Acknowledgements

We have taken efforts in successfully completing this project. However, it would not have been possible without the kind support and help of many individuals and the organization. We would like to extend my sincere thanks to all of them.

We are grateful to my Institution, **Ramaiah Institute of Technology** with its ideals and inspiration for having provided us with the facilities, which has made this project a success.

We acknowledge the guidance and support extended by our guide, **Mrs. Lincy Meera Mathews**, Assistant Professor, Department of ISE. Her incessant encouragement and invaluable technical support have been of immense help. Her guidance gave us the environment to enhance our knowledge and skills.

We take this opportunity to express our profound gratitude and deep regards to the Head of the Department and Professor **Dr. Vijaya Kumar B.P.** for his valuable information, exemplary guidance, monitoring and constant encouragement throughout the course of this thesis. The blessing, help and guidance given by him time to time shall carry us a long way in the journey of life on which we are about to embark.

We would like to extend my hearty gratitude to **Dr. N V R Naidu**, The Principal, for the kind support and permission to use the facilities available in the Institute. We are obliged to all the teaching and non-teaching staff members of the ISE department, for the valuable information provided by them in their respective fields. We are grateful for their cooperation during the period of our project.

We would also like to thank all the teaching and nonteaching staff of our department for their kind cooperation and assistance. Lastly, we thank everyone for their constant encouragement with this project.

Biswajit Mohanty (1MS16IS018)

I V Shyam Kumar Varma (1MS16IS029)

Mohammed Fayaz (1MS16IS039)

Yash Khakhar (1MS16IS120)

Abstract

The need to curb non communicable diseases is higher than ever, especially in a market like India where the mortalities caused by NCDs is on a steady rise. One solution to curb the same is to learn/analyze from the patient health data as to what exactly is the reason behind the rise in mortalities by NCDs.

This project works on a Diabetes dataset SHERM aims at providing just the right platform for the same. SHERM is an Electronic Health record management system aiming at providing efficient data collection techniques paralleled with better medication than being obtained today. It is an extremely user-friendly solution. SHERM is accompanied by efficient Machine Learning algorithms to run analysis upon. The data fodder for the algorithm populates itself with time, requiring no extra man power. Finally, SHERM gives customized analysis facilitating government to send medicines to the disease prone area and also for pharmaceutical companies to target just the right market for their product.

Contents

List of Figures

List of Table

Chapter 1

Introduction

1.1 Motivation

Combination of Healthcare and technology are starting to become the next big thing for mankind. Non-Communicable diseases are one of the biggest reasons for mortalities in India. Initial Survey highlighted a face that **61%** of deaths in India are due to Non-Communicable diseases (majority of those being due to Diabetes, Cancer and cardio vascular diseases).

It was enough motivation to work in this domain, having seen a couple of cases of chronic NCDs run in our own families. This resulted in lifelong medications, coupled with continuous misery.

One of the major problems with the existing solutions is NOT LEVERAGING stored data for performing analysis. In India, there is no regular system for collecting data on NCDs which can be said to be of adequate coverage or quality.

Terabytes of patient data is collected by hospitals and pharma companies, but is hardly being put to proper use for diagnosis of patient ailments.

Absence of efficient data capturing environments and mechanisms of data analysis increases the gap between the traditional methods of diagnosing and the futuristic methods of data capture and analysis for the betterment of the health sector.

Also, healthcare industry is one such industry which is still a bit hesitant in incorporating digital methods in their workflow, over the traditional pen and paper approach (Mainly PRESCRIPTIONS).

Lastly, having family members with expertise in Pharma Industry, that provided motivation to dive deep into this mammoth problem agonizing the whole world, and especially India.

1.2 Scope

The userbase for this particular software is expected to be very vast and diverse. The main targeted users of this app are – Doctors, Patients, Clinics, Hospitals, Pharmacies, Chemists and Pharmaceuticals companies.

The primary goal of this project is to make a scalable solution which can easily meet the requirements of the traditional healthcare systems and also bridge the gap between technology and healthcare.

- Doctor – The doctor can access the medical history of the patient and diagnose the patient accordingly.
- Patient – The patient can use the application to give doctor access to his health history and also can use the app to access his/her own medical prescriptions right on their phone.
- Data Analysts – The analysts can run the attached machine learning algorithms to understand the trends of the disease in the given demographic region.

1.3 Objectives

The primary objectives of this project are :-

- Leverage the patient healthcare data, which otherwise would've laid dormant in some database.
- Couple the collected data with suitable ML algorithms and extract useful information from the data.
- To give patient access to his own medical prescription history, right on his/her handheld device, at any given moment in time.
- Allowing Doctor-Patient interaction by empowering the patient to share their medical profile with doctor/clinic/hospital digitally.
- Personalized analysis of the data collected.

1.4 Proposed Model

The patient will be provided with a mobile application, using which he/she can view their medical prescriptions and grant the doctor access to their medical data.

The doctor will be provided with a web application, by which they can create a new patient, view patients basic medical profile, view patient’s old healthcare data and add prescriptions to patient’s profile

The analyst will be provided with a dashboard wherein he/she can view information of importance to them, like the analysis performed.

In the backend, this system runs ML algorithms on the diabetes data, so as to produce reports as per analyst needs.

1.5 Organization of Report

In order to explain the developed system, the following sections are covered:

- **Literature Review** describes the study of the existing systems and techniques taken into account prior to development of the proposed system.
- **System Analysis and Design** provides a detailed walk through of the software engineering methodology adopted to implement the model, an overview of the system and the modules incorporated into the system
- **Modelling and Implementation** provides a deeper insight into the working of the model. The various modules and their interactions are depicted using relevant descriptive diagrams.
- **Testing** the model to ensure bug/error free model along with the **Results** obtained. **Discussion** then provides detailed analysis on quality assurance measures.
- **Conclusion** about the Results obtained after successfully running the model and **Future Scope** of the model is highlighted.

Chapter 2

Literature Review

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood sugar. Hyperglycaemia, or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels.

In 2014, 8.5% of adults aged 18 years and older had diabetes. In 2016, diabetes was the direct cause of 1.6 million deaths and in 2012 high blood glucose was the cause of another 2.2 million deaths. WHO aims to stimulate and support the adoption of effective measures for the surveillance, prevention and control of diabetes and its complications, particularly in low and middle-income countries. To this end, WHO:

- Provides scientific guidelines for the prevention of major noncommunicable diseases including diabetes;
- Develops norms and standards for diabetes diagnosis and care;
- Builds awareness on the global epidemic of diabetes, marking World Diabetes Day (14 November); and
- Conducts surveillance of diabetes and its risk factors.

According to infographics produced by The Centers for Disease Control and Prevention (CDC), it illustrates the estimates for diabetes, prediabetes, the cost of diabetes (dollars, risk of death, medical costs), specifics about type 1 and type 2 diabetes and risk factors for type 2 diabetes.

A Markov model was constructed by Elbert S. Huang, Anirban Basu, Michael O'Grady and James C. Capretta [1] simulating individuals' movement across different BMI categories, the incidence of diabetes and screening, and the natural history of diabetes and its complications over the next 25 years. Prevalence and incidence of obesity and diabetes and the direct spending on diabetes care and complications are projected. The study population is 24- to 85-year-old patients characterized by the Centers for Disease Control and Prevention's National Health and Nutrition Examination Survey and National Health Interview Survey.

A new approach to population health was proposed by Narges Razavian, Saul Blecker, Ann Marie Schmidt, Aaron Smith-McLallen, Somesh Nigam, and David Sontag [2], in which data-driven predictive models are learned for outcomes such as type 2 diabetes. This approach enables risk assessment from readily available electronic claims data on large populations, without additional screening cost. Proposed model uncovers early and late-stage risk factors.

