BACHELOR OF SCIENCE IN
COMPUTER SCIENCE AND ENGINEERING

BRAC UNIVERSITY

Inspiring Excellence

Heart Failure Risk Prediction and Medicine Recommendation System using Exploratory Analysis and Big Data Analytics

Authors
Sumaya Habib (15101129)
Maisha Binte Moin (15201003)
Sujana Aziz (15101019)

Supervisor
Mr. Hossain Arif
Assistant Professor
Department of CSE

A thesis submitted to the Department of CSE
in partial fulfillment of the requirements for the degree of
B.Sc. Engineering in CSE

Department of Computer Science and Engineering
BRAC University, Dhaka - 1212, Bangladesh

December 2018

We would like to dedicate this thesis to our loving parents …

Declaration

It is hereby declared that this thesis /project report or any part of it has not been submitted elsewhere for the award of any Degree or Diploma.

Authors:

Sumaya Habib
Student ID: 15101129

Maisha Binte Moin
Student ID: 15201003

Sujana Aziz
Student ID: 15101019

Supervisor:

Mr. Hossain Arif
Assistant Professor, Department of Computer Science and Engineering
BRAC University

December 2018

The thesis titled Heart Failure Risk Prediction and Medicine Recommendation System using Exploratory Analysis and Big Data Analytic
Submitted by:
Sumaya Habib, ID: 15101129
Maisha Binte Moin, ID: 15201003
Sujana Aziz, ID: 15101019
of Academic Year 2018 has been found as satisfactory and accepted as partial fulfillment of the requirement for the Degree of Bachelor of Science in Computer Science and Engineering

1.	Mr. Hossain Arif Assistant Professor BRAC University	Chairman
2.	Dr. Md. Abdul Mottalib Professor and Chairperson BRAC University	Member
3.		Member
4.		Member
5.		Member

Acknowledgements

We would like to acknowledge and express our gratitude to our supervisor Mr. Hossain Arif for his guidance and time. We would also like to appreciate him for encouraging us in this field of research and assisting us throughout the process.

Next, we would like to acknowledge Mr. Kalyan Banik, Alumni, Department of Computer Science and Engineering (CSE), BRAC University, for providing us with valuable insights on our research topic.

Abstract

With the ever increasing population of the world, diseases and their possibilities are also increasing at an alarming rate. As time passes by, diagnosing diseases and providing appropriate treatment at the right time has become quite a challenge. Heart diseases, for one, have been a major cause of death worldwide. Therefore, this research has been focused on finding an efficient way to predict the chances of a heart failure and accordingly, recommend appropriate medicines to aid cardiologists in quicker decision making. The research includes finding the correlations or associations between the various attributes of the dataset by utilizing the standard techniques of exploratory analysis and hence using the attributes suitably to predict the chances of a heart failure, as well as the medicine recommendations. A comparative study has also been included which shows the various attained accuracy rates of different machine learning algorithms including - Logistic Regression, Naïve Bayes, Decision Tree, Linear SVC, Random Forest, and Gradient Boosting Classifier. The Apache Spark framework has been used in order to make the system capable of handling big data.

Table of Contents

List of Figures

List of Tables

Nomenclature

Acronyms / Abbreviations

AHA	American Heart Association
ANN	Artificial Neural Network
API	Application Programming Interface
AUC	Area Under the ROC Curve
BP	Back Propagation
CAD	Coronary Artery Disease
CDC	Center for Disease Control
CVD	Cardiovascular Disease
DT	Decision Tree
EDA	Exploratory Data Analysis
EHR	Electronic Health Record
GBC	Gradient Boosting Classifier
HDFS	Hadoop Distributed File System
HF	Heart-Failure
HNB	Hidden Naïve Bayes
HPCC	High Performance Computer Cluster
ID3	Iterative Dichotomiser 3

Nomenclature

IDE	ntegrated Development Environment
IHD	Ischemic Heart Disease
IQR	Interquartile Range
K-NN	K-Nearest Neighbor
LDA	Latent Dirichlet Allocation
LR	Logistic Regression
ML	Machine Learning
NB	Naïve Bayes
NIDDM	Non-Insulin-Dependent Diabetes Mellitus
NN	Neural Network
RF	Random Forest
ROC	Receiver Operating Characteristic
SVC	Support Vector Machine Classifier
SVM	Support Vector Machine
WHO	World Health Organization

Chapter 1

Overview

1.1 Introduction

According to the data published by World Health Organization (WHO), in the year 2017, 14.31% of deaths in Bangladesh were caused by Coronary Heart Diseases [25]. In fact, the latest (2018) report provided by WHO has marked Heart Disease as one of the biggest causes of death all around the World [26]. With the growing population in countries like Bangladesh, it has become very difficult to provide proper treatment at the right time. This is mainly due to the reason that the medical test reports are usually manually analyzed by Cardiologists before they can reach a reasonable conclusion. It is, undoubtedly, a time consuming process. Therefore, we intended on designing a system that would analyze the patient's medical test reports and provide Cardiologists with the chances of a Heart Failure, and even suggest a medicine list that would be most suitable for the patient. The idea was to create a system that would facilitate Cardiologists in taking quicker decisions so that more patients can receive treatments within a shorter period of time.

We have used an Austrian data set which originally contains 43 attributes, including information such as the patient's age, gender, existing diseases, smoking habits, and the prescribed medicines for each individual (see Section 2.1). A significant portion of the research had been invested on finding the correlations or associations between these attributes and selecting the attributes which would be a stronger candidate for aiding the predictions. Various techniques of exploratory analysis such as basic statistical graphs, heatmaps, etc. had been used for a better understanding of the data set. Once done with the exploratory analysis, the selected attributes were fed into various classification algorithms. The algorithms included: Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, Linear Support Vector Machine Classifier (SVC), and Gradient Boosting Classifier.

One of the main aims of our research was to create a system that would be able to handle big data. We have used the Apache Spark framework for that purpose. Apache Spark is an open-source cluster-computing framework which can interface with Hadoop Distributed File System (HDFS) alongside some other standard distributed file systems. We used the python application programming interface (API) - PySpark, for running the machine learning (ML) algorithms on the Spark framework.

We, therefore, aimed to prototype a model that would be able to provide valuable insights on the huge amount of patient data that is rapidly being generated all around the world.

1.2 Related Work

Several researches have already been conducted on similar topics which provide rich guidelines on the basic idea of fusing Computer Science with Medical Science. Results of various surveys showed the impact of big data in the field of healthcare [10]. The research also included a comparison section of various tools and techniques including Hadoop, Storm, Spark, and High Performance Computer Cluster (HPCC), which provide solutions to the challenges imposed by big data.

Naïve Bayesian Classification technique was used to develop decision support in Heart Disease Prediction System by Priyanga and Naveen [12]. With unknown input sample, Naïve Bayes Weighted Approach system classified into various sections and traced the risk of heart disease. The system was developed with 13 attributes provided by the user. The hidden knowledge could be extracted from related data.

An intelligent system for predicting heart disease risks using three data mining techniques, namely - Naïve Bayes, Decision Tree and Neural Networks has also been designed, in which Naïve Bayes proved to be the most effective one, followed by Neural Networks and Decision Tree [9].

