**Symbiosis Skills and Professional University**

# Kiwale, Pune

**PROJECT REPORT**

**On**

### **“Heart Disease Prediction”**



**Submitted by**

Nikhilraj Rindhe

**( 2001207064 )**

# DA-Batch-II

**Under the Guidance of**

Mr. Kushal Sharma

# STUDENT DECLARATION AND ATTESTATION BY TRAINER

This is to declare that this report has been written by me. No part of the report is plagiarized from other sources. All information included from other sources have been duly acknowledged. I aver that if any part of the report is found to be plagiarized, I shall take full responsibility for it.

Signature of student

**Name of student:** Nikhilraj Rindhe

Registration Number: 2001207064

Signature of trainer

**Name of trainer: Mr. Kushal Sharma**

**CERTIFICATE**

This is to certify that the report entitled, “**Heart Disease Prediction**” submitted by “Nikhilraj Rindhe” to Symbiosis Skills and Professional University, Pune, Maharashtra, India, is a record of bonafide. Project work carried out by him under my supervision and guidance and is worthy of consideration for the completion of certificate course in “Data Associate”.

Signature of Trainer

Name of Trainer: Mr. Kushal Sharma Date: 29/ 04/ 2021

Supervisor Supervisor

Date:

# ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to our Mentor Mr. Kushal Sharma for his guidance and motivation throughout the course. I am able to complete and present this project today, because of his tenacity and deadlines. Thank you, sir, for pushing me to my limits every day.

Special thanks to our ‘People’s Man’ who kept us motivated throughout the journey, our ‘Action man’ who accepted every new challenge and experimented every possibility that led to solution, and our ‘Idea man’ for saving the day.

I would like to thank everybody who mentored and motivated us throughout the journey.

**I NDEX**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SR. No.** | **Index** |  |  |  | **Page No.** |
| 1 | **Introduction** | |  |  |  |
|  | 1.1 | Problem Definition | |  | 6 |
|  | 1.2 | Motivation | |  | 6 |
|  | 1.3 | Objectives | |  | 7 |
| 2 | **Related Works** | | |  | 8 |
| 3 | **Methodology** | | |  |  |
|  | 3.1 | Dataset Description | |  | 9-10 |
|  | 3.2 | Algorithms used | |  | 11-12 |
|  | 3.3 | Evaluation Method | |  | 12 |
| 4 | **Proposed Solution** | | |  |  |
|  | 4.1 | Architecture | |  | 13 |
|  | 4.2 | Implementation | |  | 14-15 |
|  | 4.3 | Results and Discussion | |  | 16 |
| 5 | **Future Scope** | | |  |  |
|  | 5.1 | Automation | | | 17 |
|  | 5.2 | Generalized Models for Various Diseases | | | 17 |
| 6 | Learning from the Project | | |  | 18 |
| 7 | References | | |  | 19 |

# CHAPTER 1: INTRODUCTION

Heart Disease is one of the leading diseases with high morbidity and mortality rates. Even in 2020, when the World suffered from the deadliest Covid-19 virus. Still, according to the study published in January 2021, 397,042 cardiovascular deaths were reported in US only [8]. Cardiovascular disease (CVD) are disorders related to heart and blood vessels. One of the most common CVD is Coronary Heart Disease (CHD) where blood vessels to the heart are damaged. Heart Attacks are caused by blockage in the blood vessels due to deposition of calcium or any other masses making the blood vessel narrower.

## **Problem Definition**

The major challenge in heart disease is detection. There are methods that have high accuracy towards finding the location and severity of CHD but are really expensive. The best method available for detection is Coronary Angiogram. However, the problem lies in the type of method. Angiogram is an invasive method i.e.; you have to drill a hole in your body to get a diagnosis. There are other non-invasive methods but they are only 30-70% efficient.

So, we need a non-invasive method with better results. In the era of InfoTech and Bio-technology, we can use data to make Machine Learning models which can draw insights beyond human reasoning. These data driven tools along with non-invasive methods can be used to draw an early diagnosis. The earlier we are able to predict the possibility of having heart disease, the better healthcare can be provided and eventually the mortality rates would decrease.

## **Motivation**

Machine Learning has emerged as one of the most promising fields. For the past decade, Machine Learning has proved its potential by making huge contributions to the development of mankind.

Using Machine Learning, we can make the machine learn by its own experiences. With that, the machine can analyze and try to find relations between features that are strongly related to possibility of having heart disease. Along with Machine Learning, Big Data Tools can be used to make the solution scalable and more efficient. Using Big data tools, you can process huge quantity and variety of data in a fraction of seconds. With the same goal in mind, we designed a system that can accurately predict the possibility of having heart disease.

## **Objectives**

The main objectives of developing this project are:

* + 1. To develop machine learning model to predict possibility of heart disease with various classification models.