Research was done by Schmittdiel JA, Dyer WT, Marshall CJ, Bivins R [3] on predictors of clinical outcomes that usually focuses on the impact of individual patient factors, despite known relationships between neighborhood environment and health. Retrospective cohort study of all 157,752 patients aged 18 years or older from Kaiser Permanente Northern California with laboratory-defined prediabetes (fasting plasma glucose, 100 mg/dL-125 mg/dL, and/or glycated hemoglobin, 5.7%-6.4%). It was assessed whether census data on education, income, and percentage of households receiving benefits through the US Department of Agriculture's Supplemental Nutrition Assistance Program (SNAP) was associated with diabetes development using logistic regression controlling for age, sex, race/ethnicity, blood glucose levels, and body mass index.

The Liverpool Eye and Diabetes Study (LEADS) was conducted by Gerald Liew, Vincent W Wong, Mercy Saw, Tania E Tsang, Tim Nolan, Stephen Ong and I-Van Ho [4], which is a cross-sectional population-based study of patients with type 1 and type 2 diabetes in a multi-ethnic region of Sydney, Australia, to determine the population prevalence of OCT-defined DME, how this varies by ethnicity and association with systemic factors. This report describes the rationale, methodology and study aims.

A descriptive study was conducted by Chiyembekezo Kachimanga, Katie Cundale, Emily Wroe, Lawrence Nazimera, Arnold Jumbe, Elizabeth Dunbar and Noel Kalanga [5] examining 3 screening programmes for NCDs in Neno, Malawi, that were implemented from June 2015 to December 2016. The NCD screening models were integrated into existing platforms, utilizing regular mass screening events in the community, patients awaiting to be seen in a combined NCD and HIV clinic, and patients awaiting treatment at outpatient departments (OPDs).

The authors John W. Stanifer, Charles R. Cleland , Gerald Jamberi Makuka, Joseph R. Egger, Venance Maro, Honest Maro, Francis Karia, Uptal D. Patel, Matthew J. Burton and Heiko Philippin [6] have conducted a stratified, cluster-designed, serial cross-sectional household study from 2014–2015 in the Kilimanjaro Region, Tanzania. A three-stage cluster probability sampling method was used to randomly select individuals. To estimate prevalence, individuals were screened for glucose impairment, including diabetes, using hemoglobin A1C. Individuals were also screened for hypertension and obesity, and to assess for potential complications, individuals with diabetes were assessed for retinopathy, neuropathy, and nephropathy.

The measurement of SUA levels suggested by Parul Thakur, Ashwini Kumar, Pradeep Kumar Patra and Awanish Kumar [7] could play a valuable role as predictor in early type 2 diabetics as well as a potent antioxidant therapeutic. This work is a small demographic study in the patients of Chhattisgarh to find a relation between the SUA level and diabetes parameters in diabetics and controls. This was a small sample set case–control study. Patients were divided into two groups, namely, control (n = 25) and type 2 diabetics (n = 30). Biochemical estimation of parameters was performed using commercially available enzymatic kits. It was found that Plasma glucose, serum triglyceride, glycated hemoglobin and creatinine were higher in diabetic patients than in controls.

A cross sectional study was conducted by Julio Baldisserotto, Luciane Kopittke, Fulvio Borges Nedel, Silvia Pasa Takeda, Claunara Schilling Mendonça, Sérgio Antonio Sirena, Margarita Silva Diercks, Lena Azeredo de Lima and Belinda Nicolau [8] which presents data from a longitudinal research. 3784 adults were randomly selected from the registry of a health service in Porto Alegre, Brazil. The eligibility criteria were: confirmed diagnosis of hypertension and/or diabetes, consulted at least once in the prior 3 years and 18 years of age or older. Home data collection consisted of a questionnaire with information on demographic, medical history, lifestyle and socio-economic factors.

Chapter 3

System Analysis and Design

3.1 Workflow

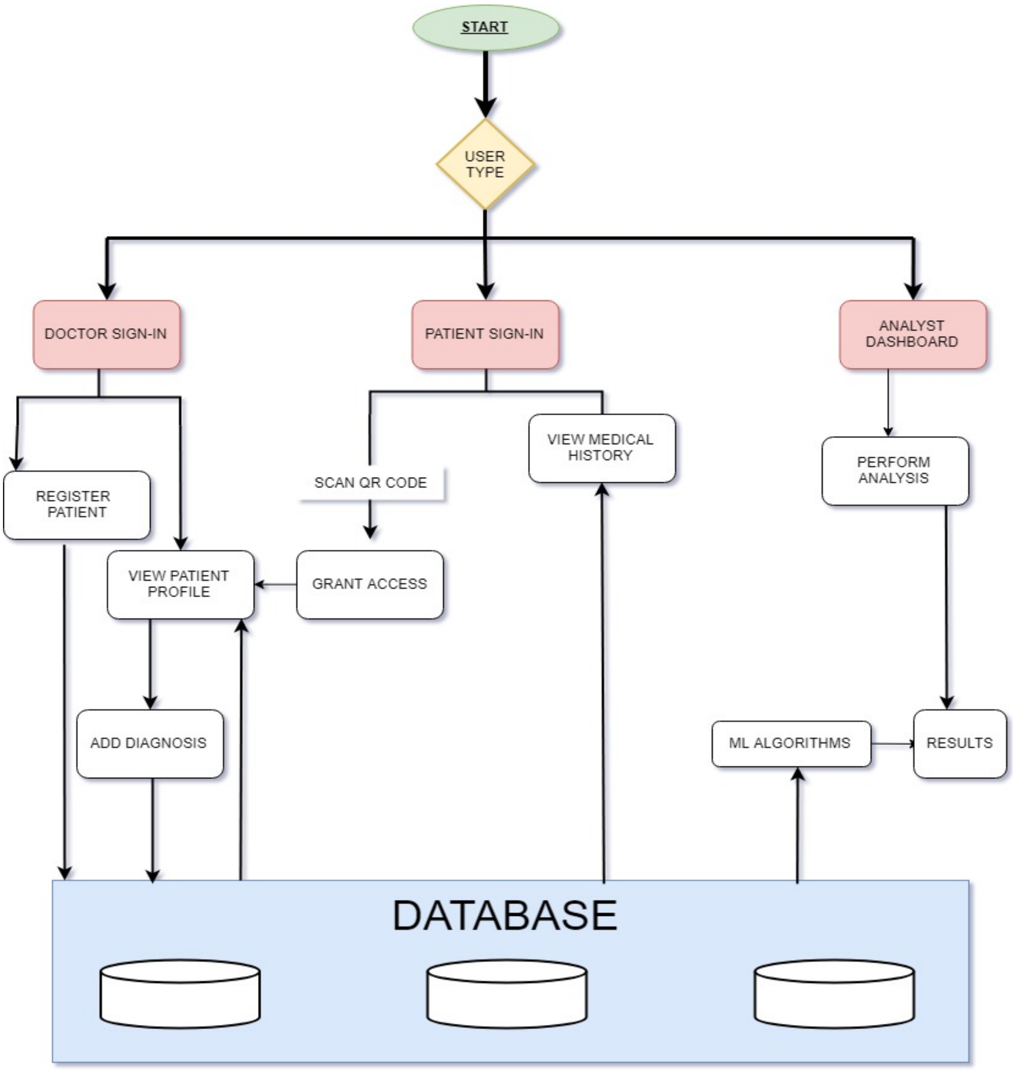


Figure 3.1: Workflow Diagram

Figure 3.1 shows the Wokflow diagram of the system, with a clear demarcation between the different stakeholders of the system.

Starting with the Doctor, the doctor has access to a web portal, where he/she can sign in. Upon sign-in a landing page/Dashboard is rendered, wherein he can see the list of patients under him. Point to be noted here is the fact that the patient name will only appear on his screen if the patient has scanned the QR code on his system with his mobile app. Once the name is there and the doctor clicks on the name, the doctor can see the basic patient profile and also previous diagnosis performed on the patient.

