I. INTRODUCTION

A. Project aim and objectives:

The main aim of this project is to predict heart disease using a machine learning algorithm and compare predicted data with the analysis tool. Heart disease is a major cause of death all over the world. The incidence of heart disease is increasing dramatically. Symptoms of heart disease differ from gender to gender. The most common heart disease symptoms are chest pain, nausea, indigestion, heartburn stomach ache, etc. [1]. In a traditional method, doctors can't detect heart disease instantly. they need a physical examination and medical image to know the current state of the patient. Moreover, it needs experience and comprehensive knowledge of prediction. [2]. Machine learning has been broadly used in the healthcare industry for detecting and predicting various diseases using different data models. As we know if early detection is possible, it can save a life, time, and money. Predicting heart disease is a classification problem. Hence, logistic regression has been selected, it will produce a better result for data analysis. Two popular data science platforms or tools Python and R will be used. Cleveland heart disease dataset will be collected, prepared, visualized, and analyzed for the logistic regression prediction model. The outcome of the prediction will lead to a comparative discussion between python and R data analysis tools.

B. Dataset:

The name of the Dataset is “Cleveland Heart disease” data set, it is collected from the UCI repository website (https://archive.ics.uci//ml/datasets/heart+disease) from an ongoing study on heart disease conducted by the Hungarian Institute of Cardiology and the University of Zurich. Dataset was collected from UCI because it is one of the largest and most reliable sources for collecting datasets. Students, Educators, and Researchers all over the world consider UCI or UC Irvine machine learning repository as a primary source of machine learning datasets. UCI is continuously updating its repository. Due to its authenticity and reliability, most people prefer UCI for collecting datasets.

This heart dataset consists of 303 separate data, 13 features, and one target variable. To visualize the dataset as a data frame, see the appendix Figure:1

C. Business Value of the Project:

Heart disease is one of the main causes of death all over the world [2]. It is a complex task for doctors to detect heart disease accurately and efficiently. This machine learning heart diseases prediction system has been developed to help doctors, medical administrators to detect heart disease. The analytical data will help to decrease the number of tests that need to be taken by the patient. This system can add value to the medical and health care industry.

D. Analytical Technique and Platforms:

The whole data preparation, exploration, visualization, and analysis has been carried out on both Python and R environments using the Logistic Regression machine learning technique. Predicting heart disease is an instance of supervised machine learning and it is also a classification problem. Hence, the Logistic regression supervised machine learning technique is appropriate for this project to predict heart disease. Python and R are excellent open-source programming languages for a data science project. Moreover, they have rich libraries and functionality for data preparation, visualization, and analysis. Machine learning techniques like logistic regression works perfectly on these platforms. Thus, it provides more accuracy than other techniques.