

**MACHINE LEARNING BASED APPROACH FOR PREDICTING DIABETES IN
YOUNG PEOPLE OF BANGLADESH**

BY

**PRODIPTO ROY DIPTO
ID: 172-15-9689**

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Zerin Nasrin Tumpa
Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Mr. Masud Rabbani
Lecturer
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

MAY 2021

APPROVAL

This Project titled “**Machine Learning Based Approach for Predicting Diabetes in Young People of Bangladesh**”, submitted by **Prodipto Roy Dipto** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **31st May, 2021**.

BOARD OF EXAMINERS



Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman



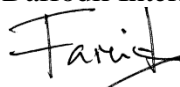
Gazi Zahirul Islam
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Raja Tariqul Hasan Tusher
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dr. Dewan Md. Farid
Associate Professor
Department of Computer Science and Engineering
United International University

External Examiner

DECLARATION

I hereby declare that, this project has been done by me under the supervision of **Zerin Nasrin Tumpa, Lecturer, Department of CSE, Daffodil International University**. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



Zerin Nasrin Tumpa
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:



Mr. Masud Rabbani
Lecturer
Department of CSE
Daffodil International University

Submitted by:



Prodipto Roy Dipto
ID: 172-15-9689
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes me possible to complete the final year project successfully.

I am really grateful and wish my profound indebtedness to **Zerin Nasrin Tumpa, Lecturer**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of “*Machine Learning*” to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to **Dr. Touhid Bhuiyan, Professor and Head**, Department of CSE, for his kind help to finish my project and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank my entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patience of my parents.

ABSTRACT

Diabetes is one of the deadly chronic diseases that occurs when the sugar level in the blood increases abnormally due to the absence of insulin hormone. Untreated Diabetes could lead a human to his/her death. In Bangladesh, the threat of Diabetes is really a matter of concern and people of all ages and gender are suffering equally. My research focused on young people who are under the age of 36 in Bangladesh. By using Machine Learning I have built a model which can predict the possibility of having or not having Diabetes. The model that I have built was trained by previous data of diabetic and non-diabetic patients. These data were collected from Bangladesh. On experiment, these data were processed and analyzed by various data pre-processing techniques. Then some classic Machine Learning algorithms like Logistic Regression, Random Forest, K-Nearest Neighbors and Naive Bayes were used for building the model and the performance of each of them was measured using metrics like Prediction Accuracy on the testing and training data, Confusion Matrix, Sensitivity, Precision, F1 score, Recall, Specificity, ROC and AUC. Overall Random Forest performed better than others. So, the Random Forest model was chosen for the prognosis of the disease and to demonstrate the use of the model. For that, I have built a web and an android application that will take required input data from a user according to the model and predict whether the user has Diabetes or not.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of Examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
CHAPTER	
CHAPTER 1: INTRODUCTION	1-3
1.1 Introduction	1
1.2 Motivation	1-2
1.3 Objective	2
1.4 Research Questions	3
1.5 Expected Outcome	3
CHAPTER 2: BACKGROUND	4-9
2.1 Preliminaries/Terminologies	4-5
2.2 Related Works	6-8
2.3 Comparative Analysis and Summary	8
2.4 Scope of the Problem	8-9
2.5 Challenges	9

CHAPTER 3: RESEARCH METHODOLOGY	10-17
3.1 Research Subject and Instrumentation	10
3.2 Dataset	10-11
3.3 Statistical Analysis	11-14
3.4 Proposed Methodology	15
3.5 Implementation Requirements	16-17
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	18-38
4.1 Experimental Result and Analysis	18-30
4.2 Discussion	30-32
4.3 Implementation and Deployment	32-38
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	39
5.1 Impact on Society	39
5.2 Ethical Aspects	39
5.3 Sustainability Plan	39

CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH 40-41

6.1 Summary of the Study	40
6.2 Conclusions	40
6.3 Implication for Further Study	40-41

REFERENCES 42-43

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.3.1: Target Class Distribution	12
Figure 3.3.2: Gender-Target Class Distribution	12
Figure 3.3.3: Correlation Matrix	14
Figure 3.4.1: Proposed Methodology	15
Figure 4.1.1: Confusion Matrix	19
Figure 4.1.2: ROC and AUC	22
Figure 4.1.3: Sigmoid Function Graph	23
Figure 4.1.4: Confusion Matrix of Logistic Regression Model	24
Figure 4.1.5: Working principle of Random Forest	25
Figure 4.1.6: Confusion Matrix of Random Forest Model	26
Figure 4.1.7: K-Nearest Neighbors	27
Figure 4.1.8: Confusion Matrix of KNN model	28
Figure 4.1.9: Confusion Matrix of Naive Bayes model	30
Figure 4.2.1: ROC curve of the algorithms	31
Figure 4.3.1: Feature Selection	32
Figure 4.3.2: Implementation and Deployment Process	33
Figure 4.3.3: Dashboard Feature (Web)	34
Figure 4.3.4: Dashboard Feature (Android)	34
Figure 4.3.5: Prediction Feature (Web)	35
Figure 4.3.6: Prediction Feature (Android)	35
Figure 4.3.7: Predicted Result (Web)	36
Figure 4.3.8: Predicted Result (Android)	36
Figure 4.3.9: Nearby Hospitals Feature (Web)	37
Figure 4.3.10: Nearby Hospitals Feature (Android)	37
Figure 4.3.11: Find Doctors Feature (Web)	38
Figure 4.3.12: Find Doctors Feature (Android)	38

LIST OF TABLES

TABLES	PAGE NO
Table 1: Dataset Attributes and Value Type	11
Table 2: Correlation	13
Table 3: Performance of Logistic Regression model	23
Table 4: Performance of Random Forest model	25
Table 5: Performance of KNN model	28
Table 6: Performance of Naive Bayes model	30
Table 7: Performance of the algorithms	31

CHAPTER 1

Introduction

1.1 Introduction

Diabetes occurs due to the lack of enough insulin production by the pancreas. Because of this, the sugar level in the blood increases abnormally. Since the discovery of insulin in 1922 by Dr. Frederick Banting, the treatment of Diabetes had been improved rapidly over time. But even today it's not always possible to determine the presence of diabetes in the early stage. Untreated Diabetes could lead to serious health complications and other diseases like coronary heart disease and stroke [1]. Diabetes also increases the risk of cognitive impairment and dementia [2]. So, this is certainly not a disease that we can ignore. This is where Machine Learning comes useful. We are living in a time which is known as the age of data revolution. But data is only fruitful when we process and put them into work. Machine learning does this exact same thing. Machine learning or ML is a sub-field of Artificial Intelligence or AI that performs intelligent computing tasks based on training and experience without any human intervention. So, if I can process the previous data of diabetic patients and apply Machine Learning to them then my ML model can predict whether a person will have Diabetes or not in the future which may save the life of that person and the lives of countless others as well. The purpose of this research is to build a Machine Learning model that can make a prediction about the possibility of Diabetes in young people of Bangladesh who are under the age of 36.

1.2 Motivation

According to WHO, there are 422 million diabetic patients worldwide. In Bangladesh, we have 8.3 million people suffering from Diabetes and half of them don't even know that they have Diabetes [3]. Due to excessive amount of carbohydrate food intake and frequent consumption of junk foods [4] a huge number of young people in Bangladesh are having Diabetes or on the verge of having Diabetes. What makes this matter worse is that since

Bangladesh is a developing country, the majority of our population lives under the poverty line. The expenditure for proper treatment of Diabetes in Bangladesh is increasing [5]. As I have discussed earlier that untreated Diabetes could lead to other diseases and health complications. So, detection is a crucial step to consider in terms of both health and economic prospects. This is why I thought to build a Machine Learning based model that can take data from a user and make a prediction whether the user has a possibility of having Diabetes or not. Diabetes at a young age is very much unfortunate although not unusual. The future of a country or this world for that matter relies on young people. If a substantial amount of these young people suffers from diseases like Diabetes, then the future is definitely uncertain. So doing something about Diabetes is definitely a logical thing to do and early detection of this disease is the first step I can take and Machine Learning is the cornerstone of this detection. And as a Machine Learning researcher, that's what motivated me to do this research.

1.3 Objective

The objectives of this research are as follows:

- i. Build a system that can predict the possibility of having Diabetes at an early stage under the age of 36 using Machine Learning.
- ii. Contribute and explore the existing knowledge in the field of Machine Learning and Diabetes.
- iii. Run a little survey on young people of Bangladesh who are under the age of 36 for understanding the severity of the situation.
- iv. Analyze the data using Data Science techniques for extracting new knowledge.
- v. Apply various Machine Learning algorithms and analyze their results.
- vi. Build a model for prognosis using the best performing algorithm.
- vii. Deploy the model using a web and a mobile application.

1.4 Research Questions

“How can we detect Diabetes at an early stage in young people of Bangladesh so that patients can take necessary precautions against serious health complications and lead a better life?”, that’s the sole question of this research and the answer I’ve got as a Computer Science student is by Machine Learning.

1.5 Expected Outcome

After this research work, I will be able to:

- i. Build a Machine Learning model for the purpose of predicting the possibility of having Diabetes or not.
- ii. Present new data on the overall situation of Diabetes under the age of 36 in Bangladesh.
- iii. Present some statistical analysis on Diabetic patients in Bangladesh.
- iv. Use the model for prognosis of Diabetes using the web and mobile application technologies.

CHAPTER 2

Background

2.1 Preliminaries/Terminologies

2.1.1 Diabetes

Diabetes Mellitus, also known as Diabetes, is a metabolic condition caused by a high level of glucose sugar in the blood. Normally in the human body, the insulin hormone moves sugar from the blood into the cells for producing energy. Because of Diabetes, either the pancreas becomes unable to produce an adequate amount of insulin or the cells of the body become unable to use the insulin effectively that the pancreas makes. Diabetes can be of 3 types:

- i. **Type 1 Diabetes:** This type of Diabetes occurs when pancreas stop producing enough insulin. This is caused due to the loss of beta cells.
- ii. **Type 2 Diabetes:** In this case, the cells of body do not respond to the insulin even though insulin is being produced.
- iii. **Gestational Diabetes:** Pregnant women develop this type of Diabetes and occurs because of high blood sugar levels during pregnancy.

2.1.2 Artificial Intelligence

The ability of a computer or machine to learn from examples and experience to imitate the capabilities of the human mind is known as Artificial Intelligence or AI. With the help of huge data that are stored online thanks to the internet and great computing power thanks to the brilliance of scientists and engineers, Artificial Intelligence is becoming the driving force of the future of human civilization. In recent years, from mastering the game of Go which was previously thought to be decades away [6] to unfold the structure of protein

which has been a grand challenge in biology for 50 years [7], Artificial Intelligence has demonstrated its amazing potential in the field of intelligent computing. That's why the landscape of biomedical research and healthcare is gradually shifting because of Artificial Intelligence as well [8]. There are lots of applications and domains in which Artificial Intelligence is used or has the potential to be used.