Coming to Patient, the patient has access to a mobile application, which requires patient sign-in. Once the patient has signed in, the patient can perform two actions. They can view their medical history, wherein all the diagnosis performed by Doctors will be visible for their own reference. Also, they can scan QR code on Doctor's Dashboard whilst visiting them.

The last stakeholder of the system is Analyst, who can view all the data in the system and run analysis on the same.

3.2 Data Description

The dataset that was chosen for this project was chosen from "Mendeley Data" [9] .This diabetes dataset consists of 6156 tuples and a total of 67 attributes, with the target value of HbA1c . There are 19 categorical attributes like HLL_category, PHN_category, CI_category, Alcohol_Amount_category, BP_category etc. There are 44 other features like Sex, Age, Height, Weight, BMI, Smoking Status, Diastolic BP etc. The ethnicity of this dataset is Japanese and is collected over a time span of 3 years.

3.3 Preprocessing of Data

The class variable "HbA1c_category" had a value from (1,2,3,4,5). Since we are interested in binary classification, if the value is 1, then the person is deemed as non-diabetic and if the value is either 2,3,4 or 5, then the person is deemed as diabetic.

The dataset consists of a lot of NULL values, which can lead to a bias in the results. As a result, the dataset was cleaned by dropping unnecessary attributes and also by dropping the tuples with null value of one or more attribute.

As a consequence of the above operation the number of tuples in the dataset are now reduced to 5267, and the number of attributes to 53. The processed dataset is stored in a new file named "diabetes_cleaned.csv".

Of the 53 attributes obtained, upon running RFE (Recursive Feature Elimination), attributes of most significance were found to be - Weight, BMI, Taking medication for hypertension, exercise more than 30 mins, waist_Circumference, physical_Activity, quick_Walking, Age.

Also, it was noticed that the dataset is skewed, that is it consists more tuples with target value as 0 and less tuples with target value as 1. Fitting this data lead to very low recall value, which is extremely undesirable. To tackle this issue, ADASYN (Adaptive Synthetic Sampling Approach for Imbalanced Data) algorithm [10] was run, which removed the skew in the dataset.

The final dataset used for training and testing purposes is thus stored in a file named “diabetes_cleaned_balanced.csv”.

3.4 Performance Check Measures

In this project performance measures such as F1-Score, accuracy, recall and precision are used. The following elements were used to calculate the above measures: TP (true positive), TN (true negative), FP(false positive) and FN(false negative).

- TP=positive values correctly recognized as positive.
- TN=negative values correctly recognized as negative.
- FP=negative values recognized as positive.
- FN=positive values recognized as negative.

Recall

It is the ratio of correctly predicted positive values out of all positive samples. It is also known as True Positive Rate and is given by:

Precision

It is the ratio of correctly identified positive samples as positive out of all positive predictions. It is also called as positive prediction value and is given by:

F1-Score

It is a parameter used to obtain balance between Recall and Precision value.

Accuracy

Accuracy is the simplest measure. It is the ratio between the number of correctly classified test inputs AND the total number of test inputs.

3.5 Technical Stack

Firebase	Open Source low latency Database Dependency
ReactJS [11]	Front-End library
Ionic Framework [12]	Development of mobile app for the patient
GCP	For employing their OCR APIs
Sci-Kit Learn	Python’s well documented Open Source ML Library
QR Code	For Access Control

Table 1: Tech-Stack used for the development of the project

Chapter 4

Modelling and Implementation

Implementation is the process of converting the designed system architecture into working modules where it is made sure that all the functional and non-functional requirements are met.

4.1 Algorithms

4.1.1 Naive Bayes Classifier

Bayes theorem works on conditional probability. Conditional probability is the probability that something will happen, given that something else has already occurred. Using the conditional probability, we can calculate the probability of an event using its prior knowledge.
Below is the formula for calculating the conditional probability. [13]

Where :

- P(H) is the probability of hypothesis H being true. This is known as the prior probability.
- P(E) is the probability of the evidence(regardless of the hypothesis).
- P(E|H) is the probability of the evidence given that hypothesis is true.
- P(H|E) is the probability of the hypothesis given that the evidence is there.

Pseudocode:

- Create a frequency table for all the features against the different classes.
- Draw the likelihood table for the features against the classes.
- Calculate the conditional probabilities for all the classes.
- Calculate the maximum probability.

Pros and Cons of Naive Bayes classifier

Pros:

- Naive Bayes Algorithm is a fast, highly scalable algorithm.
- Naive Bayes can be used for Binary and Multiclass classification.
- Naive Bayes is a choice for Text Classification problems. It's a popular choice for spam email classification.

Cons:

- It considers all the features to be unrelated, so it cannot learn the relationship between features.
- It has a 'Zero conditional probability Problem', for features having zero frequency the total probability also becomes zero.There are several sample correction techniques to fix this problem such as "Laplacian Correction."

4.1.2 Random Forest

Random forests [14], also known as random decision forests, are a popular ensemble method that can be used to build predictive models for both classification and regression problems. Ensemble methods use multiple learning models to gain better predictive results — in the case of a random forest, the model creates an entire forest of random uncorrelated decision trees to arrive at the best possible answer.

Pros :

- Random forests have less variance than a single decision tree. It means that it works correctly for a large range of data items than single decision trees.
- Random forests are extremely flexible and have very high accuracy.
- They also do not require preparation of the input data. You do not have to scale the data.
- It also maintains accuracy even when a large proportion of the data are missing.

Cons :

- The main disadvantage of Random forests is their complexity. It is much harder and time-consuming to construct than decision trees.
- It also requires more computational resources and are also less intuitive. When there is a large collection of decision trees it is hard to have an intuitive grasp of the relationship existing in the input data.
- In addition, the prediction process using random forests is time-consuming than other algorithms.

The pseudo code for random forest algorithm can split into two stages.

- Random forest creation pseudo code.
- Pseudo code to perform prediction from the created random forest classifier.

First, let's begin with random forest creation pseudo code

Random Forest pseudo code:

- Randomly select "k" features from total "m" features, where k << m.
- Among the "k" features, calculate the node "d" using the best split point.
- Split the node into daughter nodes using the best split.
- Repeat 1 to 3 steps until "l" number of nodes has been reached.
- Build forest by repeating steps 1 to 4 for "n" number times to create "n" number of trees.

Random forest prediction pseudo code:

To perform prediction using the trained random forest algorithm uses the below pseudocode.

- Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target)
- Calculate the votes for each predicted target.
- Consider the high voted predicted target as the final prediction from the random forest algorithm.

4.1.3 Logistic Regression

It's a classification algorithm, that is used where the response variable is categorical. The idea of Logistic Regression [15] is to find a relationship between features and probability of particular outcome.

How it works:

Logistic Regression measures the relationship between the dependent variable (our label, what we want to predict) and the one or more independent variables (our features), by estimating probabilities using it’s underlying logistic function.

These probabilities must then be transformed into binary values in order to actually make a prediction. This is the task of the logistic function, also called the sigmoid function. The Sigmoid-Function is an S-shaped curve that can take any real-valued number and map it into a value between the range of 0 and 1, but never exactly at those limits. This values between 0 and 1 will then be transformed into either 0 or 1 using a threshold classifier.