Another approach focused on the usage of bagging algorithm on heart disease prediction which was a merge of combinational processes or algorithms that could create room for improvement in the accuracy of the predictive performance through its learning and testing techniques. It was proved to be better than other decision trees by comparing between the resulting accuracies over a single model. That was because rather than considering new data set each time, this algorithm works on one or few data sets and progresses with learning methods [20].

Thomas and Princy proposed a system which predicted heart diseases using two modules: the classifier module and the prediction module. The classifier module used K-Nearest Neighbor (KNN) algorithm for classification and training, after which the second module found the risk rate using Iterative Dichotomiser 3 (ID3) algorithm. The mentioned data set consisted of 3 types of attributes: input, key, and prediction attributes [19].

A decision support system for heart disease classification has also been presented based on Support Vector Machine (SVM) and Artificial Neural Network (ANN). Simple Support Vector Machine (SSVM) algorithm has been used to determine the support vectors in a fast and iterative manner. For selecting the important and relevant features and discarding the irrelevant and redundant ones, ANN algorithm for back propagation was used to maximize classification accuracy. The results of the 5-class classification problem indicated an increase in the overall accuracy using the optimal feature subset. SVM classified the heart disease data into two classes which showed the presence or absence of a heart disease with 80.41% accuracy. ANN classified the data into 5 categories of heart diseases with 97.5% accuracy [5].

While studying recent publications of emerging technologies which play a role in the healthcare sector, Miotto et al. discussed the latest applications of deep learning in the said sector which focused on biomedical data obtained from clinical imaging, EHRs, genomes, etc. A tabular summary of the work done in this field using various deep learning architectures was also provided. They discussed the challenges faced, such as - data volume, temporality of the data, and the domain complexity; and scopes, such as - feature enhancement, federated inference, model privacy, incorporation of expert knowledge, and so on [8].

Soni et al. did a comparative study of data mining techniques proposed in recent years for heart disease prediction. The comparison conducted on the Cleveland data set using their Intelligent Heart Disease Prediction System showed that Decision Tree (89%) out performed the other classification techniques, whereas Bayesian Classifier (86.53%) and ANN (85.53%) also provided satisfactory results. They implemented Multilayer Perceptron Neural Networks for comparison. Finally, their research showed that the results improve significantly with the use of genetic algorithm by reducing the data set to an optimal one [17].

A fusion of genetic algorithm and neural networks using major risk factors has also been studied as a technique for heart disease prediction. This approach provided better results than the traditional Neural Network approach as it used the global optimization advantage of genetic algorithm for the initialization of NN weights. The data set which they used was customized from surveys and studies from Centre for Disease Control (CDC) and Prevention, and American Heart Association (AHA). The data set considered a total of 12 risk factors including age, heredity, diabetes, high cholesterol, and smoking [1].

Comparisons on data mining techniques by dividing them into two types of model-based approaches, such as Single Model Approach and Hybrid Model Approach have also been studied. Single Model Approaches gave the following accuracy results: Decision tree (76%), Association Rule (55%), K-NN (58%), SVM (86%), ANN (85%), Naïve Bayes (69%); whereas the Hybrid model provided a maximum accuracy of 96%. The research concluded that the hybrid approach produced the best results [13].

Srinivas et al. studied algorithms for the extraction of significant patterns from heart disease data. They suggested that in order to improve their work further, other data mining techniques such as time-series, clustering, association rule, etc. could also be incorporated [18].

Application of Hidden Naïve Bayes (HNB) classifier for the diagnosis of heart disease was used for testing its performance on the Statlog data set. Experimental results showed that HNB ensured comparatively better performance than other approaches. Proposed model exhibited data discretization and IQR filtering to enhance efficiency of HNB. The model recorded highest accuracy of 100% compared to Naïve Bayes classification model. HNB model assisted trustworthy and more authentic decision support system for automatic prediction of heart diseases [6].

Identifying Cardiovascular Disease (CVD) is a complex task that requires to be done in a short period of time, efficiently and also in an easy automation process. Every doctor cannot be equally trained or skillful in terms of each and every specialty and in lots of places people do not have access to specialists. So in this study, the design of a system that could efficiently discover the rules to predict the risk level of patients based on the given parameter about their health had been covered. The rules could be prioritized according to the user's specifications and needs. The performance of the system was tested and marked in terms of classification precision and results pointed that the system had a great potential in predicting the heart disease risk level [15].

Raihan et al. developed a simple approach to predict risk of developing Ischemic Heart Disease (IHD) using android or smartphone. An Android based prototype software was created by combining medical data acquired from patients with IHD. The data was collected from a set of patients and were interrelated with the risk factors like Hypertension, Diabetes, Dyslipidemia, Smoking, family background, Obesity, Stress, etc. which might suggest underlying non detected IHD. The data was mined with data mining technologies. Risk classifications included low, medium and high [14].

The design and implementation of a worldwide medicine recommendation system framework that applies data mining technologies to the recommendation system had been proposed by Bao and Jiang [2]. The medicine recommendation system consisted of database system module, data preparation module, recommendation model module, model evaluation, and data visualization module. They researched medicine recommendation algorithms of the SVM, Back Propagation (BP) Neural Network and ID3 Decision Tree algorithm on the diagnosis data. Lastly, on the open data set, SVM recommendation model was chosen for the medicine recommendation system to get a better trade among model accuracy, efficiency, and scalability and suggest an error checking process to ensure the accuracy and service quality.

Sodsee and Komkhao suggested an evidence-based medical recommendation system where the efficiency of medical recommendations was based on the diagnosis data and evidences correlated with reliable algorithms. These algorithms such as neural network, fuzzy theories, support vector machine, data mining techniques were used to extract the medical recommendations from data sets. These were based on physical examination of diagnosed patients [16].

Another study was based on an instance-based algorithm which permitted to explore big data set by making pairwise relations between patients with metric-free method. Here each data set had ranks. By clumping these ranking orders, it was possible to conceptualize data based on each set, holding subsets of similar and correlated patients. The paper also demonstrated a visualizing tool for utilizing a database of diabetic patients. This module of a recommendation system applied the denoted algorithm to enhance data, structure patients, create relations between individuals and give recommendations [4].

Chapter 2

Exploratory Analysis and Data Set Preprocessing

2.1 Data Set

We have used the Austrian medical data set available from [11] for our research. The data set originally contained 45 attributes and 2300 rows. The attributes included: study, physician-ID, patient-ID, age, gender, migrant, migration-class, CAD-fam-history, hyperlipidemia, periph-artery-disease, physical-activity, hypertension, NIDDM, smoking, CAD-year, post.cor, chron-lung-disease, previous-stroke, depression, any-angina, angina-episodes-week, nitro-week, angina-class, heart-failure, nyha, heart-rate, atrial-fibrillation, pacemaker, crt-defib, platelet-inhibitors, platelet-inhibitors-dual, oral-anticoagulants, ace-at2, lipid-lowering, diuretics, beta-blockers, beta-blockers-dosis, calciumantagonists, ivabradin, nicorandil, molsidomin, antiarrhythm, long-acting-nitrates, ranolazin, and glycosides.

