**INDEX**

**TITLE PAGE NO**

**CHAPTER 1: INTRODUCTION 1**

**1.1 Introduction 1**

**1.2 Objectives of Research 1**

**1.3 Problem Statement 1**

**CHAPTER 2:REVIEW OF LITERATURE 1**

**CHAPTER 3: DATA COLLECTION 2**

**CHAPTER 4: METHODOLOGY 3**

**4.1 Exploratory data analysis 3**

**4.1.1 Tables 4**

**4.2 Statistical Techniques and data visualization 5**

**4.3 Data modeling using supervised ML techniques 6**

**CHAPTER 5: Findings and Suggestions 7**

**CHAPTER 6: CONCLUSION 8**

**PREDICTION OF READMISSION TO HOSPITALS FOR HEART ATTACKS**

**1.INTRODUCTION**

**1.1. Introduction:-**

The Heart is one of the main organs of the human body. It pumps blood through the blood vessels of the circulatory system. The circulatory system is extremely important because it transports blood oxygen and other materials to the different organs of the body.  
Among various life-threatening diseases, heart disease has garnered a great deal of attention in medical research. The diagnosis of heart disease is a challenging task, which can offer automated prediction about the heart condition of patient so that further treatment can be made effective. The diagnosis of heart disease is usually based on signs, symptoms and physical examination of the patient.

A major challenge faced by health care organizations, such as hospitals and medical centres, is the provision of quality services at affordable costs.[1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5863635/) The quality service implies diagnosing patients properly and administering effective treatments. The available heart disease database consists of both numerical and categorical data. Before further processing, cleaning and filtering are applied on these records in order to filter the irrelevant data from the database.

**1.2. Objectives of research:-**

The main objective of this research is to develop a prototype health care prediction system using random forest algorithm.

The system can discover and extract hidden knowledge associated with disease database (heart attack) from a historical heart disease database.

It can answer complex queries for diagnosing disease and thus assist healthcare practitioners to make intelligent clinical decisions.

By providing effective treatments, it also helps to reduce treatment costs. To enhance visualization and ease of interpretation. It displays the results in tabular and .pdf forms.

**1.3. Problem statement:-**

Disease such as cancer can be detected and the stage can also be predicted by training dataset with pictures of cancer cells.Similarly heart disease can be predicted based on aspects like cholesterol, angina, heart rate etc.

Mitigating risk-of-readmission of Congestive Heart Failure (CHF) patients within 30 days of discharge is important because such readmissions are not only expensive but also critical indicator of provider care and quality of treatment. Accurately predicting the risk-of-readmission may allow hospitals to identify high-risk patients and eventually improve quality of care by identifying factors that contribute to such readmissions in many scenarios.

**2. Review of Literature:-**

There are numerous works has been done related to disease prediction systems using different data mining techniques and machine learning algorithms in medical centres.

K. Polaraju , proposed Prediction of Heart Disease using Multiple Regression Model and it proves that Multiple Linear Regression is appropriate for predicting heart disease chance. The work is performed using training data set consists of 3000 instances with 13 different attributes which has mentioned earlier. The data set is divided into two parts that is 70% of the data are used for training and 30% used for testing. Based on the results, it is clear that the classification accuracy of Regression algorithm is better compared to other algorithms.

Marjia, developed heart disease prediction using KStar, j48, SMO, and Bayes Net and Multilayer perception using WEKA software. Based on performance from different factor SMO and Bayes Net achieve optimum performance than KStar, Multilayer perception and J48 techniques using k-fold cross validation. The accuracy performances achieved by those algorithms are still not satisfactory. Therefore, the decision to diagnosis disease, S. Seema focuses on techniques that can predict chronic disease by mining the data containing in historical health records using Naïve Bayes, Decision tree, Support Vector Machine(SVM) and Artificial Neural Network(ANN). A comparative study is performed on classifiers to measure the better performance on an accurate rate. From this experiment, SVM gives highest accuracy rate, whereas for diabetes Naïve Bayes gives the highest accuracy. Ashok Kumar Dwivedi et al, [10] recommended different algorithms like Naive Bayes, Classification Tree, KNN, Logistic Regression, SVM and ANN. The Logistic Regression gives better accuracy compared to other algorithms.

