**TABLE CONTENT**

**TITLE PAGE NO**

ACKNOWLEDGEMENT ii

SYNPOSIS iii

Chapter 1 Introduction 2

Chapter 2 Objectives 3

Chapter 3 Data set 4

Chapter 4 Software Requirement Specification 5

Chapter 5 Methodology 6

Chapter 6 Result and Validation 9

Chapter 7 Conclusion 17

Bibliography

References

**CHAPTER 1: INTRODUCTION**

* Heart disease is a term covering any disorder of the heart.
* **Fast facts on heart disease**
* One in every four deaths in the U.S. is related to heart disease.
* [Coronary heart disease](https://www.medicalnewstoday.com/articles/184130.php), [arrhythmia](https://www.medicalnewstoday.com/articles/8887.php), and myocardial infarction are some examples of heart disease.
* Heart disease might be treated with medication or surgery.
* [Quitting smoking](https://www.medicalnewstoday.com/articles/241302.php) and exercising regularly can help prevent heart disease
* **Symptoms**
* Common symptoms include chest pain, breathlessness, and heart palpitations. The chest pain common to many types of heart disease is known as [angina](https://www.medicalnewstoday.com/articles/8886.php), or angina pectoris, and occurs when a part of the heart does not receive enough oxygen
* **Causes**
* Heart disease is caused by damage to all or part of the heart, damage to the coronary arteries, or a poor supply of nutrients and oxygen to the organ.
* **Treatment**
* Medication
* Surgery
* **Prevention**
* Eat a balanced diet
* Exercise regularly
* Maintain a healthy [body weight](https://www.medicalnewstoday.com/info/obesity/how-much-should-i-weigh.php) for your height

**CHAPTER 2: OBJECTIVES**

Improve **cardiovascular** health and quality of life through prevention, detection, and treatment of risk factors for **heart attack** and stroke; early identification and treatment of **heart attacks** and strokes; prevention of repeat **cardiovascular events**; and reduction in deaths from **cardiovascular disease**

**CHAPTER 3: DATASETS**

<https://www.kaggle.com/mazharkarimi/heart-disease-and-stroke-prevention>

Heart Prediction dataset:

This dataset was collected and made available by “National Institute of Heart and stroke-prevention” as part of the Pima Indians Heart Database. This dataset has 304 rows and 14 columns.

The following are the significant features used

* Age
* Sex
* Cp
* Trestbps
* Chol
* Fbs
* Restecg
* Thalach
* Exang
* Oldpeak
* Slope
* Ca
* Thal
* target

Several constraints were placed on the selection of these instances from a larger database. Using this dataset, we will build a logistic regression machine learning algorithm to predict whether or not a patient has heart problem.

**CHAPTER 4: SOFTWARE REQUIREMENT SPECIFICATION**

* Hardware requirements

Processor : Intel(R) Core(TM) i5 CPU

CPU : 4 GB (Minimum)

* Soft Requirements

Operating System: Window 2010

Development Kit : Python 3, Jupyter Notebook, Anaconda

**CHAPTER 5: Methodology**

**(a). Data Extraction**

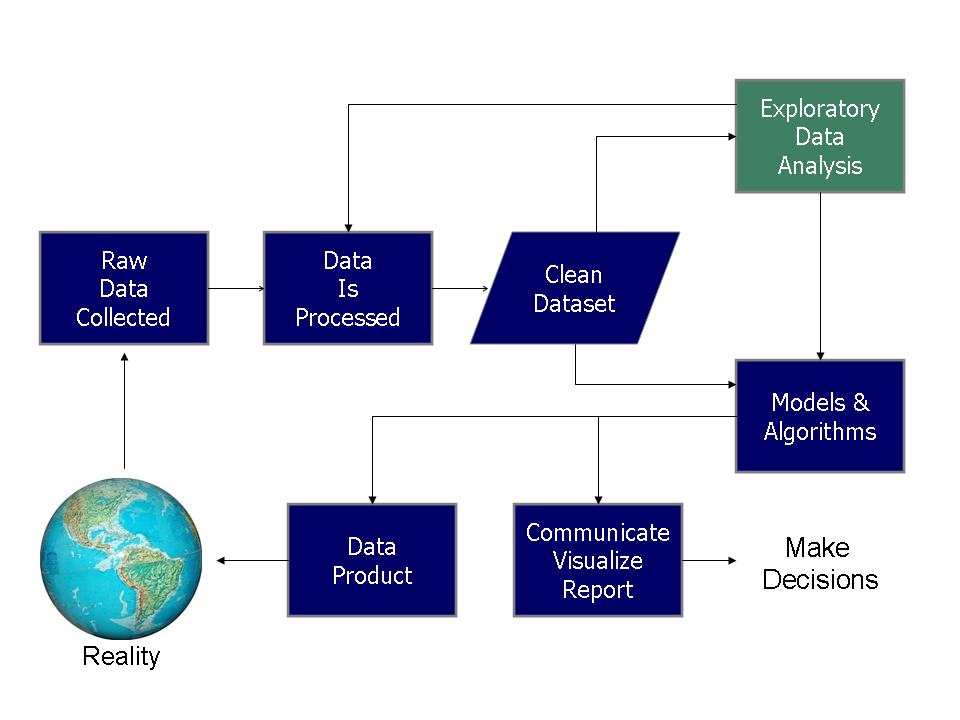
 **Data extraction** is a process that involves retrieval of **data** from various sources. Frequently, companies **extract data** in order to process it further, migrate the **data** to a **data** repository (such as a **data** warehouse or a **data** lake) or to further analyze it. It's common to transform the **data** as a part of this process.

