Diabetic Prediction System with Machine Learning

BY

Md. Golam Rasul
182-15-2126

Abdul Kayum
181-15-2062

AND

Md. Nazmul Alom Siddki Shuvo
182-15-2198

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

Supervised By
Mr. Ohidujjaman
Assistant Professor
Department of Computer Science & Engineering
Daffodil International University

Co-Supervised By
Md. Sabab Zulfiker
Lecturer
Department of Computer Science & Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh
May 2022

APPROVAL

This Project titled “Diabetic Prediction System with Machine Learning”, submitted by Md. Golam Rasul, Abdul Kayum ,Md. Nazmul Alom Siddki Shuvo 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 18.05.2022.

BOARD OF EXAMINERS

Senior Lecturer **Internal Examiner**
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Senior Lecturer **Internal Examiner**
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Associate Professor, **External Examiner**
Department of Computer Science and Engineering (CSE)
Jahangirnagar University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Mr. Ohidujjaman Assistant Professor Department of Computer Science & Engineering** Daffodil International University. We 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:

Mr. Ohidujjaman
Assistant Professor
Department of Computer Science & Engineering
Daffodil International University

Co-Supervised by:

Md. Sabab Zulfiker
Senior Lecturer
Department of Computer Science & Engineering
Daffodil International University

Golam Rasul

Md Golam Rasul
ID:182-15-2126
Department of Computer Science & Engineering
Daffodil International University

Shuvo

Md. Nazmul Alom Siddki Shuvo
ID:182-15-2198
Department of Computer Science & Engineering
Daffodil International University

Abdul Kayum

Abdul Kayum
ID:181-15-2062
Department of Computer
Science & Engineering
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Mr. Ohidujjaman, Assistant Professor**, Department of Computer Science & Engineering, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our 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 draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Mr. Ohidujjaman, Md.Sabab Zulfiker and Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

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

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Diabetes is a medical word that refers to a variety of illnesses that cause damage to the kidneys, heart, and liver. In recent years, scientific discoveries have enabled doctors to treat this illness utilizing a variety of ways, which they have done. In recent years, artificial intelligence and machine learning have gained importance as tools for improving medical treatment and medical research in general, and medicine in particular. Because diabetes is a potentially fatal illness, it necessitates the application of machine learning to forecast when it will appear in the first place. Machine learning methods, applications, and algorithms may be used in tandem to predict the onset of diabetes. The machine-learning algorithm generates a certain result.and that algorithm surpasses all other algorithms. More complex and algorithms that are dynamic need to be used to get the best results, and these algorithms include Random Forest, Nev Base, Decision Tree, K-Nearest Nebert (KNN).

Table of Contents

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

In general, diabetes mellitus or diabetes is the deficiency in the human body of a hormone called insulin or high levels of glucose/blood sugar. Diabetes is a familiar problem in today's rich, poor, developing, and underdeveloped countries of the world. According to the World Health Organization (WHO), 1.6 million people worldwide die of diabetes each year. Hyperglycemia is another name for blood pressure in medical terms[1]. The hormone insulin from the stomach of the human body absorbs glucose and provides the necessary energy to the cells. Diabetes mellitus can be caused in two ways, first by heredity and second by insulin deficiency. Failure to supply the glucose-insulin hormone, which is needed by the human body's cells, leads to insulin deficiency and elevated glucose levels in the human body, leading to diabetes. The function of insulin is to transfer blood glucose from cell to cell[2].

Blood glucose supplies all the energy to the cells. If the pancreas does not make enough insulin, the cells cannot absorb glucose, and blood clots cause hyperglycemia. This results in high blood glucose/blood sugar levels in the human body and various symptoms such as loss of appetite, frequent urination, excessive thirst for water, dry throat. The normal amount of glucose in a human body is 70 to 99 mg per deciliter. If the amount of glucose in a human body is between 100 and 125 mg, then it is considered as primary diabetics. But if the glucose level is above 126 mg, it is considered diabetes. People with diabetes gradually develop complications of heart disease, diabetes disease, and neurological disease. Diabetes is a long-term problem and there is no permanent solution to this problem. Problems with diabetes, eyes, legs, and nerves for Micro vascular complications damage small blood vessels. Although diabetes cannot be completely controlled, it is possible to control diabetes if it can be detected quickly. It is possible to prevent diabetes through proper diet and regular physical exercise[3]. Through daily physical exercise, a person with pre-diabetes can lose weight and maintain proper blood glucose levels. Reliable and advanced medical travel equipment has been developed using data mining and machine learning.

A machine learning process is data mining where a large amount of data is stored in a statistical and database system and from which a data pattern can be discovered.
According to Nvidia: Machine learning uses a variety of algorithms to learn and make predictions from parsed data[4].

1.2 Motivation

Bangladesh has a high number of persons who have Diabetes. About 9 per cent of the inhabitants of Bangladesh were afflicted by Diabetes. Illness of Diabetes is a severe concern in Bangladesh. The main cause of Diabetes is Obesity, overweight, and a sedentary lifestyle. majority of the food we consume is transformed into sugar known as glucose and absorbed into our circulation. In Bangladeshi, people consume tons of food that includes glucose. As a consequence, Stroke, heart attack, chronic diabetes disease, neuropathy, visual impairment, and amputations are all probable outcomes. Thus, you may develop a system based on Machine Learning to predict Diabetes. It would be highly useful for the people. People may be treated early-stage and conscious of Diabetes.

1.3 Rationale of the Study

There are three forms of diabetes. Type-1 diabetes causes the body to produce no insulin, while type-2 diabetes causes the body to produce or use insulin ineffective and insulin sensitivity reduces during pregnancy, resulting in gestational diabetes. If diabetes can be diagnosed and predicted early, it can be cured by proper treatment very quickly. in the actual world, people must test a variety of things in the lab to determine whether or not they have diabetes, which is a time-consuming process. As a result, a model that can predict diabetes Disease at any stage has been suggested and trained using a relevant dataset. our main goal, is for Patients, can provide data and have their diabetes disease predicted at home using a web interface. It will be beneficial to doctors and patients with diabetes Disease.

1.4 Research Questions

There are numerous questions that might be raised in relation to this study. To make this study more concise, a set of questions was extracted from several persons.

- **Why was diabetes prediction the study's goal?**

  Diabetes is a major medical illness that affects people all over the world. Diabetes is becoming more prevalent every day, affecting people not only in Bangladesh but around the world. There are several forms of diabetes, the most deadly of which being type 2. Type-2 diabetes has a negative impact on the immune system and can lead to mortality. So, if diabetes can be predicted at an early stage, it can aid ineffective treatment and the observance of many norms and regulations. That is why the study's purpose was to predict diabetes.

- **Why use machine learning? Is it trustworthy?**

  Machine learning is a popular tool for producing various predictions. Using a massive amount of data, a model may train itself and predict any outcome. In a medical dataset, a machine learning approach can easily predict diabetic disease. The world is entering a moment of modernization right now. These prime points were nothing more than a term with some mathematical reasoning ten years ago, when Artificial Intelligence and Machine Learning were still in their infancy. Artificial Intelligence, on the other hand, now powers half of the world's technology. As a result, even if it is already dependable, appropriate practice and increased precision in this department might improve its reliability.

- **What is the primary reason for using a web interface?**

  People can use a web interface to visually display output. It is the most effective method of displaying the outcome. We create a machine learning-based system and then save the project. We can quickly display results via the internet. A web interface can be accessed from anywhere in the world. We obtain the essential information from the user and enter it into the machine. Through the web interface, the machine can forecast the output. As a result, it is a practical method of predicting diabetes.

- **What are the benefits of using distinct algorithms?**

  Our primary objective is to develop a diabetes prediction system that is as accurate as possible. Basically, we're looking for an algorithm that will offer us the least amount of errors and the maximum level of accuracy out of all the methods. If just one method is chosen, finding the optimal algorithm is impossible since no one knows which algorithm will best suit the dataset.

1.5 Expected Outcome

The major subject or desired outcome has changed several times over this research period. It contributes to a better understanding of the study's precise conclusion. This research has the potential to know diabetes at its earliest stages while also identifying the root cause of the disease. Doctors and analysts may be able to determine the age at which people develop diabetes. A complete internal assessment of diabetic disease can be revealed for the sake of medical analysis using accurate calculations and algorithms. The last strategy employed in this study was the creation of a web-based system that aids in the prediction of production.

1.6 Report Layout

Six separate chapters are covered in this study in order to make the research report more compact and efficient for readers and researchers.

Table 4.1: Report Layout

Chapter	Description
Chapter 1	Chapter 1 This section serves as a vital introduction to this research project. This is about diabetes and its symptoms. This chapter explains the research objective, the reason for this study, pertinent research questions, expected results, and total management information, as well as financial issues.
Chapter 2	Chapter 2 contains the complete description regarding the background of this investigation, such as machine learning system, classification information, and associated work based on this research study. Comparative analysis breadth of this issue statement is also explained in this chapter with perceived obstacles.
Chapter 3	Chapter 3 includes the descriptive information regarding methodology suggested system and system architecture for this research project. Algorithmic specifics for each employed algorithm is presented from mathematical scratch to the current state is explained.
Chapter 4	Chapter 4 presents the whole result analysis for the outcome of each stage. In ends with the highest accuracy score with the best algorithm, Jaccard score, cross-verified score, confusion matrix, classification report. ROC-AUC curve is also given for each algorithm. Misclassification, Mean Absolute Error and Mean Squared Error is covered in the concluding portion of this chapter.

Chapter 5	Chapter 5 highlights This study's impact on society with Ethical Concerns is critical for every effective research endeavor. The final component of this chapter addresses the sustainability of this research endeavor.
Chapter 6	Chapter 6 illustrates the future scope of this research activity where it is briefly discussed as the expansion of this research study. This chapter closes the complete study report with a beneficial conclusion where essential results of this research are quickly explained.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

Diabetic disease has long been a major problem, as discussed in this chapter. As a result, there is a lot of painful history and many losses in the background of this condition. So much work has gone into preventing diabetes and saving lives. The basics and background of this condition have been explored in this chapter. This chapter discusses some relevant work on this subject. Finally, a comparison study has been provided to show how much the current work has been affected.

