TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING

Khwopa College Of Engineering
Libali, Bhaktapur
Department of Computer Engineering

A PROPOSAL ON

DIABETES PREDICTION

Submitted in partial fulfillment of the requirements for the degree

BACHELOR OF COMPUTER ENGINEERING

Submitted by

Bikesh Sitikhu	KCE/074/BCT/016
Ram Katwal	KCE/074/BCT/031
Sabin Suwal	KCE/074/BCT/035
Sachit Kumar Shrestha	KCE/074/BCT/036

Under the Supervision of
Er.Dinesh Gothe
Department Of Computer Engineering

Khwopa College Of Engineering
Libali, Bhaktapur
2020-21

TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING

Khwopa College Of Engineering
Libali, Bhaktapur
Department of Computer Engineering

A PROPOSAL ON

DIABETES PREDICTION

Submitted in partial fulfillment of the requirements for the degree

BACHELOR OF COMPUTER ENGINEERING

Submitted by

Bikesh Sitikhu	KCE/074/BCT/016
Ram Katwal	KCE/074/BCT/031
Sabin Suwal	KCE/074/BCT/035
Sachit Kumar Shrestha	KCE/074/BCT/036

Under the Supervision of
Er.Dinesh Gothe
Department Of Computer Engineering

Khwopa College Of Engineering
Libali, Bhaktapur
2020-21

Certificate of Approval

The undersigned certify that the final year project entitled **"Diabetes Prediction"** submitted by Bikesh Sitikhu, Ram Katwal, Sabin Suwal and Sachit Kumar Shrestha to the Department of Computer Engineering in partial fulfillment of requirement for the degree of Bachelor of Engineering in Computer Engineering. The project was carried out under special supervision and within the time frame prescribed by the syllabus.
We found the students to be hardworking, skilled, bona fide and ready to undertake any commercial and industrial work related to their field of study and hence we recommend the award of Bachelor of Computer Engineering degree.

.....
Er. Dinesh Gothe
(Project Supervisor)

.....
Prof. Dr. Sashidhar Ram Joshi
(External Examiner)
Central Campus Pulchowk, Institute of Engineering
Lalitpur, Nepal

.....
Er. Shiva K. Shrestha
Head of Department
Department of Computer Engineering, KhCE

Copyright

Acknowledgement

We take this opportunity to express our deepest and sincere gratitude to our supervisor Er. Dinesh Man Gothe, for his insightful advice, motivating suggestions, invaluable guidance, help and support in successful completion of this project and also for his constant encouragement and advice throughout our Bachelor's program. We are deeply indebted to our teacher Er.Bindu Bhandari for boosting our efforts and morale by their valuable advices and suggestion regarding the project and supporting us in tackling various difficulties.
In Addition, we also want to express our gratitude towards Er. Dinesh Man Gothe for providing the most important advice and giving realization of the practical scenario of the project.

Bikesh Sitikhu	KCE/074/BCT/016
Ram Katwal	KCE/074/BCT/031
Sabin Suwal	KCE/074/BCT/035
Sachit Kumar Shrestha	KCE/074/BCT/036

Abstract

Diabetes Prediction System is the web based system to be used in medical field. This idea is inspired because there are lack of awareness about diabetes disease among people at this world.Diabetes can cause many worse diseases such as heart failure, nerve damage, eyes problem and another organ failure. Next, the early diagnosis can prevent the disease become more worse. This system are build to do early diagnosis. In this system, doctors as user can predict their patients condition whether they will having the diabetes disease or not based on their records. Next, the another user can also access this system to get early diagnosis. Then, to if they want get more accurate result, they can refer to specialists. So, this system main modules are consist of user and administrator. The user will provide the details about their health condition and personal information to get the results.Then, the questionnaires is purposed and answers can be collected .Next, the system will provide early diagnosis result after do the calculation and generate result based on the input. The system will use rule based algorithms. Many algorithms like Decision Tree, KNN, Random Forest can be used for powering prediction the disease. Django will develop into web where there is system will predict the diabetes. Also,the tips and information about Diabetes sections also available to be viewed by users. Through that, user able gain knowledge about Diabetes.

Keywords: *Decision Tree Algorithm, K-Nearest Neighbors, Machine Learning, Random Forest*

Contents

List of Tables

List of Figures

List of Symbols and Abbreviation

SVM	Support Vector Machine
JS	JavaScript
API	Application Programming Interface
KNN	K-Nearest Neighbours
OOP	Object Oriented Programming