Most of these attributes contained categorical binary data represented by either a “yes” or a “no”. The only attributes with numerical data were: the IDs, age, CAD-year, angina-episodes-week, nitro-week, angina-class, nyha, heart-rate, and beta-blockers-dosis. Of the attributes which contained categorical data - smoking, calciumantagonists, beta-blockers, physical-activity, and gender contained values other than a “yes” or a “no”, to represent their respective sub-categories.

We have consulted with physicians as to what these attributes referred to, along with some research carried out by ourselves. Brief descriptions of some of the crucial columns consisting of information regarding diseases and medicines are provided below:

1. **Angina:** Chest pain that is linked to heart disease.

2. **Atrial fibrillation:** Abnormal heart rhythm characterized by rapid and irregular beatings of the atria.

3. **Heart failure:** Caused when the heart is unable to pump sufficiently to maintain blood flow.

4. **Hyperlipidemia:** Abnormally high levels of lipo-proteins in the blood.

5. **Hypertension:** Persistent elevated blood pressure in the artery.

6. **NIDDM:** Non-Insulin Dependent Diabetes Milletus

7. **Peripheral Artery Disease:** Narrowing of peripheral arteries restricting blood supply to leg muscles.

8. **Previous Stroke:** Occurrence of previous stroke.

9. **Chronic Lung Disease:** Term for a variety of persistent lung disorders.

10. **Ace at2:** Regulator for treating hypertension and heart failure.

11. **Antiarrhythm:** To treat cardiac arrhythmias characterized by unstable, disorganized, abnormal and asynchronous heartbeats.

12. **Beta blockers:** Treats heart conditions and hypertension.

13. **Calciumantagonists:** Relaxes blood vessels and increases supply of blood and oxygen to the heart.

14. **Diuretics:** Medicine for High Blood Pressure.

15. **Glycosides:** Important agents in the drug therapy of advanced cardiac heart failure.

16. **Ivabradin:** Keeps the heart rate levelled and reduces angina pain.

17. **Lipid lowering:** For patients with hyperlipidemia.

18. **Long Acting Nitrates:** For proper blood flow and treating angina.

19. **Molsidomin:** Used for treatment of coronary heart disease.

20. **Nicorandil:** Reduces painful signs of heart disease, for adults who cannot take beta blockers and calciumantagonists.

21. **Platelet inhibitors:** Decreases platelet aggregation.

22. **Ranolazin:** Used to treat chronic angina.

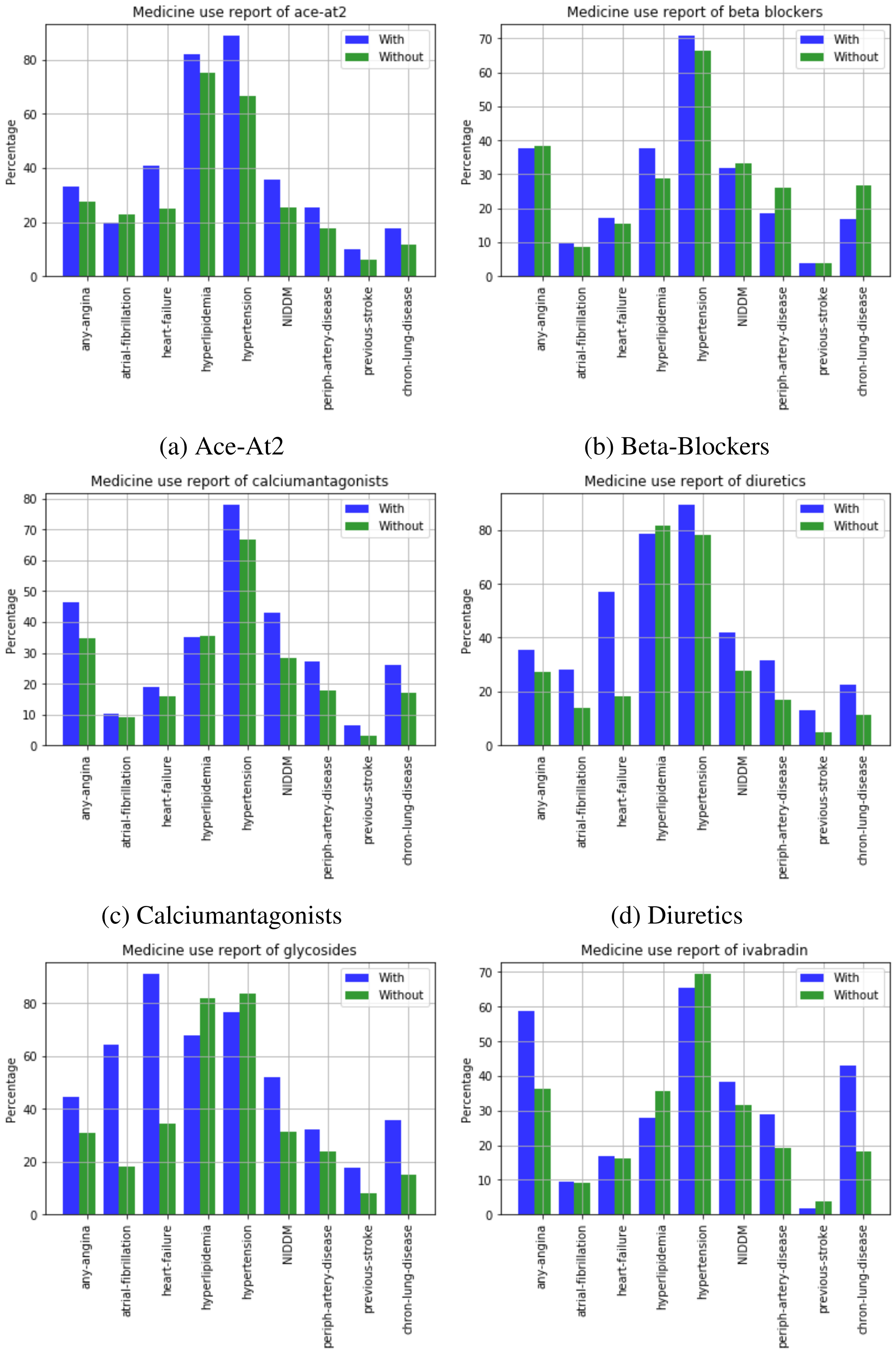
2.2 Exploratory Analysis

Exploratory Data Analysis (EDA) is the strategy for skimming data sets and seeing what information they contain. This is often done through data visualization techniques such as - statistical graphs and plots. It basically helps summarizing the properties of the data sets [23]. The standard techniques used in EDA includes: Box Plot, Histogram, Multi-vari Chart, Run Chart, Pareto Chart, Scatter Plot, Stem-and-leaf Plot, Parallel Coordinates, Odds Ratio, and much more.