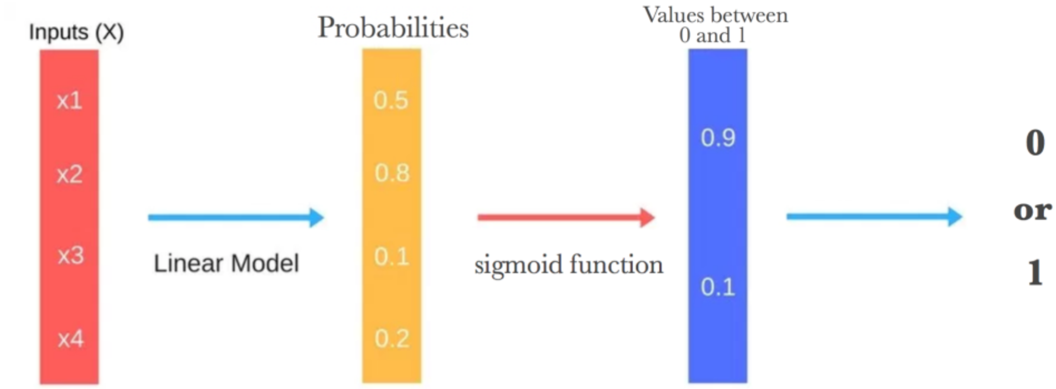


Figure 4.1 : Working of Logistic Regression

Figure 4.1 illustrates the steps that logistic regression goes through to give you your desired output.

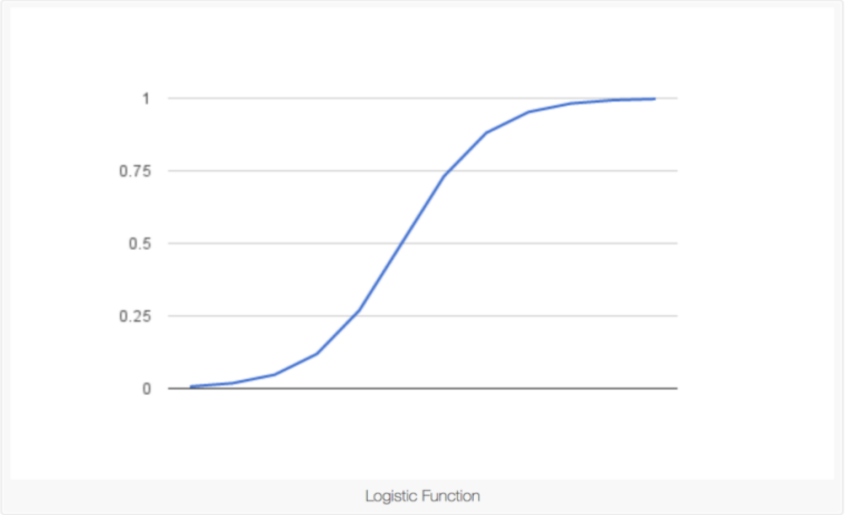


Figure 4.2 : Sigmoid Function

Figure 4.2 shows how the logistic function (sigmoid function) looks like.

Pros and Cons associated with Logistic Regression

Pros:

- It is incredibly easy to implement and very efficient to train.
- Logistic Regression is also a good baseline that can be used to measure the performance of other more complex Algorithms.
- It provides probability score for observations.
- Logistic Regression works better when attributes are removed that are unrelated to the output variable as well as attributes that are very similar (correlated) to each other.

Cons:

- Doesn't handle large number of categorical features/variables well.
- It requires transformation of non-linear features.

4.1.4 XG Boost

XGBoost [16], short for "Extreme Gradient Boosting", was introduced by Chen in 2014. Since its introduction, XGBoost has become one of the most popular machine learning algorithm. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

GBM divides the optimization problem into two parts by first determining the direction of the step and then optimizing the step length. Different from GBM, XGBoost tries to determine the step directly by solving:

for each x in the data set. By doing -order Taylor expansion of the loss function around the current estimate x), we get;

L(y,+≈ L(y,+

where) is the gradient, same as the one in GBM, and is the Hessian (second order derivative) at the current estimate:

f(x)=

The two reasons to use XGBoost are also the two goals of the project:

1. Execution Speed - Its execution is really fast when compared to other implementations of gradient boosting.
2. Model Performance.

How it works:

In XG Boost, model is fit on the gradient of loss generated from the previous step. In XG Boost, the gradient boosting algorithm is modified so that it works with any differentiable loss function.

4.1.5 K-Nearest Neighbors

The K-nearest neighbors (KNN) [17] algorithm is a type of supervised machine learning algorithms. It is a lazy learning algorithm since it doesn't have a specialized training phase. Rather, it uses all of the data for training while classifying a new data point or instance. KNN is a non-parametric learning algorithm, which means that it doesn't assume anything about the underlying data. This is an extremely useful feature since most of the real world data doesn't really follow any theoretical assumption e.g. linear-separability, uniform distribution, etc.

How it works:

The intuition behind the KNN algorithm is one of the simplest of all the supervised machine learning algorithms. It simply calculates the distance of a new data point to all other training data points. The distance can be of any type e.g Euclidean or Manhattan etc. It then selects the K-nearest data points, where K can be any integer. Finally, it assigns the data point to the class to which the majority of the K data points belong.

Suppose you have a dataset with two variables, which when plotted, looks like the one in the Figure 4.3.

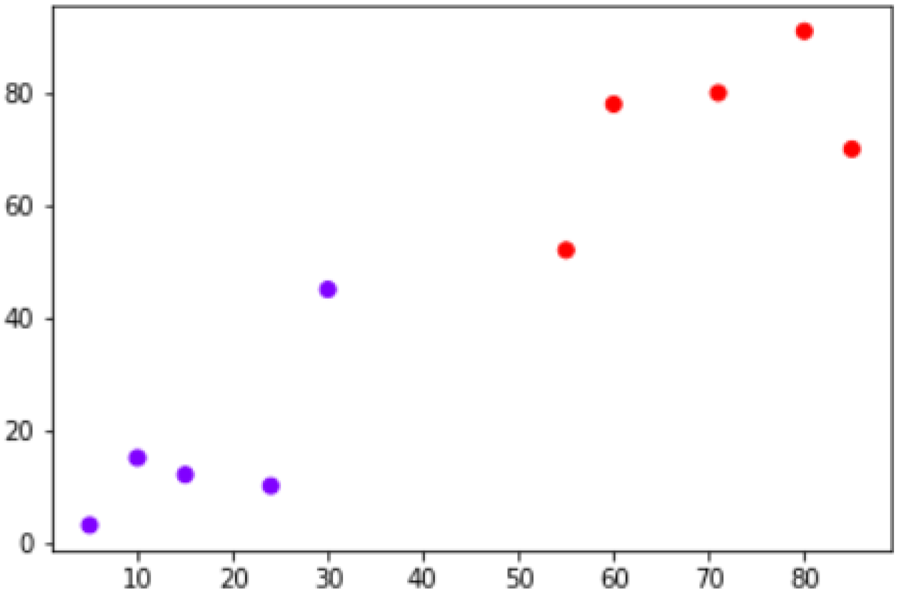


Figure 4.3 : Example Dataset for KNN

The task is to classify a new data point with 'X' into "Blue" class or "Red" class. The coordinate values of the data point are x=45 and y=50. Suppose the value of K is 3. The KNN algorithm starts by calculating the distance of point X from all the points. It then finds the 3 nearest points with least distance to point X. This is shown in the figure below. The three nearest points have been encircled.

