VISVESVARAYA TECHNOLOGICAL UNIVERSITY
JNANA SANGAMA, BELAGAVI– 590018.

INTERNSHIP REPORT ON

"HEART DISEASE ANALYSIS USING MACHINE LEARNING"

Submitted in Partial fulfillment of the requirements for Eighth semester of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

For the academic year 2020-21

Submitted By

Mithali M (1DB17CS075)

Under the Guidance of

Prof. YASHASWINI.B.M

Department of Computer Science and Engineering

DON BOSCO INSTITUTE OF TECHNOLOGY

KUMBALAGODU, BENGALURU – 560074

ACKNOWLEDGEMENT

Here by I am submitting the Internship report on "**heart disease analysis using machine learning**", as per the scheme of Visvesvaraya Technological University, Belgaum.

In this connection, I would like to express my deep sense of gratitude to my beloved institution Don Bosco Institute of Engineering and also I like to express my sincere gratitude and indebtedness to **Dr. Hemadri Naidu T, Principal, DBIT, Bangalore**.

I would like to express my sincere gratitude to **Prof.B.S.UMASHANKAR** Professor and Head of Dept. of Computer Science and Engineering, for providing a congenial environment to work in and carryout my seminar.

I consider it my cardinal duty to express the deepest sense of gratitude to internship guide, "**Mrs. Yashaswini B M**", Asst. Professor, Department of Computer Science and Engineering, DBIT, Bangalore for his constant help and support extended towards me during the course of the project.

Finally, I am very much thankful to all the teaching and nonteaching members of the Department of Computer Science and Engineering, my seniors, friends and my parents for their constant encouragement, support and help throughout completion of report.

MITHALI M

(1DB17CS075)

VISVESVARAYA TECHNOLOGICAL UNIVERSITY
JNANA SANGAMA, BELAGAVI-590018.

DON BOSCO INSTITUTE OF TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
Kumbalagodu, Mysuru Road, Bengaluru -560074.

Certificate

Certified that the project work entitled **"Heart Disease Analysis using Machine Learning"** carried out by **Mithali M (1DB17CS075)** the bonafied students of **Don Bosco Institute of Technology, Bengaluru,** submitted in partial fulfillment of seventh semester examination of **Bachelor of Engineering** in **Computer Science & Engineering** of the **Visvesvaraya Technological University, Belagavi,** during the year 2020-2021. It is certified that all corrections/suggestions indicated for internal assessment have been incorporated in the report and deposited in the departmental library. The Project report has been approved as it satisfies the academic requirement in respect of Internship Project work prescribed for the said degree.

Signature of Guide

_____

Prof. Yashaswini B M,
Assistant Professor,
Department of CSE,
DBIT, Bengaluru.

Signature of H.O.D

_____

Prof. Umashankar B. S
Professor & Head,
Department of CSE,
DBIT, Bengaluru.

Signature of Internal Examiner

_____

Signature of External Examiner

_____

ABSTRACT

The main aim of this project is to predict if a person has heart disease or not using supervised machine learning. This is a classification problem and hence classification algorithms are used for this model. There are many classification algorithms. Here a few of those algorithms are implemented and the best one is chosen for predicting if a person has heart disease or not based on the performance of the algorithm. The basic machine learning techniques and process is used to solve this problem effectively.

The most widely used Python library for machine learning algorithms called Scikit Learn library is used to implement the machine learning algorithms. Pandas and Numpy libraries are used to read the dataset and manipulate the data. Matplotlib and Seaborn libraries are used to visualize the data by plotting different graphs for better understanding of data. The data is split into training and testing data. Each of the algorithms are trained with training data and tested using the test data. The predicted results are then evaluated using accuracy metrics and hence determine its performance. The model which performs best is selected for future prediction of heart disease of a person.

CONTENTS

CHAPTER 1

INTRODUCTION

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. Machine Learning is used anywhere from automating mundane tasks to offering intelligent insights, industries in every sector try to benefit from it. But there are more examples of ML in use.

- Prediction - Machine learning can also be used in the prediction systems. Considering the loan example, to compute the probability of a fault, the system will need to classify the available data in groups.
- Image Recognition - Machine learning can be used for face detection in an image as well. There is a separate category for each person in a database of several people.
- Speech Recognition - It is the translation of spoken words into the text. It is used in voice searches and more. Voice user interfaces include voice dialing, call routing, and appliance control. It can also be used a simple data entry and the preparation of structured documents.
- Medical diagnoses - ML is trained to recognize cancerous tissues.
- Financial industry and trading - companies use ML in fraud investigations and credit checks.

According to Arthur Samuel, Machine Learning algorithms enable the computers to learn from data, and even improve themselves, without being explicitly programmed.

Machine learning (ML) is a category of an algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available.

Machine learning can be classified into 3 types of algorithms:

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning

In Supervised learning, an AI system is presented with data which is labeled, which means that each data tagged with the correct label. The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

Types of Supervised learning:

- **Classification**: A classification problem is when the output variable is a category, such as "red" or "blue" or "disease" and "no disease".
- **Regression**: A regression problem is when the output variable is a real value, such as "dollars" or "weight".

Data Summary

The original database contains 76 attributes and information from 4 different hospitals. This subset of dataset is the most widely used and contains 14 attributes and only information from the Cleveland hospital. The goal is to predict if the person has heart disease or not.