2.2 Related Works

The most accurate and widely used method of diagnosing diabetes is the oral glucose tolerance test or OGTT. In this method, the patient has to take a blood glucose test once in the morning on an empty stomach, then another two hours after drinking 75 grams of glucose sherbet. This method can accurately diagnose diabetes and pre-diabetes. But many often get in trouble for testing. Because, you have to give a blood sample in the morning without eating for at least eight hours, you have to give blood twice and sometimes you have to rest or sit in the lab for two hours. Many people do not want to drink sweet solutions on it. Scientists were therefore looking for a simpler method. In different parts of the world, therefore, tests called HBA1C are used to diagnose diabetes. This indicates an average of several months of blood sugar. This test can be done at any time of the day and a blood sample has to be given only once. [5]

According to the guidelines of the American Diabetic Association, if the HBA1C value is below 5.6, it can be considered normal. If it is more than 7.5, it will be considered as having diabetes. If this value is between 5.6 and 7.5, it should be considered as pre-diabetes or pre-diabetes.[6] However, the HBA 1C test should be done in the prescribed manner in a good quality laboratory. If someone has a hereditary blood disorder or has a problem with blood clotting, this test may not give the right result. This test is not used to diagnose gestational diabetes.[7]

Regular diabetes screening should be done after the age of 40 if you are overweight, have a family history of diabetes, and have other risks. For this purpose oral glucose tolerance test or HBA1C test can be done[8]

This study paper deals with Diabetes Mellitus At first glance, it demonstrates the dangers of Diabetes Mellitus. This illness affects a large number of people. Diabetes Mellitus can be caused by a variety of factors, including age, lack of exercise, poor discipline, and so on. Diabetes patients have a higher chance of ailments such as heart attacks, eye difficulties, and other complications, according to this study. This research also demonstrates that there are several current methods for predicting, but that their categorization and prediction accuracy are not particularly spectacular. In this study, we offer a diabetes prediction model that includes external parameters for improved categorization. In this work, we employed a variety of machine learning techniques, including Logistic Regression, which had a 96 % accuracy, and Adobos' classifier, which had a 98.8 %accuracy. We can come up with a lot of ideas from here [9].

This research paper shows that Machine learning and data analytics can be used to improve the accuracy of a patient's first diabetes prognosis. Here it is: the goal of the project is to achieve the maximum level of accuracy feasible. This research argued that we can improve accuracy by using a larger portion of the dataset for training and a smaller portion of the dataset for testing. We can also notice that three specific machine learning methods are employed, the most suited of which are support vector machines, logistical regression, and artificial neural network diabetes, which will be explained later in this section [10].

This paper deals with the significance of Decision Support Systems (DSS) and the important diabetes conditions this illness is responsible for countless fatalities across the world today. According to the International Development Association (IDA), 382 million people would be infected by the illness by 2035. Based on machine learning techniques and deep learning methodologies, this research provides a DSS for diabetes prediction. For machine learning, we looked at the most often used classifiers. On the other hand, we used a completely Convolutional Neural Network (CNN) to predict and detect diabetes patients in the Random Forest (RF) and Support Vector Machine (SVM) and Deep Learning (DL) models. A total of 500 samples were found to be non-diabetic, whereas the remaining 268 were diabetes individuals. We did a side-by-side comparison of of ML and Dl-based algorithms here, showing that RF was more effective. The prediction accuracy for SVM came up to 65.38%, while the DL method reached 76.81% [11].

This paper deals with designing a prediction system of blood glucose that can also be used as a continuous glucose monitoring device .They needed to improve the rate of false warnings on current equipment as quickly as possible because it was so high. The Kalman filter, which can anticipate glucose for up to two hours, is assisting them. It demonstrates that characteristics such as weight, lifestyle, and activity have a significant impact on the accuracy with which glucose concentration can be predicted. It makes use of a technique for decoupling noise from glucose levels. The MAD percent values grow as the prediction window length increases, according to the data. The outcome of the prediction reveals that patients should employ narrow prediction windows during the process to avoid hyperglycemia and hypoglycemia. This research might be expanded to include dealing with people on a one-on-one basis [12].

This paper deals with a non-invasive method to monitor glucose levels because the existing method is painful and has the risk of infection because it involves finger puncturing. For measuring blood glucose, Near-infrared LED was positioned over the fingertip. It was usually predicted based on the voltages that had been received and assessed. The glucose levels can be transferred to a smartphone and shown via an app for Android or IOS. Validation and calibration were staged for the prototype. In comparison to the prick approach, the designed method had a percentage inaccuracy of 7.20%. The glucose reading can be considered accurate if the percentage difference stays within 20%, according to the "Clark Error Grid," which is used to evaluate the clinical accuracy of a glucose sensor. As a result, there is some competition between the anticipated glucose readings and the sensor voltage signals [13].

This paper shows that diabetes mellitus is a metabolic disease, and its multiple causes are called chronic hyperglycemia. Some of the symptoms of this disease are weight loss, polyuria, polyphagia, and polydipsia. Prolonged processed foods, such as carbohydrates, fats, and sweets, can lead to chronic complications, leading to progressive disease and hypo functional and there is also damage to the organs of the human body. Diabetic ketoacidosis (DKA), the hypertonic hyperglycemic syndrome, is one of the diseases that can be severe or under stress. Then it can be said that there is no good research on diabetes and age limit. We can see in this paper the different methods of diagnosing diabetes. These include the highest accuracy of SVM subject to the Confusion Matrix Assessment test that we will understand later in this paper. As a result, these studies need to be updated regularly [14].

2.3 Comparative Analysis

this paper of Deepti Sisodia et al. claims to have created a model for properly predicting diabetes in its early stages in their publication. They're running an algorithm on the Pima Indians Diabetes Database dataset. This dataset is not unique, and many writers have already used it to this data set, which has a high level of accuracy. They're utilizing three categorization approaches to evaluate a variety of variables. To predict diabetes, they employ Decision Tree, SVM, and Naive Bayes algorithms. They reach a maximum accuracy of 71 % using Naive Bayes. When compared to other current algorithms, the suggested system performs poorly. On the other hand, The author of this research, achieved a high accuracy of 96.83% in Naive Bayes, with an AUC score of 97.37 %, a Cross-validation score of 94.0%, a Jaccard score of 94.74%, a Cross-validation score of 94.0%, Jaccard score of 94.74% which is better than the previous studies .and other things we collected unique dataset from Bangladeshi hospital.[9]

In this research of Nongyao Nai-arun et al. predicted the risk of diabetes mellitus with the help of algorithms. They're using Decision Tree, Artificial Neural Networks, Logistic Regression, and Naive Bayes methods to create a classification model. They are also online programs that display the outcome of risk prediction that is difficult to use. They're gathering information from the Sawanpracharak Regional Hospital. There are a lot of null values in this collection. They are utilizing Decision Tree (DT) with an accuracy of 85.090 %, Logistic Regression (LR) with an accuracy of 82.308%, and Naive Bayes (NB) with an accuracy of 81.010 %. In comparison to another author of the work, accurate's performance is poor. However, the authors of this work achieve great accuracy by employing a unique dataset. Decision tree accuracy was 93.65%, Logistic regression was 100%, Random forest was 100%, and Naive Bayes was 94.77 %.[10]

the study of Shankaracharya et al. looks at how artificial networks and machine learning algorithms may be used to identify diabetes early. Using the Pima Indian dataset, they found high accuracy in predicting the correct diabetes diagnosis. They use the algorithms (SVM) support vector machine, (ANFIS) adaptive neuro-fuzzy inference system, (RBF) radial (K-NN) K-nearest-neighbour, (LVQ) learning vector quantization, (LDA) linear discriminant analysis, (ME) mixture of experts, and (MME) modified mixture of experts to predict diabetes-like symptoms. They use Jacobs' proposed way to obtain great precision (ME). Using this dataset, they are able to obtain high accuracy. They also didn't use a web interface to display the results.[11]

2.4 Scope of the Problem

Bangladeshis consume a lot of carbohydrate-rich foods. Carbohydrate-based foods put us at a higher risk of developing diabetes. Bangladesh also has other things to offer. Another cause of diabetes illness is a sedentary lifestyle and a large gap between the number of diabetic patients and doctors. Diabetes patients in Bangladesh are also reasoning due to increased age, overweight/obesity, hypertension, and the highest wealth quintile. According to a WHO estimate, diabetes affects 13.90 million people in Bangladesh, accounting for 3% of all fatalities of all ages. As a result, diabetes must be treated at an early stage. The fatality rate from diabetes will drop if we can predict it early and treat it properly. Another thing that people are aware of is that persons with diabetes and diabetic patients will have fewer complications.

2.5 Challenges

The biggest challenge of this study is finding the relevant variables to use when assessing whether or not someone has diabetes. The removal of all null values from the dataset that has been gathered is a difficulty.

This might be a lengthy procedure. It was also difficult to create an algorithm that was appropriate for the scenario. The scientists used various algorithms to train the dataset and then picked the methods with the highest accuracy for diagnosing diabetic illness from the training data. One of the most difficult jobs for the team was to create an intuitive user interface for this system, which allows anybody to enter data and anticipate arguments at any time.

CHAPTER 3
RESEARCH METHODOLOGY & SYSTEM ARCHITECTURE

3.1. Introduction

Before starting research on any particular things planning and technique are crucial. We need to design a mechanism for locating a certain solution. The study topic is discussed in this chapter. The algorithms that were utilized to solve the problem were then detailed, followed by a description of the process with a graphic representation for easier comprehension. For better visibility, a system architecture was also shown.

3.2 Research Subject

The basic goal of the research is to identify an issue to study and then discover a solution to that problem. Diabetes is a prevalent condition that is inextricably linked to the ageing process. It is a chronic illness that might deteriorate with time. Many individuals die each year as a result of diabetes. If any sort of diabetes disease can be predicted at an early stage, the patient can be treated and have diabetes under control. That is why it is critical to focus on diabetes and develop a model that can predict diabetes at various ages. A web interface which will also be user-friendly and people will be able to get their report at home.

3.3 Machine Learning Techniques

A machine learning system is described as one that is self-contained and capable of continuously obtaining and integrating data for the goal of making choices. It is possible to learn from past events, make analytical observations, and employ other strategies to construct a system that is continually improving. Machine learning techniques come in a wide range of sizes and forms.