**3. Data Collection:-**

The data are collected from a standard dataset that contains 303 records. The 15 parameters, such as age, sex, chest pain type (CP), and cholesterol (chol), with some domain values associated with them, considered to predict the probability of heart disease.

The collected data were used to create a structured database system. The pre-processing was done by identifying the associated fields and removing all the duplications. After that, all the missing values were filled, and the data were coded according to the domain value.

After applying neural networks on training dataset, the results show that there are zero FN or FP entries suggesting that the system predicts heart disease with 100% accuracy.

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

One file has been "processed", that one containing the Cleveland database. All four unprocessed files also exist in this directory.

Attribute Information:

Only 13 attributes used:

1. Age: age of the patient in years

2. Sex: gender of the patient male:0 female:1

3. Cp : Chest Pain rated in between 0 to 5

4. trestbps : resting blood pressure in mmHg on admission to hospital

5. chol : cholesterol in mmHg

6. fbs : fasting blood sugar( >120) true=1 false=0

7. restecg: resting ECG results

8. thalach: maximum heart rate acheived

9. exang: exercise induced angina true=1 false=0

10. oldpeak :ST depression induced by exercise relative to rest

11. slope : slope of peak exercise ST segment

12. ca: calcium level and the number of major vessels (0-3) colored by fluoroscopy

13. thal : thalassemia levels normal=3 fixed defect=6 reversible defect=7

The variables chosen to control for patient demographic and illness severity were gender, age, race, admission source, discharge disposition, primary diagnosis medical specialty of the admitting physician, and time spent in hospital. Values of these variables and their distribution in the dataset.

1. Number inpatient: Number of inpatient visits of the patient

2. Time in hospital: number of days between admission and discharge

3. Number of diagnoses: diagnosis of heart disease (angiographic disease status)

-- Value 0: < 50% diameter narrowing

-- Value 1: > 50% diameter narrowing

4. Readmitted: Days to inpatient readmission. values: “NO” if the patient was readmitted in more than 30 days, and “NO” for no record of readmission.”YES” if the patient was readmitted in less than 30 days.

**4. Methodology:-**

The 30-day readmission rate is defined as the number of AMI admissions (as defined above) for which there was at least one subsequent hospital admission within 30 days divided by the total number of AMI admissions between January and November. That is, when a patient is discharged from the hospital with a principal diagnosis of AMI, they are followed for 30 days in the data. If any readmission to the same or different hospital occurs during this time period, the admission is counted as having a readmission.

If a patient was transferred to a different hospital on the same day or was transferred within the same hospital, the two events were combined as a single stay and the second event was not counted as a readmission.

That is, transfers were not considered a readmission. In the case of AMI admissions for which there was more than one readmission in the 30-day period, the data presented in this Statistical Brief reflect the characteristics, costs, and associated diagnoses of the first readmission.