Chapter 1

Introduction

1.1 Background Introduction

Diabetes is a chronic, metabolic disease characterized by elevated levels of blood glucose or blood sugar, which leads over time to serious damage to the heart, blood vessels, eyes, kidneys and nerves. The most common is type 2 diabetes, usually in adults, which occurs when the body becomes resistant to insulin or doesn't make enough insulin. In the past three decades the prevalence of type 2 diabetes has risen dramatically in countries in all income levels. Type 1 diabetes, once known as juvenile diabetes or insulin-dependent diabetes, is a chronic condition in which the pancreas produces little or no insulin by itself. For people living with diabetes, access to affordable treatment, including insulin, is critical to their survival. There is a globally agreed target to halt the rise in diabetes and obesity by 2025.
About 422 million people worldwide have diabetes, the majority living in low and middle income countries, and 1.6 million deaths are directly attributed to diabetes each year. Both the number of cases and the prevalence of diabetes have been steadily increasing over the past few decades. People for a long time suffered from different diseases that in some cases have been able to diagnose diseases and offer them the solution in order to enhance it, but unfortunately, sometimes, due to lack of diagnosis of symptoms in patients for a long time may even threaten the life of the patient. Therefore, many studies have been done in the field of predicting for several diseases toe the extent that today's human take advantage of decision supports models and smart method to predict. One of the decision support models application is in the medical field and diagnosis of illnesses such as diabetes. Deferment in the diagnosis and prediction of diabetes due to insufficient control of blood glucose increases macro vascular and capillaries difficulties risk, ocular diseases and kidney failure.
So we proposed different models like Random Forest, Naive Bayes, KNN and Decision Tree to predict diabetes that can be useful and helpful for doctors and practitioners. In this research, we used the following attributes: Number of pregnancies, PG Concentration (Plasma glucose at 2 hours in an oral glucose tolerance test), Diastolic BP (Diastolic Blood Pressure (mm Hg)), Tri Fold Thick (Triceps Skin Fold Thickness (mm)), Serum Ins(2 hours Serum Insulin (mu U/ml)), BMI(Body Mass Index: (weight in kg/(height in m)2), DP Function (Diabetes Pedigree Function), Age (years), Diabetes (whether or not the person has diabetes).
Since diabetes is a long-lasting disease and import permanent damage to the limbs

and vital organs in the body, using various classification algorithms of machine learning approaches can enhance the detection methods and disease control which will be of a great help to the physicians. According to the Diabetes Research Center, it has been shown that early diagnosis of patients at risk can prevent 80 percent of lasting complications of type II diabetes or deferred them.

1.2 Problem Statement

Diabetes mellitus is a common disease that affects a vast majority of the people in many parts of the world. Under normal conditions blood glucose concentration is maintained within a narrow range of 72-126mg/dL (Yu and Hui, 2005). The disordered regulation of glucose metabolism that results in diabetes is usually due to a deficiency of insulin release from the pancreas and/or a reduced response to insulin.

The normal identifying process is that patients need to visit a diagnostic center, consult their doctor, and sit tight for a day or more to get their reports. These processes largely reduce the number of patients died of diabetes. There are many factors that influence diabetes and the different aspects of the disease, and quite often it is unfeasible to take all in consideration. The feature selection method addresses this problem by selecting a few factors that are the most influential for each particular case. Sometimes, doctor may not able to get predicted the early stage of diabetes which may result a creator of different kinds of diseases like heart attack, blindness, kidney diseases, etc. However, the incidence of diabetes mellitus is increasing and that although there is evidence that the complications of diabetes can be prevented. Moreover, every time they want to get their diagnosis reports, they have to waste their money in vain.

With the rise of Machine Learning approaches we have the ability to find a solution to this issue, we have developed a system using data mining which has the ability to predict whether the patient has diabetes or not. Data mining is a process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. It is utilized to find new examples, find comparable connections among information, co-relations between information, this can find answers for the issues, creating rules from old information, settling on best choices of ad lib the business arrangements, finding concealed information design from leaving datasets, expectation of future yield, i.e. practices and patterns.

Furthermore, predicting the disease early leads to treating the patients before it becomes critical. Data mining has the ability to extract hidden knowledge from a huge amount of diabetes-related data. Because of that, it has a significant role in diabetes research, now more than ever. The aim of this research is to develop a system which can predict the diabetic risk level of a patient with a higher accuracy. This research has focused on developing a system based on three classification methods namely, Support Vector Machine, Logistic regression and Decision Tree.

1.3 Objective

The main aim of this project is:

1. To develop a system that can predict diabetes with a higher accuracy,
2. To evaluate the data set for the early prediction of the disease,
3. To provide an effective and easy system for all kinds of people,
4. To understand the concept of machine learning and data mining.

Chapter 2

Literature Review

Diabetes or diabetes mellitus is a metabolic disorder (metabolic) in the body. This disease destroy the ability to produce insulin in the patient's body or the body develops resistance to insulin the and consequently the produced insulin cannot achieve its normal job. The main role of the produced insulin is to decrees blood sugar by different instruments. There are two key types of diabetes. In Type I diabetes, obliteration of beta pancreatic cells damage insulin construction and in type II, there is a progressive insulin confrontation in the body and ultimately may yield to the obliteration of pancreatic beta cells and faults in insulin production. In type II diabetes, it is known that genetic issues, obesity and lack of physical activity have a vital part in a person [1].
Even though the precise cause of type I diabetes is unidentified, issues that may indicate a greater risk comprise the followings [2]:

1. Family history. A person risk upsurges if his parent or sibling has history of type I diabetes.

2. Environmental factors. Situations for example contact with a viral illness probably play some role in type I diabetes.

3. The existence of harmful immune system cells. Occasionally family members of a person with type I diabetes are examined for the existence of diabetes autoantibodies. If a person has these autoantibodies, he/she has a chance of increased risk for evolving type I diabetes. Nonetheless not every person who has these autoantibodies gets diabetes.

