

CARDIAB
CARDIOVASCULAR DISEASE & DIABETES PREDICTION
USING ML

Dissertation submitted to
Shri Ramdeobaba College of Engineering &
Management, Nagpur in partial fulfillment of
requirement for the award of
degree of

Bachelor of Engineering

In

COMPUTER SCIENCE AND ENGINEERING

By

Shefali Jindal, Tithi Agrawal, Abbas Husain,
Atharva Nimbalwar, Shubham Saboo

Guide

Prof. Heena Agrawal
Dept. of Computer Science & Engineering



Dept. of Computer Science and Engineering
Shri Ramdeobaba College of Engineering & Management, Nagpur
440 013

(An Autonomous Institute affiliated to Rashtrasant Tukdoji Maharaj Nagpur University
Nagpur)

April 2021

SHRI RAMDEOBABA COLLEGE OF ENGINEERING & MANAGEMENT,
NAGPUR

(An Autonomous Institute Affiliated to Rashtrasant Tukdoji Maharaj Nagpur University
Nagpur)

Department of Computer Science & Engineering

CERTIFICATE

This is to certify that the Thesis on **“CARDIAB- Cardiovascular diseases and Diabetes prediction using ML”** is a bonafide work of Shefali Jindal, Tithi Agrawal, Abbas Husain, Atharva Nimbawar and Shubham Saboo submitted to the Rashtrasant Tukdoji Maharaj Nagpur University, Nagpur in partial fulfillment of the award of a Bachelor of Engineering, in Project 1- CSP360 has been carried out at the Department of Computer Science & Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur during the academic year 2020-2021.

Date: April 27, 2021

Place: Nagpur

Prof. Heena Agrawal
Project Guide
Department of Computer Science
& Engineering

Prof. M. B. Chandak
H. O. D.
Department of Computer Science
& Engineering

Dr. R. S. Pande
Principal

DECLARATION

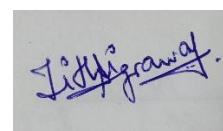
I, hereby declare that the thesis titled “**CARDIAB- Cardiovascular diseases and Diabetes prediction using ML**” submitted herein, has been carried out in the Department of Computer Science & Engineering of Shri Ramdeobaba College of Engineering & Management, Nagpur. The work is original and has not been submitted earlier as a whole or part for the award of any degree / diploma at this or any other institution / University.

Date: April 27, 2021

Place: Nagpur



A07 Shefali Jindal



A14 Tithi Agrawal



A32 Abbas Husain



A41 Atharva Nimbalwar



A76 Shubham Saboo

Approval Sheet

This thesis/dissertation/report entitled CARDIAB- Cardiovascular diseases and Diabetes Prediction using ML by Shefali Jindal, Tithi Agrawal, Abbas Husain, Atharva Nimbalwar & Shubham Saboo is approved for the degree of Bachelor of Engineering in Computer Science & Engineering.

Name & signature of Supervisor(s) Name & signature of External Examiner(s)

Name & signature RRC Members

Name & signature of HOD

Date: April 27, 2021

Place: Nagpur

ACKNOWLEDGEMENTS

We take the opportunity to thank everyone related, directly or indirectly, for the completion of this project successfully.

Firstly, we would take this opportunity to express our deepest gratitude and thanks to Dr. R. S. Pande, Principal, Shri Ramdeobaba College of Engineering and Management, Dr. Manoj B Chandak, Head of the Department of Computer Science and Engineering and Prof. Heena Agrawal, Project guide for their undivided support, morally and physically, assistance, guidance, tolerance, stimulating guidance, continuous encouragement and supervision throughout the course of present work which proved to be invaluable as to completion of our project.

Deepest thanks and appreciation to our colleagues and the team members who spent nights and days, for their cooperation, encouragement, constructive suggestion and full of support for the project completion, from the beginning till the end. Every team member has put their best efforts in every possible way for Innovation, Creativity and Enthusiasm in the project. We have collectively come forward in challenging situations and overcome them with all support and spirit.

We would like to heartily thank all the respected staff members and our friends who have helped, inspired & supported us to carry out the project. We are highly obliged to work on this thesis named “CARDIAB- CVD & Diabetes prediction using machine learning”.

ABSTRACT

Arduous times like today's require utmost medical attention by people of majorly all age groups. Cardiovascular diseases, a class of diseases that involve the heart or blood vessels, and diabetes, a disease that occurs when the blood glucose level is too high, hold a powerful significance in the effect of Covid-19 disease. Clinical studies have also reported that pre-existing cardiovascular diseases are linked with worse outcomes and increased risks of death in patients with Covid-19. Also, people with diabetes have been identified as being at increased risk of serious illness from Covid-19.

The focus of our project is to help users determine the chances of a person having diabetes or a CVD. The application includes a web portal which will take various health attributes like BMI, blood pressure, insulin, age, drinking habits, smoking habits, etc. as inputs. This will go through ML algorithm and display the likelihood of the person having diabetes or CVD. The portal will also display necessary precautions to be taken if the chances are high and general health tips otherwise. Eventually, a Generate Report button will be present on the results page and the person's report will be mailed to him/her as a result.