2.1.3 Machine Learning

Machine Learning is a branch of AI or Artificial Intelligence. It's the principle of computers learning from previous data, recognizing patterns, and making decisions with little to no help from humans. The combination of Computational Statistics and Data Science is what Machine Learning or ML derived from. In healthcare, the use of Machine Learning is very promising and effective [9]. There are various Machine Learning algorithms. These algorithms can be categorized into 3 types.

- i. **Supervised Learning:** In Supervised Learning the model gets trained on a labelled dataset or the type of dataset that has both input and output parameters. Some of the Supervised Learning algorithms are Linear Regression, Naive Bayes, Decision Tree etc.
- ii. **Unsupervised Learning:** In Unsupervised Learning data is provided without output parameter. The model itself has to find out the way of learning to categorize the data. Example: K-Means Clustering.
- iii. **Reinforcement Learning:** In Reinforcement Learning an agent learns in an interactive environment using input from its own behaviors and experiences. The model learns by rewards and punishments as indication for right and wrong actions.

2.2 Related Works

Following is the summary of some related research works that were relevant and useful for my research.

M. Al Helal *et al.* [10] have worked on PIMA Indians dataset and applied Random Forest, Decision Tree, K-Nearest Neighbor, Naive Bayes and Perceptron algorithms for predicting Diabetes using. They have done a good amount of data pre-processing. Among those mentioned algorithms, Random Forest gave the best accuracy of 73%. So, they have managed to achieve somewhat an average prediction accuracy. That's why I looked for other works that have achieved much better accuracy. S. K. Dey *et al.* [11] have applied K-Nearest Neighbor, Support Vector Machine, Gaussian Naive Bayes, and Artificial Neural Network on the PIMA dataset. Gaussian Naive Bayes gave the best accuracy of 76.25%. They have also built a web application using the model that they have created. Still the result is not very satisfactory. So, I kept looking for others. B. Pranto *et al.* [12] again created their Machine Learning model based on PIMA dataset but this time they applied this model to Bangladeshi patients. They have also collected data from Kurmitola General Hospital, Dhaka. They have applied Naive Bayes, Random Forest, Decision Tree and K-Nearest Neighbors on their data and Random Forest outperformed the others with 78% accuracy. Their result is not bad at all. But it is still not eligible for becoming a prognosis purpose model. A. Uddaula *et al.* [13] have worked on PIMA dataset which has 768 records and 9 features. In this research, the researchers have adopted a pretty unusual and good approach for predicting diabetes. Six meta classifier algorithms were used for their work. They were Ada Boost M1, Attribute Selected Classifier, Multiclass Classifier Updatable, Logit Boost, Bagging, and Filtered Classifier. After applying some other Data Mining techniques, Multiclass Classifier Updatable achieved the highest performance with a win-rate of 80%. They have measured some other performance metrics on which some other algorithms did a better job. This is a good result indeed. But since I'm doing my research on Bangladeshi people, PIMA dataset is not a suitable one for me. So, I looked for the research works that were conducted on Bangladeshi data. N. S. Khan *et al.* [14] made an mHealth application for predicting diabetes by machine learning. They collected

around 191 usable responses from a survey categorizing People's age, BMI (Body Mass Index), gender, HbA1c and ancestral diabetic history. They applied only one prediction algorithm Naive Bayes and got an accuracy of 64% in predicting diabetes. This is certainly a very poor accuracy. So, I kept looking for the research works that have achieved good results on Bangladeshi data. M. F. Faruque *et al.* [15] collected data of 200 patients from the diagnostic of Medical Centre Chittagong (MCC). They categorized patients' age and weight into 3 categories and applied K-Nearest Neighbors, Decision Tree, Support Vector Machine and Naive Bayes. Decision Tree achieved the best accuracy of 73.5%. This result is also not very satisfactory. K. C. Howlader *et al.* [16] have collected their data from Noakhali, Bangladesh. That dataset contained 220 patients' data. Four classifiers that are based on Decision Tree such as J48, CDT, REPTree and NBtree were applied to their data. They have also analyzed some models that were based on Decision Tree were applied with the strategies called pruned and unpruned. CDT (unpruned) shows the best accuracy which is 96.78%. That's a very good result for a prognosis purpose model but the number of data is not sufficient enough. N. Jahan *et al.* [17] predicted diabetes risk level by Machine Learning with factor scoring and also proposed an android application. They had collected 555 patient responses with 9 attributes which are age, gender, blood pressure (mm Hg), exercise, BMI (weight in Kg/(height in m)²), stress level, genetic, sleeping hrs, Smoking. After that, they applied 3 Machine Learning algorithms which are Decision tree, Multilayer perception and IBK. From cross-checking up to 12 fold the IBK algorithm gives about 98.73% prediction accuracy. This is an outstanding result. But I tried to push to the limit and looked for much more improved ones with more data. S. S. Rahman *et al.* [18] have worked on a diabetes detection system by Machine Learning-based approach. They collected data of 6219 individuals from different hospitals with 5 attributes and apply supervised KNN and Unsupervised K Means. By doing so they found out that the SeqNum attribute is an extra attribute that was not needed and gained accuracy of 99.78% by KNN algorithm and concluded that supervised KNN is better than unsupervised K Means. T. Le Minh *et al.* [19] have used "Early stage diabetes risk prediction dataset" which is available on the UCI Machine Learning repository. They have applied Support Vector Machine, Logistic Regression, Random Forest Classifier, Decision Tree, K-Nearest Neighbors, Naive Bayesian Classifier, GWO-MLP and APGWO-MLP. APGWO-MLP got the best

©Daffodil International University

accuracy of 97%. The last four research works achieved an outstanding prediction accuracy. But they were not conducted on young people of Bangladesh and some of them have some other limitations. So, this is the area and topic I picked up for this research.

2.3 Comparative Analysis and Summary

I intended to do my research on Bangladeshi people. Among the research works that I've studied, I've classified them into two categories: ones that were done on PIMA Indians Diabetes dataset which is present in the UCI Machine Learning Repository, and the other ones which were conducted on Bangladeshi people. PIMA Indians Diabetes dataset is a very reliable and famous dataset. A lot of research works were done on this dataset. But there are some limitations of this dataset in terms of my research. First of all, the records on this dataset were collected from female patients and there are some null values on some important features like 'Glucose Level' and 'BMI'. My intention is to do a research work that will be applicable to any person regardless of his/her gender. Since I wanted to do research on young people of Bangladesh the PIMA Indians dataset was not suitable for my work. That's why I was looking for research works that were done on Bangladeshi data and I found some. These research works helped me a lot in terms of Bangladeshi Diabetic patients. But none of them focused particularly on young people. This is where I think my research work becomes relevant.

2.4 Scope of the Problem

After reviewing all the above research works it is clearly understandable that Diabetes prediction and detection using Machine Learning techniques is one of the most sought-after research topics. We can also understand that the job isn't completely done yet, especially if we consider this fact in the context of Bangladesh. The prevalence of Diabetes in Bangladesh is increasing [20]. I believe that there is still a lot more room for research on this topic. So, I have taken a different approach. I wanted to find out how much young

people in our country are suffering from Diabetes and what I can do about it as a Machine Learning researcher.

2.5 Challenges

The biggest challenge that I faced was in terms of collecting data on Diabetic patients. I collected data from Bangabandhu Sheikh Mujib Medical University, Dhaka (BSMMU) for which I had to take permission from the Director of the Hospital. I had to apply through my university department in which a lot of time and energy was consumed. After spending a week in the hospital, I could only manage a handful amount of data. Also due to the COVID-19 pandemic, I couldn't collect data spontaneously and I had to do the survey by taking a lot of health risks.

CHAPTER 3

Research Methodology

3.1 Research Subject and Instrumentation

I'm doing this research based on the previous data of diabetic and non-diabetic patients. So, I collected these data and prepared a dataset. Then for statistically analyzing these data and creating the model I have used **Python** as the programming language and **Google Colab** and **Jupyter Notebook** as the programming environment. For the implementation and deployment of the model I have used **Python** and its framework **Flask**, **PHP**, **JavaScript**, and **Java** as the programming language and **Visual Studio Code** and **Android Studio** as the programming environment or IDE.

3.2 Dataset

I have collected the data from **Bangabandhu Sheikh Mujib Medical University, Dhaka (BSMMU)**. The total number of records in my dataset is 1036. There are 16 attributes or features, out of which 15 are independent features and 1 dependent feature named **“Class”**. In other words, the **“Class”** feature is the target variable that defines whether a person has Diabetes or not. The **“Age”** feature is the only feature that has numeric/discrete type values. The rest of them contain categorical values. The attributes and their value type are given in the following table.

TABLE 1: DATASET ATTRIBUTES AND VALUE TYPE

Attribute	Value Type
Age	Numeric/Discrete
Gender	Male(1)/Female(0)
Polyuria	Yes(1)/No(0)
Polyphagia	Yes(1)/No(0)
Polydipsia	Yes(1)/No(0)
Weakness	Yes(1)/No(0)
Sudden Weight Loss	Yes(1)/No(0)
Itching	Yes(1)/No(0)
Visual Blurring	Yes(1)/No(0)
Irritability	Yes(1)/No(0)
Delayed Healing	Yes(1)/No(0)
Muscle Stiffness	Yes(1)/No(0)
Partial Paresis	Yes(1)/No(0)
Obesity	Yes(1)/No(0)
Alopecia	Yes(1)/No(0)
Class	Yes(1)/No(0)

3.3 Statistical Analysis

3.3.1 Data Pre-processing

- i. Checking null values is the first step I've taken in the data pre-processing phase. Null values are one of the greatest obstacles in the way of creating a good model. There are many ways for removing this problem. One of them is to take the mean value of the feature that has null values. But the dataset that I have has no null values.
- ii. In the next step of data pre-processing, I had checked whether my dataset is imbalanced or not. A dataset with a huge margin of unequal target class distribution is called an imbalanced dataset. The model that we create from this type of dataset

becomes bias towards a particular class. But my dataset also passed this test as well. There are 636 individuals who have Diabetes and 400 who don't.