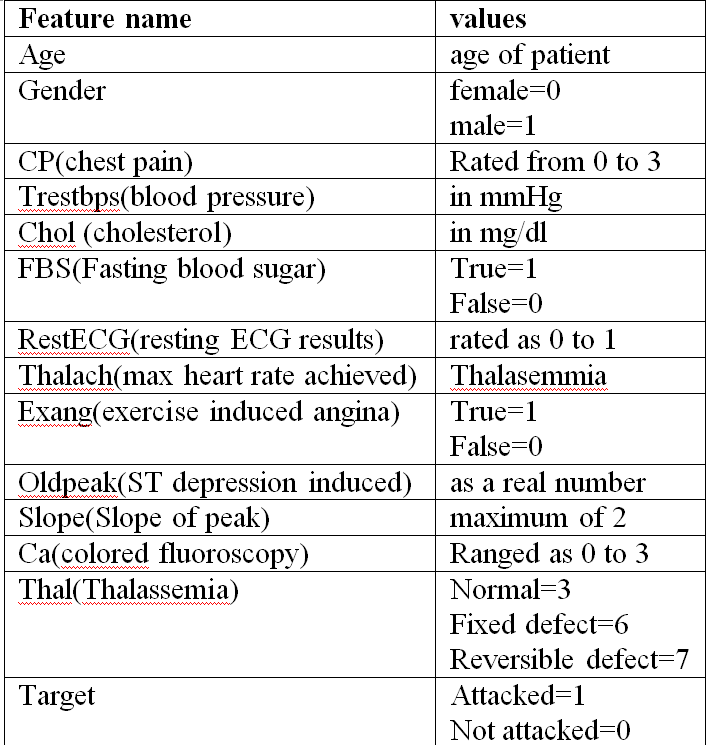
**4.1 Exploratory data analysis:-**

Preliminary Analysis and the Final Dataset: The original database contains incomplete, redundant, and noisy information as expected in any real-world data. There were several features that could not be treated directly since they had a high percentage of missing values. Weight attribute was considered to be too sparse and it was not included in further analysis. Old Peak was removed since it had a high percentage of missing values and it was not considered relevant to the outcome. Medical specialty attribute was maintained, adding the value “missing” in order to account for missing values. Large percentage of missing values of the weight attribute can be explained by the fact that prior to a structured format.

The preliminary dataset contained multiple inpatient visits for some patients and the observations could not be considered as statistically independent, an assumption of the logistic regression model. We thus used only one encounter per patient; in particular, we considered only the first encounter for each patient as the primary admission and determined whether or not they were readmitted within 30 days.

To summarize, our dataset consists of hospital admissions of length between one and 14 days that did not result in a patient death or discharge to a hospice. Each encounter corresponds to a unique patient diagnosed with heart attack, although the primary diagnosis may be different. During each of the analyzed encounters, lab tests were ordered and medication was administered.