Keywords: *Cardiovascular Disease, Diabetes, Machine Learning, KNN, Decision Trees, SVM, Random Forest Classifier, Logistic Regression, Accuracy*

TABLE OF CONTENTS

Chapter	Page No.
1. Introduction	1
2. Literature Review	2
3. Methodology	5
4. Implementation	8
4.1. Frontend [Web Application]	8
4.2. Development [ML Training Models]	8
4.3. Integration and Deployment	9
4.4. UML and DFD Diagrams	10
4.5. Working	15
5. Results	21
6. Conclusion	22
7. References	23
8. Appendix	24

List of figures and tables

Figure	Page No.
1. Use Case Diagram	10
2. Activity Diagram	11
3. State-chart Diagram (Frontend)	12
4. Data flow diagram	12
5. Sequence diagram	14
6. Home Page	15
7. CVD Page	15
8. CVD Result Page	16
9. CVD Report Generation Page	16
10. CVD Report	17
11. Diabetes Page	18
12. Diabetes Result Page	18
13. Diabetes Report Generation Page	19
14. Diabetes Report	20
15. Normal BMI Range	24
16. Normal Cholesterol Range	24
17. Normal Blood Pressure Range	25
18. Normal Glucose Range	25

Table

1. Accuracies of the two models on different algorithms (before hyperparameter tuning)	22
2. Accuracies of the two models on different algorithms (after hyperparameter tuning)	23

CHAPTER 1

INTRODUCTION

As we navigate through these challenging times, the most important thing is to stay healthy and recognize those who are at risk. As per a recent Lancet study, one in five people with pre-existing comorbidities or non-communicable diseases (NCDs) are at higher risk of severe Covid if they're infected. India has more than 180 million people suffering from cardiovascular disease, diabetes, and cancer, indicating the country must deal with the double burden of Covid and NCDs.

Specifically, for diabetes, unique health considerations and guidance are emerging. While people with diabetes are equally (if not more) likely at Covid risk, it can cause more severe symptoms and complications in some people living with diabetes, because the body's ability to fight off an infection is compromised.

Statistics reveal that cardiovascular diseases account for an estimated 31% deaths worldwide whereas an estimated 1.5 million deaths were directly caused by diabetes as of 2019. All these point to the world as it existed before two years, i.e., to the pre-covid times. As stated above, these two diseases have proven to exacerbate the COVID conditions prevailing globally. There are two major evidences for this claim.

The first is that pre-existing heart conditions, such as damaged heart muscle or blocked heart arteries, weaken the body's ability to survive the stress of the illness. A person with a vulnerable heart is more likely to succumb to the effects of fever, low oxygen levels, unstable blood pressures, and blood clotting disorders — all possible consequences of COVID-19 — than someone previously healthy.

A second explanation relates to poor underlying metabolic health, which is more common in those with heart disease. Poor metabolic health refers to diseases such as type 2 diabetes or prediabetes and obesity, which themselves cause inflammation and risk of blood clots, compounding the effects of COVID-19 and increasing the likelihood of devastating complications of COVID-19.

Hence, these diseases cannot be neglected. Moreover, as these diseases have several contributory risk factors, they are difficult to identify. CARDIAB will predict the probability of a person having a heart disease or diabetes. From the various

contributing risk factors, some of the most significant ones have been shortlisted and put as inputs in various machine learning algorithms like logistic regression, random forest classification, k nearest neighbor, etc. Accuracies of all these algorithms were compared and then the best algorithm was chosen.

CHAPTER 2

LITERATURE REVIEW

A quite significant amount of work related to the diagnosis of Cardiovascular heart disease and diabetes using machine learning algorithms has motivated this work. This chapter contains a brief literature survey of the same. A brief comparison of the existing work and our efficient contribution of the same is presented.

The Diabetes Model

1. From (Mahesh Barale et al 2016)

- They removed the cases with more than one value missing and then finally input this data on 534 out of 768 values which is very less for model training. Also, about 30% of the data is removed.
- We have worked on 757 out of 768 cases and removed only about 1.4% of the data. The missing values were input according to the skewness of the data by replacing the NaN values with mean or median to make it more consistent.

2. From (Naveen Kishore et al 2020)

- This research paper has illustrated the training of model using 4 algorithms with the following accuracies: SVM 73.43%, Random Forest 74.4%, KNN 71.3% and Logistic Regression 72.39%.
- We have implemented 4 algorithms which are result into better accuracy for each of the following algorithms: SVM 78.94%, Random Forest 75.65%, KNN 75% and Logistic Regression 77.63%.

The CVD Model

1. From (V. V. Ramalingam et al 2018)

- This research paper highlights the importance of feature selection and feature engineering but it uses a dataset (the Cleveland dataset) which consists of less than 400 records.
- The dataset which we have used uses 70000 records. This improves the efficiency of our model.

2. From (Adil Hussain Seh et al 2019)

- In this research paper the authors have removed the outliers using known values

of those attributes and highlighted the difference in the results. They achieved higher accuracies with the help of feature selection but their model seems to be over-fitted.

- We have added an additional attribute (BMI) which helped us in achieving a greater accuracy and it proved to be an important attribute even after feature selection.

3. From (G.Subbalakshmi et al 2011)

- They developed a Decision Support in Heart Disease Prediction System (DSHDPS). It is implemented as web-based questionnaire application.
- We are generating a PDF of the report and directly mailing it to the patients.

CHAPTER 3

METHODOLOGY

Machine Learning is a process to feed machine enough data to train and predict a possible outcome using the algorithms at bay. There are three types of machine learning algorithms supervised learning algorithms, unsupervised learning algorithms and reinforcement algorithms. The algorithms like K- nearest neighbor, Support vector machine, logistic regression, random forest classifier and decision tree are used which are type of supervised learning algorithm.

Machine learning algorithms used:

1. K-Nearest Neighbors (KNN): KNN is a non-parametric machine learning algorithm. The KNN algorithm is a supervised learning method. This means that all the data is labelled and the algorithm learns to predict the output from the input data. It performs well even if the training data is large and contains noisy values. The data is divided into training and test sets. The train set is used for model building and training. A k- value is decided which is often the square root of the number of observations. Now the test data is predicted on the model built. There are different distance measures. For continuous variables, Euclidean distance, Manhattan distance and Minkowski distance measures can be used.
2. Support Vector Machines (SVM): Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.
3. Random Forest Classifier: Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean

prediction (regression) of the individual trees. Random decision forests correct for decision trees habit of overfitting to their training set.

4. **Logistic Regression:** In statistics, the logistic model is a widely used statistical model that, in its basic form, uses a logistic function to model a binary dependent variable; many more complex extensions exist. In regression analysis, logistic regression is estimating the parameters of a logistic model; it is a form of binomial regression. Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail, win/lose, alive/dead or healthy/sick; these are represented by an indicator variable, where the two values are labelled "0" and "1".