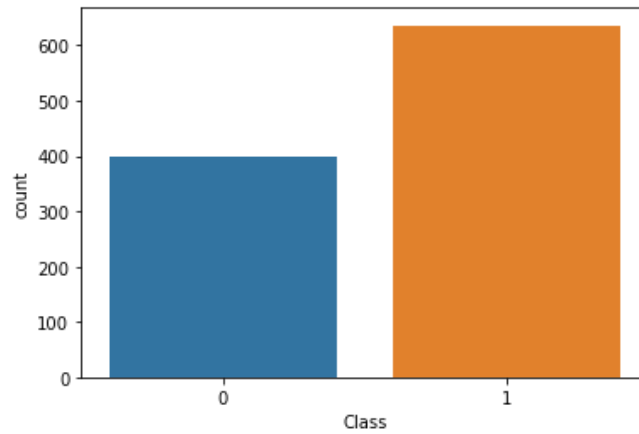


Figure 3.3.1: Target Class Distribution

iii. I have also checked the Gender-Target class distribution and it came out that women are more vulnerable to diabetes than men are.

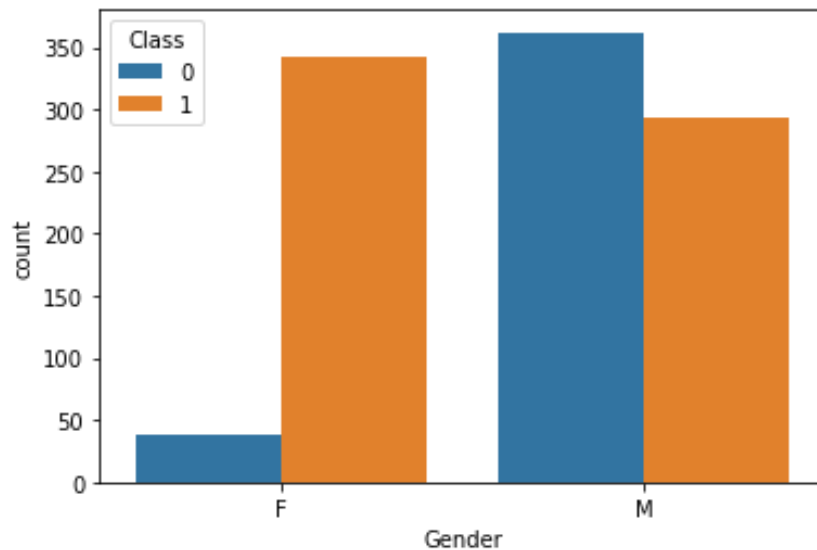


Figure 3.3.2: Gender-Target Class Distribution

- iv. Finding correlations between the target feature and independent features is very important in terms of data pre-processing and even for prediction accuracy in the long run. It's also very important for better understanding the features in the dataset. From the following Correlation table, it is obvious that **“Polyuria”** and **“Polydipsia”** are the most important features, which is understandable and makes sense in the case of Diabetes. That's why I looked for the correlation.

TABLE 2: CORRELATION

Attribute	Correlation value
Age	0.146441
Gender	-0.447259
Polyuria	0.671037
Polyphagia	0.346275
Polydipsia	0.647668
Weakness	0.247889
Sudden Weight Loss	0.434658
Itching	-0.016573
Visual Blurring	0.248805
Irritability	0.301744
Delayed Healing	0.049937
Muscle Stiffness	0.119061
Partial Paresis	0.436106
Obesity	0.073669
Alopecia	-0.273059

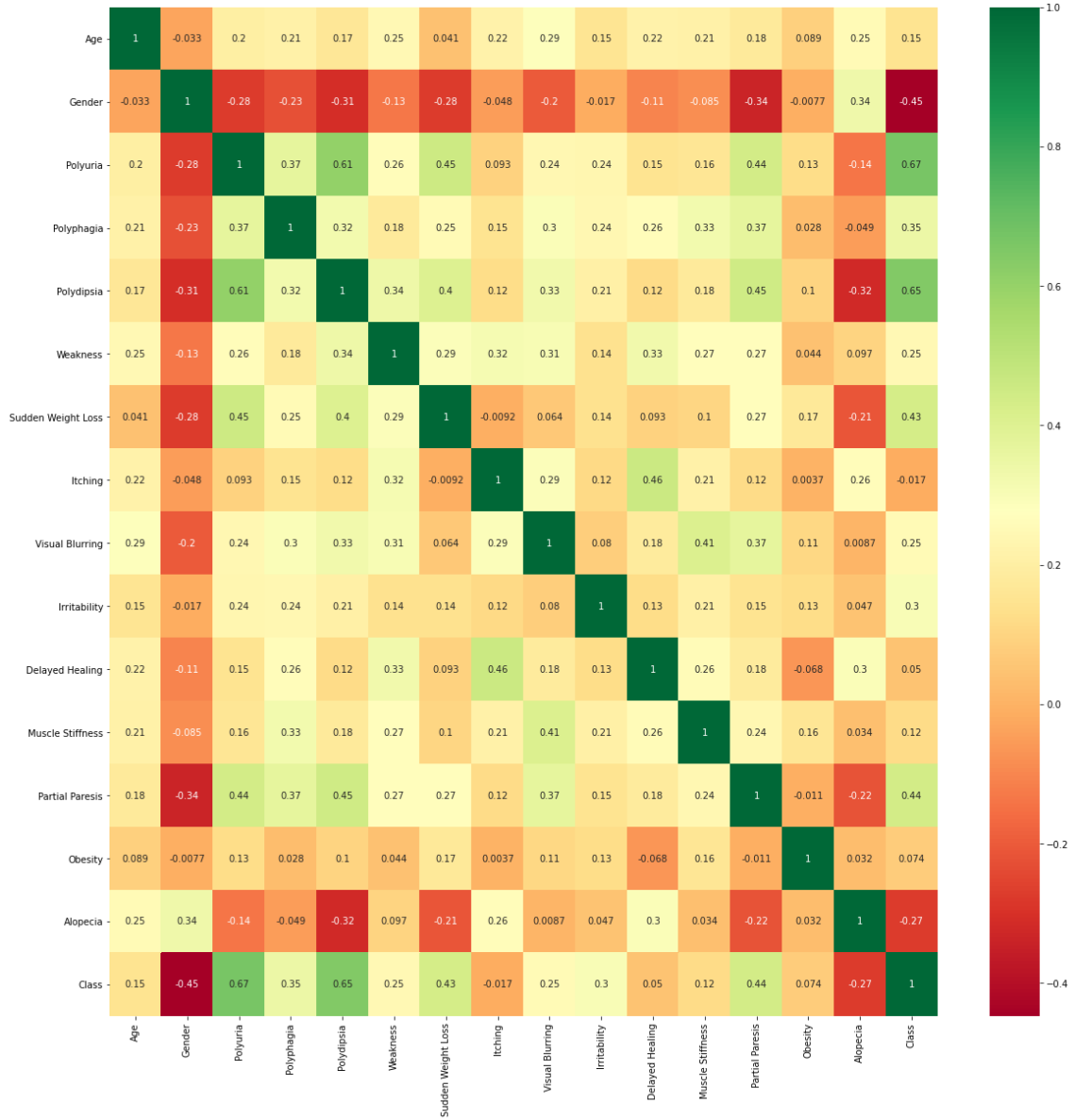


Figure 3.3.3: Correlation Matrix

3.4 Proposed Methodology

The methodology that I followed can be described in the following points.

- i. First, I collected the data and prepared the dataset.
- ii. Then I applied various data pre-processing techniques to make the dataset suitable for the algorithms.
- iii. After pre-processing, I applied some Machine Learning algorithms that I've mentioned later.
- iv. I measured the performance of each of the algorithms.
- v. I created the model using the best-performing classifier.
- vi. Finally, I deployed the model by a web and a mobile application.

The following illustration is the proposed methodology which I followed for building the model and deploying it.

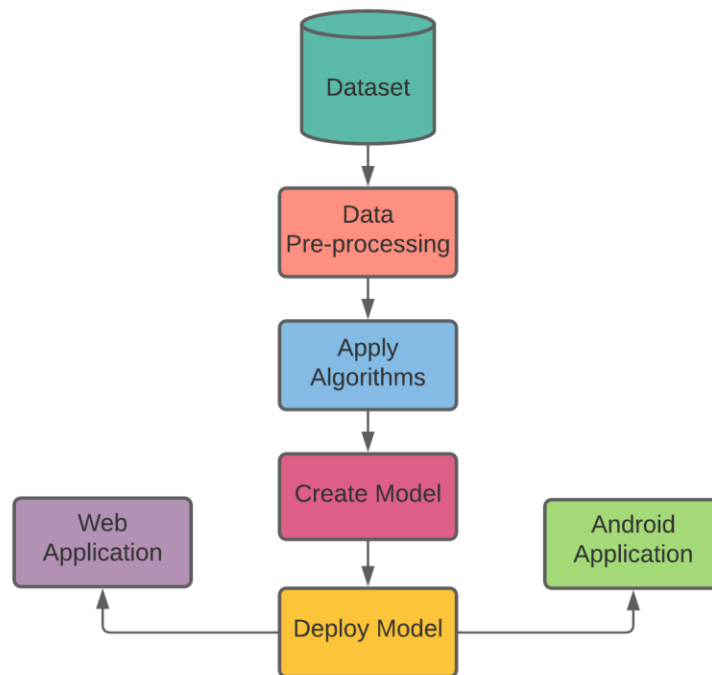


Figure 3.4.1: Proposed Methodology

3.5 Implementation Requirements

3.5.1 Python

Python is the type of programming language that is high-level, interpreted, object-oriented and has dynamic semantics. From building web applications to AI research, Python is used everywhere in today's world. For Machine Learning and AI, it is the primary choice for engineers and researchers around the world. I also used Python for this research project.

3.5.2 Jupyter Notebook

Jupyter Notebook is a type of web application designed for the tasks that are needed for Machine Learning and Data Science. Researchers all over the world directly or indirectly used it for machine learning, data analysis, data visualization, statistical analysis etc. Most of the experiments that I did for this research were carried out in Jupyter Notebook.

3.5.3 Google Colab

Google Colab or Colaboratory is a research product created by Google that allows us to write python code on the browser. It's a great tool for data analysis and visualization. It allows us to use the great computing resources from Google, such as GPUs and TPUs, for free. Since it does not require any setup or tedious process of installing Python packages, I also used it for my experiments whenever I was on the go.

3.5.4 Flask

Flask is a type of web framework that we know as a micro framework and the programming language that it is written on is Python. For deploying the Machine Learning model and expose it as an API, I have used Flask.

3.5.5 PHP

PHP is an object-oriented server scripting language that is mostly used in web development. For connecting my web application to the database and creating an API for the database, I have used PHP.

3.5.6 JavaScript

JavaScript is a programming language that has its use everywhere. Right now, it's the most popular programming language in the world. It's mostly used for web development. I have used JavaScript for building the frontend of my web application and connecting to the flask API where my Machine Learning model is deployed.