4. Geography. Some countries, like Sweden, have bigger rates of type I diabetes.

Researchers don't completely comprehend why certain people develop pre-diabetes and type II diabetes and others don't. It's sure that some factors upsurge the risk like [2]

1. Weight. The more fatty tissue you have, the more resilient a person cells to insulin.

2. Inactivity. The less energetic a person is, the more a person has risk. Physical activity assists a person control of his/her weight, consumes glucose as energy and makes a person cells more sensitive to insulin.

3. Family history. A person risk upsurges if his parent or sibling has history of type II diabetes.

4. Race. Even though it's uncertain why, people of specific races are at higher risk.

5. Age. A person risk upsurges as he/she gets older. This may be because a person has a habit to exercise less, lose muscle mass and add weight as he/she gets older. Nonetheless type II diabetes is likewise growing among children, youths and adults.

6. Gestational diabetes. If a person developed gestational diabetes when she was pregnant, her risk of emerging pre-diabetes and type II diabetes far ahead upsurges. If she gives birth to a baby weighing more than 4 kilograms, she is also at risk of type II diabetes.

7. Polycystic ovary syndrome. For females, having polycystic ovary syndrome increases the risk of getting diabetes.

8. High blood pressure. Having blood pressure more than 140/90 millimeters of mercury (mm Hg) is connected to an augmented risk of type II diabetes.

9. Abnormal cholesterol and triglyceride levels. If a person has low levels of high- density lipoprotein, or good cholesterol, his/her risk of type II diabetes is going to be higher. Triglycerides are additional type of fat passed in the blood. A person with greater levels of triglycerides has an augmented risk of type II diabetes.

One of the important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of disease like diabetes. During this work, three machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on Pima Indians Diabetes Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of 76.30 % using the Naive Bayes classification algorithm [3].

Chapter 3

Requirement Analysis

3.1 SOFTWARE REQUIREMENT

Our project Diabetes Prediction System requires Python, Django, Postgresql, TensorFlow, JavaScript, Keras which are described below;

3.1.1 Python

Python is a high-level interpreted programming language for general-purpose programming created by Guido van Rossum which as released in 1991. Python has design philosophy that emphasizes code readability, notably using significant white spaces. It provides constructs that enable clear programming in small and large scales. The programming language features dynamic type system and automatic memory management with support to multiple programming paradigms including Object Oriented Imperative, Functional and Procedural. The programming language has large comprehensive standard library.

3.1.2 Django

Django is a web framework written in python. We use django to create the interface for the diabetes prediction engine that will be created by training he data using different algorithms.

3.1.3 PostgreSQL

PostgreSQL, also known as Postgres, is a free and open-source relational database management system emphasizing extensibility and SQL compliance. It was originally named POSTGRES, referring to its origins as a successor to the Ingres database developed at the University of California, Berkeley. In our project, we use this database to feed the data into django system.

3.1.4 Keras

Keras is an open source neural network library written in Python which is capable of running on top of TensorFlow, Microsoft Cognitive Toolkit, or Theano. Actual model preparation of different models is done using Keras machine learning

framework. We 5use Keras functional API to generate models as well as train the model on different annotated data.

3.1.5 JavaScript

JavaScript often abbreviated as JS, is a high-level, interpreted programming language. It is a language which is also characterized as dynamic, weakly typed, prototype-based and multi-paradigm.

3.2 FUNCTIONAL REQUIREMENT

1. Member should be gives input to different field for prediction of diabetes.
2. Every field only accept numeric data.
3. The website shall have a home page that list the purpose of software.
4. The website home page shall list of all page which we can render.
5. A 'Contact Us' page shall provide the name and contact information of executive and member can comment their messages.
6. A 'About Us ' page shall provide the description about diabetes.
7. Dataset should be used to train different ML algorithms.
8. More than one Dataset can be used to train algorithms.
9. Member should be gives input to different field for prediction of diabetes.
10. Prediction of diabetes should be either True of False.
11. Result is displayed on screen.
12. Different analysis can be performed on database.
13. Only the authorized member should be able to predict diabetes.
14. Member should be able to view their report.
15. Member should be able to update their account.

3.3 NON-FUNCTIONAL REQUIREMENT

These requirements are not needed by the system but are essential for the better performance of sentiment engine. The points below focus on the non-functional requirement of the system.

3.3.1 Operational Requirement

3.3.1.1 Accessibility

Accessible for all kinds of people.

3.3.1.2 Availability

People can easily fill out the form anytime to obtain the results as soon as possible.

3.3.1.3 Confidentiality

Any information provided by the users is protected

3.3.1.4 Efficiency

The results shown by the system are of highly correct.

3.3.1.5 Security

Security best practices shall be used.

3.3.2 Revision Requirement

3.3.2.1 Flexibility

The parameters can be added or deleted or upgraded as per necessity.

3.3.2.2 Modifiable

Pages shall be editable in cornerstone.

3.3.3 Transition Requirement

3.3.3.1 Install ability

Easily available as a websites.

3.3.3.2 Portability

It is available in multiple operating systems as a website.

3.4 FEASIBILITY STUDY

The following points describes the feasibility of the project.

3.4.1 Economic Feasibility

The total expenditure of the project is just computational power. The dataset and computational power required for the project are easily available. Dataset is found from the local news site and computational power using our own laptop alongside Google Colaboratory cloud computing service so, the project is economically feasible.