Library Functions used for hyperparameter tuning:

1. **GridSearchCV:** In GridSearchCV approach, machine learning model is evaluated for a range of hyperparameter values. This approach is called GridSearchCV, because it searches for best set of hyperparameters from a grid of hyperparameters values. For example, if we want to set two hyperparameters C and Alpha of Logistic Regression Classifier model, with different set of values. The grid search technique will construct many versions of the model with all possible combinations of hyperparameters, and will return the best one.
2. **RandomizedSearchCV:** RandomizedSearchCV solves the drawbacks of GridSearchCV, as it goes through only a fixed number of hyperparameter settings. It moves within the grid in random fashion to find the best set hyperparameters. This approach reduces unnecessary computation.

Cross Validation Classification Metrics:

1. **Confusion Matrix-** It gives us a matrix as output and describes the complete performance of the model where, TP: True Positive, FP: False Positive, FN: False Negative and TN: True Negative.

		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Confusion Table

2. Accuracy for the matrix can be calculated by taking average of the values lying across the main diagonal. It is given as:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

3. F1 score-It is used to measure a test's accuracy. F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is as well as how robust it is. F1 Score tries to find the balance between precision and recall. Mathematically, it is given as:

$$F_1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

4. Precision: It is the number of correct positive results divided by the number of positive results predicted by the classifier. It is expressed as-

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}}$$

5. Recall: It is the number of correct positive results divided by the number of all relevant samples. In mathematical form it is given as-

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

CHAPTER 4

IMPLEMENTATION

The application comprises of four phases: frontend, development, integration and deployment. This chapter discusses each phase in detail. Also, the various UML and DFD diagrams are presented with their significance with respect to our model. Next, we display the working of the application through snapshots of the real-time application.

4.1.Frontend [Web application]

- 4.1.1. HTML (Hypertext Markup Language) defines the structure and contents of a webpage – where things go, how they are laid out, and what’s on the page. We first start by understanding the HTML document Structure. The second step is to be familiar with CSS (Cascading Style Sheets) which defines the styling/presentation of a web page and the elements on it. The next step is to link the HTML and CSS files together, pick a design and customize the application with them. The following step comprises of adding content, images, choosing the right font, choosing the right color schemes and etc. The next step is to create additional pages for the application.
- 4.1.2. Materialize CSS is a UI component library which is created with CSS, JavaScript and HTML. We use it as a design language which combines the classic principles of successful design along with innovation and technology. It is used to construct attractive, consistent, and functional web pages and web apps while adhering to modern web design principles such as browser portability, device independence, and graceful degradation.

4.2.Development [ML Training Models]

- 4.2.1. Dataset: Diabetes Dataset is taken from Kaggle which is contributed by UCL Machine Learning named as Pima Indians Diabetes Database, which consists of 768 female patients who are at least 21 years old. This is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset.

- 4.2.2. Data-Preprocessing: In this the Inconsistent data, that is all the missing values were replaced by Nan values and then according to the skewness of the Distribution curve it was replaced by mean or median. Data Visualization: Outliers were detected using Box-plot Analysis and those out of the range were then removed. Correlation Matrix and Heat-map were used to look for the feature with least correlation values with the Outcome variable & Analysis of Feature Importance using their coefficients showed how each variable affects Outcome. So, features having less correlation values with outcome and less feature importance were removed by keeping the accuracy of the model in mind.
- 4.2.3. Model training: Logistic Regression, KNN, Random Forest Classifier, Support Vector Machine. After training the diabetes dataset, the model with best accuracy was found to be Support Vector Machine, with an accuracy of 78.94%. After training the CVD dataset, the model with best accuracy was found to be K nearest neighbor classifier, with an accuracy of 73.25%.
- 4.2.4. The model with best accuracy was then tuned by hyper-parameters using GridSearchcv but the accuracy was found to be decreasing even for different types of iterations, which means the model was over-fitting. So, we applied Hyper- parameter tuning to other algorithms as well and the model with best accuracy was found to be Logistic Regression using RandomizedSearchcv with an increase in accuracy.
- 4.2.5. For the diabetes dataset, ROC curve was plotted having an area of 0.84 under the curve. Confusion Matrix was constructed with: True Positives 96, True Negatives 24, False Positives 9, False Negatives 23, Cross Validated Accuracy mean 76.1%, Cross Validated Precision 70.92%, Cross-validated Recall 53.46% and Cross- validated F1-score 60.64%.

4.3.Integration and Deployment

- 4.3.1. Generation of pickle file from the ML model and integrating it with the frontend using Flask: In machine learning, while working with scikit learn library, we need to save the trained models in a file and restore them in order to reuse it to compare the model with other models, to test the model on a new data. The saving of data is called

Serialization, while restoring the data is called Deserialization. We use Pickle string module that implements a fundamental, but powerful algorithm for serializing and de-serializing a Python object structure. `pickle.dump` is used to serialize an object hierarchy, `pickle.load` to deserialize a data stream.

- 4.3.2. Deployment of web application on Heroku: The first step is to create a Simple structure for our project with some basic files. The next step is to choose a version control system and to place our code in a development platform in a repository. The most popular version control system is Github. Then upload all the application files to Github. The following step comprises of linking the Github Repository with the Heroku Platform. After Linking the repository, we must configure and import all the necessary libraries, after which we get a link to access our application.
- 4.3.3. Report generation using Google Forms and formation of database. The lab technician will be required to fill all the details of each patient in a form. This is linked with AutoCrat and will generate a PDF report of the patient. The report will be mailed to the patient. Along with this, the complete data of all the users will be accumulated at one place simultaneously.

4.4. UML and DFD diagrams.

4.4.1. Use Case Diagram: Use case diagrams are used to gather the requirements of a system including internal and external influences. These requirements are mostly design requirements. Hence, when a system is analyzed to gather its functionalities, use cases are prepared and actors are identified. Use case diagram depicts all the actors involved in this project and their functionalities throughout the project.