3.3.1 Supervised Learning

Supervised learning is a method for developing artificial intelligence that involves training a computer system on input data that has been labelled for a certain output. Supervised machine learning approaches employ labelled samples of previous occurrences to create predictions about the future. The learning approach develops an inferred function from a study of a well-known training dataset in order to provide predictions about the output values. After then, there are no boundaries to how much training the system can handle. When the learning algorithm compares its output to what was intended, it may detect mistakes and make necessary adjustments to the model. This thesis uses different supervised learning approaches to classify diabetic diseases at the initial stage.

3.4 Classification Techniques

Classification is a type of data analysis in which models defining relevant data classes are extracted. It's the most extensively used and popular machine learning technology. Under supervised learning, such models, called classifiers, may predict categorical class labels. The predictions are unordered and discrete. The classifier can't provide an intermediate value. A classifier is used to determine if an image contains a picture of a dog or a cat, for example. Either "dog" or "cat" will be the forecast. The classifier can't provide an intermediate value. On labelled data, classification learning techniques may be applied. In classification learning, there are two types of data. One type is referred to as training data, while the other is referred to as test data. The model is built using training data, and the model is validated using test data. There are two phases to the categorization process.

3.4.1 Learning

An appropriate approach and training data are used to develop a classifier during the learning phase, which is subsequently assessed against the actual world. A classifier is simply a set of rules that may be applied to a range of different scenarios in various settings when a classification algorithm and training data are combined.

3.4.2 Classification

Following the learning phase, the classifier or model established during the prediction phase may be used to forecast which class of unknown data will be predicted. This section uses test data to determine whether a model's predictions are accurate.

3.5 Algorithmic details

Five of the best Machine Learning algorithms were employed to conduct this study project. In general, an algorithm may be described as an ordered set of instructions that tells computer software how to change a set of input data into useable data. Statistics are facts, and valuable information is any knowledge that is beneficial to humans, robots, or algorithms. Machine learning algorithms function in a similar way, using a flow and some mathematics. In general, not all Machine Learning Algorithms have the same mathematical transformation. This research study, on the other hand, contains the most essential machine learning algorithms, as well as relevant algorithmic processes in the overall system design.

3.5.1 Logistic Regression

We can predict a binary result of 0 or 1 using Logistic Regression techniques. By incorporating a multitude of variables rather than just one, these models may be able to solve issues that are significantly more challenging. As the Y-axis moves from 0 to 1, the X-axis moves from 0 to 1. This is due to the fact that the sigmoid function always utilizes these two values as the maximum and minimum, which is excellent for classifying data into two categories.

By computing the sigmoid function of X, this system obtains a probability (between 0 and 1) of an observation belonging to one of the two categories. The sigmoid function is represented by the formula from equation (i)

The sigmoid function has the following formula:

$$\mathrm{Sigmoid(x)} = \frac{1}{1 + \mathrm{e}^{-\mathrm{x}}} \dots\dots\dots (i)$$

Therefore, for Logistic Regression the equation becomes equation (ii).

$$\mathrm{P} = \frac{1}{1+\mathrm{e}^{-(\mathrm{b}' + \mathrm{b}'' \mathrm{x})}} \ldots \ldots \ldots \quad (ii)$$

3.5.2 K-Nearest Neighbors

By employing a simple supervised machine learning methodology, the k-nearest neighbors' method (KNN) may be used to solve both classification and regression issues. KNN is easy to

comprehend and put into practice. The core theorem of KNN is the Euclidean distance. Because the dataset is divided into two classes, KNN is used in this classification. The real formula for the K-Nearest Neighbor method is derived from equation (iii).

$$\mathrm{D}(\mathrm{x_i}, \mathrm{x_j}) = \sqrt{(\mathrm{x_i} - \mathrm{x_j})^2 + (\mathrm{y_i} - \mathrm{y_j})^2} \ldots \ldots \ldots (iii)$$

3.5.3 Gaussian Naive Bayes

Gaussian Naive Bayes is a variant of Naive Bayes that allows for Gaussian normal distribution and continuous data. Naive Bayes is a set of supervised algorithms for classification machines based on the Bayes theorem. The categorizing method is simple yet effective. When working with continuous data, it's common to assume that the values for each class follow a normal (or Gaussian) distribution. The entire formalization approach for the Gaussian Nave Bayes algorithm is obtained in equation (iv).

$$\mathrm{P}(\mathrm{x_i}|\mathrm{y}) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(\mathrm{x_i} - \mu_y)^2}{2\sigma_y^2}\right) \ldots \ldots \ldots (iv)$$

3.5.4 Decision Tree

Decision trees can be used to solve classification and regression problems in supervised learning. They are, nevertheless, most often utilized to solve categorization difficulties. Internal nodes represent dataset attributes, branches represent decision rules, and each leaf node represents the conclusion in this tree-structured classifier. To do so, you'll need to utilize the Decision Tree to create a training model that uses fundamental decision rules to predict the class or value of input variables based on the training dataset.

$$\sigma = \sqrt{\frac{\sum_{b-1}^{B}\left(f_b(x') - \hat{f}\right)^2}{B-1}} \dots \dots \dots (v)$$

In the equation (vii), S means the current state and is the likelihood of an event, i in state S or the percentage of class i in the node state S.

3.5.5 Random Forest

Random forests, also known as random decision-making forests, are an ensemble learning approach that uses numerous decision-making trees to classify, regress, and do other tasks. For classification problems, the random forest output is the class picked by the majority of trees. For regression tasks, the mean or average forecast of the individual trees is returned.

$$\sigma = \sqrt{\frac{\sum_{b-1}^{B}\left(f_b(x') - \hat{f}\right)^2}{B-1}} \dots \dots \dots (vi)$$

In equation (viii), B is is a perimeter free of charge. The average prediction error for each sample of training X'', which is just the trees with no X' in their bootstrap sample. After a few trees have been fitted, the training and test errors decline.

3.6 proposed system

After taking into account all of the aforementioned algorithms, the desired system may now be proposed. A system diagram, such as the shown in Figure 3.1, is better appropriate for understanding the system's exact operation and Figure 3.2 is understanding Web Interface.

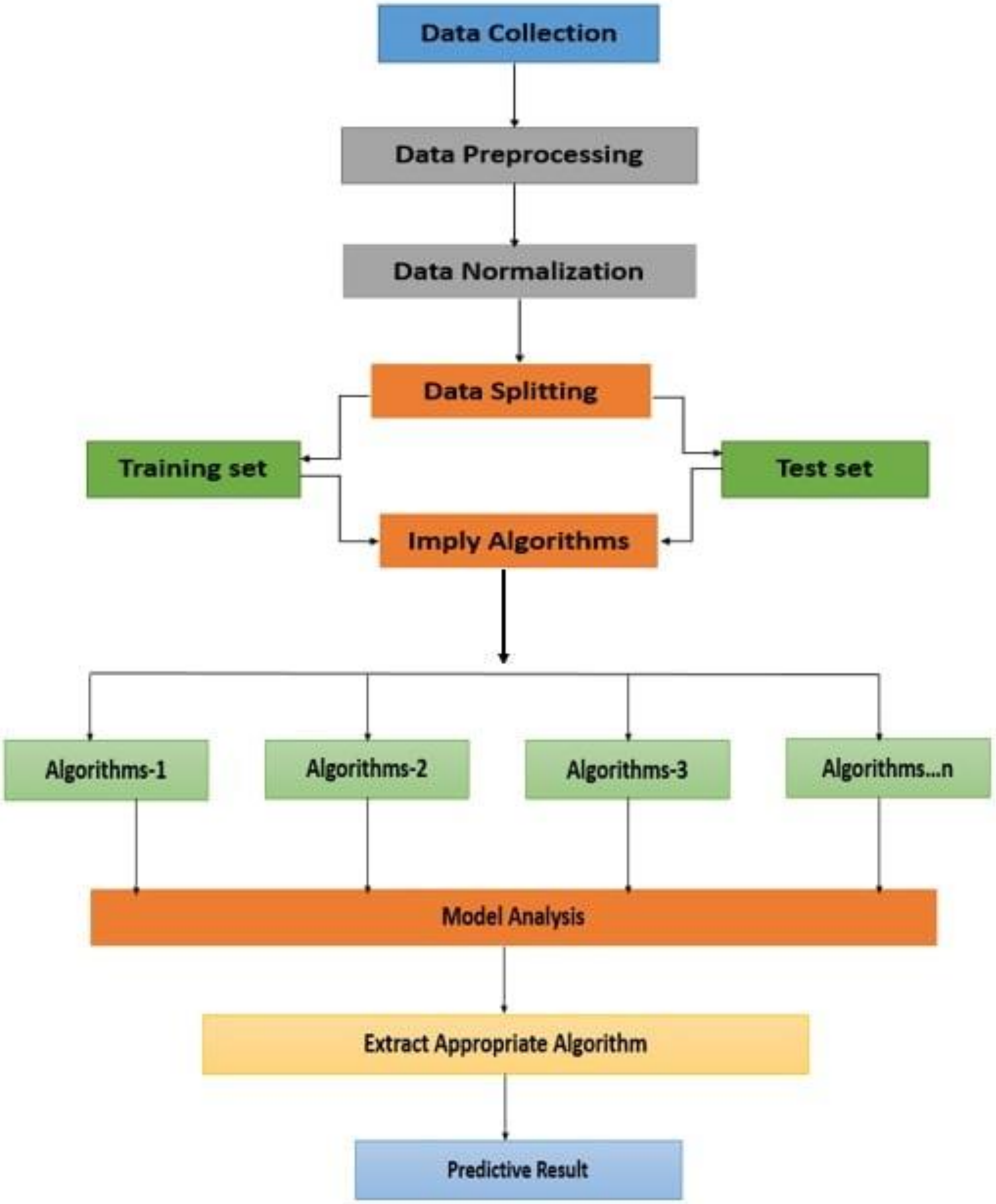

Figure 3.1: Proposed Method to Predict diabetes Disease

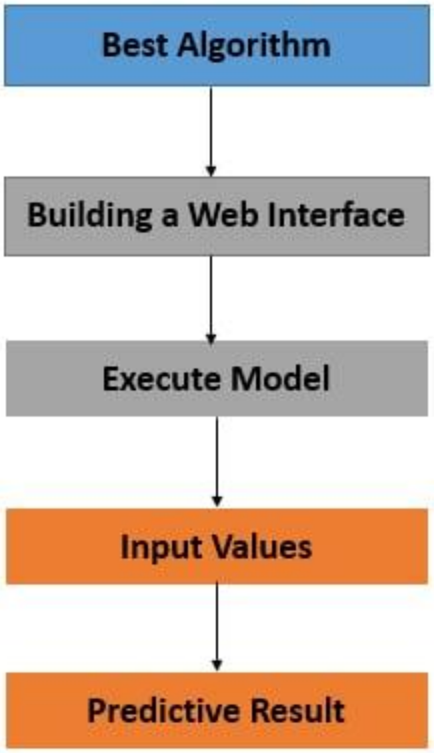

Figure 3.2: Proposed Method to Web Interface

3.6.1 Data Collection

For predicting diabetes disease, need a unique real-life dataset. For this purpose, we collected necessary data from Bangladeshi hospitals. Then, all the data was converted to a single CSV file for analysis and understanding. Following that, all of the data was combined into a single CSV file for analysis and comprehension.