3.4.2 Technical Feasibility

The project will be trained with lots of labelled data-sets. Preparing the data has its own complexity but the data sets provided by kaggle makes easy available of necessary data. Creating the project will be challenging but is feasible. Training huge dataset takes a lot of computational power.

3.4.3 Operational Feasibility

The project can be operational just after training from the datasets using different Machine Learning algorithms. Once the datasets get trained, the implementation get easy for analysis and prediction of diabetes. Thus, the project is operationally feasible.

Chapter 4

System Design and Architecture

We developed a diabetes prediction Classifier which takes various features data as its input and process it to predict diabetes.

4.1 USE CASE DIAGRAM

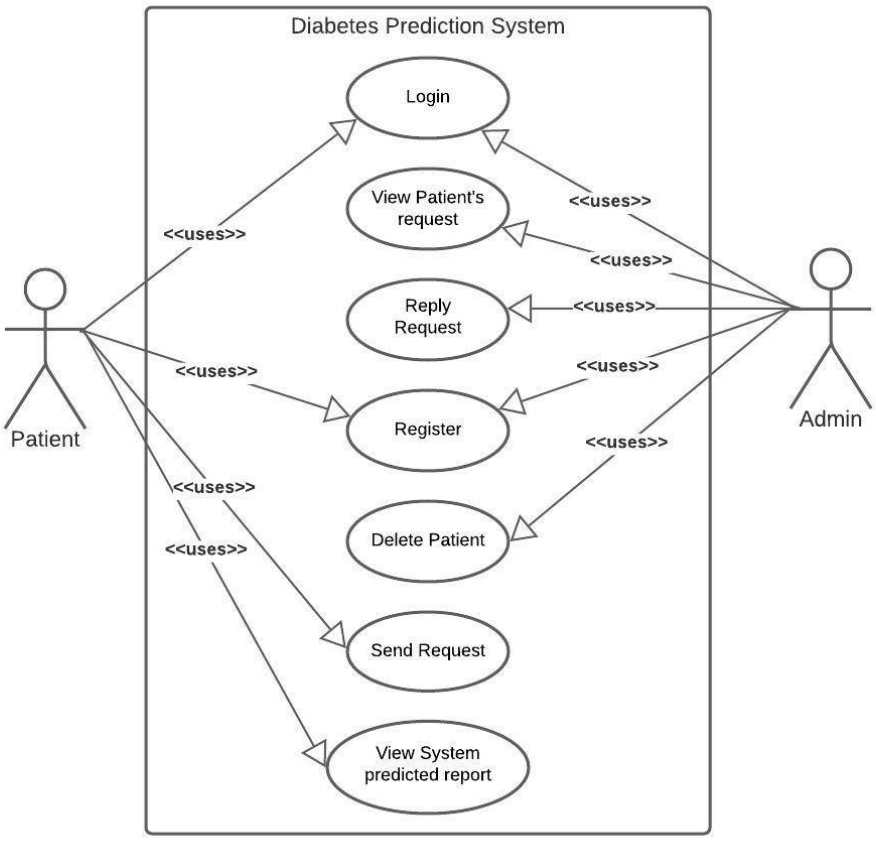


Figure 4.1: System Use Case Diagram

4.2 System Block Diagram

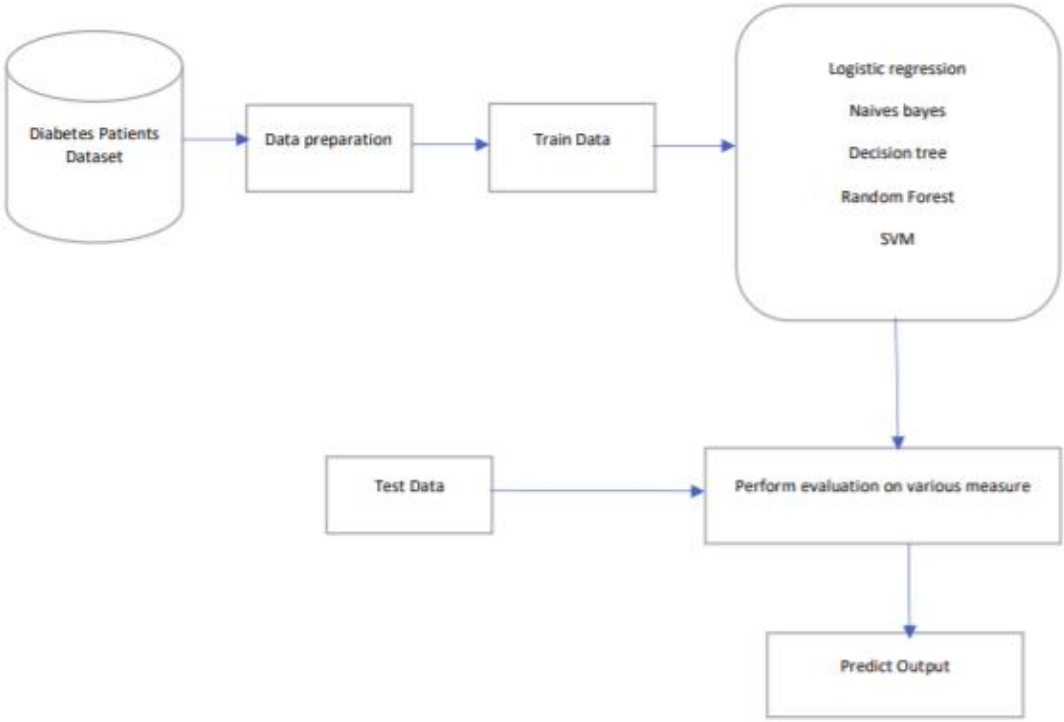


Figure 4.2: System Block Diagram

4.3 Sequence Diagram

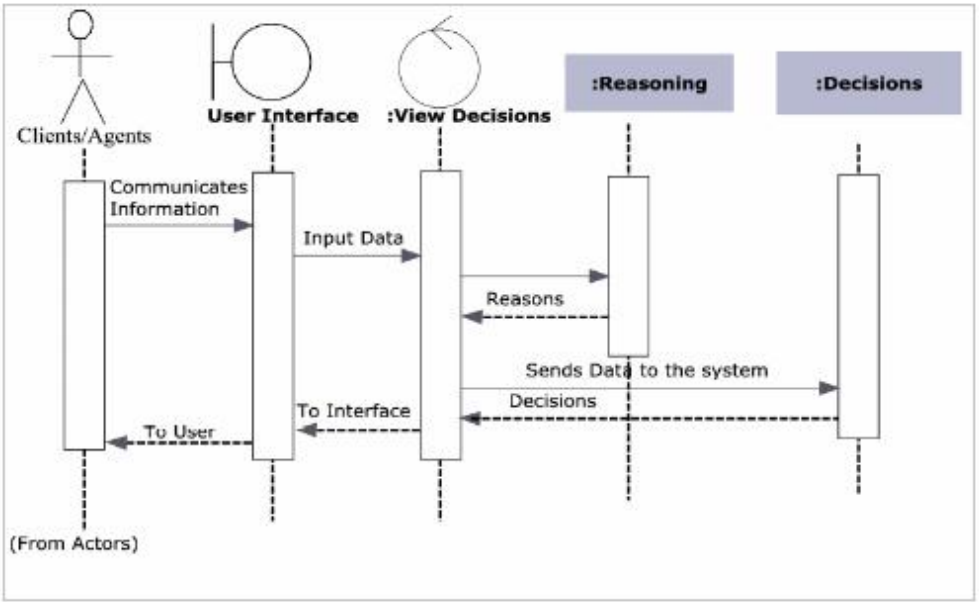


Figure 4.3: Sequence Diagram

Chapter 5

Methodology

5.1 SOFTWARE DEVELOPMENT APPROACH

Software prototyping is the activity of creating prototypes of software applications, i.e., incomplete versions of the software program being developed. Prototyping has several benefits: the software designer and implementer can get valuable feedback from the users early in the project. The purpose of a prototype is to allow users of the software to evaluate developers' proposals for the design of the eventual product by actually trying them out, rather than having to interpret and evaluate the design based on descriptions.

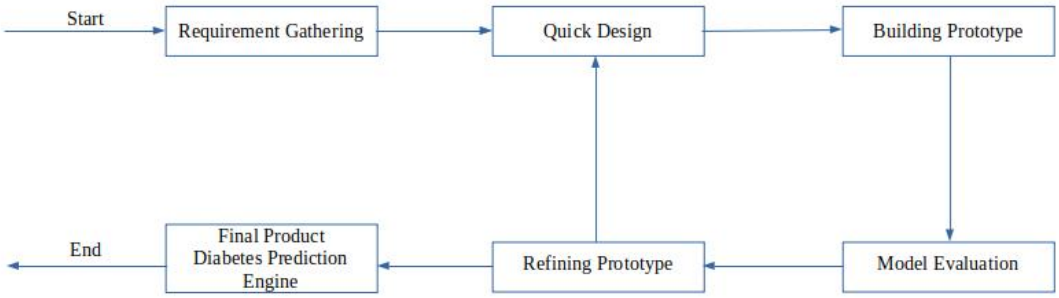


Figure 5.1: Prototype Model for Software Development