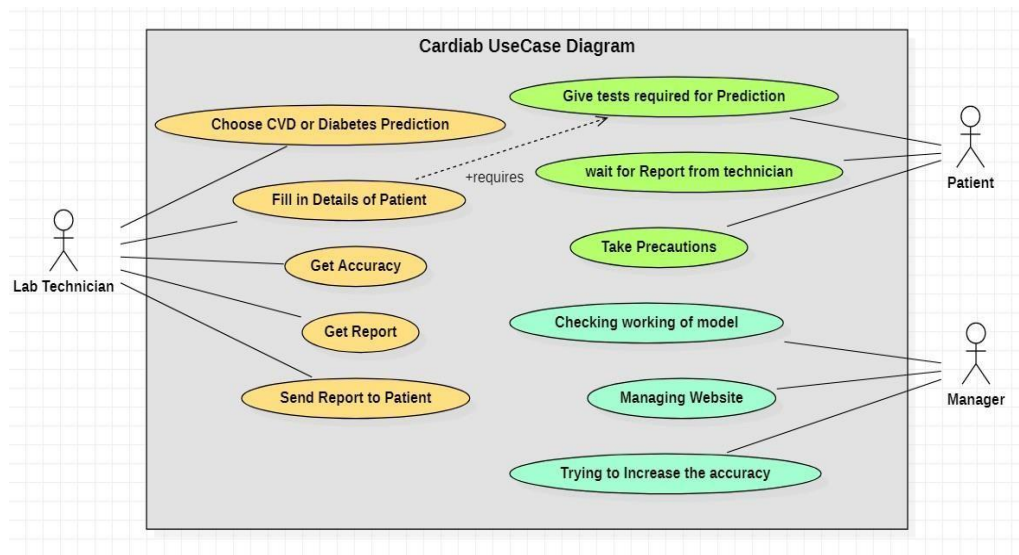


Figure 1: Use Case Diagram

4.4.2. Activity Diagram: An activity diagram is a behavioral diagram. It portrays the control flow from a start point to a finish point showing the various decision paths that exist while the activity is being executed. The given activity diagram depicts the front end of our application i.e., what user will be seeing and what all options the user will be provided with, user will have to choose from those options and fill in the values to the attributes to get to the result.

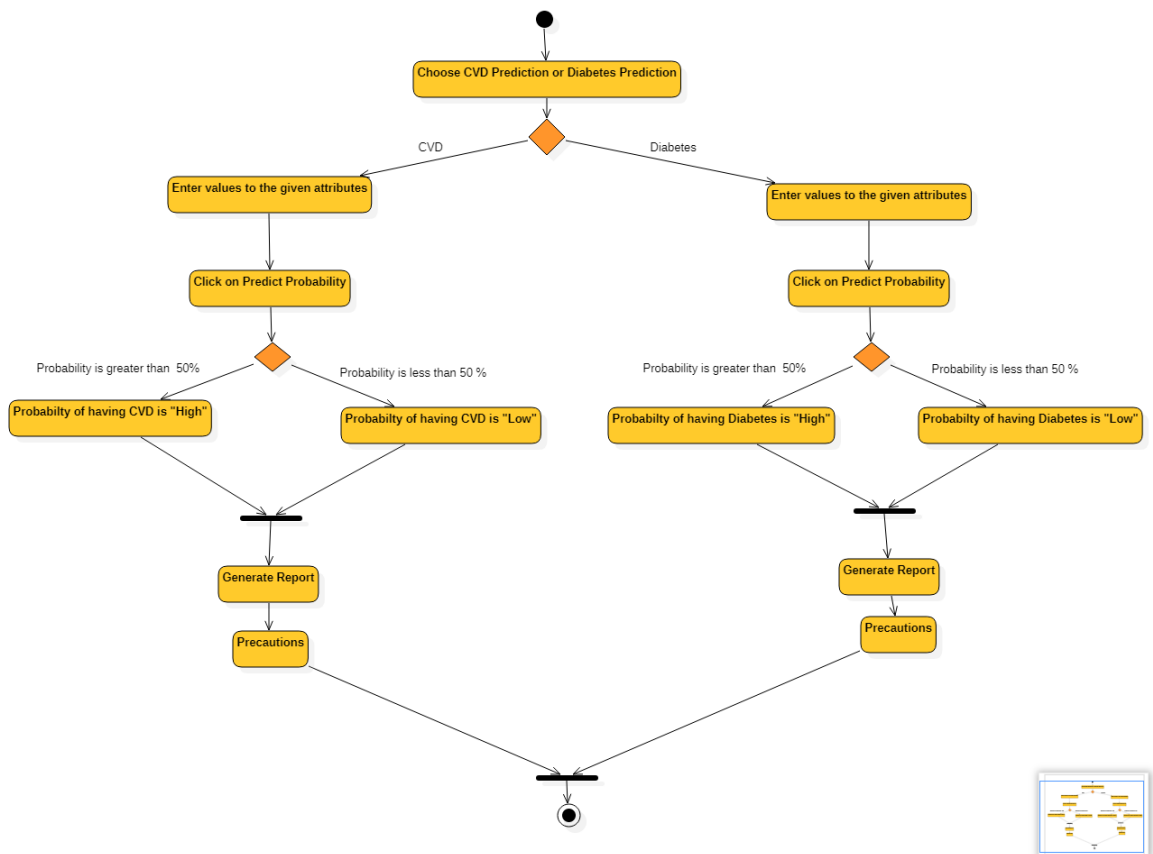


Figure 2: Activity Diagram

4.4.3. State-chart Diagram: It describes different states of a component in a system. The states are specific to a component/object of a system. Our state-chart diagram explains the frontend view/phase of the system i.e., what user will be seeing and what all options the user will be provided with, user will have to choose from two predict probability buttons and then fill in the values of the attributes to get to the desired result. Then the user can generate the report of the result displayed which will be sent to his entered email id. Also, users can take necessary precautions regarding their health conditions by clicking on the Tips and precautions button.