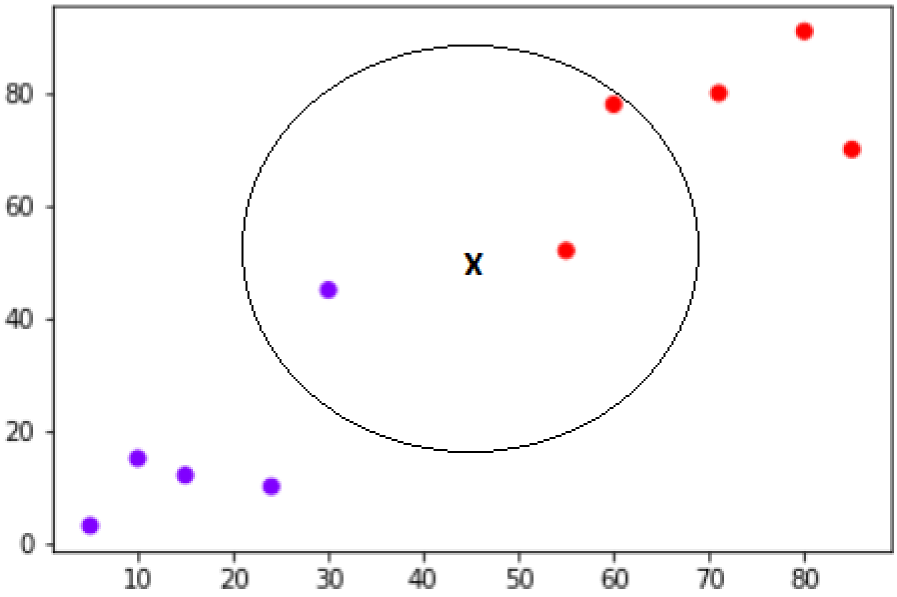


Figure 4.4 : Classification through KNN

The final step of the KNN algorithm is to assign new point to the class to which majority of the three nearest points belong. From the Figure 4.4 we can see that the two of the three nearest points belong to the class "Red" while one belongs to the class "Blue". Therefore, the new data point will be classified as "Red".

Pros and Cons associated with KNN

Pros:

- It is extremely easy to implement

- As said earlier, it is lazy learning algorithm and therefore requires no training prior to making real time predictions. This makes the KNN algorithm much faster than other algorithms that require training e.g SVM, linear regression, etc.
- Since the algorithm requires no training before making predictions, new data can be added seamlessly.
- There are only two parameters required to implement KNN i.e. the value of K and the distance function (e.g. Euclidean or Manhattan etc.)

Cons:

- The KNN algorithm doesn't work well with high dimensional data because with large number of dimensions, it becomes difficult for the algorithm to calculate distance in each dimension.
- The KNN algorithm has a high prediction cost for large datasets. This is because in large datasets the cost of calculating distance between new point and each existing point becomes higher.
- Finally, the KNN algorithm doesn't work well with categorical features since it is difficult to find the distance between dimensions with categorical features.

4.2 Use Case Diagram

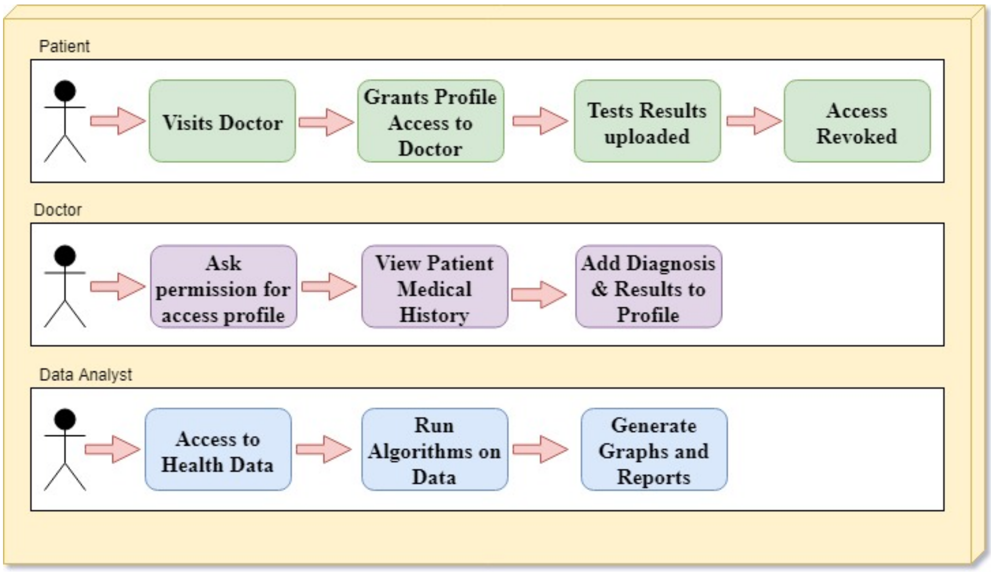

Patient
Visits Doctor
Grants Profile Access to Doctor
Tests Results uploaded
Access Revoked
Doctor
Ask permission for access profile
View Patient Medical History
Add Diagnosis & Results to Profile
Data Analyst
Access to Health Data
Run Algorithms on Data
Generate Graphs and Reports

Figure 4.5: Use Cases Diagram

This project has majorly three stakeholders:

- Patients
- Doctors
- Data Analysts

When the patients visit the doctor, he has to scan the QR Code of the doctor. In this way the patient grants the access to his medical history to the doctor. Various medical tests which occur at diagnosis are also uploaded to the profile. Once the patient finishes his visit, he revokes access of his data.

The doctor on the other hand, after getting access to the patient's medical data can review it and diagnose him accordingly. He adds the diagnosis in the patient's profile.

Data Analysts are the stakeholders which receive all the data collected except the personal data of patients like name, address, contact number. On this collected data various ML algorithms are run to provide data analytics. They can generate various graphs and reports like demographic spread of a disease, diabetic vs non diabetic people, attributes contributing to diabetic case etc.