Data Set Characteristics:	Multivariate	**Number of Instances :**	303	**Area:**	Medical
Attribute Characteristics:	Numeric	**Number of Attributes :**	14	**Year Donated:**	1988
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	3305

License:

UCI Machine Learning Repository

Creators:

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

Donor:

David W. Aha (aha '@' ics.uci.edu) (714) 856-8779

PROBLEM STATEMENT

The heart disease dataset contains different medical information about the patient. It has 303 instances with 14 variables each. The dataset is good for classification task. The model can be used to predict if the person/patient has heart disease or not.

Since I have to predict if the patient has heart disease or not, that is, yes or no and in binary it is 0 or 1. Hence this is binary classification problem and can perform various different algorithms like logistic regression, decision tree random forests, etc and differentiate between the models and analyze their performances.

Here I am performing various different Classification algorithms like Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, K Neighbors Classifier, Support Vector Machine, Artificial Neural Network using Scikit learn and trying to differentiate between the models and analyze their performances. Based on the performance of the different models, I will be choosing the best performing algorithm for this problem out of all the algorithms.

CHAPTER 2

LITERATURE SURVEY

Machine learning can be classified into 3 types: Supervised Learning, Unsupervised Learning and Reinforcement Learning.

- **Supervised machine learning:** Supervised learning is where you have input variables and an output variable and we use an algorithm to learn the mapping function from the input to the output.
- **Unsupervised machine learning**: Unsupervised learning is where you only have input data and no corresponding output variables.
- **Reinforcement learning**: It is a technique that enables an agent to learn in an interactive environment by trial and error using feedback from its own actions and experiences.

Supervised learning problems can be further grouped into regression and classification problems.

- **Classification**: A classification problem is when the output variables are discrete values, such as “red” or “blue” and “disease” or “no disease”.
- **Regression**: A regression problem is when the output variables are continuous real values, such as “dollars” or “weight”.

Important Python libraries for Machine Learning

i. **Scikit-learn:** It is a library in Python that provides many unsupervised and supervised learning algorithms. It is a module for machine learning built on top of SciPy.

ii. **Matplotlib:** It is one of the most popular and oldest plotting libraries in Python which is used in Machine Learning. In Machine learning, it helps to understand the huge amount of data through different visualizations.

iii. **Seaborn:** It is a plotting library that offers a simpler interface, sensible defaults for plots needed for machine learning, and most importantly, the plots are aesthetically better looking than those in Matplotlib.

iv. **Pandas:** It is one of the tools in Machine Learning which is used for data cleaning and analysis. It has features which are used for exploring, cleaning, transforming and visualizing from data.

v. **Numpy:** NumPy library is an important foundational tool for studying Machine Learning. Many of its functions are very useful for performing any mathematical or scientific calculation. As it is known that mathematics is the foundation of machine learning, most of the mathematical tasks can be performed using NumPy.

Types of classification algorithms:

- Logistic Regression: In this algorithm, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function.
- Decision Tree: Given a data of attributes together with its classes, a decision tree produces a sequence of rules that can be used to classify the data.
- Random Forest: It is a meta-estimator that fits a number of decision trees on various sub-samples of datasets and uses average to improve the predictive accuracy of the model and controls over-fitting.
- Gradient Boosting: It combines many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting.
- K-Nearest Neighbors: It is a type of lazy learning as it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the k nearest neighbors of each point.
- Support Vector Machine: It is a representation of the training data as points in space separated into categories by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.
- Artificial Neural Network: An ANN is based on a collection of connected units or nodes called artificial neurons. It is used for modeling non-linear problems and to predict the output values for given input parameters from their training values.

CHAPTER 3

OBJECTIVES

- To obtain the dataset and preprocess the data, that is, cleaning the data and make sure the data is proper.
- To visualize and analyze the data through data visualization using matplotlib and seaborn libraries to plot graphs.
- Split the data for training and testing with 70 percent or more data for training the algorithm.
- To perform various different Classification algorithms like Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, K Neighbors Classifier, Support Vector Machine and Artificial Neural Network using Scikit learn.
- Try to differentiate between the models.
- Analyze their performances based on accuracy metrics.
- Based on the performance of the different models, choose the best performing algorithm for predicting if a person has heart disease or not.

CHAPTER 4

SYSTEM REQUIREMENT SPECIFICATION

Requirements specification is a specification of software requirements and hardware requirements required to do the project.

Software Requirements Specification

Software Requirements are the software resources that are need to do the project work. These resources are installed on a computer in order to provide functions, services, hardware accessing capabilities to do the project. In our project we used the following software resources.