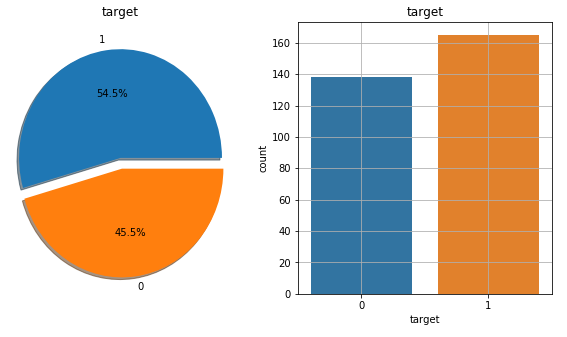
**(b). Data Preprocessing**

**Data preprocessing** is a **data** mining technique that involves transforming raw **data** into an understandable format. Real-world **data** is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors.

**(c) . Data Visualizations**

**Data visualization** is the graphical representation of information and **data**. By using visual elements like charts, graphs, and maps, **data visualization** tools provide an accessible way to see and understand trends, outliers, and patterns in **data**.





FLOW DIAGRAM

View Data

Select Dataset

Start

Classification using LR

Data Preprocessing

Result Generation

Prediction

Encoding Categorical Data

Feature Scaling

Splitting Dataset into Training and Test Data set

Missing Data Removal

**Proposed system**

Data Preprocessing

Feature Engineering

Import dataset

Evaluation

Classification Algorithm

Making Predicting

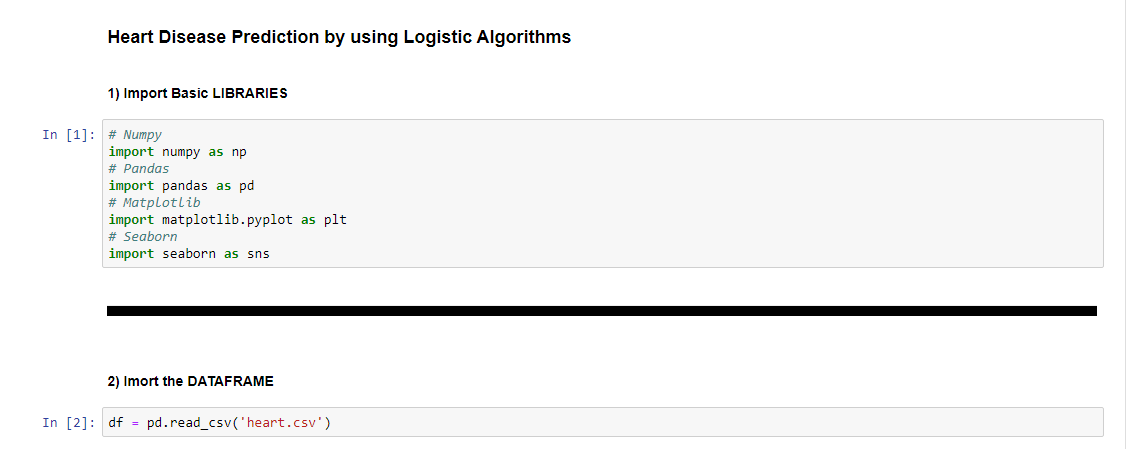
s

**CHAPTER 6: Result and Validations**

**Evaluation metrics of logistic regression model whens**

**Min Max scalar is applied**

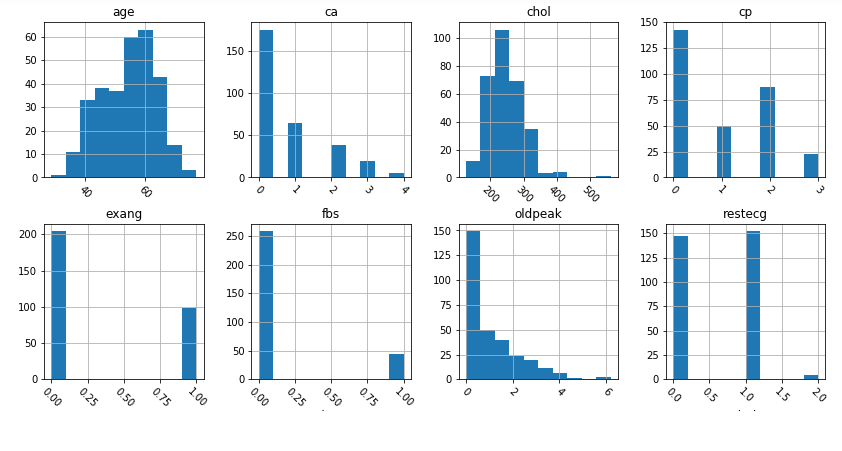
1. Data Extraction

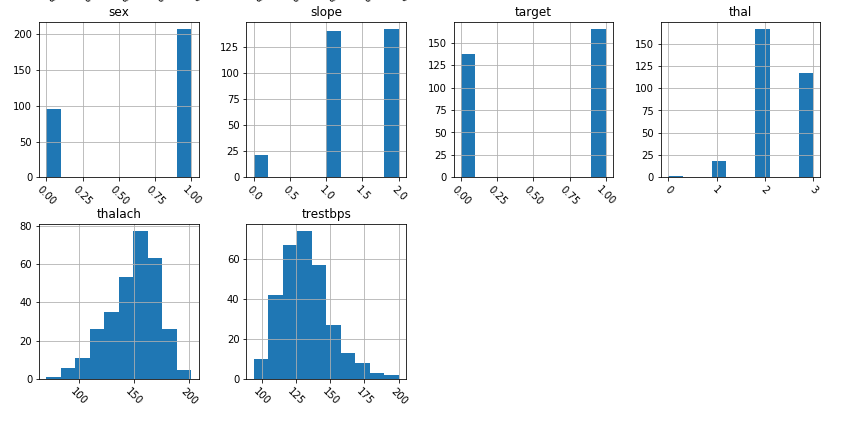


## EXPLORATORY DATA ANALYSIS

EDA is a phenomenon under data analysis used for gaining a better understanding of data. Exploratory Data Analysis or (EDA) understands the data sets by summarizing their main characteristics often plotting them visually. ... Through the process of EDA, we can ask to define the problem statement or definition on our data set which is very important.

