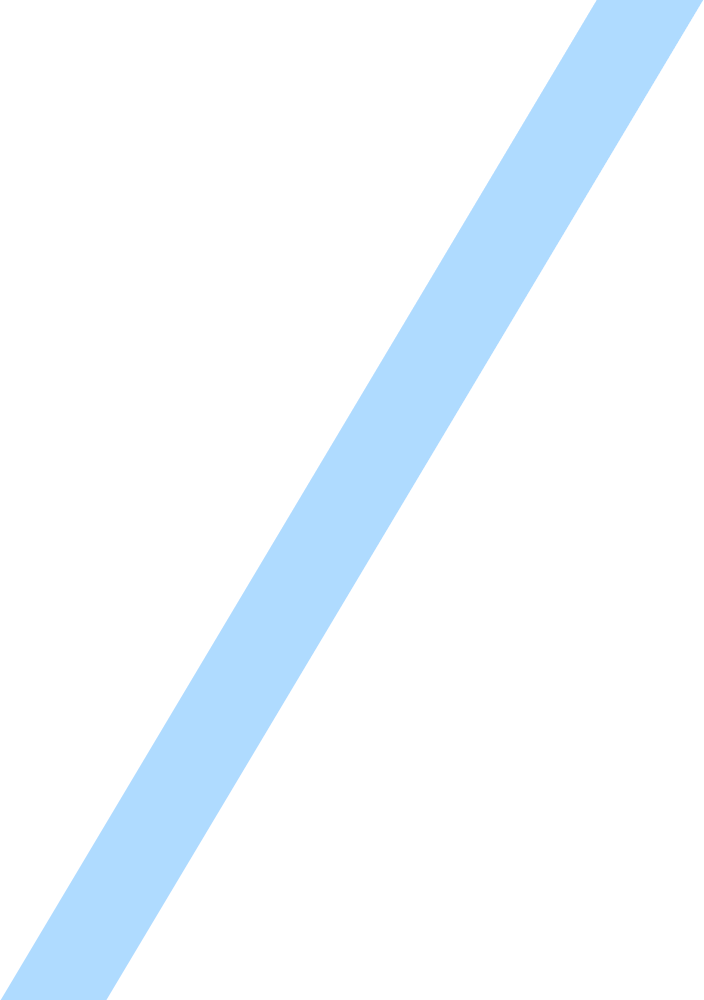
|  |
| --- |
| Heart Disease Prediction using Machine Learning Modelling & Deployment |

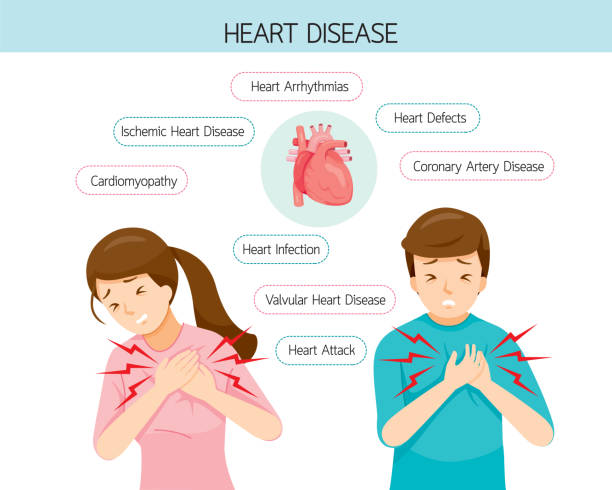
|  |
| --- |
| Electrical & Computer Engineering & Computer Science (ECECS) |

|  |  |
| --- | --- |
| SPRING 22 |  |



Contents

Heart Disease Prediction Using Machine Learning (Includes Modeling & Deployment)