1.a. To check the reliability of Big Data tools and implement a proper model for the proposed solution.

* + 1. To determine the significant features based on patient report which strongly relate to possibility of having heart disease.

The flow of the paper is as follows:

Chapter 2 is a summarized section for past related works performed in the field.

Chapter 3 discusses the methodology used for performing the project. It gives a brief idea about the algorithms and evaluation metrics used in the project

Chapter 4 is a detailed discussion on the proposed solution. It summarizes the complete architecture and design specification for the project

Chapter 5 is an overview about the past chapter and proposes future scope of the project.

# CHAPTER 2: RELATED WORKS

With the contributions from researchers and developers to the medical field, these are the most significant works done in the field:

Marjia et al, [1] proposed a heart disease prediction system using KStar, J48, SMO, and Bayes Net and Multilayer perceptron using WEKA software. Based on performance SMO and Bayes Net gave the better accuracy than other algorithms.

M. Marimuthu et al,[2] In an experiment carried out using Cleveland dataset for heart diseases they concluded that Naïve Bayes and Random Forest gave the optimum accuracy of 91%.

S. Seema et al,[3] used different techniques to predict chronic diseases by processing the historical health records using Naïve Bayes, Decision tree, Support Vector Machine (SVM) and Artificial Neural Network (ANN).

Ashok Kumar Dwivedi et al, [4] performed a competitive study using Logistic Regression, Naive Bayes, Classification Trees, KNNs, SVM and ANN. The Logistic Regression gives better accuracy compared to other algorithms.

Using the UCI dataset, experiments were carried out using 5 supervised machine learning models were trained. [6]

# CHAPTER 3: METHODOLOGY:

# 3.1 Dataset Description:

The dataset is publicly available on the UCI Repository [5], which is a famous dataset for research and academia projects. It provides patient information which includes over 303 records and 14 attributes. The attributes include: age, sex, cp, trestbps, cholestrol, fbs, restecg, thalach, exang,

oldpeak, slope, ca, thal, and pred ranging from 0 to 4, where 0 is absence of heart disease. The features are explained well in Towards Data Science article [7].

The features are explained below:

1. Age: displays the age of the individual.
2. Sex: displays the gender of the individual using the following format:

1 = male

0 = female

1. Chest-pain type: displays the type of chest-pain experienced by the individual using the following format:

1 = typical angina

2 = atypical angina

3 = non — anginal pain

4 = asymptotic

1. Resting Blood Pressure: displays the resting blood pressure value of an individual in mmHg (unit).
2. Serum Cholesterol: displays the serum cholesterol in mg/dl (unit)
3. Fasting Blood Sugar: compares the fasting blood sugar value of an individual with 120mg/dl.

If fasting blood sugar > 120mg/dl

then: 1 (true)

else: 0 (false)

1. Resting ECG: displays resting electrocardiographic results

0 = normal

1 = having ST-T wave abnormality

2 = left ventricular hyperthrophy

1. Max heart rate achieved: displays the max heart rate achieved by an individual.
2. Exercise induced angina:

1 = yes

0 = no

1. ST depression induced by exercise relative to rest: displays the value which is an integer or float.
2. Peak exercise ST segment:

1 = upsloping

2 = flat

3 = down sloping

1. Number of major vessels (0–3) coloured by fluoroscopy: displays the value as integer or float.
2. Thal: displays the thalassemia:

3 = normal

6 = fixed defect

7 = reversible defect

1. Diagnosis of heart disease: Displays whether the individual is suffering from heart disease or not:

0 = absence

1, 2, 3, 4 = present.