3.6.2 Dataset

For implementing different Machine Learning Algorithms, we need to merge all the data sets into a single CSV file. Machine learning algorithms need lots of data that is not the presence of missing values to predict something. As a result, we gathered a raw and unique dataset.

3.6.3 Data Pre-processing

If the dataset contains lots of missing values, qualitative data must be converted. First of all, the qualitative data were converted into quantitative data. Then there's the issue of filling in the blanks. All the missing values were replaced with the mean value. Fortunately, our dataset has no presence of missing value.

3.6.4 Data Normalization

Normalization is transforming numeric columns to a standard scale while keeping the value ranges intact. For improved accuracy, the independent variable (X) was then standardized.

3.6.5 Data Splitting

The data set must be separated into two parts: train and test before any machine learning method can be used. The model was tested using 25% of the data and trained with the other 75%. The particular model may be trained to predict anything using the Training half of the dataset and the Testing part can be used to see how accurately the data is being predicted.

3.6.6 Imply Algorithms

Selecting algorithms is most important for acquiring the best accuracy. So for importing algorithms, we choose the best 5 algorithms. The names of the five algorithms are Logistic Regression, Decision Tree(DT), Random forests, K-nearest Neighbors (KNN), Gaussian Naive Bayes.

3.6.7 Model Analysis

The data was turned into tables after measuring the Confusion Matrix, Accuracy Score, Jaccard Score, Cross Validated Score, AUC Score, Misclassification, Mean Absolute Error, and Mean Squared Error for all of the algorithms. The confusion matrix summarizes the accuracy with which the data is anticipated. The accuracy score, also known as the Jaccard Score, Cross Validated Score, or AUC Score, is a proportion of the predicted data's accuracy. The algorithms' error rate is determined by misclassification, mean absolute error, and mean squared error.

3.6.8 Extract Appropriate Algorithm

By measuring and observing all of the essential findings from the tables, the best method was discovered. In the dataset, the extracted method has the highest accuracy and lowest error rates. First and foremost, an appropriate algorithm must be designed to make effective use of the dataset. In this instance, it's better to use a variety of algorithms as models and then pick the best one. Various analytical criteria, such as accuracy score, Jaccard score, cross-validation score, AUC score, and so on, were utilized to determine the most effective method in this study K-nearest Neighbors (KNN), Logistic Regression, Random forest is the best method that is appropriate for the diabetes dataset in this study. they received the highest marks in all of the criteria listed above.

3.6.9 Creating Model for Web Interface

The authors created a web-based interface after extracting the best algorithm. A "pickle" was used to connect the interface to the best algorithm. Here Using the Pickle module to serialize items in Python is a simple task. Machine learning algorithms may be serialized and stored in a file using the pickling process. In this situation, the writers serialized the chosen model into a ".sav" file or model. To make a model, you'll need an object that implements that algorithm. The trained dataset is then utilized to train the model using the models.fit() method. The model is ready to use after being trained with the proper algorithm. The model is saved to a file and loaded as a model in the previous phase. The preceding step saves the model to a file and then loads it as a new object named pickled_model. The loaded model is then used to compute the accuracy score and make predictions on previously unknown (test) data.

3.5.10 Building a Web Interface

The authors used the "flask" module in Python to create a web interface. To design a user-friendly interface, basic HTML and CSS were also required. The pickle file was linked to the website in the website's backend.

3.6.11 Execute Model

The model needs to be saved in a folder when created. This is where the pickle comes in. Pickle is a Python module that allows users to decode and serialize Python object structures. As a result, Python objects are converted to byte streams, which may be stored in files or databases, used to preserve program state between sessions, and even transmit data over the internet. Then the model is needed to be dumped in python using pickle. For that pickle. dump() is used. Inside the function, the model's name is denoted, and then a file is opened in binary write mode, and a location is given. This is how the model is created and spread in a folder.

3.6.12 Input Values

Flask is a Python programming interface that helps to make an interface. When the model is created, with the help of flash, we create an interface to predict diabetes diseases. In addition, flash is a Django framework that is easy to code to build a simple web application.

3.6.13 Predictive Result

The user interface makes it easy for anyone to enter data onto the website in order to predict diabetic illnesses. After the user has successfully entered blood glucose, systolic blood pressure (mm/hg), diastolic blood pressure (mm/hg), Insulin, BMI, diabetes pedigree function, and age, the website will display the results. The website will provide two results, one of which will be favourable and the other will be negative.

3.7 System Architecture

To improve the real formation in machine learning technique and web implementation, a system architecture formation is clearly needed to indicate the full project system. A basic system design is shown in Figure 3.2, which is a wider representation of the proposed system.

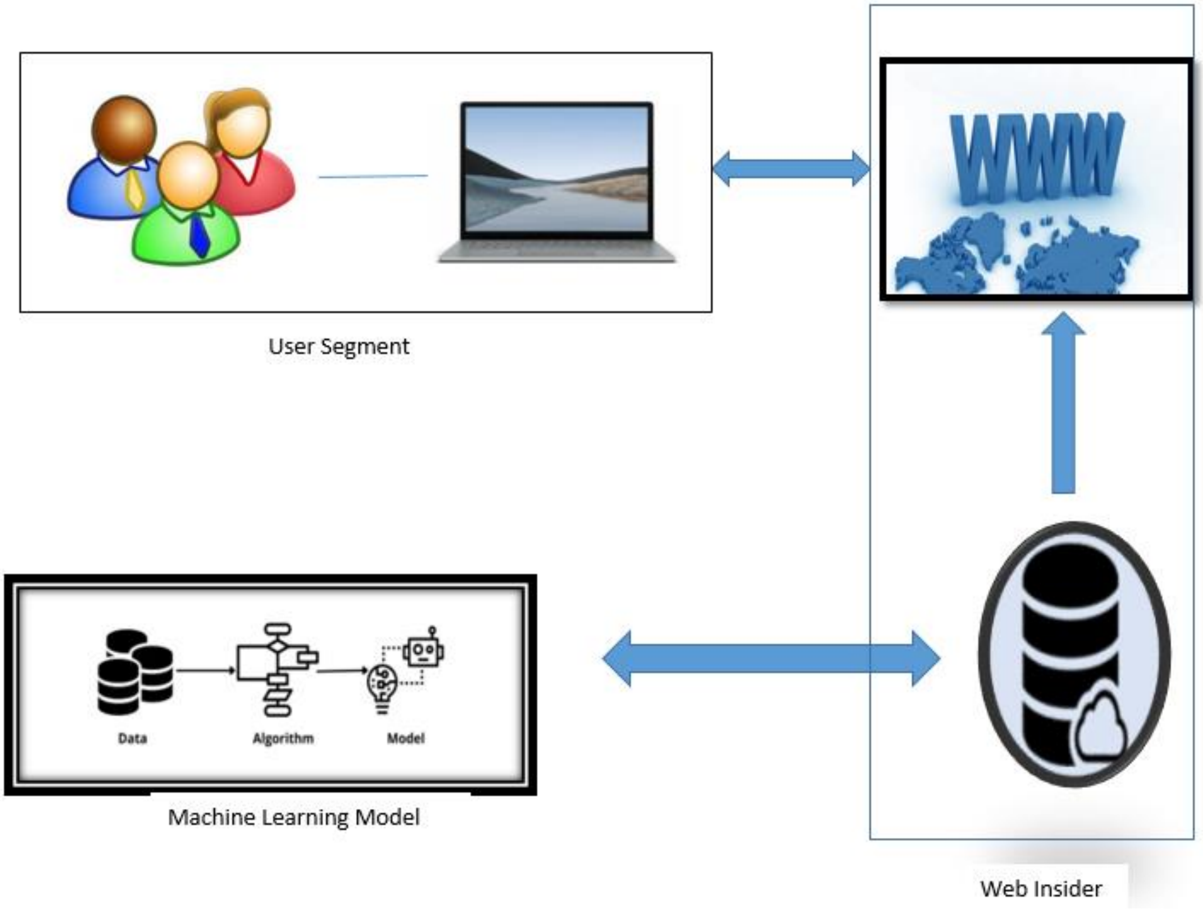

Figure 3.3: System Architecture

3.7.1 User Segment

The web interface shows a user whether diabetes disease or not. Here a user will enter the attribute values of the diagnosis report using a device like a laptop. The user segment has been designed very simply. The interface is very user-friendly so that anyone can use the user segment. Just entering the value and pressing the submitted value is needed, and the machine will predict the current condition.

3.7.2 Web Insider

The authors used the "flask" module in Python to create a web interface. To design a user-friendly interface, basic HTML and CSS were also required. The pickle file was linked to the website in the website's backend. The user interface makes it easy for anyone to enter the essential information onto the website in order to predict diabetes disease. The website will be positive if a patient has the diabetic disease. If the patient is likely to have the expected outcome, the website will show a negative outcome.