- Operating System: Windows 7 or higher version
- Python version 2 or higher
- IDE: Anaconda – Jupyter Notebook or Google colab

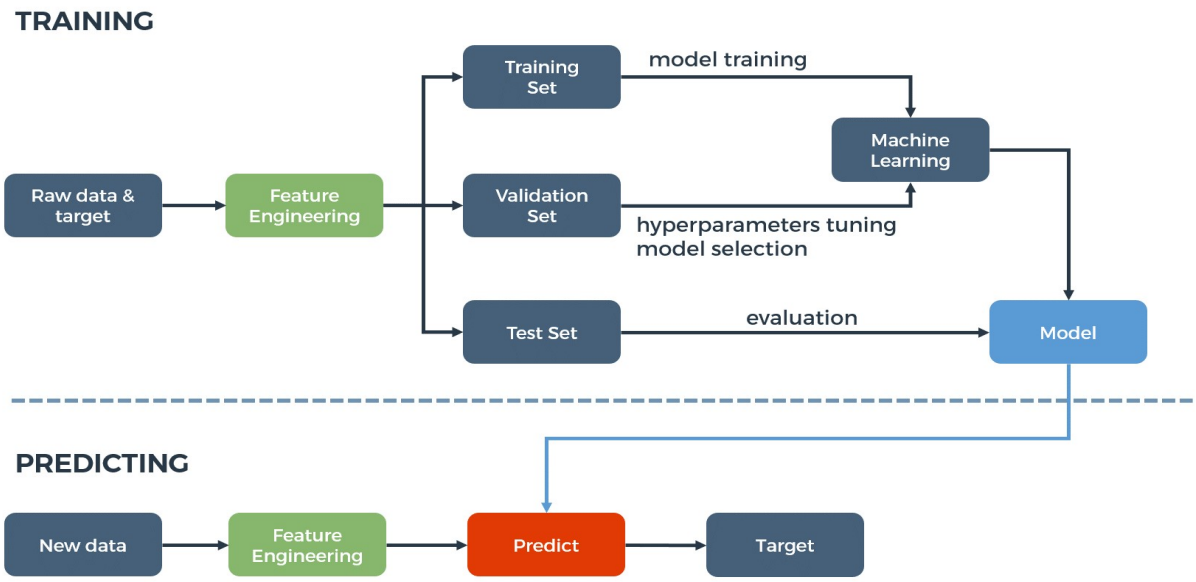
Hardware Requirements Specification

Hardware Requirements are the hardware resources that are need to do the project work. These resources are a computer resource provides functions and services to do the project. Hardware resources required for our project are shown below.

- Processor: Intel Corei5+
- RAM: >=2GB

CHAPTER 5

SYSTEM ARCHITECTURE



- This system architecture helps predict if a person has heart disease or not efficiently.
- Once we get the dataset, that is, the raw data and target, we need to preprocess that data since there will be missing values and inappropriate data.
- To understand the data we can perform data visualization to help us better in training our model.
- After preprocessing the data, feature engineering is performed in order to choose the appropriate and important attributes which has the greatest influence for the correct prediction.
- To train the model data can be split into training set, validation set and test set. Training set is used for training the machine learning model, validation set is used for hyper parameter tuning of selected model and test set is used for evaluating the model.
- When new data is given to predict, then the model which performed best while training is does the job to predict the target.

CHAPTER 6

METHODOLOGY

The tasks that were carried out are as follows:

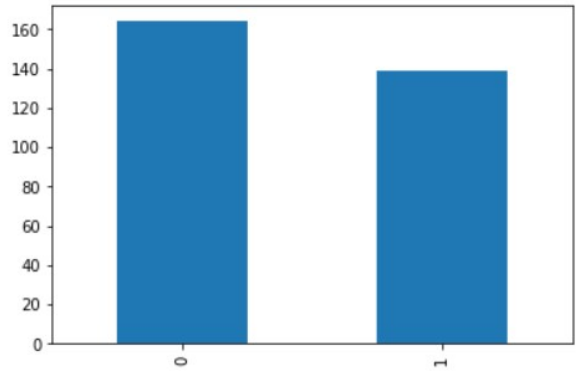
- Data preprocessing: It is the first and crucial step while creating a machine learning model. The dataset we obtained is not completely accurate and error free. So data processing is required for cleaning the data and making it suitable for a machine learning model which will increase the accuracy and efficiency of a machine learning model.
- Data visualization: This helps in exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. It is used to express and demonstrate key relationships in plots and charts. These plots give a qualitative understanding of the data.
- Feature selection: It is done to reduce the number of input variables to those that are believed to be most useful to a model in order to predict the target variable.
- Splitting the dataset for training and testing: 70% of data for training and 30% of data for testing.
- Building the model: Need to train the algorithm with training data and make predictions using the test data.
- Evaluating the model: The last step is to evaluate the performance and efficiency of the algorithm. This step becomes useful to compare it with different algorithms working on the same dataset to see which algorithm performs the best.

For this project I have applied seven different types of classification algorithms listed below:

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest Classifier
4. Gradient Boosting Classifier
5. K Neighbors Classifier
6. Artificial Neural Network
7. Support Vector Machine