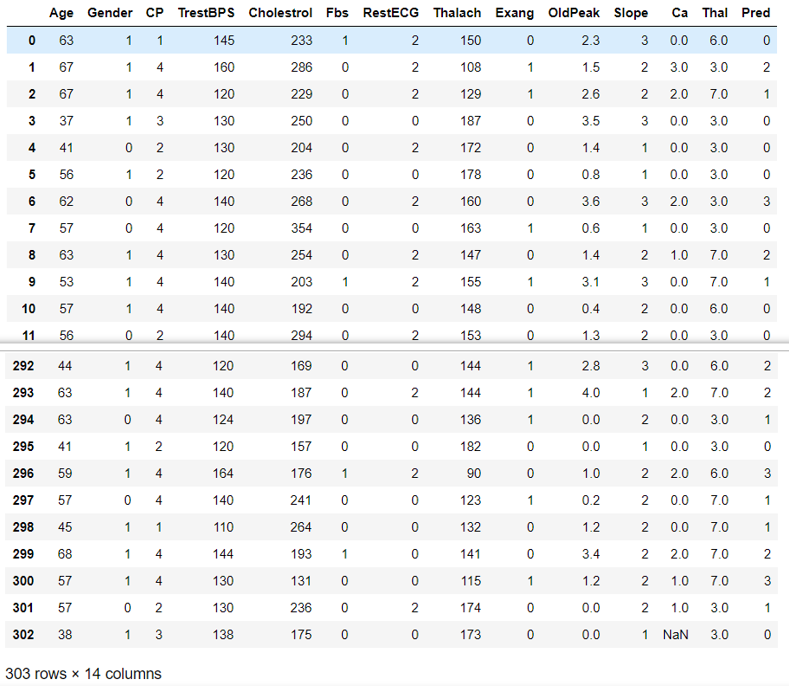


Figure 1 Cleveland Dataset

The data set is in csv (Comma Separated Value) format which is further prepared to data frame as supported by Pandas library in python. Fig.1. shows the dataset in using Pandas Dataframe.

This data is directly loaded in databases (MySQL and HDFS) for Machine Learning using Apache Spark.

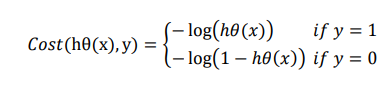
# 3.2 Algorithms used:

We are using Apache Spark ML library for the Classification Models. We are using 5 algorithms.

1. Logistic Regression:

Logistic Regression is a supervised classification algorithm. It is a predictive analysis algorithm based on the concept of probability. It measures the relationship between the dependent variable (Pred) and the one or more independent variables (features) by estimating probabilities using underlying logistic function (sigmoid function). Sigmoid function is used as a function to limit the hypothesis of logistic regression between 0 and 1 (squashing) i.e., 0 ≤ hθ (x) ≤ 1.

In logistic regression cost function is defined as:



Logistic Regression relies highly on the proper presentation of data. The dependent variable must be categorical in nature. The independent variable should not have multi-collinearity.

1. Support Vector Machines (SVM):

Support Vector Machine or SVM is a Supervised Learning algorithm, which is used primarily for Classification. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM is different from other perceptron-based Classification models. It takes decision by looking at the extreme cases i.e., support vectors.

1. Naïve Bayes:

Naïve Bayes algorithm is a supervised learning algorithm, which is based on**Bayes theorem. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.**

  It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Each feature individually contributes to identify that it is a class without depending on each other.

The formula for Bayes Theorem:



**Where, P(A|B) is Posterior probability**: Probability of hypothesis A on the observed event B, **P(B|A) is Likelihood probability**: Probability of the evidence given that the probability of a hypothesis is true, **P(A) is Prior Probability**: Probability of hypothesis before observing the evidence, **P(B) is Marginal Probability**: Probability of Evidence.

1. Decision Trees:

It is a tree-structured classifier, where**internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.** **Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.**

There are two popular techniques for Attribute Selection Method (ASM), which are:

* Information Gain:

Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.

Information Gain= Entropy(S)- [(Weighted Avg) \* Entropy (each feature)]

Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no)

A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first.

* Gini Index:

Gini index is a measure of impurity or purity used while creating a decision tree in the CART (Classification and Regression Tree) algorithm. An attribute with the low Gini index is preferred over a high Gini index to build tree.

Gini Index= 1- ∑j(Pj2) where, Pj is probability of classes.

1. Random Forest:

Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

# 3.3 Evaluation Method:

For the evaluation of our output from our training the data, the accuracy was analyzed “Confusion matrix”.

A confusion matrix, also known as an error matrix, is a table that is often used to describe the performance of a classification model (or “classifier”) on a set of test data for which the true values are known. It allows easy identification of confusion between classes.

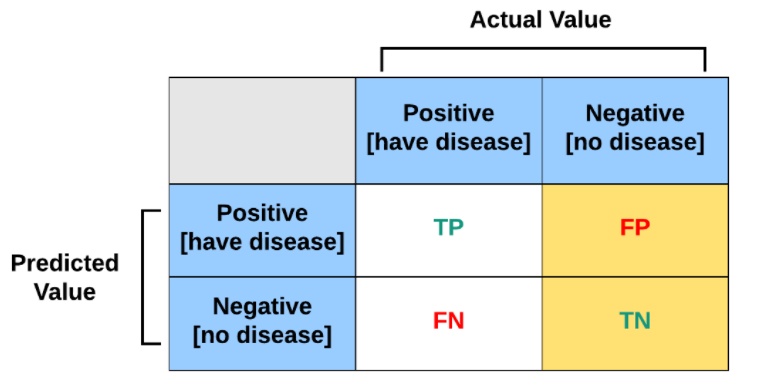


Figure 2 Confusion Matrix

We are using Accuracy Score as the evaluation method, which can be calculated as:

Accuracy = TP + TN/ (TP +TN+FN +FP)

# CHAPTER 4: Proposed Solution

# 4.1 Architecture:

The Dataflow is as follows:

1. Our Data Sources are Patient Reports which will have all the features as per Cleveland Dataset []. We will collect the data using Google Forms, which will save data in Excel Workbooks (CSV files).
2. These csv files will be loaded to MySQL Database for Data Querying.
3. Later this data is exported to Apache Hadoop (Hadoop Distributed File Storage) using Sqoop (ETL tool).
4. Apache Spark will access HDFS database for Data Analysis and store the results back in HDFS
5. Data is then used for Reporting using Tableau/Power BI.