2.2.1 Understanding the Data Set

We started out by plotting bar graphs to see the frequency distribution of each individual attributes with the other attributes (see Fig. 2.1). Some of the inferences made on the medicine columns on the basis of these graphs include:

- **Ace-At2** has a positive correlation with Hypertension, Heart-Failure, NIDDM, and Periph-Artery-Disease.
- **Beta-Blockers** have a positive correlation with Hyperlipidemia and Hypertension; and a negative correlation with Periph-Artery-Disease.
- **Calciumantagonists** have a positive correlation with Any-Angina, Hypertension, and NIDDM.
- **Diuretics** have a positive correlation with Heart-Failure, Atrial-Fibrillation, Periph-Artery-Disease, Previous-Stroke, Any-Angina, and NIDDM.
- **Glycosides** have a positive correlation with Atrial-Fibrillation, Heart-Failure, NIDDM, and Any-Angina; and a negative correlation with Hyperlipidemia.
- **Ivabradin** has a positive correlation with Any-Angina and Periph-Artery-Disease.
- **Lipid-Lowering** has a positive correlation with Hyperlipidemia, NIDDM, Hypertension, and Periph-Artery-Disease; and a negative correlation with Atrial-Fibrillation and Heart-Failure.
- **Long-Acting-Nitrates** have a positive correlation with Any-Angina and NIDDM; and a negative correlation with Atrial-Fibrillation, Hyperlipidemia, and Heart-Failure.
- **Molsidomin** has a positive correlation with Any-Angina, Heart-Failure, Previous Stroke and Periph-Artery-Disease.



(a) Ace-At2

(b) Beta-Blockers

(c) Calciumantagonists

(d) Diuretics

(e) Glycosides

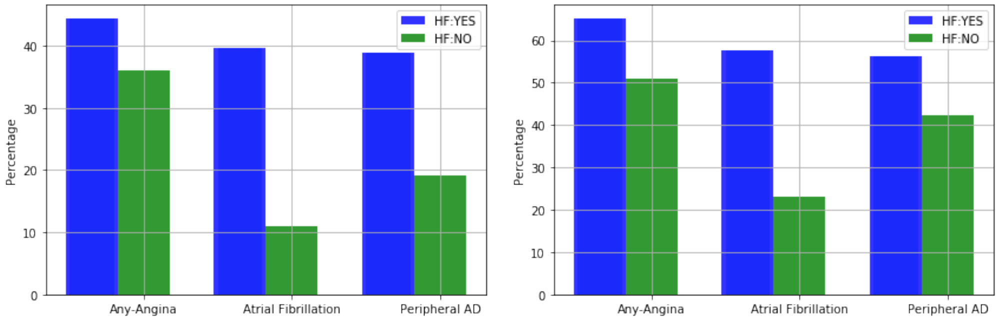
(f) Ivabradin

Fig. 2.1 Yes/No Frequency Distribution With Respect to Diseases and Symptoms for Some of the Medicines

- **Nicorandil** has a positive correlation with Any-Angina, NIDDM, and Periph-Artery-Disease; and a negative correlation with Hyperlipidemia.
- **Platelet-Inhibitors** have a positive correlation with Hyperlipidemia and a negative correlation with Atrial-Fibrillation.
- **Antiarrhythm** has a positive correlation with Atrial-Fibrillation, Heart-Failure, and Previous-Stroke.

Since Heart-Failure was our prime target attribute, we have plotted a different set of bar graphs for visualizing the relationship between Heart-Failure and other diseases and symptoms, separately for males and females. Inferences made on these graphs (see Figures 2.2 and 2.3) include:

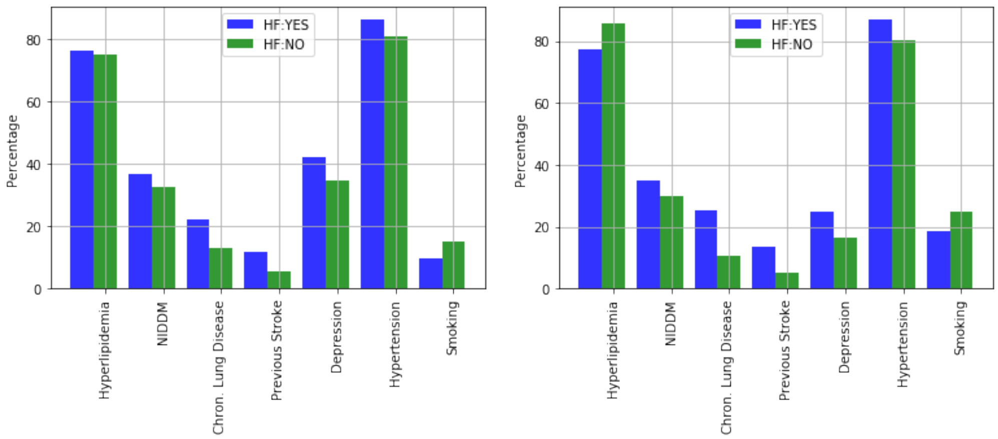
- In case of both males and females most of the patients who had a record of Heart-Failure also had Atrial-Fibrillation.
- Most patients (male and female) with and without Heart-Failure report had Angina. So the relation between Angina and Heart-Failure can be considered as weak, from these graphs.
- Whereas almost 50% of females with a Heart-Failure did not have Peripheral Artery Disease, the distinction is lower for male patients. Therefore, although we might infer that females with Peripheral Artery Disease are likely to have a Heart Failure, this cannot be said for the male patients.
- As for the other diseases, both male and female patient reports showed similar trends.
- Female patients had a higher frequency of depression than male patients.



(a) Report for Female Patients

(b) Report for Male Patients

Fig. 2.2 Heart-Failure Relation with the Three Main Attributes Related to the Heart



(a) Report for Female Patients

(b) Report for Male Patients

Fig. 2.3 Heart-Failure Relation with Other Diseases and Symptoms

2.2.2 Association Between Attributes

Although the graphs provided us with a primary insight on the data set, we still needed stronger evidences to confirm these hypotheses. For that we have applied various techniques to figure out which attributes are strongly associated with each other. Since most of the columns were categorical, the techniques used included:

- **Pearson Correlation Coefficient:** It is any statistical hypothesis test where the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. This test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories [21].
- **Spearman's Rank Correlation Coefficient:** Spearman's rank correlation coefficient, or Spearman's rho, is a non-parametric measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function [24].

Both Pearson-R and Spearman-R coefficients range from -1 to 1, to show positive linear correlation to negative linear correlation and a 0 to indicate zero-correlation. The problem with these methods is that they consider the attribute values as ordinal data. Since most of the attributes that we were concerned with were nominal, *correlation* was not quite the correct term that we were looking for. What we really wanted to see was how much the nominal attributes were *associated* with each other. The technique used for that purpose is:

- **Cramér's V:** Cramér's V (sometimes referred to as Cramér's phi) is a measure of association between two nominal variables, giving a value between 0 and +1 (inclusive). It is based on Pearson Correlation Coefficient [22]. Pearson Coefficients tell us whether the two variables are significantly related or not, but it does not tell us exactly how significant the relation is. This additional information is provided by Cramér's V. Fig. 2.4 demonstrates the calculated Cramér's V coefficients for some of the attributes. In consideration of the values obtained, we decided on a threshold value of 20%.