5.2 DATA COLLECTION

[4] This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.
The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on

5.3 DATA PREPARATION

[4] After the data has been collected, it was processed to make into correct format so that the neural networks can understand both the inputs and outputs.

5.3.1 StandardScaler and LabelEncoder

5.3.1.1 Standard Scaler

Standardize features by removing the mean and scaling to unit variance :
Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the set. Mean and standard deviation are then stored to be used on later data using the transform method,
Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data (e.g. Gaussian with 0 mean and unit variance).

5.3.1.2 Label Encoder

Encode labels with value between 0 and n_classes-1. Below we encode the data to feed properly to our algorithm Now, we can compute correlation matrix

5.3.2 X and Y

The patient records were represented as a data frame of features and a class label in the feature extraction step. Categorical features were encoded with numerical values for analysis.
We define X and y :
Def X and Y

$$X = data.drop('Outcome', 1)//features$$

$$Y = data['Outcome']//classlabel$$

5.3.3 Model Performance

To measure the performance of a model, we need several elements :
This part is essential

- Confusion matrix : also known as the error matrix, allows visualization of the performance of an algorithm :
    1. true positive (TP) : Diabetic correctly identified as diabetic
    2. true negative (TN) : Healthy correctly identified as healthy
    3. false positive (FP) : Healthy incorrectly identified as diabetic
    4. false negative (FN) : Diabetic incorrectly identified as healthy
- Metrics :
    1. Accuracy : (TP +TN) / (TP + TN + FP +FN)