Data Visualization

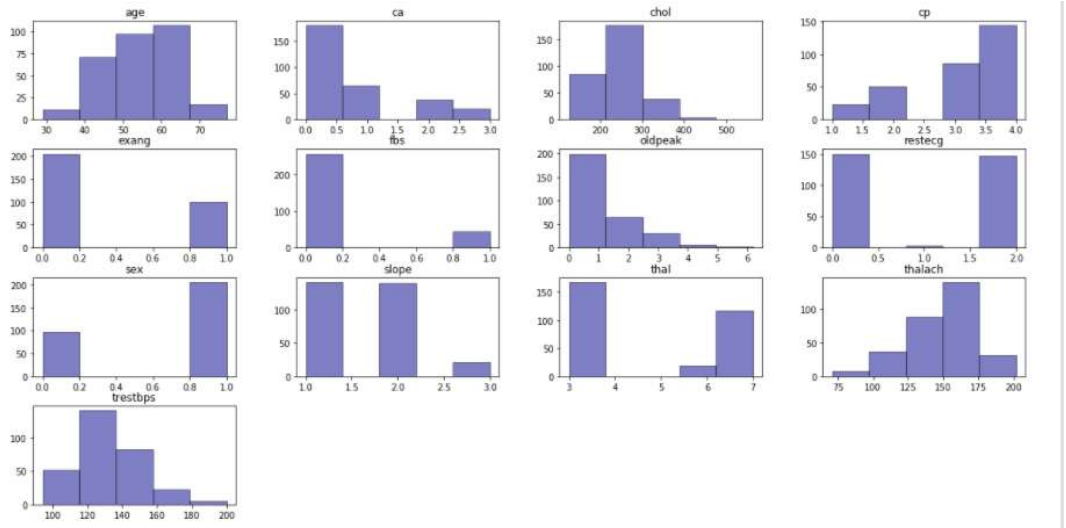
Visualization using bar graph:



This graph tells us how many patients have heart disease and how many patients don't have heart disease. From the graph we can say that in the dataset there are more patients with heart disease.

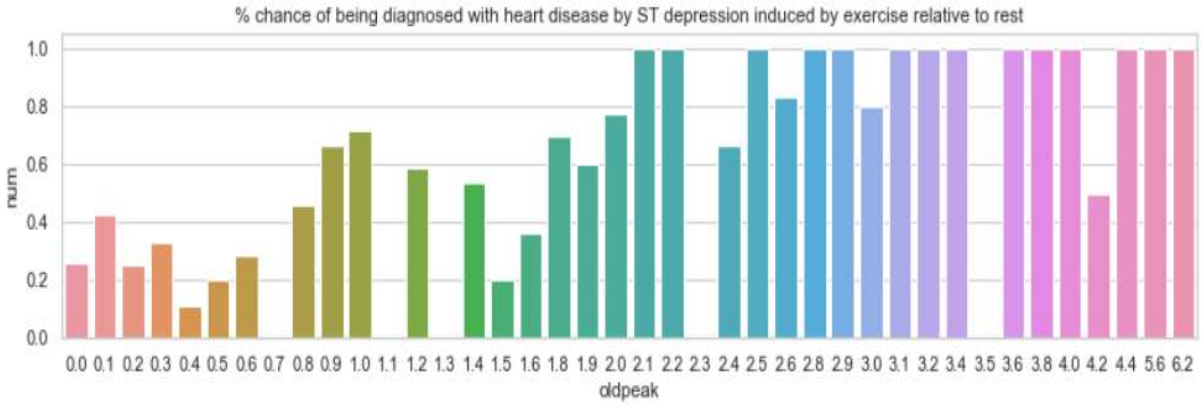
Outcome	count
Heart disease (0)	164
No heart disease (1)	139

Visualization using histogram:



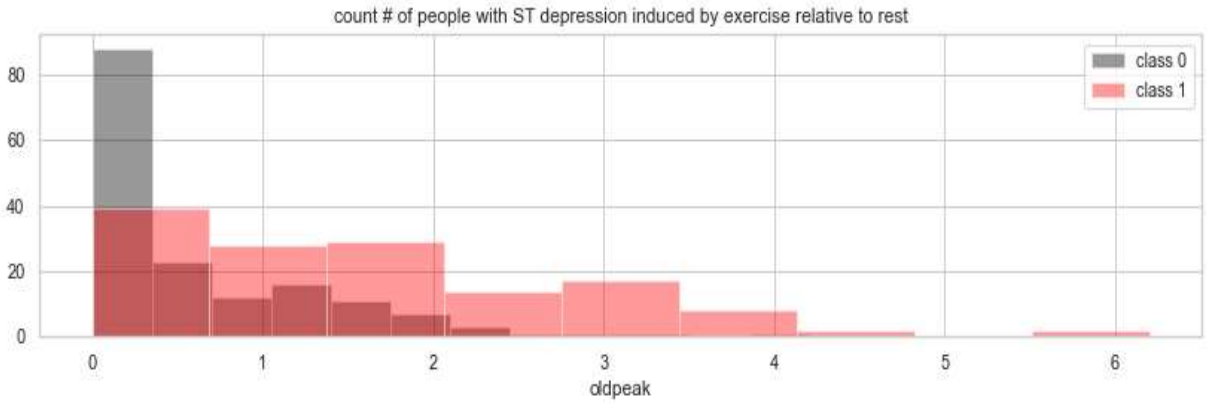
This graph plots histogram for each feature of the dataset to visualize the frequency at each uniform distribution of the feature values. There are 13 features for the dataset and each of the features has numerical data and hence histogram gives clear visualization on its frequency.