4.3 Sequence Diagram

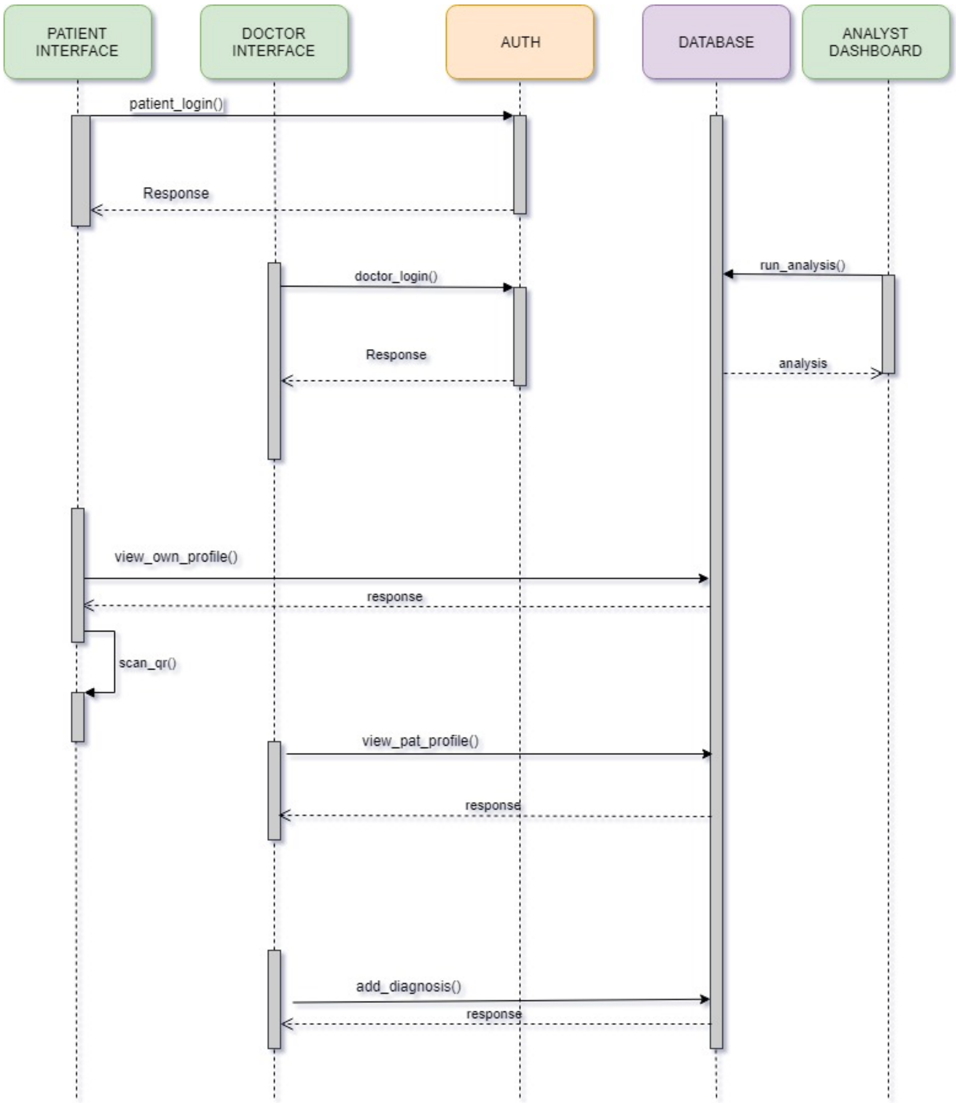


Figure 4.6: Sequence Diagram

Description: The Figure 4.6 shows the object interactions arranged in time sequence. The patient will login to his mobile app, and also will the doctor, to his web app. The patient can view their own profile. They can scan QR code on Doctors dashboard so as to grant access. Once the QR code has been scanned only then can the doctor view the patient profile. Once patient profile is viewed the doctor can also add diagnosis to the patient's profile, which will be appended in the common database.

A completely independent stakeholder, analyst also has the access to the database, from where he/she can fetch the data and run algorithms so as to make sense of the data, which otherwise would've laid dormant in the database.

Chapter 5

Testing, Results and Discussion

5.1 Testing

Testing is one of the ways of assessing the system which helps to detect the quality of the software, the methods we follow and evaluate the expected output and the actual input.

The project involved application of several different machine learning algorithms, namely – Random Forest, Naïve Bayes, Nearest Neighbour, Logistic Regression, XGBoost, making it obligatory to test all the algorithms applied. For testing purposes, standard sklearn.train_test_split is used. The dataset was split into 25% testing data and 75% training data.

5.2 Results and Discussion

The results of the various applied algorithms are presented here with the help of various performance metrics such as accuracy, precision, recall and F1-Score in the form of graphs.

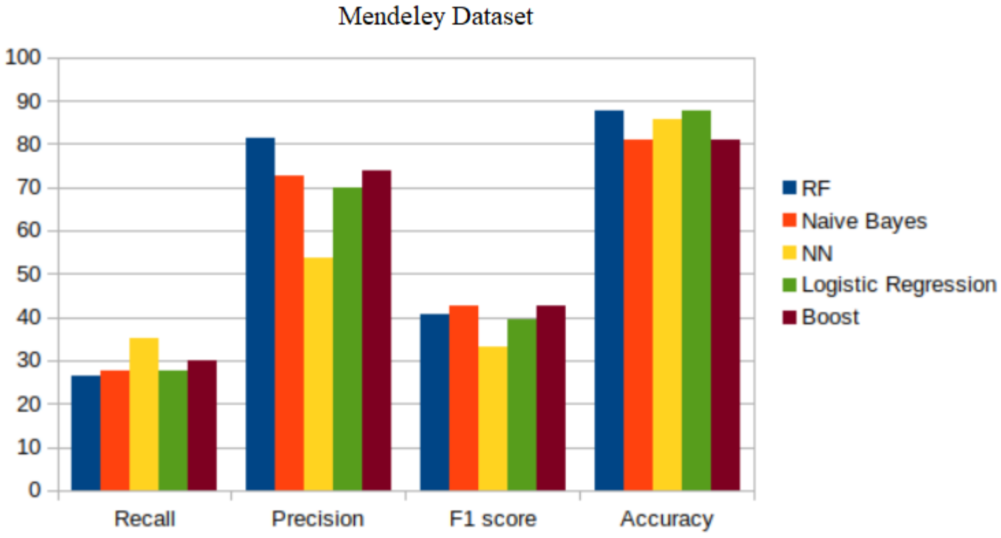


Figure 5.1- Results after applying algorithms on unbalanced dataset

Figure 5.1 shows the results obtained after applying the algorithms on unbalanced dataset. From Figure 5.1 it is evident that Nearest Neighbour algorithm yields the best Recall Value (i.e. the value of our interest).

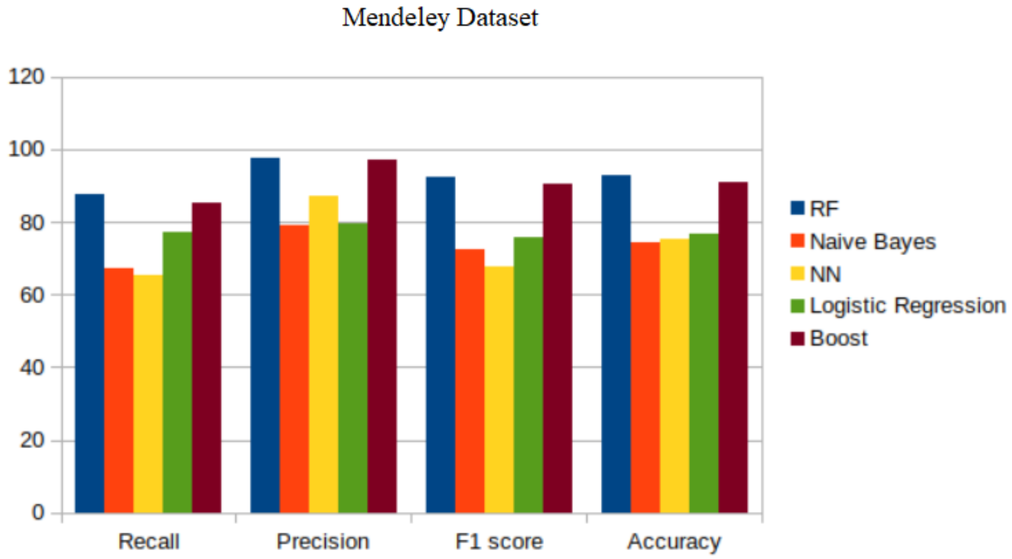


Figure 5.2 - Results after applying algorithms on balanced dataset

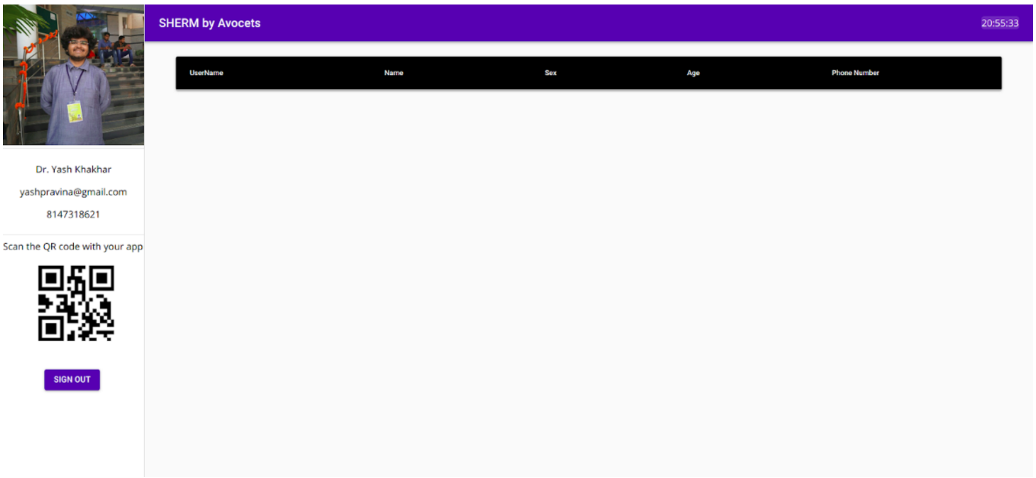
Figure 5.2 shows the results obtained after applying the algorithms on balanced dataset. Data Balancing was done using Adaptive Synthetic (ADASYN) sampling approach. From figure 5.2 it is evident that Random Forest yields the best results with high recall value.