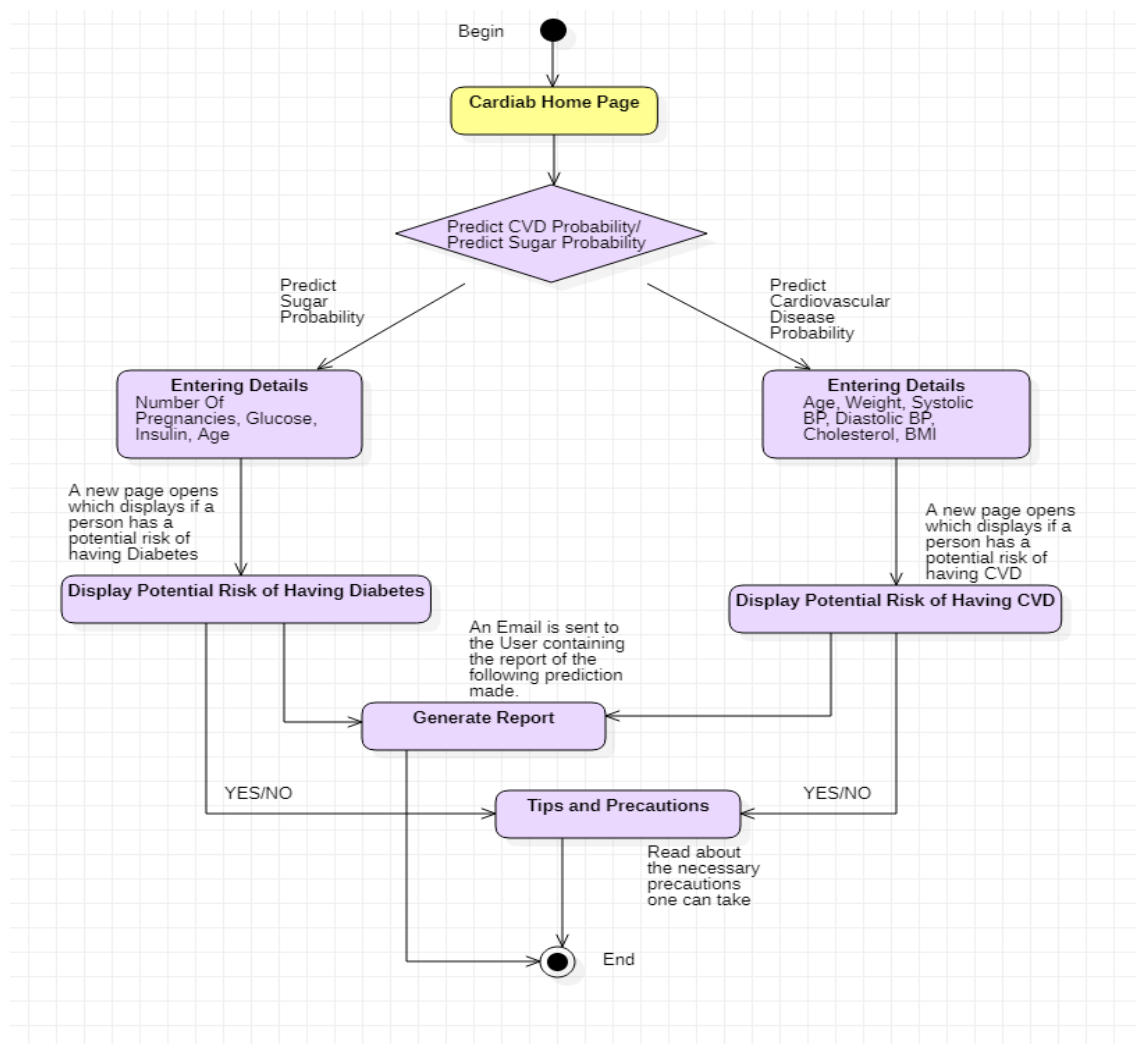


Figure 3: State-chart Diagram: Frontend

4.4.4. Data Flow Diagram: DFD graphically represents the functions, or processes, which capture, manipulate, store, and distribute data between a system and its environment and between components of a system. The visual representation makes it a good communication tool between User and System designer. Next is the dataflow diagram for the complete project which shows that data is used from the dataset and undergoes Data Preprocessing, Model training, testing and tuning. The user data which is new data is passed through the model to predict an outcome. This data is then used to generate a report which is then sent via email.