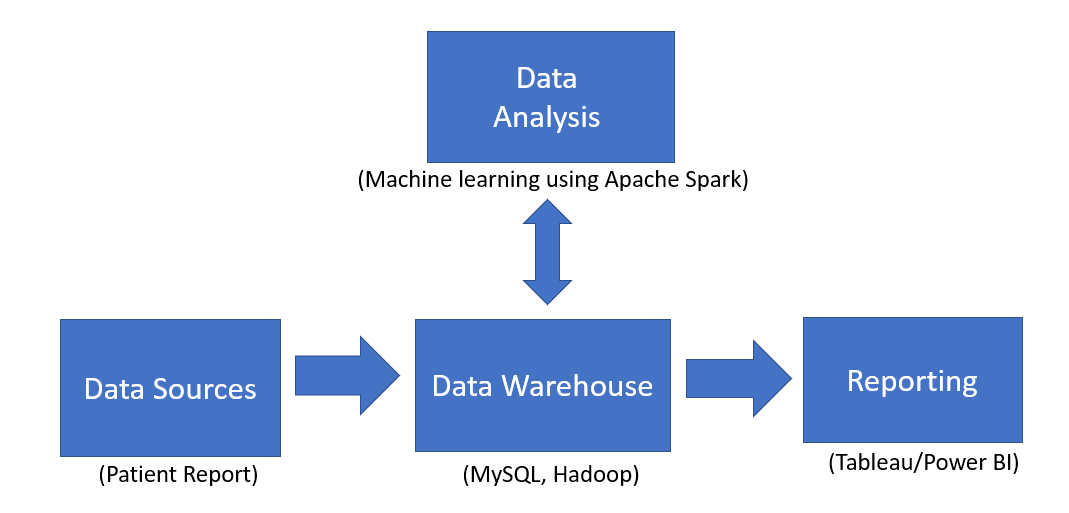


Figure 3 Dataflow of Project

The dataflow was designed to keep up with 6Vs Big Data to make it more scalable and robust.

# 4.2 Implementation:

The implementation details are shared according to the dataflow mentioned in section 3.1.

1. Data Sources:

With the help of Google Forms we are collecting the data in CSV files. The CSV files are then uploaded to MySQL database.

1. MySQL database Description:

We have created a database named “HeartA”, which contains table “cleveland” which has Cleveland Dataset [3] mentioned in section 3.1. The data is loaded with the CSV file mentioned in the MySQL Source file (.sql).