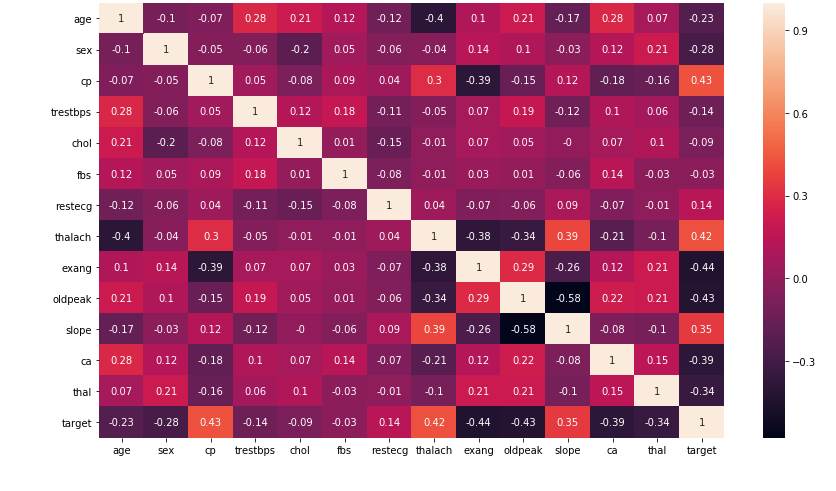
### Visualizing Histogram Grid

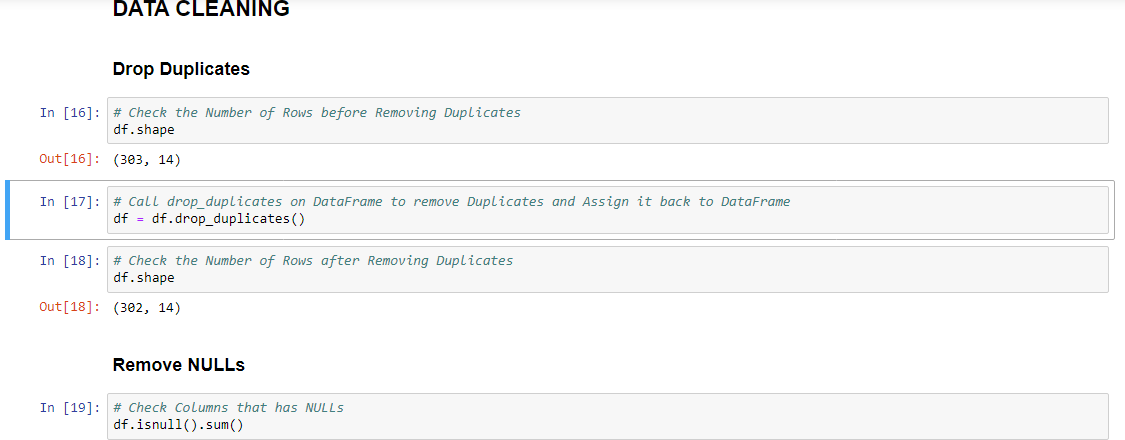




### Correlation

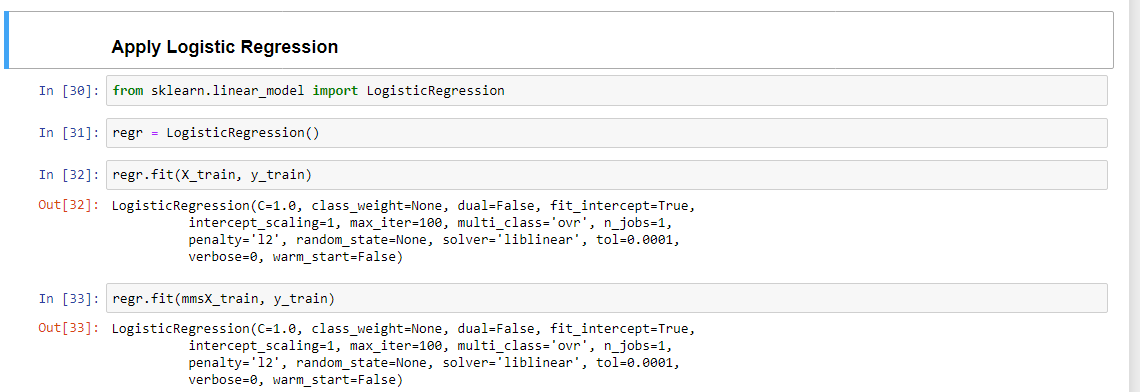
Numpy implements a corrcoef() function that returns a matrix of correlations of x with x, x with y, y with x and y with y. We're interested in the