3.5.7 Java

Java is the type of programming language that is object-oriented and class-based and it is one of the most beloved languages of all time for building software. When Android OS was first introduced, Java was the only best option for Google to build applications for it. Even today, most android apps are written in Java. Almost the whole of my android application is built on Java.

3.5.8 Visual Studio Code

Visual Studio Code, also known as VS Code is a code editor that is hugely popular among developers because of its many developer-friendly features. It is developed by Microsoft. For writing the codes of my web application, I have used it most of the time.

3.5.9 Android Studio

Android Studio provides a development environment for Android OS. It is the official Integrated Development Environment or IDE recommended by Google. It is developed by Google and JetBrains. I have developed my whole android application on Android Studio.

CHAPTER 4

Experimental Results and Discussion

4.1 Experimental Results and Analysis

After all the data pre-processing and dividing the data into training and testing dataset I have applied four different Machine Learning algorithms named **Logistic Regression**, **Random Forest**, **K-Nearest Neighbors** and **Naive Bayes** and measured their performances on different metrics like **Prediction Accuracy** on the testing and training data, **Confusion Matrix**, **Sensitivity**, **Precision**, **F1 score**, **Recall**, **Specificity**, **ROC** and **AUC**. First, I would like to explain what these performance metrics mean, then I would discuss how the algorithms that I used performed against these metrics.

4.1.1 Performance Metrics

The performance metrics that I used for evaluating the models are discussed in this section.

4.1.1.1 Classification Accuracy

Prediction or Classification Accuracy is one of the most sought-after performance metrics in Machine Learning. Its calculation is done by the formula given below.

$$\text{Classification Accuracy} = \frac{\text{Number of Correct Prediction}}{\text{Total Number of Predictions made}}$$

4.1.1.2 Confusion Matrix

Confusion Matrix is a performance measurement metric with four different combinations of predicted and actual values. It's a matrix that defines and also helps to define the overall performance of a Machine Learning model. There are four combinations that can be derived from predicted and actual values. These are:

- i. **True Positive:** In this combination, both the value that is actual and the value that is being predicted are true.
- ii. **True Negative:** In this combination, both the value that is actual and the value that is being predicted are false.
- iii. **False Positive:** In this combination, the actual value is false but the value that is predicted is true.
- iv. **False Negative:** In this combination, the actual value is true but the value that is predicted is false.

	<i>Predicted TRUE</i>	<i>Predicted FALSE</i>
<i>Actual TRUE</i>	TRUE POSITIVE	FALSE NEGATIVE
<i>Actual FALSE</i>	FALSE POSITIVE	TRUE NEGATIVE

Figure 4.1.1: Confusion Matrix

4.1.1.3 Precision

Precision is out of all positive predictions how many of them our model got it right. Its calculation is done by the formula given below.

$$\text{Precision} = \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Positive}}$$

4.1.1.4 Recall

Recall is out of all actual positive values how many of them our model got it right. Its calculation is done by the formula given below.

$$\text{Recall} = \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Negative}}$$

4.1.1.5 F1 Score

F1 Score is defined as the following equation. It's used for finding the balance between Precision and Recall.

$$\text{F1} = 2 \times \frac{\textit{Precision} \times \textit{Recall}}{\textit{Precision} + \textit{Recall}}$$

4.1.1.6 Sensitivity

Sensitivity is used for the evaluation of a model's ability to predict true positive of each target class. Its calculation is done by the formula given below.

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

4.1.1.7 Specificity

Specificity is used for the evaluation of a model's ability to predict true negatives of each target class. Its calculation is done by the formula given below.

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

4.1.1.8 ROC and AUC

Receiver Operating Characteristic or ROC curve is a type of graph that depicts the performance of an ML model over every threshold of classification. There are two parameters that are plotted on this graph:

- i. **True Positive Rate (TPR):** This is defined as the formula given below.

$$\text{TPR} = \frac{TP}{TP + FN}$$

- ii. **False Positive Rate (FPR):** This is defined as the formula given below.

$$\text{FPR} = \frac{FP}{FP + TN}$$

AUC or **Area under the Curve** on the other hand measures the whole 2D area that rests behind the ROC curve. AUC value ranges from 0 to 1. A model which has 100% prediction accuracy has an AUC of 1.0 and one which has 0% prediction accuracy has an AUC of 0.0.

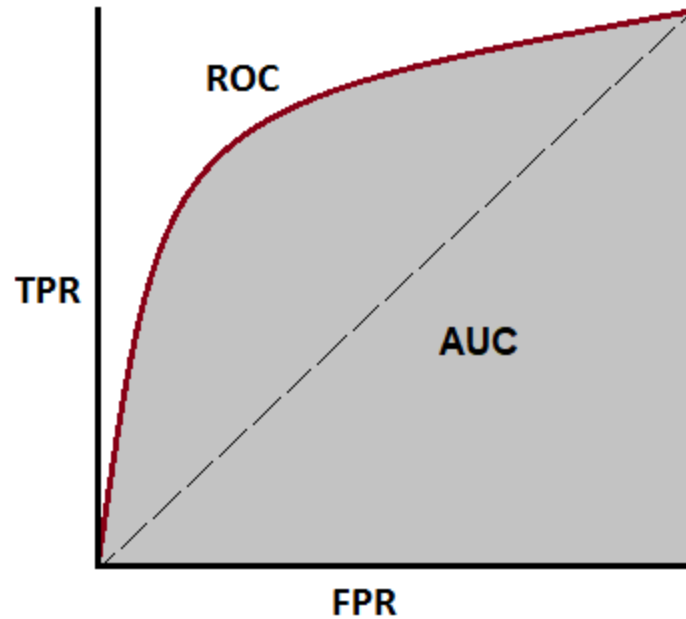


Figure 4.1.2: ROC and AUC

4.1.2 Logistic Regression

Logistic Regression is the type of ML algorithm that uses the supervised learning principle. It is used for the problems that are of classification types. The classification is done in Logistic Regression by **Sigmoid curve** which is calculated by a cost function called “**Sigmoid Function**”. The logistic regression hypothesis suggests that the cost function be limited to a value between 0 and 1 that is $0 \leq h_{\theta} \leq 1$. The **Sigmoid Function** formula is expressed as $f(x) = \frac{1}{1 + e^{-x}}$ and thus the formula for hypothesis of Logistic Regression is as follows.

$$h_{\theta}(X) = \frac{1}{1 + e^{-(\beta_1 + \beta_0 X)}}$$

where,

$\beta_1 + \beta_0 X$ is *Linear equation*

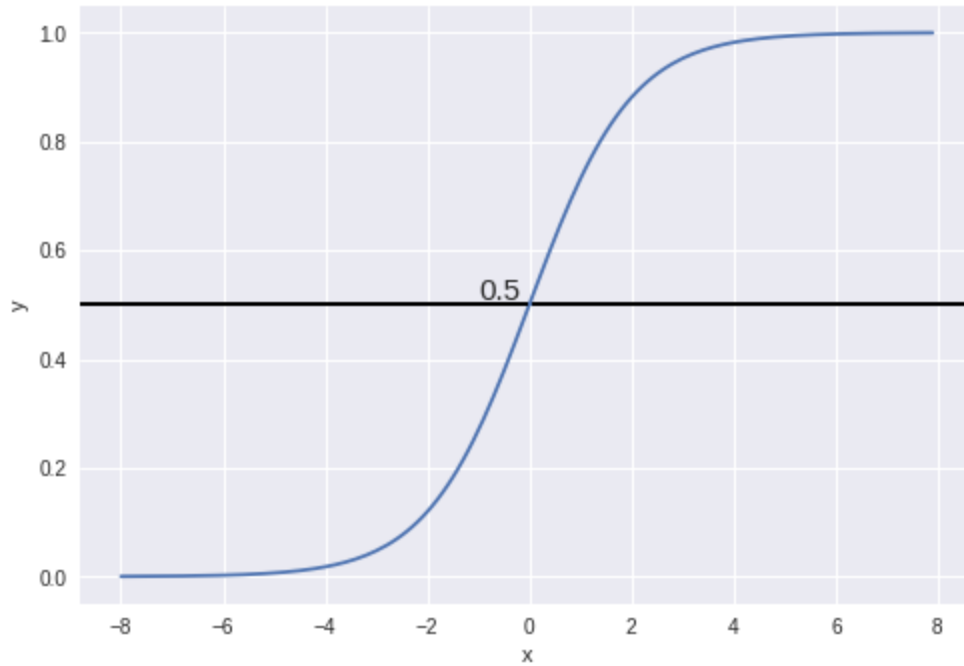


Figure 4.1.3: Sigmoid Function Graph

4.1.2.1 Results

After applying **Logistic Regression**, the results that I've got are given below.

TABLE 3: PERFORMANCE OF **LOGISTIC REGRESSION** MODEL

Accuracy on testing data	Accuracy on training data	Precision	Recall	F1 Score	AUC	Sensitivity	Specificity
94.23%	92.02%	93.93%	96.87%	95.38%	93.43%	96.87%	90.00%

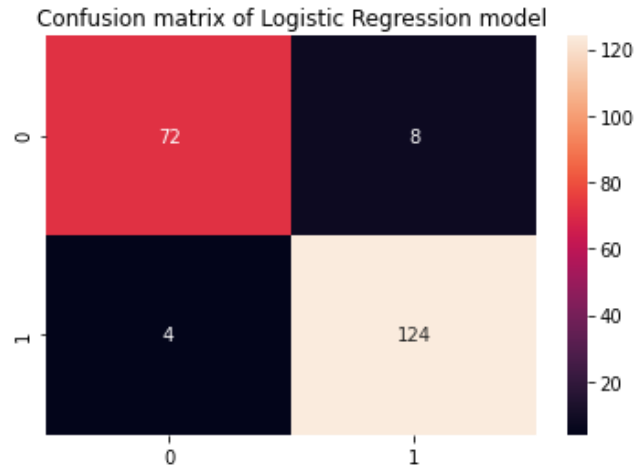


Figure 4.1.4: Confusion Matrix of Logistic Regression Model

4.1.3 Random Forest

Random Forest is the type of ML algorithm that uses the supervised learning principle. It's used for the problems that are of both classification and regression types. Random Forest uses **Ensemble Learning** technique. Ensemble Learning is a method that uses multiple learning algorithms for boosting predictive performance. In this case, Random Forest creates multiple decision trees on the samples of data. It then collects the prediction result from each one of them and eventually, it predicts the final output on the basis of the majority number of votes of predictions. To summarize the working mechanism of Random Forest the algorithm that it follows is given below.

4.1.3.1 Algorithm

Step 1: Randomly choose n samples from the training set.

Step 2: Grow a decision tree from each of the samples. At each node:

Step 2.1: Randomly select d features.

Step 2.2: Split the node using the best split feature.

Step 3: Repeat step 1 to 2 k times.