Appendix

Business Understanding 2

Data Understanding 4

Data Preparation 7

Machine Learning Modeling 8

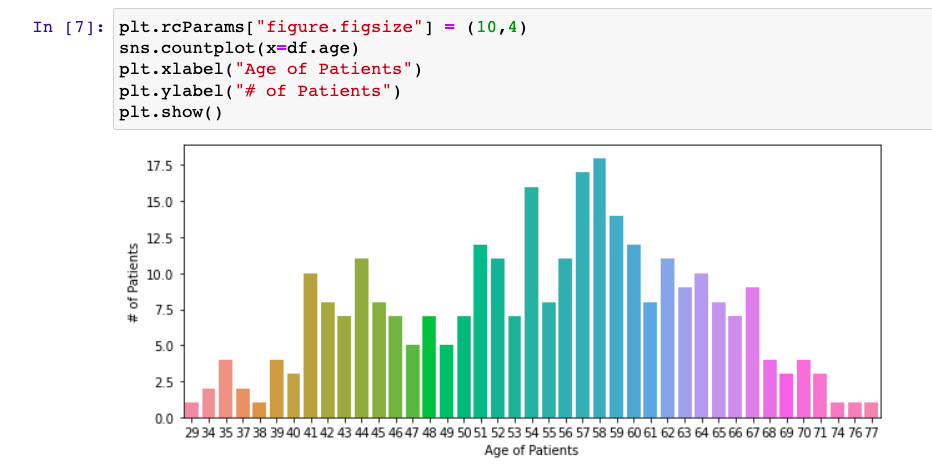
Evaluation Metrics 9

Model Deployment 10

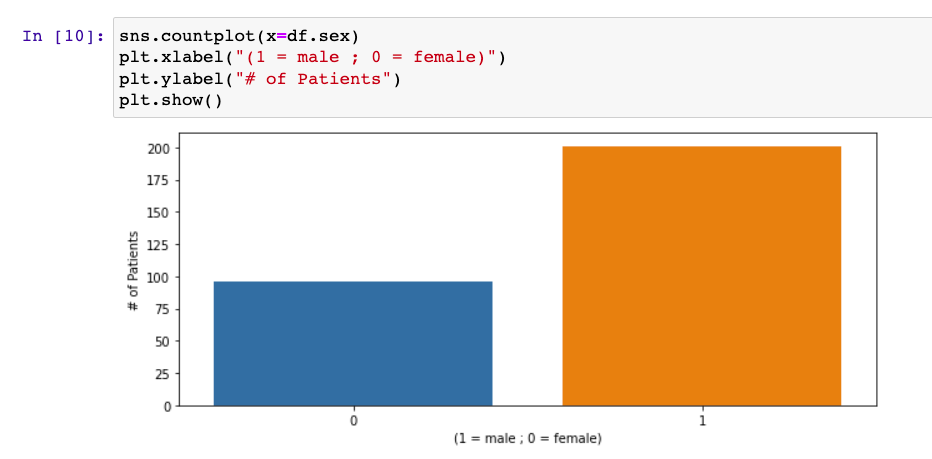
Contributions/References 10

|  |  |  |
| --- | --- | --- |
| Business Understanding / Executive Summary Heart Disease is one of leading causes of death in the United States, high blood pressure, blood cholesterol, and smoking are the key factors contributing to the disease. About half of the population in the country has one of the above risk factors. Healthcare is expensive in United States which makes it difficult for patients to get the disease diagnosed at proper time to take preventive measures. It will be helpful if we build automated systems capable of making prediction about possible heart disease using patients’ historical data. With the help of machine learning, we can accomplish the same and obtain accurate and reliable results as you will see in the project.    **Team Members:**   1. Deepak Singh   **Questions?**  **Contact:**  [dsing8@unh.newhaven.edu](mailto:dsing8@unh.newhaven.edu) / [deepak.singh.ea@gmail.com](mailto:deepak.singh.ea@gmail.com) | | |
| **Heart Disease ML Modelling & Deployment Process** | | |
| **Highlights of Project**  The idea of the project is to provide end-users to try out different models with hyper parameter tuning to make accurate predictions on heart disease classification dataset. We have deployed numerous metrics to evaluate the model such as accuracy and F1 score, area under the curve, confusion matrix and correlation heatmap to understand the results clearly and concisely. The web application was tested locally and deployed on the web using tools such as Stream Lit and Git for seamless integration from development to production setting. Project Link: GitHub Link for my Project:  <https://github.com/deepaksinghea/finalprojectdataeng>  Web application for my Project: <https://share.streamlit.io/deepaksinghea/finalprojectdataeng/main/app.py> Submitted on: April 29, 2022  **Data Understanding:**  The feature description has been provided below:    The dataset is relatively equally balanced with healthy and diseased patients as seen in the visualization below. So, we don’t have to deal with data imbalance which heavily affects machine learning model performance. |  |  |