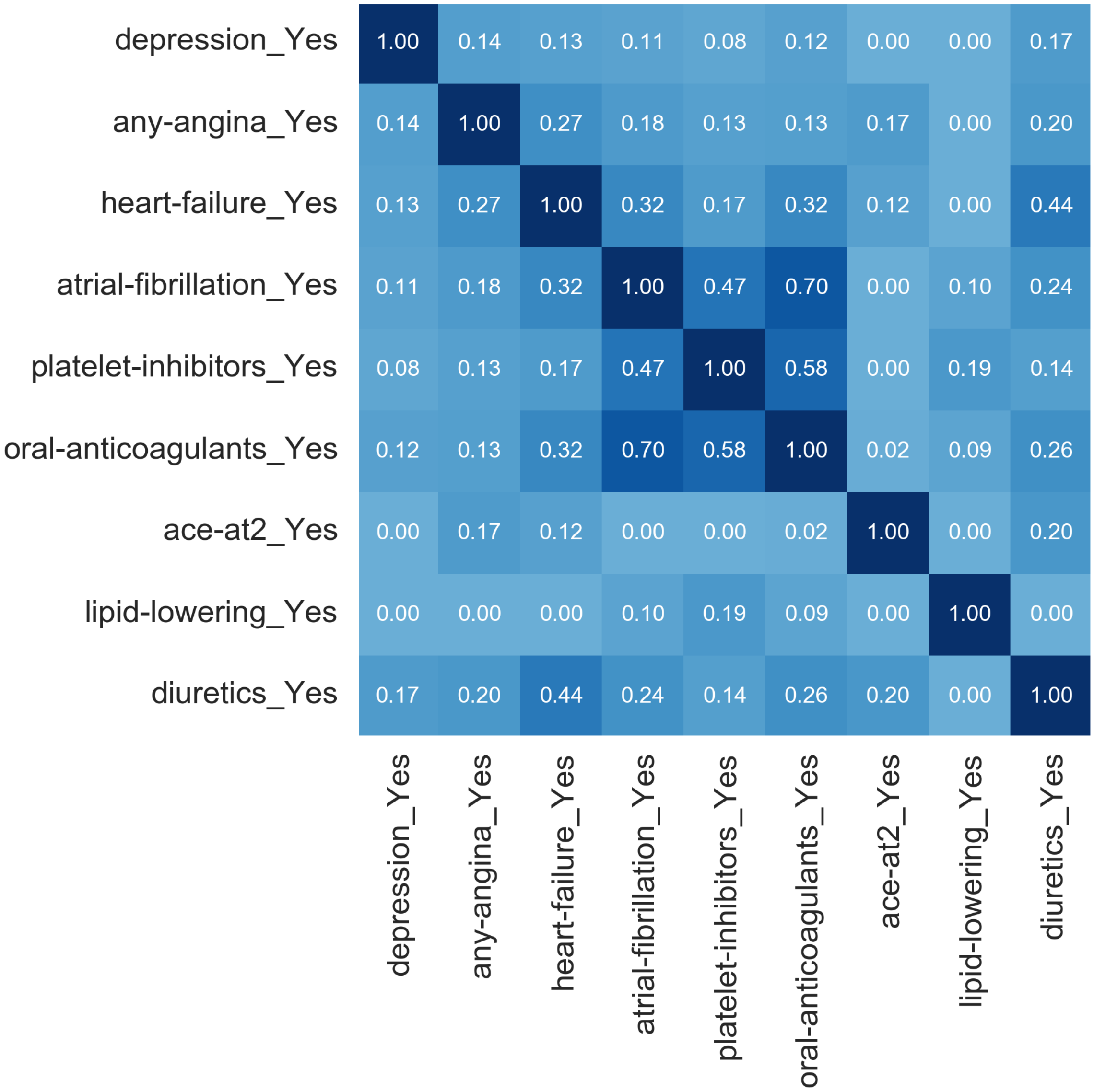


Fig. 2.4 Heatmap Showing the Cramer's V Coefficients of Some of the Attributes

Inferences based on the Cramér's V coefficients included:

- **Heart-Failure** has a strong association with Peripheral Artery Disease, Chronic Lung Disease, Angina, Atrial Fibrillation, Oral Anti-Coagulants, Diuretics and Glycosides.
- **Platelet Inhibitors** have a strong association with Atrial Fibrillation, Oral Anti-Coagulants and Glycosides.
- **Oral Anti-Coagulants** have a strong association with Gender, Heart-Failure, Atrial Fibrillation, Platelet Inhibitors, Diuretics and Glycosides.
- **Ace-At2** has a strong association with Hypertension and Diuretics.
- **Lipid Lowering** has a strong association with Hyperlipidemia.
- **Diuretics** have a strong association with Peripheral Artery Disease, Hypertension, Chronic Lung Disease, Angina, Heart Failure, Atrial Fibrillation, Oral Anti-Coagulants, Ace-At2 and Ivabradin.
- **Beta Blockers** have a strong association with Lipid Lowering, Atrial Fibrillation, Platelet Inhibitors, Oral Anti-Coagulants and Ivabradin.
- **Calciumantagonists** have a strong association with Diurectics.
- **Ivabradin** has a strong association with Diuretics and Beta Blockers.
- **Nicorandil** has a strong association with Diuretics, Angina and Oral Anti-Coagulants.
- **Molsidomin** has a strong association with Previous Stroke, Angina, Age and Heart Rate.
- **Long Acting Nitrates** have a strong association with Angina, Hyperlipidemia and Diuretics.
- **Glycosides** have a strong association with Atrial Fibrillation.

2.3 Dataset Preprocessing

Based on the definition of each individual attribute and the purpose of our research, we created a smaller subset of the original dataset by eliminating the attributes: study, physician-ID, patient-ID, migrant, migration-class, post-cor, and nyha.

2.3.1 Handling Missing Data

From Table 2.1, we can see that 23 of the selected attributes had at least 20% or more of the data missing. We could not directly eliminate these attributes since most of them were vital for the purpose of our research. The only columns dropped from this survey were: nitro-week, platelet-inhibitors-dual, CAD-year, beta-blockers-dosis, CAD-fam-history, angina-episodes-week, smoking, physical-activity, crt-defib, pacemaker and angina class.

Table 2.1 Percentage of Missing Data for Each Attribute

Attribute	Percentage	Attribute	Percentage
calciumantagonists	0.000	hyperlipidemia	57.460
beta-blockers	0.000	crt-defib	57.503
age	0.695	ace-at2	57.721
heart-rate	1.304	physical-activity	57.721
hypertension	1.739	glycosides	57.808
NIDDM	2.001	molsidomin	58.025
nicorandil	3.392	oral-anticoagulants	58.199
chron-lung-disease	3.915	ranolazin	58.199
depression	4.393	previous-stroke	58.330
long-acting-nitrates	4.524	smoking	58.504
ivabradin	4.828	platelet-inhibitors-dual	58.504
periph-artery-disease	4.915	CAD-year	59.983
any-angina	14.702	heart-failure	60.244
angina-class	21.096	antiarrhythm	60.853
beta-blockers-dosis	30.187	diuretics	63.245
pacemaker	56.633	angina-episodes-week	70.987
lipid-lowering	56.633	CAD-fam-history	73.032
atrial-fibrillation	56.851	nitro-week	88.212
platelet-inhibitors	57.460		

For the rest of the attributes, we also could not just drop the missing rows since that would drastically decrease our data size. Therefore, we used Logistic Regression Classifier to predict the missing using the 13 attributes with the least missing data rate. We used One-Hot Encoding before proceeding with the algorithm. Table 2.2 shows the iteration accuracy rates of each of the predicted attributes.