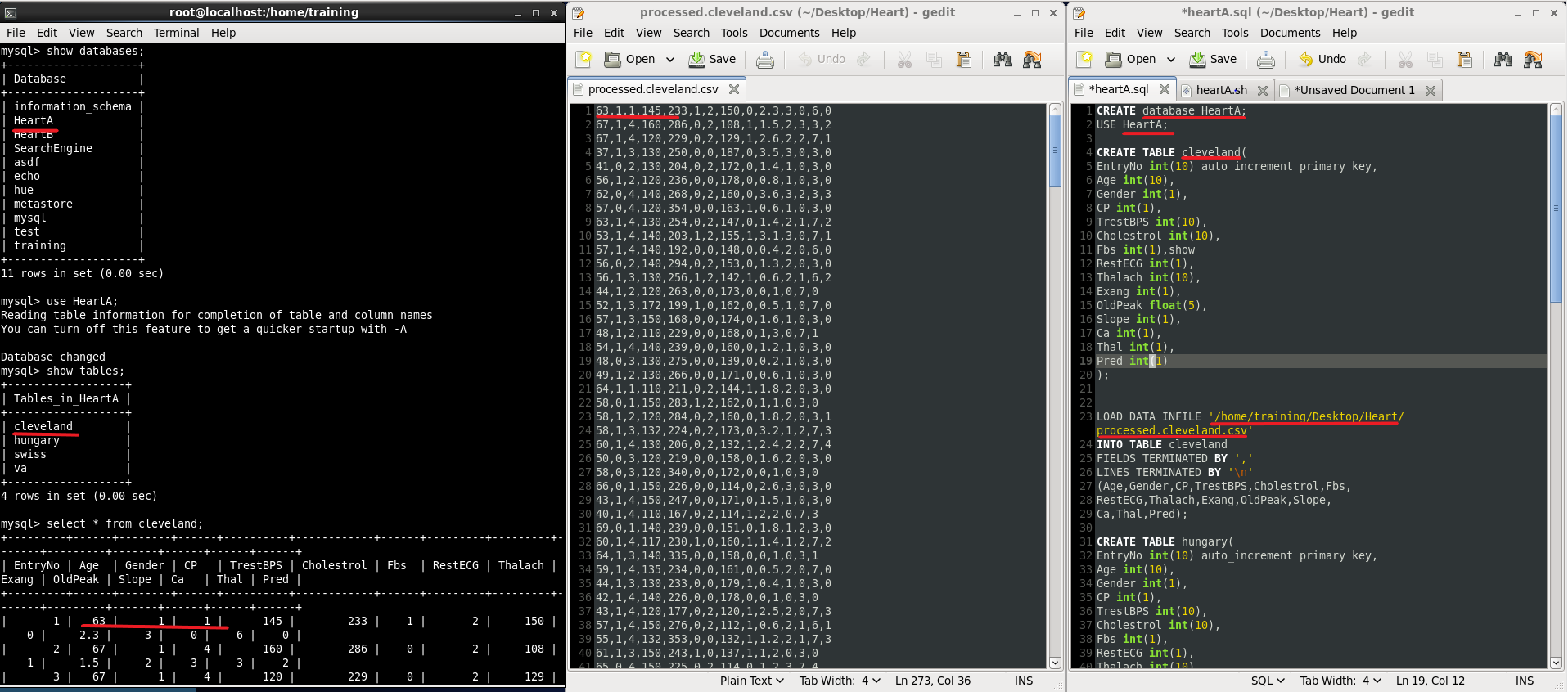


Figure 4 MySQL database and source files

1. MySQL to HDFS using Sqoop ETL tool:

With the use of Sqoop, I have transferred the cleveland table from MySQL to HDFS at /HeartA/.