3.7.3 Machine Learning Model

By measuring and observing all of the necessary findings from the tables, the best method was discovered. In the dataset, the extracted method has the highest accuracy and lowest error rates. It was important to create a web-based interface after extracting the best algorithm. A "pickle" was used to connect the interface to the best algorithm. Here serializing objects in python is accomplished by using the Pickle library. The pickling technique can serialize machine learning algorithms and store them in a file. In this case, the selected model was serialized into a “.sav” file or model.

CHAPTER 4
EXPERIMENTAL RESULTS & DISCUSSION

4.1 Introduction

The results are important for any type of research. All results at this stage are shown in the table. This chapter details the importance of data acquisition, data usage, and features for understanding diabetics datasets. The Confusion Matrix table is used to show the results of different algorithms. Accuracy, Jacquard score, cross-verified score, AUC score, ROC curve are shown in different graphs. One more table has been created to know the details better. Lastly, the erroneous data is sorted into a table. The project has a web implementation, so the output display is given in this chapter as a screenshot.

4.2 Experimental Results

After using the machine learning code, each technique proved its own accuracy, which is crucial for predicting chronic diabetic patients. Examine each possible score for algorithmic application and technique in the exam outcomes analytical section.

4.2.1 Data Acquisition

The data for the set was gathered from a group of Dhaka Medical College students who have been lobbying for diabetics at various occasions. For the model, about 250 data samples were employed. This diabetes dataset has seven significant factors, including seven numerical variables and a goal variable in each sample (class). A large number of present values are found in the data. The dataset is summarized in Table 4.2.

Table- 4.2: Data Acquisition & Null Percentage

Attribute	Scale	Data Type	Missing Value
Blood Glucose	mg / dl	Numerical	0
Blood Pressure	mm / hg	Numerical	0
Insulin	IU / mL	Numerical	0
BMI	mm	Numerical	0
Diabetics pedigree	%	Numerical	0

Age	Age in years	Numerical	0
Outcome		Numerical	0

4.2.2 Data Utilization

By coding each subject independently, it was made easier to handle the data of a computer system. Where normal and abnormal RBC and pc levels are given as 1 and 0, respectively. As a consequence, yes/no responses were classified as 1 and 0, with 1 indicating yes and 0 indicating no. To compute the appetite value, the good appetite value was assigned to 1 and the poor appetite value was set to 0. Each sample was separated by a random number between 1 and 250. There was not a single data sheet that was empty. For a better understanding of the dataset, data details for seven characteristics were extracted. Calculations, averages, standard deviations, minimums, 25%, 50%, and 75% are retrieved from the maximum dataset and are shown in Table 4.3.

Table 4.3: Dataset Description

	Age	Blood Glucose	Systolic Blood Pressure	Diastolic Blood Pressure	Insulin	BMI	Diabetes Pedigree Function	Diabetes
count	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000
mean	38.752000	178.480000	131.684000	87.568000	45.208000	26.156948	0.949534	0.468000
std	19.618103	55.321782	21.501474	11.098924	63.055123	12.401522	1.359631	0.499976

min	15.000000	120.000000	100.000000	50.000000	0.000000	0.000000	0.000000	0.000000
25%	21.000000	130.000000	120.000000	80.000000	0.000000	20.110000	0.000000	0.000000
50%	30.000000	140.000000	125.000000	90.000000	0.000000	21.735000	0.252500	0.000000
75%	57.000000	240.000000	140.000000	95.000000	120.000000	38.200000	1.951000	1.000000
Max	77.000000	280.000000	180.000000	120.000000	200.000000	50.920000	15.000000	1.000000

4.3 Result & Discussion

In this study, diabetes individuals had positive values, whereas non-diabetics had negative values. The confusion matrix has been used to demonstrate particular outcomes and assess the performance of machine learning algorithms. Table 4 displays the confusion matrix template for several algorithm types.

4.3.1 Confusion Matrix

The confusion matrix is critical for confirming the real dependent outcomes. The Confusion matrix is a N x N matrix that analyzes a classification model's performance. The confusion matrix assesses the machine learning model's accuracy. As a result, an algorithm builds a model that highlights any errors. It will be easy to determine accuracy, memory, and accuracy with the use of binary equations. It is critical to understand the four basic components that are utilized to compute various assessment systems. True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The confusion matrix for each method is shown in Tables 4.4 and 4.5.

Table 4.4: Confusion Matrix

Confusion Matrix		
Predicted Class	**True Positive (TP)**	**False Positive (FP)**
	False Negative (FN)	**True Negative (TN)**

Table 4.5: Confusion Matrix for Algorithms

Algorithm	**Confusion Matrix**			**Confusion Matrix Percentage**	
		DT	**NonDT**	**DT**	**NonDT**
KNN	**DT**	**25**	**0**	**40%**	**0%**
	NonDT	**0**	**38**	**0%**	**60%**
Logistic Regression	**DT**	**25**	**0**	**40%**	**0%**
	NonDT	**0**	**38**	**0%**	**60%**
Random Forest	**DT**	**25**	**0**	**40%**	**0%**

	NonDT	**0**	**38**	**0%**	**60%**
Naive Bayes	**DT**	**25**	**0**	**40%**	**0%**
	NonDT	**2**	**36**	**3%**	**57%**
Decision Tree	**DT**	**24**	**1**	**38%**	**2%**
	NonDT	**3**	**35**	**5%**	**35%**

A. True Positive (TP)

By classification a little positive is good and the percentage value is calculated. Here in KNN, Logistic Regression, Random Forest, Naive Bayes,40% of the value were True Positive (TP) values. Followed by Decision Tree with 38%.

B. True Negative (TN)

Negative tuples are positive tuples that the classifier mistakenly classified. Here in KNN, Logistic Regression, Random Forest, Naive Bayes, 0% of the value were True Negative (TN) values. Followed by Naive Bayes and Decision Tree with 3% and 5%.

C. False Positive (FP)

The classifier mistakenly labeled negative tipples as positive. FP can be used to represent such a connection. Here in KNN, Logistic Regression, Random Forest, 0% of the value were True Negative (TN) values. Followed by a Decision Tree with 2%.

D. False Negative (FN)

The classifier misidentified these positive tuples as negative. It is represented by the letter FN..Here in KNN, Logistic Regression, Random Forest, 60% of the value were True Negative (TN) values. Followed by Naive Bayes and Decision Trees with 57% and 35%.

E. Precision

Precision is a measurement of how much detail is provided. Precision in numbers, on the other hand, is defined as the total number of significant decimal or other digits. Precision equations are measured mathematically using the formula -

$$Precision = \frac{TP}{TP+FP} \dots \dots \dots \dots \dots \dots (vii)$$

F. Recall

It is a completion metric used in machine learning. The ratio of relevant instances is defined as a relevant example. The mathematical formulas used to calculate the Recall equation are as follows:

$$Recall = Sensitivity \ = \frac{TP}{TP + FN} \dots \dots \dots \dots \dots \dots (viii)$$

G. F1-Measure

The weighted harmonic mean is used to evaluate the accuracy and recall of a test, and this value is known as the F measure. The mathematical formulas for measuring the F1-Measure equation are as follows:-

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall} \dots \dots \dots \dots \dots \dots (ix)$$

H. Accuracy

The percentage of test set tuples correctly classified by a classifier on a given test set is known as its accuracy. Accuracy equation mathematical formula -

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

4.3.2 Classification Report

Machine learning uses the report as a statistic to measure system performance. Model accuracy, withdrawal, F1 score, and support are all demonstrated using it. Model precision, withdrawal, F1 score, and support are all displayed. The accuracy, memory, and F1-score and accuracy percentage figures are shown in Table 4.6 for each algorithm.

Table 4.6: Classification Report

Algorithm	Class	Precision	Recall	F1-Score	Accuracy (%)
KNN	**DT**	**1**	**1**	**1**	**100**
	NonDT	**1**	**1**	**1**	
	Macro Avg.	**1**	**1**	**1**	
	Weighted Avg.	**1**	**1**	**1**	
	DT	**1**	**1**	**1**	**100**

Logistic Regression	**NonDT**	**1**	**1**	**1**	
	Macro Avg.	**1**	**1**	**1**	
	Weighted Avg.	**1**	**1**	**1**	
Random Forest	**DT**	**1**	**1**	**1**	**100**
	NonDT	**1**	**1**	**1**	
	Macro Avg.	**1**	**1**	**1**	
	Weighted Avg.	**1**	**1**	**1**	
Naïve Bayes	**DT**	**0.93**	**1**	**0.96**	**97**
	NonDT	**1**	**0.95**	**0.97**	
	Macro Avg.	**0.96**	**0.97**	**0.97**	
	Weighted Avg.	**0.97**	**0.97**	**0.97**	
Decision Tree	**DT**	**0.89**	**0.96**	**0.92**	**94**
	NonDT	**0.97**	**0.92**	**0.95**	
	Macro Avg.	**0.93**	**0.94**	**0.93**	
	Weighted Avg.	**0.94**	**0.94**	**0.94**	

4.4 Result Analysis

All possible aspects are calculated (accuracy, withdrawal, F1-measurement, accuracy) after going through the analysis of the result. Find out which of these algorithms performed best and which corner performed less than the others.

4.4.1 Accuracy

Algorithms' most useful feature is their capacity to create precise and accurate measurements. Data is used to determine how well it will operate. The KNN algorithm, Logistic Regression, and Random Forest algorithms are the most trustworthy algorithms in this algorithm, while the decision tree method is the least accurate. The KNN method, Logistic Regression, and Random Forest Algorithm demonstrate that gradient machines are capable of being implemented in a powerful and scalable manner. For boosted tree algorithms, we're pushing the limits of computer power. It was created with the primary purpose of improving model performance and increasing the speed with which computers can analyze data. Figure 4.1 and Table 4.7 show the accuracy chart and percentage of each algorithm used to make predictions in this model. The accuracy chart and percentage of each algorithm used to create predictions in this model are shown in Figure 4.1 and Table 4.7.