Comparing Figure 5.1 and Figure 5.2, application of a sampling algorithm on the dataset generated far more superior results than not applying it.

The values obtained after application of the best approach, i.e. by applying RandomForest on the dataset generated after Sampling: -

- **Accuracy - 92.6**
- **Precision - 97.51**
- **Recall - 87.57**
- **F1 - 92.27**

5.3 INTERFACES



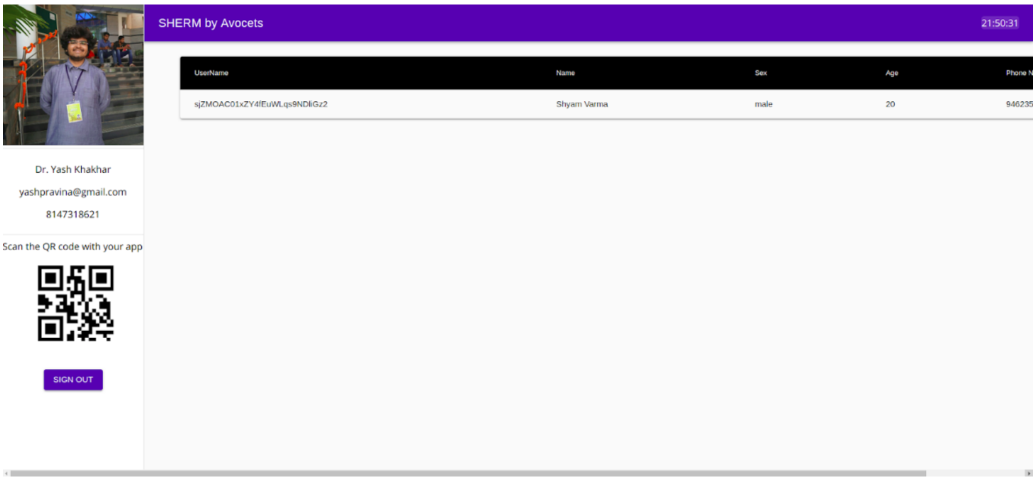
UserName	Name	Sex	Age	Phone Number

Dr. Yash Khakhar
yashpravina@gmail.com
8147318621

Scan the QR code with your app

Figure 5.3 - Doctor Dashboard post Sign in

Figure 5.3 shows screenshot of the doctor dashboard after signing in. Notable features include the QR code in the bottom left corner, which will be scanned by the patient while visiting the doctor.



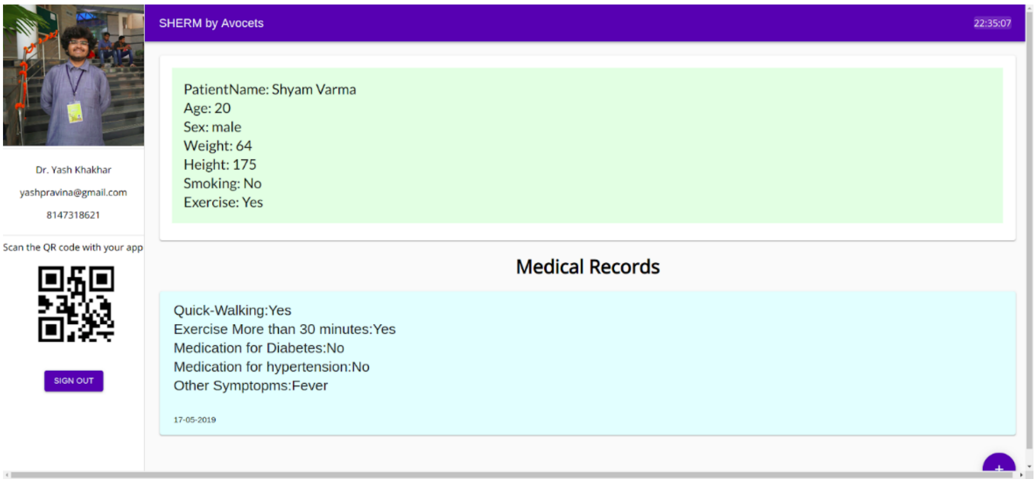
UserName	Name	Sex	Age	Phone N
sjZMOAC01xZY4fEuWLqs9NDliGz2	Shyam Varma	male	20	946235

Dr. Yash Khakhar
yashpravina@gmail.com
8147318621

Scan the QR code with your app

Figure 5.4 - Doctor dashboard state after patient scans QRCode

Figure 5.4 shows the dashboard state after a visiting patient scans the QR code. The diagram shows that the patient with name “Shyam Varma” is visiting the doctor, and thus his details are visible.



SHERM by Avocets 22:35:07

Dr. Yash Khakhar
yashpravina@gmail.com
8147318621

Scan the QR code with your app

SIGN OUT

PatientName: Shyam Varma
Age: 20
Sex: male
Weight: 64
Height: 175
Smoking: No
Exercise: Yes

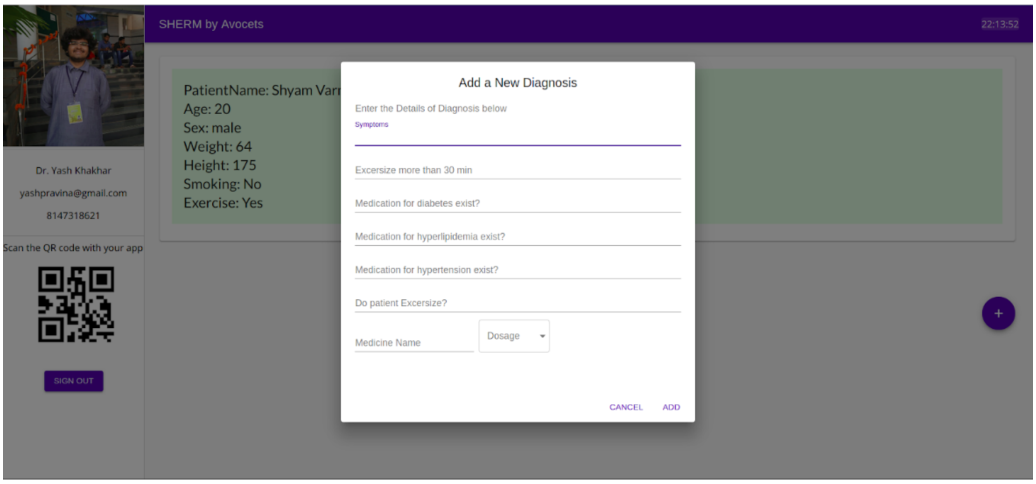
Medical Records

Quick-Walking:Yes
Exercise More than 30 minutes:Yes
Medication for Diabetes:No
Medication for hypertension:No
Other Symptopms:Fever

17-05-2019

Figure 5.5 - Doctor Dashboard after new prescription is added

Figure 5.5 shows the state after the doctor clicks on the patient name. The doctor can see previous medical records of the patient and can also add a prescription of his/her own to the patients profile.



SHERM by Avocets 22:13:52

Dr. Yash Khakhar
yashpravina@gmail.com
8147318621

Scan the QR code with your app

SIGN OUT

PatientName: Shyam Varı
Age: 20
Sex: male
Weight: 64
Height: 175
Smoking: No
Exercise: Yes

Add a New Diagnosis

Enter the Details of Diagnosis below

Symptoms

Excersize more than 30 min

Medication for diabetes exist?

Medication for hyperlipidemia exist?

Medication for hypertension exist?

Do patient Excersize?