Step 4: Make prediction based on the majority votes provided by the trees.

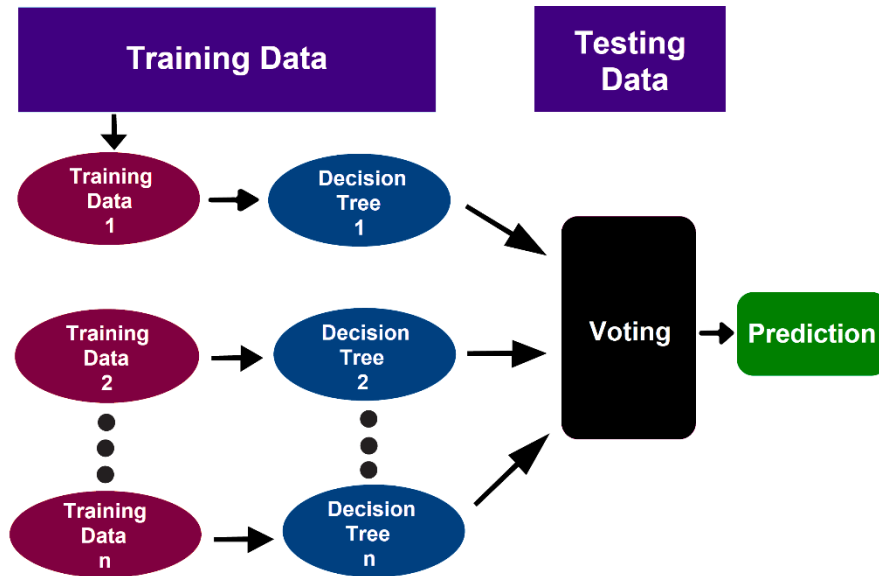


Figure 4.1.5: Working principle of Random Forest

4.1.3.2 Results

After applying **Random Forest**, the results that I've got are given below.

TABLE 4: PERFORMANCE OF **RANDOM FOREST** MODEL

Accuracy on testing data	Accuracy on training data	Precision	Recall	F1 Score	AUC	Sensitivity	Specificity
99.03%	100.00%	100.00%	98.43%	99.21%	99.21%	98.43%	100.00%

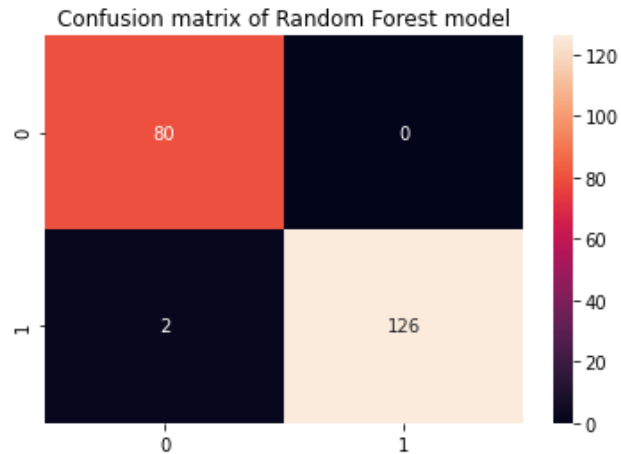


Figure 4.1.6: Confusion Matrix of Random Forest Model

4.1.4 K-Nearest Neighbors

K-Nearest Neighbors or KNN is the type of ML algorithm that uses the supervised learning principle and we use it primarily for the problems that are of classification types. KNN compares the similarity between the available data points and places the new data point into the most similar available category. That means when new data appears KNN can classify it into a well-suited category. Now in KNN, there are two important things that need to be calculated.

- i. First the value of K is calculated. K indicates the number of nearest neighbors that are considered while classifying a new data point.
- ii. The distance from each nearest neighbor is then calculated. This calculation can be done by the **Euclidean Distance** formula.

$$\text{Euclidean Distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

An algorithm is given below to summarize the working principle of **K-Nearest Neighbors**.

4.1.4.1 Algorithm

Step 1: *Select the number K of the neighbors*

Step 2: *Calculate the Euclidean distance of K number of neighbors*

Step 3: *Take the K nearest neighbors as per the calculated Euclidean distance.*

Step 4: *Among these k neighbors, count the number of the data points in each category.*

Step 5: *Assign the new data points to that category for which the number of the neighbor is maximum.*

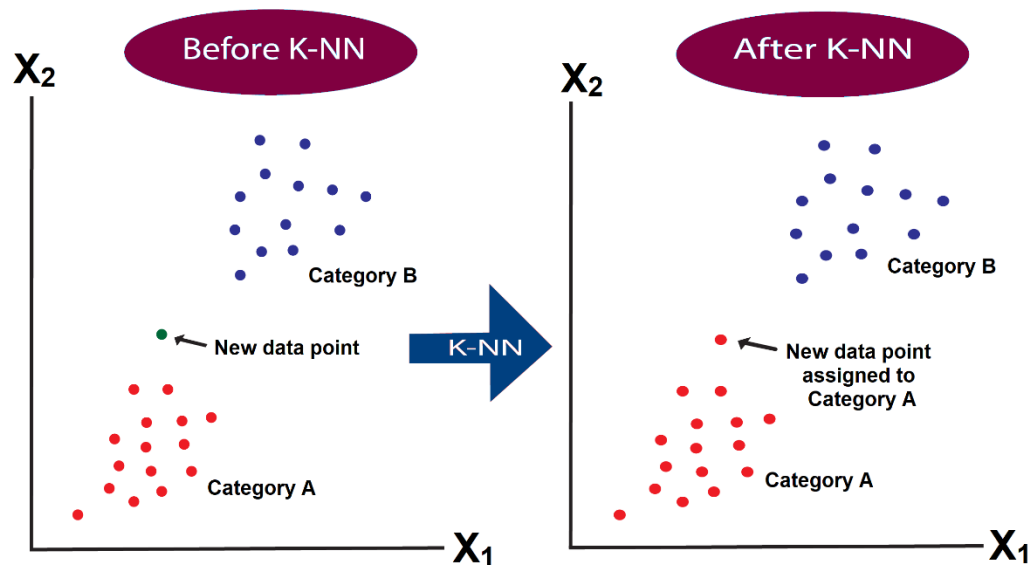


Figure 4.1.7: K-Nearest Neighbors

4.1.4.2 Results

After applying **K-Nearest Neighbors** to my dataset the results that I've got are given below.

TABLE 5: PERFORMANCE OF **KNN** MODEL

Accuracy on testing data	Accuracy on training data	Precision	Recall	F1 Score	AUC	Sensitivity	Specificity
94.71%	92.87%	99.15%	92.18%	95.54%	95.46%	92.18%	98.75%

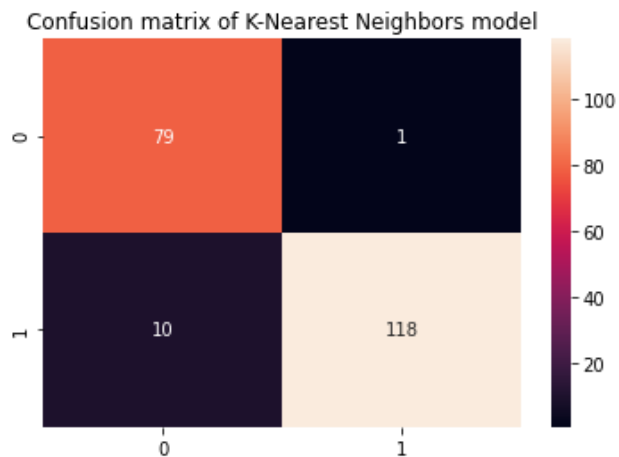


Figure 4.1.8: Confusion Matrix of KNN model

4.1.5 Naive Bayes

Naive Bayes is the type of ML algorithm that uses supervised learning principle. It is used for the problems that are of classification types. The working mechanism of Naive Bayes is derived from **Bayes Theorem**. It predicts by calculating an object's probability. Since the principle of Naive Bayes is rooted in Bayes Theorem, so a brief introduction to it is necessary. With prior knowledge, the Bayes Theorem is used to calculate the likelihood of

a hypothesis. It uses conditional probability principle. The formula for Bayes' Theorem is given below.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where,

$P(A|B)$ *is* **Posterior probability**,

$P(B|A)$ *is* **Likelihood probability**,

$P(A)$ *is* **Prior probability**,

$P(B)$ *is* **Marginal probability**.

Now, the working principle of Naive Bayes can be expressed by the following algorithm.

4.1.5.1 Algorithm

Step 1: *Convert the given dataset into frequency tables.*

Step 2: *Generate Likelihood table by finding the probabilities of given features.*

Step 3: *Use Bayes theorem to calculate the posterior probability.*

4.1.5.2 Results

After applying **Naive Bayes** to my dataset, the results that I've got are given below.

TABLE 6: PERFORMANCE OF NAIVE BAYES MODEL

Accuracy on testing data	Accuracy on training data	Precision	Recall	F1 Score	AUC	Sensitivity	Specificity
87.98%	86.83%	94.01%	85.93%	89.79%	88.59%	85.93%	91.25%

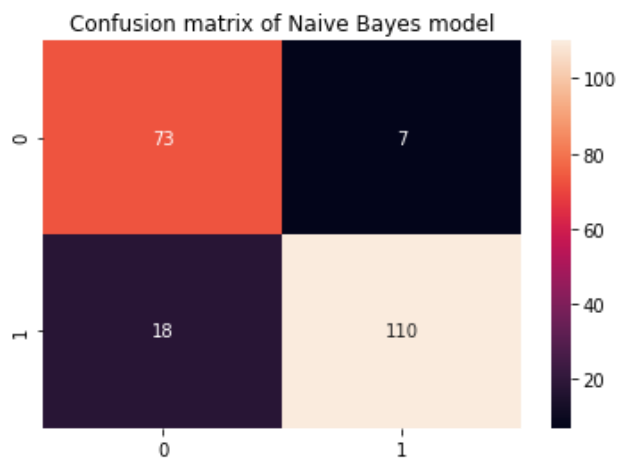


Figure 4.1.9: Confusion Matrix of Naive Bayes model

4.2 Discussion

From the above analysis, I am summarizing the performance of the algorithms that I used in the following table and ROC graph.