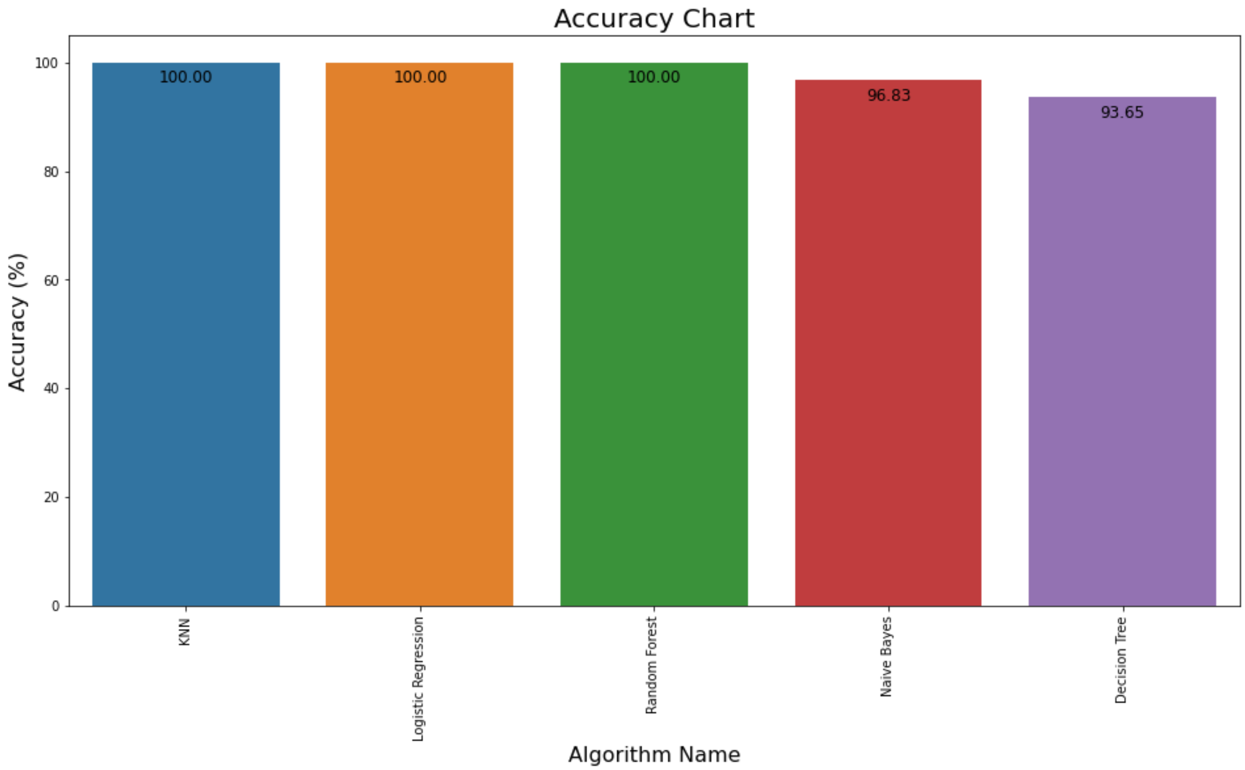

Figure 4.1: Accuracy Chart

4.4.2 Jaccard Score

The jacquard score is calculated by matching the similarities and variations of the sample set. In terms of the ratio of Intersection to Union, they are equivalent. There is a way to compare the similarity of two finite sample sets mathematically by using the Jaccard coefficient, which is defined as the intersection size divided by the union size. Equation (xiv), Figure 4.2 and Table 4.7 shows the accuracy chart and percentage of each algorithm used to predict in this model.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \ldots \ldots \ldots (xi)$$

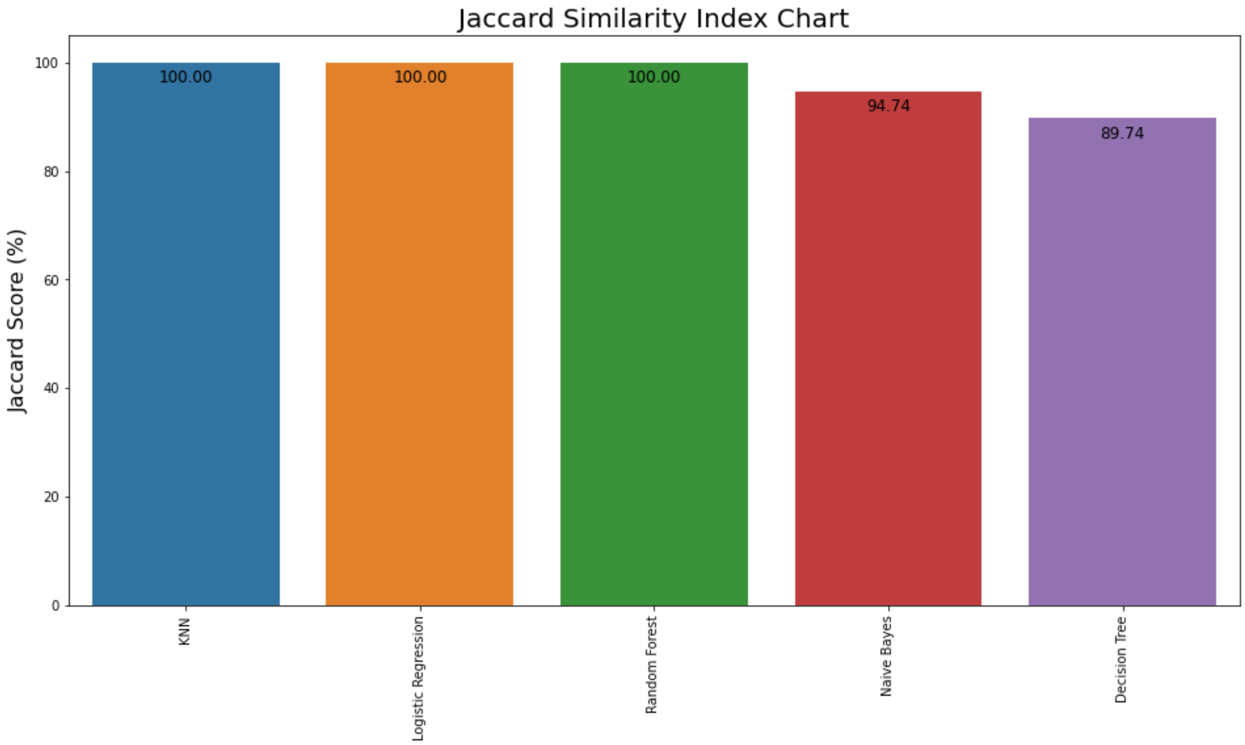

Figure 4.2: Jaccard Score chart

4.4.3 Cross Validated Score

The statistical approach of cross-validation is used to assess the skill of machine learning models. Cross-validation begins by shuffling the data and splitting it into k folds. Then, k models are fitted to (k-1)/k of the data, and 1/k of the data is assessed. The final score is generated by averaging the outcomes of each evaluation, and the resultant model is then fitted to the whole dataset for implementation. Figure 4.3 and Table 4.7 show the cross-validated score chart and proportion of each method used to forecast in the model.

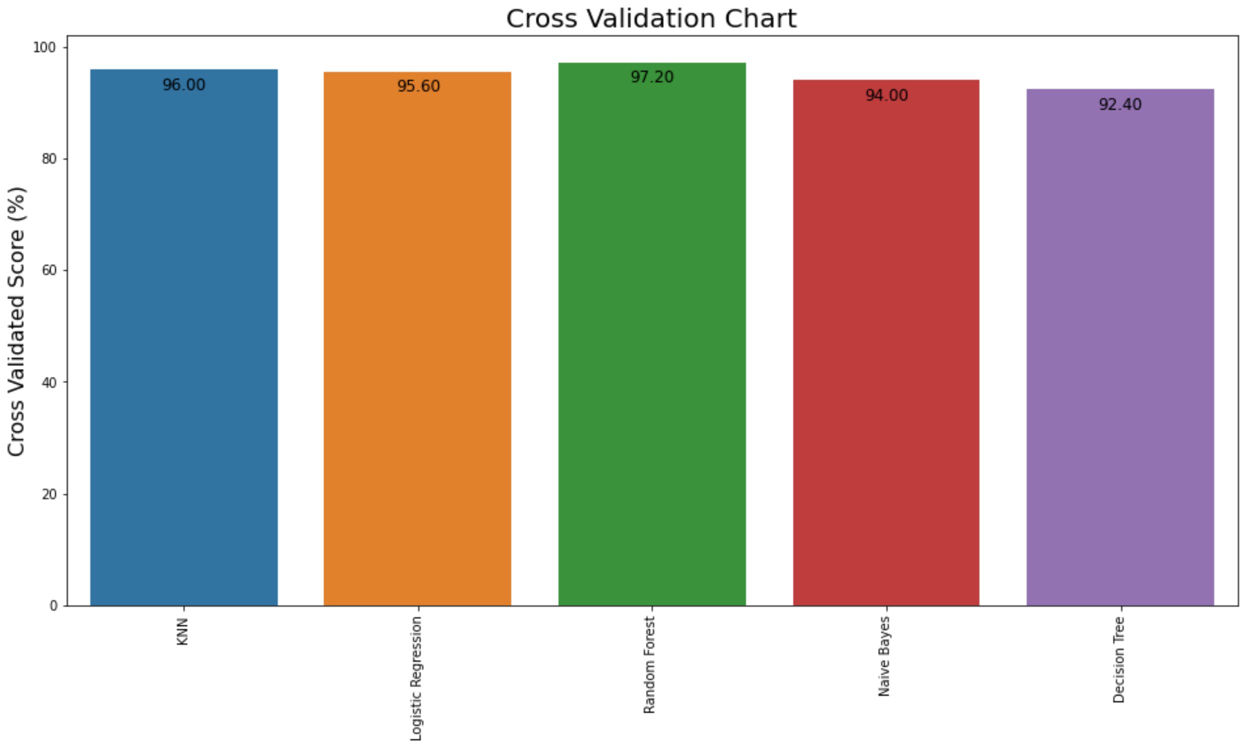

Figure 4.3: Cross Validation Score

4.4.4 AUC Score

A system's viewpoint categorization levels may be computed using this performance metric in conjunction with machine learning applications, which can be valuable in a variety of contexts. The AUC may be calculated by comparing a percentile of random positive situations in which the model performs significantly better to a percentile of random negative cases in which the model performs significantly worse. This number might have four different values, with one being the biggest possible value. The values range from 0 to 1, with 0 representing the lowest possible value. As shown in Figure 4.4 and Table 4.8, models that make 100% erroneous predictions have an accuracy of zero, whereas models that make 100% right predictions have an accuracy of one.