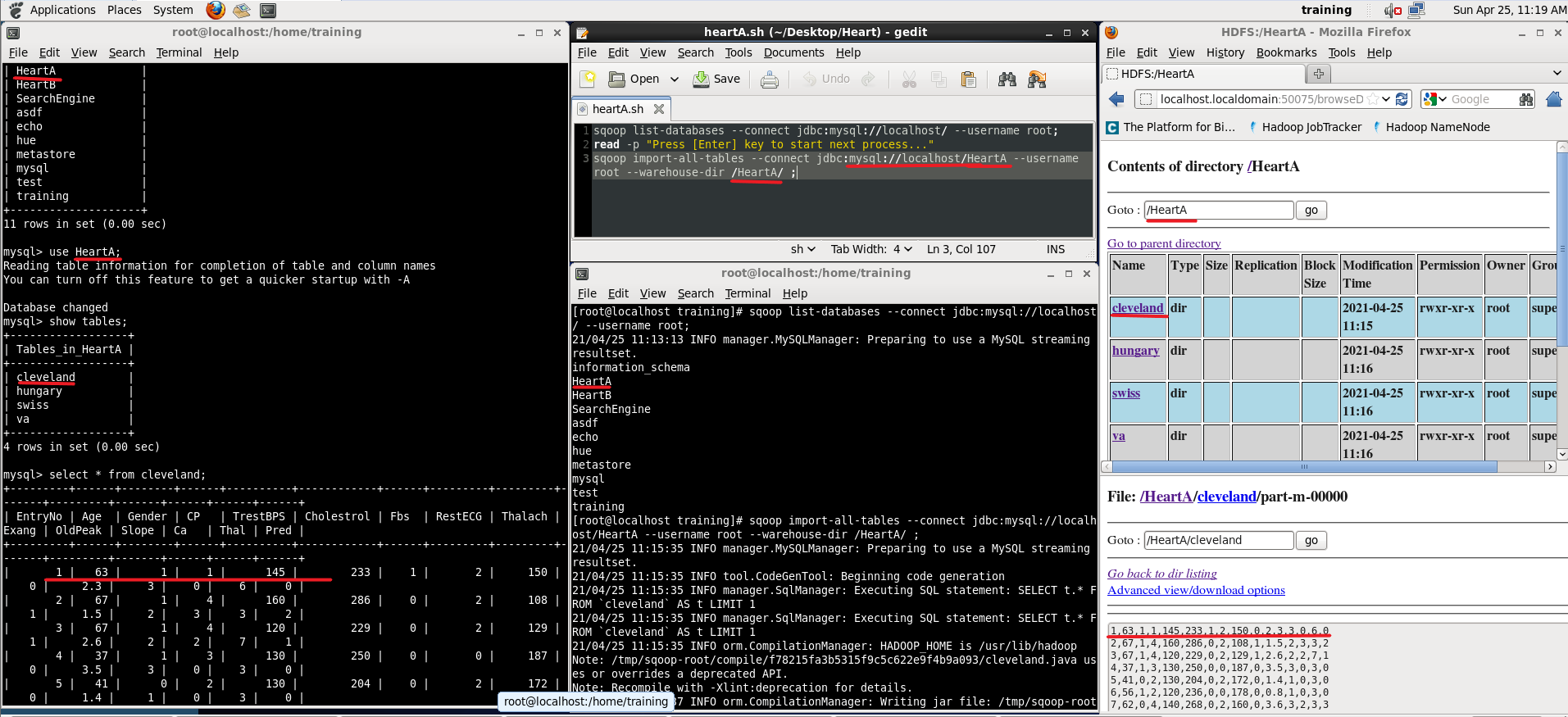


Figure 5 MySQL to HDFS using Sqoop

1. Machine Learning:

With Apache Spark ML library, we created a Classification Model using Logistic Regression Algorithm. By splitting the dataset in 70-30 ratio, we trained our Model with 70% of Cleveland Dataset and the rest was used for testing our model. For every patient record, our model will try to predict the possibility of having heart disease.

The code files are in GitHub Repository.

1. Reporting:

With the help of Business Intelligence Tools like Tableau and Power BI, we created interactive dashboards which help us visualize the data and draw insights from it.

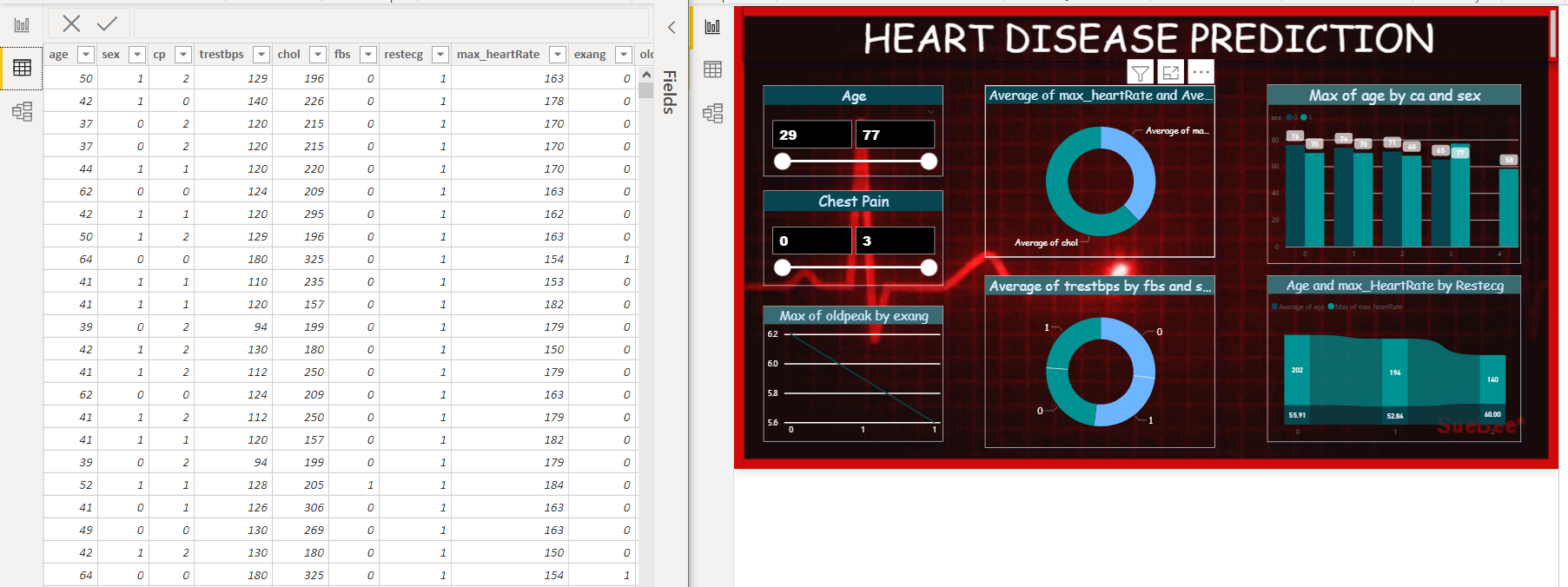


Figure 6 Power BI Dashboard

Tableau Dashboard can be accessed at Tableau Public website.