**4.1.1 Tables:-**

****

**4.2 Statistical Techniques and Data Visualization:-**

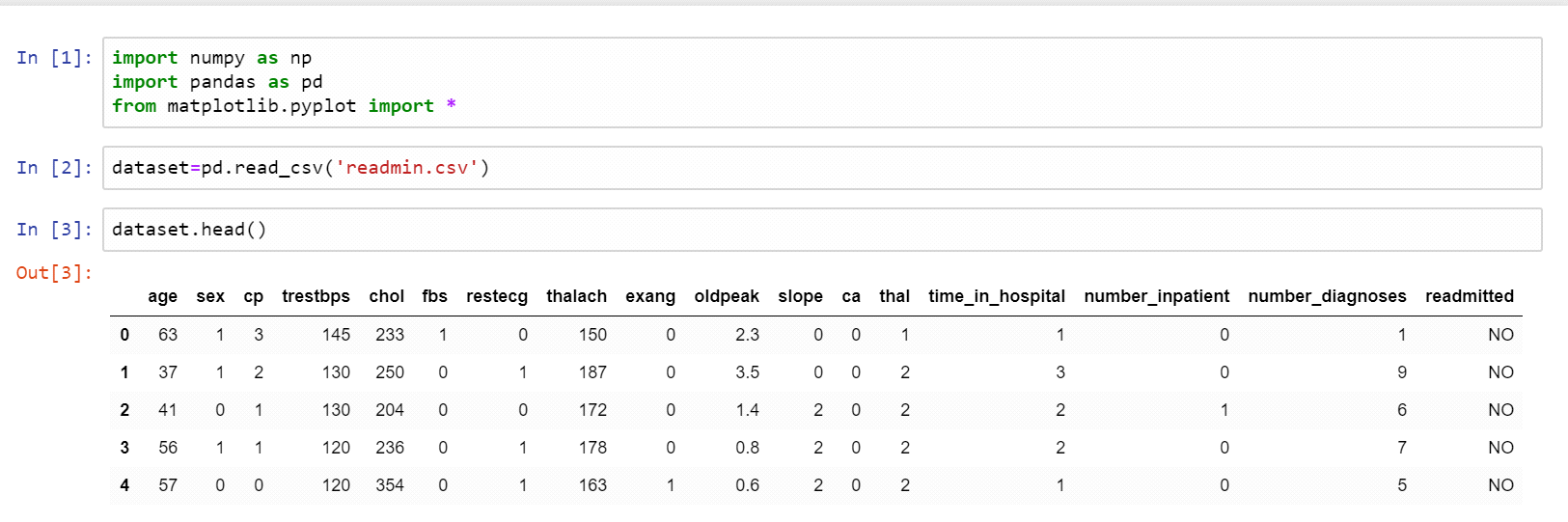
**Statistical Methods:**

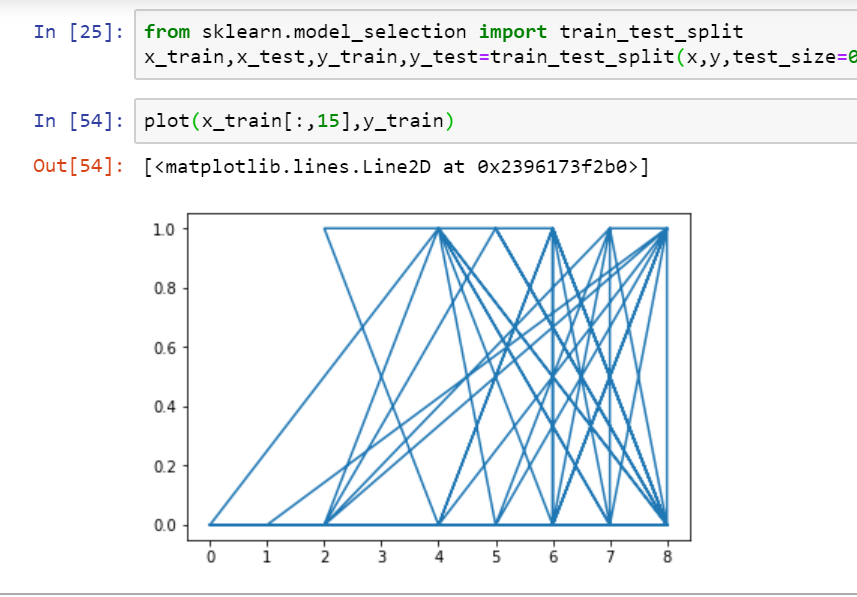
The unit of our analysis is an encounter; however, in order to keep the observations independent, we only analyzed one encounter per patient. After preliminary analysis and taking into account the amount of data, the significance level was determined by a 𝑃 value of less than 0.01.

To assess whether the candidate covariates were significantly associated with readmission, we created the model in four steps. Each step was followed by tests for significance of variables with higher degree of freedom, an analysis of deviance table, and sensitivity analysis which was done by removing one variable at the time and looking at changes.

Considering the prediction of re admission into hospitals with a heart attack, it comes under the supervised machine learning model. And this comes into a categorical supervised ML. Here the readmission is not a continuous value to show the probability of prediction, but a classification type of supervised ML. The classes or the categories on which prediction is done are readmission and no readmission. The classification algorithms are Logistic Regression, SVM, Random Forest, Decision Tree and KNN.