TABLE 7: PERFORMANCE OF THE ALGORITHMS

Algorithm	Accuracy on testing data	Accuracy on training data	Precision	Recall	F1 Score	AUC	Sensitivity	Specificity
Random Forest	99.03%	100.00%	100.00%	98.43%	99.21%	99.21%	98.43%	100.00%
KNN	94.71%	92.87%	99.15%	92.18%	95.54%	95.46%	92.18%	98.75%
Logistic Regression	94.23%	92.02%	93.93%	96.87%	95.38%	93.43%	96.87%	90.00%
Naive Bayes	87.98%	86.83%	94.01%	85.93%	89.79%	88.59%	85.93%	91.25%

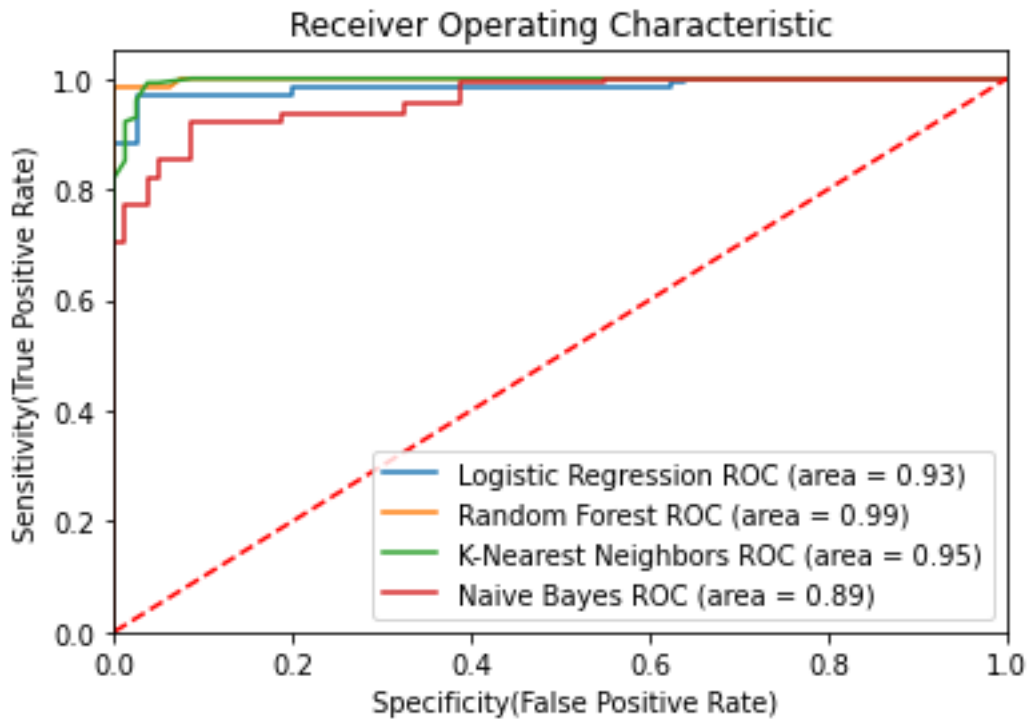


Figure 4.2.1: ROC curve of the algorithms

From the above tabular representation, ROC graph and the result analysis of each of the algorithms that I have used, it is obvious that Random Forest classifier outperformed the other algorithms in all of the performance metrics. It did an outstanding job in the **Sensitivity** metric which is very important in the case of building a model that will be used in healthcare and prognosis.

4.3 Implementation and Deployment

After doing all the above result analysis, for demonstrating the use of the model I have deployed it by a web and an android application. For doing this implementation and deployment I have selected the top 10 features using **Feature Selection** technique represented in the following figure according to the model created by Random Forest since that was the best performing algorithm.

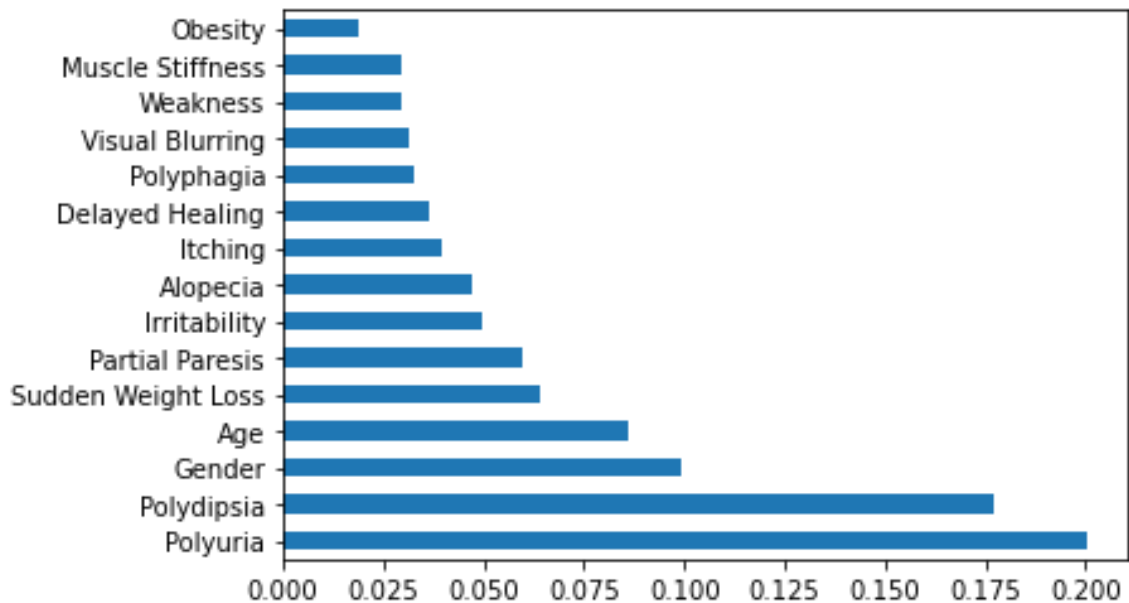


Figure 4.3.1: Feature Selection

Using these top dataset features I have built the prognosis model and implemented and deployed it as the applications. The Implementation and Deployment process is represented in the following figure.

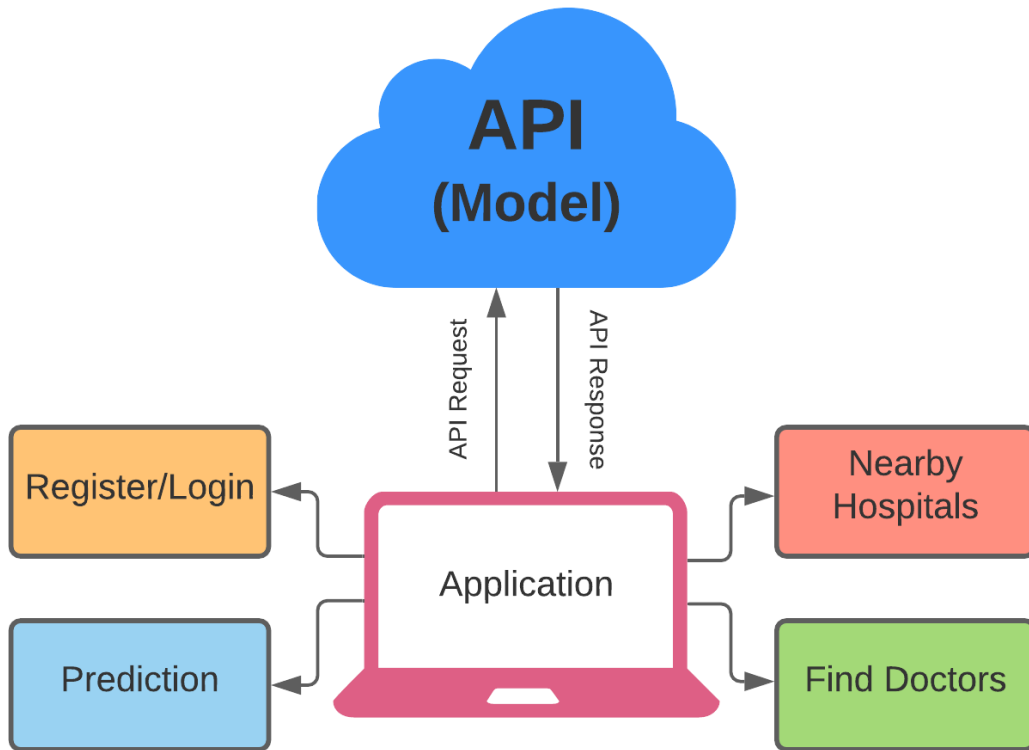


Figure 4.3.2: Implementation and Deployment Process

The features and working mechanism of these applications are discussed below.

- i. Users can login to their profile by their “username” and “password”.
- ii. In their profile dashboard they will find 3 features: **Prediction**, **Nearby Hospitals** and **Find Doctors**.

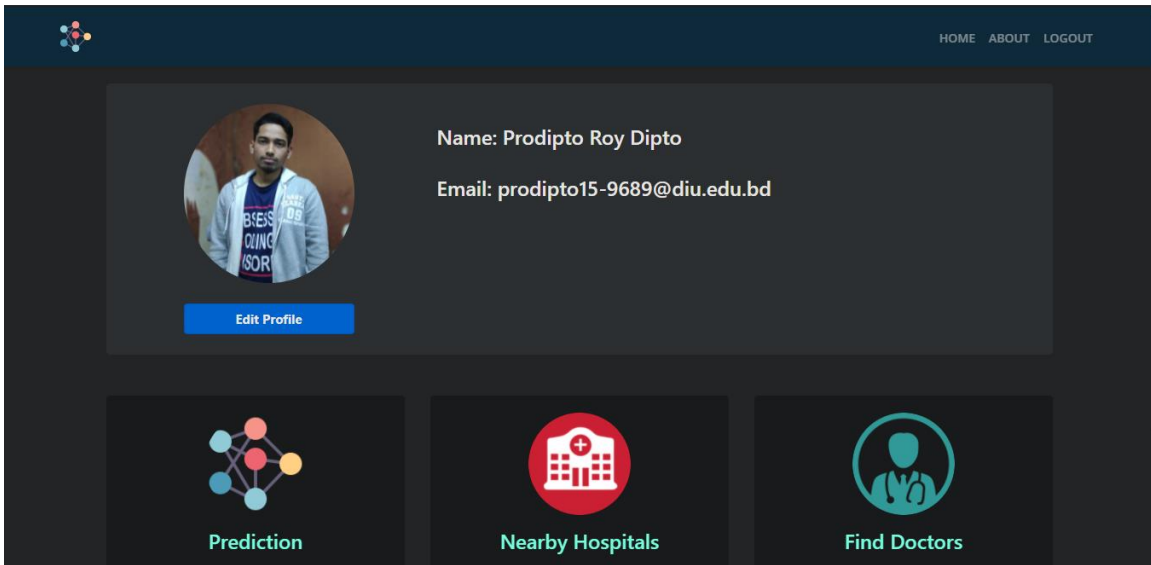


Figure 4.3.3: Dashboard Feature (Web)

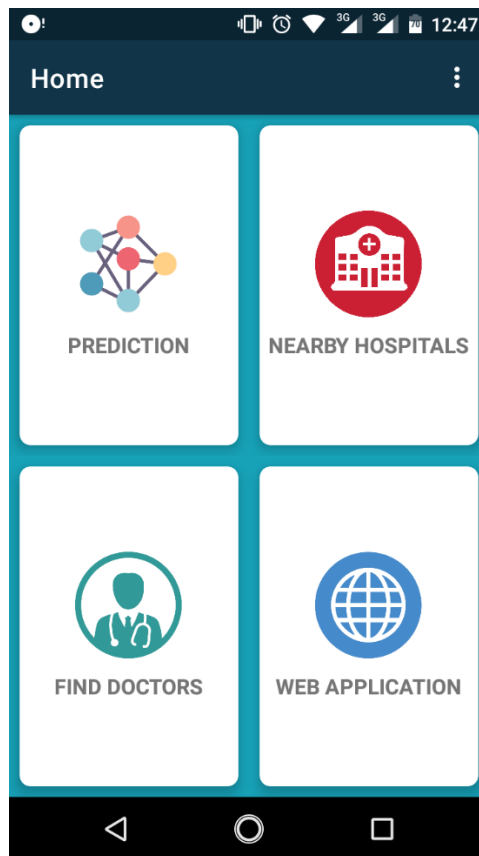


Figure 4.3.4: Dashboard Feature (Android)

- iii. In the **Prediction** feature users can enter their required data according to my model and get their predicted result.