Visualization using bar plot by grouping:



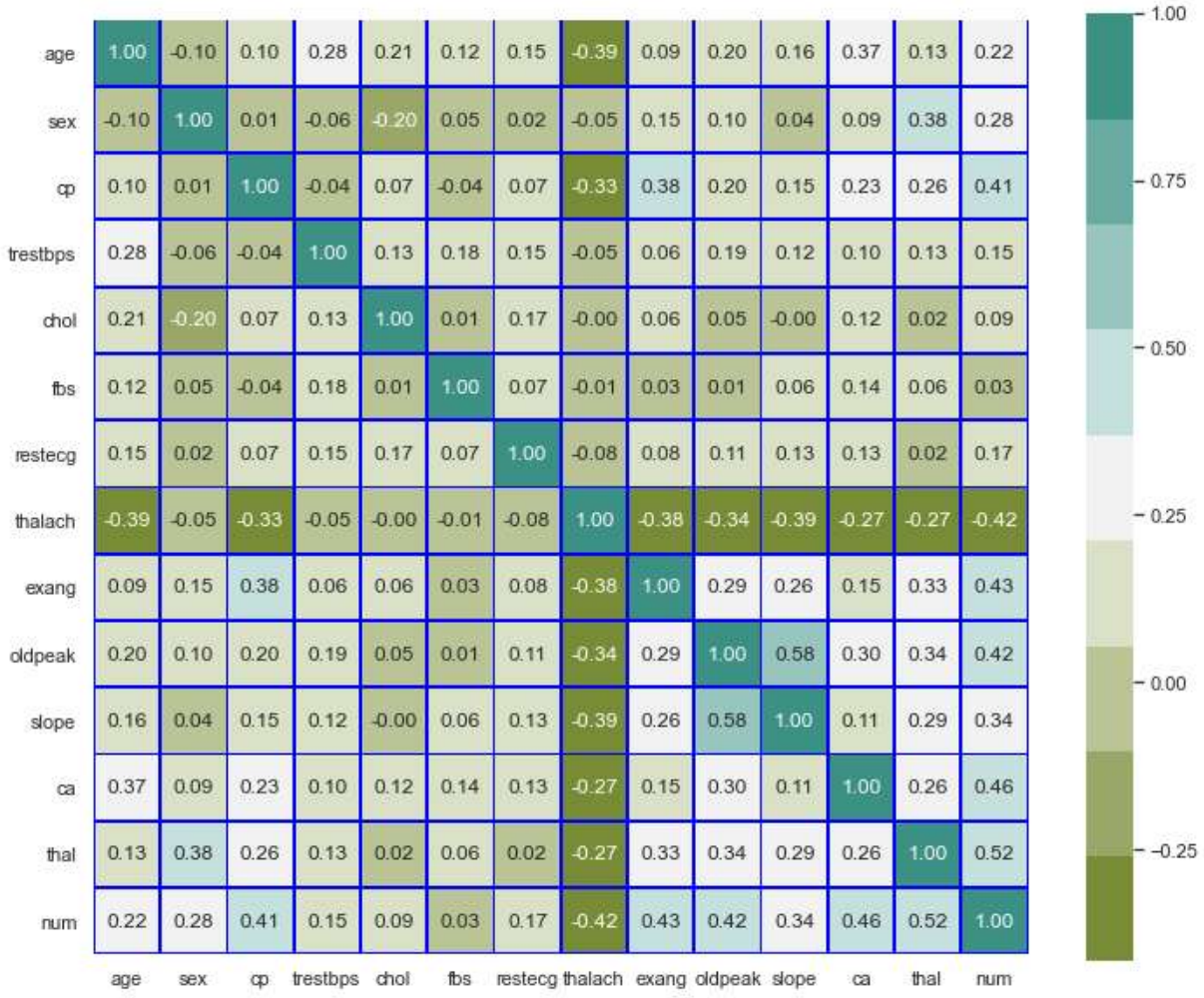
This is bar graph grouped by oldpeak (ST depression induced by exercise relative to rest) with Outcome (num) column's mean to find the percentage chance of being diagnosed with heart disease by ST depression induced by exercise relative to rest.

Visualization using two distplots:



In this graph there are two distplots where one is on patients with heart disease (class 0) and another one on patients without heart disease (class 1) with oldpeak (ST depression induced by exercise relative to rest) to count the number of people with ST depression induced by exercise relative to rest.

Visualization using heatmap:



This is graph of heatmap of the correlation matrix of the dataset. In this graph we can see the correlations between the variables of the dataset using correlation matrix. It gives a better understanding of the relationship between the variables.

CHAPTER 7

RESULTS

1. Logistic Regression

Accuracy:

The Accuracy score obtained by the Logistic Regression is-> 92.307 %

Classification Report:

classification report:

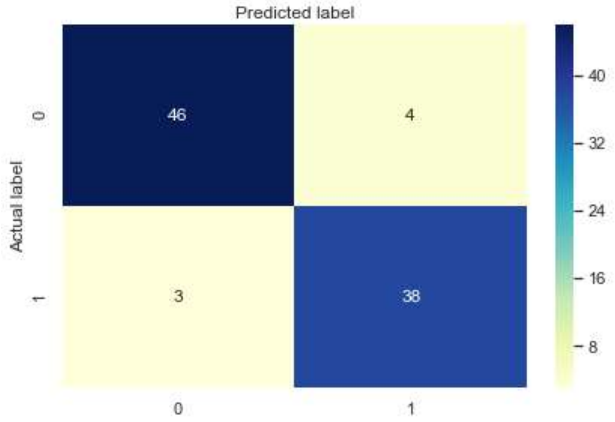
	precision	recall	f1-score	support
0	0.94	0.92	0.93	50
1	0.90	0.93	0.92	41
accuracy			0.92	91
macro avg	0.92	0.92	0.92	91
weighted avg	0.92	0.92	0.92	91