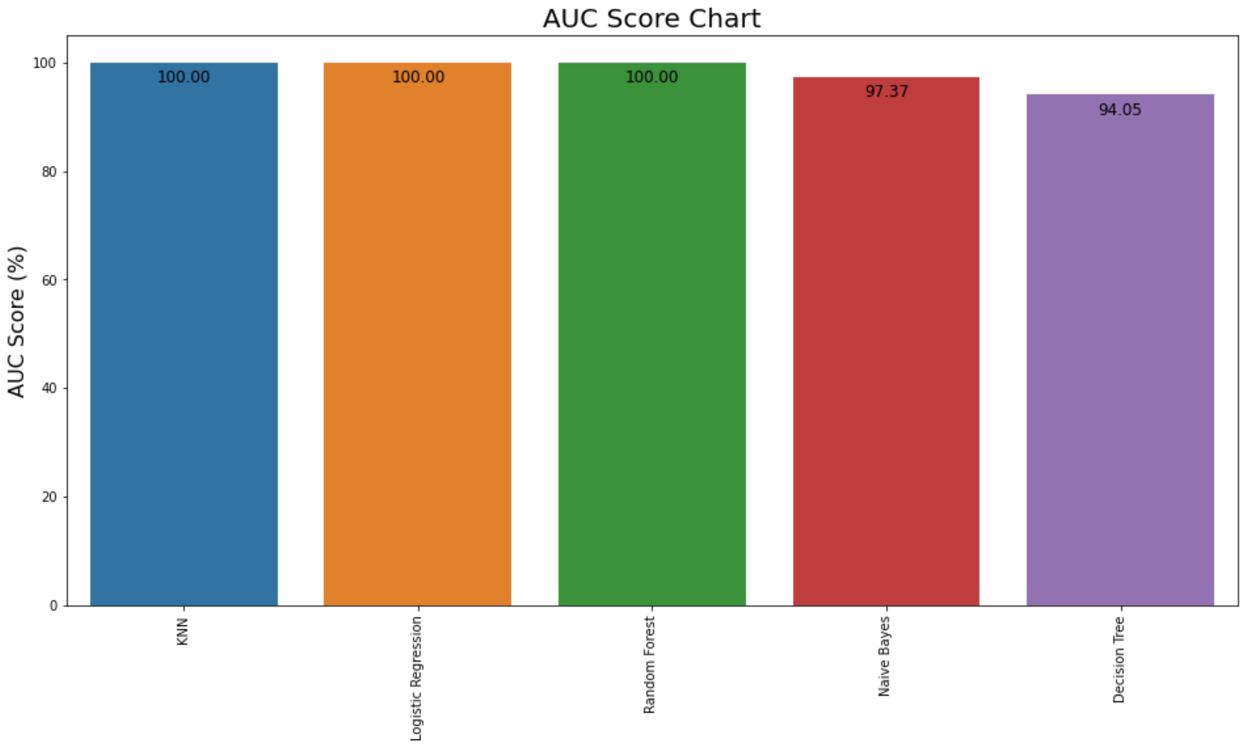

Figure 4.4:AUC Score chart

4.4.5 ROC Curve

ROC analysis is a critical approach for assessing diagnostic test performance and, more broadly, the accuracy of a statistical model that categorizes people into one of two groups: sick or non-diseased. One of the most useful applications of ROC curve analysis is as a simple graphical tool for demonstrating the accuracy of a medical diagnostic test. This crucial curve score is depicted in Figure 4.5.

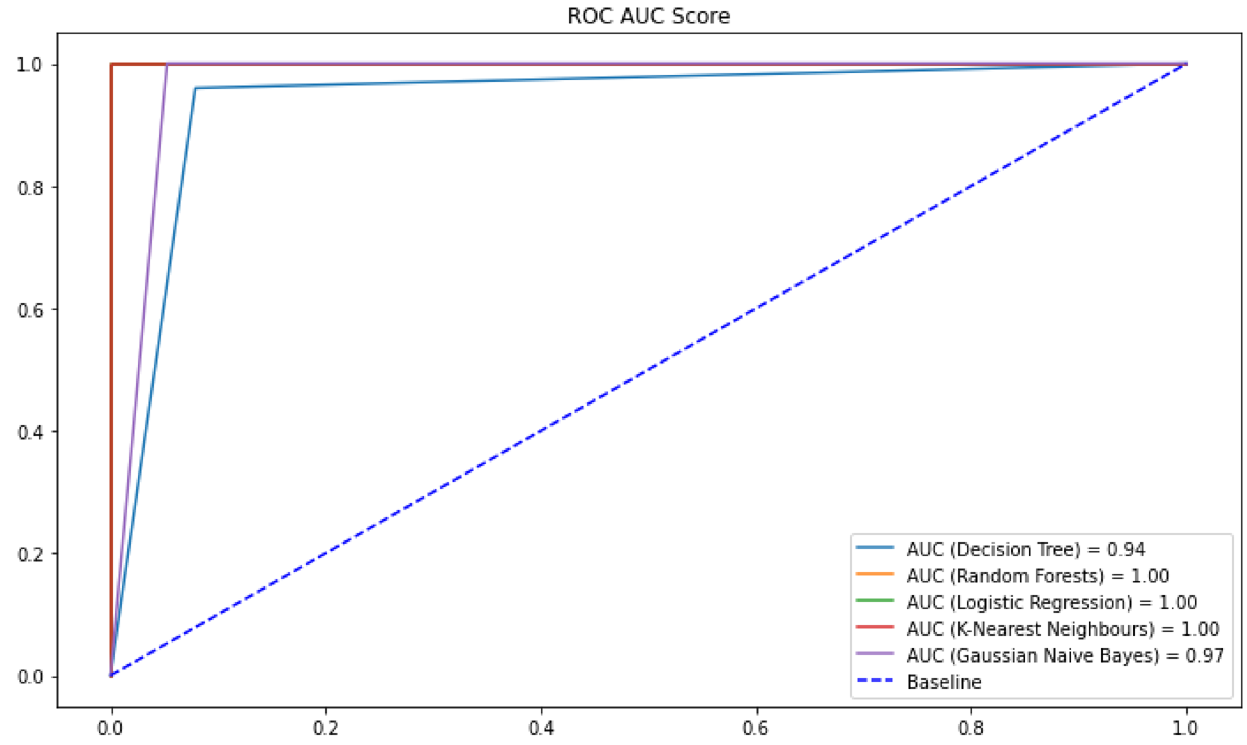

Figure 4.5: Roc Curve

After finishing the whole study, the model with the highest accuracy was Random Forest, which had an accuracy score of 100 percent, a Jaccard score of 100 percent, a Cross Validated score of 97.20 percent, and an AUC value of 100 percent. Table4.7 summarizes the accuracy.

Table 4.7:Accuracy, Jaccard, Cross Validated and AUC Score

Algorithm Name	Accuracy Score (%)	Jaccard Score (%)	Cross Validated Score (%)	AUC Score (%)
KNN	**100**	**100**	**96**	**100**
Logistic Regression	**100**	**100**	**95.60**	**100**
Random Forest	**100**	**100**	**97.20**	**100**
Naive Bayes	**96.83**	**94.74**	**94**	**97.37**
Decision tree	**93.65**	**89.74**	**92.40**	**94.05**

4.4.6 Misclassification & Error

Errors come to the fore when verifying the accuracy of an algorithm. After classification errors, complete quality errors and average square errors can be solved by a machine learning model. If there is a classification error, the whole data site becomes an incorrect data set so that the results of the algorithm are not correct. All classes, groups, or classes of a variable have the same error rate if incorrect classification occurs.

Average absolute error (MAE) is the average of all absolute errors of measurement. The formula of mean absolute error represents as

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |x_i - x| \dots \dots \dots (xii)$$

Average absolute error (MAE) is the average of all absolute errors of measurement. The average presents as the source of absolute error. The formula of Mean Squared Error represents as

$$MSE = \frac{1}{n}\sum_{i=1}^{n} |y_i - y|^2 \dots \dots \dots (xiii)$$

Table 4.8 shows the misclassification, mean absolute error and mean square error in the algorithms. The Error rate for the KNN, Logistic Regression, Random Forest, was also less for Misclassification with 0%, Mean Absolute Error with 0% and Mean Squared Error 0%.

Table 4.8: Misclassifications & Errors

Algorithm Name	Misclassification (%)	Mean Absolute Error (%)	Mean Squared Error (%)
KNN	0.00	0.00	0.00
Logistic Regression	0.00	0.00	0.00
Random Forest	0.00	0.00	0.00
Naïve Bayes	3.17	3.17	3.17
Decision Tree	6.35	6.35	6.35

4.5 Web Implementation

The development of machine learning models has been carried out precisely and successfully. Appropriate algorithms have been developed and implemented. It is now time to demonstrate how this model is implemented in the web interface described in the Architecture section.

4.5.1 Web Interface

Chronic diabetes Disease will be predicted using the Flask approach, as explained in Chapter 3. As previously stated in Chapter 3, a web interface based on the "Flask" concept will be built. A very effective and functional website has been established for the purpose of ensuring that this job is accomplished effectively. That Web Interface is shown in the Figure 4.5.