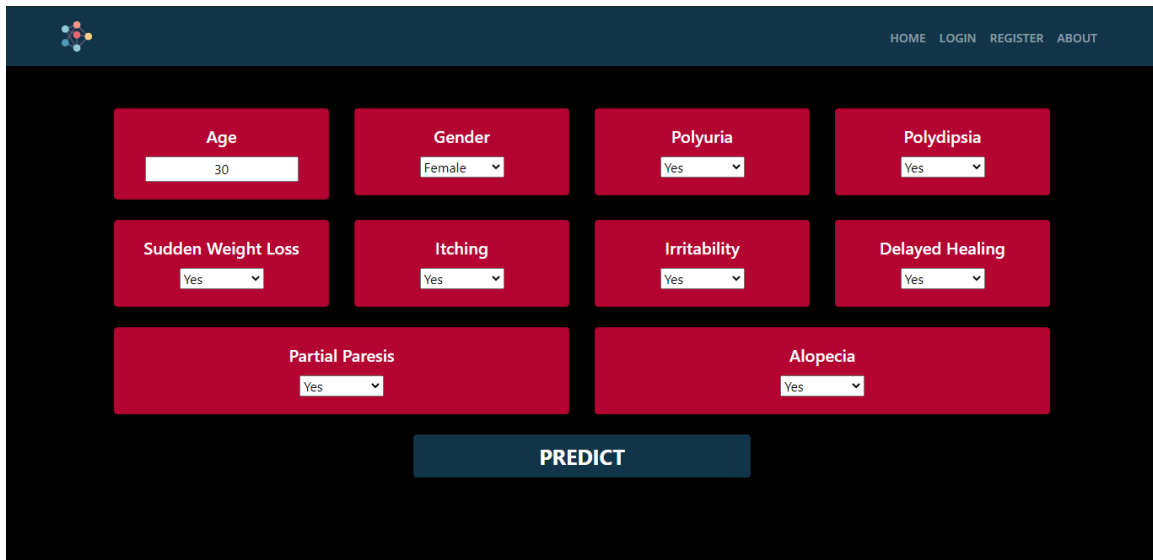


Figure 4.3.5: Prediction Feature (Web)

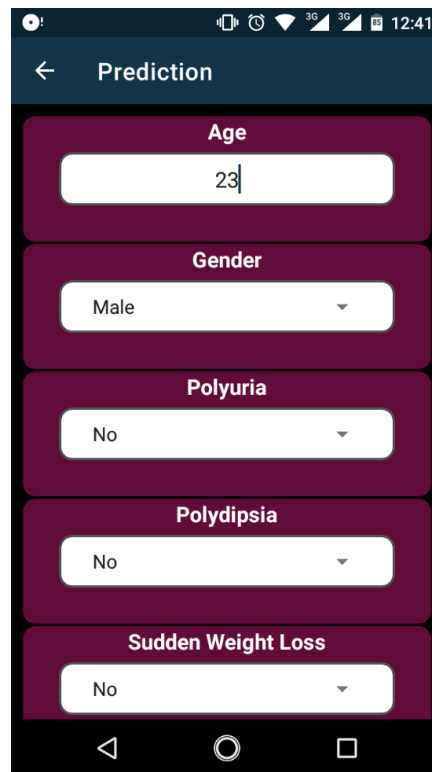


Figure 4.3.6: Prediction Feature (Android)

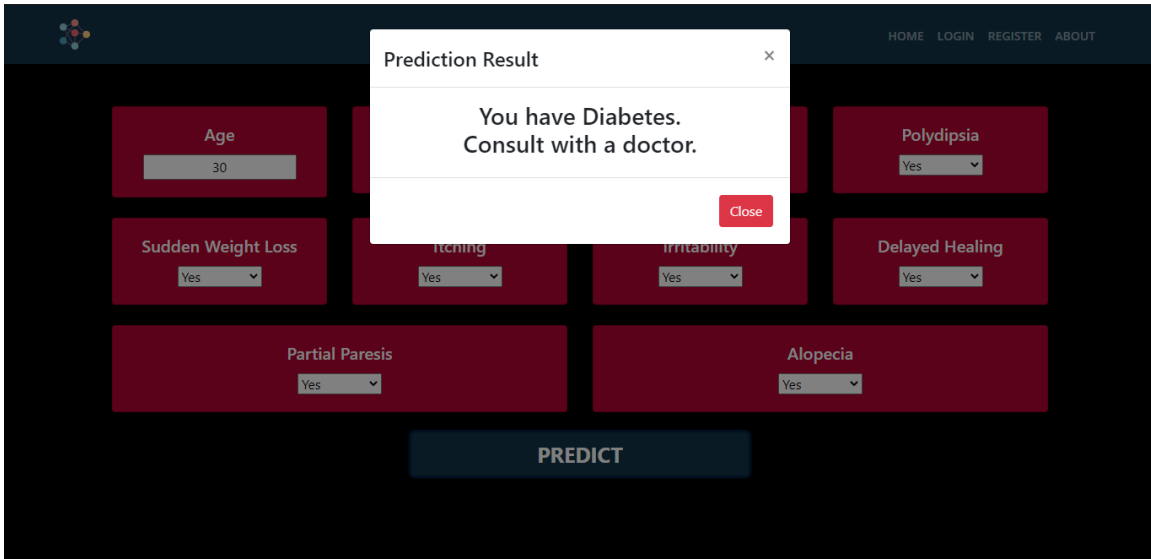


Figure 4.3.7: Predicted Result (Web)



Figure 4.3.8: Predicted Result (Android)

- iv. In the **Nearby Hospitals** feature users can enter the area name in which they want to find the hospitals and can see the hospitals in that area using Google Maps.

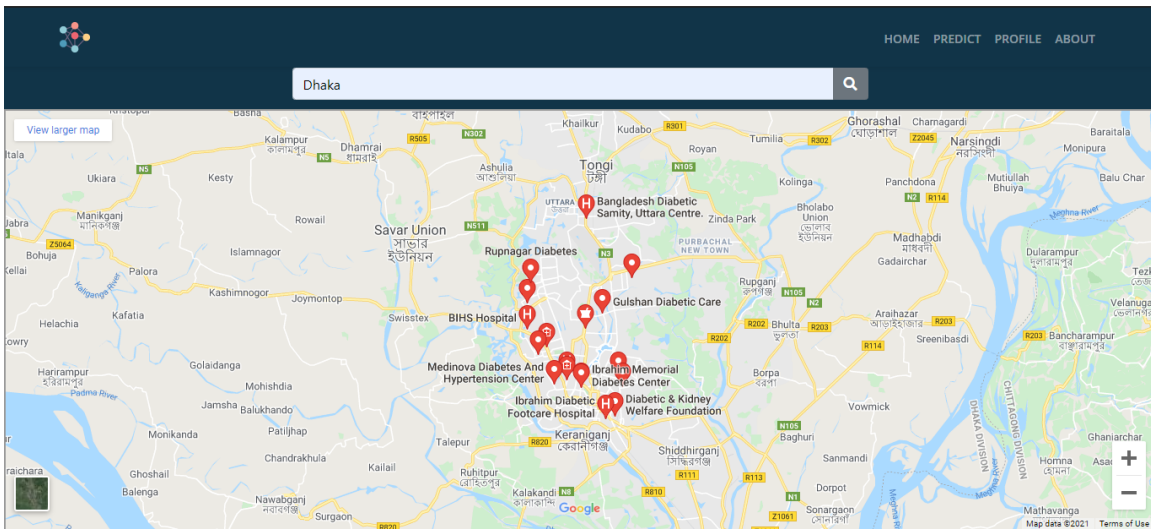


Figure 4.3.9: Nearby Hospitals Feature (Web)

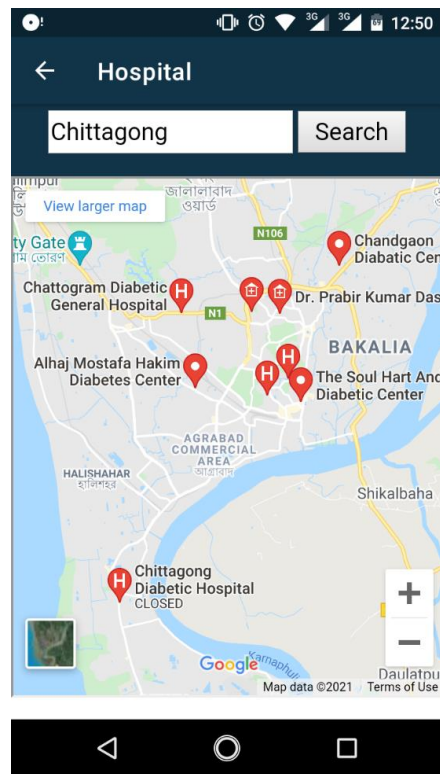
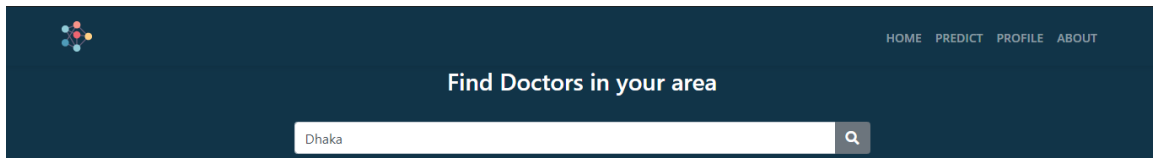


Figure 4.3.10: Nearby Hospitals Feature (Android)

- v. In the **Find Doctors** feature users can enter an area name and find out the doctor names, hospital names and their contact numbers.