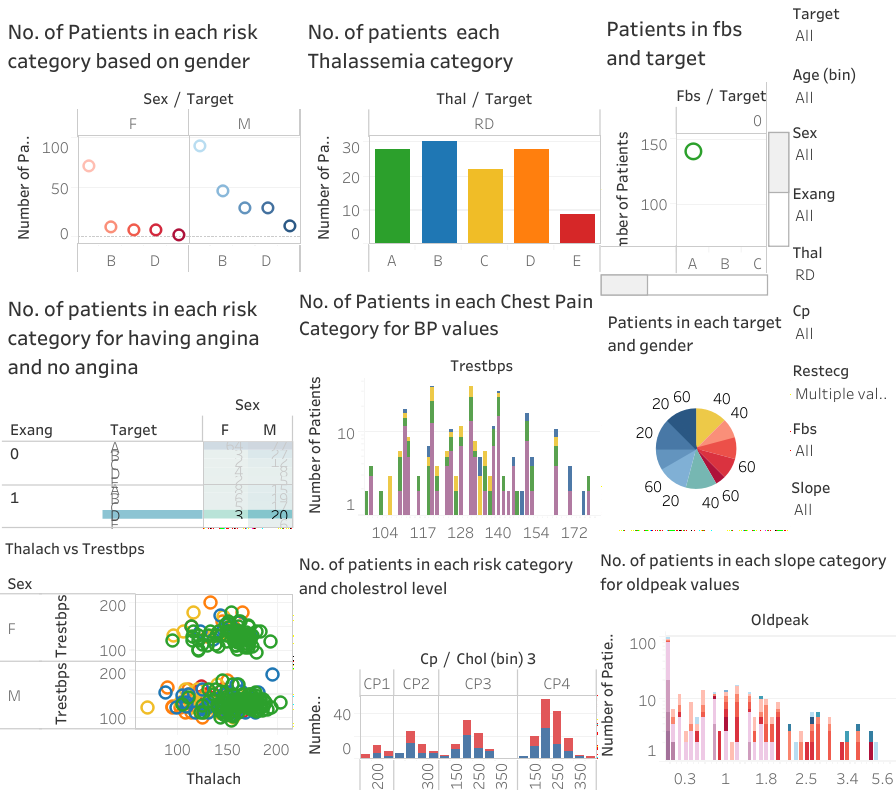


Figure 7 Tableau Dashboard

# 4.3 Results and Discussion:

After Training the Models for each Classification Model, we calculated their Accuracy Scores.

1. Logistic Regression:

Logistic Regression was able to perform with a accuracy of 86% and 80% with the training set and test set respectively.

1. SVM:

SVM produced a accuracy score of 92% for training set and 80% for test set.

1. Naïve Bayes:

Naïve Bayes algorithm was 86% accurate for training set and 78% for test set.

1. Decision Trees:

Decision Trees were 100% accurate for training set and 78% for test set.

1. Random Forest:

Random Forest was 99% accurate for training set and 78 % for test set.

For the Training Set, the Best Classification model was Decision Tree (100% Accuracy)

but it was not able to perform well for Test Set. So, we can conclude that this model was overfitted.

SVM was able to perform well for Test Set (80% Accuracy) with 92% accuracy for Training Set. This makes it the Best Fit model for the data.

Along with the Confusion Matrix, another Evaluation Parameter is available in Apache Spark ML lib called Binary Classification Evaluator. The results for Logistic Regression are:

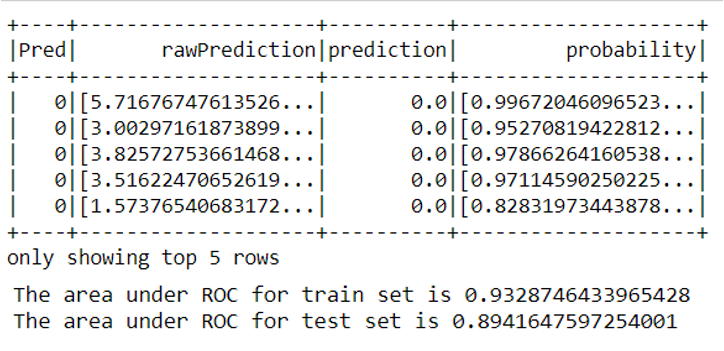


Figure 8 Logistic Regression Results

The average accuracy for all models can be assumed as 80%. If we assume, that the model is only 80% accurate. That means out of 5 people, we were able to correctly for 4 persons. With these results, we can deploy this as a Pre-Diagnosis Tool for Heart Disease. With increasing number of records and increase in Quality of data that is analyzed, accuracy of the model will also increase.

The 1 person out of 5, that we are wrongly predicting. This is an important part of the project. As this wrongly predicted person, has to be an outlier. In more medical terms, we can say that this person is an exception. Let’s take an example, if this person is suffering from any disease other than/along with heart disease, this patient’s data will deviate from the normal range of features. And as a result, model will predict wrong. These cases are important to find out as they are to be reviewed by the Doctors to be double assured.