**Data Visualization:**



****

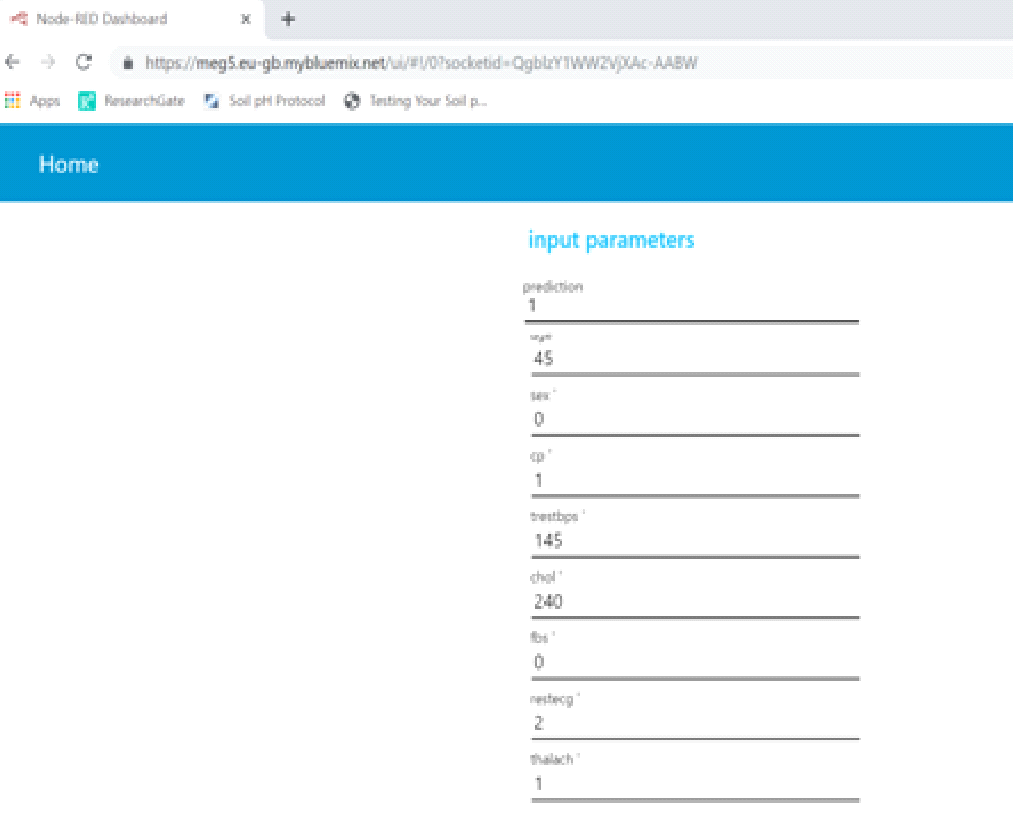
**4.3 Data modelling using Supervised Machine Learning Techniques:-**