Table 2.2 Predicted Missing Attributes and Their Accuracy Rates

Attribute	Iteration Accuracy
Heart-Failure	0.80
Lipid-Lowering	0.94
Atrial-Fibrillation	0.92
Platelet-Inhibitors	0.88
Ace-At2	0.94
Hyperlipidemia	0.94
Glycosides	0.88
Molsidomin	0.90
Oral-Anticoagulants	0.88
Ranolazin	1.00
Previous-Stroke	0.92
Diuretics	0.86
Antiarrhythm	1.00

After filling out the missing values, we dropped the columns - Antiarrhythm and Ranolazin because their unnaturally high accuracy rates were owing to the fact that most of their rows had a "No" classification.

After taking all these measures, our final processed dataset had a total of 25 attributes and 1614 rows (without any missing values). In order to check the reliability of this semi-synthesized dataset, we ran the Cramer’s V on it and got the coefficients as shown in Fig. 2.5. Comparing this new Heatmap with the one obtained before, it was seen that although the values changed (due to an increase in data size), the relative association between the attributes remained very similar. Therefore, we concluded that this new dataset was in fact, consistent with the original dataset and thereafter, proceeded with the ML algorithms.

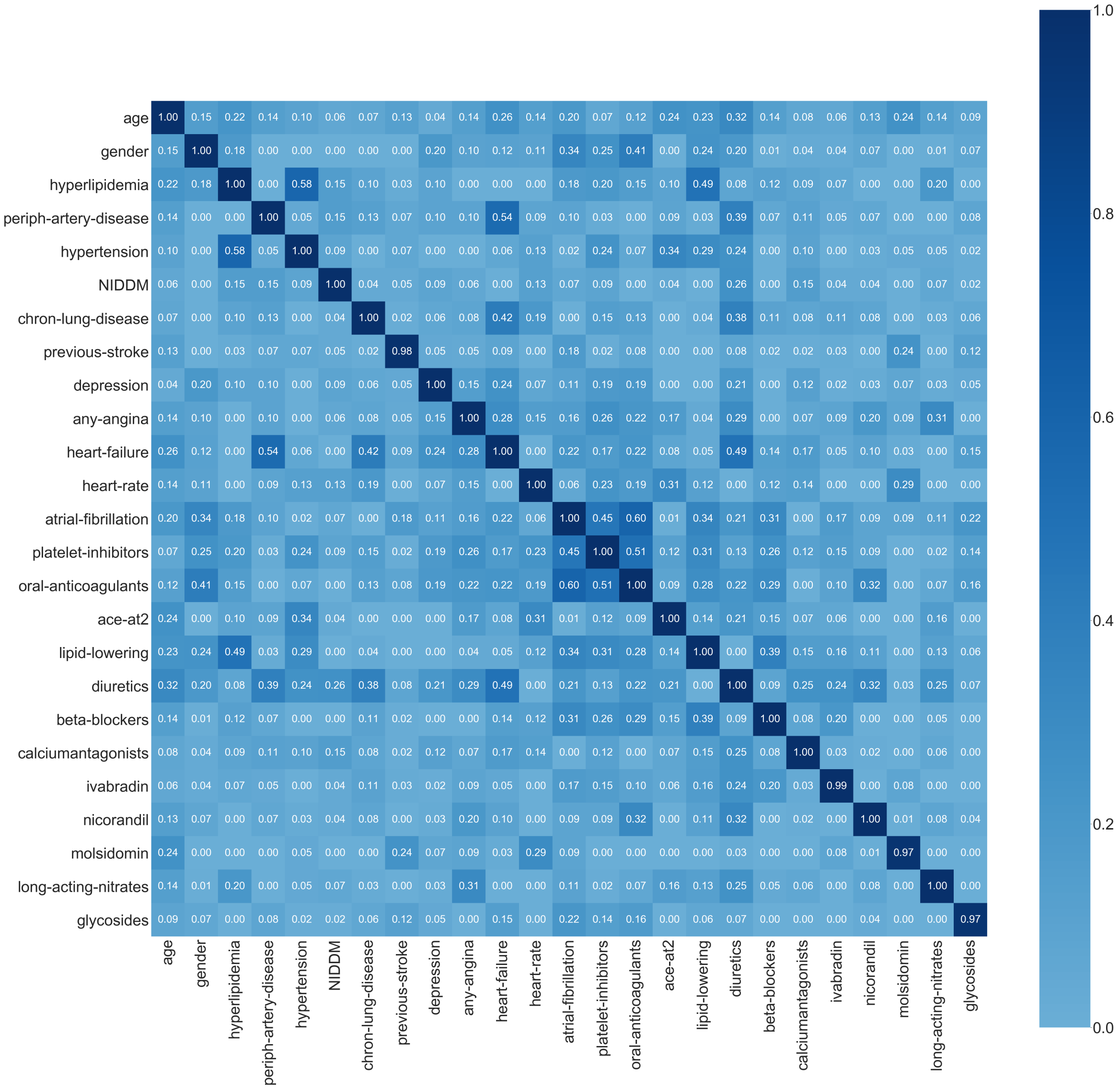


Fig. 2.5 Heatmap Showing the Cramer's V Coefficients of the New Dataset

Chapter 3

Methodology and Results

3.1 System Design

Based on all our findings from the Exploratory Analysis, we listed out the attributes that we could use to predict Heart-Failure risk and the appropriate Medicines (see Table 3.1). From Table 3.1 we can clearly see that some of our labels have dependencies on previous labels. Fig. 3.1 demonstrates the simplified block diagram of our system to better understand the model.

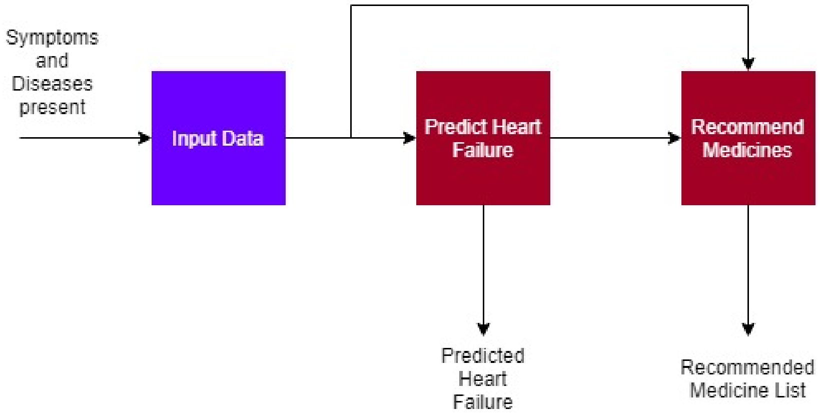


Fig. 3.1 Simplified Block Diagram of Our System

Table 3.1 Label and Features for Predictive System

Label	Features
heart-failure	age, periph-artery-disease, chron-lung-disease, any-angina, atrial-fibrillation
platelet-inhibitors	gender, hyperlipidemia, hypertension, any-angina, heart-rate, atrial-fibrillation
oral-anticoagulants	gender, atrial-fibrillation, heart-failure, any-angina, platelet-inhibitors
ace-at2	age, hypertension, heart-rate
lipid-lowering	age, gender, hyperlipidemia, hypertension, atrial-fibrillation, platelet-inhibitors, oral-anticoagulants
diuretics	age, gender, periph-artery-disease, hypertension, NIDDM, chron-lung-disease, depression, any-angina, heart-failure, atrial-fibrillation, ace-at2, oral-anticoagulants
beta-blockers	atrial-fibrillation, platelet-inhibitors, lipid-lowering oral-anticoagulants
calciumantagonist	diuretics
ivabradin	diuretics, beta-blockers
nicorandil	any-angina, oral-anticoagulants, diuretics
molsidomin	age, previous-stroke, heart-rate
long-acting-nitrates	hyperlipidemia, any-angina, diuretics

3.2 Methodology

As stated earlier, we have used Logistic Regression, Naïve Bayes (Bernoulli's Naïve Bayes to be more precise), Decision Tree, Random Forest, Linear SVC and Gradient Boosting Classifier for our comparative study.