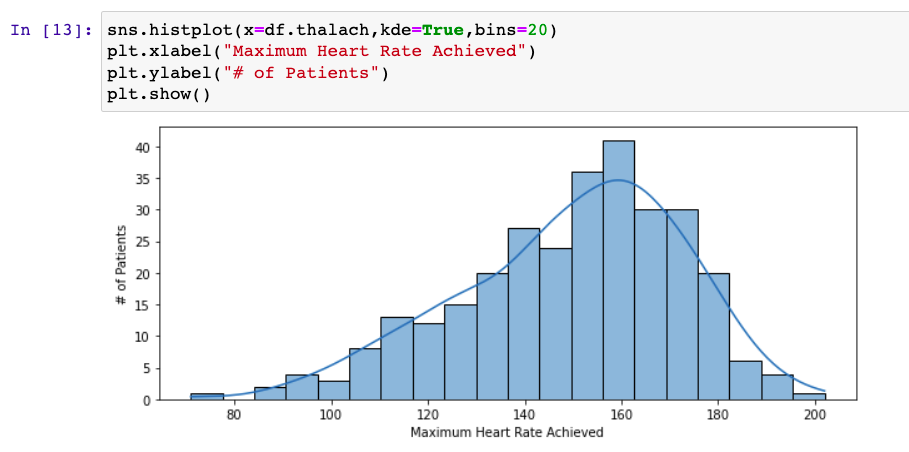
The dataset has the maximum number of patients from age group 50 to 60 as seen in the visualization.



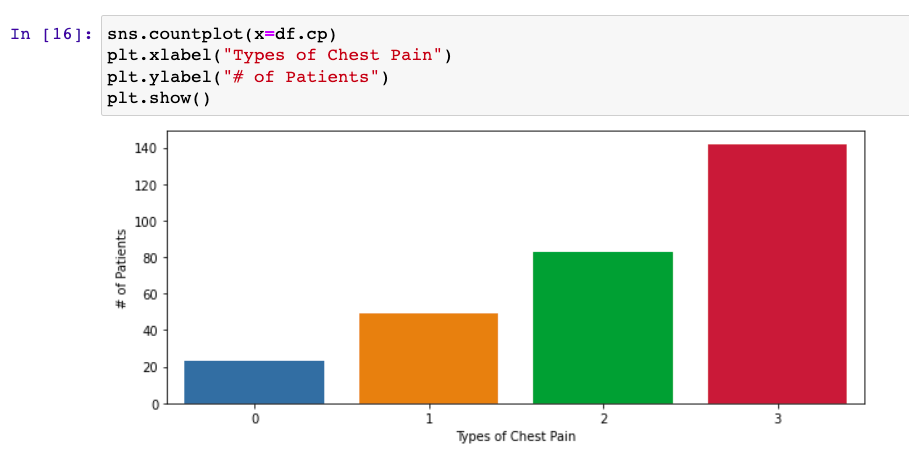
The dataset has relatively more males (~200) than female (~90) as seen in the visualization.



The distribution of maximum heart rate of the patients has been provided below using the visualization.



The different types of chest pain patients have is shown below using visualization. (Type III > Type II > Type I > Type 0)



|  |
| --- |
| Technical Report |

|  |  |
| --- | --- |
| **Data Preparation**  The feature importance can be understood using correlation plot provided below where +1 is strong positive correlation and -1 is strong negative correlation. This helps us in making better decision for feature selection. |  |
| We have no missing data, as seen below. We would have used imputation techniques to fill out missing values if we had any features with the same.    Preview of the dataset: |

Machine Learning Modelling:

For the project, we tried 4 machine learning approach (Logistic Regression, Support Vector Machines, Random Forest, K Nearest Neighbors). We have also used hyper parameter tuning for improved model performance. The accuracy metrics is provided below:

|  |  |
| --- | --- |
| Model | Accuracy Score |
| SVM | 53% |
| K Nearest Neighbors | 61% |
| Logistic Regression | 73% |
| Random Forest | 75% |

The best performing model came out to be Random Forest Classifier.

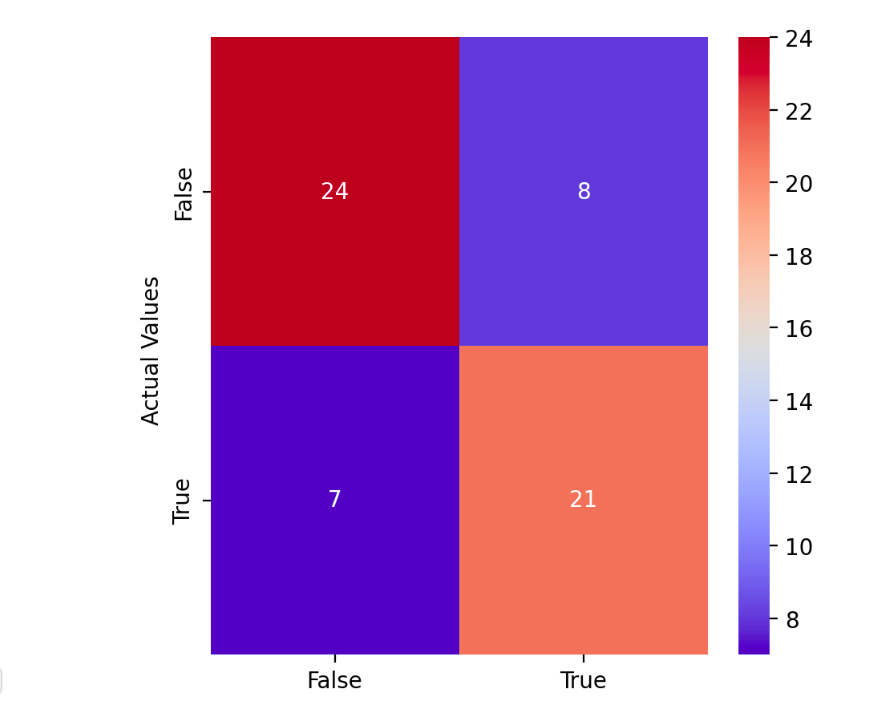
Code Block for building classifier and adding hyper parameters using Scikit-Learn.