2. Precision : TP / (TP + FP)
3. Recall : TP / (TP + FN)
4. F1 score : 2 x ((Precision x Recall) / (Precision + Recall))

5.3.4 Scores Table

Table 5.1: Scores Table

Classification	**Label Encoder**
Diabetes	1
UnDiagnosed Diabetes	0

5.4 TRAINING

After data has been prepared the data is fed into various models described in in this report. Various models' evaluation was carried out after training the model. The data will be fed into the training system as shown in the figure below;

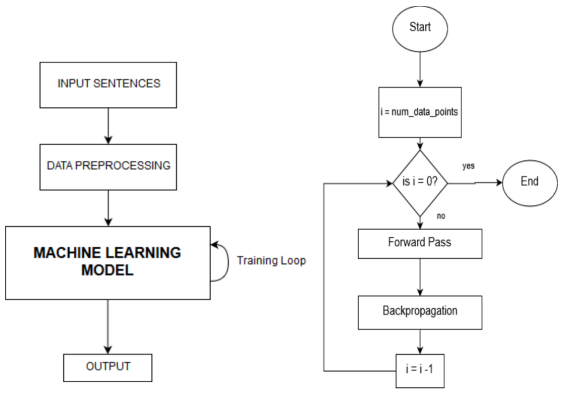


Figure 5.2: Basic Machine Learning Training Process

5.5 ALGORITHMS AND MODELS

As above mentioned, various algorithms and models were developed for building Diabetes Prediction System. Following description describes the various algorithms used in the system development:

5.5.1 Random Forest

The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. It is the most powerful and supervised algorithm than others. As its name suggests it creates the forest of datasets. This algorithm can be used as classifier as well as in regression. To build up multiple

decision trees, it uses algorithms like Information Gain, GIGI Index approach and other decision tree algorithms.

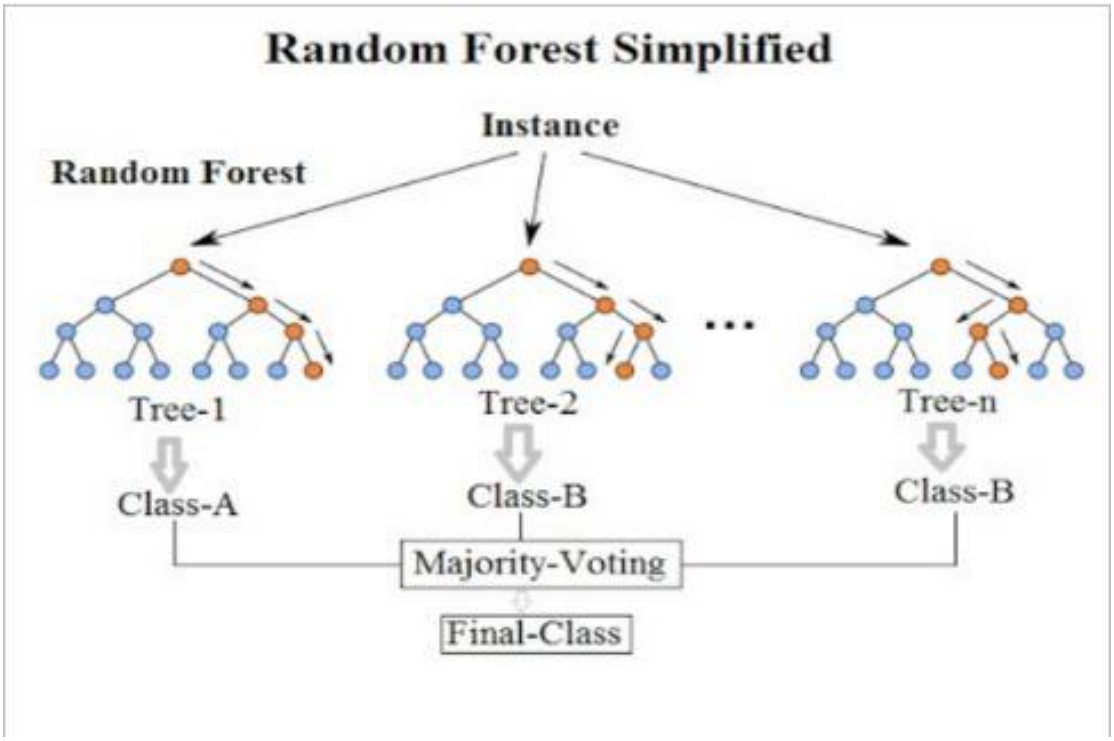


Figure 5.3: Simple Random Forest

To classify a new object based on attributes each tree gives a classification and the tree votes for that class. The forest chooses the classification having the most votes of all the trees in the forest. In case of regression, it takes the average of all the outputs by different trees. It identifies the medical diseases analyzing the patient's medical records and on its majority it clearly separates the person as diabetic or non-diabetic.
Random Forest handles the missing values and maintains the accuracy for missing data. As a difference from Decision tree it doesn't overfit the model and has the power to handle the large dataset with higher dimensionality.