3.2.1 Classifiers

1. **Logistic Regression Classifier**

   Logistic Regression was first developed by statistician David Cox in 1958 [3], and is a regression analysis used for predicting categorical, binary or ordinal variables. It is a supervised machine learning technique and is basically a probabilistic classification model [7].

   Binary Logistic Regression is a form of binomial regression, which estimates the parameters of a logistic function. It is a widely used regression analysis and is especially appropriate when the dependent variable is binary. The dependent variable is usually represented by "0" or "1", indicating the absence or presence of a certain attribute or feature.

2. **Naïve Bayes Classifier**

   Naïve Bayes Classifiers belong to the family of probabilistic classifiers based on the Bayes theorem. It is called "naive" because of its assumption that there is a strong dependency between the concerned features [7].

   The three main types of Naïve Bayes Classifier includes Gaussian Naïve Bayes, Bernoulli's Naïve Bayes and Multinomial Naïve Bayes.

   Gaussian Naïve Bayes comes in handy when dealing with continuous data. It assumes that the continuous values are distributed according to the Gaussian distribution.

   Multinomial Naïve Bayes is typically used for document classification.

   Bernoulli's Naïve Bayes, on the other hand, deals with binary data.

3. **Support Vector Machine Classifier**

   It is a set of supervised learning methods used for classification, regression and outlier detection. It tries to find a hyper plane that can effectively divide the given training data into two parts. The major advantage of support vector machines is effectiveness in high dimensional spaces. Also it uses a subset of training points in the decision function called support vectors, so it is also memory efficient. The one drawback in SVM is when training data is highly unbalanced, resulting model tends to perform well on majority data but perform bad on minority data.

4. **Decision Tree Classifier**

   Its goal is to create a model that predicts the value of a target variable based on several input variables. Decision trees used in data mining are of two main types: Classification tree and Regression tree.

5. **Random Forest Classifier**

   It is a meta estimator that fits a number of decision tree classifiers on different sub samples of the dataset and use averaging method to improve the predictive accuracy and control over-fitting. Each tree in the ensemble is built from bootstrap sample from the training set. When splitting a node during the construction of the tree, the split that is chosen is no longer the best split among all features. Instead, the split that is picked is the best split among a random subset of the features. As a result of this randomness, the bias of the forest usually slightly increases compare to the bias of a single non-random tree.

6. **Gradient Boosting Classifier**

   It is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

3.2.2 Model Evaluation Techniques

Described below are some of the standard model evaluation criteria:

- **Accuracy:** The number of correct classifications divided by the total number of classifications.
- **Precision:** A measure of exactness or quality of a model. It is the total number of true positives divided by the total number of true positive and false positive. It is also known as the specificity of the model.
- **Recall:** A measure of completeness or quantity of the model. It is the total number of true positive divided by the total number of true positive and false negative. It is also known as the sensitivity of the model.
- **F1-Score:** It is a harmonic mean of Precision and Recall.
- **AUC:** ROC curves determine the performance of a classifier without considering distribution or error costs.

3.2.3 Hyperparameter Tuning

Our data set was randomly was divided into two parts: 80% for Training and Validating the Model; and 20% for Testing. The data separated for training and validation where then used to tune the models, to figure out the optimum parameters for each classifier. The validation parameters were set to 10 fold cross-validation, with a maximum of 50 iterations. It means that - in a single iteration the training data was randomly split into 10 approximately equal divisions and each division was used once, for testing the accuracy, while the rest were used to train the model with some specific parameters. 50 such iterations meant that the model was trained and tested 500 times before selecting the parameters for which the model obtained the highest accuracy rate. Table 3.2 enlists some of the evaluation readings obtained during the hyperparameter tuning stage.

After the best model was selected for each classifier, they were once again tested on the entire data set, with a cross-validation of 10 folds. Table 3.3 enlists the average evaluation readings obtained during the testing stage.

3.2.4 Results

We can clearly see from the Tables 3.2 and 3.3 that the model accuracy is highest for Logistic Regression and Decision Tree, followed by Gradient Boosting Classifier. The difference between accuracy rates of Model Performance and Test Performance is the least for GBT, whereas it is the highest for DT. Despite the fact that RF is an upgradation of DT, the overall model performance turned out to be better for DT, than for RF. NB and SVC, on the other hand, showed similar model performance. Their test scores were also quite close.

The high precision value of LR, DT, GBC and RF, respectively, shows that their total correct positive predictions outnumbered their total incorrect positive predictions. The high recall value of DT, LR, GBC and RF, respectively, shows that they could accurately predict most of the positive outcomes. The f1-score is almost identical to the recall value of the classifiers. On the other hand, DT, LR, GBC and RF also excelled in terms of accuracy. However, AUC-ROC turned out to be highest for RF.

The test evaluation scores are relatively lower than the model evaluation scores, with the exception of AUC-ROC. Which signifies that the models could better distinguish between the two classes in the testing phase, than in the modelling phase.

The test performance usually signifies how well the model performs on unexplored data. In our case, however, since the data size was not that big, a small portion of the test data were unexplored. In any case, the best performer among the six algorithms turned out to be Logistic Regression, followed by Gradient Boosting Classifier and Decision Tree.

Therefore, our System could predict the chances of a heart-failure and recommend medicines at the maximum model accuracy rate of 90% and test accuracy rate of 88%.

Table 3.2 Model Performance

Algorithms	Accuracy	Precision	Recall	F1-Score	AUC
LR	0.91	0.92	0.91	0.91	0.88
NB	0.83	0.83	0.84	0.83	0.74
DT	0.91	0.92	0.92	0.92	0.87
RF	0.89	0.89	0.89	0.89	0.89
SVC	0.83	0.83	0.84	0.83	0.78
GBC	0.90	0.90	0.90	0.90	0.86

Table 3.3 Test Performance

Algorithms	Accuracy	Precision	Recall	F1-Score	AUC
LR	0.88	0.84	0.72	0.76	0.93
NB	0.86	0.82	0.61	0.70	0.92
DT	0.75	0.83	0.70	0.75	0.90
RF	0.89	0.82	0.76	0.79	0.92
SVC	0.88	0.82	0.72	0.76	0.93
GBC	0.89	0.81	0.75	0.78	0.93

3.2.5 Handling Big Data

After selecting the most suitable model for our system, it was finally coded using PySpark (distribution 2.4.0), which is Python's API for Apache Spark. Apache Spark is an open source cluster computing framework which was designed for dealing with huge amounts of data. The ML library in PySpark currently supports all the aforementioned classifiers along with Multilayer Perceptron Classifier and One Vs Rest Classifier. It also supports clustering models such as Gaussian Mixture, K-Means, Bisecting Means, Latent Dirichlet Allocation (LDA), Distributed LDA, and Power Iteration Clustering. There is also a rich module for the various evaluation techniques for hyperparameter tuning and model evaluation.