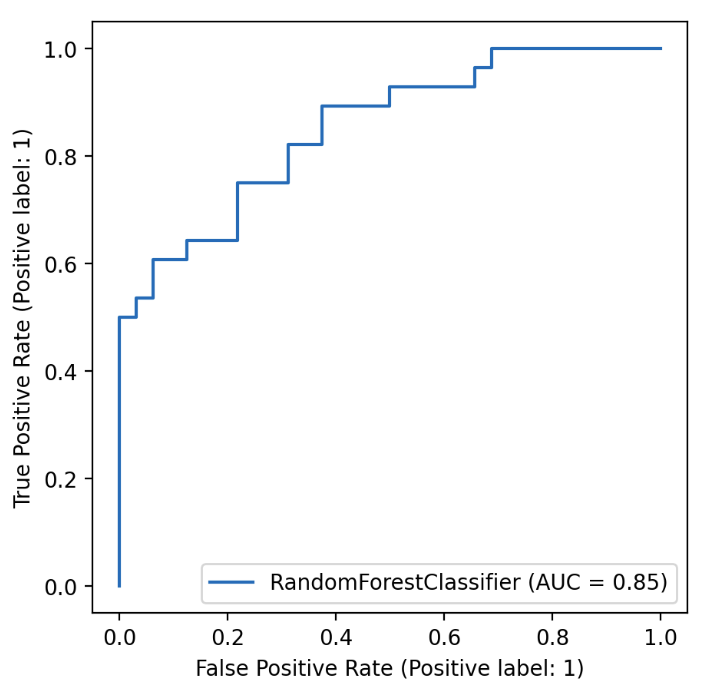
Evaluation Metrics:

Other performance metrics are provided below:

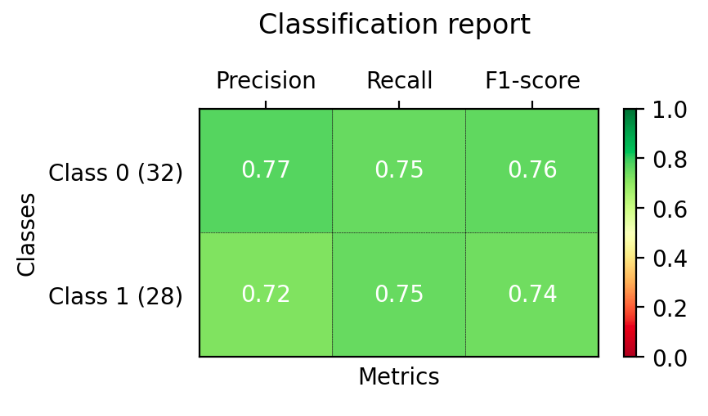
1. Confusion Matrix: 45 samples classified correctly.



1. ROC Curve: AUC value of 0.85 is excellent, value ranges from 0 to 1.



1. Classification Report: Class 0 has 76% F1 score and Class 1 has 74% F1 score.



Model Deployment:

Web Application Deployed!!

Git Version Control

Stream Lit Web Framework

Model Building and Metrics

## 

Web Application URL: <https://share.streamlit.io/deepaksinghea/finalprojectdataeng/main/app.py>

We used Scikit Learn for building and evaluating ML models, then we push our changes to version control system like Git, then we develop a web application for end-users to interact which is later deployed to the web.

## Contributions/References

GitHub Link for my Project: <https://github.com/deepaksinghea/finalprojectdataeng>

Web application for my Project: <https://share.streamlit.io/deepaksinghea/finalprojectdataeng/main/app.py>

Dataset Link: <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>

Thank you!