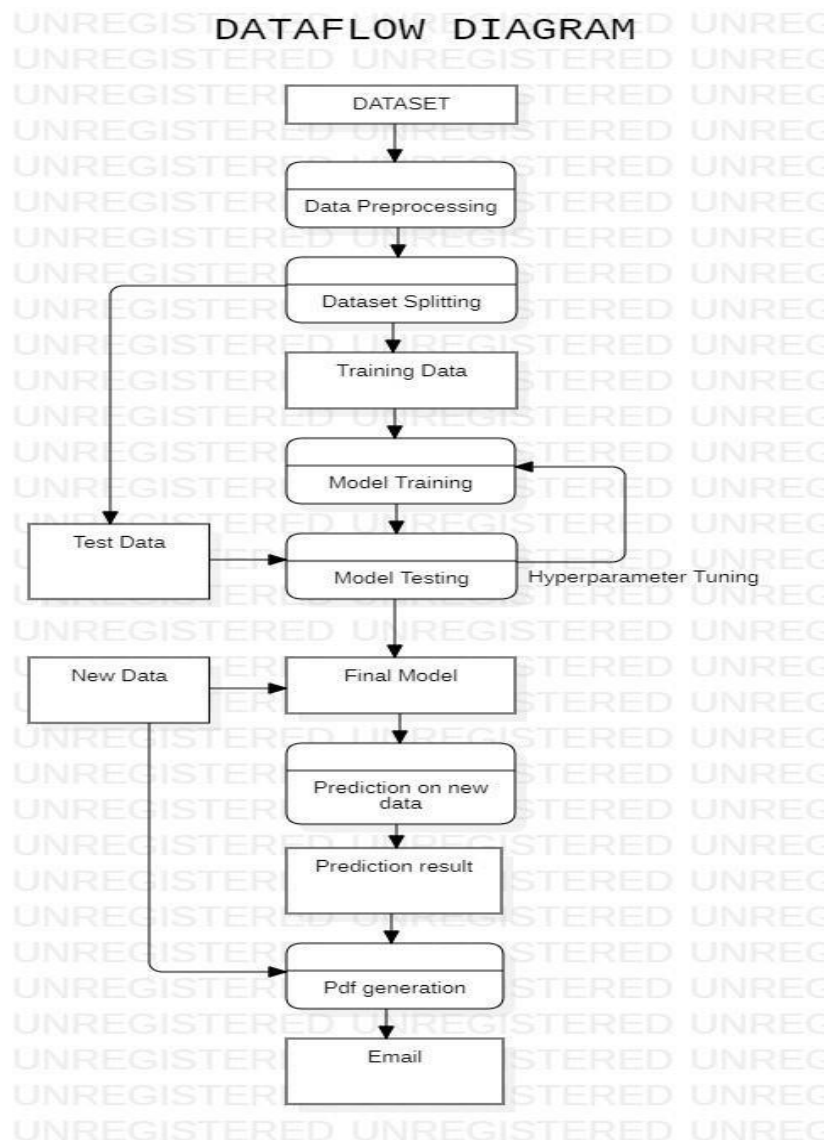


Figure 4: Data Flow Diagram

4.4.5. Sequence Diagram: A sequence diagram simply depicts interaction between objects in a sequential order i.e., the order in which these interactions take place. Sequence diagrams describe how and in what order the objects in a system function. These diagrams are widely used by businessmen and software developers to document and understand requirements for new and existing systems. Sequence Diagram depicts the steps involved associated with actors and the sequence in which they are executed.

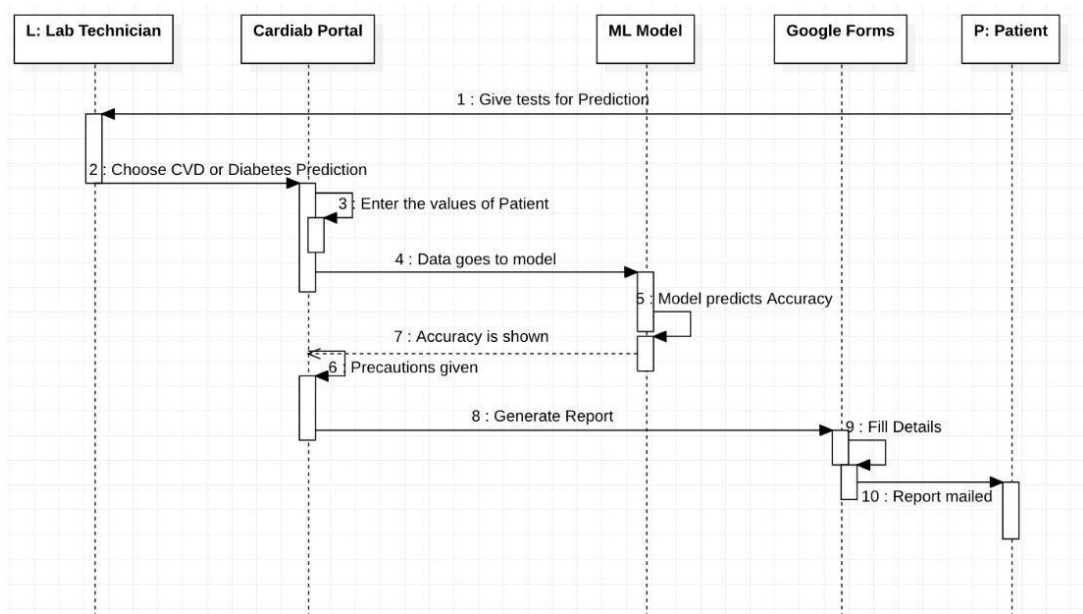


Figure 5: Sequence Diagram

4.5. Working

This sub-section demonstrates the working of the CARDIAB web application. First, the user is presented with the home page. There will be two options, one for CVD and one for Diabetes prediction. On selecting one, the user will be required to enter desired health attributes. On submission, the chance of the person having a CVD or Diabetes will be displayed. On clicking the 'Generate Report' button, user will be directed to the Report Generation Form where on entering the desired values, a PDF copy of the report will be sent to the entered email id.