Since we did not have access to big data, we could not re-evaluate our model. But it has been designed with the appropriate tools which can, in fact, work with huge data.

The tools we used simply included the Anaconda distribution, for downloading all the necessary libraries and APIs, and the PyCharm integrated development environment (IDE).

Python's Seaborn, Matplotlib, Numpy, and Pandas library was used during the EDA phase. The sklearn library has also been used while predicting the missing values and evaluating the models.

Chapter 4

Conclusion

With the aim to help Cardiologists take quicker decisions, we have tried to build a System that would predict the chances of a Heart-Failure based on the patients' reports and also suggest a medicine list that would be most suitable for the patient. For our research we have used an Austrian medical data set which originally contained 45 attributes and 2300 rows. After completing Exploratory Analysis and Data Preprocessing, we were left with a total of 25 attributes, 1614 rows and a list of attributes that would be the best candidates for each predictive sub-system. Next, we used these features to predict our target attributes, namely - Heart Failure, Platelet Inhibitor, Oral Anticoagulant, Ace-At2, Lipid-Lowering, Diuretic, Beta-Blocker, Calciumantagonist, Ivabradin, Nicorandil, Molsidomin and Long Acting Nitrates. The algorithms used included Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, Linear SVC and Gradient Boosting Classifier. Best model and test performance was attained by Logistic Regression, followed by Gradient Boosting Classifier and Decision Tree. These algorithms were tested on the Apache Spark framework, since we wanted our System to be compatible with big data.

Since we did not have access to Big Data, we could not re-evaluate our model further to check for over or under fitting. We, therefore, look forward to exploring bigger data sets to further polish our model and figure out a mechanism to create a feedback system so that the more predictions or recommendations it make, the more "learned" it becomes.

References

[1] Amin, S. U., Agarwal, K., and Beg, R. (2013). Genetic neural network based data mining in prediction of heart disease using risk factors. In *Information & Communication Technologies (ICT), 2013 IEEE Conference on*, pages 1227–1231. IEEE.

[2] Bao, Y. and Jiang, X. (2016). An intelligent medicine recommender system framework. In *Industrial Electronics and Applications (ICIEA), 2016 IEEE 11th Conference on*, pages 1383–1388. IEEE.

[3] Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society (JSTOR)*, 20(02):215–242.

[4] Falip, J., Aït-Younes, A., Blanchard, F., Delemer, B., Diallo, A., and Herbin, M. (2017). Visual instance-based recommendation system for medical data mining. *Procedia computer science*, 112:1747–1754.

[5] Gudadhe, M., Wankhade, K., and Dongre, S. (2010). Decision support system for heart disease based on support vector machine and artificial neural network. In *Computer and Communication Technology (ICCCT), 2010 International Conference on*, pages 741–745. IEEE.

[6] Jabbar, M. and Samreen, S. (2016). Heart disease prediction system based on hidden naïve bayes classifier. In *International Conference on Circuits, Controls, Communications and Computing (I4C), Bangalore*.

[7] Madhavan, S. (2015). *Mastering Python for Data Science*. Packt Publishing Ltd.

[8] Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. (2017). Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*.

[9] Palaniappan, S. and Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. In *Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on*, pages 108–115. IEEE.

[10] Patel, J. A. and Sharma, P. (2014). Big data for better health planning. In *2014 International Conference on Advances in Engineering and Technology Research (ICAETR)*, pages 1–5. IEEE Unnao.

[11] Pichlhöfer, O., Maier, M., Badr-Eslam, R., Ristl, R., Zebrowska, M., and Lang, I. M. (2017). Clinical presentation and management of stable coronary artery disease in austria. https://figshare.com/articles/Clinical_presentation_and_management_of_stable_coronary_artery_disease_in_Austria/4925702. [Online; accessed 30-November-2018].

[12] Priyanga, P. and Naveen, N. (2017). Web analytics support system for prediction of heart disease using naive bayes weighted approach (nbwa). In *2017 Asia Modelling Symposium (AMS)*, pages 21–26. IEEE.

[13] Purusothaman, G. and Krishnakumari, P. (2015). A survey of data mining techniques on risk prediction: Heart disease. *Indian Journal of Science and Technology*, 8(12).

[14] Raihan, M., Mondal, S., More, A., Sagor, M. O. F., Sikder, G., Majumder, M. A., Al Manjur, M. A., and Ghosh, K. (2016). Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design. In *Computer and Information Technology (ICCIT), 2016 19th International Conference on*, pages 299–303. IEEE.

[15] Saxena, K., Sharma, R., et al. (2015). Efficient heart disease prediction system using decision tree. In *Computing, Communication & Automation (ICCCA), 2015 International Conference on*, pages 72–77. IEEE.

[16] Sodsee, S. and Komkhao, M. (2013). Evidence-based medical recommender systems: A review. *Int J Inf Process Manag*, 4(6):114–20.

[17] Soni, J., Ansari, U., Sharma, D., and Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, 17(8):43–48.

[18] Srinivas, K., Rani, B. K., and Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering (IJCSE)*, 2(02):250–255.

[19] Thomas, J. and Princy, R. T. (2016). Human heart disease prediction system using data mining techniques. In *Circuit, Power and Computing Technologies (ICCPCT), 2016 International Conference on*, pages 1–5. IEEE.

[20] Tu, M. C., Shin, D., and Shin, D. (2009). Effective diagnosis of heart disease through bagging approach. In *Biomedical Engineering and Informatics, 2009. BMEI'09. 2nd International Conference on*, pages 1–4. IEEE.

[21] Wikipedia contributors (2018a). Chi-squared test — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Chi-squared_test&oldid=868667934. [Online; accessed 30-November-2018].

[22] Wikipedia contributors (2018b). Cramér's v — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Cram%C3%A9r%27s_V&oldid=870865993. [Online; accessed 30-November-2018].

[23] Wikipedia contributors (2018c). Exploratory data analysis — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Exploratory_data_analysis&oldid=871191702. [Online; accessed 30-November-2018].

[24] Wikipedia contributors (2018d). Spearman's rank correlation coefficient — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Spearman%27s_rank_correlation_coefficient&oldid=869374150. [Online; accessed 30-November-2018].

[25] World Health Organization (2017). World health rankings. https://www.worldlifeexpectancy.com/bangladesh-coronary-heart-disease. [Online; accessed: 02-December-2018].

[26] World Health Organization (2018). World heart day. https://www.who.int/cardiovascular_diseases/world-heart-day/en/. [Online; accessed: 02-December-2018].