values of correlation of x with y (so position (1, 0) or (0, 1)).

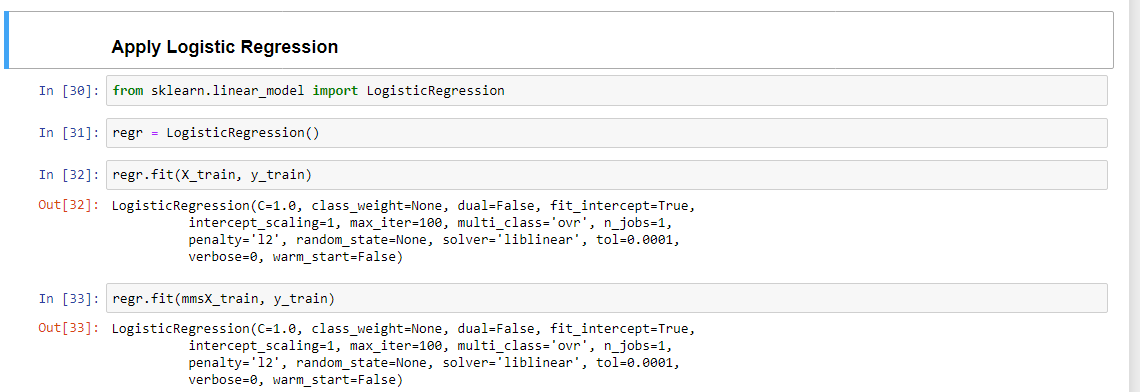




## CLASSIFICATION ALGORITHMS

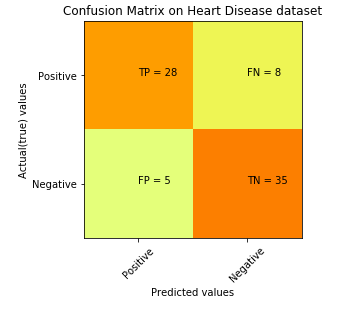
## Classification is technique to categorize our data into a desired and distinct number of classes where we can assign label to each class.