4.5.1. This is the Home Page of CARDIAB. It has two buttons, one for CVD prediction and one for diabetes.

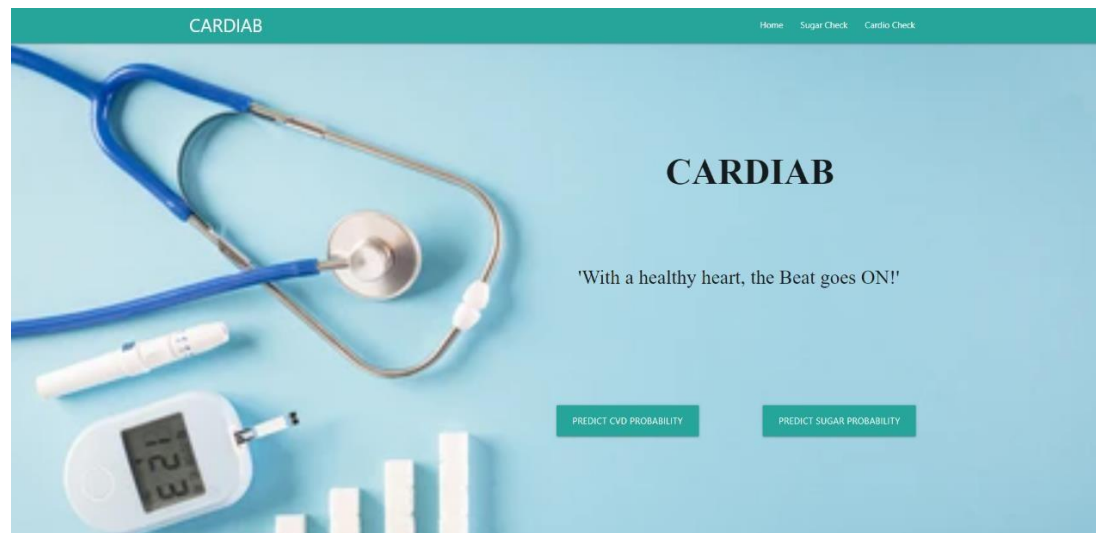


Figure 6: Home Page

4.5.2. This is the CVD prediction page. User will enter the desired health attributes here.

Figure 7: CVD Page


4.5.3. This is the result page for CVD prediction.

Cardio Vascular Disease Prediction


Home Sugar Check Cardio Check

Probability	Percentage
Yes	96.0 %


GENERATE REPORT




Know your risk.



Eat a healthy diet.




Be physically active.



Watch your weight.

Figure 8: CVD Result Page

4.5.4. This is the report generation page for CVD. The user will fill this form to generate the report of the patient.



Cardiovascular Disease Report Form

After you fill this form, the report of the patient will be sent to the patient.

*Required

Email address *

mindspeak2k17@gmail.com

Name of Lab *

JJ Thompson

Name of Lab Technician *

JJ Thompson

Name of Patient *

Rahul

Figure 9: CVD Report Generation Page

4.5.5. This is the report which the patient receives on the entered mail id.

CVD Report



Name of Lab: JJ Thompson

Name of Lab Technician: JJ Thompson

Name of Patient: Rahul

Age: 60 years

Gender: Male

Patient's Health Attributes

Weight (kg)	95
Systolic BP (mmHg)	140
Diastolic BP (mm Hg)	90
Cholesterol	3
BMI	30

General Values

Total Cholesterol		
Desirable [1]	Borderline High [2]	High [3]
Less than 200	200-239	240 and higher
Blood Pressure		
Category	Systolic BP (mm Hg)	Diastolic BP (mm Hg)
Normal	Less than 120	Less than 80
Elevated	120-129	Less than 80
High	130 or higher	80 or higher

Chance of the patient having a CVD: 86 %

Figure 10: CVD Report

4.5.6. This is the Diabetes prediction page. User will enter the desired health attributes here.

Diabetes Prediction

Home

Sugar Check

Cardio Check

Diabetes Prediction

Predict the probability of having Diabetes

Pregnancies

5

Glucose (mg/dl)

139

Insulin (mcU/ml)

140

DiabetesPedigreeFunction

0.411

Age (years)

26

PREDICT PROBABILITY

© Group 6 RCOEM

Figure 11: Diabetes Page

4.5.7. This is the result page for Diabetes prediction.

Diabetes prediction


Home

Sugar Check


Cardio Check

Probability	Percentage
No	47.0 %


GENERATE REPORT




Expand your palate's palette.



Less sugar, more water.




Move more, sit less.



Get enough rest.

Figure 12: Diabetes Result Page

4.5.8. This is the report generation page for Diabetes. The user will fill this form to generate the report of the patient.



The form is titled "Diabetes Report Form" and includes a decorative floral header. Below the title, a note states: "After you fill this form, the report of the patient will be sent to the patient." A red asterisk indicates required fields. The form contains four input fields: "Email address *" with the value "mindspeak2k17@gmail.com", "Name of Lab *" with the value "JJ Thompson", "Name of Lab Technician *" with the value "JJ Thompson", and "Name of Patient *" with the value "Sakshi".

Diabetes Report Form

After you fill this form, the report of the patient will be sent to the patient.

**Required*

Email address *

mindspeak2k17@gmail.com

Name of Lab *

JJ Thompson

Name of Lab Technician *

JJ Thompson

Name of Patient *

Sakshi

Figure 13: Diabetes Report Generation Page

4.5.9. This is the report which the patient receives on the entered email id.

Diabetes Report



Name of Lab: JJ Thompson

Name of Lab Technician: JJ Thompson

Name of Patient: Sakshi

Age: 26 years

Gender: Female

Patient's Health Attributes

No. of pregnancies	5
Blood Glucose (mg/DL)	139
Serum Insulin (mIU/L)	140
DiabetesPedigreeFunction	0.411

General Values

2-hour Blood Glucose (mg/DL)		
Normal	Impaired Glucose	Diabetic
120-140	140-160	200+
2-hour Serum Insulin (mIU/L)		
Normal	16-166	

Chance of the patient having diabetes: 47 %

Figure 14: Diabetes Report

OUTPUT: Machine learning algorithms are categorized as being supervised or unsupervised. A supervised learning algorithm uses the past experience to make predictions on new or unseen data. This study uses classification technique to produce a more accurate predictive model. Initially since the data is inconsistent, we are performing data cleaning and preprocessing in two steps:

1. Identifying the missing values and 0's, and replacing it with mean or median (this depends on the skewness of the distribution graph).
2. Removing Outliers: These refer to the infeasible values of the attributes.

The next step is data visualization and feature selection with the help of pair plot, histogram, etc. After visualizing data with the help of correlation matrix, the attributes which are least significant were removed (feature selection). Then we split the data into training data and testing data. On training data, we apply different machine learning algorithms to find out the best model suited for this dataset by comparing the accuracy of each model. The model is then tested on testing data, and hyperparameter tuning is done to improve efficiency of the model.

RESULTS

Supervised machine learning algorithms namely, K nearest neighbor, support vector machines, logistic regression and random forest classifier are trained on two different datasets. Efficiency of the result is successfully improved by hyperparameter tuning. The accuracies of both the models are given below.

Table 1: Accuracies of the two models on different algorithms (before hyperparameter tuning):

Algorithm	CVD Model	Diabetes Model
KNN	73.16%	75%
SVM	73.15%	78.94%
Logistic Regression	72.59%	77.63%
Random Forest	69.47%	75.65%

To improve the results hyperparameter tuning was done on the models with the highest accuracies for both the datasets.