5.5.2 Naive Bayes

[**?**] Naive Bayes Classifier: Naive Bayes is a classification technique with a notion which defines all features are independent and unrelated to each other. It defines that status of a specific feature in a class does not affect the status of another feature. Since it is based on conditional probability it is considered as a powerful algorithm employed for classification purpose. It works well for the data with imbalancing problems and missing values.
Naive Bayes is a machine learning classifier which employs the Bayes Theorem. Using Bayes theorem posterior probability P(C—X) can be calculated from P(C),P(X) and P(X—C).
P(C—X)=((P(X—C)*P(C))/(P(X))
P(X—C) = predictor class's probability.
P(C) = class C's probability being true.

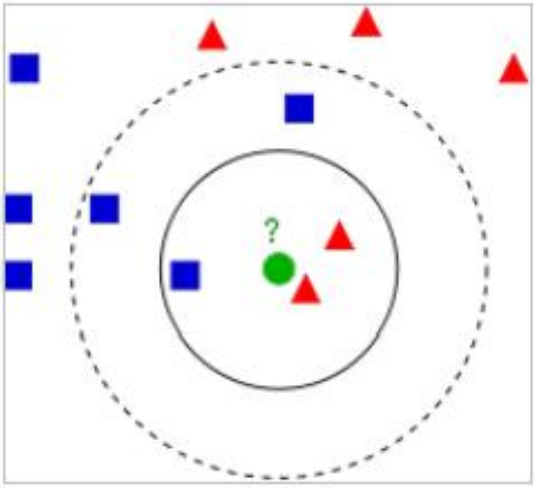
P(X) = predictor's prior probability.
Let us consider example.
C=yes/no
X= features (pregnancies=2, bloodpressure=75, BMI=33.66, insulin=9 etc)
Therefore, P(C—X) = (P(X1—C)*P(X2—C)* P(X3—C)* P(X4—C)* P(C))/(P(X1)* P(X2)* P(X3)*P(X4) Where, P(C—X) = target class's posterior probability .
P(X1/C)=(total no of pregnany of count =2 with predicted yes)/(total no of pregnancy of count =2)
P(X1)=(total no of pregnancy of count =2)/(sum of all pregnancies)
P(C)=(total no of (yes or no) in Outcomes)/(sum of total no of yes and no)
Simarly calculate the value of other features
Finally, the value is predict by comparing value of probability of yes or no.
A=(p(yes/X))/(p(yes/X)+p(no/X))
B=(p(no/X))/(p(yes/X)+p(no/X))
If A is greater than B then ,the patiet suffer from diabetes else no.
Generate a model using naive bayes classifier in the following steps:

- Create naive bayes classifier
- Fit the dataset on classifier
- Perform prediction

Generate a model using naive bayes classifier in the following steps:
Suppose there is no tuple for a risky loan in the dataset, in this scenario, the posterior probability will be zero, and the model is unable to make a prediction. This problem is known as Zero Probability because the occurrence of the particular class is zero.
The solution for such an issue is the Laplacian correction or Laplace Transformation. Laplacian correction is one of the smoothing techniques. Here, you can assume that the dataset is large enough that adding one row of each class will not make a difference in the estimated probability. This will overcome the issue of probability values to zero.
P(x)=(total no of X+1)/(total no of all data + total no of all data)

5.5.3 K-Nearest Neighbours(KNN) Algorithm

It is one of the simplest supervised machine learning algorithm mostly used for classification. This is a lazy learner algorithm because is doesn't learn a discriminate function from the training set. KNN algorithm is based on feature similarity that performs classification using KNN classifier. This algorithm classifies a data point based on how its neighbors are classified. Based on the similarity measure, it classifies all the new cases and stores all available cases. K in KNN is a parameter that refers to the number of nearest neighbors to include in the majority voting process i.e. diabetic or non-diabetic one. Choosing the right value of k is a process called parameter tuning and is important for better accuracy of diabetes prediction.
Example of KNN classification:



The test sample (green dot) should be classified either to blue squares or to red triangles. If k = 3 (solid line circle) it is assigned to the red triangles because there are 2 triangles (majority) and only 1 square inside the inner circle. If k = 5 (dashed line circle) it is assigned to the blue squares (3 squares vs. 2 triangles inside the outer circle) due to majority of blue squares.
Hence for choosing the right value of k, the following two points must be considered:

- Sqrt(n), where n is the total number of data points.
- Odd value of k is selected to avoid confusion between two classes i.e. diabetic or non-diabetic.

This algorithm works using the Euclidian Distance formula i.e. it calculates the Euclidian distance of unknown datapoint from all the points in the dataset. The Euclidian distance formula is given below.
Mathematically,

$$Distance(d) = sqrt((x-a)^2 + (y-b)^2)$$

,Where (a,b) and (x,y) are the respective coordinates of two neighboring datapoints.
KNN algorithm is used when the data is labeled i.e. specific and noise free. Also the KNN is used when the dataset is small as KNN is a lazy learner algorithm. For large dataset Support Vector Algorithm (SVM) is used to classify the dataset.

KNN algorithm example
Suppose we have a input data with X= features (pregnancies(Pg)=2, bloodpressure(Bg)=75, BMI(BMg)=33.66, insulin(Ig)=9 etc). And we have to calculated distance between each row with respective column of featues using distance formula.
D1=(Pg-P1)+(Bg-B1)+(BMg-BM1)+..... outcome=1
D2=(Pg-P2)+(Bg-B2)+(BMg-BM2)+.....outcome=0
.....
Then we order the distance in ascending order and find it first K nearest neighbourhood. the value is predicted on the basis of majority of Outcomes of neighbourhood.