Confusion Matrix:

[[46 4]

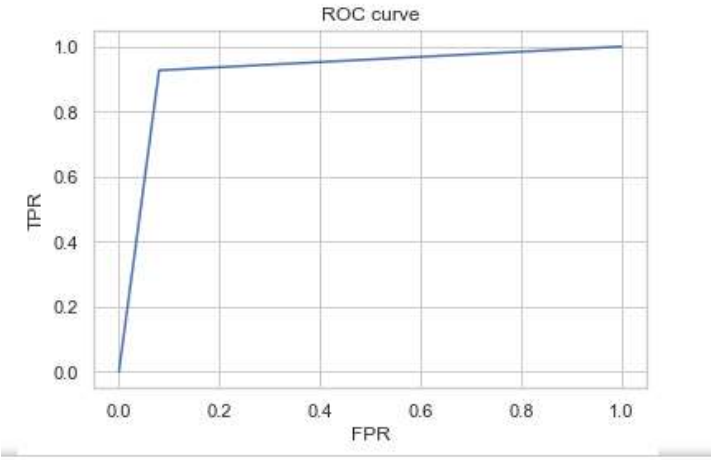
[3 38]]

Confusion Matrix using Heatmap:



ROC_AUC_SCORE: 0.9234146341463415

ROC Curve:



2. Decision Tree Classifier

Accuracy:

The Accuracy score obtained by the Decision Tree Classifier is-> 75.82 %

Classification report:

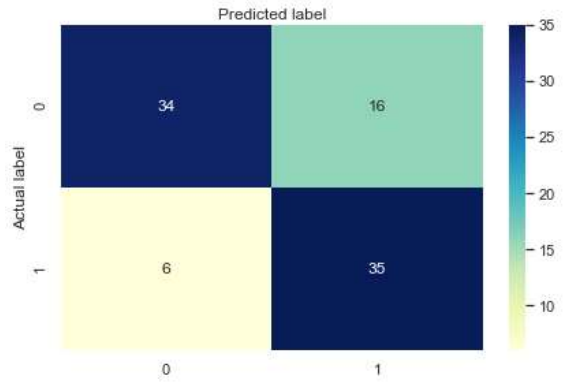
classification report:

	precision	recall	f1-score	support
0	0.85	0.68	0.76	50
1	0.69	0.85	0.76	41
accuracy			0.76	91
macro avg	0.77	0.77	0.76	91
weighted avg	0.78	0.76	0.76	91

Confusion matrix:

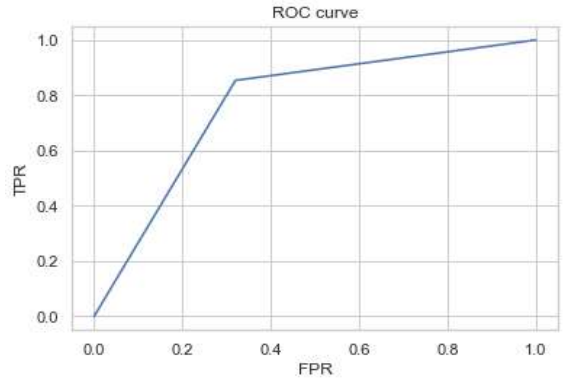
[[34 16]
[6 35]]

Confusion Matrix using Heatmap:



ROC_AUC_SCORE : 0.7668292682926828

ROC Curve:



3. Random Forest Classifier

Accuracy:

The Accuracy score obtained by the Random Forest Classifier is-> 84.61 %

Classification report:

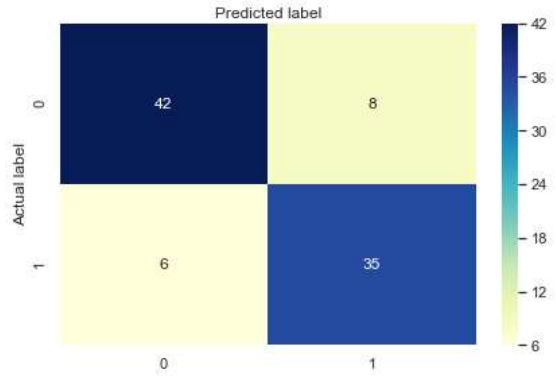
classification report:

	precision	recall	f1-score	support
0	0.88	0.84	0.86	50
1	0.81	0.85	0.83	41
accuracy			0.85	91
macro avg	0.84	0.85	0.85	91
weighted avg	0.85	0.85	0.85	91

Confusion matrix:

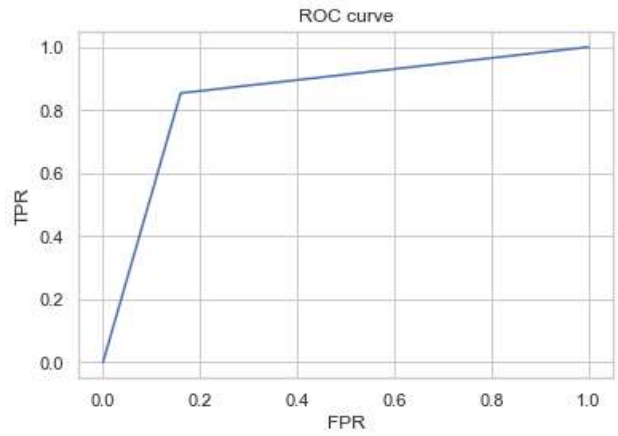
[[42 8]
[6 35]]