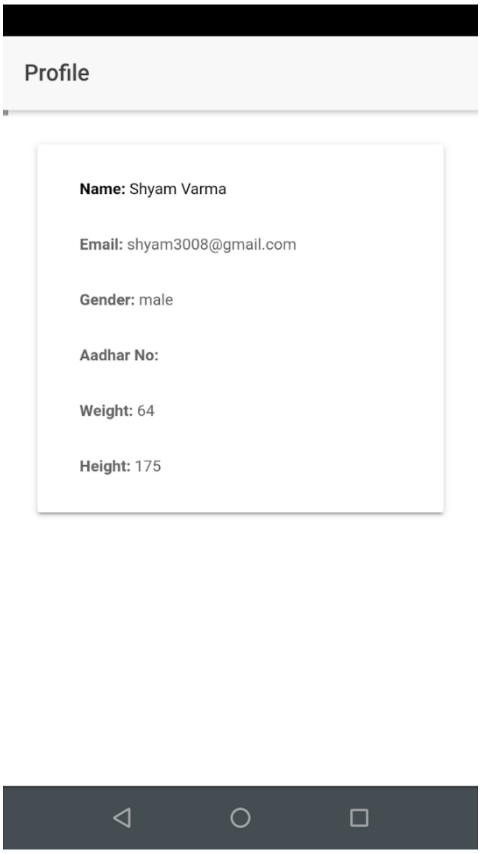
Medicine Name

Dosage

CANCEL ADD

Figure 5.6 - Doctor Dashboard to add a new prescription

Figure 5.6 shows the form using which the doctor can enter a new diagnosis to the patients profile.



Profile

Name: Shyam Varma

Email: shyam3008@gmail.com

Gender: male

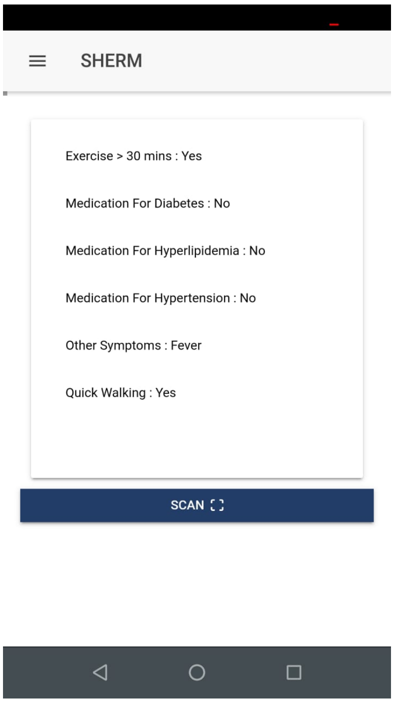
Aadhar No:

Weight: 64

Height: 175

Figure 5.7 - Patient Mobile App after Login

Figure 5.7 shows the patient app screenshot once the patient has logged in into his account.



Exercise > 30 mins : Yes

Medication For Diabetes : No

Medication For Hyperlipidemia : No

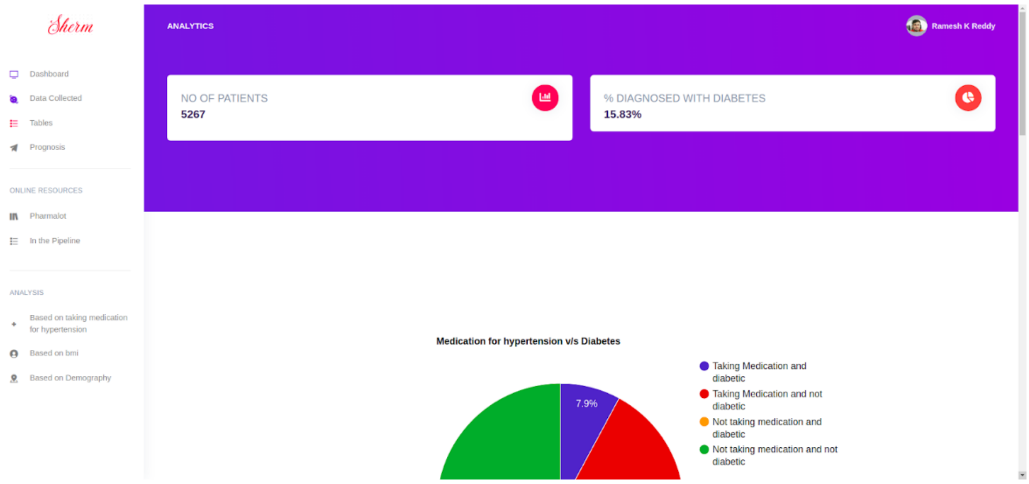
Medication For Hypertension : No

Other Symptoms : Fever

Quick Walking : Yes

Figure 5.8 - Patient App after new diagnosis is added

Figure 5.8 shows the app screenshot of all the medical records of the patient. From this page the patient also gets an option to scan QR code, using which he/she can scan the QR code on the doctors dashboard,


Sherm
Dashboard
Data Collected
Tables
Prognosis
ONLINE RESOURCES
Pharmalot
In the Pipeline
ANALYSIS
Based on taking medication for hypertension
Based on bmi
Based on Demography
ANALYTICS
Ramesh K Reddy
Medication for hypertension v/s Diabetes
7.9%
Taking Medication and diabetic
Taking Medication and not diabetic
Not taking medication and diabetic
Not taking medication and not diabetic

NO OF PATIENTS
5267

% DIAGNOSED WITH DIABETES
15.83%

Figure 5.9 - Analyst Dashboard

Figure 5.9 shows the analyst dashboard wherein the analyst can go through all the data and also go through the results obtained after application of Machine Learning Algorithms.

Chapter 6

Conclusion and Future Work

In the present world, patient data stored in most hospitals is not being judiciously used for betterment, rather lays dormant in their database, which if used properly can help tackle this colossal problem.
The solution provided in this project, aimed at resolving the above stated problem. Applying Machine Learning algorithms on the data, led to a lot of insightful results, all of which can be used for the betterment of overall community health.

Since this was just an attempt done in limited time, there are a lot of challenges to be addressed in near future. They are :-

- Dynamic running of the model – The algorithms will run on the fresh data collected everyday and will generate daily reports.

- Make differentially-abled friendly – Being on the biggest patient base, it would be extremely helpful if differentially abled people could benefit from the project. This can be accomplished by using technologies such as RFID.

- Modifying software so as to make it of use to different classes of stakeholders – As of now the software is specific to Hospitals, but modifications can lead to the indication of software with fitness industries as well as wellness industries.

- Adding OCR in the system – As of now, the doctors are expected to make patient health data entry through their system, which breaks the traditional doctor-patient relationship (i.e. this pen-based prescription approach). This problem can be tackled by using OCR, wherein the prescription written by doctor, will be scanned and converted to digital form.

References

[1] http://care.diabetesjournals.org/content/32/12/2225

[2] https://www.liebertpub.com/doi/full/10.1089/big.2015.0020

[3] https://www.ncbi.nlm.nih.gov/pubmed/30296398

[4] https://bmjopen.bmj.com/content/9/1/e021884

[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5610274/

[6] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0164428

[7] http://www.aihbonline.com/article.asp?issn=23218568;year=2017;volume=-7;issue=3;spage=124;epage=129;aulast=Thakur

[8] https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-016-32-30-7

[9] https://data.mendeley.com/datasets/g7gzwd4c7h/1

[10] https://sci2s.ugr.es/keel/pdf/algorithm/congreso/2008-He-ieee.pdf

[11] https://reactjs.org/docs/getting-started.html

[12] https://ionicframework.com/docs/v3/

[13] https://www.hackerearth.com/blog/machine-learning/introduction-naive-b-ayes-algorithm-codes-python-r/

[14] https://www.analyticsvidhya.com/blog/2015/06/tuning-random-fo-rest-mo-del/

[15] https://www.experfy.com/blog/the-logistic-regression-algorithm

[16] https://machinelearningmastery.com/develop-first-xgboost-model-python-scikit-learn/

[17] https://stackabuse.com/k-nearest-neighbors-algorithm-in-python-and-sciki-t-learn/