ID	Image	Name	Hospital	Area	Contact
101		Dr. A	Square Hospitals Ltd	Dhanmondi, Dhaka	01666666666
102		Dr. B	LABAID Hospital	Dhanmondi, Dhaka	01777777777
105		Dr. E	BSMMU	Dhaka	01444444444
106		Dr. G	Samorita Hospital Ltd.	Dhaka	01111111111

Figure 4.3.11: Find Doctors Feature (Web)

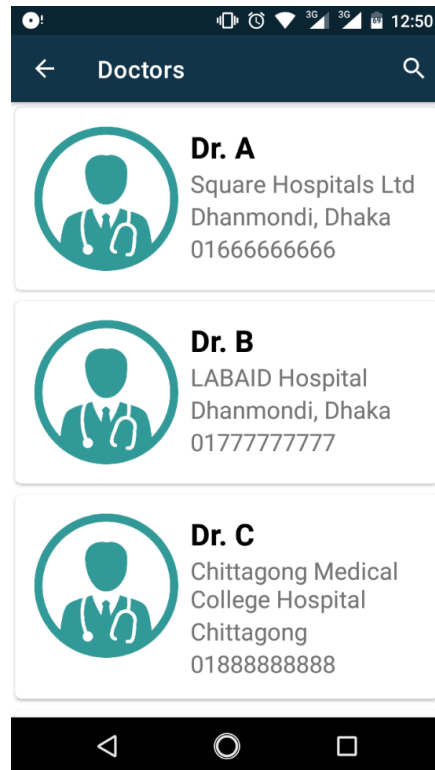


Figure 4.3.12: Find Doctors Feature (Android)

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

I think my research will have a good impact on society if I consider it in the context of Bangladesh. A big portion of Bangladeshi people are suffering from Diabetes and a good number of them are young people. By using Machine Learning, detecting Diabetes at an early stage could be life-saving and beneficial for prolonging the lifespan of an individual in the long run. In Bangladesh, the majority of the population is not so concerned about their health. So, if they can check their Diabetes status easily by an intelligent computer system that I have built in this research without going to the doctor frequently, I think it will impact a lot in the long run in the society of Bangladesh.

5.2 Ethical Aspects

The research that I have done is completely ethical. The data that I have collected were collected with full consent from each individual and these data were only used for this research. More importantly, the overall project that I have done will be used for the goodwill of mankind. So I think there is no ethical concern about this research.

5.3 Sustainability Plan

I started this research with a long-term plan. I think I have completed many of these plans. Since I have done this research almost alone in the middle of a pandemic, I also couldn't do it properly the way I thought. But if I can overcome these problems in the future then this project will definitely be improved.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for Future Research

6.1 Summary of the Study

The prevalence of Diabetes is enormous in Bangladesh and young people are no exception. My study also came up with the same fact. Detecting Diabetes at an early stage is very crucial for the healthy life that a diabetic patient need. For this purpose, I built a Machine Learning based model. In the model building process, four classic Machine Learning algorithms were used. The best performing model which was Random Forest model was deployed as a web and an android application as well.

6.2 Conclusions

This research is a big part of the process of fulfilling my Bachelor's degree. When I started this research, I didn't know much about Artificial Intelligence and Machine Learning and how they are used in the medical and healthcare sectors. Over time I have learned a lot while doing this work and started to enjoy the field of AI. Since I am enjoying it, I am still learning and improving. I hope that this research will be useful in the field of Diabetes and Machine Learning research and in the long run the people of Bangladesh will be benefited.

6.3 Implications for Further Study

The doors are opened. More thorough and profound research is demanded in the field of Diabetes and Machine Learning, particularly in the context of Bangladesh. Even the research that I have done has a lot more room for improvement. A great volume of data should be collected since Machine Learning works amazingly well with a good amount of

data. Other advanced Machine Learning algorithms and more sophisticated AI techniques like Artificial Neural Network, Deep Learning can be used for building a much more intelligent and efficient model. The model can be deployed at a production level with more features and advanced technologies.

REFERENCES

- [1] N. Sarwar, P. Gao, S. R. K. Seshasai, R. Gobin, S. Kaptoge, E. D. Angelantonio, E. Ingelsson, D. A. Lawlor, E. Selvin, M. Stampfer, C. D. A. Stehouwer, S. Lewington, L. Pennells, A. Thompson, N. Sattar, I. R. White, K. K. Ray and J. Danesh, "Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies," *The Lancet*, vol. 375, 2010.
- [2] E. Saedi, M. R. Gheini, F. Faiz and M. A. Arami, "Diabetes mellitus and cognitive impairments," *World Journal of Diabetes*, vol. 7, no. 17, p. 412, 2016.
- [3] T. Desk, "Dhaka Tribune," 13 November 2020. [Online]. Available: <https://www.dhakatribune.com/health/2020/11/13/over-8-million-people-have-diabetes-in-bangladesh>.
- [4] A. Mohiuddin, "Diabetes Fact: Bangladesh Perspective," *International Journal of Diabetes Research*, vol. 2, no. 1, pp. 14-20, 2019.
- [5] S. M. S. Islam, A. Lechner, U. Ferrari, M. Laxy, J. Seissler, J. Brown, L. W. Niessen and R. Holle, "Healthcare use and expenditure for diabetes in Bangladesh," *BMJ Global Health*, vol. 2, no. 1, 2017.
- [6] D. Silver, A. Huang and C. Maddison et al., "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, p. 484–489, 2016.
- [7] A. Senior, R. Evans and J. Jumper et al., "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, no. 7792, p. 706–710, 2020.
- [8] K.-H. Yu, A. L. Beam and I. S. Kohane, "Artificial intelligence in healthcare," *Nature Biomedical Engineering*, vol. 2, p. 719–731, 2018.
- [9] A. Callahan and N. H. Shah, "Machine Learning in Healthcare," *Elsevier*, pp. 279-291, 2017.
- [10] M. A. Helal, A. I. Chowdhury, A. Islam, E. Ahmed, M. S. Mahmud and S. Hossain, "An Optimization Approach to Improve Classification Performance in Cancer," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Cox'sBazar, Bangladesh, 2019.
- [11] S. K. Dey, A. Hossain and M. M. Rahman, "Implementation of a Web Application to Predict," in *2018 21st International Conference of Computer and Information Technology (ICCIT)*, 2018.

- [12] B. Pranto, S. M. Mehnaz, E. B. Mahid and I. Mahmud, "Evaluating Machine Learning Methods for Predicting Diabetes among Female Patients in Bangladesh," *Information*, vol. 11, no. 8, p. 374, 2020.
- [13] M. A. Uddaula, M. A. - A. Hossain, M. K. Hossen and A. A. Marouf, "Implications of Meta Classifiers for Onset Diabetes Prediction," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 9, no. 5, 2020.
- [14] N. S. Khan, M. H. Muaz, A. Kabir and M. N. Islam, "Diabetes Predicting mHealth Application Using Machine Learning," in *2017 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*, Dehradun, India, 2017.
- [15] M. F. Faruque, Asaduzzaman and I. H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2019.
- [16] K. C. Howladar, M. S. Satu, A. Barua and M. A. Moni, "Mining Significant Features of Diabetes Mellitus Applying Decision Trees: A Case Study In Bangladesh," *BioRxiv*, p. 481994, 2018.
- [17] N. Jahan, A. Islam and A. A. Mamun, "Machine Learning With Factor Scoring To Predict Diabetes Risk Level In Bangladesh," *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, vol. 9, no. 2, 2020.
- [18] S. S. Rahman, H. Mahmud, M. Talukder, A. Daria and S. Akhtar, "A Machine Learning Based Approach for Diabetes Detection and Care in Bangladesh," *Gyancity Journal of Engineering and Technology*, vol. 4, no. 2, pp. 21-28, 2018.
- [19] T. M. Le, T. M. Vo, T. N. Pham and S. V. T. Dao, "A Novel Wrapper–Based Feature Selection for Early Diabetes Prediction Enhanced With a Metaheuristic," *IEEE Access*, vol. 9, pp. 7869-7884, 2020.
- [20] T. Biswas, A. Islam, L. Rawal and S. Islam, "Increasing prevalence of diabetes in Bangladesh: a scoping review," *Elsevier*, vol. 138, pp. 4-11, 2016.

Defense Final Summer 2021 Last

ORIGINALITY REPORT

10%

SIMILARITY INDEX

8%

INTERNET SOURCES

2%

PUBLICATIONS

6%

STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	4%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	Submitted to TechKnowledge Student Paper	1%
4	link.springer.com Internet Source	1%
5	Ankur Saxena, Shivani Chandra. "Artificial Intelligence and Machine Learning in Healthcare", Springer Science and Business Media LLC, 2021 Publication	<1%
6	"Implications of Meta Classifiers for Onset Diabetes Prediction", International Journal of Innovative Technology and Exploring Engineering, 2020 Publication	<1%
7	doctorpenguin.com Internet Source	<1%

8	scholar.ppu.edu Internet Source	<1 %
9	Submitted to Indian Institute of Information Technology, Allahabad Student Paper	<1 %
10	lup.lub.lu.se Internet Source	<1 %
11	Daniel Alves de Brito Filho, Rinaldo Artes. "Application of bayesian additive regression trees in the development of credit scoring models in Brazil", Production, 2018 Publication	<1 %
12	www.windowcentral.com Internet Source	<1 %
13	"Chapter 4 Local Properties of Differentiable Mappings", Springer Science and Business Media LLC, 2007 Publication	<1 %
14	Tuan Le Minh, Thanh Vo Minh, Tan Nhat Pham, Son Vu Truong Dao. "A Novel Wrapper – Based Feature Selection for Early Diabetes Prediction Enhanced with a Metaheuristic", IEEE Access, 2020 Publication	<1 %
15	Submitted to Sim University Student Paper	<1 %

16	vcetputtur.ac.in Internet Source	<1 %
17	"Biometric Recognition", Springer Science and Business Media LLC, 2017 Publication	<1 %
18	Dalia F. Elmansy. "Distortion Discovery: A Framework to Model, Spot and Explain Tumor Heterogeneity and Mitigate its Negative Impact on Cancer Risk Assessment", Cold Spring Harbor Laboratory, 2021 Publication	<1 %
19	Submitted to Texas A&M University - Commerce Student Paper	<1 %
20	erepository.uonbi.ac.ke Internet Source	<1 %
21	journals.sagepub.com Internet Source	<1 %
22	studentsrepo.um.edu.my Internet Source	<1 %

Exclude quotes Off

Exclude matches Off

Exclude bibliography Off