5.5.4 Decision Tree Algorithm

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. Decision Tree performs a test on the feature or attributes. The root node is called decision node to which the dataset is given. Decision Tree can be used as a classifier as well as regression. Decision trees perform classification without requiring much computation so that it is used to classify a person a diabetic or non-diabetic. It is able to handle both continuous and categorical variables. This algorithm provides a clear indication of which fields are most important for prediction or classification.

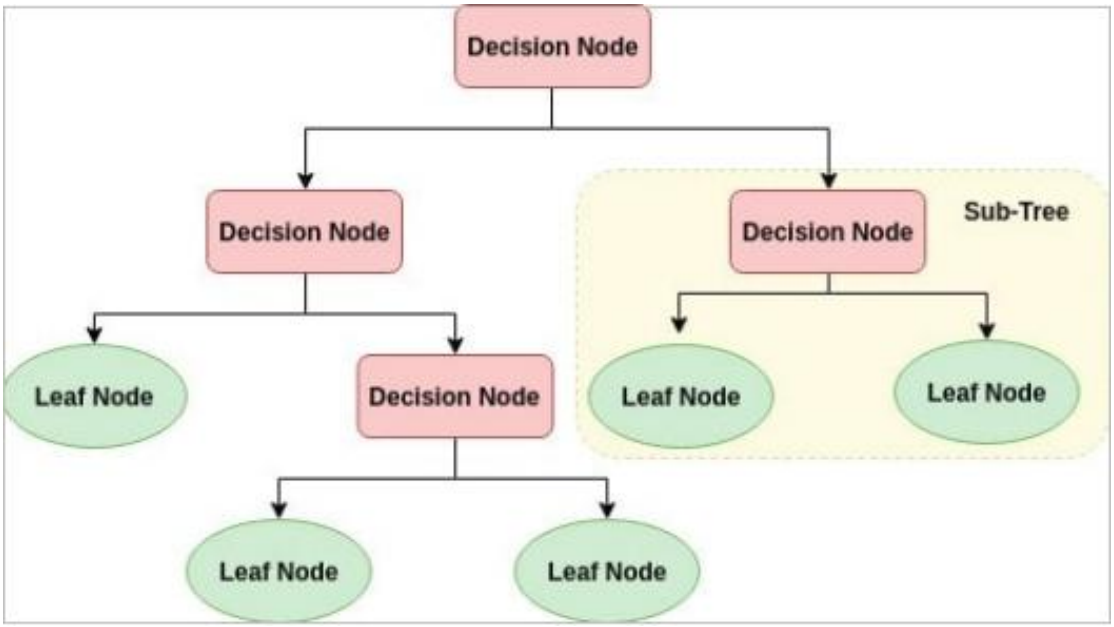


Figure 5.4: Decision Tree

Decision tree builds classification or regression models in the form of a tree structure. It breaks set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision node has two or more branches. Leaf node represents a classification or decision that gives the final report of a patient to be diabetic or not. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.

Decision Tree Calculation

Entropy Calculation

Entropy=-(P/(P+N))*log(P/(P+N)) -(N/(P+N))*log(N/(P+N)) where P =positive and N=negative

Calculate average information:

I(attribute)=sum (Pi+Ni)/(P+N) entropy(A)

Calculate information gain:

gain=entropy(s)- I(attribute)

Algorithm

1. compute the entropy for Dataset
2. for every attribute/feature
3. pick the highest gain attribute
4. Repeat until we get the desired result

Chapter 6

Expected Outcomes

One of the important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of disease like diabetes. In database, there are 762 data with 8 features of diabetes are stored. The system for labelling diabetes has been completed. The input data from the database is successfully retrieved, then preprocessed and sent to the model for training. Various models for diabetes classification was trained and the result of the training are showing in the table below:

Table 6.1: Accuracy of different algorithms

Algorithm	Accuracy Score
Logistic Regression	77.82%
Random Forest	75.57%
Navies Baise	76.22%
SVM	78.84%
KNN	75.97%
Decision Tree	77.22%

The best model was found to be SVM model with testing accuracy 78.84% while other model has less accuracy. We can also increase accuracy by changing various variable, due to lo number of datapoints for added complexity in the model, they started overfitting. To solve the problem of overfitting of the model we tried different measures such as dropout, regularization and increasing data. But increasing data was not an option for us due to limited time.

Bibliography

[1] K. G. M. M. Alberti and P. Z. Zimmet, "Definition, diagnosis and classification of diabetes mellitus and its complications. part 1: diagnosis and classification of diabetes mellitus. provisional report of a who consultation," *Diabetic medicine*, vol. 15, no. 7, pp. 539–553, 1998.

[2] H. Temurtas, N. Yumuşak, and F. Temurtas, "A comparative study on diabetes disease diagnosis using neural networks," *Expert Syst. Appl.*, vol. 36, pp. 8610–8615, 05 2009.

[3] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia computer science*, vol. 132, pp. 1578–1585, 2018.

[4] J. Sun, "The study of pima indian diabetes," 10 2016.