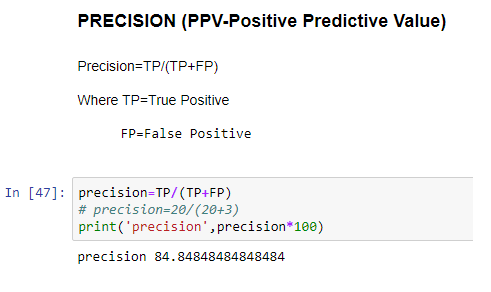


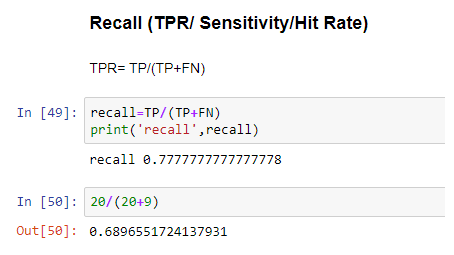


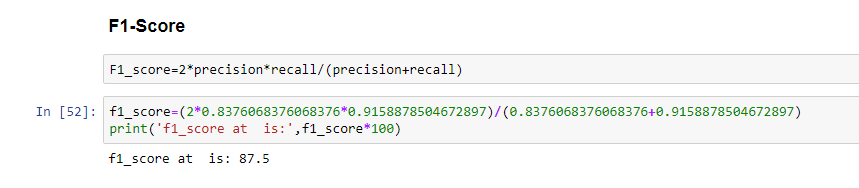
### Evaluation Metrics of the Model





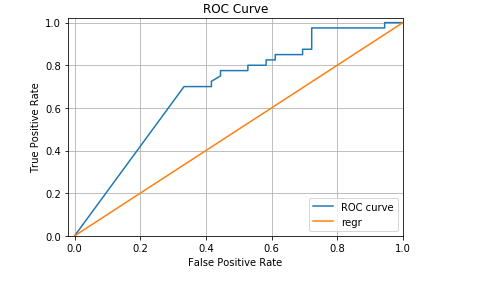






### ROC Curve (Receiver Operating Characteristics) curve

The **ROC curve** is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection in **machine learning**. ... The **ROC curve** is thus the sensitivity as a function of fall-out.



**CHAPTER 7: CONCLUSION AND FUTURE WORK**

The factors which may lead to **heart disease** are obesity, high cholesterol, smoking, increase in blood pressure, diabetes and other factors. Separate the data regarding the patients related to **heart diseases**. We train the data as per proposed **algorithm** of machine Learning by **using logistic regression**.

Experimental results determine the adequacy of the designed system with an achieved accuracy of **82.89 %** using the logistic regression classiﬁcation algorithm in **80%-20%.**

In future, the work can be extended and improved for the automation of heart analysis including some other machine learning algorithms.

**BIBLIOGRAPHY**

1. [www.google.co.in](http://www.google.co.in)

2. [www.pypi.python.org](http://www.pypi.python.org)

3. [www.anaconda.org](http://www.anaconda.org)

4. Data set

5. [www.stackoverflow.com](http://www.stackoverflow.com)

**REFERENCES**

1. Baldi, P. and Brunak, S. (2002). Bioinformatics: A Machine Learning Approach. Cambridge, MA: MIT Press.
2. Baldi, P., Frasconi, P., Smyth, P. (2003). Modeling the Internet and the Web - Probabilistic Methods and Algorithms. New York: Wiley.
3. Bishop, C. M. Neural Networks for Pattern Recognition. New York: Oxford University Press (1995).
4. Chakrabarti, S. (2003). Mining the Web, Morgan Kaufmann.
5. Cohen, P.R. (1995) [Empirical Methods in Artificial Intelligence](http://babs.cs.umass.edu/emai.html). Cambridge, MA: MIT Press.
6. Cowell, R.G., Dawid, A.P., Lauritzen, S.L., and Spiegelhalter,D.J. (1999). Graphical Models and Expert Systems.Berlin: Springer.
7. Vapnik, V. N. [The Nature of Statistical Learning Theory](https://books.google.com/books?id=EqgACAAAQBAJ&printsec=frontcover#v=snippet&q=%22empirical%20risk%20minimization%22%20OR%20%22structural%20risk%20minimization%22&f=false) (2nd Ed.), Springer Verlag, 2000.