# CHAPTER 5: FUTURE SCOPE:

There are two major ideas that we wanted to work on before completion of this project. Those are as follows:

# 5.1 Automation:

All the instructions were executed manually, though most part of it was done with script files and source files. But we still felt the need to automate it to have very less or no human requirement for the operation of the project.

The areas where we could automate the system are as follows:

1. CSV to MySQL files: With the use of MySQL Scheduler, we can schedule importing new data out of CSV file which gets new entries per submit of Patient Report. This new data can be imported at regular intervals.
2. MySQL to HDFS: This can be easily done by executing Sqoop Jobs with any scheduler.
3. HDFS to MySQL: With every update in HDFS storage, by reviewing the log files, we can attempt to reflect same changes in MySQL database.
4. Tableau/Power BI Data Streaming: We can create a continuous stream of data to reflect the changes in database to Tableau, to make it more interactive and dynamic in true sense.

For all these changes, we are proposing the use of AWS Storage and Instances, to make the complete infrastructure in one place. So, that the changes can be reflected with no or minimal delay.

Using Hue Manager, we can strengthen the Security and Scheduling of the Data. With the use of Oozie or Any Scheduler, we can set a pipeline of tasks to execute them recurringly at regular intervals.

# Generalized Model for Diseases:

As this model is pushed towards Deployment, we can request access to more features of Patient Reports and try to draw insights to other diseases as well. Well Documented Datasets and Quality of data in those datasets can be a gamechanger for further advances in medical field.

# CHAPTER 6: Learnings from the Project:

During this period of project, we were able to discover new technologies and make decisions on what to use in the implementation. We tried using two or more technologies at every step of the project.

The technologies used in the project were:

1. Google forms and Spread Sheets (CSV files): We used these for the reason of being easily accessible to all technologies in each step.
2. MySQL and HDFS databases: There are two databases used in the project with almost identical data in both the databases. We created MySQL database for the project for the sole reason of fast querying. We also tried using Hive for the same step, but for the reason of keeping data independently for querying and Learning. We used MySQL and Hadoop, where Hadoop data is used for Apache Spark Machine Learning.
3. Sqoop : Sqoop is an excellent and minimal ETL (Extract Transform Load) Tool. Sqoop has no UI interface to have a drag and drop feature, but with its wide variety of input-control options it can be used with ease.
4. Apache Spark: It is a unified computational technology which greatly enhances the speed of computation along with the use of Diverse resources. Apache Spark is great tool for learning and has the capability of running seamlessly in real scenarios.
5. Tableau and Power BI: As Tableau dashboards can directly be shared using Tableau Public. It was preferred over Power BI. On the other hand, Power BI gives access to MySQL and other options to import data from, which are not there in Tableau Public.

After discussion, we decided to make both dashboards to have both the advantages.

With the further progress in mind, we learned few other technologies:

1. AWS (Amazon Web Service): AWS offers customizable storage, computing facilities and many other options as per customer needs. With AWS cloud storage, data and system accessible in one place.
2. Hue Manager: It provides all the software available in one place where you can set the security settings and also schedule tasks to implement and monitor from a single point.

The fields of Big Data and Machine Learning are used hand in hand to store and analyze the ever-growing data. The project was a great experience to both these fields.

# CHAPTER 7: REFRENCES

[1] Marjia Sultana, Afrin Haider, “Heart Disease Prediction using WEKA tool and 10-Fold cross-validation”, The Institute of Electrical and Electronics Engineers, March 2017.

[2] Marimuthu Muthuvel, “A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach”, 2018 in International Journal of Computer Applications.

[3] Dr. S.Seema Shedole, Kumari Deepika, “Predictive analytics to prevent and control chronic disease”, https://www.researchgate.net/punlication/316530782, January 2016.

[4] Ashok kumar Dwivedi, “Evaluate the performance of different machine learning techniques for prediction of heart disease using ten-fold cross-validation”, Springer, 17 September 2016

[5] Heart Disease Data Set at UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Heart+Disease (Retrieved on March 26th, 2021)

[6] Cleveland Clinic Heart Disease Dataset at Kaggle.

https://www.kaggle.com/aavigan/cleveland-clinic-heart-disease-dataset (Retrieved on April 26th, 2021)

[7] Shubhankar Rawat, Heart Disease Prediction article at TowardsDataScience. https://towardsdatascience.com/heart-disease-prediction-73468d630cfc

[8] US Cardiovascular Deaths During the COVID-19 Pandemic

https://www.acc.org/latest-in-cardiology/journal-scans/2021/01/11/20/27/cardiovascular-deaths-during-the-covid19