Confusion Matrix using Heatmap:



ROC_AUC_SCORE: 0.8468292682926828

ROC Curve:



4. Gradient Boosting Classifier

Accuracy:

The Accuracy score obtained by the Random Forest Classifier is-> 85.71 %

Classification report:

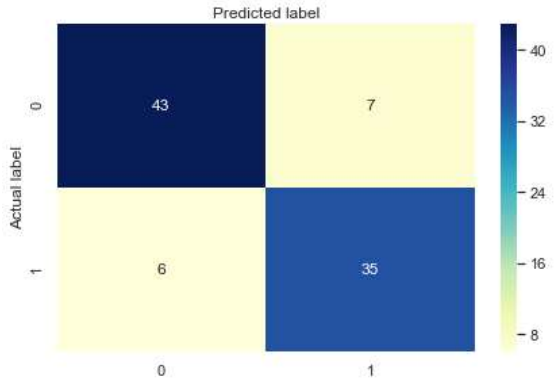
classification report:

	precision	recall	f1-score	support
0	0.88	0.86	0.87	50
1	0.83	0.85	0.84	41
accuracy			0.86	91
macro avg	0.86	0.86	0.86	91
weighted avg	0.86	0.86	0.86	91

Confusion matrix:

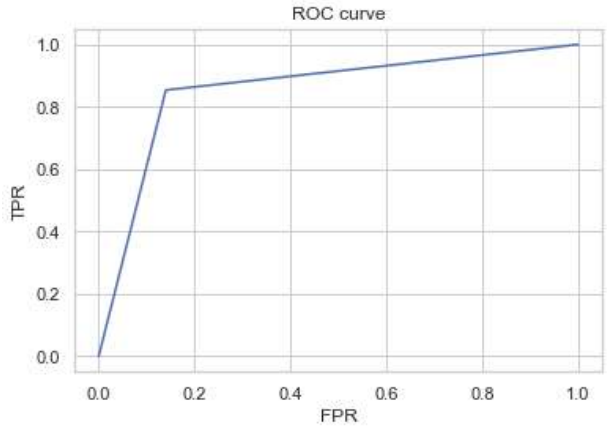
[[43 7]
[6 35]]

Confusion Matrix using Heatmap:



ROC_AUC_SCORE: 0.8568292682926829

ROC Curve:



5. K Neighbors Classifier

Accuracy:

The Accuracy score obtained by the Random Forest Classifier is-> 74.72 %

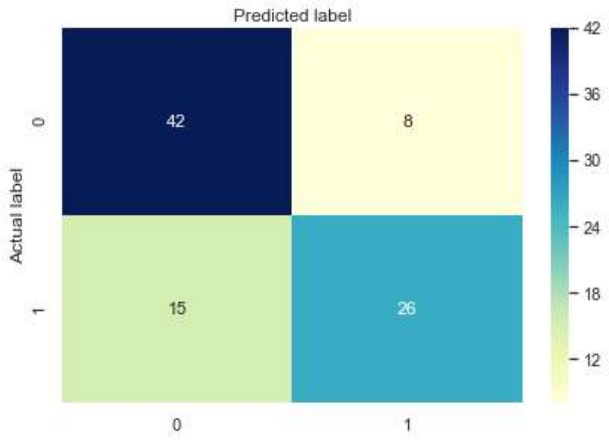
Classification report:

```
classification report:

              precision    recall  f1-score   support

           0       0.74      0.84      0.79        50
           1       0.76      0.63      0.69        41

    accuracy                           0.75        91
   macro avg       0.75      0.74      0.74        91
weighted avg       0.75      0.75      0.74        91
```

Confusion matrix:

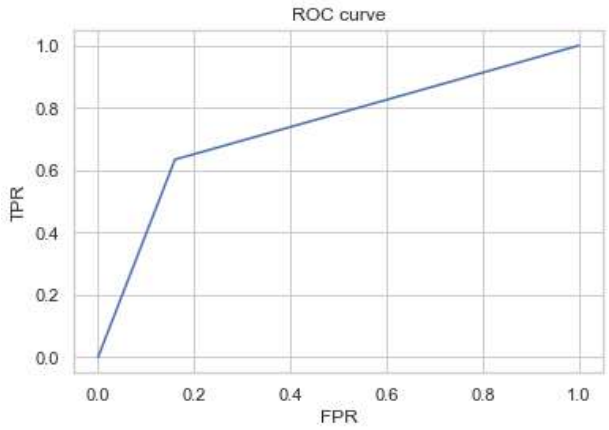
[[42 8]
[15 26]]

Confusion Matrix using Heatmap:



ROC_AUC_SCORE : 0.7370731707317074

ROC Curve:



6. Artificial Neural Network using Sklearn

Accuracy:

The Accuracy score obtained by the Random Forest Classifier is-> 89.01 %

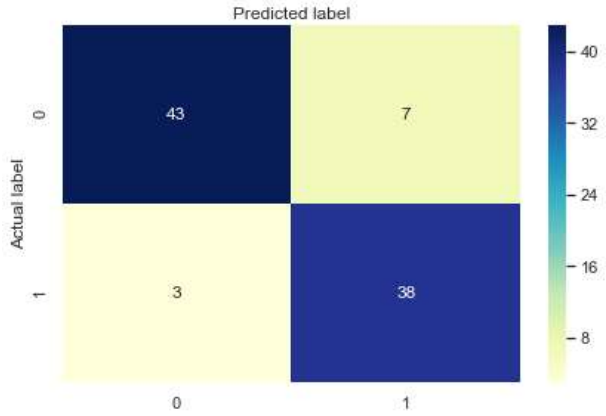
Classification report:

```
classification report:

              precision    recall  f1-score   support

           0       0.93      0.86      0.90        50
           1       0.84      0.93      0.88        41

    accuracy                           0.89        91
   macro avg       0.89      0.89      0.89        91
weighted avg       0.89      0.89      0.89        91
```

Confusion matrix:

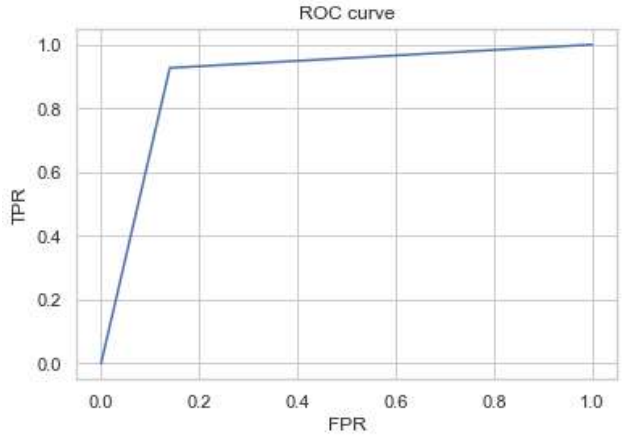
[[43 7]
[3 38]]

Confusion Matrix using Heatmap:



ROC_AUC_SCORE: 0.8934146341463415

ROC Curve:



7. Support Vector Machine

Accuracy:

The Accuracy score obtained by the Random Forest Classifier is-> 91.21 %

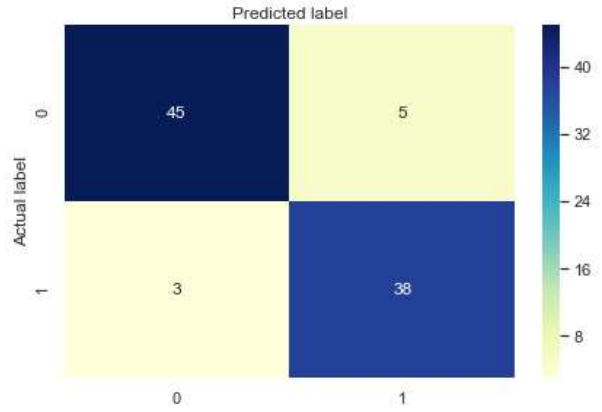
Classification report:

```
classification report:

              precision    recall  f1-score   support

           0       0.94      0.90      0.92        50
           1       0.88      0.93      0.90        41

    accuracy                           0.91        91
   macro avg       0.91      0.91      0.91        91
weighted avg       0.91      0.91      0.91        91
```

Confusion matrix:

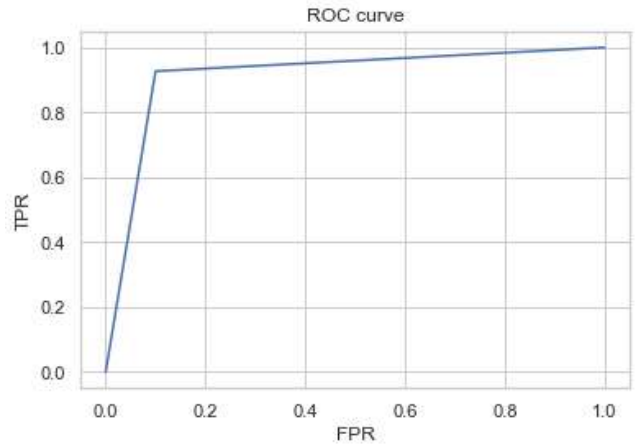
[[45 5]
[3 38]]

Confusion Matrix using Heatmap:



ROC_AUC_SCORE : 0.9134146341463415

ROC Curve:



CONCLUSION

The problem is a Binary Classification problem and there are many supervised classification machine learning algorithms that can be applied. Here seven machine learning algorithms have been applied and tested: logistic regression, decision tree, random forest, gradient boosting, k neighbors, SVM and ANN.

From the entire classification algorithms used, Logistic Regression algorithm gives the highest accuracy of 92 % and highest ROC_AUC_SCORE of 0.923. SVM also gave a very good accuracy of 91% and ROC_AUC_SCORE is 0.913. So, to predict if the person/patient has heart disease or not Logistic Regression or SVM is the best Machine learning algorithm that can be used for this binary classification problem.

Bibliography

- https://scikit-learn.org/stable/supervised_learning.html#supervised-learning
- https://matplotlib.org/stable/tutorials/introductory/sample_plots.html
- Machine Learning, Tom Mitchell, McGraw Hill
- Machine Learning and Deep Learning with Python, Scikit-learn, and TensorFlow 2, 3rd Edition
- https://seaborn.pydata.org/introduction.html
- https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/