- The accuracy of SVM for Diabetes decreased to 75% on hyperparameter tuning. Hence, for diabetes the logistic regression model was tuned and the final accuracy came out to be 78.94%.
- For the CVD dataset, the KNN model was successfully tuned to an accuracy of 73.25%.

Table 2: Accuracies of the two models on best algorithms (after hyperparameter tuning):

Models	Algorithm	Accuracy
CVD Model	KNN	73.25%
Diabetes Model	Logistic Regression	78.94%

After rigorous training of the models, logistic regression for the diabetes dataset and KNN for the cardiovascular dataset are used to predict the probabilities of a person having the disease. The model is deployed on Heroku and the report generation is also done successfully.

CONCLUSION

The Covid-19 pandemic has made us all realize the importance of our health and life more than ever. The fact that people who have pre-existing medical conditions like cardiovascular diseases and diabetes have a greater risk of having severe effects is very worrisome considering a huge section of population already having these diseases. Our application is aimed at predicting the probability of a person having a CVD or diabetes and will help the patients to take the necessary precautions irrespective of the result. It uses two different datasets for the prediction of CVD and Diabetes respectively. Out of Logistic Regression, SVM Classifier, Random Forest Classifier and K-Nearest Neighbor Classifier, we got the highest accuracy using KNN Classifier for the CVD dataset and Logistic Regression for the Diabetes Dataset. The model is deployed on Heroku. The result was then used to generate a lab report, which can later be referred by the patients and doctors. The future scope for the project includes increasing the accuracy of the predictions to make the application more reliable and accurate.

REFERENCES

- [1] Barale, Mahesh & Shirke, Digambar. (2016). Cascaded Modeling for PIMA Indian Diabetes Data. *International Journal of Computer Applications*. 139. 1-4. 10.5120/ijca2016909426.
- [2] Naveen Kishore G, V.Rajesh, A.Vamsi Akki Reddy, K.Sumedh, T.Rajesh Sai Reddy. (2020) Prediction Of Diabetes Using Machine Learning Classification Algorithms. *ISSN 2277-8616*
- [3] Subbalakshmi, G. & Ramesh, K. & Rao, M.. (2011). Decision Support in Heart Disease Prediction System using Naive Bayes. *Ind. J. Comput. Sci. Eng. (IJCSE)*. 2. 170-176.
- [4] Seh, Adil Hussain. (2019). A Review on Heart Disease Prediction Using Machine Learning Techniques.
- [5] Ramalingam, V V & Dandapath, Ayantan & Raja, M. (2018). Heart disease prediction using machine learning techniques: A survey. *International Journal of Engineering & Technology*. 7. 684. 10.14419/ijet.v7i2.8.10557.
- [6] Yu, W., Liu, T., Valdez, R., Gwinn, M., Khoury, M.J., 2010. Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes. *BMC Medical Informatics and Decision Making* 10. doi:10.1186/1472-6947-10-16 | arXiv:arXiv:1011.1669v3
- [7] Sellappan Palaniappan, Rafiah Awang, Intelligent Heart Disease Prediction System Using Data Mining Techniques, 978-1-4244-1968- 5/08/\$25.00 ©2008 IEEE.
- [8] Andreas G. K. Janecek ,WilfriedN.Gansterer and Michael A.Demel,||On the Relationship Between Feature Selection and Classification Accuracy||,JMLR: Workshop and Conference Proceedings 4: 90-105
- [9] Marvin L. Brown and John F. Kros, Data Mining and the Im-pact of Missing Data, *Industrial Management & Data Systems*, Volume 103, ISS: 611–621, (2003).
- [10] Bamnote, M.P., G.R., 2014. Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. *Advances in Intelligent Systems and Computing* 1, 763–770 | doi:10.1007/978-3-319-11933-5.

APPENDIX

The desirable values of health attributes used in both the datasets are give below.

1. BMI

Nutritional status	BMI (kg/m ²)
Underweight	<18.5
Normal range	18.5-22.9
Overweight	23-24.9
Obese I	25-29.9
Obese II	>30

Figure 15: Normal BMI Range

2. Cholesterol

National Cholesterol Education Program Cholesterol Guidelines			
	Desirable	Borderline High	High
Total Cholesterol	Less than 200	200 - 239	240 and higher
LDL Cholesterol (the "bad" cholesterol)	Less than 130	130 - 159	160 and higher
HDL Cholesterol (the "good" cholesterol)	50 and higher	40 - 49	Less than 40
Triglycerides	Less than 200	200 - 399	400 and higher

Figure 16: Normal Cholesterol Range

3. Blood Pressure

Blood Pressure Categories



BLOOD PRESSURE CATEGORY	SYSTOLIC mm Hg (upper number)		DIASTOLIC mm Hg (lower number)
NORMAL	LESS THAN 120	and	LESS THAN 80
ELEVATED	120-129	and	LESS THAN 80
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 1	130-139	or	80-89
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 2	140 OR HIGHER	or	90 OR HIGHER
HYPERTENSIVE CRISIS (consult your doctor immediately)	HIGHER THAN 180	and/or	HIGHER THAN 120

heart.org/bplevels

Figure 17: Normal Blood Pressure Range

4. Glucose

Plasma glucose test	Normal	Prediabetes	Diabetes
Random	Below 11.1 mmol/l Below 200 mg/dl	N/A	11.1 mmol/l or more 200 mg/dl or more
Fasting	Below 5.5 mmol/l Below 100 mg/dl	5.5 to 6.9 mmol/l 100 to 125 mg/dl	7.0 mmol/l or more 126 mg/dl or more
2 hour post-prandial	Below 7.8 mmol/l Below 140 mg/dl	7.8 to 11.0 mmol/l 140 to 199 mg/dl	11.1 mmol/l or more 200 mg/dl or more

Figure 18: Normal Glucose Range