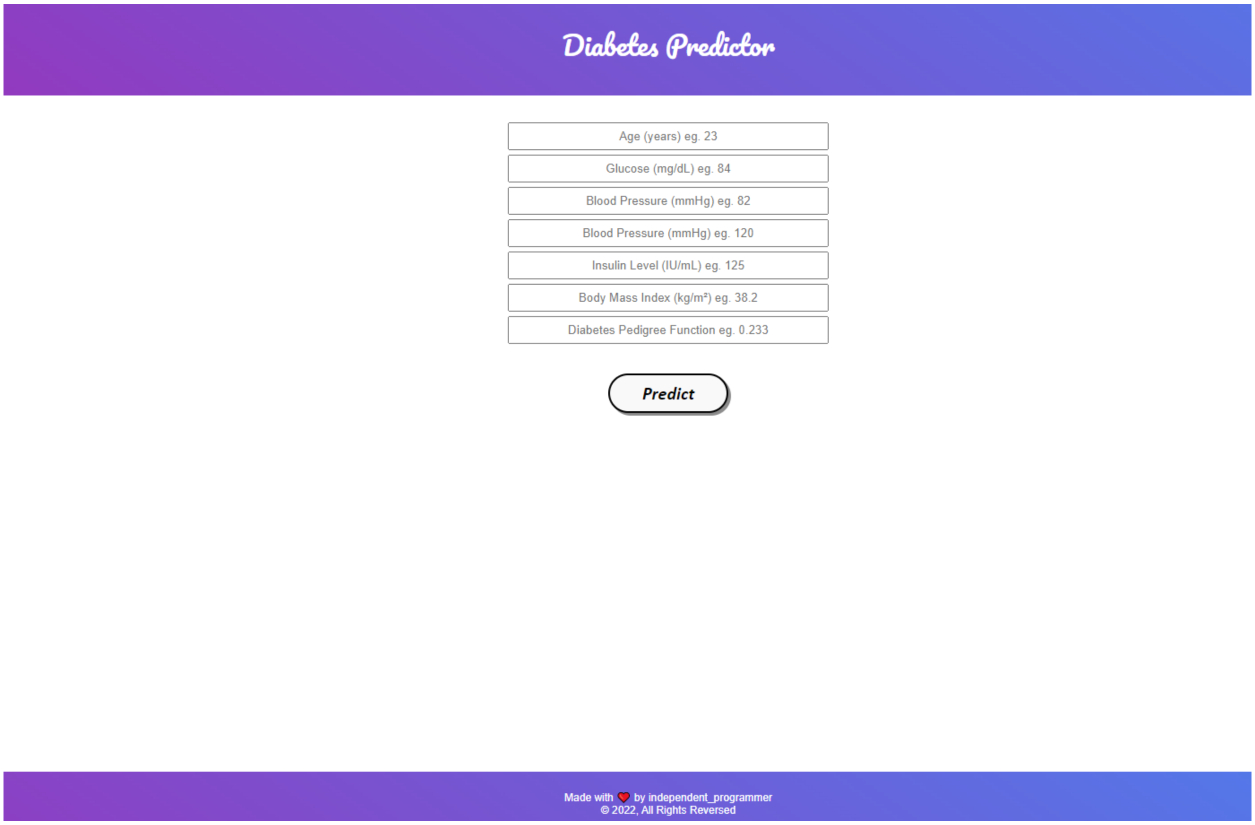
Diabetes Predictor
Age (years) eg. 23
Glucose (mg/dL) eg. 84
Blood Pressure (mmHg) eg. 82
Blood Pressure (mmHg) eg. 120
Insulin Level (IU/mL) eg. 125
Body Mass Index (kg/m²) eg. 38.2
Diabetes Pedigree Function eg. 0.233
Predict
Made with ❤ by independent_programmer
© 2022, All Rights Reversed

Figure 4.5: Web Interface

CHAPTER 5
IMPACT ON SOCIETY & SUSTAINABILITY

5.1 Introduction

If a project has been developed, the influence of such a project on society ought to be examined and studied. Here in this chapter, the influence of the diabetes Disease project has been described in three sections. The impact on society is explained how this project will affect society in a good manner. Then comes ethical considerations. Here the ethical element was covered well for understanding how this initiative can aid the patients. And at the end, the project sustainability was examined. How this initiative might develop in the future and aid more individuals is addressed.

5.2 Impact on Society

This study has a substantial societal effect. Because people today all have busy lives, travelling to the hospital to guarantee that individuals are not infected with any ailment takes a long time. An interface has been designed where individuals may input the data they need to forecast diabetes Disease and obtain a rapid answer. It is unsafe to go to the hospital for examinations today as individuals all over the globe are at risk for COVID-19. It would be pretty beneficial for us if anybody could analyze their report at home. A survey on patients and physicians is continuing concerning this study activity and its usefulness. And it is pretty clear that the survey report will indicate a favourable consequence on the influence on society and patients-doctors as well.

5.3 Ethical Aspects

There is no need for someone to travel to a diagnostic hospital for any kind of test. They can get a basic understanding of diabetic disease at home. They have the ability to foresee on their own. Because a web interface has been developed, all data will be stored in the database, and the model will become more important over time. Because this is a machine learning-based project, it is now unwise to depend only on this model. Nothing can be predicted perfectly by a computer. It will take some time. When the database has a million records, the model will be more robust, and it is anticipated that it will be able to predict accurately over time. It will be unnecessary to visit a diagnostic hospital in the future if Artificial Intelligence and the Internet of Things can be integrated with a database. People will be able to predict diabetes disease at home before seeing a doctor in their area.

5.4 Sustainability

Because of its innovative strategy, which allows people to use a website to analyze their diabetic disease, this research has high long-term viability. The website now relies primarily on machine learning algorithms to predict diabetes Disease. However, in the future, there are numerous possibilities with deep learning, artificial intelligence, and the Internet of Things that might deliver more accurate results for this project. As a result, in terms of sustainability, this project might be used for a variety of future activities. It's even possible to create a mobile app. The only diabetic disease is discussed in this article. However, it makes use of machine learning technologies and a web-based approach. Uploading an image of a disease using a web interface may potentially be recognized in the future.

CHAPTER 6
FUTURE SCOPE & CONCLUSION

6.1 Introduction

At this stage, the future plans of the project are discussed which can be used as a machine learning method. This project will help a company reach the pinnacle of development and increase efficiency. This chapter has a clear and accurate conclusion. Finally, a list of references was published.

6.2 Implication for Further Study

Diabetes patients can be used properly in hospitals as a website foundation. An interface has been developed for this purpose that can accurately and easily identify any diabetic patient. In the future, some websites will be created based on the Internet of Things and will be able to serve the public through the Internet. At the sit on home, the diabetic patient will be able to complete all the necessary information on the website and get tested for diabetes. This will allow a person to know the amount of diabetes. Because it is an Internet of Things-based web that will always provide accurate and precise results. Over time, the model will be able to reach people more efficiently, accurately, easily, and more conveniently. In the future, the project may also include neural network algorithms and artificial intelligence.

6.3 Recommendations

If an abnormal phase is found at a diagnostic center instead of the normal result, the system may be aware of that particular condition. Here are some of the key features of chronic diabetes: a new step for patients with chronic diabetes is frequent urination (polyuria), Frequent thirst (polydipsia), Frequent loss of appetite (polyphagia), Weight loss etc. The normal range of polyuria in women is 0.6-1.1 mg / dL and in men 0.7-1.3 mg / dL. If the amount of protein in the urine increases then kidney problems occur later. The level of protein in normal human urine is 0 to about 8 mg / dL. Some indications are provided in the early stages of chronic diabetes. Some precautions can be taken in the early stages, such as changing exercise habits, drinking more, and following a doctor's advice.

6.4 Conclusion

Chronic diabetes has been diagnosed and treated early. Diabetes is a long-term disease that can not be easily cured. Many laboratory tests can be done to diagnose chronic diabetes. In this project, using the interface, diabetes patients can enter the requirements and get the diabetes report. Five machine learning algorithms are given data set training to find the most suitable model using a data set and are evaluated based on their accuracy, Jacquard score, cross-validated score, and AUC score. Leading the competition in KNN, Linear Regression, Random Forest Performance compared to other divisions. One hundred percent of the results of these three algorithms are available. A complete website for this project will later be made accessible to any hospital. A person does not have to leave their home to receive a diabetic report because they are accessible online and will be available 24 hours a day, seven days a week.

REFERENCES

[1] Hasan and Md Kamrul, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access,* pp. 76516-76531, 2020.

[2] Khanam, Jobeda Jamal and Simon Y. Foo, "A comparison of machine learning algorithms for diabetes prediction.," *ICT Express,* pp. 432-439, 2021.

[3] Sonar, Priyanka and K. JayaMalini, "Diabetes prediction using different machine learning approaches," *Computing Methodologies and Communication (ICCMC),* pp. 367-371, 2019.

[4] Fazakis and Nikos, "Fazakis, Nikos, et al. "Machine learning tools for long-term type 2 diabetes risk prediction.," *IEEE Access,* pp. 103737-103757, 2021.

[5] Hasan and Md Kamrul, "Diabetes prediction using ensembling of different machine learning classifiers," IEEE Access, pp. 76516-76531, 2020.

[6] Khanam, Jobeda Jamal and Simon Y. Foo, "A comparison of machine learning algorithms for diabetes prediction.," ICT Express, pp. 432-439, 2021.

[7] Sonar, Priyanka and K. JayaMalini, "Diabetes prediction using different machine learning approaches," Computing Methodologies and Communication (ICCMC), pp. 367-371, 2019.

[8] Fazakis and Nikos, "Fazakis, Nikos, et al. "Machine learning tools for long-term type 2 diabetes risk prediction.," IEEE Access, pp. 103737-103757, 2021.

[9] A. a. V. V. Mujumdar, "Diabetes prediction using machine learning algorithms," in Procedia Computer Science,, 2019.

[10] P. C. -. I. TN Joshi, "Diabetes prediction using machine learning techniques," *academia.edu,* pp. 9-13, 2018.

[11] A. J. J. R. A Yahyaoui, "A decision support system for diabetes prediction using machine learning and deep learning techniques," in *researchgate.net*, 2019.

[12] H. BadeeaAhmed, "Effects of External Factors in CGM Sensor Glucose Concentration Prediction," *Effects of external factors in CGM sensor glucose concentration prediction,* vol. 102, pp. 623-629, 2016.

[13] E. S. M. A. MAA Rahmat, "IoT-Based Non-invasive Blood Glucose Monitoring," *utem.edu.my,* vol. 9, pp. 71-75, 2017.

[14] O. Daanouni, "Predicting diabetes diseases using mixed data and supervised machine learning algorithms," *acm.org,* pp. 1-6, 2019.

[15] Sisodia, "Prediction of diabetes using classification algorithms," *Procedia computer science,* pp. 1578-1585, 2018.

[16] N Nai-arun and R Moungmai, "Comparison of Classifiers for the Risk of Diabetes Prediction," *Procedia Computer Science,* pp. 132-142, 2015.

[17] Shankaracharya and Devang Odedra, "Computational intelligence in early diabetes diagnosis: a review," *The review of diabetic studies: RDS,* p. 252, 2010.