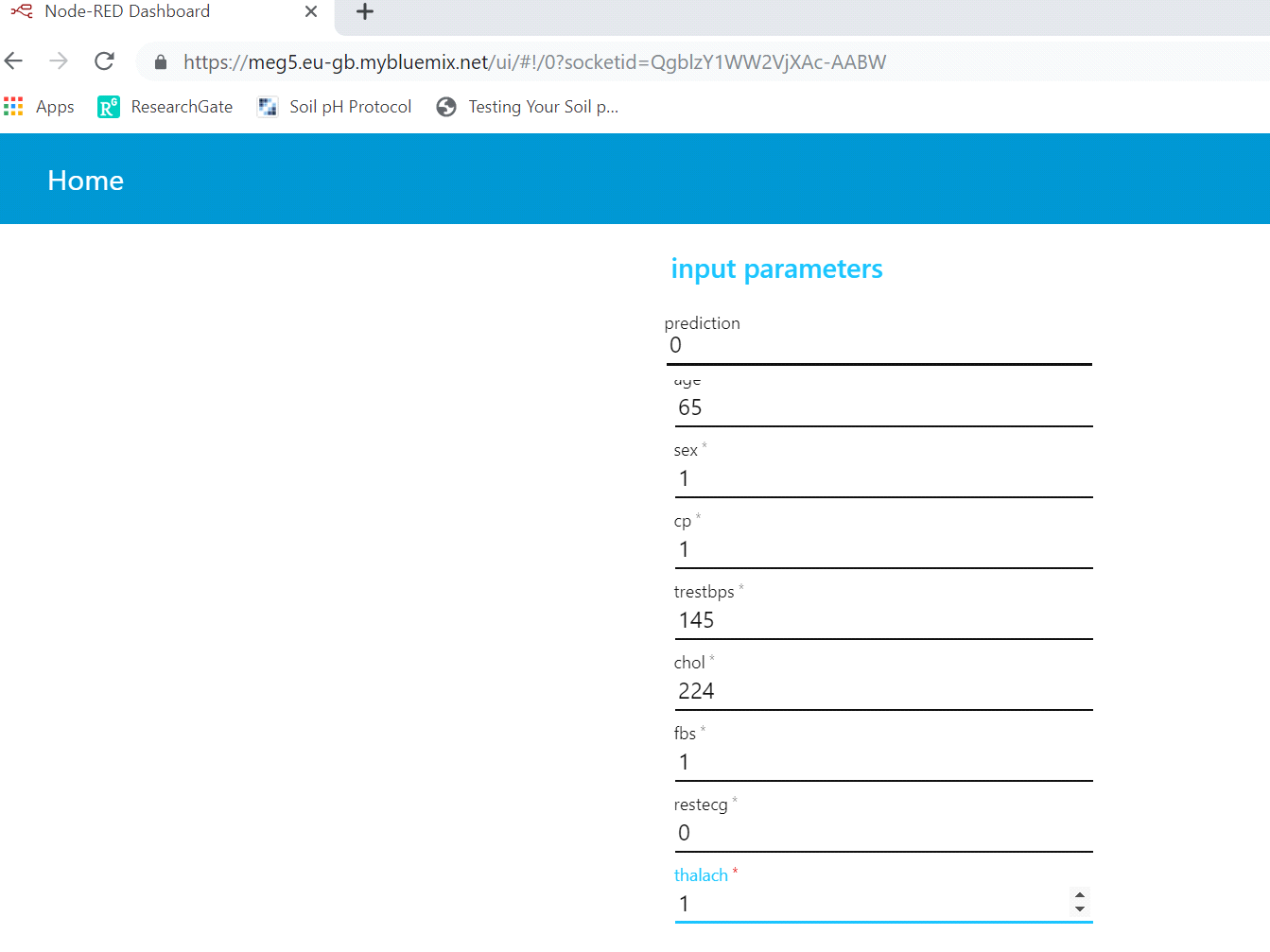
In this, we made a detailed investigation on prediction accuracy rate of heart diseases using different supervised machine learning techniques, which will pave the way for researchers to choose the efficient technique(s) in order to design and develop clinical decision support systems that predicts the occurrence of Heart attack.

Machine learning techniques integrate a variety of supervised machine learning techniques and algorithms that helps to develop Decision Support Systems (DSS) in the healthcare domain that involve large datasets and variables. These techniques and algorithms are used to predict the possibility of heart diseases in healthy people with the help of supporting documents such as electronically generated health records and history of heart disease patients obtained from collected datasets. Supervised machine learning techniques include classification and regression algorithms such as Support Vector Machines (SVM),Random Forest, K-Nearest Neighbour, Linear Regression, Support Vector Regression, Decision Trees, Ensemble Methods and Neural Networks

**5. Findings and Suggestions:-**

* <https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c>
* <https://www.mayoclinic.org/diseases-conditions/heart-attack/symptoms-causes/syc-20373106>
* <https://www.the-hospitalist.org/hospitalist/article/121417/predicting-30-day-readmissions>





**6. Conclusion:-**

In conclusion, the decision to obtain a prediction of readmitting heart attack patient is a useful predictor in the development of strategies to reduce readmission rates and costs for the care of individuals with heart attacks. For instance, our analysis showed that the profile of readmission differed significantly in patients where the setting of a primary diagnosis, when compared to those with a primary circulatory disorder. While readmission rates remained the highest for patients with circulatory diagnoses, readmission rates for patients with heart attacks appeared to be associated with the decision.